

Analyse génétique de la composition protéique & des aptitudes fromagères du lait de vache prédites à partir des spectres moyen infrarouge

Marie-Pierre Sanchez

▶ **To cite this version:**

Marie-Pierre Sanchez. Analyse génétique de la composition protéique & des aptitudes fromagères du lait de vache prédites à partir des spectres moyen infrarouge. Génétique animale. Université Paris Saclay (COmUE), 2019. Français. ⟨NNT : 2019SACLA008⟩. ⟨tel-02434974⟩

HAL Id: tel-02434974

https://pastel.hal.science/tel-02434974v1

Submitted on 10 Jan 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

NNT : 2019SACLA008

Analyse génétique de la composition protéique & des aptitudes fromagères du lait de vache prédites à partir des spectres moyen infrarouge

Thèse de doctorat de l'Université Paris-Saclay
préparée à AgroParisTech (Institut des Sciences et Industries du Vivant et de l'Environnement)

École doctorale n°581 Agriculture, alimentation, biologie, environnement et santé (ABIES)

Spécialité de doctorat: Génétique Animale

Thèse présentée et soutenue à Paris, le 15 mai 2019, par

Marie-Pierre Sanchez

Composition du Jury :

Xavier Rognon Professeur à AgroParisTech, Paris	Président
Nicolas Gengler Professeur à l'Université de Liège, Gembloux	Rapporteur
Sandrine Lagarrigue Professeure à Agrocampus-Ouest, Rennes	Rapporteur
Gwenola Tosser-Klopp Directrice de recherche à l'INRA, Toulouse	Examinateur
Mickaël Brochard Responsable R&D à Umotest, Ceyzériat	Examinateur
Didier Boichard Directeur de Recherche à l'INRA, Jouy-en-Josas	Directeur de thèse

Thèse de doctorat

A Lou et Fleur,

« Choisissez un travail que vous aimez

et vous n'aurez pas à travailler un seul jour de votre vie. »

Confucius

Remerciements

Ces trois années ont été riches et elles vont assurément marquer un tournant dans ma carrière. J'ai beaucoup appris, y compris sur moi-même et cela n'aurait pas été possible sans l'aide, le soutien et la bienveillance de mon entourage.

Tout d'abord, merci à toi Didier Boichard, pour l'accueil que tu m'as réservé dans ton équipe, la confiance que tu m'as accordée et tes encouragements vers cette thèse. Sans tout ça, je n'aurais jamais osé me lancer dans cette aventure. Merci pour la qualité (et la quantité) de ton implication durant ces trois ans, je mesure toute la chance que j'ai eue ! Plus qu'un directeur de thèse, tu as été un vrai mentor pour moi.
Tu as toute ma reconnaissance et toute mon admiration.

Merci à Claire Rogel-Gaillard, directrice de l'unité GABI, qui a accepté que je me lance dans une thèse en plus de mes fonctions d'ingénieur d'études.

J'adresse mes remerciements aux personnes qui me font l'honneur d'être membres du jury :
- Nicolas Gengler, professeur à l'Université de Liège à Gembloux et Sandrine Lagarrigue, professeure à AgroCampus à Rennes, en tant que rapporteurs ;
- Gwenola Tosser-Klopp, directrice de recherche à l'INRA à Toulouse et Mickaël Brochard, responsable R&D à Umotest à Ceyzériat, en tant qu'examinateurs ;
- Xavier Rognon, professeur à AgroParisTech, en tant que président.

Merci à tous les membres de mon comité de thèse pour leurs conseils et leur bienveillance : Pascal Croiseau, Agnès Delacroix-Buchet, Cécile Laithier, Rachel Rupp et Etienne Verrier.

Ce travail n'aurait pu se faire sans les projets *PhénoFinlait* et *From'MIR* qui ont chacun impliqué un très grand nombre de personnes. Pour le volet « protéines » de *PhénoFinlait*, je remercie tout particulièrement Guy Miranda, Marion Ferrand et Patrice Martin. Merci également à tous les membres du comité de pilotage *From'MIR* (ses réunions ont été pour moi l'occasion de découvrir quelques facettes de la Franche-Comté) et en particulier, Valérie Wolf et Cécile Laithier qui ont su porter le projet avec beaucoup d'enthousiasme et qui ont toujours été très réactives à mes demandes ; Daniel Pourchet, Philippe Grosperrin et Nicolas Gaudillière pour leur implication dans ce projet ; Agnès Delacroix-Buchet pour les réponses immédiates à toutes mes questions fromagères, pour ses conseils de lecture et pour sa veille bibliographique ; Mohammed El Jabri pour notre collaboration sur les équations MIR et enfin Stéphanie Minéry (trop vite partie vers d'autres aventures) et Mickaël Brochard, animateurs successifs de l'action « génétique », pour toute la confiance que vous m'avez accordée et pour votre extrême gentillesse.

Merci à Rassel Rahman et à Maurane Beaumont, stagiaires que j'ai encadrés durant cette thèse et qui ont respectivement travaillé sur les variants protéiques et sur la sélection des critères fromagers. Et merci aux co-encadrants Sébastien Fritz, Mickaël Brochard et Didier Boichard.

Merci à l'ensemble de l'équipe G2B pour son accueil chaleureux lors de mon arrivée dans l'équipe, quelques années avant le début de cette thèse (2012). Merci à tous ceux qui m'ont encouragée dans cette thèse et qui m'ont apporté leur aide et leur soutien d'une manière ou

d'une autre : Mekki Boussaha pour son appui sur les séquences, Sébastien Taussat pour son travail sur les imputations, Alexis Michenet pour ses scripts R bien avant Think R, Yuliaxis Ramayo-Caldas pour son regard averti sur l'AWM, Thierry Tribout pour son aide précieuse dans l'évaluation Single Step, Vincent Ducrocq pour la relecture de certains des articles de cette thèse… et les autres, nombreux, que je ne peux pas tous citer.

Et parce qu'on ne côtoie pas les collègues que pour le travail, je pense … à tous les moments agréables passés avec les sportifs, copains et copines de gym, Zumba et course, ces séances ont été un vrai exutoire pour moi, surtout durant les longues semaines de rédaction … et à tous les autres bons moments, ils sont forcément nombreux après tout ce temps.

J'ai une pensée toute particulière pour les collègues et amis de longue date, merci de toutes vos attentions, petites et grandes, ça compte beaucoup. Mes voisins, Seb et Pascal, j'ai été très discrète ces derniers mois mais je vais me rattraper. Aurélie, Gilles, Sophie Mo, Stéphanie, Thierry, merci pour votre gentillesse sans bornes. Sophie, Stéphanie, toujours partante pour faire les pipelettes à la Pipelote ou ailleurs ! Thierry, tu es très présent depuis très longtemps, merci pour tout !

Et pour finir, merci Hugues, Lou et Fleur ! En acceptant les longues journées passées à Jouy et le travail à la maison, vous avez été d'un grand soutien.
Merci à toi Hugues d'être si présent pour les filles et d'accepter sans réserve que je m'investisse autant dans mon travail.
Merci à vous Lou et Fleur de faire notre bonheur. Vous êtes si faciles du haut de vos 15 et 12 ans, vous êtes et serez toujours ma plus grande fierté.

Table des matières

Table des matières 7

Liste des tableaux 12

Liste des figures 14

Liste des abréviations 16

Introduction générale 21

1. Contexte bibliographique 25

1.1. Le lait 25
1.1.1. Production de lait 25
1.1.2. Composition du lait 26
1.1.2.1. Glucides 26
1.1.2.2. Lipides 26
1.1.2.3. Protéines 28
1.1.2.4. Minéraux et autres constituants 30
1.1.2.5. Cellules somatiques et flore microbienne 30
1.2. La transformation du lait en fromage 32
1.2.1. La production de fromage en France 32
1.2.2. Zoom sur la région Franche-Comté 34
1.2.3. Les technologies fromagères 34
1.2.4. Fromages à pâte pressée cuite au lait cru, type Comté 35
1.2.4.1. Préparation du lait 36
1.2.4.2. Coagulation 36
1.2.4.3. Egouttage 37
1.2.4.4. Affinage 38
1.2.5. Critères de fromageabilité du lait et méthodes de mesure 38
1.2.5.1. Paramètres d'acidification 38
1.2.5.2. Paramètres de coagulation 39
1.2.5.3. Rendements de laboratoire 41
1.3. La spectrométrie moyen infrarouge 43
1.3.1. Principe de la spectrométrie MIR 43
1.3.2. Du spectre MIR au phénotype 45
1.3.2.1. Standardisation des spectres MIR 45
1.3.2.2. Population de calibration 46
1.3.2.3. Equations de prédiction 46
1.3.2.4. Précision des équations de prédiction 46
1.3.3. Les spectres MIR et la production laitière bovine 47

1.4. Programmes de recherche *PhénoFinlait* et *From'MIR* 49
1.5. Facteurs de variation de la composition et de la fromageabilité du lait 52
1.6. Méthodes d'analyse génétique 54
1.6.1. Modélisation de la performance 54
1.6.2. Paramètres génétiques 55
1.6.2.1. Définitions 55
1.6.2.2. Modélisation des performances 56
1.6.2.3. Méthodes d'estimation 58
1.6.3. Gènes et variants 59
1.6.3.1. Régions du génome avec des effets sur les caractères quantitatifs (QTL) 59
1.6.3.2. Puces bovines et génotypages 60
1.6.3.3. Projet 1000 génomes bovins et imputations sur la séquence 60
1.6.3.4. Détection des QTL 62
1.6.3.5. Identification des gènes et variants candidats 62
1.6.4. Sélection 63
1.6.4.1. Objectif de sélection en race Montbéliarde 63
1.6.4.2. Méthodes d'évaluation génétique & génomique 64

2. Du spectre MIR au phénotype 71

2.1. Prédiction de la composition protéique du lait dans le projet *PhénoFinlait* 71
2.2. Prédiction de la fromageabilité du lait dans le projet *From'MIR* 73
2.2.1. Population de calibration : échantillons et analyses de référence 73
2.2.2. Spectres MIR 75
2.2.3. Equations de prédiction : test des méthodes bayésiennes 77
2.2.3.1. Méthodes bayésiennes testées 77
2.2.3.2. Résultats : comparaison des précisions des équations 79
2.2.4. Précision des paramètres de fromageabilité du lait dans le projet *From'MIR* 81
2.2.5. Spectres et prédictions MIR dans le projet *From'MIR* 83
2.2.5.1. Spectres MIR collectés 83
2.2.5.2. Traitement des spectres MIR collectés 83
2.2.5.3. Prédictions de la composition fine du lait et des aptitudes fromagères 84
2.3. Bilan du chapitre 2 86

3. Paramètres génétiques 89

3.1. Paramètres génétiques de la composition protéique du lait 91
Journal of Dairy Science 2017. 100:6371-6375 91
3.2. Paramètres génétiques de la fromageabilité et de la composition du lait 99
Journal of Dairy Science 2018. 101:10048–10061 99
3.3. Héritabilités au cours de la première lactation 115
3.4. Corrélations génétiques entre les trois premières lactations 121
3.5. Corrélations génétiques avec les caractères en sélection 123

3.6. Bilan du chapitre 3 125

4. Gènes et variants candidats 131

4.1. Variants génétiques des lactoprotéines en races Montbéliarde, Normande et Holstein 132
4.1.1 Caractérisation des variants protéiques à partir des génotypes aux SNP 132
4.1.2. Inventaire et fréquences des variants protéiques dans les trois races 134
4.1.3. Evolution des fréquences des variants protéiques dans les trois races 135
4.1.4. Effets des variants protéiques sur les caractères de production laitière 136
4.2. Détection de QTL pour la composition protéique (puce 50K) 139
Journal of Dairy Science 2016. 99:8203–8215. 139
4.3. GWAS pour la composition protéique à l'échelle de la séquence 155
Genetics Selection Evolution 2017. 49:68. 155
4.4. Confirmation des mutations candidates 173
Journal of Dairy Science 2018. 101:10076–10081 173
4.5. GWAS et réseaux de gènes pour la composition et la fromageabilité du lait 181
Genetics Selection Evolution 2019. 51:34. 181
4.6. Bilan du chapitre 4 203

5. Vers une sélection génomique des aptitudes fromagères 207

5.1. Estimation de la précision d'une évaluation génomique 207
5.1.1. Matériel et méthodes 207
5.1.1.1. Données et caractères analysés 207
5.1.1.2. Constitution des populations d'apprentissage et de validation 208
5.1.1.3. Méthode et logiciel d'évaluation génomique utilisés 210
5.1.1.4. Modèles testés 210
5.1.1.5. Densités et pondérations des marqueurs 211
5.1.1.6. Estimation de la précision des valeurs génomiques 213
5.1.2. Résultats & discussion 213
5.1.2.1. Comparaison des modèles 213
5.1.2.2. Comparaison des densités et pondérations des SNP 214
5.2. Estimation du progrès génétique réalisé sur les caractères fromagers 217
5.2.1. Matériel et méthodes 217
5.2.1.1. Modèle et estimation des valeurs génétiques 217
5.2.1.2. Animaux et génotypages 218
5.2.2. Résultats 218
5.3. Etudes des potentialités de sélection des caractères fromagers 220
5.3.1. Concertation avec un groupe d'experts 221
5.3.2. Simulation de scénarios de sélection 221
5.3.2.1. Scénarios de sélection 222
5.3.2.2. Estimation des réponses à la sélection 223

5.3.2.3. Résultats 223
5.4. Bilan du chapitre 5 226

6. Discussions, conclusions et perspectives 231

6.1. Apports de *PhénoFinlait* et *From'MIR* 231
6.2. Qualité des fromages produits à partir d'un lait plus fromageable 233
6.3. Les suites de *From'MIR* 234
6.3.1. Au niveau régional 235
6.3.2. Au niveau national 235
6.3.3. Et au-delà 235
6.4. Quel impact économique aurait un lait plus fromageable ? 236
6.5. Sélection sur la fromageabilité dans les autres pays 236
6.6. Les spectres MIR du lait 237
6.6.1. Un outil de phénotypage fin pour la sélection 237
6.6.2. La génétique du spectre MIR 238
6.6.3. Remarques sur les prédictions MIR 240
6.7. Identification des mutations causales 241
6.7.1. Vers une meilleure connaissance du génome bovin et de son fonctionnement 241
6.7.2. Analyses d'expression pangénomiques et analyses fonctionnelles ciblées 243

Quelques remarques finales… 246

Bibliographie additionnelle * 249

Liste des publications entre 2016 et 2019 265

Résultats du travail de thèse 265
Articles scientifiques 265
Communications à des congrès / séminaires 266
Encadrement de stages 267
Autres résultats sur la période 2016-2019 268
Articles scientifiques 268
Communications à des congrès / séminaires 268
Encadrement de stages 270

Annexes 273

Annexe 1 - Distributions des caractères prédits par spectrométrie MIR dans le projet *From'MIR* 273
Critères fromagers 273
Composition protéique 274
Composition en acides gras 275

Composition en minéraux, citrate et lactose 276
Annexe 2 - Courbes de lactation et héritabilités estimées par régression aléatoire 277
Composition protéique 277
Composition en acides gras 278
Composition en minéraux, citrate et lactose 279
Annexe 3 – Variants génétiques des lactoprotéines 280
Annexe 4 – Restitution et diffusion des résultats *From'MIR* 281
Programmes des journées de restitution 281
Quelques exemples d'articles, fiche technique et Newsletter du projet From'MIR 282
L'éleveur laitier - n°272 – juillet-août 2018 282
La Terre de chez nous – 29 juin 2018 283
Réussir lait – 12 juin 2018 284
La revue laitière française – n°782 – Juin 2018 285
Fiche technique et Newsletter n°5 From'MIR 286

Résumé 298

Abstract 298

Liste des tableaux

Tableau 1.1. Proportions relatives des principaux acides gras de la matière grasse du lait de vache 27

Tableau 1.2. Proportions relatives des principales protéines du lait de vache 28

Tableau 1.3. Composition des micelles de caséines 29

Tableau 1.4. Nombre de spectres MIR, de vaches avec spectres et avec génotypes dans les projets PhénoFinlait et From'MIR 51

Tableau 2.1. Teneurs en protéines (% lait) pour les trois races bovines du programme PhénoFinlait : moyennes ± écart-types et qualité de la prédiction MIR 72

Tableau 2.2. Teneurs en protéines (% de protéines) pour les trois races bovines du programme PhénoFinlait : moyennes ± écart-types 73

Tableau 2.3. Caractéristiques des protocoles appliqués pour les deux technologies fromagères 74

Tableau 2.4. Noms et descriptions des paramètres fromagers mesurés pour les deux types de technologie pâte pressée cuite (PCC) et pâte molle (SC) 75

Tableau 2.5. Précisions des équations de prédiction (R^2_{CV} = R^2 calculé par validation croisée) développées avec les méthodes bayésiennes et la meilleure méthode PLS 79

Tableau 2.6. Performances des équations de prédiction développées sur 416 laits (246 individuels + 100 troupeaux + 70 cuves) pour les 24 paramètres fromagers 82

Tableau 2.7. Prédictions des caractères fromagers et de la composition en protéines à partir des spectres MIR des vaches Montbéliarde du projet From'MIR 84

Tableau 2.8. Prédictions de la composition en acides gras, minéraux et citrate et lactose à partir des spectres MIR des vaches Montbéliarde du projet From'MIR 85

Tableau 2.9. Matrice des corrélations entre les neuf critères fromagers les mieux prédits 86

Tableau 3.1. Héritabilités (diagonale) et corrélations génétiques (hors diagonale) des caractères de composition et de fromageabilité du lait mesurés sur les trois premières lactations (L1, L2 et L3) [couleur bleue d'autant plus foncée que la corrélation est forte] 122

Tableau 3.2. Description des caractères en sélection 123

Tableau 3.3. Corrélations génétiques entre les critères de fromageabilité du lait en première lactation et sept caractères en sélection 124

Tableau 4.1. Gènes des lactoprotéines avec SNP non synonymes et polymorphes présents sur la puce EuroG10K et noms des variants protéiques associés 132

Tableaux & Figures

Tableau 4.2. Effets comparés de certains variants protéiques sur les caractères de production laitière 137

Tableau 5.1. Description des modèles utilisés pour l'évaluation génomique 211

Tableau 5.2. SNP sélectionnés à partir des résultats GWAS et % de variance génétique associée à chaque caractère dans le scénario QTL 212

Liste des figures

Figure 1.1. Les deux types de modèles décrivant la structure de la caséine de micelle 29

Figure 1.2. Utilisation du lait pour la fabrication des produits laitiers en France en 2017 32

Figure 1.3. Proportions de fromages produits en France en 2017 par catégorie de technologie 33

Figure 1.4. Processus de transformation des fromages frais, à pâte molle, à pâte pressée, à pâte persillée et fondus 35

Figure 1.7. Diagramme et paramètres de coagulation obtenus avec le Formoptic 41

Figure 1.9. Principe de la spectrométrie 44

Figure 1.10. Effet de la « Piece Direct Standardisation », d'après Grelet et al. (2015) 45

Figure 1.11. Partenaires des projets PhénoFinlait (a) et From'MIR (b) 49

Figure 1.12. Courbes de lactation, d'après Legarto et al. (2014), projet PhénoFinlait 52

Figure 1.13. Effets du stade de lactation sur le rendement en extrait sec (%)* 53

Figure 1.15. Précision théorique (R) des valeurs génomiques (GEBV) en fonction de la taille de la population de référence et de l'héritabilité du caractère 67

Figure 2.1. Spectres MIR (absorbances pour 446 longueurs d'onde) du lait de deux individus 76

Figure 2.2. Matrice des corrélations des absorbances calculées sur les 246 laits individuels. 76

Figure 2.3. Résultats d'une analyse en composantes principales (ACP) sur les spectres MIR (absorbances / 446 longueurs d'onde) des 246 laits individuels 77

Figure 2.4. R^2 calculé par validation croisée (R^2_{CV}) pour les équations de prédiction des 24 paramètres fromagers en fonction de la méthode utilisée 80

Figure 3.1. Variations du phénotype (en bleu) et héritabilités (en rouge) entre 7 et 350 jours de lactation pour les neuf paramètres fromagers, le taux protéique (TP), le taux butyreux (TB) et le calcium 116

Figure 3.2. Variances génétique, de l'environnement permanent et résiduelle estimées entre 7 et 350 jours de lactation pour les neuf paramètres fromagers, les taux protéique (TP), butyreux (TB) et de calcium 118

Figure 3.3. Proportions relatives de la variabilité génétique expliquée par les trois premières valeurs propres (VP) de la matrice des variances-covariances entre jours de lactation 119

Figure 3.4. Trajectoires de corrélations génétiques entre 7 et 350 jours de lactations pour le rendement en extrait sec (CY_{DM}) mesuré à 7, 92, 178, 264 et 350 jours 120

Tableaux & Figures

Figure 4.1. Arbres de décision pour la caractérisation des variants protéiques dans les quatre gènes CSN1S1, CSN2, CSN3 et PAEP (pour le gène CSN2, avec les SNP disponibles, il n'a pas été possible de distinguer les variants A1, C et F). 133

Figure 4.2. Fréquence des variants protéiques dans les trois races Holstein (HOL), Montbéliarde (MON) et Normande (NOR) 134

Figure 4.3. Evolution des fréquences des variants protéiques par année de naissance des taureaux de race Montbéliarde (MON), Normande (NOR) et Holstein (HOL) 136

Figure 5.1. Constitution des populations d'apprentissage (APP) et de validation (VAL) à partir des 19 564 vaches From'MIR Umotest phénotypées et génotypées 209

Figure 5.2. CD estimés dans la population de validation pour les 4 modèles 213

Figure 5.3. CD estimés dans la population de validation pour le modèle CTL3 et les 3 scénarios 214

Figure 5.4. Pentes de la régression des performances corrigées (Y_{CORR}) sur les valeurs génétiques (VGE) pour le modèle CTL3 et les 3 scénarios 215

Figure 5.5. Estimation des effets des SNP dans la région du gène DGAT1 pour le rendement en extrait sec CY_{DM} et les 3 scénarios 50K, 50K+ et QTL 216

Figure 5.6. Nombre de taureaux et de vaches de race Montbéliarde (Umotest) génotypés par année de naissance 218

Figure 5.7. Courbes de l'évolution génétique réalisée par année de naissance des taureaux nés entre 2005 et 2018 et des vaches nées entre 2008 et 2018 pour les 11 caractères fromagers 219

Figure 5.8. Différences entre les valeurs génétiques estimées entre 2018 et 2005 pour les taureaux et entre 2018 et 2008 pour les vaches pour les 11 caractères fromagers (en écart-type génétique du caractère) 220

Figure 5.9. Réponses à la sélection sur ISU, ISU-COMP et ISU-FROM estimées sur les index fromagers et les index des autres caractères (en écart-type génétique) 224

Figure 6.1. Comparaison des approches indirecte (a) et directe (b) pour estimer les valeurs génétiques (EBV) à partir des spectres MIR, d'après Bonfatti et al. (2017b) 239

Liste des abréviations

50K : puce à ADN 50K (54 609 SNP)

a : fermeté à une fois le temps de prise

a2 : fermeté à deux fois le temps de prise

ACP : analyse en composantes principales

ADN : acide désoxyribonucléique

AG : acides gras

ANR : agence nationale de la recherche

AOP : appellation d'origine contrôlée

APP : population d'apprentissage

ARN : acide ribonucléique

aseQTL : *allele specific expression QTL*

ASTRE : approches sociales et travail en élevage

AWM : *association weight matrix*

BLUP : *best linear unbiased prediction*

C14:0 : acide myristique

C16:0 : acide palmitique

C18:0 : acide stéarique

C18:1 : acide oléique

C4-C10 : somme des acides gras C4 à C10

C4-C12 : somme des acides gras C4 à C12

Ca : calcium

CAL : population de calibration

CAS : index génétique des caséines

CASDAR : compte d'affectation spéciale pour le développement agricole et rural

CCS : comptage des cellules somatiques

CD : coefficient de détermination

CNAOL : conseil national des appellations d'origine laitière

CNE : confédération nationale de l'élevage

CNIEL : centre national interprofessionnel de l'économie laitière

CNV : *copy number variation*

CV_{gen} : coefficient de variation génétique

CY : *cheese yield*

CY_{DM} : rendement en extrait sec

$CY_{FAT\text{-}PROT}$: rendement en matière sèche utile

CY_{FRESH} : rendement frais

DL : déséquilibre de liaison

DM : développement de la mamelle

DYD : *daughter yield deviation*

ECEL : entreprise de conseil en élevage

ENILBio : école nationale d'industrie laitière et des biotechnologies

eQTL : *expression QTL*

ES : entreprise de sélection

ET_G : écart-type génétique

EuroG10K : puce à ADN du consortium Eurogenomics

FER : index de fertilité vache

FERG : index de fertilité génisse

GBLUP : *genomic best linear unbiased prediction*

GEBV : *genomic estimation of breeding value*

GO : *gene ontology*

GWAS : *genome wide association study*

h^2 : héritabilité

HD : puce à ADN haute densité (777 962 SNP)

HOL : Holstein

HSSGBLUP : *hybrid single step genomic best linear unbiased prediction*

IGP : indication géographique protégée

INAO : institut national de l'origine et de la qualité

Indel : insertion - délétion

INEL : index économique laitier

INRA : institut national de la recherche agronomique

ISU : index de synthèse racial

IVIA1 : intervalle vêlage – première insémination artificielle

K : potassium

K10/RCT : inverse de la vitesse de coagulation

L1 : première lactation

L2 : deuxième lactation

L3 : troisième lactation

LC-MS : chromatographie liquide couplée à la spectrométrie de masse

LD : puce à ADN *low density* (6909 marqueurs)

LDLA : *linkage disequilibrium and linkage analysis*

LGF : longévité fonctionnelle

LO : longueur d'onde

LV : nombre de variables latentes

MACL : mammites cliniques

MAF : *minor allele frequency*

Mg : magnésium

MG : matière grasse

MIR : moyen infrarouge

MO : index de synthèse « morphologie »

MON : Montbéliarde

MP : matière protéique

MSU : matière sèche utile

MUNSAT : *monounsaturated fatty acids*

Na : sodium

NOR : Normande

OS : organisme de sélection

P : phosphore

pb : paire de bases

PCC : *pressed cooked cheese* (pâte pressée cuite)

pH_0 : pH du lait à l'ensemencement

PLS : *partial least squared*

PUNSAT : *polyunsaturated fatty acids*

QTL : *quantitative trait locus*

R^2_{cal} : coefficient de détermination calculé dans la population de calibration

R^2_{cv} : coefficient de détermination calculé par validation croisée

R^2_{val} : coefficient de détermination calculé dans la population de validation

r_a : corrélation génétique

RCT : *rennet coagulation time*

REML : *restricted maximum likelihood*

REPRO : index de synthèse « fertilité »

RFPLS : *random forest partial least squared*

RMSE : *root mean squared error*

RPD : *residual prediction deviation*

RR : *random regression*

RZE : règlement zootechnique européen

SAT : *saturated fatty acids*

SC : *soft cheese* (pâte molle)

SCS : score des cellules somatiques

SNP : *single nucleotide polymorphism*

sQTL : *splicing expression QTL*

SS : *single step*

SS-GBLUP : *single step genomic best linear unbiased prediction*

STMA : index de synthèse « santé mamelle »

SYNTLAIT : index de synthèse laitier

TB : taux butyreux

TP : taux protéique

UE : union européenne

UFC : unité formant colonie

UNSAT : *unsaturated fatty acids*

URFAC : union régionale des fromages d'appellation d'origine Comtoise

UVEPLS : *uninformative variable elimination partial least squared*

VAL : population de validation

VEP : *variant effect predictor*

VGE : valeur génétique estimée

VP : vecteur propre

VTRA : vitesse de traite

Σ CN : caséines totales

Σ PS : protéines sériques totales

α-LA : alpha-lactalbumine

αs1-CN : caséine alpha s1

αs2-CN : caséine alpha s2

β-CN : caséine beta

β-LG : beta-lactoglobuline

κ-CN : caséine kappa

Introduction générale

Introduction générale

Le lait des ruminants, et en particulier celui des bovins, est utilisé pour fabriquer du fromage depuis les débuts de la domestication, soit depuis le Néolithique (5000 ans avant J.C.). Le lait peut être transformé en fromage à partir d'un simple ajout de présure, naturellement présente dans la caillette (estomac) des veaux non sevrés. Sans doute découvert de manière fortuite au départ, ce geste qui mime la digestion du veau, entraine la coagulation du lait (séparation du caillé ou fromage et du lactosérum ou « petit lait ») et transforme une denrée liquide périssable en un fromage, beaucoup plus facile à stocker et à conserver et aussi beaucoup plus digeste. La transformation du lait en fromage est également caractérisée par une acidification par transformation d'une partie du lactose en acide lactique par des bactéries. Cette acidification joue un rôle essentiel dans la texture du fromage et dans sa conservation. Le reste du lactose du lait est évacué dans le lactosérum lors de sa transformation, le fromage peut donc être consommé par les personnes intolérantes au lactose, ce qui représente environ les deux tiers des adultes de la population mondiale actuelle. Aujourd'hui dans le monde, plus d'un tiers du lait produit par les vaches est transformé en fromage. La France, avec sa très grande diversité de fromages (1200 variétés), est un des pays qui produit, consomme et exporte les plus grandes quantités de fromage.

Derrière le processus de transformation du lait en fromage, qui peut paraitre simple, se cachent des mécanismes physico-chimiques complexes liés à la composition du lait et notamment à la composition en protéines. En fonction de la composition du lait qu'elle produit, une vache peut donc fournir un lait plus ou moins fromageable, *i.e.* plus ou moins apte à la transformation. Les caractères de composition ou de fromageabilité du lait sont difficiles à mesurer avec les méthodes de référence, ce qui rend impossible leur étude à grande échelle. La spectrométrie moyen infrarouge (MIR) qui permet de prédire la composition de la matière organique et notamment celle du lait, offre une alternative aux mesures longues et coûteuses.

Les deux projets de recherche *PhénoFinlait* et *From'MIR* ont été mis en place dans le but d'exploiter la spectrométrie MIR pour prédire et étudier la composition fine et la fromageabilité (aptitudes physico-chimiques) du lait à grande échelle dans les principales races bovines laitières françaises. Ce travail de thèse s'inscrit dans ces deux projets, il a pour ambition de réaliser une étude génétique approfondie de ces caractères en vue de mieux comprendre leur

déterminisme génétique et d'étudier les possibilités de sélectionner les vaches pour qu'elles produisent des laits plus fromageables.

Ce manuscrit s'articule autour de six chapitres, le premier est une synthèse bibliographique et une présentation du contexte, les quatre suivants présentent l'ensemble des études réalisées pour ce travail de thèse et le sixième et dernier chapitre est une discussion générale.

Chapitre 1 - Dans le premier chapitre, nous posons le contexte bibliographique des caractères étudiés (la composition et les aptitudes fromagères du lait), de l'outil utilisé pour prédire ces caractères (la spectrométrie MIR) et des méthodes et outils disponibles pour étudier le déterminisme génétique et les possibilités de sélection de ces caractères.

Chapitre 2 - Le deuxième chapitre décrit les travaux réalisés pour prédire les aptitudes fromagères du lait à partir des spectres MIR.

Chapitre 3 - Dans le troisième chapitre, nous décrivons l'ensemble des analyses effectuées pour estimer les paramètres génétiques de la composition et de la fromageabilité du lait.

Chapitre 4 - Le chapitre 4 est dédié aux travaux de recherche et d'identification des gènes, des variations dans ces gènes et des réseaux de gènes impliqués dans le déterminisme de ces caractères.

Chapitre 5 - Nous étudions ensuite dans le chapitre 5 les possibilités de sélectionner les vaches pour qu'elles produisent un lait plus fromageable.

Chapitre 6 - Enfin, dans le 6ème et dernier chapitre, nous nous attachons à faire une discussion générale autour l'ensemble des résultats obtenus.

Chapitre 1

Contexte bibliographique

1. Contexte bibliographique

1.1. Le lait

1.1.1. Production de lait

Parce qu'il contient de nombreux nutriments dont certains sont essentiels (eau, glucides, protéines, lipides, minéraux, vitamines), le lait est un aliment complet et il occupe une place prépondérante dans l'alimentation humaine. En 2016, 826 milliards de litres de lait, dont 82% de vache, ont été produits dans le monde. Ce volume ne cesse d'augmenter, en raison notamment d'un accroissement régulier de la demande des pays émergents. L'Union Européenne (U.E. à 28 pays) réalise 24% de la production et 60% des exportations mondiales. Avec 25,8 milliards de litres de lait de vache produit en 2016, la France est le deuxième pays producteur en Europe après l'Allemagne (32,7 milliards de litres) et avant le Royaume-Uni (15,5 milliards de litres) (FranceAgriMer, 2018).

Le cheptel bovin laitier français compte près de 3,8 millions de vaches laitières réparties dans environ 58 000 exploitations sur tout le territoire. La France dispose d'une grande diversité de races bovines mais l'essentiel du lait est produit par les trois races Prim'Holstein, Montbéliarde et Normande qui représentent respectivement, 64%, 19% et 9% du cheptel bovin laitier français. Le reste (environ 8%) se répartit entre 6 races d'importance régionale (Abondance, Tarine, Brune, Simmental, Pie Rouge, Jersey), 4 races laitières à petits effectifs (Bretonne Pie Noir, Vosgienne, Flamande, Bleue du Nord) et une race allaitante mais avec un petit effectif à la traite (Salers) (Idele et CNE, 2018).

Comme tous les mammifères, une vache entre en lactation après avoir mis bas, le lait est alors sécrété dans la mamelle. Une forte proportion des constituants du lait sont synthétisés directement par les cellules épithéliales de la glande mammaire à partir de précurseurs d'origine sanguine qui proviennent entre autres de l'alimentation de la vache, les autres étant prélevés directement dans la circulation sanguine. La race, le stade de lactation et l'état sanitaire de la vache ont un effet sur la composition du lait. En moyenne, un litre de lait (densité = 1,030) est constitué d'eau (902 g) et d'un extrait sec (128 g) qui se décompose en 48 g de glucides, 37 g de lipides, 34 g de protéines et 9 g de sels minéraux.

1.1.2. Composition du lait

La plupart des éléments bibliographiques de ce paragraphe sont tirés de Alais (1984), Couvreur et Hurtaud (2007) et Léonil *et al.* (2013).

1.1.2.1. Glucides

Le seul glucide libre présent en quantité importante dans le lait est le lactose (48g/L). C'est un disaccharide, composé de glucose et de galactose. Du glucose et du galactose libres, issus de la dégradation du lactose, sont également retrouvés en quantité négligeable (0,1g/L). Le lactose est synthétisé par la glande mammaire à partir du glucose sanguin. Les pressions osmotiques du lait et du sang sont en équilibre, de sorte que toute diminution (augmentation) de la concentration en lactose sera compensée par une augmentation (diminution) des autres éléments solubles et notamment des minéraux. Les teneurs en lactose sont relativement peu variables, contrairement à d'autres constituants du lait.

1.1.2.2. Lipides

La fraction lipidique du lait est en moyenne de 37g/L. La matière grasse proprement dite est constituée de triglycérides (98%) mais on trouve également dans le lait des phospholipides, des stérols et des vitamines liposolubles (< 2%). Un triglycéride est constitué de trois acides gras (AG) et d'une molécule de glycérol. Un AG est lui-même constitué d'une chaîne hydro-carbonée (éléments -CH2- ou -CH=) et de deux groupements à ses extrémités : méthyle (-CH3) et carboxyle (-COOH). Plus de 400 AG différents ont été identifiés dans le lait de vache mais seuls 14 d'entre eux sont présents à plus de 1% dans la matière grasse du lait (Tableau 1.1). Ils se caractérisent par :

- La longueur de la chaîne carbonée
    - AG courts (2 à 10 atomes de C)
    - AG moyens (12 à 17 atomes de C)
    - AG longs (18 atomes de C et plus)
- Le degré de saturation des liaisons C
    - AG saturés (pas de double liaison)
    - AG mono-insaturés (une double liaison)
    - AG poly-insaturés (au moins deux doubles liaison)
- La position de la deuxième double liaison sur la chaîne carbonée, en plus de celle située sur le $9^{ème}$ C

  - ω3 = oméga 3 (double liaison sur le 3ème C)
  - ω6 = oméga 6 (double liaison sur le 6ème C)
- Les isomères de position : *cis* si la double liaison modifie le plan de la molécule (la majorité des AG), *trans* sinon.

Les AG peuvent être synthétisés *de novo* par la mamelle (tous les AG courts et moyens, une partie des longs) ou prélevés dans la circulation sanguine (la majorité des AG longs, d'origine alimentaire ou adipeuse). La matière grasse est le constituant du lait le plus variable et les proportions relatives des AG dépendent beaucoup de l'alimentation de la vache. Néanmoins, les parts relatives des AG saturés et insaturés sont assez stables, 2/3 et 1/3 respectivement. Parmi les AG insaturés, on retrouve surtout des AG mono-insaturés (env. 95%), les AG poly-insaturés étant peu abondants dans le lait. Les AG les plus fréquents sont l'acide palmitique (C16:0) et l'acide oléique (C18:1), ils représentent 30 et 25% des AG totaux respectivement (***Tableau 1.1***).

Tableau 1.1. *Proportions relatives des principaux acides gras de la matière grasse du lait de vache*

	Acides gras	Nomenclature	Moyennes (%)
Saturés 63%	Butyrique	C4:0	3,6
	Caproïque	C6:0	2,3
	Caprylique	C8:0	1,3
	Caprique	C10:0	2,7
	Laurique	C12:0	3,3
	Myristique	C14:0	10,7
	Pentadécanoïque	C15:0	1,2
	Palmitique	C16:0	27,6
	Stéarique	C18:0	10,1
	Arachidique	C20:0	0,2
Monoinsaturés 30%	Myristoléique	C14:1	1,4
	Palmitoléique	C16:1	2,6
	Oléique	C18:1	26,0
Polyinsaturés 4,2%	Linoléique	C18:2 ω6	2,5
	α-linolénique	C18:3 ω3	1,4
	Arachidonique	C20:4 ω6	0,3

La matière grasse est pour l'essentiel sous forme globulaire à l'état d'émulsion. Les globules gras, dont le diamètre peut varier entre 0,2 et 1,5 µm, sont assemblés dans les cellules épithéliales mammaires. Les petits globules (< 1 µm) sont plus nombreux (80%) mais ils représentent seulement 10% du volume total des globules gras. Un globule gras est composé d'un noyau (triglycérides à bas point de fusion), d'une zone intermédiaire (triglycérides à haut

point de fusion) et d'une membrane (lipoprotéines et phospholipides). Les caractéristiques physico-chimiques de la membrane ont un rôle déterminant sur la stabilité de l'émulsion.

1.1.2.3. Protéines

Les matières azotées du lait de vache (34g/L de lait) sont retrouvées sous forme de protéines (95%), ou pour une faible part, sous forme d'urée, de peptides ou d'acides aminés. La majorité des protéines (80%) sont des caséines et 20% sont des protéines solubles. Les caséines sont phosphorylées, les protéines sériques ne le sont pas. Parmi les protéines totales, les caséines les plus abondantes sont les caséines αs1 et β (env. 30% de chaque type), puis les caséines κ (10%) et αs2 (8%) et enfin la caséine γ (2,4%) qui est un sous-produit de protéolyse de la caséine β. Les protéines solubles, appelées aussi protéines sériques, sont principalement la β-lactoglobuline (env. 10%), l'α-lactalbumine (3,7%), les immunoglobulines (2,1%), la sérum albumine (1,2%) et la lactoferrine (0,7%). Toutes les caséines, la β-lactoglobuline et l'α-lactalbumine sont synthétisées par les cellules épithéliales de la glande mammaire, les autres protéines sériques sont d'origine sanguine (***Tableau 1.2***).

Tableau 1.2. *Proportions relatives des principales protéines du lait de vache*

	Protéines	Nomenclature	Moyennes (%)
Caséines CN - 80%	Caséine αs1	αs1-CN	30,6
	Caséine αs2	αs2-CN	8,0
	Caséine β	β-CN	28,4
	Caséine κ	κ-CN	10,1
	Caséine γ	γ-CN	2,4
Protéines sériques PS - 20%	β-lactoglobuline	β-LG	10,1
	α-lactalbumine	α-LA	3,7
	Immunoglobulines	Ig	2,1
	Sérum albumine	SA	1,2
	Lactoferrine	LF	0,7

Les caséines sont des phosphoprotéines plus ou moins phosphorylées. Les caséines αs1, αs2 et β ont des charges négatives qui leur permettent de fixer le calcium et de s'associer entre elles par l'intermédiaire de ponts phospho-calciques. La caséine κ se distingue des autres caséines par la présence de glucides, elle peut être plus ou moins glycosylée. Les propriétés spécifiques des différentes caséines leur permettent de s'organiser en un complexe stable appelée **micelle**. La micelle de caséine est une particule sphérique de 50 à 500 nm de diamètre qui se compose de 92% de caséines (αs1, αs2, β et κ en proportions relatives 3:1:3:1) et de 8% de sels minéraux (90% de phosphate de calcium et 10% d'ions magnésium et citrate) (***Tableau 1.3***).

Tableau 1.3. *Composition des micelles de caséines*

	Composé	g/100 g micelle
Caséines 92%	Caséine αs1	33
	Caséine αs2	11
	Caséine β	33
	Caséine κ	11
	Caséine γ	4
Minéraux 8%	Phosphate inorganique	4,3
	Calcium	2,9
	Magnésium	0,2
	Citrate	0,5

La structure de la micelle de caséine n'est aujourd'hui encore pas clairement établie et deux grands types de modèles ont été proposés : le modèle à sous-unités (Walstra, 1999) et le modèle à structure ouverte (De Kruif et Holt, 2003) (***Figure 1.1***).

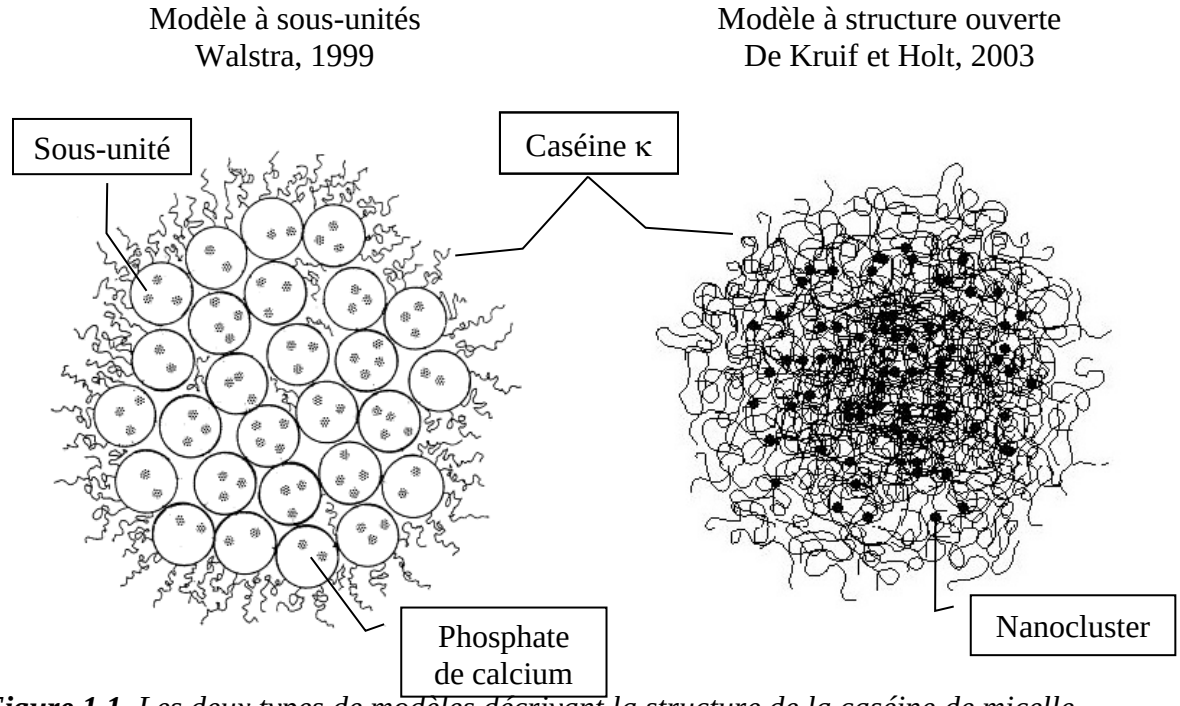


Figure 1.1. *Les deux types de modèles décrivant la structure de la caséine de micelle*

Dans le modèle à sous-unités, la micelle est une sphère composée de sous-unités protéiques, également sphériques (10 à 20 nm de diamètre), dont la cohésion est assurée par des interactions hydrophobes. Deux types de sous-micelles coexistent, celles constituées de caséines αs et β, situées au cœur de la micelle et celles constituées de caséines αs et κ, retrouvées à la surface de la micelle. Dans le modèle décrit par Walstra (1999), le phosphate de calcium formerait des microgranules à l'intérieur des sous-micelles.

Le deuxième modèle, à structure ouverte, décrit la micelle comme un réseau de « nanoclusters ». Un nanocluster est une structure de 2,3 nm, composée de calcium stabilisé par les interactions entre les groupements phosphorylés des caséines αs et β. La caséine κ interagit avec les autres caséines par son extrémité hydrophobe et forme, comme dans le modèle à sous-unités, une couche chevelue à la surface qui stabilise la structure colloïdale de la micelle.

Ces modèles sont sensiblement différents mais les deux décrivent la caséine κ à la périphérie et lui confèrent ainsi un rôle dans la stabilité colloïdale de la micelle. La déstabilisation de la structure colloïdale qui peut être induite par l'acidification du lait ou l'action des enzymes protéolytiques (chymosine) est le processus de base qui permet de transformer le lait en fromage (**coagulation**).

La fraction protéique soluble se compose de toutes les protéines autres que les caséines, parmi lesquelles la β-lactoglobuline et l'α-lactalbumine sont les plus abondantes. Contrairement aux caséines, ces protéines ne coagulent pas sous l'effet des enzymes protéolytiques, on les retrouve donc dans le **lactosérum**, qui est la phase liquide obtenue après coagulation du lait.

1.1.2.4. Minéraux et autres constituants

La fraction minérale est d'environ 9g/L de lait. Les minéraux sont présents dans le lait à l'état d'ions ou de sels non dissociés. Les plus abondants sont le potassium (1,5g/L), le calcium (1,2g/L), le phosphore (0,9g/L), le sodium (0,45g/L) et le magnésium (0,12g/L). On trouve également du chlore (1,15g/L) et du citrate (1,7g/L) ainsi que des oligo-éléments (Zn, Fe, Cu, Mn, Se…).

Les minéraux, bien qu'en quantité moindre par rapport aux autres constituants du lait, sont très importants d'un point de vue nutritionnel et technologique. Comme décrit précédemment, une partie de la matière minérale est sous forme colloïdale, *i.e.* associée aux molécules de caséines dans les micelles (64% du calcium, 46% du phosphate inorganique, 32% du magnésium et 7% du citrate) et l'autre partie est en solution sous forme d'ions (Na^+, K^+, Cl^-) ou de sels (phosphate et citrate). Les teneurs en minéraux sont donc étroitement liées aux concentrations en protéines.

1.1.2.5. Cellules somatiques et flore microbienne

La composition du lait en cellules somatiques dépend beaucoup du statut infectieux de la mamelle. Une mamelle saine contient des cellules peu nombreuses (lymphocytes, neutrophiles

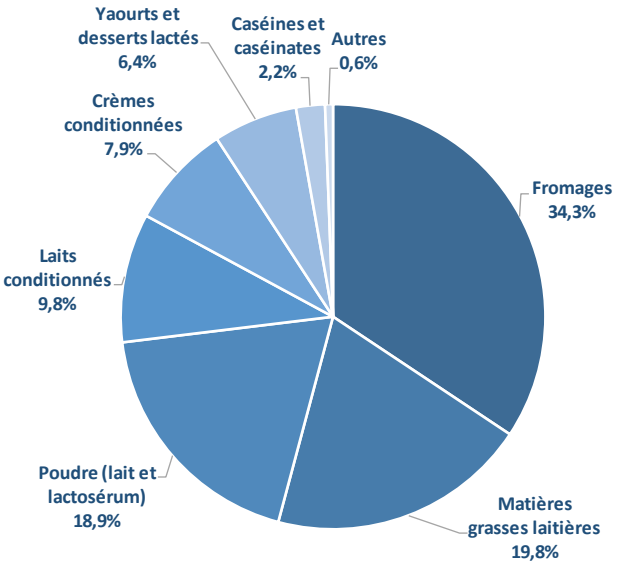
et macrophages, ainsi que des cellules épithéliales). En cas de mammite, le nombre de cellules somatiques augmente considérablement du fait de l'arrivée massive de neutrophiles.

On a longtemps cru que le lait d'une mamelle saine était stérile. On sait maintenant que ce n'est pas vrai et qu'il existe une flore commensale. Toutefois, en cas de mammite (ou infection mammaire) qui peut être provoquée par différents types de bactéries (staphylocoques, streptocoques, *Escherichia coli*…), le nombre de germes augmente considérablement. Par ailleurs, dès sa sortie du trayon de la mamelle, le lait est contaminé par la flore présente sur la peau des trayons, le matériel et l'environnement de traite. D'autres sources de contamination existent ensuite durant le stockage et le transport du lait. Un lait cru peut donc contenir une flore microbienne riche et diversifiée (bactéries, levures, moisissures…).

1.2. La transformation du lait en fromage

1.2.1. La production de fromage en France

En raison de son instabilité qui lui confère une faible aptitude à la conservation, le lait a subi des transformations dès les débuts de la domestication. Une étude britannique, relativement récente, montre qu'en Europe, les premiers agriculteurs du Néolithique (env. 5000 ans avant J.C.) fabriquaient et consommaient déjà du fromage (Salque *et al.*, 2013). Le lait ainsi transformé devenait non seulement une denrée non périssable mais également un aliment plus digeste pour beaucoup d'êtres humains de l'époque, intolérants au lactose. Au départ empiriques, les procédés ont évolué avec la compréhension des mécanismes physico-chimiques de transformation. Aujourd'hui en France, plus de 90% du lait de vache produit est transformé, en fromage (34,3% de la production totale), mais aussi en beurre et matière grasse (19,8%), poudres (18,9%), crème (7,9%) et yaourts et desserts lactés (6,4%)…(CNIEL, 2018) (***Figure 1.2***).



En équivalent M.S.U., matières sèches utiles du lait : protéines et matière grasse (CNIEL, 2018)

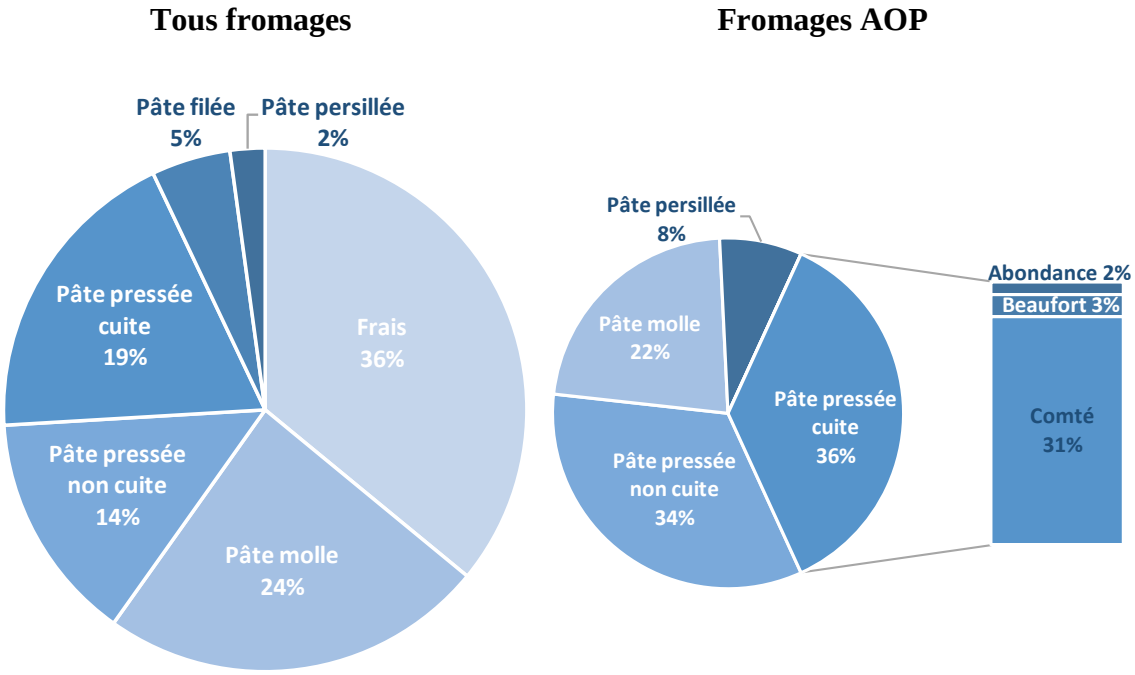
Figure 1.2. *Utilisation du lait pour la fabrication des produits laitiers en France en 2017*

Un produit laitier ne peut être qualifié de fromage que s'il est le résultat d'un processus de transformation particulier. Selon le décret n°2007-628 du 27 avril 2007 relatif aux fromages et spécialités fromagères : « *La dénomination "fromage" est réservée au produit fermenté ou non, affiné ou non, obtenu à partir des matières d'origine exclusivement laitière suivantes : lait, lait*

partiellement ou totalement écrémé, crème, matière grasse, babeurre, utilisées seules ou en mélange et coagulées en tout ou en partie avant égouttage ou après élimination partielle de la partie aqueuse. ».

En 2017, 1,7 millions de tonnes de fromage ont été produits sur le sol français, ce qui représente environ 17% de la production européenne et près de 9% de la production mondiale. Les français sont les plus grands consommateurs de fromage avec 27,2 kg de fromage consommé en moyenne par habitant et par an et la France est le plus grand pays exportateur de fromage en valeur (3 milliards d'euros), l'Allemagne exportant les plus grandes quantités (CNIEL, 2018).

Qualifiée parfois de « pays aux 1000 fromages », la France produit une large diversité de fromages fabriqués selon cinq grandes catégories de technologie : les fromages frais (type petit-suisse), les pâtes molles à croûte fleurie (type camembert traditionnel) et à croûte lavée (type munster), les pâtes pressées non cuites (type cantal) et cuites (type comté), les pâtes filées (type Mozzarella) et les pâtes persillées (type bleu de Gex) (***Figure 1.3***).



Sources CNAOL (2018) et CNIEL (2018)

Figure 1.3. *Proportions de fromages produits en France en 2017 par catégorie de technologie*

Parmi les fromages français à base de lait de vahe, 28 ont reçu une appellation d'origine protégée (**AOP**). Cette appellation, délivrée conjointement par l'INAO (Institut National de l'Origine et de la Qualité) et l'UE, garantit l'origine géographique du fromage (de la production

jusqu'à l'affinage), un savoir-faire reconnu et un cahier des charges spécifique. Les fromages AOP sont donc reconnus comme des aliments de qualité, typiques et élaborés dans le respect de l'environnement et du bien-être animal. Dans la filière fromagère, ils représentent environ 10% de la production totale, soit 167 338 tonnes, et plus d'un quart du chiffre d'affaires. Les fromages AOP sont en majorité des fromages à pâtes pressées cuite (36%) ou non cuite (34%) (CNAOL, 2018).

1.2.2. Zoom sur la région Franche-Comté

Le Comté, produit en Franche-Comté et plus précisément dans les départements du Jura, du Doubs et dans l'est de l'Ain, est le premier fromage AOP de France. Il bénéficie dès 1958 de l'Appellation d'Origine Contrôlée, devenue AOP en 1996. Avec 52 727 tonnes produites en 2017, le Comté représente à lui seul environ un tiers de la production totale des fromages AOP (***Figure 1.3***) et des fromages au lait cru. Il est le fromage franc-comtois le plus connu et parmi les plus consommés en France. La région Franche-Comté abrite d'autres fromages AOP fabriqués à partir de lait cru : le Morbier, le Mont d'Or (ou vacherin du Haut-Doubs) et le Bleu de Gex-Haut-Jura (15 620 tonnes en 2016, soit env. 10% de la production des fromages AOP au lait de vache).

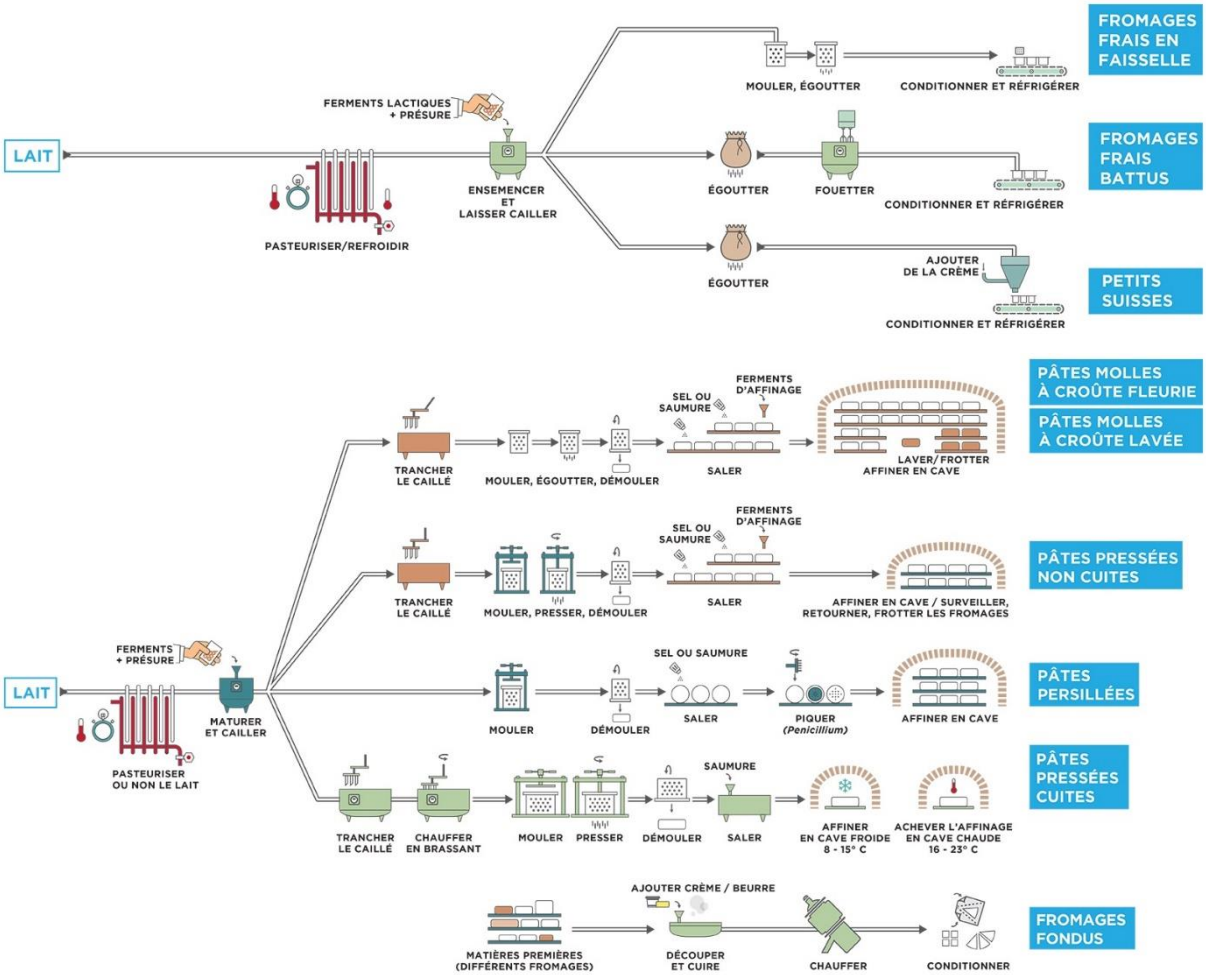
1.2.3. Les technologies fromagères

Il existe une très grande variété de fromages qui peuvent être classés en grandes familles (Eck, 1984, Mahaut *et al.*, 2000) :

1. les fromages frais ;
2. les fromages à pâte molle à croûte fleurie ou à croûte lavée ;
3. les fromages à pâte pressée non cuite ou cuite ;
4. les fromages à pâte persillée ;
5. les fromages fondus.

La fabrication du fromage comprend généralement quatre grandes étapes : 1) préparation du lait (standardisation / traitement thermique), 2) coagulation par voie enzymatique (présure) et/ou lactique (ferments), 3) égouttage et moulage (sauf pour les fromages fondus) et 4) salage et affinage (sauf pour les fromages frais et fondus). La ***Figure 1.4*** décrit l'enchainement des étapes nécessaires à la fabrication des différents types de fromage. Pour mieux comprendre les

mécanismes impliqués dans chacune des phases de la fabrication du fromage, nous détaillerons dans le paragraphe suivant les étapes qui conduisent à la fabrication des fromages au lait cru à pâte pressée cuite du type Comté.



https://www.produits-laitiers.com/le-circuit-de-fabrication-du-fromage/

Figure 1.4. *Processus de transformation des fromages frais, à pâte molle, à pâte pressée, à pâte persillée et fondus*

1.2.4. Fromages à pâte pressée cuite au lait cru, type Comté

Issue d'un savoir-faire ancestral, la fabrication du Comté obéit aujourd'hui à un cahier des charges bien précis. L'INAO (https://www.inao.gouv.fr/produit/4496) définit le mode de production : « *Le Comté est fabriqué de manière artisanale, à partir de lait cru de vaches de race Montbéliarde ou Simmentale française. Les vaches sont nourries exclusivement à partir d'herbe et de foin. Les produits fermentés sont interdits dans l'alimentation du troupeau laitier. Le système d'exploitation est extensif.* » et le mode d'élaboration : « *Le délai de fabrication est limité à 24h. L'ensemencement en ferment lactique reste essentiellement naturel. Le caillé est*

finement découpé, chauffé et brassé à 53°C pendant au moins 30 min. Le fromage est mis sous presse puis est ensuite démoulé, salé, frotté et retourné régulièrement. L'affinage de 4 mois minimum, permet une longue maturation de la pâte développant toute son onctuosité et sa palette d'arômes. ».

Pas moins de 420 L de lait sont nécessaires pour produire une meule de Comté (40-45 kg et 55-75 cm de diamètre). La fabrication du Comté est le résultat d'un savoir-faire et d'un travail concerté entre l'éleveur, le fromager et l'affineur. Les éleveurs de vaches de race Montbéliarde, pour l'essentiel, mettent en commun le lait dans des petites coopératives, appelées aussi fruitières, situées dans la zone de production (rayon de 25 km maximum). Le fromager transforme le lait en meules de Comté dans une cuve en cuivre, chaque cuve pouvant contenir l'équivalent de 12 meules au maximum. Les meules sont ensuite confiées à l'affineur dans la fruitière si elle contient une cave ou dans une cave d'affinage dédiée. Durant l'affinage, les meules, stockées sur des planches d'épicéa, sont retournées et frottées au sel régulièrement pendant au moins 4 mois. Chacune des étapes de fabrication est détaillée dans ce paragraphe.

1.2.4.1. Préparation du lait

Le lait utilisé pour la fabrication du Comté est un **lait cru**. Selon le règlement européen 853/2004, le lait cru est : « *le lait produit par la sécrétion de la glande mammaire d'animaux d'élevage et non chauffé à plus de 40°C, ni soumis à un traitement d'effet équivalent* ». Afin de satisfaire le rapport gras / sec exigé par le cahier des charges AOP Comté (« *45 à 54g de matière grasse pour 100g de fromage après complète dessiccation* »), le lait utilisé peut être partiellement écrémé.

1.2.4.2. Coagulation

Le lait cru subit une coagulation mixte via l'ajout de ferments lactiques et de présure.

a - Coagulation acide par ensemencement

Le lait est chauffé à 32°C puis ensemencé en ferments lactiques pour favoriser sa maturation. L'ajout de **ferments lactiques** a pour conséquence une acidification biologique, ce qui entraine une déstructuration des micelles de caséine (diminution des charges négatives et donc des répulsions électrostatiques et solubilisation du calcium et du phosphore). Lorsque le lait atteint

un pH de 4,6, un gel se forme. Ce gel est perméable et très friable à cause d'un manque de structuration du réseau.

b - Coagulation enzymatique par emprésurage

L'ajout de **présure**, solution extraite de la caillette des veaux non sevrés qui contient des enzymes protéolytiques (chymosine), permet d'obtenir un gel plus structuré. Pour aboutir à des temps de prise et de durcissement courts, des doses élevées d'enzymes protéolytiques sont ajoutées dans le lait (20 à 40 mL / 100 L de lait). La structuration du gel se déroule en trois phases, chacune étant sous l'influence de mécanismes plus ou moins connus.

i. La phase primaire ou enzymatique entraine l'hydrolyse de la caséine κ entre les 105ème (Phe) et 106ème (Met) acides aminés. Le caséinomacropeptide hydrophile qui correspond au fragment 106-169 et qui joue un rôle dans la stabilité de la micelle est ainsi libéré, il est retrouvé dans le lactosérum après hydrolyse.

ii. La phase secondaire, phase de coagulation proprement dite, commence lorsqu'environ 80% de la caséine κ est hydrolysée. La micelle perd son caractère hydrophile et la formation du gel a lieu grâce aux liaisons hydrophobes et électrostatiques qui se forment entre les micelles modifiées.

iii. Durant la phase tertiaire, les micelles se réorganisent via la formation de liaisons phosphocalciques et probablement de ponts disulfures.

L'étape de coagulation dure 30-35 min à 32-33°C.

1.2.4.3. Egouttage

Le caillé obtenu par coagulation est ensuite tranché, chauffé (35 à 50 min à 53-56°C) et brassé (40 à 60 min) pour un temps total de travail en cuve d'environ 2h30. Ces différentes opérations permettent d'obtenir un caillé avec une teneur élevée en matière grasse. L'étape de pressage dure une 20aine d'heures. La croûte est grainée par le dernier pressage en toile sèche après 4 à 5 retournements. Le salage se fait traditionnellement en saumure saturée en cave froide durant 24h.

1.2.4.4. Affinage

La meule de Comté ainsi créée est ensuite affinée dans une cave à température et hygrométrie contrôlée durant au moins 4 mois :

i. 4 à 6 semaines en cave froide (12-14°C) avec frottage au sel à l'aide d'un chiffon (emmorgeage) avec 2 ou 3 retournements par semaine ;
ii. 8 semaines environ en cave chaude (16-18°C, humidité relative 90-95%) avec le même traitement qu'en cave froide ;
iii. 4 semaines minimum (et souvent beaucoup plus) en cave froide (12-14°C) avec un lavage de la croûte et un retournement par semaine.

1.2.5. Critères de fromageabilité du lait et méthodes de mesure

La quantité et la qualité du fromage obtenu dépendent bien sûr de la technologie de transformation utilisée mais aussi des qualités intrinsèques du lait. Les caractéristiques **physico-chimiques** du lait, directement liées à sa composition, orientent notamment ses aptitudes technologiques aux différentes étapes de la transformation du lait en fromage. Un lait dit « fromageable » est un lait qui présente de bonnes aptitudes à l'**acidification** (phase de préparation du lait), à la **coagulation** (phase de coagulation) et qui permet d'obtenir des bons **rendements fromagers** (phase d'égouttage). Ces trois critères technologiques peuvent être mesurés au laboratoire par différentes méthodes, sur des quantités de lait limitées. Un lait peut donc être caractérisé pour sa fromageabilité sans la fabrication complète du fromage.

1.2.5.1. Paramètres d'acidification

Les paramètres d'acidification du lait peuvent être évalués en mesurant l'évolution du pH du lait après ajout de ferments lactiques. Le système CINAC (Corrieu *et al.*, 1988) a été développé pour suivre en continu l'activité acidifiante des ferments lactiques. Après ensemencement, le pH du lait est enregistré durant 20h et de nombreux paramètres sont mesurés sur la courbe de décroissance du pH (***Figure 1.5***) :

- **pH_0** = pH du lait au moment de l'ensemencement (en unité pH, upH) ;
- **$AR_{170\text{-}230min}$** = taux d'acidification du lait entre 170 et 230 min après l'ensemencement (upH/h) ;
- **pH_{10h}** = pH du lait 10h après l'ensemencement (upH) ;

- **$T\Delta pH_{0.08}$** = temps pour atteindre une chute de pH de 0,08 à partir de pH_0 (h) ;
- **$T\Delta pH_{6\text{-}4.8}$** = temps pour passer de pH 6 à pH 4,8 (h) ;
- **$\Delta pH_{16\text{-}20h}$** = décroissance du pH entre 16 et 20h après l'ensemencement (upH) ;
- **pH_{20h}** = pH du lait 20h après l'ensemencement (upH).

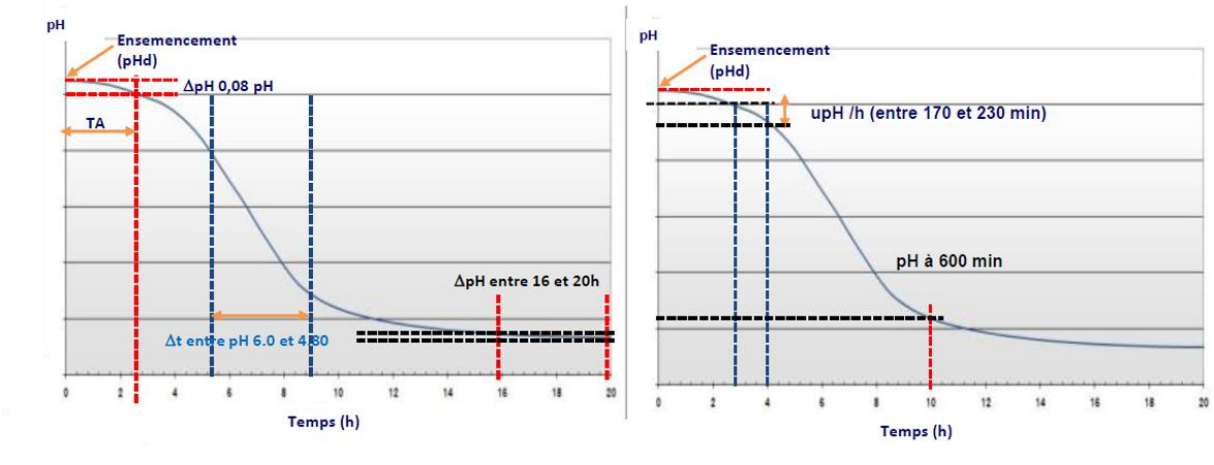


Figure 1.5. Courbe de décroissance du pH obtenue avec le système CINAC

1.2.5.2. Paramètres de coagulation

Les paramètres de coagulation classiquement mesurés sont le temps de coagulation, appelé aussi temps de prise, et les paramètres de raffermissement du caillé (vitesse et fermeté). Le temps de prise se définit comme le temps nécessaire à la formation d'un gel dans un récipient contenant un lait emprésuré (Alais, 1984). La vitesse d'évolution de la fermeté du gel est beaucoup plus variable que le temps de prise et la fermeté du gel au tranchage est déterminante pour la suite de la transformation.

Différentes méthodes sont utilisées pour évaluer les paramètres de coagulation, elles sont basées sur des dispositifs mécanique, vibratoire, ultrasonique ou optique. Une des méthodes les plus couramment utilisées est la « lactodynamographie ». Elle permet d'enregistrer la viscosité du lait à température fixe après ajout de présure.

Le Formagraph (Foss Electric A/S, Hillerød, Danemark), décrit par Cipolat-Gotet *et al.* (2012), est le lactodynamographe le plus largement utilisé depuis de nombreuses années notamment parce qu'il permet de réaliser des mesures sur plusieurs échantillons de lait en même temps. Un pendule est immergé dans le lait qui est soumis à des oscillations linéaires. Le pendule, immobile avant coagulation, se met en mouvement lorsque la viscosité du lait augmente.

L'amplitude du mouvement est donc d'autant plus forte que le gel se raffermit. L'appareil fournit en sortie un diagramme typique en forme de fourche sur lequel il est possible de mesurer différents paramètres, notamment le temps de coagulation (Rennet Coagulation Time ou RCT), le temps de raffermissement pour obtenir une fermeté standard de 20 mm (k_{20}) et la fermeté du gel 30 min après emprésurage (a_{30}) (***Figure 1.6***). Dans les premières versions du Formagraph, les sorties étaient imprimées sur du papier photographique et les mesures étaient réalisées à la main.

Le Formoptic est une version améliorée du Formagraph (Chr. Hansen, Hørsholm, Danemark et ENILBio, Poligny, France). Il contient un capteur optique qui enregistre le mouvement du pendule et le convertit en tension électrique. Ces mesures sont informatisées et un logiciel spécifique recrée le diagramme de coagulation. La fermeté étant proportionnelle à la tension enregistrée (un indice de fermeté, IF = 10 volts), le Formoptic fournit un diagramme de fermeté du gel en fonction du temps (***Figures 1.6*** et ***1.7***).

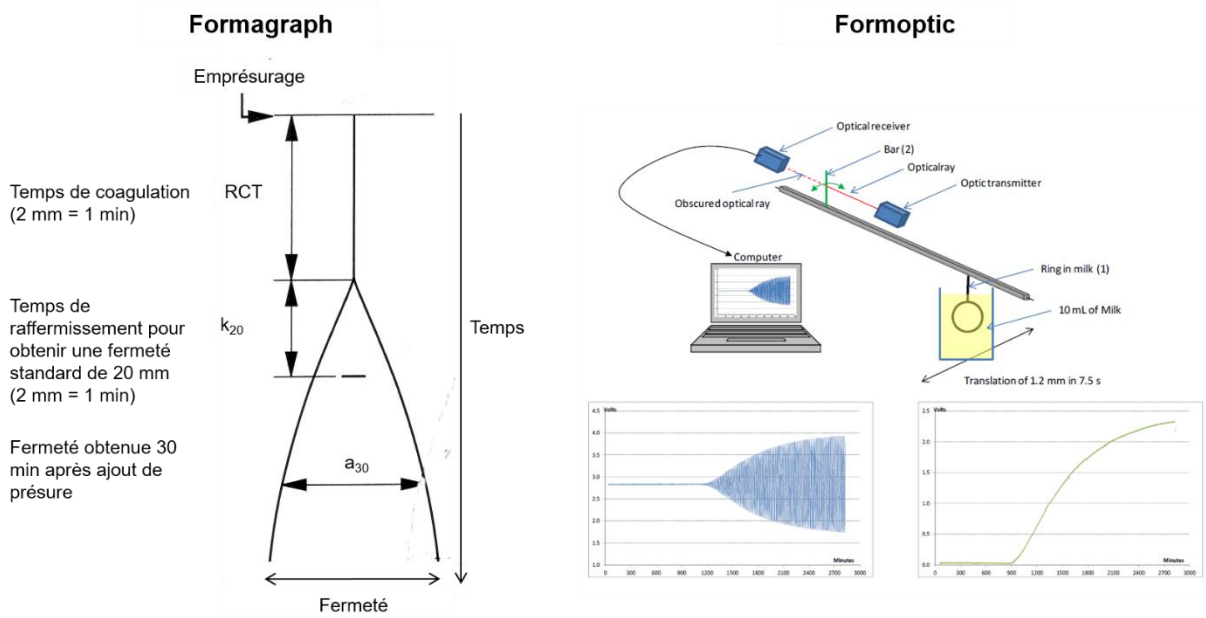


Figure 1.6. *Diagrammes et paramètres de coagulation obtenus avec les instruments Formagraph et Formoptic*

Le diagramme fourni par le Formoptic (***Figure 1.7***) permet de mesurer les paramètres suivants :

- **RCT** = temps de prise (*Rennet Coagulation Time*), soit le temps nécessaire pour obtenir 0,5 IF (min);
- **aR** = fermeté du gel à une fois le temps de prise, *i.e.* IF obtenu en reportant une fois le temps de prise (IF) ;

- **a2R** = fermeté du gel à deux fois le temps de prise, *i.e.* IF obtenu en reportant deux fois le temps de prise (IF) ;
- **K10** = temps nécessaire pour obtenir 10 IF à partir du temps de prise (min) ;
- **K10/RCT** = rapport entre K10 et RCT, détermine l'**inverse** de la vitesse d'organisation du gel, la vitesse est donc d'autant plus rapide que le ratio est faible ;
- **Tg10** = pente de la courbe entre 7 et 13 IF, elle exprime la vitesse moyenne d'organisation du gel, d'autant plus rapide que Tg10 augmente ;
- **Tg10/RCT** = rapport entre Tg10 et RCT.

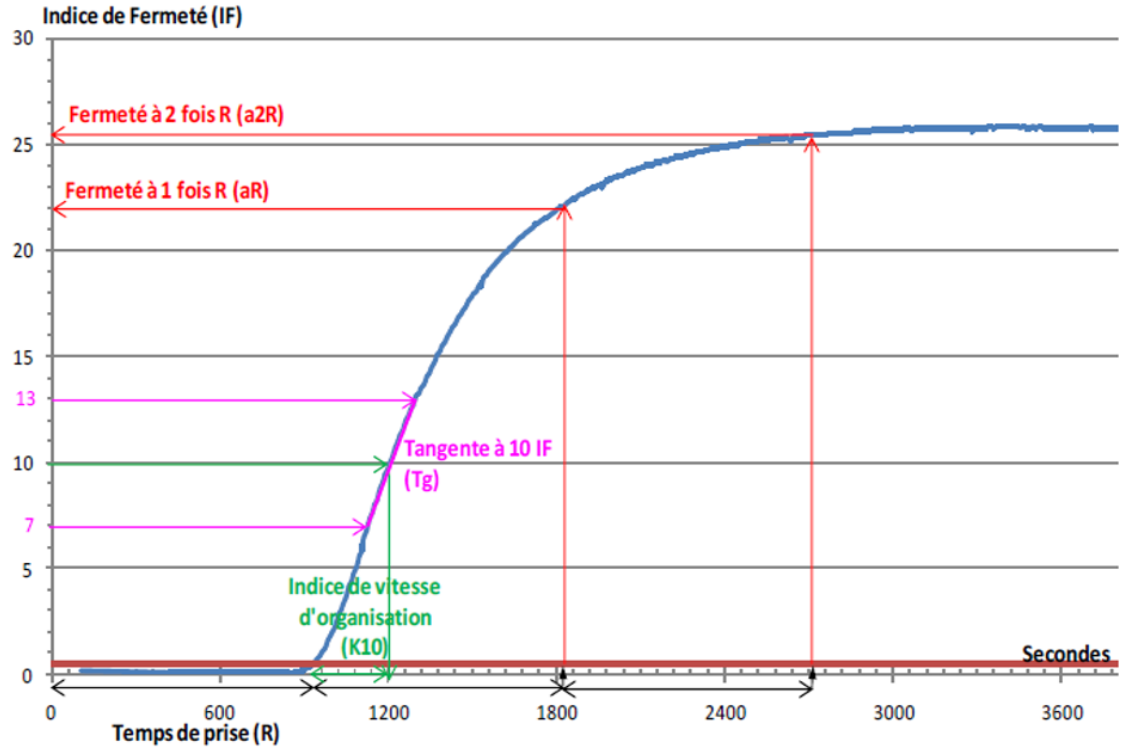


Figure 1.7. *Diagramme et paramètres de coagulation obtenus avec le Formoptic*

1.2.5.3. Rendements de laboratoire

Hurtaud *et al.* (1995) ont développé une méthode pour mesurer les rendements de laboratoire après centrifugation (2700 *g*, 15 min) d'un échantillon de lait coagulé (50 mL, pH 6,6, 32°C). La centrifugation a pour effet de séparer les phases solide (caillé) et liquide (sérum) et la matière sèche (ES pour extrait sec) peut être extraite de chacune des phases. Le caillé, le sérum et leurs extraits secs sont pesés pour calculer les rendements en appliquant les formules suivantes :

- Rendement frais en % (*fresh cheese yield*) :

$$\mathbf{CY_{FRESH}} = 100 \text{ x } \left(\frac{\text{g caillé}}{\text{g lait}}\right)$$

- Rendement en extrait sec (*cheese yield in dry matter*) en % :

$$\mathbf{CY_{DM}} = 100 \text{ x } \left(1 - \frac{\text{g ES sérum}}{\text{g ES lait}}\right)$$

- Rendement en matières protéique et grasse en g/kg (*fat and protein cheese yield*) :

$$\mathbf{CY_{FAT-PROT}} = (\text{TP} + \text{TB}) \text{ x} \left(\frac{\text{g lait}}{\text{g caillé}}\right)$$

avec **TP** = taux protéique et **TB** = taux butyreux.

Toutes ces mesures de laboratoire sont très utiles car elles permettent d'apprécier la fromageabilité d'un lait pour une technologie fromagère donnée sans avoir à fabriquer le produit fini. Néanmoins, ces mesures restent longues et coûteuses à mettre en œuvre et elles ne peuvent donc être effectuées que sur un nombre limité d'échantillons de lait. Des alternatives à ces mesures, plus faciles à réaliser et peu onéreuses et donc pouvant être réalisées à grande échelle, existent.

1.3. La spectrométrie moyen infrarouge

La spectrométrie moyen infrarouge (**MIR**) qui repose sur l'interaction entre la matière et les ondes électromagnétiques permet de déterminer la composition chimique d'un échantillon et donc de prédire les caractères qui sont directement liés à cette composition.

1.3.1. Principe de la spectrométrie MIR

La spectrométrie (ou spectroscopie) est l'étude des spectres électromagnétiques. Un spectre électromagnétique est l'ensemble des rayonnements classés par longueur d'onde (ou fréquence / énergie), de 0 à l'infini en théorie, de manière continue. Le spectre est découpé en plusieurs domaines qui dépendent de la longueur d'onde et du type de phénomènes physiques qui entraine l'émission de ce type d'ondes. La partie visible par l'œil humain ne correspond qu'à une toute petite partie de ce spectre (380 à 780 nm) (***Figure 1.8***).

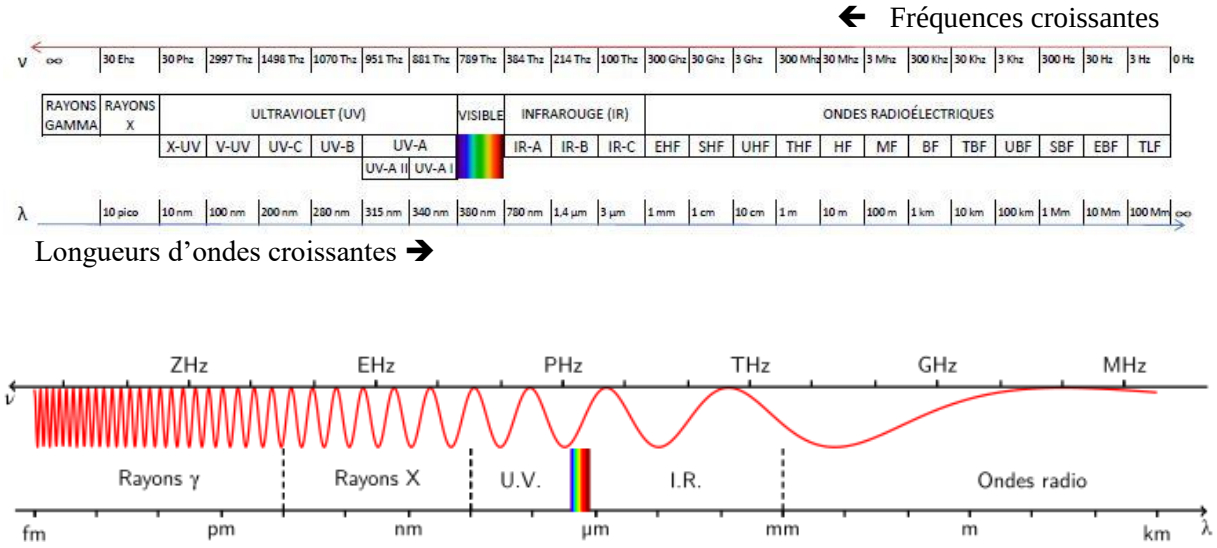


Figure 1.8. *Spectre électromagnétique*

Toute matière soumise à un rayonnement électromagnétique et donc à une source d'énergie (chaleur, lumière, micro-ondes, rayons X…) absorbe une partie de cette énergie. L'absorption s'accompagne de mouvements des molécules de la matière, d'autant plus marqués que la quantité d'énergie fournie est élevée (*i.e.* fréquences élevées et longueurs d'ondes petites). Les atomes vibrent ou tournent (rayons infrarouge), les électrons des couches de valence (rayons visibles ou ultra-violets) ou des couches profondes (rayons X) s'agitent et les noyaux des atomes peuvent se désintégrer (rayons γ). Tous ces phénomènes physiques dépendent des propriétés des atomes et des molécules qui constituent la matière.

La composition chimique d'un échantillon de matière organique, le lait par exemple, peut être caractérisée par la spectrométrie moyen infrarouge (MIR) qui correspond aux longueurs d'ondes comprises entre 2500 et 25000 nm. Dans cette gamme de longueurs d'ondes, les mouvements des molécules (vibrations ou rotations) sont associés à une plus ou moins grande absorption de l'énergie fournie. Connaissant la quantité d'énergie absorbée par la matière pour chacune des longueurs d'ondes du spectre, il est possible de caractériser les groupements fonctionnels, les liaisons chimiques et la structure des molécules qui constituent la matière organique. Un spectromètre à transformée de Fourier qui permet d'enregistrer simultanément l'absorption sur une gamme étendue de longueurs d'ondes, fournit directement un spectre d'absorption MIR par simple éclairage d'un échantillon de lait avec de la lumière infrarouge (***Figure 1.9***). Une molécule est caractérisée par plusieurs pics du spectre, chaque pic correspondant à une liaison moléculaire.

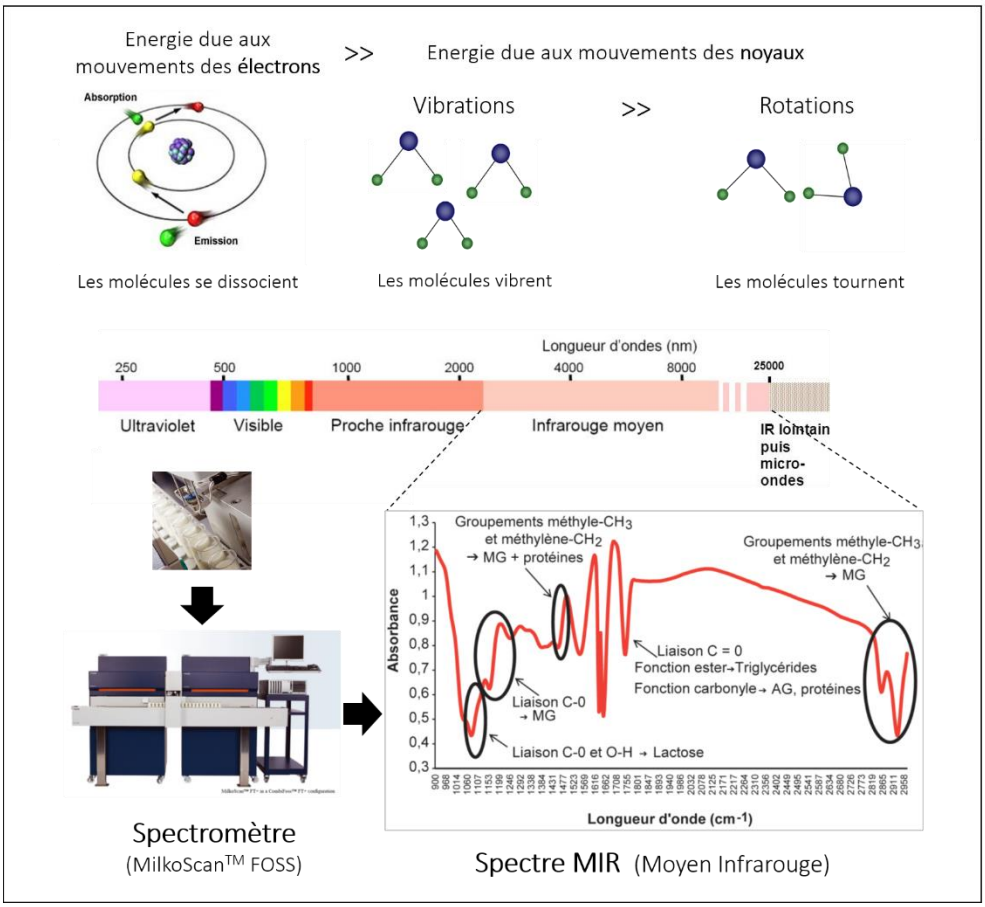


Figure 1.9. *Principe de la spectrométrie*

1.3.2. Du spectre MIR au phénotype

Le spectromètre fournit des spectres MIR bruts qu'il faut traiter et confronter à des données de référence avant de pouvoir déduire la composition du lait ou tout autre phénotype lié à cette composition.

1.3.2.1. Standardisation des spectres MIR

Tout d'abord, il existe plusieurs fournisseurs de spectromètres à transformée de Fourier, Foss® et Bentley® étant les plus répandus en France, et plusieurs modèles pour chaque fournisseur. Par ailleurs, les différentes machines d'un même modèle fournissent des spectres différents. Il peut en être de même pour la même machine au cours du temps. Pour analyser les spectres MIR produits dans des conditions différentes, une standardisation est nécessaire, à partir de spectres obtenus avec des échantillons témoins. Le projet européen *OptiMIR* (2011-2016) qui regroupait 17 partenaires de six pays européens (Allemagne, Belgique, France, Irlande, Luxembourg et Royaume-Uni) visait à créer une base de données rassemblant les spectres MIR de l'ensemble des partenaires. Dans le cadre de ce projet, Grelet *et al.* (2015) ont développé la procédure de standardisation dite « *Piecewise Direct Standardisation* » (Wang *et al.*, 1991). Cette procédure consiste à définir un appareil de référence (*master*), un FT6000 de Foss® dans le cas d'*OptiMIR*, et à corriger les spectres fournis par les autres appareils (*slaves*) (***Figure 1.10***).

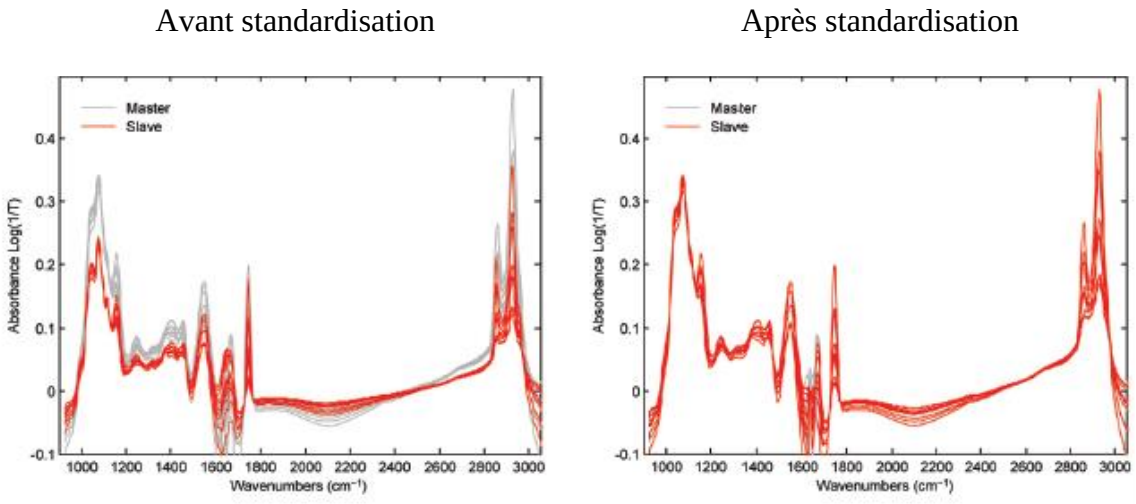


Figure 1.10. *Effet de la « Piece Direct Standardisation », d'après Grelet et al. (2015)*

De plus, les constructeurs des spectromètres recommandent de ne retenir que les longueurs d'onde informatives du spectre *i.e.* celles qui ne correspondent pas à l'absorption des molécules d'eau (446 pour Foss® par exemple).

1.3.2.2. Population de calibration

Pour prédire un phénotype à partir d'un spectre MIR, il est nécessaire de constituer une base de données avec des échantillons analysés à la fois par spectrométrie MIR et par la méthode de référence de mesure du phénotype. La population de calibration doit être de taille suffisante et aussi représentative que possible de la variabilité du phénotype de la population dans laquelle on veut prédire le phénotype. Il faut donc avoir une bonne connaissance du phénotype et veiller à échantillonner des laits qui couvrent toute la gamme des facteurs de variation du phénotype.

1.3.2.3. Equations de prédiction

Les données de la population de calibration (spectres et phénotypes mesurés par la méthode de référence) sont ensuite analysées via des techniques mathématiques (**chimiométrie**) qui incluent la sélection des longueurs d'onde, le prétraitement des spectres et le modèle statistique. Même en réduisant les données spectrales aux 446 longueurs d'onde informatives (hors absorption de l'eau), elles restent hautement dimensionnées et nécessitent l'utilisation de méthodes statistiques spécifiques. La méthode PLS (*Partial Least Square Regression*) est une des méthodes les plus classiquement utilisées pour développer des équations de prédiction à partir des spectres MIR (Soyeurt *et al.*, 2011). Elle permet de réduire la dimension du jeu de données et elle peut être couplée à un algorithme de sélection de variables (Ferrand-Calmels *et al.*, 2014). Elle donne de bons résultats pour prédire notamment les composants majeurs du lait (taux butyreux, taux protéique, lactose…). Des méthodes de régression bayésiennes, développées à l'origine pour des analyses génétiques (Meuwissen *et al.*, 2001), ont été plus récemment appliquées aux spectres MIR. Elles peuvent donner, dans certains cas, de meilleurs résultats que les méthodes PLS (Ferragina *et al.*, 2015).

1.3.2.4. Précision des équations de prédiction

La précision de la prédiction à partir d'un spectre MIR est grandement influencée par le type de phénotype à prédire. Le phénotype peut être direct, *i.e.* quantifié directement dans le spectre (ex. taux protéique), ou indirect, *i.e.* lié à un phénotype quantifié dans le spectre (ex. rendement fromager). Un phénotype indirect est en général moins bien prédit qu'un phénotype direct.

Toutefois, la perte de précision peut être au moins en partie compensée en optimisant la population de calibration (taille et sélection des échantillons) et en appliquant les méthodes chimiométriques les plus adaptées.

La précision d'une équation est généralement estimée sur une population de validation différente de la population de calibration (également appelée population d'apprentissage). La population de validation permet de comparer la prédiction à un phénotype non utilisé pour la prédiction. En cas de **validation externe** (validation la plus rigoureuse), l'échantillon de validation est complètement indépendant de l'échantillon de calibration. Lorsque cet échantillon indépendant n'est pas disponible, on peut réaliser une validation interne, souvent une **validation croisée**. Dans une validation interne croisée d'ordre *n,* la population mesurée de taille *N* est divisée en *n* sous-ensembles exclusifs, l'équation est établie à partir des données de *N*(*n-1)/n* échantillons puis appliquée sur les *N/n* échantillons laissés de côté. Cette opération est répétée *n* fois, en omettant à chaque fois un des sous-ensembles. A l'issue de cette opération, on dispose des *N* prédictions que l'on peut comparer aux *N* phénotypes mesurés.

Pour comparer les valeurs prédites et les valeurs vraies d'un phénotype donné et donc juger de la qualité d'une équation de prédiction, on calcule plusieurs paramètres et notamment le coefficient de détermination **R^2** qui est le carré de la corrélation entre les valeurs prédites et les valeurs vraies. R^2 peut varier entre 0 et 1 et une équation est d'autant plus précise que le R^2 est proche de 1. Karoui *et al.* (2006) et Coppa *et al.* (2010) qualifient une équation de « pauvre » si le $R^2 < 0,66$, approximative si $0,66 \leq R^2 < 0,81$, bonne si $0,81 \leq R^2 < 0,90$ et enfin, excellente si le $R^2 \geq 0,90$. Les autres paramètres souvent utilisés sont l'écart-type d'erreur de la prédiction (**RMSE** pour *Root Mean Squared Error*) qui combine à la fois précision et biais, l'**erreur relative** (RMSE / moyenne des données de référence) et le **RPD** (*Residual Prediction Deviation*) qui est le rapport entre l'écart-type des données de référence et le RMSE.

1.3.3. Les spectres MIR et la production laitière bovine

Parce que c'est une méthode rapide et peu coûteuse, la spectrométrie MIR est utilisée depuis de nombreuses années (début des années 1970) dans les laboratoires d'analyse du lait, à des fins de paiement ou de contrôle laitier (Biggs, 1978). Plusieurs dizaines de millions d'analyses sont réalisées chaque année en France. Aujourd'hui, un spectromètre de type Foss (Foss Electric A/S, Hillerød, Danemark) est capable d'analyser jusqu'à 300 échantillons de lait par heure et de mesurer, pour chaque analyse, l'absorbance pour 1060 longueurs d'onde. Les spectres MIR

sont donc utilisés en routine et à grande échelle pour prédire les composants qui entrent dans le calcul du paiement du lait (taux protéique, taux butyreux), d'autres composants comme le lactose, l'urée et le citrate ou encore les corps cétoniques ou le point de congélation du lait. De plus, les spectres MIR produits dans le cadre du contrôle laitier sont maintenant souvent stockés et constituent donc potentiellement une source de données pour du phénotypage à haut débit. En effet, toute nouvelle équation peut être appliquée à des données historiques pour produire un phénotype nouveau à grande échelle.

1.4. Programmes de recherche *PhénoFinlait* et *From'MIR*

Les potentialités de la spectrométrie MIR pour le phénotypage fin à grande échelle ont conduit à la mise en place de nombreux programmes de recherche visant à prédire de nouveaux caractères à partir des spectres MIR, à l'instar du projet européen *OptiMIR* évoqué plus haut. En France, les programmes ***PhénoFinlait*** (2008-2013) et ***From'MIR*** (2015-2018) ont successivement été financés principalement par l'ANR (Agence Nationale de la Recherche) et Apis-Gene pour le premier, par le CASDAR (Compte d'Affectation Spéciale pour le Développement Agricole et Rural), le CNIEL (Conseil National Interprofessionnel de l'Industrie Laitière) et l'URFAC (Union Régionale des Fromages d'Appellation d'origine Comtois) pour le second, et ils ont tous deux bénéficié du soutien de nombreux partenaires (***Figure 1.11***).



Figure 1.11. *Partenaires des projets PhénoFinlait (a) et From'MIR (b)*

De tels projets reposent sur un partenariat large et l'intégration de données et de technologies diverses. Les principaux acteurs du projet *From'MIR* sont détaillés ci-après.

- Les trois entreprises de conseil en élevage (ECEL) de Franche Comté, CEL25-90 (coordinateur du projet), Eva-Jura et CEL de Haute Saône ont fourni les spectres MIR, les prédictions issues du projet *OptiMIR*, les informations variées relatives aux élevages, et ont réalisé tous les prélèvements nécessaires. Ces entreprises ont pour métier le conseil dans plus de 3000 élevages, en s'appuyant sur une activité de phénotypage très active. A ce titre, elles fournissent de nombreuses données de routine présentes dans la base nationale également utilisées à des fins de sélection.

- Umotest (Union Montbéliarde de Testage) est une des trois et la plus importante entreprise de sélection en race Montbéliarde. Jusqu'en 2010, elle testait environ 120 taureaux sur descendance chaque année pour en commercialiser les meilleurs. Depuis la sélection génomique (voir §1.6.4), elle produit et génotype environ 1600 mâles par an pour en sélectionner et commercialiser 80 sous forme de paillettes d'insémination. Umotest conduit également une politique de génotypage femelle intra troupeau très active depuis le début de la sélection génomique (plus de 300 000 femelles génotypées à ce jour) ce qui a permis entre autres d'étendre la population de référence Montbéliarde aux femelles. C'est une partie de ces typages qui est utilisée dans le projet *From'MIR*. Umotest a également été la première entreprise française à intégrer un laboratoire de sexage de la semence.

- L'INRA et l'ENILBio (école nationale d'industrie laitière et des biotechnologies) de Poligny ont réalisé les travaux de phénotypage de l'aptitude fromagère, à l'échelle de l'individu, du troupeau et de la cuve de fromagerie, à partir des échantillons fournis par les ECEL et ont apporté leur expertise fromagère.

- L'Institut de l'Elevage (Idele) a assuré l'aide au montage et à la coordination du projet et a contribué à l'élaboration des plans d'échantillonnage, la construction des équations MIR et à l'analyse des facteurs de variation des caractéristiques fromagères des laits.

- L'INRA de Jouy-en-Josas a réalisé des travaux de caractérisation fine des laits de référence du projet, ainsi que les travaux d'analyse génétique des données MIR (travaux présentés dans cette thèse).

Les programmes *PhénoFinlait* et *From'MIR*, de grande envergure et complémentaires, avaient pour ambition commune de réaliser du phénotypage fin à haut débit à partir de spectres MIR, pour la composition fine du lait en acides gras et en protéines dans *PhénoFinlait* et pour la fromageabilité du lait dans *From'MIR*. Le programme *PhénoFinlait* a regroupé des acteurs des trois filières bovine (races Montbéliarde, Normande et Holstein), ovine et caprine, répartis sur tout le territoire français tandis que le programme *From'MIR* s'est concentré sur la filière AOP de la région Franche-Comté et sur la race bovine Montbéliarde. Par ailleurs, une partie des vaches du projet *PhénoFinlait* a été génotypée dans le cadre du projet tandis que le projet *From'MIR*, plus récent, a pu bénéficier des génotypages réalisés pour la sélection génomique (***Tableau 1.4***).

Tableau 1.4. *Nombre de spectres MIR, de vaches avec spectres et avec génotypes dans les projets PhénoFinlait et From'MIR*

Projet	Race	# Spectres MIR	# Vaches avec spectres MIR	# Vaches avec génotypes
PhénoFinlait	Montbéliarde	637 419	100 217	2967
	Normande	117 318	27 025	2737
	Holstein	152 147	32 714	2306
From'MIR	Montbéliarde	6 670 769	410 622	19 862

Pour la composition fine et les propriétés fromagères du lait de vache, les objectifs des deux programmes étaient :

1) de développer des équations de prédiction de ces caractères et de les appliquer à l'ensemble des spectres MIR produits dans le cadre du contrôle laitier et collectés pour les deux projets ;

2) d'étudier les effets des facteurs physiologiques et d'élevage sur ces caractères ;

3) d'analyser le déterminisme génétique de ces caractères en estimant les paramètres génétiques et en recherchant les régions du génome impliqués dans le déterminisme de ces caractères ;

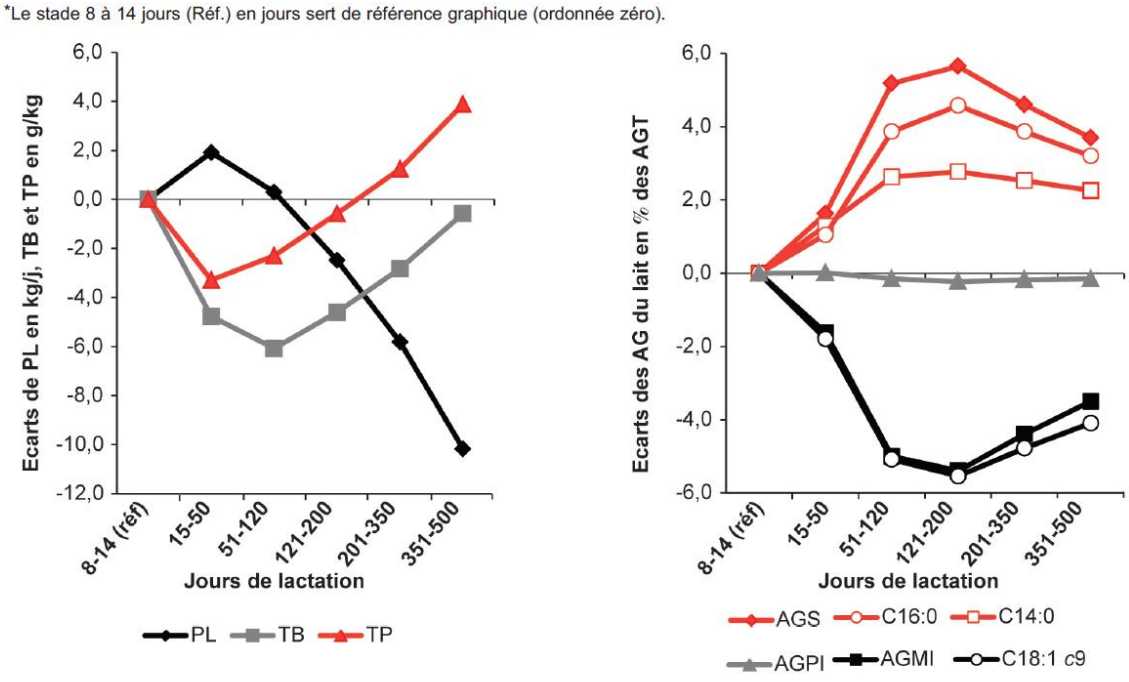
4) d'aboutir à des outils de pilotage de ces caractères par la génétique ou par la conduite du troupeau.

1.5. Facteurs de variation de la composition et de la fromageabilité du lait

La composition et les aptitudes fromagères du lait dépendent du statut physiologique, des conditions d'élevage et du potentiel génétique de la vache qui produit le lait.

Les principaux facteurs de variation non génétiques, étudiés dans les projets *PhénoFinlait* et *From'MIR*, sont le rang et le stade de lactation de la vache ainsi que son régime alimentaire et plus généralement le troupeau dans lequel elle se trouve (Gelé *et al.*, 2014, Legarto *et al.*, 2014, Gaudillière *et al.*, 2018).

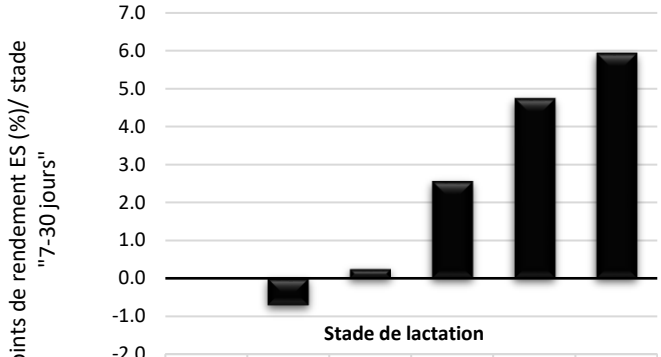
Le stade de lactation a un effet très fort sur la composition et la fromageabilité du lait. Sur des données du projet *PhénoFinlait*, Legarto *et al.* (2014) observent en effet une diminution de la concentration en protéines et en matière grasse jusqu'au pic de lactation puis un réenrichissement du lait après le pic jusqu'au tarissement (***Figure 1.12***). Les 50 premiers jours sont très différents du reste de la lactation, comme l'illustre par exemple la forte proportion d'acides gras monoinsaturés du fait de la forte mobilisation corporelle de la vache.



* Evolution de la production laitière (PL), des taux butyreux (TB) et protéique (TP) et du profil en acides gras (saturés, AGS ; monoinsaturés AGMI ; polyinsaturés AGPI), du lait de vaches Holstein (n = 125 514) en fonction du stade de lactation

Figure 1.12. *Courbes de lactation, d'après Legarto et al. (2014), projet PhénoFinlait*

Dans le projet *From'MIR*, Gaudillière *et al.* (2018) ont également montré un fort impact du stade de lactation sur les rendements fromagers et les paramètres de coagulation. Le rendement en extrait sec, par exemple, diminue légèrement jusqu'au pic de lactation pour ensuite ré-augmenter jusqu'à la fin de la lactation (***Figure 1.13***).



	7 à 30 j	30 à 60 j	60 à 100 j	100 à 200 j	200 à 300 j	300 à 500 j
paramètre estimé	0	-0.7	0.2	2.6	4.7	5.9

* Corrigé des autres facteurs, d'après Gaudillière *et al.* (2018)

Figure 1.13. *Effets du stade de lactation sur le rendement en extrait sec (%)**

L'effet de l'alimentation, et de l'élevage en général, est également important. Rappelons par exemple qu'une ration plus énergétique augmente le taux protéique, qu'une ration plus cellulosique favorise le taux butyreux et les acides gras saturés et polyinsaturés tandis que les concentrés tendent à diminuer le taux butyreux et augmenter la proportion globale d'acides gras insaturés. Les effets du régime alimentaire de la vache sur la composition du lait, peuvent se répercuter sur les aptitudes fromagères (Hurtaud *et al.*, 2009). En adaptant le régime alimentaire de la vache, il est donc possible de moduler, dans une certaine mesure, la composition et les aptitudes fromagères de son lait. Toutefois, les cahiers des charges de certains fromages AOP peuvent être restrictifs en ce qui concerne l'alimentation. En zone AOP Comté par exemple, les vaches doivent être nourries exclusivement à partir d'herbe et de foin comme fourrages, sans ensilage et avec une quantité limitée de concentrés.

Pour les caractères qui ont une composante génétique forte, comme les taux butyreux et protéique du lait par exemple, la sélection est relativement facile, sur toutes les populations de vaches, quelles que soient les contraintes d'élevage. Les gains relatifs espérés sont modérés à chaque génération (du fait d'un coefficient de variation génétique assez faible) mais la sélection présente l'avantage de se transmettre d'une génération à l'autre rendant possible le cumul du progrès génétique au fil du temps.

1.6. Méthodes d'analyse génétique

Les analyses génétiques requérant des dispositifs puissants, la spectrométrie MIR nous permet donc aujourd'hui d'étudier le déterminisme génétique et d'envisager la sélection de nouveaux caractères, tels que la composition et la fromageabilité du lait, qui ne pouvaient pas être mesurés facilement auparavant.

1.6.1. Modélisation de la performance

Pour les analyses génétiques, le modèle mathématique de base qui permet de décrire une performance est un modèle **linéaire** (combinaison linéaire des effets) dans lequel la performance *P* est modélisée par la somme des effets de milieu *M* et des effets génétiques *G* :

$$P = M + G \quad (1)$$

En pratique, on cherche à estimer les effets de milieu identifiés et enregistrés (les autres effets de milieu étant par définition non estimables) ainsi que les effets génétiques additifs des gènes qui sont ceux exploitables en sélection car transmis à la descendance. Les effets de milieu non identifiés ou non enregistrés et les effets génétiques non additifs liés aux interactions entre allèles d'un même gène (dominance) ou de gènes différents (épistasie) sont inclus dans l'effet résiduel, encore appelé erreur du modèle, que l'on cherche à minimiser.

Plus généralement, pour un vecteur **y** des observations (une observation par animal), (1) devient :

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Za} + \mathbf{e} \quad (2)$$

Sous l'hypothèse d'un **modèle polygénique infinitésimal**, les effets génétiques additifs **a** résultent de la somme d'effets petits et nombreux, et sont supposés distribués normalement, de variance $\mathbf{A}\boldsymbol{\sigma}_{\boldsymbol{a}}^{\mathbf{2}}$. La **variance génétique additive** $\boldsymbol{\sigma}_{\boldsymbol{a}}^{\mathbf{2}}$ est estimée à partir de la ressemblance (covariance) entre apparentés et du degré d'apparentement entre animaux (**matrice de parenté A**), que l'on déduit classiquement de la connaissance de la généalogie des individus et plus récemment des génotypes pour un grand nombre de marqueurs moléculaires répartis sur le génome (voir §1.6.3). De même, après prise en compte des effets de milieu identifiés qui affectent identiquement les performances « contemporaines » (même troupeau, même année,

même numéro ou stade de lactation…), les effets résiduels **e** sont supposés résulter de nombreux évènements aléatoires. Ils sont donc considérés indépendants entre individus et distribués normalement, de **variance résiduelle** $\mathbf{I}\sigma_e^2$. Par opposition, les effets de milieu **β** sont généralement considérés comme fixés, aucune hypothèse n'étant faite sur leur distribution. Les matrices **X** et **Z** du modèle (2) sont les matrices d'incidence qui relient les observations aux vecteurs des effets de milieu **β** et des effets génétiques additifs **a**, respectivement. Un modèle qui combine des effets fixés et des effets aléatoires est qualifié de modèle **mixte**. Le plus souvent, le modèle utilisé estime les valeurs génétiques des individus qui réalisent les performances et on parle de modèle **animal** (par opposition au modèle père par exemple qui estimerait les valeurs génétiques des pères à partir des performances de leurs filles).

En fonction des données utilisées (une ou plusieurs observations par animal) et des effets que l'on cherche à estimer, le modèle (2) pourra se décliner de différentes façons et être utilisé pour estimer les paramètres génétiques (§1.6.2), rechercher les régions du génome qui ont un effet sur les caractères (§1.6.3) et estimer les valeurs génétiques ou génomiques des individus (§1.6.4).

1.6.2. Paramètres génétiques

1.6.2.1. Définitions

La sélection exploite la variance génétique additive d'une population, *i.e.* la variabilité due aux effets additifs des gènes qui est transmissible aux descendants et qui peut être estimée directement à partir du modèle (2). L'**héritabilité** (**h^2**), qui est la proportion de la variance phénotypique d'origine génétique additive, permet de prédire si l'amélioration génétique par sélection sera efficace. Comprise entre 0 et 1, elle est d'autant plus élevée que les performances d'animaux apparentés se ressemblent. Un caractère est qualifié de peu héritable si $h^2 < 0{,}20$, modérément héritable si $0{,}20 < h^2 < 0{,}40$ et fortement héritable si $h^2 > 0{,}40$.

Pour estimer les effets indirects de la sélection ou pour évaluer les possibilités de sélectionner plusieurs caractères simultanément, on estime aussi les **corrélations génétiques** (**r_a**). La corrélation génétique est définie comme la corrélation entre les valeurs génétiques additives d'un même individu pour deux caractères. Sa valeur, qui varie entre -1 et +1, permet de quantifier le lien génétique entre deux caractères A et B. Indépendamment du sens (< 0 ou > 0), la corrélation peut être favorable si l'amélioration génétique de A est accompagnée d'une

amélioration génétique de B ou défavorable dans le cas contraire. On utilise dans ce cas un modèle mixte **multi-caractères** dont les effets fixés de milieu sont propres à chaque caractère et dont les distributions des effets aléatoires sont multinormales. Cela conduit à inclure les covariances dans les calculs (pour un même individu ou pour deux individus apparentés) pour les effets aléatoires du modèle. Un modèle multi-caractères est donc plus coûteux en ressources informatiques mais il permet d'estimer les corrélations génétiques entre plusieurs caractères tout en améliorant la précision des paramètres pour l'ensemble des caractères (les observations d'un caractère apportent une information qui va aider à estimer les effets génétiques et les effets de milieu d'un autre caractère).

1.6.2.2. Modélisation des performances

Une vache peut avoir plusieurs lactations et pour une lactation donnée, la performance est mesurée à chaque contrôle, soit généralement une fois par mois. Les modèles de description des données utilisés peuvent être plus ou moins complexes en fonction des hypothèses choisies. Ainsi, le phénotype inclus dans le modèle peut être la performance élémentaire de la vache à chaque contrôle, ou une performance agrégée par lactation. Selon les cas, les effets pris en compte dans le modèle ne sont pas les mêmes et le nombre de performances par vache pour estimer sa valeur génétique varie également. Un modèle de type **lactation** est plus simple à mettre en œuvre. D'autre part, il ne fait aucune hypothèse sur le déterminisme des contrôles élémentaires qu'il intègre dans une moyenne, mais sa validité repose sur un protocole de contrôle rigoureux. Le caractère analysé est généralement plus héritable que par contrôle puisque la composante résiduelle d'une moyenne est plus faible que celle de chacune des composantes. Au contraire, un modèle de type **contrôles élémentaires** permet une modélisation plus fine, en particulier des effets de milieu avec un effet troupeau x jour de contrôle qui permet de comparer entre eux tous les animaux produisant le même jour.

Lorsque l'animal ne réalise qu'une seule performance (par exemple dans le modèle **lactation** où seule la première lactation est analysée), on peut estimer les (co)variances à partir du modèle (2) et en déduire l'héritabilité h^2 de chaque caractère ainsi que la corrélation génétique r_a entre deux caractères à l'aide des formules suivantes : $h^2 = \frac{\hat{\sigma}_a^2}{\hat{\sigma}_a^2 + \hat{\sigma}_e^2}$ et $r_a = \frac{\hat{\sigma}_{a1,a2}}{\hat{\sigma}_{a1}.\hat{\sigma}_{a2}}$ avec $\hat{\sigma}_{a1,a2}$ l'estimation de la covariance génétique additive entre les caractères 1 et 2 et $\hat{\sigma}_{a1}$ et $\hat{\sigma}_{a2}$ les estimations des écart-types génétiques additifs pour les caractères 1 et 2, respectivement.

Dans un modèle **contrôles élémentaires**, chaque vache a plusieurs observations par lactation, le phénotype peut être modélisé selon deux types de modèles différents qui dépendent des hypothèses sous-jacentes.

i. Un premier modèle, le plus simple, suppose que le déterminisme génétique est constant le long de la lactation (mêmes variances, corrélations génétiques égales à 1). Comme les performances d'un même individu sont plus corrélées entre elles que ne le suppose l'héritabilité, on introduit un effet propre à l'animal mais non transmissible, dit effet d'environnement permanent, également de variance constante. Un tel modèle qui suppose donc que les effets ne varient pas au cours de la lactation est dit **modèle à répétabilité**. L'effet individuel de la vache est donc partitionné en un effet génétique additif $\mathbf{a}$ et un effet d'environnement permanent $\mathbf{p}$:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Za} + \mathbf{Wp} + \mathbf{e} \quad (3)$$

avec $\mathbf{p} \sim \mathbf{N}(\mathbf{0}, \mathbf{I}\sigma_p^2)$ le vecteur des effets de l'environnement permanent, $\mathbf{W}$ la matrice d'incidence, $\boldsymbol{\sigma_p^2}$ la variance de l'effet de l'environnement permanent et $\mathbf{I}$ la matrice identité. Les autres effets sont les mêmes que ceux décrits dans le modèle (2). On déduit du modèle (3), l'héritabilité h^2 et la répétabilité t : $h^2 = \frac{\hat{\sigma}_a^2}{\hat{\sigma}_a^2+\hat{\sigma}_p^2+\hat{\sigma}_e^2}$ et $t = \frac{\hat{\sigma}_a^2+\hat{\sigma}_p^2}{\hat{\sigma}_a^2+\hat{\sigma}_p^2+\hat{\sigma}_e^2}$ ainsi que la corrélation génétique entre deux caractères (*idem* modèle lactation (2)).

ii. On peut supposer que les effets (de milieu, génétique, environnement permanent et résiduel) peuvent varier au cours de la lactation. Dans ce modèle, à chaque jour correspond un caractère et les caractères sont d'autant plus corrélés qu'ils sont exprimés à des temps plus proches. Cette trajectoire de caractère peut se modéliser à partir de plusieurs (k) composantes avec un **modèle de régression aléatoire.** La valeur génétique *a(j)* le jour *j* varie donc de façon continue au cours de la lactation sous la forme suivante : $a(j) = \sum_{i=1}^{k} c_i(j)\, a_i$ avec a_i la valeur génétique de l'animal pour la composante *i* (*i*=1, *k*) et $c_i(j)$ le coefficient (connu) ne dépendant, pour une composante donnée *i*, que du stade de lactation *j*. Le modèle comprend les mêmes effets $\boldsymbol{\beta}$, $\mathbf{a}$, $\mathbf{p}$ et $\mathbf{e}$ que le modèle (3) mais les effets fixés et aléatoires sont modélisés en fonction du jour de lactation *j* :

$$\mathbf{y(j)} = \mathbf{X\beta(j)} + \mathbf{Za(j)} + \mathbf{Wp(j)} + \mathbf{e(j)} \quad (4)$$

Les variances $\boldsymbol{\sigma_a^2}$, $\boldsymbol{\sigma_p^2}$ et $\boldsymbol{\sigma_e^2}$ varient donc au cours de la lactation. Par exemple, la variance génétique au stade *j* est égale à $\sigma_a^2(j)$= **c**(j)' **V** **c**(j), avec **V** la matrice (*k*,*k*) des variances-covariances des valeurs génétiques a_i.

Ce modèle est plus généralement présenté en explicitant les *k* valeurs génétiques des composantes (et les *k'* effets d'environnement permanent) et en incluant les coefficients de régression dans les matrices d'incidence des effets :

$$y_i(j) = \boldsymbol{x_i b} + \sum_{l=1}^{k} c_l(j)\, a_{il} + \sum_{l=1}^{k'} d_l(j)\, p_{il} + e_{ij}$$

avec $y_i(j)$ la performance de l'animal *i* le jour *j*, $\boldsymbol{b}$ l'ensemble des effets fixés du modèle (dont certains dépendant du jour *j*), $\boldsymbol{x_i}$ le vecteur d'incidence correspondant, a_{il} la valeur génétique de l'animal *i* pour la composante *l*, p_{il} l'effet d'environnement permanent de l'animal *i* pour la composante *l*, $c_l(j)$ et $d_l(j)$ les coefficients pour la composante *l* et le jour *j* (identiques ou non) pour les effets génétiques et d'environnement permanent.

Avec un modèle de régression aléatoire, on n'obtient donc pas des valeurs uniques pour les paramètres génétiques mais des courbes de variances, de covariances, d'héritabilité $h^2(j)$, de répétabilité $t\,(j)$ et de corrélations génétiques $r_a(j)$ en fonction du jour de lactation *j* :

$$h^2(j) = \frac{\hat{\sigma}_a^2(j)}{\hat{\sigma}_a^2(j)+\hat{\sigma}_p^2(j)+\hat{\sigma}_e^2(j)},\ t\,(j) = \frac{\hat{\sigma}_a^2\,(j)+\hat{\sigma}_p^2(j)}{\hat{\sigma}_a^2(j)+\hat{\sigma}_p^2(j)+\hat{\sigma}_e^2(j)} \text{ et } r_a(j) = \frac{\hat{\sigma}_{a1,a2}(j)}{\hat{\sigma}_{a1}(j).\hat{\sigma}_{a2}(j)}$$

1.6.2.3. Méthodes d'estimation

La méthode de référence pour estimer les composantes de variance dans un modèle mixte est la méthode dite REML (pour REstricted Maximum Likelihood ou maximum de vraisemblance restreinte). Les estimations des paramètres du modèle sont celles qui maximisent la fonction de vraisemblance des données, marginalement aux effets fixés.

Plusieurs algorithmes ont été proposés, sans calcul de dérivées de la vraisemblance (DF-REML), ou utilisant les dérivées premières ou les dérivées secondes. L'algorithme dit AI-REML (pour Average Information REML) calcule les dérivées premières et utilise comme proxy des dérivées secondes la moyenne de l'information observée et de l'information espérée, cette moyenne étant plus simple à calculer que ses deux composantes. Il présente de bonnes performances, en particulier pour de gros jeux de données. Cet algorithme, implémenté dans le

logiciel WOMBAT (Meyer, 2007), est celui que nous avons utilisé pour estimer les paramètres génétiques de la composition et des aptitudes fromagères du lait.

L'estimation des paramètres génétiques, qui nous permettent de prédire l'efficacité de la sélection d'un caractère dans une population donnée, est donc une étape indispensable avant la mise en place d'une sélection. Les méthodes d'estimation sont toutefois peu puissantes et nécessitent des jeux de données importants. Les projets *PhénoFinlait* et *From'MIR*, qui exploitent les spectres MIR produits en routine et offrent donc de grands dispositifs (grand nombre de vaches avec phénotypes et généalogie), vont nous permettre d'estimer les paramètres génétiques de la composition et les aptitudes fromagères du lait de manière très précise.

1.6.3. Gènes et variants

1.6.3.1. Régions du génome avec des effets sur les caractères quantitatifs (QTL)

Comme nous venons de le voir, les modèles d'estimation des paramètres génétiques reposent sur le modèle polygénique infinitésimal qui suppose qu'un caractère quantitatif est sous l'influence d'un grand nombre de gènes non identifiés, chacun ayant un petit effet sur le caractère. Ce modèle, utilisé efficacement depuis de très nombreuses années en génétique quantitative, cache une réalité plus complexe. Dans les années 1990, le développement des cartes génétiques avec des marqueurs moléculaires (séquences nucléotidiques polymorphes, que l'on sait génotyper, et cartographiées sur le génome) de plusieurs espèces d'élevage, a permis de localiser des régions du génome ou **QTL** (*Quantitative Trait Loci*) avec des effets parfois forts sur les caractères quantitatifs. Les méthodes de cartographie de QTL exploitent le déséquilibre de liaison (**DL**) entre deux *loci* proches, *i.e.* les associations entre les allèles transmis ensemble au cours de la méiose (sans recombinaison). Si les marqueurs sont suffisamment nombreux et bien répartis sur le génome, certains d'entre eux sont en DL avec les mutations responsables des effets (causales) et permettent ainsi de les localiser. Les premiers QTL de production laitière ont pu être identifiés de cette façon dans l'espèce bovine (Bovenhuis et Weller, 1994, Georges *et al.*, 1995, Spelman *et al.*, 1996) et de nombreux autres QTL ont ensuite été décrits. Cependant, à quelques exceptions près, comme la mutation *K232A* dans le gène *DGAT1* qui a un effet très fort sur le taux butyreux notamment (Grisart *et al.*, 2002), très peu de gènes et de polymorphismes causaux dans ces gènes ont pu être identifiés par la suite, probablement à cause de la densité réduite des marqueurs microsatellites utilisés à l'époque, limitant la résolution des analyses.

1.6.3.2. Puces bovines et génotypages

Des puces à ADN pangénomiques qui contiennent des marqueurs **SNP** (*Single Nucleotide Polymorphism*) sont disponibles depuis une dizaine d'années. Les SNP sont des polymorphismes liés à la variation d'un seul nucléotide sur l'ADN et sont bi-alléliques. Ils présentent l'avantage d'être très nombreux (plusieurs dizaines de millions) et répartis sur l'ensemble du génome et ce, dans toutes les espèces. Dans l'espèce bovine, des puces avec plusieurs densités de SNP coexistent. Parmi elles, on retrouve les puces BovineSNP50™ (54 609 SNP, **puce 50K**), BovineHD™ (777 962 SNP, **puce HD**) et BovineLD™ (6909 SNP, **puce LD**) mises sur le marché par la société Illumina (Illumina Inc, San Diego, USA) en 2008, 2011 et 2012, respectivement. La puce LD, de moindre coût, a été créée de façon à permettre la reconstitution *in silico* (ou imputation décrite dans le paragraphe suivant) assez précise des génotypes 50K dans toutes les populations bovines (Boichard *et al.*, 2012a).

En Europe, cette puce LD a évolué ces dernières années vers la puce **EuroG10K** créée par et pour le consortium *Eurogenomics* qui regroupe huit pays européens (France, Pays-Bas, Allemagne, Danemark, Suède, Finlande, Pologne et Espagne). La nouvelle puce basse densité contient, en plus des 6909 SNP, deux autres parties : une partie commune à tous les pays, constituée de SNP génériques 50K et HD pour améliorer la qualité de l'imputation et de plusieurs centaines de SNP prédictifs ou causaux de la littérature ; une partie privative pour chacun des pays du consortium, dite aussi partie « recherche » ou « custom » (Boichard *et al.*, 2018). Chaque pays est libre de mettre dans sa partie privative les variants de son choix. Le coût attractif de la puce EuroG10K qui permet son utilisation à grande échelle pour la sélection génomique (voir §1.6.4) et sa mise à jour régulière, en font un outil particulièrement adapté pour valider les SNP identifiés dans les travaux de recherche.

Parmi les vaches avec spectres MIR des projets *PhénoFinlait* et *From'MIR*, une partie a été génotypée pour la puce 50K ou LD : 2967, 2737 et 2306 vaches de race Montbéliarde, Normande et Holstein, respectivement pour *PhénoFinlait* et 19 862 vaches de race Montbéliarde pour *From'MIR*.

1.6.3.3. Projet 1000 génomes bovins et imputations sur la séquence

Même si leur densité peut être élevée, les puces commerciales disponibles ne nous donnent pas accès à l'intégralité des variations du génome et donc pas à tous les variants causaux. Initié en 2012, le projet international « **1000 génomes bovins** » qui rassemble aujourd'hui 36 partenaires

a pour objectif le partage des données de séquences complètes (incluant toutes les variations du génome) d'animaux de l'espèce bovine (Daetwyler *et al.*, 2014, Bouwman *et al.*, 2018). Depuis le début du projet, plusieurs « runs » se sont succédés et lors de son sixième run en 2017, les données de séquences de 2333 bovins de près de 70 races ou types génétiques étaient disponibles. Cette population de référence rend possible l'**imputation** de la séquence complète des taureaux et des vaches génotypés pour les puces commerciales. L'imputation permet la reconstitution des variations bi-alléliques (variants) du génome bovin. Cette liste de variants (SNP ou insertions-délétions de quelques nucléotides (indel)) étant proche de l'exhaustivité, elle contient en théorie les variants causaux qui ont des effets sur les caractères. Notons cependant que cette phrase doit être modulée dans la mesure où il existe d'autres variants structuraux (insertions, délétions, duplications, inversions, ...), souvent négligés dans une première approche.

Pour imputer les génotypes manquants, plusieurs approches exploitant l'information apportée par la famille et / ou la population sont utilisées. L'approche familiale utilise le pedigree pour reconstituer les génotypes en appliquant les règles de transmission mendélienne. Elle s'appuie aussi sur la grande taille des segments chromosomiques transmis entre parents et produits. L'approche « population » exploite le DL entre les marqueurs proches (haplotypes) sans tenir compte directement des relations de parenté entre les individus. Cette dernière approche, implémentée dans le logiciel Minimac (Howie *et al.*, 2012), est précise mais elle présente l'inconvénient d'être relativement lente et donc difficilement applicable aux très grandes populations. Le logiciel FImpute (Sargolzaei *et al.*, 2014) utilise une approche combinée qui exploite les haplotypes partagés entre les animaux apparentés (pedigree + DL), ce qui améliore nettement les temps de calcul. La précision de l'imputation dépend i) de la méthode utilisée pour reconstituer les génotypes manquants mais aussi, ii) des relations entre la population de référence et la population à imputer, iii) de la densité des génotypes dans les deux populations, iv) des fréquences alléliques des variants imputés et v) du DL entre les marqueurs adjacents. Il convient aussi de noter que les méthodes reposant sur la construction et l'utilisation de bibliothèques d'haplotypes, comme FImpute, fournissent le génotype le plus probable, tandis que les méthodes bayésiennes utilisées par exemple dans Minimac fournissent la probabilité de chaque génotype possible ainsi que la précision d'imputation.

Grâce à l'effort de séquençage réalisé au niveau international en race Holstein et au niveau national pour les races Montbéliarde et Normande, il a été possible d'imputer les séquences complètes des vaches des projets *PhénoFinlait* et *From'MIR*.

1.6.3.4. Détection des QTL

Les animaux qui ont à la fois des génotypes (vrais ou imputés) et des phénotypes sont utilisés pour rechercher les QTL. De façon similaire aux méthodes utilisées pour l'imputation, il existe plusieurs types de méthodes de détection de QTL qui dépendent notamment de la nature de l'information utilisée. L'approche la plus communément employée, qui exploite le DL entre les marqueurs et les QTL, est l'analyse d'association ou **GWAS** pour *Genome Wide Association Study*, implémentée dans de nombreux logiciels dont GCTA (Yang *et al.*, 2011). Si la population étudiée a une structure familiale forte (familles de demi-sœurs par exemple) une analyse de type **LDLA** pour *Linkage Disequilibrium and Linkage Analysis* (Meuwissen et Goddard, 2000) permet de prendre en compte toute l'information (intra famille et populationnelle) et donc de maximiser la puissance de détection des QTL. Toutefois, cette dernière approche, plus exigeante en ressources informatiques, est très difficilement applicable aux grandes populations et aux fortes densités de marqueurs. Une analyse GWAS, qui teste l'effet individuel de chaque variant tout en incluant dans le modèle un effet polygénique aléatoire pour corriger la structure familiale de la population (estimé *via* une matrice de parenté calculée à partir du pedigree ou des marqueurs), est relativement simple et elle permet d'identifier les QTL avec une bonne résolution (Visscher *et al.*, 2017). D'autres méthodes estiment l'effet des marqueurs de manière simultanée, *e.g.* GBLUP (Legarra *et al.*, 2018) ou Bayes R (Erbe *et al.*, 2012). Ces méthodes s'affranchissent du DL à grande distance et fournissent donc généralement des positions plus précises, mais elles sont d'autant plus difficiles à mettre en œuvre que le réseau de marqueurs se densifie.

1.6.3.5. Identification des gènes et variants candidats

Une analyse GWAS réalisée au niveau de la séquence, qui contient en théorie les variants causaux, permet d'améliorer la puissance de détection des QTL et leur résolution (Daetwyler *et al.*, 2014). Toutefois, en raison du très fort DL présent dans une région QTL donnée, il est la plupart du temps difficile d'identifier un seul variant, voire le gène candidat.

Pour mieux cibler les variants candidats, il est possible de raffiner les modèles utilisés pour les GWAS. Pour les QTL partagés entre races, une analyse multi-races permet d'exploiter

l'historique des recombinaisons propre à chaque race et de bénéficier d'un DL à de plus courtes distances, et donc d'affiner la position des variants causaux (Raven *et al.*, 2014). De plus, elle permet de cumuler les effectifs et donc de gagner en puissance. Une méta-analyse entre races combinant les résultats d'analyses intra-race a des propriétés comparables (van den Berg *et al.*, 2016). Par ailleurs, pour déterminer si tous les variants significatifs d'une région QTL donnée sont en DL avec une même mutation causale, il est possible de réaliser une deuxième analyse GWAS multi-marqueurs locale en testant l'effet de chaque variant conditionnellement au variant le plus significatif du pic (Yang *et al.*, 2012). Cette dernière analyse peut nous permettre d'exclure des pics secondaires ou au contraire d'identifier plusieurs mutations causales dans une même région QTL.

Les résultats des GWAS peuvent ensuite être confrontés aux données d'annotation du génome bovin obtenues avec le logiciel Variant Effect Predictor (VEP), disponibles sur le site *Ensembl* (www.ensembl.org (McLaren *et al.*, 2016)). Parmi les variants en DL, on pourra alors choisir les **variants candidats** en fonction de leur localisation fonctionnelle et de leur effet prédit et privilégier par exemple les variants qui sont dans des gènes, voire dans des régions codantes de ces gènes. Enfin, différentes approches peuvent être utilisées pour définir des **réseaux de gènes**. Dans cette thèse, nous avons utilisé la méthode AWM pour *Association Weight Matrix* (Fortes *et al.*, 2010) afin d'exploiter les effets pléiotropiques des gènes sur les caractères de composition et de fromageabilité du lait du projet *From'MIR*. Le réseau ainsi créé permet *in fine* d'identifier les voies métaboliques et les régulateurs associés aux caractères étudiés.

1.6.4. Sélection

Pour sélectionner un caractère dans une population, on estime la **valeur génétique** des animaux candidats à la sélection pour ce caractère. Cette estimation est aussi appelée index génétique. On peut ensuite classer les candidats en fonction de leurs index et sélectionner les meilleurs pour produire la génération suivante.

1.6.4.1. Objectif de sélection en race Montbéliarde

Les valeurs génétiques sont estimées pour chaque caractère. Or, on cherche à améliorer plusieurs caractères en même temps. On définit ainsi un index de synthèse racial (**ISU** pour index de synthèse unique) qui est la combinaison linéaire des valeurs génétiques estimées pour chaque caractère et dont les coefficients sont les pondérations économiques attribuées aux

différents caractères. L'ISU, qui est le reflet de l'objectif de sélection, est actuellement propre à chaque race. En race Montbéliarde, il est calculé de la façon suivante depuis 2012 :

ISU = 0,45 SYNTLAIT + 0,145 STMA + 0,18 REPRO + 0,05 LGF + 0,05 VTRA + 0,125 MO avec SYNTLAIT, l'index économique laitier ; STMA, l'index de synthèse de santé de la mamelle ; REPRO, l'index de synthèse de la fertilité ; LGF, l'index de la longévité fonctionnelle ; VTRA, l'index de la vitesse de traite et MO, l'index de synthèse de la morphologie. Dans cette formule, chaque index est exprimé en écart-type génétique, les poids reflètent donc directement l'importance de chaque composante.

L'index économique laitier est lui-même calculé ainsi : SYNTLAIT = 1,050 (MP [kg] + 0,1 MG [kg] + 3 TP [g/kg] + 0,5 TB [g/kg]) avec les index élémentaires des matières protéique (MP) et grasse (MG) et des taux protéique (TP) et butyreux (TB). Exprimée en écart type génétique, cette formule se réécrit : SYNTLAIT = k (0,70 MP + 0,09 MG + 0,16 TP + 0,05 TB), k étant une valeur arbitraire pour standardiser la variabilité.

En raison de la forte utilisation du lait dans les filières fromagères en race Montbéliarde, l'index SYNTLAIT donne un poids beaucoup plus fort au TP (et un poids plus réduit à la MG) par rapport à l'index économique laitier calculé dans les autres races : INEL = 0,98 (MP [kg] + 0,2 MG [kg] + TP [g/kg] + 0,5 TB [g/kg]) soit en unités d'écart-type génétique : INEL = k (0,71 MP + 0,19 MG + 0,05 TP + 0,05 TB).

1.6.4.2. Méthodes d'évaluation génétique & génomique

Depuis les années 1990 et jusqu'à récemment, la méthode de référence d'estimation des valeurs génétiques, basée sur le modèle polygénique infinitésimal, était le BLUP (*Best Linear Unbiased Prediction*) appliqué à un modèle animal. Cette méthode permet d'estimer tous les effets du modèle de manière simultanée à partir des performances des animaux et des pedigrees. L'index d'un individu est estimé à partir de ses performances propres et celles de tous ses apparentés, le poids d'une performance dans l'index dépendant de l'apparentement entre l'individu réalisant la performance et l'individu à évaluer. Les apparentements sont estimés à partir du pedigree supposé connu et sans erreur. Cette évaluation dite polygénique fournit des prédictions des valeurs génétiques pour tous les animaux indépendamment de l'environnement dans lequel ils se trouvent ou de la période durant laquelle ils ont réalisé leur performance.

Toutefois, avec une évaluation polygénique type BLUP, il fallait attendre qu'un taureau laitier soit testé sur descendance, *i.e.* que ses filles aient des performances, pour pouvoir estimer sa valeur génétique de manière précise. Ce dispositif était long et coûteux. Avec l'arrivée des puces à ADN commerciales en 2008, l'**évaluation génomique**, qui permet de prédire les valeurs génétiques (ou génomiques) des animaux à partir des génotypes pour un grand nombre de marqueurs répartis sur le génome, s'est rapidement développée. Aujourd'hui, il est donc possible d'estimer la valeur génomique d'un animal dès sa naissance à partir d'un génotypage réalisé sur l'ADN d'un échantillon de sang ou de cartilage. Les jeunes animaux qui n'ont pas encore de performance peuvent alors être sélectionnés sur valeurs génomiques et, dès qu'ils sont en âge de procréer, ils peuvent être diffusés (mâles) ou utilisés pour le renouvellement (femelles).

Parce qu'elle permettait d'envisager un progrès génétique plus rapide et une meilleure efficacité sur les caractères peu héritables, la **sélection génomique** a été mise en place en France dès 2009 dans les trois races laitières nationales Holstein, Montbéliarde et Normande (Boichard *et al.*, 2012b), puis plus récemment dans les races régionales à effectifs plus limités Brune (Baur *et al.*, 2014), Abondance, Tarentaise et Vosgienne (Sanchez *et al.*, 2016b).

La méthode d'évaluation génomique de type GBLUP (*Genomic-BLUP*) est dérivée du BLUP. Elle utilise la matrice de parenté calculée à partir des génotypes aux marqueurs. Avec un nombre suffisant de marqueurs (au moins plusieurs dizaines de milliers), les coefficients de parenté entre les animaux sont estimés plus précisément qu'avec le pedigree qui permet de calculer l'espérance d'apparentement (*e.g.* à partir du pedigree, deux pleines sœurs ont un coefficient d'apparentement de 50% quel que soit le pourcentage d'allèles communs qu'elles ont reçu de leurs parents).

Généralement, dans une population, seule une partie des animaux est génotypée. Or, nous sommes intéressés par les valeurs génétiques / génomiques de l'ensemble des animaux phénotypés et / ou génotypés. Pour cela, il est possible d'effectuer les évaluations en plusieurs étapes. Une première étape consiste à réaliser une évaluation polygénique pour estimer les valeurs génétiques des animaux non génotypés. Ensuite, les effets de milieu estimés par l'évaluation polygénique sont utilisés pour corriger les performances des animaux génotypés. Enfin, les valeurs génomiques des animaux génotypés sont estimées à partir de leurs performances corrigées. Ces méthodes sont relativement faciles à implémenter mais elles

nécessitent des approximations qui entrainent une perte de précision. De plus, les évaluations polygéniques ne tiennent pas compte de la sélection génomique et les index polygéniques sont donc biaisés. Pour résoudre ces problèmes, une approche dite *Single Step* qui évalue l'ensemble des animaux du pedigree a été développée. En estimant tous les effets en même temps (de milieu, génétiques et résiduels) pour tous les animaux (génotypés et non génotypés), elle permet d'obtenir des valeurs génétiques / génomiques moins biaisées et plus précises qu'une méthode qui réalise une évaluation en plusieurs étapes.

Dans une approche Single Step GBLUP (**SS-GBLUP**), les parentés des animaux calculés à partir du pedigree (animaux non génotypés) et des génotypes (animaux génotypés) sont combinées pour estimer les valeurs génétiques des animaux (Aguilar *et al.*, 2010). A partir des index SS-GBLUP (comme pour le GBLUP), il est en outre possible d'estimer les effets des marqueurs et de fournir des équations de prédiction qui peuvent ensuite être appliquées à tout individu génotypé (non phénotypé) pour lui prédire une valeur génomique.

Les équations de prédiction des valeurs génomiques sont estimées à partir des animaux qui ont à la fois des phénotypes et des génotypes (population de référence). La précision des valeurs génomiques augmente avec i) l'héritabilité du caractère, ii) la taille de la population de référence, iii) le niveau d'apparentement entre les animaux de la population de référence et les candidats génotypés et iv) lorsque l'effectif génétique de la population diminue. La précision est définie par la corrélation (ou par son carré R^2 également appelé coefficient de détermination **CD**) entre la valeur génomique estimée et la valeur génétique vraie. La valeur génétique vraie n'étant pas connue, on estime généralement la précision de manière empirique dans une population de validation à partir de populations d'apprentissage / validation ou par validation croisée (voir précisions des équations MIR §1.3.2.4) en calculant la corrélation entre les valeurs génomiques prédites et les phénotypes réalisés, corrigés des effets de milieu. Le CD peut être estimé par le carré de cette corrélation divisé par l'héritabilité du caractère.

Goddard et Hayes (2009) ont proposé une formule analytique pour estimer la précision théorique des valeurs génomiques (GEBV = *Genomic Estimation of Breeding Values*) en fonction de la taille de la population de référence, de l'héritabilité du caractère et de l'effectif génétique de la population (***Figure 1.15***). Connaissant l'héritabilité du caractère et la taille de la population de référence, cette figure permet d'évaluer la précision attendue d'une évaluation génomique. Pour des caractères d'héritabilité moyenne, le projet *From'MIR*, qui offre une

première population de référence déjà conséquente, *i.e.* environ 20 000 vaches de race Montbéliarde, devrait donc nous permettre d'estimer des valeurs génomiques avec une précision relativement bonne (corrélation $R = \sqrt{CD}$ de l'ordre de 0,7 - 0,8, soit environ un CD de 0,6).

Pour un effectif génétique de 100, d'après Goddard et Hayes (2009)

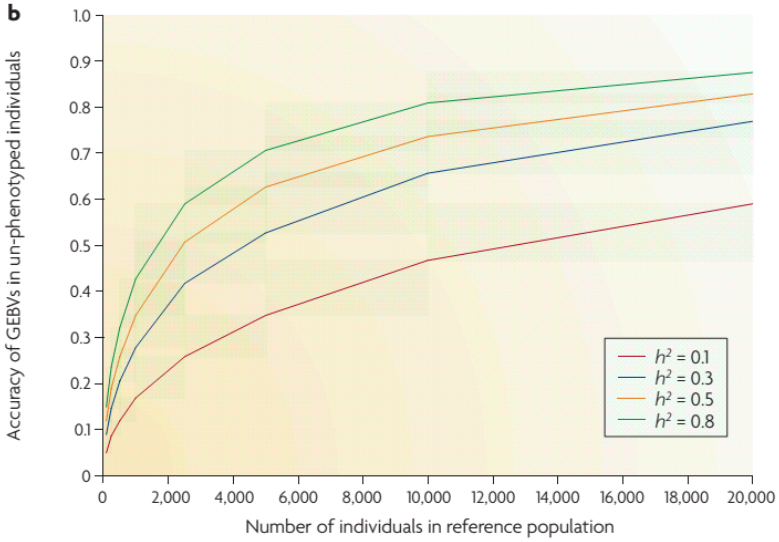


Figure 1.15. *Précision théorique (R) des valeurs génomiques (GEBV) en fonction de la taille de la population de référence et de l'héritabilité du caractère*

Chapitre 2

Du spectre MIR au phénotype

2. Du spectre MIR au phénotype

De longue date, la spectrométrie MIR du lait est utilisée en routine pour prédire les teneurs en matière grasse, matière protéique et lactose (Biggs, 1978). Au cours des années 2000, il a été montré que l'analyse des spectres MIR avec des méthodes appropriées permettait d'extraire beaucoup plus d'information et de prédire des caractéristiques nombreuses. Ainsi les spectres MIR donnent accès à des caractères de composition fine, comme les teneurs des principaux acides gras (AG), protéines ou minéraux, aussi bien qu'à des caractères complexes, tels que les aptitudes fromagères du lait (rendements fromagers, critères de coagulation et d'acidification).

En effet, dès 2006, Soyeurt *et al.* (2006) ont montré qu'il était possible de prédire la composition en AG, exprimée en pourcentage de lait ou en pourcentage de matière grasse totale du lait, à partir des spectres MIR. Quelques années plus tard, des équations ont été développées pour prédire les teneurs en protéines (De Marchi *et al.*, 2009a) et minéraux (Soyeurt *et al.*, 2009) ainsi que les aptitudes à la coagulation ou à l'acidification du lait (Dal Zotto *et al.*, 2008, De Marchi *et al.*, 2009b) et les rendements fromagers (Ferragina *et al.*, 2013). Dès les premières études, les qualités des prédictions étaient relativement bonnes pour les principaux AG (C14:0, C16:0 et C18:1 notamment) et minéraux (calcium et phosphore notamment) et les rendements fromagers tandis que les protéines et les aptitudes à la coagulation et à l'acidification du lait étaient prédites avec des précisions plus modérées. Plusieurs études plus récentes, utilisant des populations de calibration plus grandes ou des méthodes statistiques plus élaborées, ont permis d'améliorer les précisions des équations des composants fins du lait (Rutten *et al.*, 2009, Bonfatti *et al.*, 2011b, Soyeurt *et al.*, 2011, Ferrand-Calmels *et al.*, 2014, Toffanin *et al.*, 2015a, Bonfatti *et al.*, 2017a, Fleming *et al.*, 2017) et des aptitudes fromagères (Ferragina *et al.*, 2015, Visentin *et al.*, 2015). Toutefois, les qualités des prédictions des teneurs en protéines et des aptitudes à la coagulation / acidification du lait restent plus modérées que celles des rendements fromagers et des AG et minéraux du lait.

2.1. Prédiction de la composition protéique du lait dans le projet *PhénoFinlait*

Des équations de prédiction de la composition fine du lait ont été développées dans le projet *PhénoFinlait*. Au total, 450 échantillons de lait également répartis entre les races Montbéliarde, Normande et Holstein ont été analysés par spectrométrie MIR (MilkoScan FT6000, Foss Electric A/S, Hillerød, Danemark) et par chromatographie liquide couplée à la spectrométrie de masse (LC-MS) qui est la méthode de référence mise au point par l'INRA de Jouy-en-Josas

(Miranda *et al.*, 2011). Cette méthode est très résolutive puisqu'elle permet d'identifier et de quantifier chacune des lactoprotéines majeures du lait, ainsi que leurs variants génétiques et isoformes d'épissage, leurs modifications post-traductionnelles et certains produits de protéolyse. Les données ont été attribuées aléatoirement au jeu de calibration (70%) et de validation (30%). Après élimination des données aberrantes selon plusieurs critères, les jeux de calibration et de validation comprenaient respectivement 311 et 133 échantillons. Comme recommandé par le constructeur du spectromètre, seules les 446 longueurs d'onde non absorbées par les molécules d'eau ont été retenues. La méthode PLS couplée à un algorithme génétique pour sélectionner les longueurs d'onde les plus informatives a ensuite été appliquée (Ferrand *et al.*, 2012, Ferrand-Calmels *et al.*, 2014).

Des équations ont ainsi été développées pour prédire les teneurs dans le lait des six principales protéines du lait de vache, les quatre caséines (αs1, αs2, β et κ) et les deux protéines sériques (α-LA et β-LG). Les performances de ces équations (***Tableau 2.1***) sont bonnes, voire excellentes pour les caséines (R^2_{VAL} compris entre 0,82 et 0,92 et erreur relative comprise entre 3,7 et 8,4%) mais plus modérées pour les deux protéines sériques (R^2_{VAL} égal à 0,59 et 0,74 et erreur relative égale à 14,4 et 11,7% pour α-LA et β-LG respectivement).

Ces équations ont été appliquées sur 637 419 spectres MIR de 100 217 vaches Montbéliarde, 117 318 spectres MIR de 27 025 vaches Normande et 152 147 spectres MIR de 32 714 vaches Holstein collectés pour le projet *PhénoFinlait* (***Tableau 2.1***).

Tableau 2.1. *Teneurs en protéines (% lait) pour les trois races bovines du programme PhénoFinlait : moyennes ± écart-types et qualité de la prédiction MIR*

	Prédictions MIR (g/100g lait)			Equations *PhénoFinlait*	
Protéine	Montbéliarde (n=637 419)	Normande (n=117 318)	Holstein (n=152 147)	R^2_{VAL}*	Erreur relative (%)
TP	3,4±0,4	3,6±0,4	3,3±0,4	1	0,73
α-LA	0,14±0,02	0,15±0,02	0,14±0,02	0,59	14,4
β-LG	0,28±0,05	0,28±0,05	0,28±0,05	0,74	11,7
αs1-CN	0,94±0,10	0,99±0,10	0,92±0,11	0,88	4,7
αs2-CN	0,32±0,04	0,35±0,04	0,32±0,04	0,82	7,5
β-CN	1,24±0,11	1,29±0,11	1,20±0,13	0,92	3,7
κ-CN	0,33±0,05	0,35±0,05	0,31±0,05	0,80	8,4

***** R^2val = coefficient de détermination dans la population de validation

Les teneurs en protéines ont également été calculées en pourcentage de protéines totales en divisant les teneurs en protéines dans le lait par le taux protéique total de l'échantillon (***Tableau 2.2***).

Tableau 2.2. *Teneurs en protéines (% de protéines) pour les trois races bovines du programme PhénoFinlait : moyennes ± écart-types*

	Prédictions MIR (g/100g protéines)		
Protéine	Montbéliarde (n=637 419)	Normande (n=117 318)	Holstein (n=152 147)
α-LA	4,07±0,28	4,16±0,36	4,27±0,42
β-LG	8,25±1,12	7,94±1,03	8,46±1,17
αs1-CN	27,8±0,55	27,8±0,68	27,9±0,69
αs2-CN	9,53±0,30	9,89±0,33	9,69±0,39
β-CN	36,6±0,88	36,2±1,2	36,2±1,2
κ-CN	9,75±0,60	9,87±0,48	9,43±0,58

Tous les détails sur les animaux, les spectres MIR et les méthodes utilisés dans le programme *PhénoFinlait* pour prédire la composition protéique sont présentés dans deux articles qui figurent dans les chapitres 3 et 4 de ce manuscrit (Sanchez *et al.*, 2016a, Sanchez *et al.*, 2017a).

2.2. Prédiction de la fromageabilité du lait dans le projet *From'MIR*

2.2.1. Population de calibration : échantillons et analyses de référence

De la même façon, le programme *From'MIR* visait à établir des équations de prédiction pour les aptitudes fromagères du lait de vaches de race Montbéliarde en zone AOP Franche-Comté. Pour cela, des échantillons de lait individuel (250), de tanks de troupeaux (100) et de cuves fromagères (70) ont été recueillis et analysés par spectrométrie MIR (MilkoScan FT6000, Foss Electric A/S, Hillerød, Danemark) et par les méthodes de mesure de référence pour les aptitudes fromagères décrites dans le chapitre 1 (§1.2.5). L'échantillonnage des laits de mélange (troupeaux et cuves) a été réalisé en février-mars et mai-juin 2015 pour être aussi représentatif que possible de la variabilité des aptitudes fromagères en région Franche-Comté (localisation géographique, taille du troupeau, saison de vêlage et taux protéique annuel du troupeau lors de la campagne précédente). Les laits individuels ont été collectés entre janvier et juin 2016 sur des vaches à différents stades de lactation qui présentaient une large variété de génotypes aux lactoprotéines (Fang *et al.*, 2016). Sur les 250 échantillons de lait individuels recueillis, 99 provenaient de la traite du matin, 98 de la traite du soir et 53 étaient un mélange de laits prélevés en proportions relatives à la traite du matin et du soir. Quatre échantillons de laits individuels

avec comptage des cellules somatiques (CCS) > 10^6 ou avec plus de 10^5 UFC, ont été éliminés. Au total donc, 416 échantillons de lait ont été inclus dans les analyses de référence.

Les échantillons de lait frais, cru et entier ont été collectés par les techniciens des entreprises de conseil en élevage (ECEL) sous des conditions strictes d'hygiène, immédiatement refroidis à 4°C et analysés dans les 24h. Chaque échantillon a été séparé en deux aliquots. Le premier a été conservé dans du bronopol et analysé par spectrométrie MIR par le laboratoire d'analyse de CEL25-90 (ECEL du Doubs et du Territoire de Belfort). Le second a été utilisé pour mesurer les aptitudes fromagères avec les méthodes de référence par les laboratoires de l'INRA et d'ENILBio (Ecole Nationale d'Industrie Laitière et des Biotechnologies) localisés à Poligny dans le département du Jura. Avant mesure des rendements et des paramètres de coagulation, le lait a été standardisé pour le pH par ajout d'acide lactique. Les trois rendements fromagers de laboratoire ont été mesurés en duplicat selon la méthode de Hurtaud *et al.* (1995). Les paramètres de coagulation et d'acidification ont été mesurés pour deux types de technologie fromagère : pâte pressée cuite (PCC de l'anglais *Pressed Cooked Cheese*) et pâte molle (SC de l'anglais *Soft Cheese*). Chaque échantillon de lait (10 mL) a été coagulé par ajout de présure selon deux protocoles différents décrits dans le ***Tableau 2.3***.

Tableau 2.3. *Caractéristiques des protocoles appliqués pour les deux technologies fromagères*

		Pâte molle	Pâte pressée cuite
Standardisation pH	upH	6,45	6,6
Coagulation	Présure (éq. 100 L lait)	25 mL	14 mL
	Chymosine (éq. 1 L de lait)	520 mg	810 mg
	Nombre de paramètres mesurés	7	6
Acidification	Type de bactérie lactique utilisée pour l'ensemencement du lait	Mésophile	Thermophile
	Nombre de paramètres mesurés	5	3

Les paramètres de coagulation ont été mesurés après 30 min à 32°C avec un instrument Formoptic (Foss Electric A/S, Hillerød, Danemark). La décroissance du pH a été enregistrée pendant 20h avec le système CINAC (Corrieu *et al.*, 1988). Les différentes méthodes de mesures et les paramètres mesurés sont détaillés dans le chapitre 1 de ce manuscrit (§1.2.5). Au total, 24 paramètres fromagers, décrits dans le ***Tableau 2.4***, ont été mesurés.

Tableau 2.4. *Noms et descriptions des paramètres fromagers mesurés pour les deux types de technologie pâte pressée cuite (PCC) et pâte molle (SC)*

	Nom	Description
Rendements de laboratoire	CY_{FRESH}	Rendement frais (%)
	CY_{DM}	Rendement en extrait sec (%)
	$CY_{FAT-PROT}$	Rendement en matière sèche utile (%)
Paramètres de coagulation PCC	RCT_{PCC}	Temps de prise = temps nécessaire pour obtenir 0,5 IF* (min)
	$K10_{PCC}$	Temps pour obtenir 10 IF à partir du temps de prise (min)
	$K10/RCT_{PCC}$	Rapport entre K10 et RCT
	a_{PCC}	Fermeté du gel à une fois le temps de prise (IF)
	$TG10_{PCC}$	Pente de la courbe entre 7 et 13 IF
	$TG10/RCT_{PCC}$	Rapport entre Tg10 et RCT
Paramètres de coagulation SC	RCT_{SC}	Temps de prise = temps nécessaire pour obtenir 0,5 IF (min)
	$K10_{SC}$	Temps pour obtenir 10 IF à partir du temps de prise (min)
	$K10/RCT_{SC}$	Rapport entre K10 et RCT
	a_{SC}	Fermeté du gel à une fois le temps de prise (IF)
	$a2_{SC}$	Fermeté du gel à deux fois le temps de prise (IF)
	$TG10_{SC}$	Pente de la courbe entre 7 et 13 IF
	$TG10/RCT_{SC}$	Rapport entre Tg10 et RCT
Paramètres d'acidification PCC	$pH_{0\ PCC}$	pH du lait au moment de l'ensemencement (en unité pH, upH)
	$AR_{170\text{-}230min\ PCC}$	Taux d'acidification du lait entre 170 et 230 min après l'ensemencement (upH/h)
	$pH_{10h\ PCC}$	pH du lait 10h après l'ensemencement (upH)
Paramètres d'acidification SC	$pH_{0\ SC}$	pH du lait au moment de l'ensemencement (en unité pH, upH)
	$T\Delta pH_{0.08\ SC}$	Temps pour atteindre une chute de pH de 0,08 à partir de pH_0 (h)
	$\Delta pH_{16\text{-}20h\ SC}$	Variation du pH 16 à 20h après l'ensemencement (upH)
	$pH_{20h\ SC}$	pH du lait 20h après l'ensemencement (upH)
	$T\Delta pH_{6\text{-}4.8\ SC}$	Temps pour passer de pH 6 à pH 4,8 (h)

* IF = indice de fermeté

2.2.2. Spectres MIR

Les spectres MIR des 416 échantillons de lait ont été standardisés selon la méthode décrite par Grelet *et al.* (2015) et seules les 446 longueurs d'onde (**LO**) non absorbées par les molécules d'eau ont été retenues. Deux spectres MIR sont représentés sur la ***Figure 2.1***, ils ont été produits à partir de deux échantillons de lait individuel de la population de référence. Chaque spectre MIR quantifie l'absorbance d'un échantillon de lait sur la gamme de longueur d'onde du moyen infrarouge.

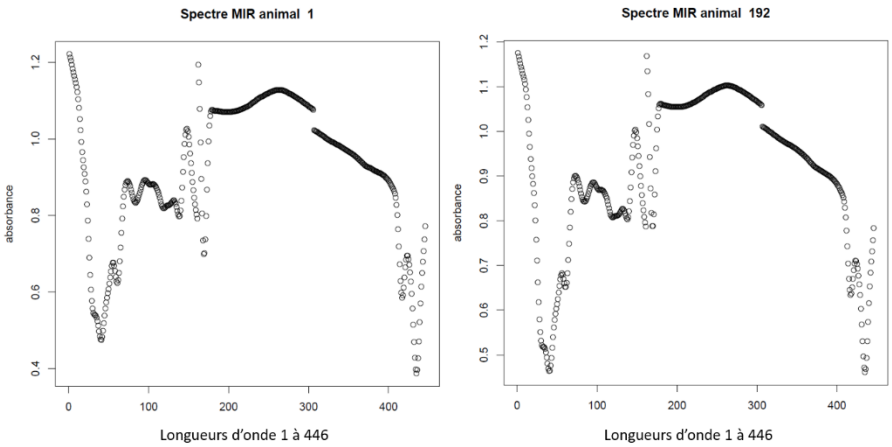


Figure 2.1. *Spectres MIR (absorbances pour 446 longueurs d'onde) du lait de deux individus*

La ***Figure 2.2*** représente la matrice des corrélations des absorbances calculées à partir des 246 spectres MIR mesurés sur les laits individuels. De nombreux blocs correspondant à des fenêtres de LO plus ou moins larges sont fortement corrélés, ce qui illustre les redondances entre absorbances au sein des spectres MIR.

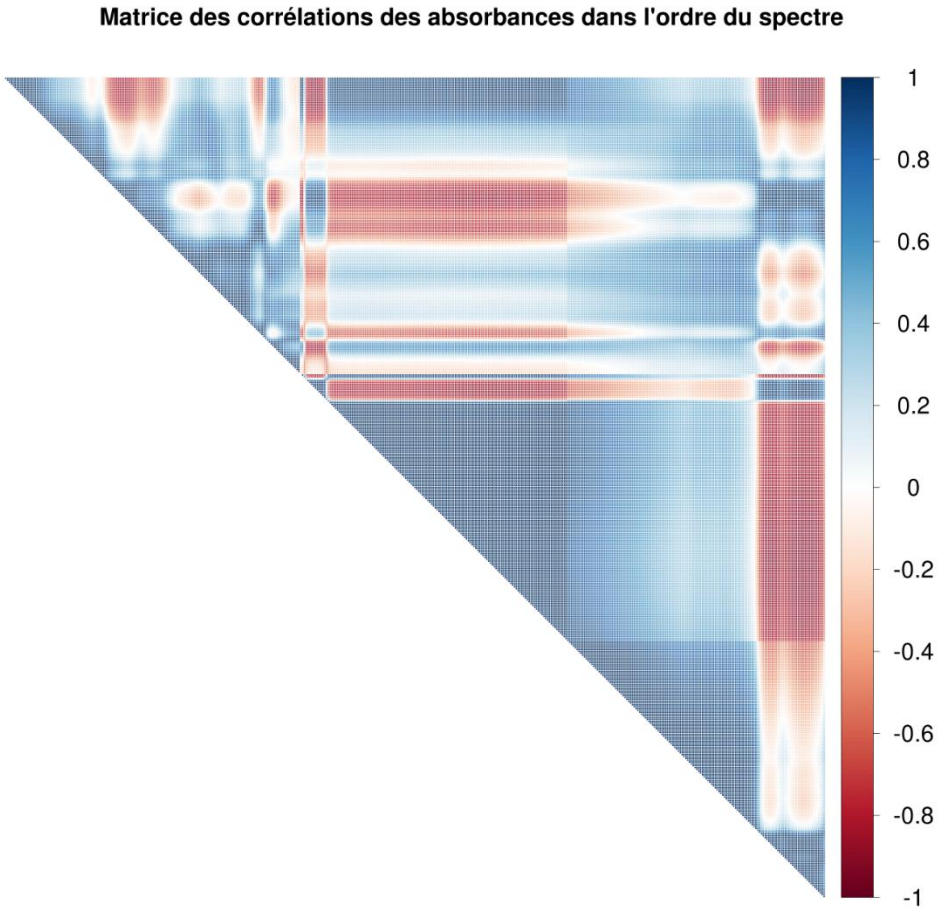


Figure 2.2. *Matrice des corrélations des absorbances calculées sur les 246 laits individuels*

Par ailleurs, une analyse en composantes principales (ACP), réalisée sur les mêmes spectres MIR (absorbances pour 446 longueurs d'onde des 246 laits individuels) révèle une forte structuration des variables. Les quatre premières valeurs propres de la matrice des corrélations expliquent plus de 97% de la variabilité entre les individus. On ne note en revanche aucune structuration des individus, ce qui suggère un bon échantillonnage de la population de référence (***Figure 2.3***).

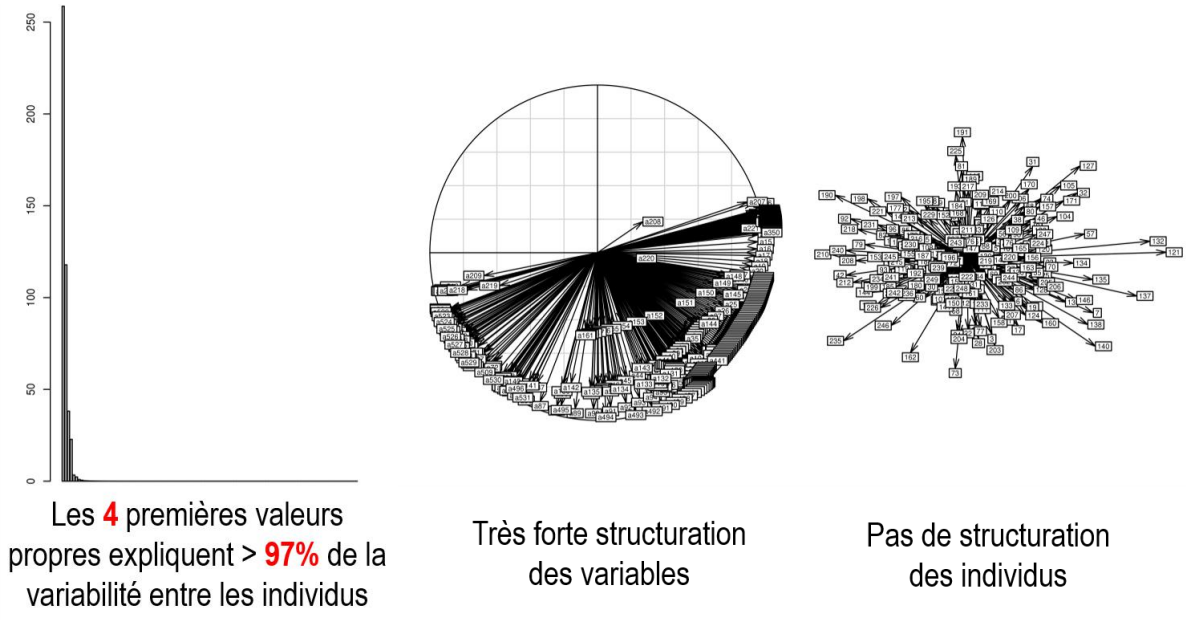


Figure 2.3. *Résultats d'une analyse en composantes principales (ACP) sur les spectres MIR (absorbances / 446 longueurs d'onde) des 246 laits individuels*

Ces résultats montrent l'intérêt d'utiliser des méthodes et modèles mathématiques qui réduisent la dimension du jeu de données.

2.2.3. Equations de prédiction : test des méthodes bayésiennes

2.2.3.1. Méthodes bayésiennes testées

Dans le cadre de ce travail de thèse, des équations de prédiction ont été développées à partir de méthodes bayésiennes et les précisions des prédictions ont été comparées à celles obtenues à partir de méthodes PLS avec ou sans sélection de variables (El Jabri *et al.*, 2017, El Jabri *et al.*, 2019). Pour cette étude, seules les données individuelles de la population de référence ont été retenues, soit 246 échantillons de lait.

Quatre méthodes de régression bayésienne ont été testées. Elles diffèrent selon les distributions *a priori* des effets des variables (LO dans le cas des spectres). Les distributions *a priori* induisent deux types d'effets, d'une part une régression des estimations (*shrinkage*), d'autre part, pour certaines, une sélection des variables incluses dans le modèle.

Deux des méthodes testées, **Bayes RR** (*Ridge Regression*) et **Bayes A**, supposent une distribution *a priori* non nulle pour les effets de toutes les LO. Dans le Bayes RR, les effets sont supposés distribués dans une loi Normale de variance constante et sont donc régressés vers la moyenne 0 (toutes les LO ont un petit effet) et ce, d'autant plus que l'échantillon analysé est petit. La méthode Bayes A (Meuwissen *et al.*, 2001) suppose également que les effets sont distribués *a priori* dans une loi Normale, avec une variance spécifique à chaque LO. Cette distribution est équivalente à une loi de Student, avec des queues plus épaisses que la loi Normale. En conséquence, les effets sont moins régressés vers la moyenne qu'avec le Bayes RR.

Dans les deux autres méthodes, **Bayes B** (Meuwissen *et al.*, 2001) et **Bayes C** (Habier *et al.*, 2011), une proportion (1-π) des effets est supposée nulle et donc seule une proportion de LO (π) a un effet non nul, les deux méthodes différant par la distribution *a priori* des effets non nuls. Les deux distributions sont normales mais la méthode Bayes B utilise la même distribution que le Bayes A (variance spécifique de chaque LO) tandis que la variance des effets est constante pour toutes les LO dans la méthode Bayes C.

Toutes ces méthodes ont été comparées avec le package R *BGLR* implémenté par Perez et de los Campos (2014). Pour chacune des méthodes et chacun des caractères, 100 000 itérations ont été réalisées. Pour estimer la distribution *a posteriori*, les 10 000 premières itérations ont été éliminées (*burn in*) et une itération sur 100 a été retenue (*thin*). En parallèle, la méthode **PLS** seule et la méthode PLS couplée à deux méthodes de sélection de variables (**UVEPLS** pour *Uninformative Variable Elimination* + PLS et **RFPLS** pour *Random Forest* + PLS) ont été testées par El Jabri *et al.* (El Jabri *et al.*, 2017, El Jabri *et al.*, 2019).

Dans les deux cas (analyses bayésiennes et PLS), les précisions des équations développées ont été calculées par validation croisée d'ordre 10 (voir §1.3.2.4). Dix réplicas ont donc été réalisés avec pour chaque réplica, 222 laits dans la population d'apprentissage et 24 laits dans la population de validation. Le coefficient de détermination (R^2), qui est le carré de la corrélation

calculée entre les valeurs vraies et les valeurs prédites par les équations dans les populations de validation (24 x 10), a été utilisé pour comparer les méthodes bayésiennes entre elles et pour comparer les méthodes bayésiennes aux méthodes PLS.

2.2.3.2. Résultats : comparaison des précisions des équations

Les précisions des équations développées à partir des méthodes bayésiennes sont indiquées dans le ***Tableau 2.5*** pour les 24 paramètres fromagers.

Tableau 2.5. *Précisions des équations de prédiction (R^2_{CV} = R^2 calculé par validation croisée) développées avec les méthodes bayésiennes et la meilleure méthode PLS*

	Régression bayésienne				Régression PLS (El Jabri *et al.*, 2017)	
Critère fromager	Bayes RR	Bayes A	Bayes B	Bayes C	Méthode*	R^2_{CV}
CY_{FRESH}	**0,772**	0,768	0,751	0,769	RFPLS	0,857
CY_{DM}	**0,859**	0,856	0,841	0,856	UVEPLS	0,916
$CY_{FAT\text{-}PROT}$	0,413	**0,415**	0,414	0,408	RFPLS	0,587
RCT_{PCC}	0,043	0,036	0,048	**0,115**	PLS	0,417
$K10_{PCC}$	0,320	0,318	0,231	**0,336**	RFPLS	0,437
$K10/RCT_{PCC}$	0,586	0,583	0,566	**0,599**	UVEPLS	0,638
a_{PCC}	0,685	0,677	0,675	**0,689**	UVEPLS	0,787
$TG10_{PCC}$	0,402	0,403	0,283	**0,407**	RFPLS	0,598
$TG10/RCT_{PCC}$	0,178	0,167	0,133	**0,201**	UVEPLS	0,638
RCT_{SC}	0,112	0,111	0,083	**0,153**	UVEPLS	0,324
$K10_{SC}$	0,335	0,338	0,168	**0,344**	UVEPLS	0,361
$K10/RCT_{SC}$	0,558	0,557	0,527	**0,582**	UVEPLS	0,653
a_{SC}	0,676	0,666	0,656	**0,688**	UVEPLS	0,805
$a2_{SC}$	0,627	0,615	0,533	**0,644**	UVEPLS	0,741
$TG10_{SC}$	0,419	0,420	0,270	**0,437**	RFPLS	0,614
$TG10/RCT_{SC}$	0,168	0,158	0,079	**0,196**	RFPLS	0,406
$pH_{0\ PCC}$	0,099	0,082	0,078	**0,176**	UVEPLS	0,616
$AR_{170\text{-}230min\ PCC}$	0,177	0,164	0,068	**0,206**	UVEPLS	0,446
$pH_{10h\ PCC}$	0,091	0,080	0,057	**0,108**	UVEPLS	0,091
$pH_{0\ SC}$	0,040	0,032	0,032	**0,075**	UVEPLS	0,440
$T\Delta pH_{0.08\ SC}$	0,042	**0,064**	0,057	0,029	UVEPLS	0,097
$\Delta pH_{16\text{-}20h\ SC}$	0,134	0,112	0,024	**0,201**	UVEPLS	0,288
$pH_{20h\ SC}$	0,097	0,090	0,033	**0,119**	RFPLS	0,090
$T\Delta pH_{6\text{-}4.8\ SC}$	0,102	0,054	0,019	**0,191**	RFPLS	0,218

* PLS = Partial Least Square ; UVEPLS = Uninformative Variable Elimination + PLS ; RFPLS = Random Forest + PLS

Les valeurs des R^2 varient fortement selon le paramètre fromager considéré et globalement, les quatre méthodes bayésiennes utilisées donnent des résultats proches. Pour les rendements fromagers et un critère d'acidification, le Bayes RR (pour CY_{FRESH} et CY_{DM}) et le Bayes A

(pour $CY_{FAT\text{-}PROT}$ et $T\Delta pH_{0.08\ SC}$) donnent des résultats légèrement plus précis que le Bayes C. Pour tous les autres caractères, les équations de prédiction développées par la méthode Bayes C sont plus précises.

Les deux dernières colonnes du ***Tableau 2.5*** donnent les résultats obtenus avec la méthode PLS la plus performante (nom de la méthode et valeur du R^2) par El Jabri *et al.* (2017). Quelle que soit la méthode bayésienne utilisée, les résultats sont presque toujours meilleurs avec une des méthodes PLS (***Figure 2.4***). La différence de précision entre les deux types de méthodes est variable selon le paramètre analysé et parfois très forte (jusqu'à 44 points de R^2 pour $pH_{0\ PCC}$).

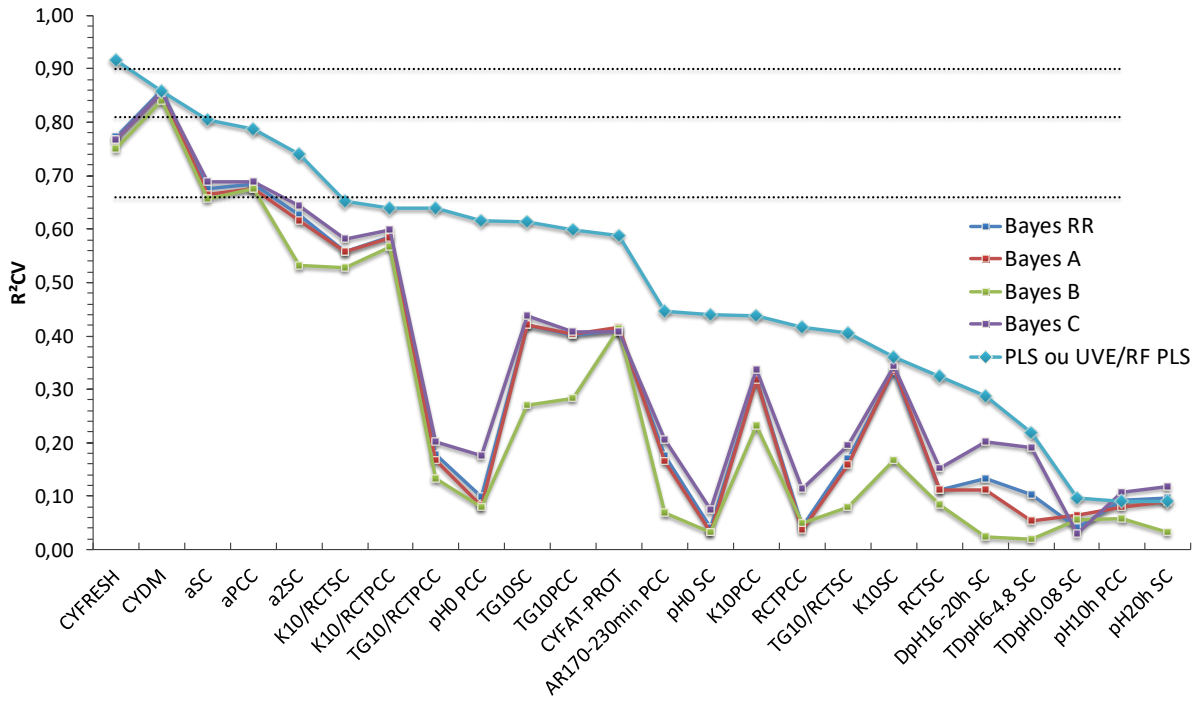


Figure 2.4. *R^2 calculé par validation croisée (R^2_{CV}) pour les équations de prédiction des 24 paramètres fromagers en fonction de la méthode utilisée*

Les méthodes bayésiennes ne permettent pas d'améliorer la précision des équations de prédiction sur le jeu de données *From'MIR*. Les méthodes PLS avec sélection de variables sont plus performantes pour tous les paramètres étudiés. Ces résultats, présentés dans une communication à un congrès (El Jabri *et al.*, 2017) et dans un article scientifique (El Jabri *et al.*, 2019) dont je suis deuxième auteur, sont différents de ceux obtenus par Ferragina *et al.* (2015). Ces auteurs, qui comparent différentes méthodes sur l'ensemble des 1060 LO du spectre, mettent en évidence la supériorité des méthodes bayésiennes sur les méthodes PLS pour les critères fromagers qu'ils ont mesurés (RCT et trois rendements) et surtout pour les

teneurs des principaux AG. Ils n'ont pas testé la méthode Bayes C mais par rapport aux deux méthodes PLS qu'ils ont utilisées, la méthode Bayes B améliore la qualité des prédictions : + 9 à 19 points de R^2 pour les AG, + 3 à 6 points pour les critères fromagers. Dans tous les cas, les valeurs de R^2 qu'ils obtiennent sont faibles à modérées : 0,28 à 0,71 pour les critères fromagers et 0,48 à 0,67 pour les AG. Une étude plus récente (Bonfatti *et al.*, 2017a) présente des résultats plus mitigés et plus cohérents avec ceux que nous obtenons dans *From'MIR*. Sur 92 caractères prédits avec différentes méthodes bayésiennes et PLS, les prédictions de 33 caractères seulement (essentiellement les teneurs en AG) sont plus précises avec les méthodes bayésiennes (Bayes C) qu'avec une méthode PLS optimisée. Dans la même étude, Bonfatti *et al.* (2017a) montrent une plus grande sensibilité des analyses PLS au pré-traitement des spectres et notamment à l'élimination des LO non informatives.

2.2.4. Précision des paramètres de fromageabilité du lait dans le projet *From'MIR*

Les équations de prédiction les plus précises sont donc celles développées à partir des méthodes PLS avec sélection de variables (RF ou UVE) pour tous les paramètres fromagers mesurés dans *From'MIR*. Néanmoins, le ***Tableau 2.5*** et la ***Figure 2.4*** montrent une grande hétérogénéité de la précision des prédictions MIR selon les paramètres étudiés. Les rendements et certains paramètres de coagulation sont bien prédits tandis que d'autres paramètres de coagulation, notamment le temps de prise (RCT), ainsi que la quasi-totalité des paramètres d'acidification sont très mal prédits à partir des spectres MIR. L'ajout des laits de mélange dans la population de référence, *i.e.* 100 laits de troupeaux et 70 laits de cuves, a permis d'améliorer sensiblement les performances des équations, en particulier pour les caractères les plus difficiles à prédire. Au final, les équations développées à l'aide des méthodes PLS sur 416 laits (246 individuels, 100 de troupeaux et 70 de cuves) ont donc été retenues et appliquées sur les spectres MIR collectées dans la région Franche-Comté.

Les caractéristiques des équations de prédiction des paramètres fromagers sont dans le ***Tableau 2.6***. Sur les 24 paramètres fromagers, neuf ont pu être prédits avec une précision modérée à bonne ($R^2_{VAL} \geq 0,54$ ou RPD $\geq 1,5$). Seuls ces 9 paramètres, en gras dans le ***Tableau 2.6***, ont été inclus dans la suite des analyses. A noter que les précisions ont été estimées dans une population de validation externe, ce qui donne des résultats plus rigoureux mais qui peuvent être sensiblement différents des résultats obtenus par validation croisée.

Tableau 2.6. *Performances des équations de prédiction développées sur 416 laits (246 individuels + 100 troupeaux + 70 cuves) pour les 24 paramètres fromagers*

	Population de calibration * (n = 291)			Population de validation (n = 125)		
Caractère	LV	R^2_{CAL}	$RMSE_{CAL}$	R^2_{VAL}	$RMSE_{VAL}$	RPD
CY_{FRESH}	**13**	**0,85**	**2,96**	**0,86**	**2,78**	**2,63**
CY_{DM}	**13**	**0,90**	**1,68**	**0,89**	**1,67**	**3,05**
$CY_{FAT\text{-}PROT}$	**18**	**0,64**	**13,49**	**0,54**	**14,50**	**1,47**
RCT_{PCC}	13	0,47	3,74	0,23	5,53	1,14
$K10_{PCC}$	10	0,51	2,64	0,37	3,05	1,25
$K10/RCT_{PCC}$	**12**	**0,74**	**0,052**	**0,62**	**0,06**	**1,63**
a_{PCC}	**12**	**0,79**	**1,16**	**0,72**	**1,44**	**1,88**
$TG10_{PCC}$	10	0,51	1,66	0,32	2,01	1,21
$TG10/RCT_{PCC}$	13	0,52	0,067	0,14	0,10	1,07
RCT_{SC}	10	0,37	2,33	0,22	2,47	1,13
$K10_{SC}$	13	0,50	1,24	0,35	1,37	1,23
$K10/RCT_{SC}$	**14**	**0,76**	**0,051**	**0,62**	**0,071**	**1,61**
a_{SC}	**13**	**0,80**	**1,14**	**0,73**	**1,52**	**1,92**
$a2_{SC}$	**13**	**0,76**	**1,01**	**0,64**	**1,37**	**1,67**
$TG10_{SC}$	11	0,51	3,08	0,44	3,21	1,33
$TG10/RCT_{SC}$	16	0,49	0,23	0,26	0,26	1,16
$pH_{0\ PCC}$	**17**	**0,68**	**0,030**	**0,65**	**0,035**	**1,68**
$AR_{170\text{-}230min\ PCC}$	15	0,56	0,024	0,30	0,031	1,19
$pH_{10h\ PCC}$	14	0,41	0,13	0,38	0,16	1,27
$pH_{0\ SC}$	16	0,60	0,041	0,46	0,059	1,35
$T\Delta pH_{0.08\ SC}$	9	0,15	0,29	0,17	0,29	1,10
$\Delta pH_{16\text{-}20h\ SC}$	17	0,42	0,10	0,18	0,11	1,10
$pH_{20h\ SC}$	6	0,17	0,075	0,04	0,074	1,01
$T\Delta pH_{6\text{-}4.8\ SC}$	18	0,51	1,01	0,27	1,21	1,16

***** Nombre de variables latentes (LV), coefficient de détermination (R^2), RMSE (*root mean squared error*) dans les populations de calibration (CAL) et validation (VAL) ; RPD (*Ratio Performance Deviation*) est le rapport entre l'écart-type calculé dans la population de calibration et l'erreur standard de la prédiction ; en gras, les neuf paramètres fromagers les plus précis.

Les précisions des équations de prédiction obtenues dans le projet *From'MIR* sont équivalentes, voire plus élevées que les précisions rapportées dans la littérature. Des études ont précédemment estimé des valeurs de R^2 comprises entre 0,67 et 0,85 pour CY_{FRESH} et CY_{DM} (Ferragina *et al.*, 2013, Bonfatti *et al.*, 2016, Grelet *et al.*, 2017), entre 0,20 et 0,52 pour la fermeté du caillé (Cecchinato *et al.*, 2009, Visentin *et al.*, 2015, Bonfatti *et al.*, 2016) et entre 0,71 et 0,79 pour le pH (Visentin *et al.*, 2015, Bonfatti *et al.*, 2016). En revanche, nous ne réussissons pas à prédire le temps de prise (RCT) alors que d'autres études rapportent des valeurs de R^2 comprises entre 0,55 et 0,69 pour ce caractère (Cecchinato *et al.*, 2009, Visentin *et al.*, 2015, Bonfatti *et al.*, 2016). Le pH du lait à l'emprésurage étant un des facteurs de variation les plus importants du RCT, la standardisation du pH du lait appliquée dans le projet

From'MIR pour réaliser les mesures des paramètres de coagulation a pu gommer une partie de la variabilité de ce caractère, le rendant ainsi difficile à prédire à partir des spectres MIR. Cependant, le temps de raffermissement du caillé peut être évalué par le paramètre K10/RCT (rapport entre K10, temps pour obtenir 10 IF, et RCT) qui est relativement bien prédit pour les deux types de fromages PCC et SC (R^2_{VAL}=0,62). Les performances des équations des mêmes paramètres de coagulation mesurés sur les deux types de fabrication sont équivalentes.

2.2.5. Spectres et prédictions MIR dans le projet *From'MIR*

Les équations de prédiction des neuf paramètres fromagers les mieux prédits ainsi que les équations de prédiction développées dans les projets *PhénoFinlait* (Ferrand *et al.*, 2012, Ferrand-Calmels *et al.*, 2014) pour la composition en protéines, *OptiMIR* (Gengler *et al.*, 2016) pour la composition en acides gras, minéraux et citrate ou par Foss (Foss Electric A/S, Hillerød, Danemark) pour le lactose ont été appliquées sur l'ensemble des spectres MIR collectés dans le projet *From'MIR*.

2.2.5.1. Spectres MIR collectés

Le jeu de données initial, résultant du stockage systématique des spectres par les entreprises de contrôle laitier depuis plusieurs années, comprenait 6 670 769 spectres MIR de 410 622 vaches Montbéliarde de la zone AOP Comté. Les échantillons de lait ont été prélevés dans le cadre du contrôle laitier entre janvier 2012 et juin 2017 et analysés par 10 spectromètres différents de type MilkoScan FT6000 (Foss Electric A/S, Hillerød, Danemark). Les spectres MIR produits ont été standardisés selon la procédure issue du projet *OptiMIR* (Grelet *et al.*, 2016).

2.2.5.2. Traitement des spectres MIR collectés

Le jeu de données a tout d'abord été fusionné avec la base de données nationale, afin de récupérer l'ensemble des variables associées aux contrôles élémentaires des vaches. Les données ont ensuite été filtrées pour un certain nombre de critères. Pour chaque vache, les contrôles réalisés en tout début (< 7 jours) et en fin de lactation (> 350 jours) ont été éliminés. Par souci d'homogénéité entre les vaches, seules les lactations complètes ou les lactations en cours au moment des derniers prélèvements (avec au moins sept contrôles) ont été retenues. Cela a eu pour effet d'éliminer toutes les lactations démarrées après décembre 2016 ainsi que les lactations démarrées avant le début du stockage. Pour éliminer les données aberrantes de chaque caractère prédit, seules les données comprises dans l'intervalle moyenne ± 4 écart-types

ont été gardées. Les laits qui présentaient des valeurs aberrantes pour le TP ou le TB ont été écartés des analyses. Après l'application de tous ces filtres, le jeu de données *From'MIR* comprenait plus de 4,8 millions de données par contrôle pour 311 613 vaches provenant de 3229 élevages Franc-Comtois (1753 élevages du Doubs et du Territoire de Belfort, 865 élevages du Jura et 611 élevages de la Haute-Saône).

2.2.5.3. Prédictions de la composition fine du lait et des aptitudes fromagères

Les effectifs, moyennes et écart-types de 54 caractères prédits à partir des spectres MIR sont dans le ***Tableau 2.7*** pour les caractères fromagers et la composition en protéines du lait et dans le ***Tableau 2.8*** pour la composition en acides gras, minéraux, citrate et lactose du lait.

Tableau 2.7*. Prédictions des caractères fromagers et de la composition en protéines à partir des spectres MIR des vaches Montbéliarde du projet From'MIR*

Abréviation	Nom complet	N	Moyenne	Ecart-type
Caractères fromagers				
CY_{FRESH} (%)	Rendement frais	4 877 908	37,6	7,5
CY_{DM} (%)	Rendement en extrait sec	4 879 203	66,9	5,0
$CY_{FAT\text{-}PROT}$ (%)	Rendement en matière sèche utile	4 863 635	187,6	21,7
$K10/RCT_{PCC}$	1 / vitesse d'organisation du gel	4 887 419	0,37	0,10
a_{PCC} (IF)	Fermeté à une fois le temps de prise	4 879 751	19,1	2,6
$K10/RCT_{SC}$	1 / vitesse d'organisation du gel	4 887 420	0,36	0,11
a_{SC} (IF)	Fermeté à une fois le temps de prise	4 879 845	19,4	2,7
$a2_{SC}$ (IF)	Fermeté à 2 fois le temps de prise	4 879 569	23,2	2,1
$pH_{0\ PCC}$ (upH)	pH lait à l'ensemencement	4 879 165	6,53	0,07
Protéines (g/100g lait)				
TP	Taux protéique	4 887 430	3,33	0,32
α-LA	Alpha-lactalbumine	4 871 648	0,13	0,02
β-LG	Beta-lactoglobuline	4 872 454	0,40	0,07
αs1-CN	Caséine αs1	4 885 933	1,08	0,10
αs2-CN	Caséine αs2	4 880 855	0,32	0,04
β-CN	Caséine β	4 881 331	0,99	0,09
κ-CN	Caséine κ	4 876 744	0,30	0,04
Σ CN	Caséines totales	4 883 658	2,70	0,26
Σ PS	Protéines sériques totales	4 875 989	0,55	0,08
Protéines (g/100g matière protéique)				
α-LA	Alpha-lactalbumine	4 883 090	4,03	0,36
β-LG	Beta-lactoglobuline	4 884 738	12,1	1,6
αs1-CN	Caséine αs1	4 886 672	32,3	0,3
αs2-CN	Caséine αs2	4 870 990	9,7	0,3
β-CN	Caséine β	4 886 336	29,7	1,2
κ-CN	Caséine κ	4 883 445	8,9	0,4
Σ CN	Caséines totales	4 885 187	80,9	1,3
Σ PS	Protéines sériques totales	4 883 358	16,5	1,7

***Tableau 2.8**. Prédictions de la composition en acides gras, minéraux et citrate et lactose à partir des spectres MIR des vaches Montbéliarde du projet From'MIR*

Abréviation	Nom complet	N	Moyenne	Ecart-type
Acides gras (g/100g lait)				
TB	Taux butyreux	4 887 430	3,72	0,50
SAT	Acides gras saturés	4 878 673	2,63	0,42
MUNSAT	Acides gras monoinsaturés	4 859 459	0,96	0,20
UNSAT	Acides gras insaturés	4 859 582	1,09	0,22
PUNSAT	Acides gras polyinsaturés	4 876 336	0,12	0,02
C4-C10	Somme des C4 à C10	4 876 843	0,43	0,06
C4-C12	Somme des C4 à C12	4 875 228	0,54	0,09
C14:0	Acide myristique	4 877 533	0,42	0,09
C16:0	Acide palmitique	4 872 841	1,08	0,25
C18:1	Acide oléique	4 856 396	0,84	0,18
C18:0	Acide stéarique	4 871 644	0,38	0,11
Acides gras (g/100g matière grasse)				
SAT	Acides gras saturés	4 887 376	70,9	5,4
MUNSAT	Acides gras monoinsaturés	4 887 313	26,0	4,6
UNSAT	Acides gras insaturés	4 887 297	29,5	5,1
PUNSAT	Acides gras polyinsaturés	4 887 420	3,3	0,7
C4-C10	Somme des C4 à C10	4 886 389	11,7	1,2
C4-C12	Somme des C4 à C12	4 887 422	14,5	1,6
C14:0	Acide myristique	4 887 429	11,2	1,9
C16:0	Acide palmitique	4 887 400	29,0	4,9
C18:1	Acide oléique	4 887 397	22,7	4,5
C18:0	Acide stéarique	4 887 105	10,3	2,8
Minéraux (mg/kg lait)				
Na	Sodium	4 214 619	364	52
Ca	Calcium	4 227 803	1161	95
P	Phosphore	4 220 168	972	84
Mg	Magnésium	4 233 129	98	8
K	Potassium	4 222 655	1451	106
Autres composés (g/kg lait)				
Citrate	Citrate	2 661 348	0,8	0,15
Lactose	Lactose	4 330 295	48,7	2,3

La matrice des corrélations entre les neuf critères fromagers les mieux prédits (***Tableau 2.9***) révèle une corrélation proche de 1 entre les rendements fromagers frais et en extrait sec, ainsi qu'entre les paramètres de coagulation mesurés pour les fromages à pâte pressée cuite d'une part et les fromages à pâte molle d'autre part. On observe également une corrélation assez forte entre les rendements frais et en extrait sec d'une part et les paramètres de fermeté du caillé (a_{PCC}, a_{SC} et $a2_{SC}$) d'autre part. Nous voyons en outre que $CY_{FAT\text{-}PROT}$ (rendement en matière sèche utile) est corrélé négativement et assez fortement avec les autres rendements. K10/RCT (inverse de la vitesse d'organisation du gel) est aussi corrélé négativement mais de façon

beaucoup plus modéré avec les paramètres de fermeté du caillé. Enfin, le pH du lait est plutôt faiblement corrélé avec les rendements et les paramètres de coagulation, excepté avec K10/RCT. Il est important de noter pour la suite des analyses que toutes les corrélations présentées dans le ***Tableau 2.9*** sont favorables, y compris les négatives. Un lait plus fromageable se caractérise en effet par des valeurs plus élevées de CY_{FRESH}, CY_{DM}, a et a2 et des valeurs plus faibles de $CY_{FAT-PROT}$ (rendement en matière sèche utile *i.e.* inverse d'un rendement fromager) et K10/RCT (inverse de la vitesse de coagulation).

Tableau 2.9. *Matrice des corrélations entre les neuf critères fromagers les mieux prédits*

	CY_{DM}	$CY_{FAT-PROT}$	$K10/RCT_{PCC}$	a_{PCC}	$K10/RCT_{SC}$	a_{SC}	$a2_{SC}$	$pH_{0\ PCC}$
CY_{FRESH}	0,98	-0,76	0,06	0,64	0,09	0,61	0,56	-0,11
CY_{DM}		-0,73	0,05	0,64	0,08	0,6	0,54	-0,10
$CY_{FAT-PROT}$			0,16	-0,38	0,16	-0,35	-0,36	0,26
$K10/RCT_{PCC}$				-0,11	0,98	-0,04	-0,07	0,56
a_{PCC}					-0,06	0,97	0,95	0,03
$K10/RCT_{SC}$						0,01	-0,03	0,54
a_{SC}							0,97	0,07
$a2_{SC}$								0,11

Nous présentons les courbes de lactation de l'ensemble des caractères dans le chapitre suivant.

2.3. Bilan du chapitre 2

Grâce à la collecte des spectres MIR et aux équations de prédiction développées dans les projets *PhénoFinlait* et *From'MIR*, nous avons à disposition des prédictions relativement précises pour un large panel de caractères mesuré sur un grand nombre de vaches (plusieurs contrôles par vache) :

1) la composition en protéines du lait dans les trois races Montbéliarde, Normande et Holstein (906 884 contrôles de 159 956 vaches) ;
2) la fromageabilité et la composition en protéines, acides gras, minéraux, citrate et lactose pour des vaches Montbéliarde de la région Franche-Comté (4,8 millions de contrôles de 311 613 vaches).

Combinés aux bases de données nationales (généalogies, données des contrôles élémentaires, génotypages…), ces jeux de données vont permettre une analyse fine et précise du déterminisme génétique de la composition et de la fromageabilité du lait dans l'espèce bovine.

Chapitre 3

Paramètres génétiques

3. Paramètres génétiques

Le lien phénotypique entre les aptitudes fromagères du lait et sa composition est connu depuis de nombreuses années, notamment avec les protéines (Wedholm *et al.*, 2006). Au cours de la dernière décennie, des études ont estimé les paramètres génétiques de la composition du lait pour les teneurs en protéines (Schopen *et al.*, 2009, Bonfatti *et al.*, 2011a, Gebreyesus *et al.*, 2016), acides gras (Garnsworthy *et al.*, 2010, Boichard *et al.*, 2014), minéraux (van Hulzen *et al.*, 2009, Buitenhuis *et al.*, 2015, Toffanin *et al.*, 2015b) et lactose (Haile-Mariam and Pryce, 2017). De même, les paramètres génétiques des aptitudes fromagères ont fait l'objet de plusieurs études, italiennes pour la plupart, pour les rendements fromagers (Bittante *et al.*, 2013, Cecchinato *et al.*, 2015, Cecchinato et Bittante, 2016, Bonfatti *et al.*, 2017c) et les paramètres de coagulation et d'acidification (Vallas *et al.*, 2010, Bonfatti *et al.*, 2011a, Cecchinato *et al.*, 2011, Bonfatti *et al.*, 2017c, Visentin *et al.*, 2017). Pour les deux types de caractères (composition et fromageabilité du lait), les paramètres génétiques ont été estimés à partir de mesures de référence ou de prédictions MIR. Les résultats sont très variables en fonction de la population étudiée et des méthodes de référence ou prédictions MIR utilisées. Cette grande variabilité reflète différentes situations : des jeux de données petits, en particulier avec les mesures de référence et des équations de prédiction MIR de précision variable. Les estimations des héritabilités sont comprises entre 0,09 et 0,41 pour les rendements fromagers, 0,12 et 0,75 pour les paramètres de coagulation et 0,06 et 0,37 pour les paramètres d'acidification (pH du lait essentiellement). Les corrélations génétiques entre les différents critères fromagers et les teneurs totales en protéines (TP) et matière grasse (TB) sont également très variables (de 0 à près de 1) selon le critère fromager et l'étude considérés (Vallas *et al.*, 2010, Bonfatti *et al.*, 2011a, Bittante *et al.*, 2013, Cecchinato *et al.*, 2015, Visentin *et al.*, 2017). En revanche, les liens génétiques entre les différents critères fromagers (rendements et paramètres de coagulation par exemple) et entre les critères fromagers et la composition fine du lait (teneurs individuelles des protéines, acides gras et minéraux) ont été peu étudiés.

Les objectifs de ce chapitre sont d'estimer les paramètres génétiques (héritabilités, répétabilités et corrélations génétiques entre caractères ou entre lactations) de la composition et des aptitudes fromagères du lait prédites à partir des spectres MIR. Plusieurs modèles, décrits au §1.6.2, ont été appliqués sur différents jeux de données pour répondre aux questions posées ci-après.

Q1 - Les caractères de composition et de fromageabilité du lait prédits par la spectrométrie MIR sont-ils héritables ? Q2 – Quels sont les liens génétiques entre les différents critères de fromageabilité, et entre fromageabilité et composition du lait ?

§3.1. Le modèle lactation (modèle 2) a été appliqué dans les trois races Montbéliarde, Normande et Holstein pour estimer les héritabilités et les corrélations génétiques des teneurs en protéines du lait en première lactation (***Article 1*** publié dans Journal of Dairy Science en 2017).

§3.2. Un modèle à répétabilité (modèle 3) a ensuite été utilisé pour estimer les héritabilités, répétabilités et corrélations génétiques des caractères de composition et de fromageabilité du lait en première lactation en race Montbéliarde (***Article 2*** publié dans Journal of Dairy Science en 2018).

Q3 - Le déterminisme génétique de la composition et de la fromageabilité du lait varie-t-il au cours de la lactation ?

§3.3. Sur le même jeu de données que celui utilisé au §3.2, nous avons appliqué un modèle de régression aléatoire (modèle 4) pour estimer les trajectoires des paramètres génétiques au cours de la première lactation.

Q4 - Les performances de composition et de fromageabilité du lait mesurées en première lactation sont-elles représentatives de la performance de la vache sur l'ensemble de sa carrière ?

§3.4. Les corrélations entre les trois premières lactations ont été estimées en race Montbéliarde à l'aide d'un modèle lactation (modèle 2) tri-caractères pour les caractères de composition et de fromageabilité du lait. Pour limiter les difficultés de calcul, ce modèle a été supposé à l'échelle de la lactation et non du contrôle.

Q5 - Existe-t-il des antagonismes génétiques entre la fromageabilité du lait et les caractères sélectionnés ?

§3.5. Les résultats des corrélations génétiques entre d'une part, les aptitudes fromagères du lait et d'autre part, les caractères actuellement sélectionnés sont finalement présentés en race Montbéliarde. Les estimations ont été obtenues par une série de modèles lactation (modèle 2) bi-caractère.

3.1. Paramètres génétiques de la composition protéique du lait

Short-communication: Genetic parameters for milk protein composition predicted using mid-infrared spectroscopy in the French Montbéliarde, Normande and Holstein dairy cattle breeds.

M. P. Sanchez[1*], M. Ferrand[†], M. Gelé[†], D. Pourchet[‡], G. Miranda[*], P. Martin[*], M. Brochard[†], D. Boichard[*]

[*] GABI, INRA, AgroParisTech, Université Paris Saclay, F-78350 Jouy-en-Josas, France

[†] Institut de l'Elevage, F-75012 Paris, France

[‡] ECEL Doubs - Territoire de Belfort, F-25640 Roulans, France

Journal of Dairy Science 2017. 100:6371-6375

https://doi.org/10.3168/jds.2017-12663

J. Dairy Sci. 100:6371–6375
https://doi.org/10.3168/jds.2017-12663
© American Dairy Science Association®, 2017.

Short communication: Genetic parameters for milk protein composition predicted using mid-infrared spectroscopy in the French Montbéliarde, Normande, and Holstein dairy cattle breeds

M. P. Sanchez,*[1] M. Ferrand,† M. Gelé,† D. Pourchet,‡ G. Miranda,* P. Martin,* M. Brochard,† and D. Boichard*
*Génétique Animale et Biologie Intégrative, INRA, AgroParisTech, Université Paris Saclay, F-78350 Jouy-en-Josas, France
†Institut de l'Elevage, F-75012 Paris, France
‡Conseil Elevage 25-90, F-25640 Roulans, France

ABSTRACT

Genetic parameters for the major milk proteins were estimated in the 3 main French dairy cattle breeds (i.e. Montbéliarde, Normande, and Holstein) as part of the PhénoFinlait program. The 6 major milk protein contents as well as the total protein content (PC) were estimated from mid-infrared spectrometry on 133,592 test-day milk samples from 20,434 cows in first lactation. Lactation means, expressed as a percentage of milk (protein contents) or of protein (protein fractions), were analyzed with an animal mixed model including fixed environmental effects (herd, year × month of calving, and spectrometer) and a random genetic effect. Genetic parameter estimates were very consistent across breeds. Heritability estimates (h^2) were generally higher for protein fractions than for protein contents. They were moderate to high for α_{S1}-casein, α_{S2}-casein, β-casein, κ-casein, and α-lactalbumin ($0.25 < h^2 < 0.72$). In each breed, β-lactoglobulin was the most heritable trait ($0.61 < h^2 < 0.86$). Genetic correlations (r_g) varied depending on how the percentage was expressed. The PC was strongly positively correlated with protein contents but almost genetically independent from protein fractions. Protein fractions were generally in opposition, except between κ-casein and α-lactalbumin ($0.39 < r_g < 0.46$) and κ-casein and α_{S2}-casein ($0.36 < r_g < 0.49$). Between protein contents, r_g estimates were positive, with highest values found between caseins ($0.83 < r_g < 0.98$). In the 3 breeds, β-lactoglobulin was negatively correlated with caseins ($-0.75 < r_g < -0.08$), in particular with κ-casein ($-0.75 < r_g < -0.55$). These results, obtained from a large panel of cows of the 3 main French dairy cattle breeds, show that routinely collected mid-infrared spectra could be used to modify milk protein composition by selection.

Received January 30, 2017.
Accepted April 29, 2017.
[1]Corresponding author: marie-pierre.sanchez@inra.fr

Key words: dairy cattle, mid-infrared spectrometry, protein composition, genetic parameters

Short Communication

In cattle, the relative proportions of proteins in milk play a key role in determining the functional properties of milk, such as clotting and cheese yield (Wedholm et al., 2006). Accurate genetic analyses of milk protein composition require large-scale studies, but reference methods such as capillary zone electrophoresis are time consuming and expensive. They have therefore only been applied to small or moderate numbers of milk samples. To date, the 2 most important studies aiming to estimate genetic parameters of milk protein composition traits measured by reference methods included 1,940 Dutch Holstein-Friesian cows (Schopen et al., 2009) and 2,167 Simmental cows (Bonfatti et al., 2011a). More recently, Gebreyesus et al. (2016) used genomic relationships between 650 Danish Holstein cows to estimate genetic parameters for milk protein composition. Mid-infrared (**MIR**) spectrometry has been shown to be useful to predict milk protein composition (De Marchi et al., 2009; Bonfatti et al., 2011b; Rutten et al., 2011; Ferrand et al., 2012; Samore et al., 2012) and offers an alternative method for large-scale analyses.

PhénoFinlait, a major project implemented to study milk in dairy cattle, sheep, and goats (Gelé et al., 2014) aimed, among other objectives, to dissect the genetic architecture of individual milk protein composition. In cattle, MIR predictive equations were derived from 450 reference samples analyzed using reverse-phase (**RP**) HPLC. The equations were applied to the MIR spectra routinely collected in Montbéliarde (**MO**), Normande (**NO**), and Holstein (**HO**) French dairy breeds (Ferrand et al., 2012). Concentrations of the 6 major milk proteins (α_{S1}-CN, α_{S2}-CN, β-CN, κ-CN, α-LA, and β-LG) were predicted with satisfactory accuracy (Sanchez et al., 2016). A genetic analysis of milk protein

Table 1. Milk protein composition: accuracy of mid-infrared (MIR) predictions (g/100 g of milk) and means ± SD as a percentage of milk or as a percentage of proteins in the Montbéliarde (MO), Normande (NO), and Holstein (HO) breeds

	Accuracy of MIR predictions[1]				g/100 g of milk			g/100 g of protein		
Trait	R^2_{val}	RE	RMSEP	RPD	MO	NO	HO	MO	NO	HO
PC[2]	1.00	0.73	0.025	14.1	3.4 ± 0.4	3.6 ± 0.4	3.3 ± 0.4	—	—	—
α-LA	0.59	14.4	0.020	1.6	0.14 ± 0.02	0.15 ± 0.02	0.14 ± 0.02	4.07 ± 0.28	4.16 ± 0.36	4.27 ± 0.42
β-LG	0.74	11.7	0.044	2.0	0.28 ± 0.05	0.28 ± 0.05	0.28 ± 0.05	8.25 ± 1.12	7.94 ± 1.03	8.46 ± 1.17
α_{S1}-CN	0.88	4.7	0.046	2.9	0.94 ± 0.10	0.99 ± 0.10	0.92 ± 0.11	27.8 ± 0.55	27.8 ± 0.68	27.9 ± 0.69
α_{S2}-CN	0.82	7.5	0.024	2.4	0.32 ± 0.04	0.35 ± 0.04	0.32 ± 0.04	9.53 ± 0.30	9.89 ± 0.33	9.69 ± 0.39
β-CN	0.92	3.7	0.044	3.5	1.24 ± 0.11	1.29 ± 0.11	1.20 ± 0.13	36.6 ± 0.88	36.2 ± 1.2	36.2 ± 1.2
κ-CN	0.80	8.4	0.038	2.2	0.33 ± 0.05	0.35 ± 0.05	0.31 ± 0.05	9.75 ± 0.60	9.87 ± 0.48	9.43 ± 0.58

[1]R^2_{val} = coefficient of determination; RE = relative error; RMSEP = root mean squared error of prediction; and RPD = ratio of prediction to deviation, calculated on MIR predictions in the validation set (n = 133) as g/100 g of milk.
[2]PC = total milk protein.

composition was therefore carried out in the 3 French dairy cattle breeds using a very large data set and for the first time in MO and NO breeds.

We herein report the estimation of genetic parameters for the 6 major milk proteins, using 133,592 test-day records in first lactation from 8,477 MO, 6,253 NO, and 5,734 HO cows.

The MIR spectra of 848,068 milk samples from 156,660 MO, NO, and HO cows were collected between November 2009 and August 2012 during the Phéno-Finlait program using MIR spectrometry with defined routine Fourier transform MIR analyses (MilkoScan FT6000, Foss Electric A/S, Hillerød, Denmark). The samples were distributed across 1,043 herds, covering a broad range of geographical locations (16 small regions) and production systems (grass or maize silage, high- or low-input, conventional or organic, and so on).

A total of 450 cow milk reference samples of the 3 breeds were analyzed using RP-HPLC. Equations were derived from these samples to predict total milk protein content (**PC**) and milk protein composition (Ferrand et al., 2012). Outliers were removed from reference data using the Grubbs test (Grubbs, 1969). Samples were then randomly assigned to either the calibration (70%) or validation (30%) set. Only the wavelengths not spoiled by water molecules were used (i.e., 446 wavelengths following the recommendations of the MilkoScan FT600 manufacturer), and the most informative wavelengths were selected using genetic algorithms (Ferrand-Calmels et al., 2014). Moreover, to improve the robustness of equations, calibration samples with a studentized residual greater than 2.58 were considered as outliers and deleted. Final calibration and validation sets contained 311 and 133 samples, respectively. Individual protein contents were predicted for the 6 main milk proteins: α-LA and β-LG whey proteins; and α_{S1}-CN, α_{S2}-CN, β-CN, and κ-CN, and expressed as grams per 100 g of milk (protein contents). Individual protein fractions, as grams per 100 g of protein, were then calculated by dividing predicted protein contents by PC. Means and standard deviations, as well as prediction accuracies obtained for milk protein contents of the validation set, are detailed in Table 1. As expected, PC was well predicted ($R^2 = 1$). The accuracy of content traits was high for caseins ($0.80 \leq R^2 \leq 0.92$) and moderate for α-LA ($R^2 = 0.59$) and β-LG ($R^2 = 0.74$). A total of 13 traits were therefore analyzed: PC, the 6 protein contents, and the 6 protein fractions.

For each trait, the phenotype of each cow was defined as the average test-day measures during the first lactation per cow. The variance components were estimated within-breed by REML with the procedure described by Meyer (1985) and implemented as in Boichard et al. (1989) with the following animal model:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Za} + \mathbf{e}, \quad [1]$$

where $\mathbf{y}$ is the vector of phenotypes; $\boldsymbol{\beta}$ a vector of fixed effects; $\mathbf{a} \sim N(0, \mathbf{A} \otimes \mathbf{G}_0)$ is the vector of random genetic effects; $\mathbf{e} \sim N(0, \mathbf{I} \otimes \mathbf{R}_0)$ is the vector of random residual effects. $\mathbf{X}$ and $\mathbf{Z}$ are incidence matrices common to all traits. $\mathbf{A}$ is the relationship matrix among individuals, $\mathbf{I}$ is the identity matrix, $\mathbf{G}_0$ is the 13 × 13 matrix of genetic covariances, and $\mathbf{R}_0$ is the matrix of residual covariances.

Only first lactation records with at least 7 test-day measurements in MO and at least 3 test-day measurements in NO and HO were included in the analyses. They corresponded to 72,561, 31,189, and 29,842 test-day record data from 8,477 MO, 6,253 NO, and 5,734 HO cows, respectively. Fixed effects included in the model were herd (944 in MO, 398 in NO, and 390 in HO), year × month of calving (12 in MO, 15 in NO, and 14 in HO), and spectrometer (1 in MO, 3 in NO, and 4 in HO). Pedigrees were traced over 3 generations and contained 23,956 individuals in MO, 17,376 in NO, and 15,895 in HO.

For all milk composition traits analyzed, heritability values were medium to high and generally higher for protein fractions than for protein contents (Table 2). Only one previous study compared heritability values of milk protein fractions and contents measured by RP-HPLC (Bonfatti et al., 2011a) and the authors also observed that protein fractions were more heritable than protein contents. For each trait, heritability estimates were similar in NO and HO and ranged from 0.27 to 0.72. Estimates were higher in MO (0.61–0.86) because of the higher test-day number in this breed (8.6) than in NO (5) and HO (5.2), resulting in a lower residual variance. In the 3 breeds, β-LG was the most heritable trait, with estimates ranging from 0.71 to 0.79 for β-LG fraction and from 0.61 to 0.86 for β-LG content. These results are consistent with those (0.80) obtained in Dutch Holstein Friesian as a percentage of protein (Schopen et al., 2009) and higher than those (0.54) found in Danish Holstein as a percentage of protein (Gebreyesus et al., 2016) or in Simmental (Bonfatti et al., 2011a) as a percentage of protein (0.34) and as a percentage of milk (0.37). Generally, heritability values obtained in our study from MIR spectra were similar or even higher than values estimated from reference analyses, except for α_{S2}-CN. For this casein, Schopen et al. (2009) have found a high heritability value (0.73), whereas in our study, estimations were weak to moderate (from 0.25 to 0.58, depending on the breed) but comparable or higher than values found by Bonfatti et al. (2011a) or Gebreyesus et al. (2016). It should be emphasized, however, that estimates based on reference measures were obtained with one unique measure, whereas our results are on a lactation level and used several measures per lactation. The loss of accuracy due to MIR prediction is compensated for by repeated measures.

Genetic correlations were similar across breeds (Table 3). Phenotypic correlations, not shown here, were comparable to the genetic correlations. As expected, PC was strongly genetically correlated with protein contents (g/100 g of milk). Higher genetic correlations were observed between PC and casein contents than between PC and whey protein contents. Conversely, PC was relatively genetically independent from protein fractions (g/100 g of protein). Only a moderate negative genetic correlation, ranging from −0.34 to −0.42 according to the breed, was observed between PC and α-LA. Protein contents were always positively correlated. Highest estimates were obtained between caseins (0.83–0.98), and the smallest estimates were found between β-LG and the other proteins. In contrast, protein fractions showed genetic correlations close to zero or negative in most cases. Only κ-CN was positively correlated with α-LA (0.39–0.46) and α_{S2}-CN (0.36–0.49). In the 3 breeds, β-LG was negatively correlated with all caseins, moderately with α_{S1}-CN, α_{S2}-CN, and β-CN (from −0.08 to −0.33), and more strongly with κ-CN (from −0.55 to −0.75). These results suggest that β-LG could share genetic regulatory pathways with caseins and in particular with κ-CN. Bonfatti et al. (2011a) consistently found positive genetic correlations between protein contents and negative ones between protein fractions. However, with MIR predictions, we generally found stronger genetic correlations than those reported by Bonfatti et al. (2011a) with milk protein composition measured by RP-HPLC. It should also be emphasized that protein fractions sum to 1 and this automatically tends to generate negative covariances between the main components.

Our results show that genetic parameters for milk protein composition predicted from MIR spectra were globally consistent with previously reported results from reference methods. We have found medium to high heritabilities for all protein contents/fractions with maximal values obtained for β-LG in spite of its medium MIR prediction accuracy ($R^2 = 0.74$). The MIR prediction accuracies are therefore high enough to undertake genetic investigations. In addition, this study, carried out at a large scale in the 3 main French dairy cattle breeds, shows that (1) the total milk protein content is relatively genetically independent from whey protein and casein fractions, (2) β-LG is highly heritable, and (3) β-LG and casein contents are positively correlated but the corresponding fractions exhibit negative genetic

Table 2. Heritability[1] values for milk protein contents (g/100 g of milk) and fractions (g/100 g of protein) in the Montbéliarde (MO), Normande (NO), and Holstein (HO) breeds

Item	Breed	PC[2]	α-LA	β-LG	α_{S1}-CN	α_{S2}-CN	β-CN	κ-CN
g/100 g of milk	MO	0.57	0.57	0.86	0.54	0.54	0.66	0.48
	NO	0.41	0.42	0.61	0.42	0.38	0.43	0.42
	HO	0.27	0.31	0.61	0.30	0.29	0.27	0.32
g/100 g of protein	MO	—	0.72	0.79	0.67	0.58	0.42	0.61
	NO	—	0.53	0.72	0.57	0.25	0.39	0.55
	HO	—	0.44	0.71	0.53	0.31	0.39	0.54

[1]Standard errors ranged from 0.02 to 0.03.

[2]PC = total milk protein.

Table 3. Genetic correlation values[1] for milk protein contents (g/100 g of milk, above diagonal) and fractions (g/100 g of protein; below diagonal) in the Montbéliarde (MO), Normande (NO), and Holstein (HO) breeds

Item	Breed	PC[2]	α-LA	β-LG	α_{S1}-CN	α_{S2}-CN	β-CN	κ-CN
PC	MO		0.72**	0.52**	0.98***	0.97***	0.99***	0.85***
	NO		0.70**	0.61**	0.99***	0.99***	0.98***	0.90***
	HO		0.73**	0.45*	0.98***	0.98***	0.98***	0.87***
α-LA	MO	−0.34*		0.18	0.69**	0.72**	0.70**	0.80***
	NO	−0.42*		0.27*	0.71**	0.69**	0.68**	0.74**
	HO	−0.38*		−0.01	0.72**	0.71**	0.75**	0.80***
β-LG	MO	0.04	−0.27*		0.51**	0.46*	0.47*	0.10
	NO	0.10	−0.33*		0.59**	0.57**	0.58**	0.31*
	HO	−0.08	−0.52**		0.39*	0.39*	0.39*	0.07
α_{S1}-CN	MO	0.04	0.01	−0.13		0.94***	0.96***	0.84***
	NO	0.25*	0.11	−0.10		0.98***	0.95***	0.89***
	HO	−0.18	−0.01	−0.19		0.96***	0.94***	0.87***
α_{S2}-CN	MO	0.00	0.05	−0.08	−0.33*		0.95***	0.88***
	NO	−0.17	−0.04	−0.14	−0.07		0.96***	0.92***
	HO	−0.26*	−0.09	−0.22*	0.04		0.95***	0.88***
β-CN	MO	−0.08	0.06	−0.33*	−0.26*	−0.37*		0.83***
	NO	−0.09	0.01	−0.25*	−0.44*	−0.51**		0.85***
	HO	0.07	0.25*	−0.20*	−0.48*	−0.37*		0.84***
κ-CN	MO	0.03	0.45*	−0.75**	0.02	0.38*	−0.11	
	NO	0.06	0.39*	−0.55**	0.02	0.49*	−0.36*	
	HO	0.17	0.46*	−0.66**	0.14	0.36*	−0.31*	

[1]Absolute genetic correlation values ($|r_g|$) are indicated by asterisks as follows: ***($|r_g| \geq 0.80$), **($0.50 \leq |r_g| < 0.80$), or *($0.20 \leq |r_g| < 0.80$); SE ranged from 0.04 to 0.08.

[2]PC = total milk protein.

correlations. Several options appear to be possible for selection (e.g., increase concentration of all proteins; increase concentration of caseins; or increase the casein/total protein ratio, leading to a decrease in the proportion of β-LG). The PC is already included in the breeding objectives of the 3 breeds. Reorienting the selection effort to casein concentration would be sound and easy to implement, at least in the breeds with a dominant cheese orientation. The interest of increasing the casein fraction and, therefore, decreasing the β-LG fraction (which is the major and most variable whey protein) should be investigated in more detail. Whatever the definition of the breeding goal, applying prediction equations to routinely collected MIR spectra provides new opportunities to modify milk protein composition and select for better cheese-making abilities.

ACKNOWLEDGMENTS

The PhénoFinlait program has received financial support from Agence Nationale de la Recherche (ANR-08-GANI-034 Lactoscan), APIS-GENE, Compte d'Affectation Spéciale Dévelopment Agricole et Rural, Centre National Interprofessionnel de l'Economie Laitière, FranceAgriMer, France Génétique Elevage, and the French Ministry of Agriculture (all from Paris, France). The authors thank the breeders who participated in PhénoFinlait; colleagues from Institut de l'Elevage and INRA who designed and coordinated farm samples and data collection; the partners of the project, laboratories, manufacturers, and DHI organizations that provided data; and the members of the PhénoFinlait scientific committee who advised and managed this program.

REFERENCES

Boichard, D., N. Bouloc, G. Ricordeau, A. Piacere, and F. Barillet. 1989. Genetic-parameters for 1st lactation dairy traits in the Alpine and Saanen goat breeds. Genet. Sel. Evol. 21:205–215. https://doi.org/10.1186/1297-9686-21-2-205.

Bonfatti, V., A. Cecchinato, L. Gallo, A. Blasco, and P. Carnier. 2011a. Genetic analysis of detailed milk protein composition and coagulation properties in Simmental cattle. J. Dairy Sci. 94:5183–5193. https://doi.org/10.3168/jds.2011-4297.

Bonfatti, V., G. Di Martino, and P. Carnier. 2011b. Effectiveness of mid-infrared spectroscopy for the prediction of detailed protein composition and contents of protein genetic variants of individual milk of Simmental cows. J. Dairy Sci. 94:5776–5785. https://doi.org/10.3168/jds.2011-4401.

De Marchi, M., V. Bonfatti, A. Cecchinato, G. Di Martino, and P. Carnier. 2009. Prediction of protein composition of individual cow milk using mid-infrared spectroscopy. Ital. J. Anim. Sci. 8(S2):399–401. https://doi.org/10.4081/ijas.2009.s2.399.

Ferrand, M., G. Miranda, S. Guisnel, H. Larroque, O. Leray, F. Lahalle, M. Brochard, and P. Martin. 2012. Determination of protein composition in milk by mid-infrared spectrometry. Pages 41–45 in Proc. International Strategies and New Developments in Milk Analysis. VI ICAR Reference Laboratory Network Meeting, Cork, Ireland.

Ferrand-Calmels, M., I. Palhiere, M. Brochard, O. Leray, J. Astruc, M. Aurel, S. Barbey, F. Bouvier, P. Brunschwig, H. Caillatt, M. Douguet, F. Faucon-Lahalle, M. Gele, G. Thomas, J. Trommenschlager, and H. Larroque. 2014. Prediction of fatty acid profiles

in cow, ewe, and goat milk by mid-infrared spectrometry. J. Dairy Sci. 97:17–35. https://doi.org/10.3168/jds.2013-6648.

Gebreyesus, G., M. S. Lund, L. Janss, N. A. Poulsen, L. B. Larsen, H. Bovenhuis, and A. J. Buitenhuis. 2016. Short communication: Multi-trait estimation of genetic parameters for milk protein composition in the Danish Holstein. J. Dairy Sci. 99:2863–2866. https://doi.org/10.3168/jds.2015-10501.

Gelé, M., S. Minery, J. M. Astruc, P. Brunschwig, M. Ferrand, G. Lagriffoul, H. Larroque, J. Legarto, P. Martin, G. Miranda, I. Palhière, P. Trossat, and M. Brochard. 2014. Phénotypage et génotypage à grande échelle de la composition fine des laits dans les filières bovine, ovine et caprine. Prod. Anim. 27:255–268.

Grubbs, F. 1969. Procedures for detecting outlying observations in samples. Technometrics 11:1–21.

Meyer, K. 1985. Maximum-likelihood estimation of variance-components for a multivariate mixed model with equal design matrices. Biometrics 41:153–165.

Rutten, M., H. Bovenhuis, J. Heck, and J. van Arendonk. 2011. Predicting bovine milk protein composition based on Fourier transform infrared spectra. J. Dairy Sci. 94:5683–5690. https://doi.org/10.3168/jds.2011-4520.

Samore, A., F. Canavesi, A. Rossoni, and A. Bagnato. 2012. Genetics of casein content in Brown Swiss and Italian Holstein dairy cattle breeds. Ital. J. Anim. Sci. 11:196–202. https://doi.org/10.4081/ijas.2012.e36.

Sanchez, M., A. Govignon-Gion, M. Ferrand, M. Gele, D. Pourchet, Y. Amigues, S. Fritz, M. Boussaha, A. Capitan, D. Rocha, G. Miranda, P. Martin, M. Brochard, and D. Boichard. 2016. Whole-genome scan to detect quantitative trait loci associated with milk protein composition in 3 French dairy cattle breeds. J. Dairy Sci. 99:8203–8215. https://doi.org/10.3168/jds.2016-11437.

Schopen, G. C., J. M. Heck, H. Bovenhuis, M. H. Visker, H. J. van Valenberg, and J. A. van Arendonk. 2009. Genetic parameters for major milk proteins in Dutch Holstein-Friesians. J. Dairy Sci. 92:1182–1191. https://doi.org/10.3168/jds.2008-1281.

Wedholm, A., L. B. Larsen, H. Lindmark-Månsson, A. H. Karlsson, and A. Andrén. 2006. Effect of protein composition on the cheese-making properties of milk from individual dairy cows. J. Dairy Sci. 89:3296–3305. https://doi.org/10.3168/jds.S0022-0302(06)72366-9.

3.2. Paramètres génétiques de la fromageabilité et de la composition du lait

Genetic parameters for cheese-making properties and milk composition predicted from mid-infrared spectrometry in a large dataset of Montbéliarde cows.

M.P. Sanchez[1*], M. El Jabri[†], S. Minéry[†], V. Wolf [‡], E. Beuvier [#], C. Laithier[†], A. Delacroix-Buchet[*], M. Brochard[‖], D. Boichard[*]

[*] GABI, INRA, AgroParisTech, Université Paris Saclay, F-78350 Jouy-en-Josas, France

[†] Institut de l'Elevage, F-75012 Paris, France

[‡] ECEL Doubs - Territoire de Belfort, F-25640 Roulans, France

[#] URTAL, INRA, F-39800 Poligny, France

[‖] Umotest, F-01250 Ceyzériat, France

Journal of Dairy Science 2018. 101:10048–10061

https://doi.org/10.3168/jds.2018-14878

J. Dairy Sci. 101:10048–10061
https://doi.org/10.3168/jds.2018-14878
© American Dairy Science Association®, 2018.

Genetic parameters for cheese-making properties and milk composition predicted from mid-infrared spectra in a large data set of Montbéliarde cows

M. P. Sanchez,*[1] M. El Jabri,† S. Minéry,† V. Wolf,‡ E. Beuvier,§ C. Laithier,† A. Delacroix-Buchet,* M. Brochard,# and D. Boichard*
*GABI, L'Institut National de la Recherche Agronomique, AgroParisTech, Université Paris Saclay, F-78350 Jouy-en-Josas, France
†Institut de l'Elevage, F-75012 Paris, France
‡Entreprise de Conseil en Elevage Doubs—Territoire de Belfort, F-25640 Roulans, France
§URTAL, L'Institut National de la Recherche Agronomique, F-39800 Poligny, France
#Umotest, F-01250 Ceyzériat, France

ABSTRACT

Cheese-making properties of pressed cooked cheeses (PCC) and soft cheeses (SC) were predicted from mid-infrared (MIR) spectra. The traits that were best predicted by MIR spectra (as determined by comparison with reference measurements) were 3 measures of laboratory cheese yield, 5 coagulation traits, and 1 acidification trait for PCC (initial pH; $pH_{0\ PPC}$). Coefficients of determination of these traits ranged between 0.54 and 0.89. These 9 traits as well as milk composition traits (fatty acid, protein, mineral, lactose, and citrate content) were then predicted from 1,100,238 MIR spectra from 126,873 primiparous Montbéliarde cows. Using this data set, we estimated the corresponding genetic parameters of these traits by REML procedures. A univariate or bivariate repeatability animal model was used that included the fixed effects of herd × test day × spectrometer, stage of lactation, and year × month of calving as well as the random additive genetic, permanent environmental, and residual effects. Heritability estimates varied between 0.37 and 0.48 for the 9 cheese-making property traits analyzed. Coagulation traits were the ones with the highest heritability (0.42 to 0.48), whereas cheese yields and $pH_{0\ PPC}$ had the lowest heritability (0.37 to 0.39). Strong favorable genetic correlations, with absolute values between 0.64 and 0.97, were found between different measures of cheese yield, between coagulation traits, between cheese yields and coagulation traits, and between coagulation traits measured for PCC and SC. In contrast, the genetic correlations between milk $pH_{0\ PPC}$ and CY or coagulation traits were weak (−0.08 to 0.09). The genetic relationships between cheese-making property traits and milk composition were moderate to high. In particular, high levels of proteins, fatty acids, Ca, P, and Mg in milk were associated with better cheese yields and improved coagulation. Proteins in milk were strongly genetically correlated with coagulation traits and, to a lesser extent, with cheese yields, whereas fatty acids in milk were more genetically correlated with cheese yields than with coagulation traits. This study, carried out on a large scale in Montbéliarde cows, shows that MIR predictions of cheese yields and milk coagulation properties are sufficiently accurate to be used for genetic analyses. Cheese-making traits, as predicted from MIR spectra, are moderately heritable and could be integrated into breeding objectives without additional phenotyping cost, thus creating an opportunity for efficient improvement via selection.
Key words: Montbéliarde, cheese-making property, mid-infrared spectrometry, genetic parameter

INTRODUCTION

More than 36% of total cow milk is processed into cheese products (International Dairy Federation, 2016), and this proportion has increased by 23% during the last decade. The milk processing industry thus stands to gain a great deal economically from the improvement of milk cheese-making properties (**CMP**). To assess CMP, different laboratory methods have been developed (reviewed in Bittante et al., 2012), but they are costly and time consuming and thus difficult to implement on a large scale. Mid-infrared (**MIR**) spectrometry has been proposed as an alternative method for prediction of various milk characteristics, including fractions of proteins, fatty acids, or minerals as well as cheese yield or coagulation properties (De Marchi et al., 2014). Mid-infrared technology is cheap and routinely used. Consequently, CMP can be measured on a large scale and, therefore, included in breeding goals for dairy cattle. However, before this approach is

Received April 4, 2018.
Accepted July 13, 2018.
[1] Corresponding author: marie-pierre.sanchez@inra.fr

generalized, it is necessary to better understand the genetic background of these traits. During the last 10 yr, several studies have estimated the genetic parameters of CMP, with data sets comprising from 892 to 17,577 reference measurements (Vallas et al., 2010; Cecchinato et al., 2011, 2015; Bittante et al., 2013; Poulsen et al., 2015) or 136,807 to 311,354 MIR predictions (Cecchinato et al., 2015; Bonfatti et al., 2017; Visentin et al., 2017). Using a micro-cheese-making procedure, some Italian studies reported low to moderate heritability for cheese yields (0.09 to 0.41; Bittante et al., 2013; Cecchinato et al., 2015; Cecchinato and Bittante, 2016; Bonfatti et al., 2017). In most other studies, coagulation properties were quantified with mechanical or optical lactodynamograph measurements after the addition of rennet. The traits most commonly analyzed were rennet coagulation time (**RCT**), time to curd firmness of 20 mm, and curd firmness 30, 45, and 60 min after chymosin addition. The heritability of some of these traits was found to be high, whereas it was low for others (heritabilities from 0.12 to 0.75). Heritability estimates of milk acidification, as assessed by milk pH measurements, were low to moderate (0.06 to 0.37). Most of these studies investigated the genetic correlations between various CMP traits as well as between CMP traits and total milk protein and fat content, but so far none have examined the genetic relationships between fine-scale milk composition traits (e.g., levels of individual proteins, fatty acids, minerals) and CMP.

The FROM'MIR project, initiated in 2015, aims to analyze CMP and milk composition traits predicted from MIR spectra in Montbéliarde cows from the Franche-Comté region; the milk from these cows is used to produce several cheeses that bear appellations of a protected designation of origin or a protected geographical indication. A set of 24 CMP was measured on 416 milk samples in which milk had been coagulated and then processed for making either soft cheese (**SC**) or pressed cooked cheese (**PCC**). Cheese-making traits included 3 laboratory cheese yields, 13 coagulation traits measured with a Formoptic instrument adapted from a mechanical Formagraph (Foss Electric, Hillerød, Denmark) by Chr. Hansen (Horsholm, Denmark) and ENILBio (Poligny, France), and 8 acidification traits measured with the Cinac system (Corrieu et al., 1988). Prediction equations were developed (El Jabri et al., 2017; Laithier et al., 2017) and applied to more than 6 million MIR spectra obtained from more than 400,000 Montbéliarde cows. To estimate the genetic relationships between CMP and milk composition, the content in proteins, fatty acids, minerals, citrate, and lactose was also predicted from MIR spectra using equations developed in previous projects—namely the French PhénoFinlait project (Ferrand et al., 2012; Ferrand-Calmels et al., 2014) and the European Optimir project (Gengler et al., 2016).

The objectives of the present study were to estimate the heritability and repeatability of the 9 CMP traits that were best predicted by MIR spectra (3 cheese yields, 5 coagulation traits, and 1 acidification trait) as well as that of milk composition traits (proteins, fatty acids, minerals, citrate, and lactose). Furthermore, we investigated the genetic relationships among different CMP and milk composition traits.

MATERIALS AND METHODS

Reference Analyses

The reference data set included 420 milk samples corresponding to 250 individual Montbéliarde cows, 100 herd tanks, and 70 dairy vats. The sampling was designed to maximize genetic diversity as well as variation in milk composition. To this end, herd and dairy vat milks were obtained from dairy farms that varied in geographical location, calving season, average annual herd protein content of the previous milking campaign, and herd size; samples were collected in February to March 2015 and May to June 2015. Individual milk samples were collected between January and June 2016 from cows at different stages of lactation, which represented a wide range of lactoprotein phenotypes (Fang et al., 2016). Of the 250 individual milk samples that were retained, 98 came from the morning milking, 99 came from the evening milking, and 53 were a mixture of 2 successive milkings in relative proportions. Milk samples were excluded if SCC were greater than 10^6 or if more than 10^5 cfu were present. Four individual samples were excluded. In total, 416 milk samples (246 from individual cows, 100 from herd tanks, and 70 from dairy vats) were included in reference analyses.

Whole fresh raw milk samples were collected under strict hygienic conditions, immediately cooled to 4°C, and analyzed within 24 h. Each sample was divided into 2 aliquots: one was preserved by bronopol and analyzed by MIR spectrometry, and the other was analyzed by reference methods of measuring milk coagulation, acidification, and curd yield.

Milk samples (10 mL) were standardized in pH with lactic acid and coagulated with rennet according to 1 of 2 protocols for SC or PCC. For SC, rennet was added at pH 6.45 at a dose of 25 mL/100 L of milk (equivalent to 520 mg of chymosin/L). For PCC, rennet was added at pH 6.60 at a dose of 14 mL/100 L of milk (810 mg of chymosin/L of milk). After 30 min at 32°C, coagulation properties were measured with a Formoptic device, which was adapted from a mechanical Formagraph (Foss Electric) by Chr. Hansen

and ENILBio to improve data computerization. As a Formagraph, the Formoptic registers the position of a pendulum immersed in milk to which rennet has been added. Only the way of recording the position of the pendulum axis is modified. With Formagraph, the position is recorded on a photosensitive piece of paper (the firmness measurements are made manually using a ruler) and is expressed in millimeters. In the case of the Formoptic, the position is automatically recorded via optical transmitters and receivers that convert the physical positions of the pendulum into voltages that are computerized and expressed as a firmness index (**FI**; FI = 10 × volts). The following coagulation traits were calculated:

1. RCT to 0.5 FI (in min) for SC ($\mathbf{RCT_{SC}}$) and PCC ($\mathbf{RCT_{PCC}}$),
2. curd firmness at RCT (in FI units) for SC ($\mathbf{a_{SC}}$) and PCC ($\mathbf{a_{PCC}}$),
3. curd firmness at 2 times RCT (in FI units) for SC ($\mathbf{a2_{SC}}$),
4. time to obtain 10 FI from RCT (in min) for SC ($K10_{SC}$) and PCC ($K10_{PCC}$),
5. ratio of K10 to RCT for SC ($\mathbf{K10/RCT_{SC}}$) and PCC ($\mathbf{K10/RCT_{PCC}}$),
6. curd organization speed (slope of the curve at 10 FI) for SC ($TG10_{SC}$) and PCC ($TG10_{PCC}$), and
7. ratio of TG10 to RCT for SC ($TG10/RCT_{SC}$) and PCC ($TG10/RCT_{PCC}$).

Laboratory curd yields (**CY**) were measured in duplicate using the method of Hurtaud et al. (1995). Coagulated milk (50 mL at pH 6.6 and 32°C) was centrifuged (2,700 × *g*, 15 min), curd and whey were weighed, and DM of whey and milk was measured. The following CY were calculated:

1. fresh CY ($\mathbf{CY_{FRESH}}$; %): CY_{FRESH} = 100 × (g of curd/g of milk).
2. CY in DM ($\mathbf{CY_{DM}}$; %): CY_{DM} = 100 × (1 − g of DM whey/g of DM milk).
3. CY in protein and fat ($\mathbf{CY_{FAT\text{-}PROT}}$; g/kg): $CY_{FAT\text{-}PROT}$ = (PC + FC) × (g of milk/g of curd). where **PC** = protein content and **FC** = fat content.

Note that CY_{FRESH} and CY_{DM} are expressed as a ratio of curd to processed milk, whereas $CY_{FAT\text{-}PROT}$ is expressed as a ratio of processed milk to curd. Therefore, these traits are expected to be inversely correlated.

Finally, acidification properties were assessed along a temperature gradient using the Cinac system (Corrieu et al., 1988). The decrease in pH was measured for 20 h after the addition of lactic bacteria (mesophilic or thermophilic strains for SC and PCC, respectively). Different cheese-specific traits were measured:

1. initial pH (in pH units; **upH**) for SC ($pH_{0\ SC}$) and PCC ($\mathbf{pH_{0\ PCC}}$),
2. acidification rate from 170 to 230 min (in upH/h),
3. pH at 10 h (in upH),
4. time to decrease by 0.08 upH from pH_0 (in h),
5. time from pH 6 to pH 4.8 (in h),
6. decrease in pH from 16 h to 20 h (in upH), and
7. pH at 20 h (in upH).

Mid-infrared spectra of the 416 milk samples were standardized by piecewise direct standardization, which matches slave-instrument spectra to master-instrument spectra (Grelet et al., 2015). Only the wavelengths not absorbed by water molecules were used (i.e., absorption was measured for 446 wavelengths following the recommendations of the manufacturer of the MilkoScan FT6000). El Jabri et al. (2017) and Laithier et al. (2017) compared different approaches [Bayesian or partial least squares (**PLS**) using random forest or uninformative variable elimination] for different data sets. They found the highest accuracy with PLS applied to the largest composite data set including individual, herd, and vat samples. An optimal number of latent variables was selected for each equation by minimizing the root mean squared error (**RMSE**) in cross-validation using 10 random segments. Coefficient of determination ($\mathbf{R^2}$) and RMSE were calculated in both calibration (70%) and validation (30%) sets randomly selected (Table 1). For all 24 CMP traits, these PLS equations were used for all further analyses, and all CMP traits had a relatively normal distribution. The accuracies of MIR predictions were estimated using the R^2 in the validation set. These values ranged from 0.04 to 0.89 depending on the CMP trait (Table 1). Of the 24 CMP traits, only 9 were predicted with relatively good accuracy (R^2 value >0.50): 3 CY traits; 5 coagulation traits, which included measurements of curd-firming time (K10/RCT for PCC and SC) and curd firmness (a for PCC and SC and a2 for SC); and 1 acidification trait for PCC ($pH_{0\ PPC}$). These 9 traits were thus retained for genetic parameter estimations, whereas all others were excluded from further analysis. It should be noted that MIR analysis generally failed to accurately predict the acidification process and succeeded only in predicting the initial pH.

MIR Spectra Predictions

The original data set of the present study comprised 6,670,769 milk samples from 410,622 Montbéliarde

Table 1. Descriptive statistics of mid-infrared (MIR) predictions and performance of prediction equations in the calibration (n = 291) and validation (n = 125) sets for cheese-making properties

		MIR predictions				Prediction equations[2]				
Item	Trait[1]	No.	Mean	SD	CV	LV	R^2_{cal}	$RMSE_{cal}$	R^2_{val}	$RMSE_{val}$
Cheese yields	CY_{FRESH}, %	1,098,841	38.3	6.8	17.8	13	0.85	2.96	*0.86*	2.78
	CY_{DM}, %	1,099,179	67.2	4.6	6.8	13	0.90	1.68	*0.89*	1.67
	$CY_{FAT\text{-}PROT}$, %	1,097,248	185.5	20.3	10.9	18	0.64	13.49	*0.54*	14.50
Coagulation traits for pressed cooked cheese	RCT, min	1,095,809	34.33	5.10	14.9	13	0.47	3.74	0.23	5.53
	K10, min	1,097,657	12.11	2.76	22.8	10	0.51	2.64	0.37	3.05
	K10/RCT	1,100,235	0.37	0.09	24.3	12	0.74	0.052	*0.62*	0.06
	a, FI	1,099,010	18.94	2.43	12.8	12	0.79	1.16	*0.72*	1.44
	TG10	1,097,904	7.49	1.89	25.2	10	0.51	1.66	0.32	2.01
	TG10/RCT	1,100,235	0.22	0.08	36.4	13	0.52	0.067	0.14	0.10
Coagulation traits for soft cheese	RCT, min	1,096,665	18.12	2.05	11.3	10	0.37	2.33	0.22	2.47
	K10, min	1,096,525	6.50	1.36	20.9	13	0.50	1.24	0.35	1.37
	K10/RCT	1,100,235	0.36	0.10	27.8	14	0.76	0.051	*0.62*	0.071
	a, FI	1,098,890	19.18	2.58	13.5	13	0.80	1.14	*0.73*	1.52
	a2, FI	1,098,766	23.04	2.03	8.8	13	0.76	1.01	*0.64*	1.37
	TG10	1,097,899	13.72	3.60	26.2	11	0.51	3.08	0.44	3.21
	TG10/RCT	1,095,689	0.81	0.27	33.3	16	0.49	0.23	0.26	0.26
Acidification traits for pressed cooked cheese	pH_0	1,098,925	6.52	0.07	1.1	17	0.68	0.030	*0.65*	0.035
	Acidification rate from 170 to 230 min, upH/h	1,099,954	0.11	0.03	27.3	15	0.56	0.024	0.30	0.031
	pH at 10h	1,095,145	5.52	0.12	2.2	14	0.41	0.13	0.38	0.16
Acidification traits for soft cheese	pH_0	1,098,563	6.57	0.07	1.1	16	0.60	0.041	0.46	0.059
	Time to reach a decrease of 0.08 upH, h	1,093,320	3.39	0.15	4.4	9	0.15	0.29	0.17	0.29
	Decrease of pH from 16 h to 20 h	1,098,757	0.17	0.11	64.7	17	0.42	0.10	0.18	0.11
	pH at 20 h	1,100,236	4.49	0.03	0.67	6	0.17	0.075	0.04	0.074
	Time from pH 6 to pH 4.8, h	1,098,099	5.84	1.29	22.1	18	0.51	1.01	0.27	1.21

[1]CY_{FRESH} = fresh curd yield; CY_{DM} = curd yield in DM; $CY_{FAT\text{-}PROT}$ = curd yield in protein and fat; RCT = rennet coagulation time at 0.5 firmness index (FI; optical sensors convert positions into voltages that are computerized and expressed as units of an FI; FI = 10 × volts); K10 = time to obtain 10 FI from RCT; a = curd firmness at RCT; TG10 = curd organization speed; a2 = curd firmness at 2 times RCT; pH_0 = pH after adding lactic acid bacteria; upH = pH units.

[2]Number of latent variables (LV), coefficients of determination (R^2), and root mean squared error (RMSE) in the calibration (cal) and validation (val) sets; R^2 values of cheese-making traits retained for genetic parameter estimations are in italics.

Table 2. Heritability (h^2), proportion of phenotypic variance explained by permanent environmental (PE) effect, repeatability (t), and coefficient of genetic variation (CV_{gen}) of cheese-making properties

Trait[1]	h^2	SE h^2	PE	SE PE	t	CV_{gen}, %
Cheese yield						
CY_{FRESH}	0.38	0.006	0.08	0.005	0.46	8.2
CY_{DM}	0.39	0.006	0.08	0.005	0.47	3.3
$CY_{FAT\text{-}PROT}$	0.37	0.006	0.06	0.004	0.43	4.4
Coagulation trait						
$K10/RCT_{PCC}$	0.42	0.009	0.21	0.007	0.63	12.0
a_{PCC}	0.47	0.007	0.13	0.006	0.60	6.3
$K10/RCT_{SC}$	0.45	0.008	0.19	0.007	0.64	13.8
a_{SC}	0.48	0.007	0.13	0.006	0.61	6.6
$a2_{SC}$	0.47	0.007	0.13	0.006	0.60	4.3
Acidification trait						
$pH_{0\ PCC}$	0.37	0.011	0.31	0.010	0.68	0.5

[1]CY_{FRESH} = fresh curd yield; CY_{DM} = curd yield in DM; $CY_{FAT\text{-}PROT}$ = curd yield in protein and fat. Coagulation and acidification traits were measured for soft cheese (SC) and pressed cooked cheese (PCC): K10 = time to obtain 10 firmness index units (FI) from rennet coagulation time (RCT); a = curd firmness at RCT; a2 = curd firmness at 2 times RCT; pH_0 = pH after adding lactic acid bacteria.

cows. Milk samples were collected between January 2012 and June 2017 and analyzed by MIR spectrometry using a MilkoScan FT6000 (Foss Electric). Cows originated from 3,246 commercial herds in the Franche-Comté region in Eastern France. The MIR spectra of these samples were obtained from 10 spectrometers and standardized according to the method of Grelet et al. (2015).

Cheese-making traits were predicted by applying equations developed in the FROM'MIR project (El Jabri et al., 2017; Laithier et al., 2017). Proteins ($0.54 < R^2 < 1$; Ferrand et al., 2012; Sanchez et al., 2017) and fatty acids ($0.84 < R^2 < 1$; Ferrand-Calmels et al., 2014) were predicted with the equations developed in the PhénoFinlait project, whereas minerals, citrate, and lactose ($0.44 < R^2 < 0.90$) used the equations of the Optimir project (Gengler et al., 2016). Nine CMP (Table 2) and 45 milk composition traits (Appendix Tables A1, A2, and A3) were included in our analyses.

Test-day records before 7 or after 350 DIM were excluded. We selected test-day records from complete lactations or from ongoing lactations with at least 7 test-day records. Lactations that started after December 2016 were excluded. For each trait, outlier values (outside the mean ±3 SD) were discarded. Furthermore, test-day records that had outlier values for PC or FC were removed from the analyses. After data editing, 4,100,261 test-day records from 237,661 cows remained.

For estimations of variance components, we retained only the first-lactation records. Age at calving of the cows ranged from 24 to 42 mo. Contemporary groups were defined as the combination of test day × herd × spectrometer, and season of calving was defined as the month × year of calving combination. Thirty classes of stage of lactation were defined as follows: one stage for 7 to 10 DIM, twenty-four 10-d stages between 10 and 250 DIM (11 to 20 DIM, 21 to 30 DIM, and so on), and five 20-d stages between 250 and 350 DIM (251 to 270 DIM, 271 to 290 DIM, and so on). The minimum number of records per class was 5, 1,000, and 500 for contemporary groups, season, and lactation stage, respectively. The final data set included 1,100,238 test-day records (7 to 13 test-day records per cow, with an average of 8.7) from 126,873 cows born between 2008 and 2014. Data were distributed among 3,086 commercial herds, 89,129 contemporary groups, 58 seasons of calving, and 30 stages of lactation.

Estimations of (Co)variance Components and Genetic Parameters

First-lactation data were analyzed using univariate and bivariate repeatability animal models. In the case of the univariate model, variance components were estimated using the AI-REML algorithm as implemented in Wombat software (Meyer, 2007) with the following linear animal model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{W}\mathbf{p} + \mathbf{e},$$

where $\mathbf{y}$ is the vector of test-day observations, $\mathbf{a} \sim N\left(\mathbf{0}, \mathbf{A}\sigma_a^2\right)$ is the vector of random additive genetic effects, $\mathbf{p} \sim N\left(\mathbf{0}, \mathbf{I}\sigma_p^2\right)$ is the vector of random permanent environmental ($\mathbf{PE}$) effects, and $\mathbf{e} \sim N\left(\mathbf{0}, \mathbf{I}\sigma_e^2\right)$ is the vector of random residual effects. $\mathbf{X}$, $\mathbf{Z}$, and $\mathbf{W}$ are incidence matrices; $\mathbf{A}$ is the relationship matrix among individuals; $\mathbf{I}$ is the identity matrix; and σ_a^2, σ_p^2, and σ_e^2 are the additive genetic, PE, and residual variances, respectively. The $\boldsymbol{\beta}$ vector included the fixed effects of herd × test day × spectrometer, stage of lactation, and

season of calving. The relationship matrix was calculated from the pedigree, which was traced over 4 generations and contained 315,661 animals. Covariances were estimated with a similar bivariate model, with the genetic, PE, and residual variances as 2 × 2 matrices.

Heritability $\left(h^2 = \frac{\hat{\sigma}_a^2}{\hat{\sigma}_a^2 + \hat{\sigma}_p^2 + \hat{\sigma}_e^2}\right)$ and repeatability $\left(t = \frac{\hat{\sigma}_a^2 + \hat{\sigma}_p^2}{\hat{\sigma}_a^2 + \hat{\sigma}_p^2 + \hat{\sigma}_e^2}\right)$ were calculated from the additive genetic $\left(\hat{\sigma}_a^2\right)$, PE $\left(\hat{\sigma}_p^2\right)$, and residual $\left(\hat{\sigma}_e^2\right)$ variance estimates for the 9 best-predicted CMP traits and 45 milk composition traits. The additive genetic correlations (r_a) were computed between CMP traits and between CMP and milk composition traits in bivariate models, $\left(r_a = \frac{\hat{\sigma}_{a1,a2}}{\hat{\sigma}_{a1}, \hat{\sigma}_{a2}}\right)$, with $\hat{\sigma}_{a1,a2}$ being the estimate of additive genetic covariance between trait 1 and trait 2 and $\hat{\sigma}_{a1}$ and $\hat{\sigma}_{a2}$ being the additive genetic standard deviation estimates for traits 1 and 2, respectively.

RESULTS

CMP

Descriptive statistics and accuracy of MIR predictions for the 24 CMP are reported in Table 1. The accuracy of MIR predictions, as estimated by the R^2 in the validation set, was high for 2 measures of cheese yield (CY_{FRESH}: 0.86; CY_{DM}: 0.89) and 2 coagulation traits (0.73 for a_{SC}; 0.72 for a_{PCC}). Mid-infrared-based predictions were moderately accurate for $a2_{SC}$ (0.64), $K10/RCT_{PCC}$ (0.62), $K10/RCT_{SC}$ (0.62), $pH_{0\,PCC}$ (0.65), and $CY_{FAT\text{-}PROT}$ (0.54). Predictions of all other CMP—8 coagulation traits, including RCT measured for both SC ($R^2 = 0.22$) and PCC ($R^2 = 0.23$), and 7 acidification traits ($0.04 < R^2 < 0.46$)—had low accuracy and consequently were not included in genetic analyses. The 9 traits best predicted by MIR spectra analysis (CY_{FRESH}, CY_{DM}, $CY_{FAT\text{-}PROT}$, a_{PCC}, a_{SC}, $a2_{SC}$, $K10/RCT_{SC}$, $K10/RCT_{PCC}$, and $pH_{0\,PCC}$) were used for further investigations of the genetic determinism of CMP.

Heritability estimates varied between 0.37 and 0.48 for the 9 CMP traits analyzed (Table 2). Coagulation traits were the most heritable (0.42 to 0.48): K10/RCT had heritability estimates of 0.42 and 0.45 for PCC and SC, respectively, whereas all curd firmness traits presented similar values (between 0.47 and 0.48). Estimates of heritability were somewhat lower for CY (0.37 to 0.39) and for $pH_{0\,PCC}$ (0.37). However, all these values were relatively high considering that they come from test-day measures. Because the data set used in our study was large, heritability estimates for all traits were highly accurate (SE <0.01; Table 2).

The coefficient of genetic variation ranged from 3.3 to 13.8% for CY and coagulation traits, and it was 0.5% for $pH_{0\,PCC}$. For all CMP traits analyzed, the additive genetic variance was larger than the PE variance. For CY, the part of phenotypic variance due to PE ranged between 0.06 and 0.08. This part was higher for coagulation traits (0.13 to 0.21) and even higher for $pH_{0\,PCC}$ (0.31).

The genetic correlations calculated between CMP traits are presented in Table 3. Fresh CY and CY_{DM} were very strongly positively correlated (0.97), whereas, as expected, the genetic correlations between either of these CY and $CY_{FAT\text{-}PROT}$ (expressed as grams of fat and protein per kilogram of cheese) were highly negative (−0.84 and −0.82, respectively). The genetic correlations between coagulation traits were also relatively high, with absolute values ranging from 0.72 to 0.80. Values of curd-firming time (K10/RCT) and curd firmness (a and a2) were genetically negatively correlated, meaning that milk with longer coagulation times had softer curds. We also found strong genetic correlations between coagulation traits measured for PCC and SC: 0.80 for curd-firming time and 0.76 to 0.77 for curd firmness, which suggests that selection for improved coagulation traits for either type of cheese would be beneficial to both PCC and SC. Moreover, CY and coagulation traits were also strongly genetically correlated. Fresh CY and CY_{DM} were both positively correlated with curd firmness and negatively correlated with curd-firming time. In contrast, the genetic correlations between $pH_{0\,PCC}$ and CY or coagulation traits were very weak (−0.08 to 0.09).

CMP and Milk Composition

Descriptive statistics, heritability, and proportion of phenotypic variance explained by PE effect for milk protein, fatty acid, and mineral, citrate, and lactose composition are reported in Appendix Tables A1, A2, and A3, respectively. On average, the milk analyzed in this study comprised 4.96% lactose, 3.73% fat, 3.35% protein, 0.83% citrate, and 0.41% minerals.

Heritability estimates of PC and FC were 0.48 and 0.36, respectively. The heritability estimates of individual milk proteins (0.35 to 0.58) were higher than those of individual fatty acids (0.14 to 0.40). Heritability estimates of Ca, P, K, and Mg were high (0.50 to 0.56), as were the values obtained for citrate (0.48) and lactose (0.47).

As shown in Table 4, milk PC was strongly genetically correlated with coagulation traits (−0.80 and −0.81 for

Table 3. Genetic correlations between cheese-making properties for soft cheese (SC) and pressed cooked cheese (PCC)

	Cheese yield		Coagulation trait					Acidification
Trait[1]	CY_{DM}	$CY_{FAT\text{-}PROT}$	$K10/RCT_{PCC}$	a_{PCC}	$K10/RCT_{SC}$	a_{SC}	$a2_{SC}$	$pH_{0\ PCC}$
CY_{FRESH}	0.97	−0.84	−0.73	0.78	−0.76	0.78	0.75	0.03
CY_{DM}		−0.82	−0.72	0.77	−0.74	0.76	0.72	−0.02
$CY_{FAT\text{-}PROT}$			0.65	−0.65	0.71	−0.67	−0.64	−0.01
$K10/RCT_{PCC}$				−0.76	0.80	−0.73	−0.72	0.09
a_{PCC}					−0.78	0.76	0.77	−0.08
$K10/RCT_{SC}$						−0.77	−0.77	0.06
a_{SC}							0.74	0.02
$a2_{SC}$								0.02
$pH_{0\ PCC}$								

[1]CY_{FRESH} = fresh curd yield; CY_{DM} = curd yield in DM; $CY_{FAT\text{-}PROT}$ = curd yield in protein and fat; K10 = time to obtain 10 firmness index units (FI) from rennet coagulation time (RCT); a = curd firmness at RCT; a2 = curd firmness at 2 times RCT; pH_0 = pH after adding lactic acid bacteria. SE varied between 0.001 and 0.038.

$K10/RCT_{PCC}$ and $K10/RCT_{SC}$, respectively, and 0.89 to 0.94 for curd firmness traits) and, to a lesser extent, with cheese yields (0.75, 0.74, and −0.52 for CY_{FRESH}, CY_{DM}, and $CY_{FAT\text{-}PROT}$, respectively). Milk FC was more genetically correlated with CY (0.87, 0.91, and −0.57 for CY_{FRESH}, CY_{DM}, and $CY_{FAT\text{-}PROT}$, respectively) than it was with coagulation traits (−0.53 to −0.47 for $K10/RCT_{PCC}$ and $K10/RCT_{SC}$, respectively, and 0.48 to 0.55 for curd firmness traits; Table 5). In contrast, both PC and FC were only weakly genetically correlated with $pH_{0\ PCC}$ (−0.11 and −0.06, respectively).

For protein and fatty acid composition, we calculated genetic correlations between proteins (or fatty acids) expressed as a percentage of milk or as a percentage of total protein (or total fat). The results obtained varied greatly depending on the expression unit. It should be noted that convergence was not completely achieved for most of the genetic correlation estimates between curd-firming time and protein, fatty acid, and mineral composition. Therefore, these results have to be considered with caution. Nevertheless, CMP traits appeared to be more strongly genetically associated with protein

Table 4. Genetic correlations between cheese-making properties for soft cheese (SC) or pressed cooked cheese (PCC) and milk protein composition[1]

	Cheese yield			Coagulation trait					Acidification
Trait[1]	CY_{FRESH}	CY_{DM}	$CY_{FAT\text{-}PROT}$	$K10/RCT_{PCC}$	a_{PCC}	$K10/RCT_{SC}$	a_{SC}	$a2_{SC}$	$pH_{0\ PCC}$
Proteins, g/100 g of milk									
PC	0.75	0.74	−0.52	−0.80	0.94	−0.81	*0.91*	0.89	−0.11
α-LA	*0.14*	*0.14*	*−0.10*	−0.19	0.20	−0.15	0.20	0.20	0.04
β-LG	0.14	0.11	0.18	*−0.32*	0.28	*−0.27*	0.26	0.25	−0.04
α_{S1}-CN	0.75	0.74	*−0.53*	−0.78	0.81	−0.82	0.80	0.79	*−0.13*
α_{S2}-CN	0.34	0.33	−0.22	−0.44	0.43	−0.42	0.43	0.41	*−0.05*
β-CN	0.71	0.71	−0.63	−0.79	0.79	−0.79	0.77	0.77	*−0.19*
κ-CN	0.34	0.34	−0.25	−0.43	0.42	−0.38	0.41	0.41	0.03
ΣWP	0.26	0.23	0.09	−0.45	0.41	*−0.40*	0.38	0.38	−0.05
ΣCN	0.75	0.74	−0.54	−0.80	*0.94*	−0.81	*0.90*	0.90	−0.12
Proteins, g/100 g of protein									
α-LA	0.09	0.09	−0.07	*−0.18*	0.22	*−0.17*	0.28	0.31	0.32
β-LG	−0.28	−0.31	0.56	*0.17*	−0.19	*0.20*	−0.21	−0.21	0.02
α_{S1}-CN	0.04	0.11	0.04	*0.17*	−0.15	*0.14*	−0.15	−0.18	−0.14
α_{S2}-CN	0.19	0.15	−0.21	*−0.41*	0.41	*−0.36*	0.39	0.44	0.02
β-CN	−0.04	−0.02	−0.34	*0.08*	−0.05	*0.05*	−0.05	−0.02	−0.21
κ-CN	0.30	0.30	−0.22	−0.41	0.42	−0.32	0.44	0.46	0.28
ΣWP	−0.26	−0.29	0.56	*0.15*	−0.17	*0.20*	−0.18	−0.18	0.02
ΣCN	0.09	0.12	−0.17	*−0.09*	0.13	*−0.01*	0.09	0.15	−0.09

[1]CY_{FRESH} = fresh curd yield; CY_{DM} = curd yield in DM; $CY_{FAT\text{-}PROT}$ = curd yield in protein and fat; K10 = time to obtain 10 firmness index units (FI) from rennet coagulation time (RCT); a = curd firmness at RCT; a2 = curd firmness at 2 times RCT; pH_0 = pH after adding lactic acid bacteria; PC = protein content; ΣWP = sum of whey proteins (α-LA + β-LG); ΣCN = sum of caseins (α_{S1}-CN + α_{S2}-CN + β-CN + κ-CN). SE varied between 0.002 and 0.022; full convergence was not achieved for values in italics.

Table 5. Genetic correlations between cheese-making properties for soft cheese (SC) or pressed cooked cheese (PCC) and milk fatty acid composition[1]

	Cheese yield			Coagulation trait					Acidification
Trait	CY_{FRESH}	CY_{DM}	$CY_{FAT\text{-}PROT}$	$K10/RCT_{PCC}$	a_{PCC}	$K10/RCT_{SC}$	a_{SC}	$a2_{SC}$	$pH_{0\ PCC}$
Fatty acids, g/100 g of milk									
FC	0.87	0.91	−0.57	*−0.53*	0.55	*−0.47*	0.51	0.48	−0.06
SFA	0.81	0.86	−0.52	*−0.47*	0.49	*−0.40*	0.45	0.42	−0.08
MUFA	*0.69*	*0.69*	−0.50	*−0.50*	0.50	*−0.49*	0.49	0.47	−0.01
UFA	0.72	0.71	−0.52	*−0.53*	0.53	*−0.52*	0.52	0.50	0.00
PUFA	0.14	0.14	−0.09	−0.12	0.13	−0.12	0.12	0.12	−0.01
ΣC4-C10	0.56	0.59	−0.42	−0.45	0.44	−0.38	0.41	0.41	−0.08
ΣC4-C12	0.69	0.71	*−0.44*	*−0.48*	*0.50*	*−0.42*	*0.46*	*0.46*	−0.11
C14:0	0.63	0.65	−0.45	−0.47	0.47	−0.41	0.43	0.42	−0.13
C16:0	0.77	0.82	−0.49	*−0.42*	0.44	*−0.36*	0.42	0.38	0.02
C18:0	0.36	0.37	−0.30	*−0.08*	*0.07*	*−0.04*	0.03	0.02	0.02
C18:1	*0.61*	*0.59*	*−0.44*	*−0.40*	*0.42*	*−0.41*	*0.41*	*0.39*	0.02
Fatty acids, g/100 g of fat									
SFA	0.32	0.37	−0.16	*−0.14*	0.16	*−0.06*	0.11	0.11	−0.13
MUFA	−0.33	−0.39	0.17	*0.14*	−0.15	*0.06*	−0.11	−0.10	0.08
UFA	−0.34	−0.40	0.17	*0.13*	−0.14	*0.05*	−0.10	−0.09	0.08
PUFA	−0.42	−0.47	0.30	*0.00*	−0.02	*−0.04*	0.01	0.04	0.03
ΣC4-C10	−0.06	−0.03	0.00	*−0.07*	0.08	*−0.07*	0.07	0.10	−0.09
ΣC4-C12	0.11	0.15	−0.02	*−0.19*	0.19	*−0.16*	0.17	0.20	−0.16
C14:0	0.21	0.24	−0.12	*−0.21*	0.20	*−0.18*	0.18	0.18	−0.19
C16:0	0.40	0.45	−0.24	*−0.15*	0.17	*−0.09*	0.17	0.14	0.12
C18:0	−0.33	−0.36	0.13	*0.39*	−0.41	*0.39*	−0.43	−0.42	0.02
C18:1	−0.38	−0.45	0.22	*0.20*	−0.22	*0.13*	−0.18	−0.17	0.10

[1]CY_{FRESH} = fresh curd yield; CY_{DM} = curd yield in DM; $CY_{FAT\text{-}PROT}$ = curd yield in protein and fat; K10 = time to obtain 10 firmness index units (FI) from rennet coagulation time (RCT); a = curd firmness at RCT; a2 = curd firmness at 2 times RCT; pH_0 = pH after adding lactic acid bacteria; FC = fat content; ΣC4-C10 = sum of fatty acids with 4 to 10 carbon atoms; ΣC4-C12 = sum of fatty acids with 4 to 12 carbon atoms. SE varied between 0.002 and 0.022; full convergence was not achieved for values in italics.

and fatty acid content when they were expressed as a percentage of milk rather than as a percentage of protein or fat. Regardless of the protein or fatty acid considered, high levels in milk were always genetically associated with better CMP (Tables 4 and 5).

This result was particularly marked for caseins—namely, α_{S1}- and β-casein. As observed for total PC, the total casein content in milk was strongly genetically correlated with coagulation traits (−0.80 and −0.81 for $K10/RCT_{PCC}$ and $K10/RCT_{SC}$, respectively, and 0.90 to 0.94 for curd firmness traits), whereas the genetic correlations between whey proteins in milk and coagulation traits were much weaker (0.15 to 0.32 in absolute values). In contrast, genetic correlations with relative amounts of proteins were more moderate (Table 4). The content of β-LG in total protein was negatively correlated with CY_{FRESH}, CY_{DM}, and curd firmness traits and positively correlated with $CY_{FAT\text{-}PROT}$ and K10/RCT.

Genetic correlations were moderate to high between CY_{FRESH} and CY_{DM} and fatty acids expressed as a percentage of milk, in particular with SFA and palmitic acid (C16:0). Coagulation traits were more genetically correlated with SFA, MUFA, and UFA than with PUFA (Table 5). For relative fatty acid contents, expressed as a percentage of total fat, high amounts of SFA and low levels of MUFA, UFA, or PUFA were associated with better CMP (higher CY_{FRESH} and CY_{DM}, lower curd-firming times, and firmer curds). However, unlike saturated short-chain (C4-C12), myristic (C14:0), and palmitic (C16:0) fatty acids, lower levels of stearic acid content (C18:0), which represents about 10% of the total fatty acids in milk, were associated with better CMP.

Mineral composition and, in particular, Ca, P, and Mg contents were moderately genetically correlated with CY_{FRESH} and CY_{DM} (0.40 to 0.58) and coagulation traits (0.37 to 0.59 in absolute values). In contrast, genetic correlations between CMP and other minerals (K and Na), citrate, or lactose were weak (−0.16 to 0.17; Table 6).

DISCUSSION

Our study is the first to report genetic parameters of CMP and milk composition in Montbéliarde cows (the second largest dairy cattle breed in France) at a very large scale (more than 1 million test-day records) and for a large number of traits. The accuracies of MIR predictions for cheese-making traits were equivalent

Table 6. Genetic correlations between cheese-making properties for soft cheese (SC) or pressed cooked cheese (PCC) and the mineral, citrate, and lactose composition of milk

	Cheese yield			Coagulation trait					Acidification
Trait[1]	CY_{FRESH}	CY_{DM}	$CY_{FAT\text{-}PROT}$	$K10/RCT_{PCC}$	a_{PCC}	$K10/RCT_{SC}$	a_{SC}	$a2_{SC}$	$pH_{0\ PCC}$
Mineral, mg/kg of milk									
Ca	0.40	0.41	−0.25	*−0.46*	0.45	*−0.37*	0.39	0.42	0.10
P	0.58	0.54	−0.44	−0.58	0.59	*−0.58*	0.54	0.54	−0.18
K	−0.09	−0.13	−0.02	*0.13*	−0.15	*0.05*	−0.13	−0.15	0.12
Mg	0.41	0.40	−0.31	*−0.54*	0.50	*−0.53*	0.42	0.40	−0.29
Na	−0.08	−0.04	0.13	*−0.09*	0.06	*−0.10*	0.12	0.06	0.17
Other compounds, g/kg of milk									
Citrate	−0.08	−0.09	0.00	*0.05*	−0.12	*0.07*	−0.16	−0.16	0.00
Lactose	0.03	−0.05	−0.04	*0.03*	0.01	*0.05*	0.00	0.07	0.04

[1]CY_{FRESH} = fresh curd yield; CY_{DM} = curd yield in DM; $CY_{FAT\text{-}PROT}$ = curd yield in protein and fat; K10 = time to obtain 10 firmness index units (FI) from rennet coagulation time (RCT); a = curd firmness at RCT; a2 = curd firmness at 2 times RCT; pH_0 = pH after adding lactic acid bacteria. SE varied between 0.009 and 0.023; full convergence was not achieved for values in italics.

to or higher than accuracies reported in other studies. For example, R^2 values were previously estimated to be between 0.67 and 0.85 for CY_{FRESH} or CY_{DM} (Ferragina et al., 2013; Bonfatti et al., 2016; Grelet et al., 2017), 0.20 and 0.52 for curd firmness (Cecchinato et al., 2009; Visentin et al., 2015; Bonfatti et al., 2016), and 0.71 and 0.79 for pH (Visentin et al., 2015; Bonfatti et al., 2016). However, the current study failed to accurately predict RCT, which had previously been predicted with R^2 values ranging from 0.55 to 0.69 (Cecchinato et al., 2009; Visentin et al., 2015; Bonfatti et al., 2016). These discrepancies might be due to differences in the equipment used to measure CMP or to the standardization of initial milk pH, which is the main factor influencing RCT. In our study, a Formoptic instrument was used and a pH standardization for measuring coagulation traits was applied. As a result, the coefficient of variation of RCT was more than 2 times smaller (11% for SC and 15% for PCC) than values reported in other studies that applied a computerized renneting meter (Cecchinato et al., 2009; Bonfatti et al., 2016) or Formagraph apparatus (Visentin et al., 2015) without pH standardization. Although we were not able to directly analyze RCT, we could assess curd-firming time using the ratio of K10 (time to obtain 10 FI) to RCT, which was predicted with R^2 of 0.62 for both PCC and SC.

Heritabilities

Despite some discrepancies that were probably due to different equipment or protocols used, we found that our estimates of heritability were globally in agreement with previously reported results. For CY, heritabilities were estimated to be between 0.13 and 0.21 using reference measurements in Brown Swiss milk (Bittante et al., 2013; Cecchinato and Bittante, 2016) and between 0.09 and 0.33 using MIR predictions in Italian Holstein, Brown Swiss, and Simmental milk (Cecchinato et al., 2015; Bonfatti et al., 2017). Depending on the study, heritability estimates of coagulation, expressed as curd-firming time, varied between 0.21 and 0.43, whereas those for curd firmness (measured at different times after addition of rennet) varied between 0.12 and 0.41 (Vallas et al., 2010; Cecchinato et al., 2011; Tiezzi et al., 2013; Poulsen et al., 2015; Cecchinato and Bittante, 2016; Bonfatti et al., 2017; Visentin et al., 2017). In these studies, acidification assessed by a single measurement of milk pH showed low heritability (0.06 to 0.27; Vallas et al., 2010; Cecchinato et al., 2011; Tiezzi et al., 2013; Bonfatti et al., 2017; Visentin et al., 2017). We found estimates of heritability very similar to or higher than those obtained using reference measurements, indicating that MIR predictions are sufficiently accurate to be used for genetic investigations. Nevertheless, it is possible that, by decreasing environmental variance, pH standardization led to an increase of CMP heritabilities.

We also found relatively high coefficients of genetic variation for CY and coagulation traits (3.3 to 13.8%) and low coefficients of genetic variation for $pH_{0\ PCC}$ (0.5%). These were all strongly consistent with values reported from reference measurements by Visentin et al. (2017) in Holstein-Friesian cows for both coagulation traits (3.1 to 11.5%) and pH (0.3%). In addition, for all CMP traits analyzed, the additive genetic variance was larger than the PE variance. Our CY traits were less influenced by PE variance than those of Cecchinato et al. (2015), whereas the PE effect on coagulation traits here was slightly higher than or similar to previous reports (Vallas et al., 2010; Tiezzi et al., 2013; Visentin et al., 2017). The effect of PE variance on $pH_{0\ PCC}$ was higher than the one published by Vallas et al. (2010) and Tiezzi et al. (2013) but similar to the effect reported by Visentin et al. (2017).

In addition to the moderate to high heritability estimates (0.37 to 0.48) obtained here for single test-day records, all these results suggest that CMP could be improved by selection. For instance, these values of heritability are higher than those found for milk yield, which has been included in breeding goals for many years and which has more than doubled in the last few decades (Ducrocq and Wiggans, 2015).

For all milk composition traits that were predicted from MIR spectra, we found heritability estimates that were very consistent with those from previous studies. For example, with regard to milk protein and fatty acid composition, our results were in complete agreement with previous studies in Montbéliarde cows (Boichard et al., 2014; Sanchez et al., 2017). For mineral composition, we found heritability values similar to those reported from reference analyses (van Hulzen et al., 2009) and MIR predictions (Govignon-Gion et al., 2015) but higher than the estimates from MIR predictions reported by Toffanin et al. (2015) and Bonfatti et al. (2017). Furthermore, in our samples the content of citrate and lactose was highly heritable, which was similar to previously reported values for citrate (Buitenhuis et al., 2013) but higher than those reported for lactose (Buitenhuis et al., 2013; Haile-Mariam and Pryce, 2017).

Genetic Correlations Between CMP

Our study also revealed high genetic correlations between different cheese-making traits, especially between different measures of CY, which corroborates the results of Bittante et al. (2013) and Cecchinato et al. (2015) obtained from the production of micro-cheese. Likewise, our observation of high genetic correlations between coagulation traits is consistent with results obtained from a computerized renneting meter (Cecchinato et al., 2011) and a Formagraph device (Visentin et al., 2017). We also found a strong genetic link between CY and coagulation traits, and the values obtained were higher than those recently reported by Cecchinato and Bittante (2016). In contrast, the genetic correlations between $pH_{0\ PCC}$ and other cheese-making traits were low, which was inconsistent with the high positive genetic correlations (0.69 to 0.94) between RCT and pH that have been reported in some studies (Vallas et al., 2010; Cecchinato et al., 2011); however, this can be explained by the pH standardization procedure we applied.

We found strong, favorable genetic correlations between CY and coagulation traits as well as between coagulation traits measured for PCC and SC cheeses. The implication of this is that the inclusion in breeding goals of a single CMP trait (e.g., the best-predicted one or the most heritable one) could generate progress for the overall cheese yield and coagulation abilities in a wide range of cheese-production systems.

Genetic Correlations Between CMP and Milk Composition

The genetic relationships between milk composition and CMP are strong. Two previous studies investigated the genetic correlations between CY and both PC and FC in Holstein, Brown Swiss, and Simmental cows (Bittante et al., 2013; Cecchinato et al., 2015) and found, as we report here, high genetic correlation estimates with both PC (0.75 to 0.93) and FC (0.84 to 0.97). In contrast, estimates of genetic correlations between milk coagulation traits and PC or FC varied widely among studies. Vallas et al. (2010) and Cecchinato et al. (2011) found very weak genetic correlations between curd-firming time and PC or FC (−0.11 to 0.19), whereas Visentin et al. (2017) reported higher values, closer to those reported in the present study for FC (−0.46) but lower for PC (−0.56). Genetic correlation estimates between curd firmness and PC or FC have been more consistent. As in our study, these values were always positive, varying between 0.35 and 0.90 for PC and between 0.25 and 0.62 for FC (Vallas et al., 2010; Cecchinato et al., 2015; Visentin et al., 2017). With pH, the slightly negative correlation estimates recovered in our study were lower than those reported in Visentin et al. (2017); these authors found low correlations between pH and PC (−0.33) or FC (−0.24).

The genetic correlations estimated between protein composition and CMP were consistent but higher than results reported in the few studies that have been conducted to date. So far, only 1 study has reported genetic correlations between milk coagulation traits and milk protein composition as a percentage of milk or of total protein (Bonfatti et al., 2011), whereas 3 other studies indicated genetic correlations between CMP and total caseins in milk (Cecchinato et al., 2011; Gustavsson et al., 2014) or with κ-casein and β-LG expressed as percentage of proteins (Gustavsson et al., 2014). Furthermore, to our knowledge, our study is the first to investigate the genetic relationships between milk fatty acid composition and CMP.

High levels of Ca, P, and Mg were associated with better CY and coagulation properties, in agreement with previous results (Amenu and Deeth, 2007). In contrast, although phenotypic associations were previously identified between coagulation properties and K, Na, citrate, or lactose (Glantz et al., 2010; Sundekilde et al., 2011; Bland et al., 2015), only low genetic correlations were observed in our study.

Cheese yields and coagulation traits were strongly genetically correlated with individual proteins (mainly κ-casein) and, to a lesser extent, with fatty acids (mainly SFA and C16:0) and with mineral (Ca, P, and Mg) composition. These relationships can be interpreted considering the mechanisms behind the cheese-making process. Caseins are aggregated in micelles that mainly comprise caseins (92%), calcium phosphate (7%), citrate (0.5%), and magnesium (0.2%). In these micelles, α_{S1}-, α_{S2}-, β-, and κ-caseins are present in relative amounts of 3:1:3:1. Most κ-casein is found on the surface of the micelles, whereas the other caseins (α_{S1}, α_{S2}, and β) are located inside. Moreover, two-thirds of total Ca, one-third of Mg, and one-half of P contents are associated with casein molecules, whereas Na and K are diffusible ions. During the cheese-making process, enzymes of rennet (chymosin) degrade the κ-casein, which causes a decrease in the colloidal stability and promotes the coagulation of micelles. Casein micelle size was found to affect the kinetics of the coagulation process. Milk with smaller casein micelles, containing higher levels of κ-casein and minerals, form firmer curds earlier (Glantz et al., 2010; Logan et al., 2014, 2015).

To a lesser extent, the size of fat globules, which is influenced by fatty acid composition (Timmen and Patton, 1988; Couvreur and Hurtaud, 2017), is also associated with milk coagulation via their interaction with the aggregation of casein micelles. However, results have not always been consistent. Logan et al. (2015) found that large fat globules, combined with small casein micelles, gave the firmest curds. Luo et al. (2017) showed that a reduction in fat globule size accelerated casein micelle aggregation during renneting and led to firmer gels. In the present study, the relative content of different fatty acids was not strongly genetically associated with coagulation properties with 1 exception: high levels of C18:0, which has been found to be associated with larger fat globules (Timmen and Patton, 1988), were genetically associated with undesirable coagulation properties (longer curd-firming time and less-firm curds). Our results are therefore in agreement with those of Luo et al. (2017) and support their observation that a reduction of fat globule size accelerated the aggregation of casein micelles during renneting and increased curd firmness.

CONCLUSIONS

This large-scale study carried out in Montbéliarde cows from the Franche-Comté cheese production area (protected designation of origin and protected geographical indication) shows that MIR predictions of cheese yields and milk coagulation properties are sufficiently accurate for genetic analyses. Cheese-making traits, as predicted from MIR spectra, are moderately heritable and could be integrated without additional cost into breeding objectives, thus promoting their efficient improvement via selection. Using genotyping data produced for routine genomic selection, additional analyses will be performed to estimate the effects of milk protein variants on CMP and to identify novel genetic variants with an effect on CMP that could be included in genomic evaluation models.

ACKNOWLEDGMENTS

This study was funded by the French Ministry of Agriculture, Agro-food, and Forest, the French Dairy Interbranch Organization (CNIEL, Paris, France), the Regional Union of Protected Designation cheeses of Franche-Comté (URFAC, Paris, France), and the Regional Council of Bourgogne-Franche-Comté (Poligny, France) under the project FROM'MIR (Besançon, France).

REFERENCES

Amenu, B., and H. Deeth. 2007. The impact of milk composition on Cheddar cheese manufacture. Aust. J. Dairy Technol. 62:171–184.

Bittante, G., C. Cipolat-Gotet, and A. Cecchinato. 2013. Genetic parameters of different measures of cheese yield and milk nutrient recovery from an individual model cheese-manufacturing process. J. Dairy Sci. 96:7966–7979. https://doi.org/10.3168/jds.2012-6517.

Bittante, G., M. Penasa, and A. Cecchinato. 2012. Invited review: Genetics and modeling of milk coagulation properties. J. Dairy Sci. 95:6843–6870. https://doi.org/10.3168/jds.2012-5507.

Bland, J. H., A. Grandison, and C. Fagan. 2015. Evaluation of milk compositional variables on coagulation properties using partial least squares. J. Dairy Res. 82:8–14. https://doi.org/10.1017/S0022029914000508.

Boichard, D., A. Govignon-Gion, H. Larroque, C. Maroteau, I. Palhiere, G. Tosser-Klop, R. Rupp, M. Sanchez, and M. Brochard. 2014. Genetic determinism of milk composition in fatty acids and proteins in ruminants, and selection potential. INRA Prod. Anim. 27:283–298.

Bonfatti, V., A. Cecchinato, L. Gallo, A. Blasco, and P. Carnier. 2011. Genetic analysis of detailed milk protein composition and coagulation properties in Simmental cattle. J. Dairy Sci. 94:5183–5193. https://doi.org/10.3168/jds.2011-4297.

Bonfatti, V., L. Degano, A. Menegoz, and P. Carnier. 2016. Short communication: Mid-infrared spectroscopy prediction of fine milk composition and technological properties in Italian Simmental. J. Dairy Sci. 99:8216–8221. https://doi.org/10.3168/jds.2016-10953.

Bonfatti, V., D. Vicario, A. Lugo, and P. Carnier. 2017. Genetic parameters of measures and population-wide infrared predictions of 92 traits describing the fine composition and technological properties of milk in Italian Simmental cattle. J. Dairy Sci. 100:5526–5540. https://doi.org/10.3168/jds.2016-11667.

Buitenhuis, A. J., U. Sundekilde, N. Poulsen, H. Bertram, L. Larsen, and P. Sorensen. 2013. Estimation of genetic parameters and detection of quantitative trait loci for metabolites in Danish Holstein milk. J. Dairy Sci. 96:3285–3295. https://doi.org/10.3168/jds.2012-5914.

Cecchinato, A., A. Albera, C. Cipolat-Gotet, A. Ferragina, and G. Bittante. 2015. Genetic parameters of cheese yield and curd nutrient recovery or whey loss traits predicted using Fourier-transform infrared spectroscopy of samples collected during milk recording

on Holstein, Brown Swiss, and Simmental dairy cows. J. Dairy Sci. 98:4914–4927. https://doi.org/10.3168/jds.2014-8599.
Cecchinato, A., and G. Bittante. 2016. Genetic and environmental relationships of different measures of individual cheese yield and curd nutrients recovery with coagulation properties of bovine milk. J. Dairy Sci. 99:1975–1989. https://doi.org/10.3168/jds.2015-9629.
Cecchinato, A., M. De Marchi, L. Gallo, G. Bittante, and P. Carnier. 2009. Mid-infrared spectroscopy predictions as indicator traits in breeding programs for enhanced coagulation properties of milk. J. Dairy Sci. 92:5304–5313. https://doi.org/10.3168/jds.2009-2246.
Cecchinato, A., M. Penasa, M. De Marchi, L. Gallo, G. Bittante, and P. Carnier. 2011. Genetic parameters of coagulation properties, milk yield, quality, and acidity estimated using coagulating and noncoagulating milk information in Brown Swiss and Holstein-Friesian cows. J. Dairy Sci. 94:4205–4213. https://doi.org/10.3168/jds.2010-3913.
Corrieu, G., H. Spinnler, Y. Jomier, and D. Picque. 1988. Automated system to follow up and control the acidification activity of lactic acid starters. French patent FR 2:629-612.
Couvreur, S., and C. Hurtaud. 2017. Relationships between milks differentiated on native milk fat globule characteristics and fat, protein and calcium compositions. Animal 11:507–518. https://doi.org/10.1017/S1751731116001646.
De Marchi, M., V. Toffanin, M. Cassandro, and M. Penasa. 2014. Invited review: Mid-infrared spectroscopy as phenotyping tool for milk traits. J. Dairy Sci. 97:1171–1186. https://doi.org/10.3168/jds.2013-6799.
Ducrocq, V., and G. R. Wiggans. 2015. Genetic improvement of dairy cattle. Pages 371–396 in The Genetics of Cattle. D. J. Garrick and A. Ruvinsky, ed. CABI, Wallingford, UK.
El Jabri, M., M. P. Sanchez, C. Laithier, E. Doutart, V. Wolf, D. Pourchet, P. Grosperrin, E. Beuvier, O. Rolet-Répécaud, Y. Gaüzère, O. Belysheva, A. Delacroix-Buchet, and D. Boichard. 2017. Bayesian regression models and variable selection methods before PLS regression. Application to the Prediction of milk cheese-making properties using infared spectral data. Pages 15–17 in Chimiometrie XVIII. AgroParisTech, Paris, France.
Fang, Z. H., M. Visker, G. Miranda, A. Delacroix-Buchet, H. Bovenhuis, and P. Martin. 2016. The relationships among bovine alpha(S)-casein phosphorylation isoforms suggest different phosphorylation pathways. J. Dairy Sci. 99:8168–8177. https://doi.org/10.3168/jds.2016-11250.
Ferragina, A., C. Cipolat-Gotet, A. Cecchinato, and G. Bittante. 2013. The use of Fourier-transform infrared spectroscopy to predict cheese yield and nutrient recovery or whey loss traits from unprocessed bovine milk samples. J. Dairy Sci. 96:7980–7990. https://doi.org/10.3168/jds.2013-7036.
Ferrand, M., G. Miranda, S. Guisnel, H. Larroque, O. Leray, F. Lahalle, M. Brochard, and P. Martin. 2012. Determination of protein composition in milk by mid-infrared spectrometry. Pages 41–45 in Proc. International Strategies and New Developments in Milk Analysis. VI ICAR Reference Laboratory Network Meeting, Cork, Ireland. ICAR, Rome, Italy.
Ferrand-Calmels, M., I. Palhiere, M. Brochard, O. Leray, J. Astruc, M. Aurel, S. Barbey, F. Bouvier, P. Brunschwig, H. Caillatt, M. Douguet, F. Faucon-Lahalle, M. Gele, G. Thomas, J. Trommenschlager, and H. Larroque. 2014. Prediction of fatty acid profiles in cow, ewe, and goat milk by mid-infrared spectrometry. J. Dairy Sci. 97:17–35. https://doi.org/10.3168/jds.2013-6648.
Gengler, N., H. Soyeurt, F. Dehareng, C. Bastin, F. Colinet, H. Hammami, M. Vanrobays, A. Laine, S. Vanderick, C. Grelet, A. Vanlierde, E. Froidmont, and P. Dardenne. 2016. Capitalizing on fine milk composition for breeding and management of dairy cows. J. Dairy Sci. 99:4071–4079. https://doi.org/10.3168/jds.2015-10140.
Glantz, M., T. Devold, G. Vegarud, H. Mansson, H. Stalhammar, and M. Paulsson. 2010. Importance of casein micelle size and milk composition for milk gelation. J. Dairy Sci. 93:1444–1451. https://doi.org/10.3168/jds.2009-2856.
Govignon-Gion, A., S. Minery, M. Wald, M. Brochard, M. Gelé, B. Rouillé, D. Boichard, M. Ferrand-Calmels, and C. Hurtaud. 2015. Paramètres génétiques du taux de calcium, prédit à partir des spectres moyen infrarouge, dans le lait des 3 principales races bovines laitières françaises. Pages 111–114 in Proc. Renc. Rech. Ruminants, Paris, France. INRA, Paris, France.
Grelet, C., J. Pierna, P. Dardenne, V. Baeten, and F. Dehareng. 2015. Standardization of milk mid-infrared spectra from a European dairy network. J. Dairy Sci. 98:2150–2160. https://doi.org/10.3168/jds.2014-8764.
Grelet, C., J. Pierna, P. Dardenne, H. Soyeurt, A. Vanlierde, F. Colinet, C. Bastin, N. Gengler, V. Baeten, and F. Dehareng. 2017. Standardization of milk mid-infrared spectrometers for the transfer and use of multiple models. J. Dairy Sci. 100:7910–7921. https://doi.org/10.3168/jds.2017-12720.
Gustavsson, F., M. Glantz, N. Poulsen, L. Wadso, H. Stalhammar, A. Andren, H. Mansson, L. Larsen, M. Paulsson, and W. Fikse. 2014. Genetic parameters for rennet- and acid-induced coagulation properties in milk from Swedish Red dairy cows. J. Dairy Sci. 97:5219–5229. https://doi.org/10.3168/jds.2014-7996.
Haile-Mariam, M., and J. E. Pryce. 2017. Genetic parameters for lactose and its correlation with other milk production traits and fitness traits in pasture-based production systems. J. Dairy Sci. 100:3754–3766. https://doi.org/10.3168/jds.2016-11952.
Hurtaud, C., H. Rulquin, M. Delaite, and R. Verite. 1995. Prediction of cheese yielding efficiency of individual milk of dairy cows—Correlation with coagulation parameters and laboratory curd yield. Ann. Zootech. 44:385–398.
International Dairy Federation. 2016. The World Dairy Situation. Bulletin 485/2016. International Dairy Federation, Brussels, Belgium.
Laithier, C., V. Wolf, M. El Jabri, P. Trossat, S. Gavoye, D. Pourchet, P. Grosperrin, E. Beuvier, O. Rolet-Répécaud, Y. Gauzère, O. Belysheva, E. Notz, and A. Delacroix-Buchet. 2017. Prediction of cheesemaking properties of Montbéliarde milks used for PDO/PGI cheeses production in Franche-Comté by mid-infrared spectrometry. Pages 15–19 in 12th International Meeting on Mountain Cheese, Padova, Italy. Padova Universtiy Press, Padova, Italy.
Logan, A., L. Day, A. Pin, M. Auldist, A. Leis, A. Puvanenthiran, and M. Augustin. 2014. Interactive effects of milk fat globule and casein micelle size on the renneting properties of milk. Food Bioprocess Technol. 7:3175–3185. https://doi.org/10.1007/s11947-014-1362-2.
Logan, A., A. Leis, L. Day, S. Oiseth, A. Puvanenthiran, and M. Augustin. 2015. Rennet gelation properties of milk: Influence of natural variation in milk fat globule size and casein micelle size. Int. Dairy J. 46:71–77. https://doi.org/10.1016/j.idairyj.2014.08.005.
Luo, J., Y. Wang, H. Guo, and F. Ren. 2017. Effects of size and stability of native fat globules on the formation of milk gel induced by rennet. J. Food Sci. 82:670–678. https://doi.org/10.1111/1750-3841.13649.
Meyer, K. 2007. WOMBAT—A tool for mixed model analyses in quantitative genetics by REML. J. Zhejiang Univ. Sci. B 8:815–821. https://doi.org/10.1631/jzus.2007.B0815.
Poulsen, N. A., A. Buitenhuis, and L. Larsen. 2015. Phenotypic and genetic associations of milk traits with milk coagulation properties. J. Dairy Sci. 98:2079–2087. https://doi.org/10.3168/jds.2014-7944.
Sanchez, M. P., M. Ferrand, M. Gelé, D. Pourchet, G. Miranda, P. Martin, M. Brochard, and D. Boichard. 2017. Short communication: Genetic parameters for milk protein composition predicted using mid-infrared spectroscopy in the French Montbéliarde, Normande, and Holstein dairy cattle breeds. J. Dairy Sci. 100:6371–6375. https://doi.org/10.3168/jds.2017-12663.
Sundekilde, U. K., P. Frederiksen, M. Clausen, L. Larsen, and H. Bertram. 2011. Relationship between the metabolite profile and technological properties of bovine milk from two dairy breeds elucidated by NMR-based metabolomics. J. Agric. Food Chem. 59:7360–7367. https://doi.org/10.1021/jf202057x.
Tiezzi, F., D. Pretto, M. De Marchi, M. Penasa, and M. Cassandro. 2013. Heritability and repeatability of milk coagulation properties predicted by mid-infrared spectroscopy during routine data recording, and their relationships with milk yield and quality traits. Animal 7:1592–1599. https://doi.org/10.1017/S1751731113001195.

Timmen, H., and S. Patton. 1988. Milk-fat globules—Fatty-acid composition, size and in vivo regulation of fat liquidity. Lipids 23:685–689.

Toffanin, V., M. Penasa, S. McParland, D. P. Berry, M. Cassandro, and M. De Marchi. 2015. Genetic parameters for milk mineral content and acidity predicted by mid-infrared spectroscopy in Holstein-Friesian cows. Animal 9:775–780. https://doi.org/10.1017/S1751731114003255.

Vallas, M., H. Bovenhuis, T. Kaart, K. Parna, H. Kiiman, and E. Parna. 2010. Genetic parameters for milk coagulation properties in Estonian Holstein cows. J. Dairy Sci. 93:3789–3796. https://doi.org/10.3168/jds.2009-2435.

van Hulzen, K. J., R. Sprong, R. van der Meer, and J. van Arendonk. 2009. Genetic and nongenetic variation in concentration of selenium, calcium, potassium, zinc, magnesium, and phosphorus in milk of Dutch Holstein-Friesian cows. J. Dairy Sci. 92:5754–5759. https://doi.org/10.3168/jds.2009-2406.

Visentin, G., A. McDermott, S. McParland, D. P. Berry, O. A. Kenny, A. Brodkorb, M. A. Fenelon, and M. De Marchi. 2015. Prediction of bovine milk technological traits from mid-infrared spectroscopy analysis in dairy cows. J. Dairy Sci. 98:6620–6629. https://doi.org/10.3168/jds.2015-9323.

Visentin, G., S. McParland, M. De Marchi, A. McDermott, M. A. Fenelon, M. Penasa, and D. P. Berry. 2017. Processing characteristics of dairy cow milk are moderately heritable. J. Dairy Sci. 100:6343–6355. https://doi.org/10.3168/jds.2017-12642.

APPENDIX

Table A1. Descriptive statistics, heritability (h^2), proportion of phenotypic variance explained by permanent environmental (PE) effect, and repeatability (t) of milk protein composition

Trait[1]	No.	Mean	SD	h^2	SE h^2	PE	SE PE	t
Proteins, g/100 g of milk								
PC	1,100,238	3.35	0.29	0.48	0.008	0.15	0.006	0.63
α-LA	1,097,749	0.13	0.02	0.41	0.053	0.18	0.044	0.59
β-LG	1,098,516	0.41	0.06	0.42	0.012	0.29	0.010	0.71
α_{S1}-CN	1,100,094	1.08	0.09	0.46	0.009	0.21	0.007	0.67
α_{S2}-CN	1,099,402	0.32	0.03	0.44	0.025	0.18	0.017	0.62
β-CN	1,099,217	0.99	0.08	0.44	0.009	0.23	0.008	0.67
κ-CN	1,098,790	0.29	0.03	0.43	0.025	0.18	0.018	0.61
ΣWP	1,099,055	0.56	0.08	0.46	0.011	0.27	0.010	0.73
ΣCN	1,099,594	2.70	0.24	0.48	0.007	0.15	0.006	0.63
Proteins, g/100 g of protein								
α-LA	1,099,358	3.97	0.34	0.35	0.006	0.094	0.005	0.44
β-LG	1,100,013	12.32	1.58	0.58	0.007	0.054	0.005	0.63
α_{S1}-CN	1,100,073	32.26	0.26	0.46	0.007	0.097	0.005	0.56
α_{S2}-CN	1,097,360	9.69	0.32	0.38	0.007	0.11	0.005	0.49
β-CN	1,099,976	29.75	1.05	0.38	0.006	0.097	0.005	0.48
κ-CN	1,099,282	8.74	0.43	0.38	0.007	0.11	0.005	0.49
ΣWP	1,100,030	16.73	1.69	0.60	0.007	0.052	0.006	0.65
ΣCN	1,099,731	80.81	1.23	0.46	0.007	0.091	0.005	0.55

[1]PC = protein content; ΣWP = sum of whey proteins; ΣCN = sum of caseins.

Table A2. Descriptive statistics, heritability (h^2), proportion of phenotypic variance explained by permanent environmental (PE) effect, and repeatability (t) of milk fatty acid composition

Trait[1]	No.	Mean	SD	h^2	SE h^2	PE	SE PE	t
Fatty acid, g/100 g of milk								
FC	1,100,238	3.73	0.46	0.36	0.006	0.081	0.005	0.44
SFA	1,099,073	2.61	0.39	0.41	0.007	0.11	0.005	0.52
MUFA	1,091,310	0.99	0.20	0.16	0.004	0.066	0.003	0.23
UFA	1,091,353	1.12	0.22	0.16	0.004	0.067	0.003	0.23
PUFA	1,098,459	0.12	0.03	0.20	0.035	0.20	0.039	0.40
ΣC4–C10	1,098,322	0.42	0.06	0.34	0.010	0.31	0.009	0.65
ΣC4–C12	1,098,394	0.52	0.08	0.37	0.008	0.21	0.007	0.58
C14:0	1,097,898	0.40	0.08	0.34	0.009	0.27	0.008	0.61
C16:0	1,098,428	1.06	0.24	0.40	0.006	0.090	0.005	0.49
C18:0	1,096,470	0.40	0.11	0.18	0.007	0.18	0.005	0.36
C18:1	1,090,057	0.87	0.18	0.14	0.004	0.070	0.003	0.21
Fatty acid, g/100 g of fat								
SFA	1,100,225	69.99	5.53	0.21	0.005	0.10	0.004	0.31
MUFA	1,100,201	26.91	4.77	0.23	0.005	0.10	0.004	0.33
UFA	1,100,200	30.44	5.20	0.24	0.005	0.11	0.004	0.35
PUFA	1,100,235	3.33	0.67	0.18	0.004	0.073	0.003	0.25
ΣC4–C10	1,099,969	11.39	1.17	0.27	0.010	0.093	0.004	0.36
ΣC4–C12	1,100,237	14.04	1.56	0.26	0.008	0.085	0.004	0.35
C14:0	1,100,238	10.84	1.93	0.19	0.009	0.062	0.003	0.25
C16:0	1,100,233	28.45	4.89	0.25	0.006	0.080	0.004	0.33
C18:0	1,100,153	10.70	2.76	0.20	0.007	0.053	0.003	0.25
C18:1	1,100,229	23.64	4.61	0.23	0.004	0.10	0.004	0.33

[1]FC = fat content; ΣC4–C10 = sum of C4 to C10 fatty acids; ΣC4–C12 = sum of C4 to C12 fatty acids.

Table A3. Descriptive statistics, heritability (h^2), proportion of phenotypic variance explained by permanent environmental (PE) effect, and repeatability (t) of milk mineral composition and citrate and lactose content

Trait	No.	Mean	SD	h^2	SE h^2	PE	SE PE	t
Mineral, mg/kg of milk								
Ca	982,586	1,161.4	92.6	0.50	0.008	0.12	0.006	0.62
P	979,649	1,007.0	77.5	0.56	0.008	0.15	0.007	0.71
K	983,218	1,473.8	104.5	0.53	0.008	0.12	0.007	0.65
Mg	982,517	100.5	7.2	0.52	0.008	0.15	0.007	0.67
Na	982,511	341.7	44.5	0.32	0.008	0.19	0.006	0.51
Other compounds, g/kg of milk								
Citrate	624,731	8.27	1.49	0.48	0.014	0.19	0.012	0.67
Lactose	1,018,276	49.6	1.95	0.47	0.007	0.14	0.006	0.61

3.3. Héritabilités au cours de la première lactation

L'étude précédente repose sur un modèle à répétabilité qui suppose que le caractère est le même tout au long de la lactation (corrélation génétique égale à 1 entre stades), et avec des variances constantes. C'est une hypothèse forte qu'il faut vérifier.

Pour avoir une estimation de l'évolution des composantes de la variance et des corrélations tout au long de la première lactation, nous avons appliqué un modèle **à régression aléatoire** (modèle 4 §1.6.2.2) sur les données utilisées avec le modèle à répétabilité dans l'article 2 du §3.3 (Sanchez *et al.*, 2018a), *i.e.* environ 1,1 million de performances (caractères de composition et de fromageabilité du lait) de 126 873 vaches primipares Montbéliardes du projet *From'MIR*.

Les effets individuels de la vache (génétique additif et environnement permanent) ont été modélisés en fonction du jour de lactation (entre 7 et 350 jours) par un polynôme de Legendre d'ordre 3. Le premier terme étant constant, le second étant une fonction linéaire du stade de lactation, ils ont une signification biologique, le premier représentant le **niveau moyen** et le second la **persistance**. On retrouve généralement cette interprétation biologique à partir des vecteurs propres de la matrice de variance-covariance entre jours de lactation : le premier vecteur propre ne contient que des termes positifs et reflète le niveau moyen ; le second vecteur propre contient des termes qui évoluent plus ou moins linéairement avec le stade et reflète la « persistance », c'est-à-dire la capacité à faire évoluer ses performances au cours de la lactation.

La variance résiduelle peut elle aussi varier et la fixer tandis que les autres variances sont susceptibles de varier conduit à des artefacts. Toutefois, une modélisation continue n'est pas simple à mettre en œuvre. Une alternative est de la fixer par intervalles, un nombre suffisant d'intervalles permettant de traduire ses variations au cours de la lactation. Ainsi, pour modéliser la variance résiduelle, dix stades de lactation ont été définis de la façon suivante :

- Stade 1 = entre 7 et 20 jours ;
- Stades 2 à 5 = tous les 20 jours entre 21 et 100 jours ;
- Stades 6 à 10 : tous les 50 jours entre 101 et 350 jours.

Les composantes de la variance au cours de la lactation ont été estimés avec le logiciel Wombat (Meyer, 2007). Les valeurs d'héritabilité et de corrélations génétiques ont ensuite été calculées par jour de lactation en appliquant les formules décrites en §1.6.2.2.

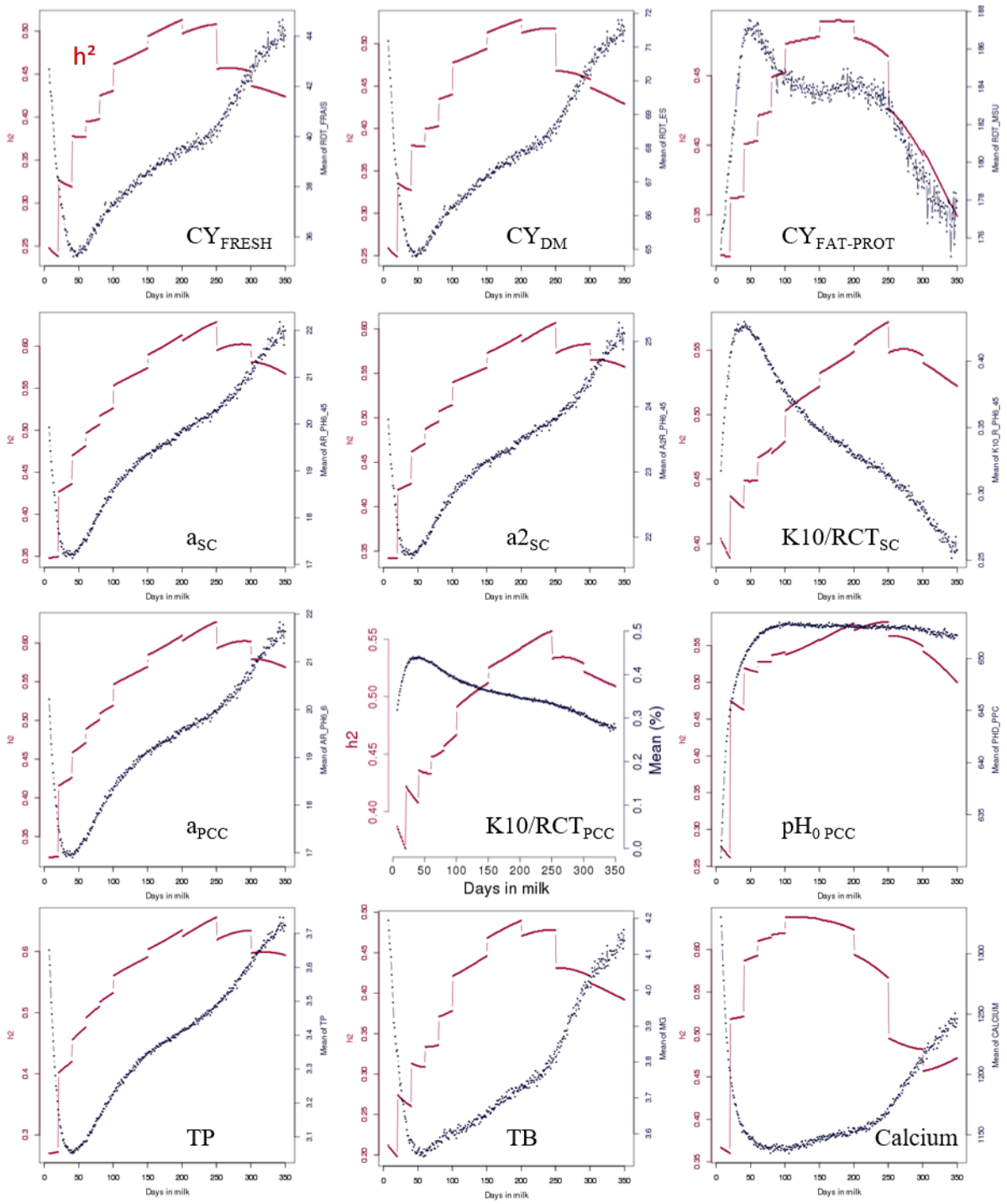


Figure 3.1. *Variations du phénotype (en bleu) et héritabilités (en rouge) entre 7 et 350 jours de lactation pour les neuf paramètres fromagers, le taux protéique (TP), le taux butyreux (TB) et le calcium*

Pour chaque caractère, la performance moyenne et l'héritabilité ont été calculées par jour de lactation (entre 7 et 350 jours). Les valeurs obtenues pour les caractères fromagers, le TP, le TB et le taux de calcium sont présentées dans la ***Figure 3.1***. Pour les autres caractères (composition en protéines, acides gras, minéraux, citrate et lactose), les courbes figurent en ***Annexe 2***. On note tout d'abord une évolution de la performance moyenne des caractères fromagers au cours de la lactation. L'évolution des caractères fromagers au cours de la lactation suit globalement celle du TP et du TB et dans une moindre mesure celle du calcium. Les rendements fromagers et les paramètres de coagulation, élevées en tout début de lactation, chutent fortement jusqu'au 50ème jour de lactation pour remonter régulièrement jusqu'en fin de lactation (350 j) et retrouver plus ou moins le niveau du début de lactation. Bien sûr, pour les caractères définis dans l'autre sens ($CY_{FAT-PROT}$ et K10/RCT), on observe la tendance inverse. Les écarts entre les valeurs extrêmes observées à 50 et 350 j de lactation sont assez importants, entre 1 et 2 écart-types phénotypiques selon le caractère fromager et environ 1,3 et 2 écart-types phénotypiques pour le TB et le TP, respectivement. Le pH du lait a un profil un peu différent, sa valeur est minimale en début de lactation (environ 6,30), elle atteint son maximum vers 100j (environ 6,55) et stagne jusqu'en fin de lactation (350j).

L'héritabilité varie également assez fortement au cours de la lactation, entre 0,25 et 0,50 pour les rendements et le TB et entre 0,30 et 0,65 pour les paramètres de coagulation et le TP. Le même profil de courbe en cloche, plus ou moins marqué, est observé pour tous les caractères de la ***Figure 3.1***. L'héritabilité est minimale en début de lactation, elle augmente ensuite pour atteindre sa valeur maximale autour de 200-250 jours de lactation et enfin, elle diminue légèrement (de manière plus prononcée pour $CY_{FAT-PROT}$) jusqu'en fin de lactation.

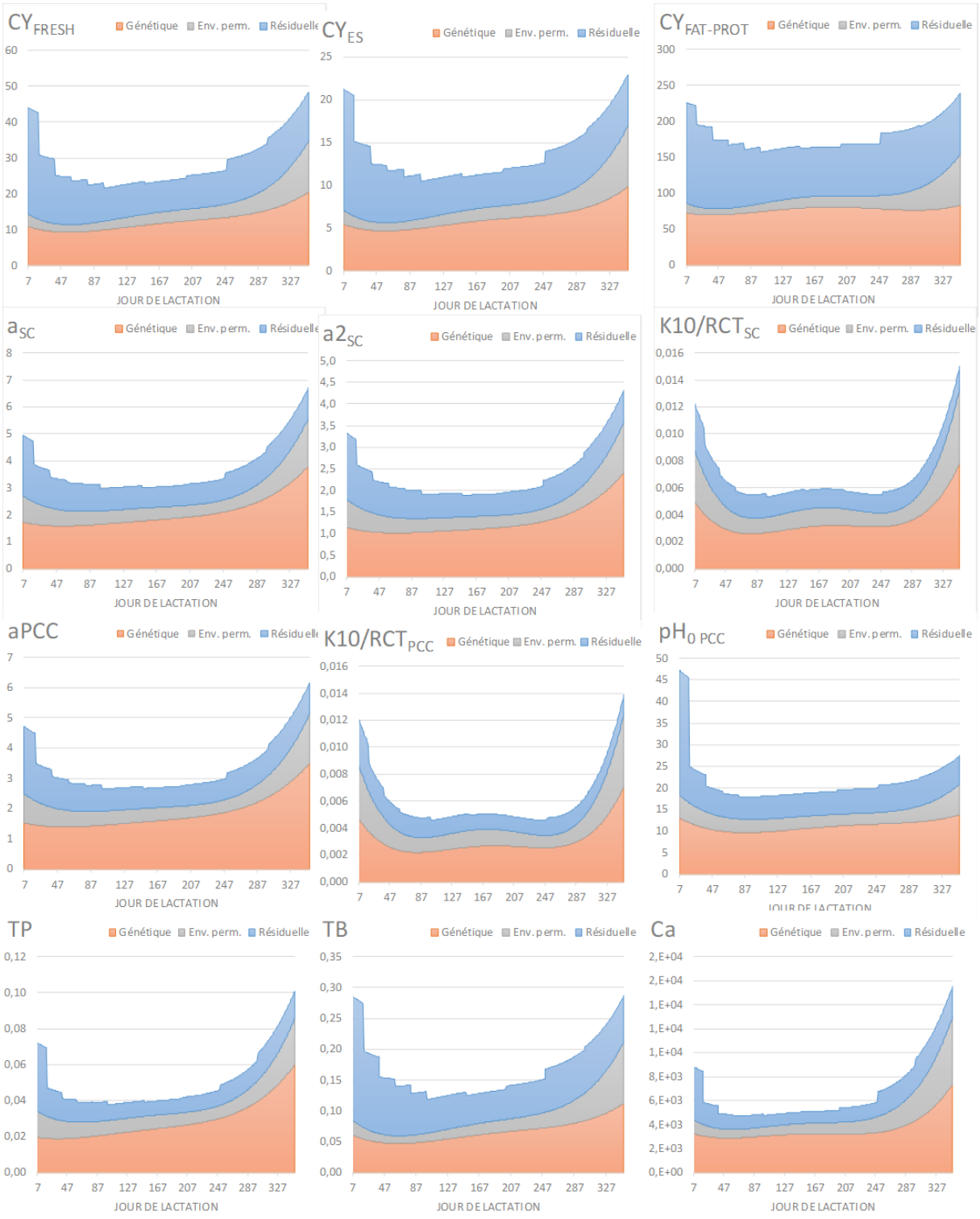


Figure 3.2. *Variances génétique, de l'environnement permanent et résiduelle estimées entre 7 et 350 jours de lactation pour les neuf paramètres fromagers, les taux protéique (TP), butyreux (TB) et de calcium*

La ***Figure 3.2*** représente l'évolution des variances au cours de la lactation pour les mêmes caractères. Les variances évoluent de manière similaire pour tous les caractères. La variance résiduelle est élevée et maximale au cours du premier stade de lactation (7-20j), en particulier pour le TP, le calcium et $pH_{0\ PCC}$. Dès le deuxième stade (21-40j), elle diminue fortement puis reste relativement stable jusqu'aux derniers stades de lactation où elle remonte légèrement. Pour tous les caractères, les variances génétiques et d'environnement permanent sont moins variables, elles ont tendance à augmenter légèrement et régulièrement pour atteindre une valeur maximale en fin de lactation (350j). Les variations de l'héritabilité au cours de la lactation sont donc principalement liées aux variations de la variance résiduelle.

Pour l'effet génétique additif, la première valeur propre de la matrice de variance-covariance, qui représente le niveau génétique moyen sur la lactation, explique en moyenne plus de 82% de la variabilité génétique des caractères (de 66 à 96% selon le caractère, ***Figure 3.3***). Ce dernier résultat montre que chaque caractère analysé est génétiquement assez homogène tout au long de la lactation. Pour des analyses génétiques, les modèles plus simples, à l'échelle de la lactation ou bien du contrôle avec un modèle à répétabilité (éventuellement en prenant en compte les différences de variance résiduelle) sont donc justifiés.

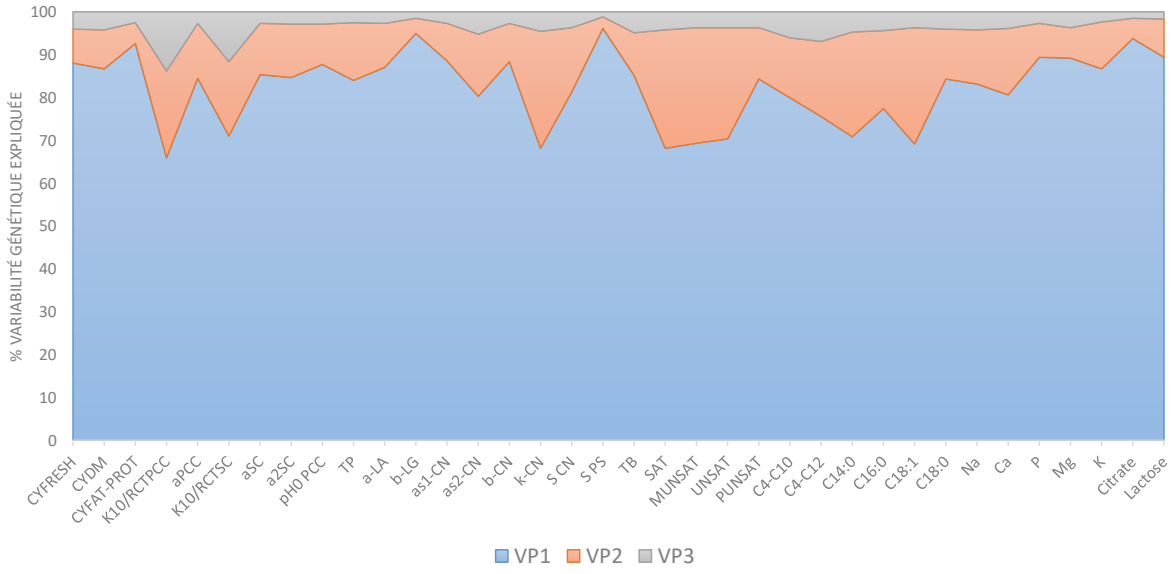


Figure 3.3. *Proportions relatives de la variabilité génétique expliquée par les trois premières valeurs propres (VP) de la matrice des variances-covariances entre jours de lactation*

La méthode de régression aléatoire permet d'estimer des corrélations génétiques entre n'importe quels jours de la lactation. Ces trajectoires de corrélations génétiques pour le rendement en extrait (CY_{DM}) entre les 7ème, 92ème, 178ème, 264ème et 350ème jours de lactation d'une part et tous les jours de la lactation (7-350j) d'autre part, sont représentées dans la ***Figure 3.4***. Toutes les corrélations génétiques sont comprises entre 0,55 et 1. Les corrélations entre stades proches sont toujours très élevées. Les courbes ont les niveaux les plus faibles pour CY_{DM} mesuré en tout début (7j) et en toute fin (350j) de lactation. Des trajectoires en cloche sont observées pour les stades intermédiaires avec des corrélations génétiques minimales en début ou en fin de lactation mais qui sont globalement élevées sur le reste de la lactation. Pour CY_{DM} mesuré à 92, 178 et 264 jours, les corrélations génétiques sont effectivement toujours supérieures à 0,85 entre 50 et 264 jours de lactation. Les mêmes profils de courbes, non présentés ici, sont observés pour les autres caractères fromagers et de composition du lait qui semblent donc être des caractères un peu différents génétiquement lorsqu'ils sont mesurés en début et en fin de lactation. Druet *et al.* (2005) obtiennent des résultats relativement similaires pour les caractères de production laitière en race Holstein avec des niveaux de corrélation globalement plus élevés sur la lactation. Toutefois, ces auteurs n'ont pas étudié la lactation au-delà de 305 jours.

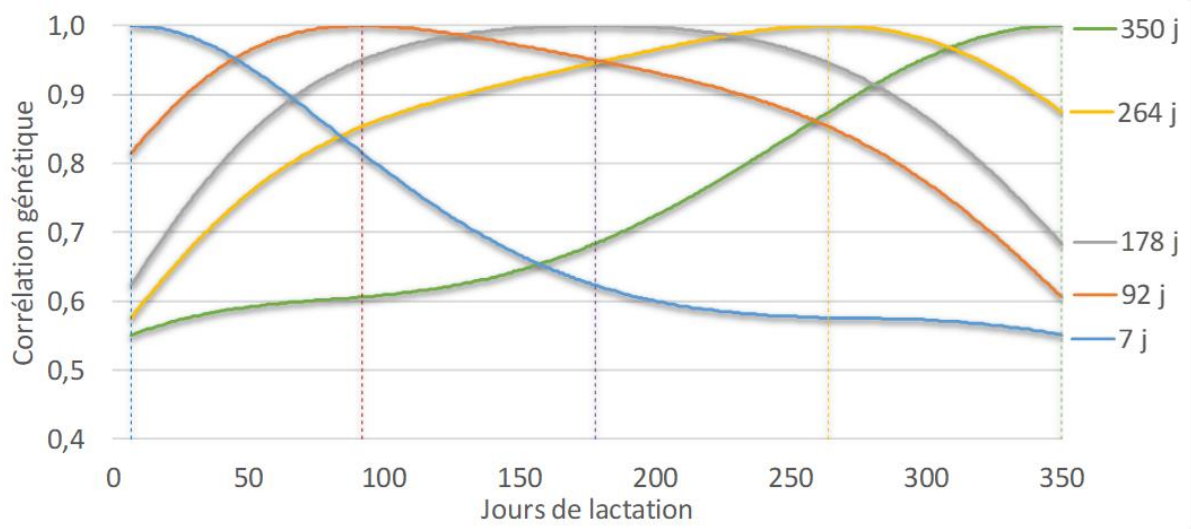


Figure 3.4. *Trajectoires de corrélations génétiques entre 7 et 350 jours de lactations pour le rendement en extrait sec (CY_{DM}) mesuré à 7, 92, 178, 264 et 350 jours*

3.4. Corrélations génétiques entre les trois premières lactations

La question posée sur l'unicité du déterminisme des caractères le long de la lactation se pose de la même façon entre lactations. Pour évaluer l'effet du rang de lactation sur le déterminisme génétique de la composition et de la fromageabilité du lait, nous avons estimé les corrélations génétiques entre les trois premières lactations. Un modèle de régression aléatoire étant difficilement applicable sur trois lactations avec les effectifs de cette étude, nous avons opté pour l'analyse de caractères définis par la moyenne des mesures à chaque contrôle, pour chacun des 54 phénotypes analysés. Chaque vache pouvait donc avoir une à trois observations. Le modèle utilisé, tri-caractères, considère pour un même phénotype un caractère différent par lactation (modèle 1 décrit au §1.6.2). En plus des effets aléatoires génétique additif et résiduel, le modèle incluait les effets fixés du nombre de contrôles (reflétant la durée de lactation ou le stade de lactation moyen), de la combinaison mois x année de vêlage et de la combinaison élevage x campagne de vêlage. Il avait été montré préalablement l'absence d'effet important de l'âge au vêlage. Le jeu de données analysé comprenait 145 861 vaches en L1 (avec un minimum de trois contrôles), parmi lesquelles 77 182 avaient également une L2 et 36 957 aussi une L3.

Les corrélations génétiques entre lactations ainsi que les valeurs d'héritabilité estimées par lactation sont dans le ***Tableau 3.1***. On note tout d'abord des valeurs d'héritabilité beaucoup plus élevées que celles estimées dans un modèle par contrôle qui sont dans les tableaux 2, A1, A2 et A3 de l'***Article 2*** du §3.3 (Sanchez *et al.*, 2018a). Elles sont très proches, quel que soit le critère fromager étudié (0,70 à 0,73 en L1 contre 0,37 à 0,48 avec un modèle par contrôle). En L1, elles varient entre 0,66 et 0,84 pour les teneurs en protéines (0,35 à 0,60 avec un modèle par contrôle), entre 0,44 et 0,72 pour les teneurs en AG (0,16 à 0,40 avec un modèle par contrôle) et entre 0,57 et 0,77 pour les minéraux, le citrate et le lactose (0,32 à 0,56 avec un modèle par contrôle). On observe par ailleurs pour presque tous les phénotypes une légère diminution de l'héritabilité entre la L1 et la L2 et entre la L2 et la L3. Les corrélations génétiques entre les lactations sont très fortes pour tous les caractères. En moyenne sur tous les caractères elles sont égales à 0,97 entre la L1 et la L2, 0,99 entre la L2 et la L3 et 0,96 entre la L1 et la L3.

Selon le modèle utilisé (par contrôle ou par lactation), les valeurs d'héritabilité peuvent donc être très différentes. Elles sont toujours beaucoup plus élevées quand elles sont estimées à l'échelle de la lactation en raison d'une variance résiduelle plus faible. En effet, la variance

résiduelle d'une moyenne est inférieure à la variance de chaque donnée élémentaire. Par contre, la variance génétique est peu différente. De ce fait, avec une réduction de son dénominateur, le coefficient d'héritabilité augmente.

Ces résultats montrent également que les gènes impliqués dans le déterminisme des caractères de composition ou de fromageabilité du lait sont les mêmes en première, deuxième et troisième lactation. D'un point de vue génétique, la performance mesurée en première lactation est donc représentative de la performance de la vache sur l'ensemble de sa carrière. Ces résultats valident l'utilisation d'un modèle à répétabilité entre lactations pour les caractères fromagers et de composition fine du lait.

Tableau 3.1. *Héritabilités (diagonale) et corrélations génétiques (hors diagonale) des caractères de composition et de fromageabilité du lait mesurés sur les trois premières lactations (L1, L2 et L3) [couleur bleue d'autant plus foncée que la corrélation est forte]*

Critères Fromagers			
CY_{FRESH}	L1	L2	L3
L1	0,699	0,980	0,972
L2		0,650	0,996
L3			0,632
CY_{DM}	L1	L2	L3
L1	0,716	0,982	0,974
L2		0,667	0,996
L3			0,649
$CY_{FAT-PROT}$	L1	L2	L3
L1	0,699	0,981	0,972
L2		0,648	0,994
L3			0,619
$K10/RCT_{SC}$	L1	L2	L3
L1	0,721	0,966	0,962
L2		0,671	0,996
L3			0,669
$K10/RCT_{PCC}$	L1	L2	L3
L1	0,709	0,964	0,960
L2		0,662	0,995
L3			0,655
a_{SC}	L1	L2	L3
L1	0,728	0,966	0,960
L2		0,682	0,996
L3			0,679
a_{PCC}	L1	L2	L3
L1	0,724	0,967	0,963
L2		0,676	0,996
L3			0,670
$a2_{SC}$	L1	L2	L3
L1	0,721	0,967	0,959
L2		0,678	0,996
L3			0,672
$pH_{0\ PCC}$	L1	L2	L3
L1	0,729	0,958	0,914
L2		0,686	0,983
L3			0,654

Protéines								
	g/100g lait			**g/100g protéines**				
TP	L1	L2	L3					
L1	0,723	0,963	0,962					
L2		0,665	0,994					
L3			0,675					
α-LA	L1	L2	L3	L1	L2	L3		
L1	0,675	0,959	0,952	0,656	0,974	0,952		
L2		0,619	0,993		0,643	0,995		
L3			0,612			0,628		
β-LG	L1	L2	L3	L1	L2	L3		
L1	0,793	0,978	0,974	0,841	0,988	0,979		
L2		0,759	0,996		0,828	0,997		
L3			0,763			0,817		
αs1-CN	L1	L2	L3	L1	L2	L3		
L1	0,724	0,964	0,962	0,739	0,979	0,965		
L2		0,666	0,995		0,740	0,994		
L3			0,676			0,736		
αs2-CN	L1	L2	L3	L1	L2	L3		
L1	0,706	0,962	0,958	0,663	0,971	0,953		
L2		0,651	0,995		0,636	0,993		
L3			0,662			0,641		
β-CN	L1	L2	L3	L1	L2	L3		
L1	0,747	0,969	0,966	0,663	0,969	0,956		
L2		0,710	0,996		0,605	0,993		
L3			0,700			0,578		
κ-CN	L1	L2	L3	L1	L2	L3		
L1	0,706	0,960	0,950	0,700	0,970	0,942		
L2		0,640	0,994		0,670	0,991		
L3			0,644			0,669		
Σ CN	L1	L2	L3	L1	L2	L3		
L1	0,726	0,963	0,960	0,734	0,972	0,951		
L2		0,665	0,995		0,694	0,993		
L3			0,673			0,641		
Σ PS	L1	L2	L3	L1	L2	L3		
L1	0,772	0,973	0,971	0,835	0,989	0,982		
L2		0,727	0,994		0,826	0,997		
L3			0,733			0,813		

Acides gras						
	g/100g lait			**g/100g matière grasse**		
TB	L1	L2	L3			
L1	0,711	0,987	0,979			
L2		0,684	0,996			
L3			0,669			
SAT	L1	L2	L3	L1	L2	L3
L1	0,720	0,987	0,975	0,507	0,967	0,931
L2		0,704	0,996		0,53	0,986
L3			0,684			0,541
MUNSAT	L1	L2	L3	L1	L2	L3
L1	0,463	0,968	0,951	0,548	0,976	0,951
L2		0,446	0,992		0,58	0,992
L3			0,459			0,574
UNSAT	L1	L2	L3	L1	L2	L3
L1	0,468	0,966	0,951	0,548	0,976	0,948
L2		0,45	0,992		0,58	0,992
L3			0,46			0,579
C4-C10	L1	L2	L3	L1	L2	L3
L1	0,711	0,990	0,979	0,621	0,971	0,950
L2		0,688	0,994		0,601	0,988
L3			0,662			0,586
C4-C12	L1	L2	L3	L1	L2	L3
L1	0,722	0,988	0,978	0,616	0,981	0,959
L2		0,704	0,995		0,611	0,988
L3			0,672			0,59
C14:0	L1	L2	L3	L1	L2	L3
L1	0,692	0,982	0,970	0,498	0,973	0,951
L2		0,677	0,996		0,514	0,995
L3			0,651			0,459
C16:0	L1	L2	L3	L1	L2	L3
L1	0,705	0,986	0,976	0,538	0,969	0,951
L2		0,688	0,995		0,543	0,989
L3			0,664			0,532
C18:1	L1	L2	L3	L1	L2	L3
L1	0,435	0,967	0,947	0,552	0,974	0,948
L2		0,427	0,991		0,585	0,992
L3			0,442			0,58
C18:0	L1	L2	L3	L1	L2	L3
L1	0,463	0,978	0,962	0,480	0,963	0,952
L2		0,464	0,991		0,461	0,987
L3			0,45			0,436
PUNSAT	L1	L2	L3	L1	L2	L3
L1	0,472	0,937	0,935	0,439	0,958	0,937
L2		0,424	0,991		0,447	0,991
L3			0,416			0,445

Minéraux			
Na	L1	L2	L3
L1	0,571	0,935	0,878
L2		0,589	0,974
L3			0,59
Ca	L1	L2	L3
L1	0,749	0,969	0,952
L2		0,685	0,994
L3			0,685
P	L1	L2	L3
L1	0,774	0,967	0,947
L2		0,704	0,995
L3			0,713
Mg	L1	L2	L3
L1	0,763	0,965	0,961
L2		0,700	0,991
L3			0,708
K	L1	L2	L3
L1	0,751	0,960	0,945
L2		0,739	0,991
L3			0,755
Lactose	L1	L2	L3
L1	0,704	0,956	0,931
L2		0,660	0,992
L3			0,634
Citrate	L1	L2	L3
L1	0,741	0,988	0,986
L2		0,750	0,995
L3			0,752

3.5. Corrélations génétiques avec les caractères en sélection

Enfin, pour étudier la liaison génétique entre les caractères fromagers et les caractères actuellement sélectionnés (***Tableau 3.2***), nous avons estimé les corrélations génétiques entre les deux types de caractères.

Tableau 3.2. *Description des caractères en sélection*

Nom du caractère	Description
LAIT	Quantité de lait par lactation
MP	Quantité de matière protéique par lactation
MG	Quantité de matière grasse par lactation
SCS	Moyenne des scores de cellules somatiques par lactation, le score par contrôle étant défini par $SCS = log_2\left(\frac{CCS}{100000}\right) + 3$ avec CCS = comptages de cellules somatiques
FERV	Fertilité post partum de la vache
DM	Développement de la mamelle
VTRA	Vitesse de traite

Les données moyennes de 135 191 vaches primipares ont été utilisées.

Pour les neuf variables fromagères, le modèle (2) a été appliqué avec les effets fixés suivants : troupeau x campagne, mois x année de vêlage et nombre de contrôles.

Pour les autres caractères mesurés dans le lait (LAIT, MP, MG et SCS), nous avons utilisé le même modèle auquel nous avons ajouté l'effet fixé de l'âge de la vache au moment du vêlage, découpé en 11 classes.

Concernant la fertilité (FERV), le développement de la mamelle (DM) et la vitesse de traite (VTRA), les variables sont dérivées du système d'indexation : les données brutes sont ajustées pour les effets non génétiques du modèle d'indexation spécifique de chaque variable. En conséquence, le modèle d'analyse de ces variables ne comprend qu'un effet année.

Tableau 3.3. *Corrélations génétiques entre les critères de fromageabilité du lait en première lactation et sept caractères en sélection*

r_a	LAIT	MP	MG	SCS	FERV	DM	VTRA	h^2
CY_{FRESH}	-0,43	-0,03	0,24	-0,08	0,01	-0,10	0,02	*0,70*
CY_{DM}	-0,42	-0,03	0,29	-0,07	0,01	-0,08	0,03	*0,71*
$CY_{FAT\text{-}PROT}$	0,32	0,05	-0,12	0,05	0,002	0,11	-0,01	*0,71*
$K10/RCT_{SC}$	0,43	-0,07	0,08	0,06	-0,06	0,10	0,07	*0,72*
$K10/RCT_{PCC}$	0,47	-0,06	0,07	0,09	-0,08	0,10	0,07	*0,71*
a_{SC}	-0,44	0,07	-0,06	-0,06	0,05	-0,09	-0,06	*0,73*
a_{PCC}	-0,46	0,06	-0,05	-0,09	0,06	-0,10	-0,05	*0,73*
$a2_{SC}$	-0,45	0,05	-0,08	-0,08	0,06	-0,10	-0,07	*0,72*
$pH_{0\ PCC}$	0,12	0,05	0,05	0,11	-0,05	0,09	-0,05	*0,74*
h^2	*0,35*	*0,30*	*0,35*	*0,22*	*0,06*	*0,47*	*0,21*	

Les corrélations génétiques les plus fortes sont observées avec la quantité de lait (***Tableau 3.3***). Elles sont défavorables avec les rendements fromagers et les paramètres de coagulation (entre 0,32 et 0,47 en valeur absolue). Génétiquement, la quantité de lait est associée à de moins bons rendements fromagers et à de moins bonnes aptitudes à la coagulation (vitesse de coagulation plus lente et caillé moins ferme), sans doute du fait de l'opposition entre quantité de lait et taux.

Dans une moindre mesure, les rendements frais et en extrait sec sont génétiquement corrélés à la quantité de matière grasse : 0,24 et 0,29, respectivement. Cela peut refléter la corrélation positive entre la MG par lactation et le TB, ce dernier étant un facteur important du rendement.

Enfin, il existe une corrélation génétique légèrement défavorable entre le caractère de morphologie DM (développement de la mamelle) et les caractères fromagers (0,10 environ en valeur absolue).

Toutes les autres corrélations génétiques entre les caractères fromagers d'une part et la quantité de matière protéique (MP), le score cellulaire (SCS), la fertilité de la vache (FERV) et la vitesse de traite (VTRA) d'autre part sont très faibles.

3.6. Bilan du chapitre 3

L'ensemble des résultats obtenus permet de répondre aux questions posées au début de ce chapitre.

Q1 - Les caractères de composition et de fromageabilité du lait prédits par la spectrométrie MIR sont-ils héritables ?

Nous avons estimé l'héritabilité de la composition et de la fromageabilité du lait à l'aide de plusieurs modèles différents et les résultats montrent que les caractères de composition et de fromageabilité du lait prédits par la spectrométrie MIR sont héritables. Pour les paramètres fromagers, les estimations d'héritabilité varient entre 0,37 et 0,48 à l'échelle du contrôle (0,70-0,73 par lactation). Les coefficients de variation génétiques varient quant à eux entre 3,3 et 13,8% pour les rendements et les paramètres de coagulation. De la même façon, pour les caractères de composition fine du lait, l'héritabilité est comprise entre 0,16 et 0,60 à l'échelle d'un contrôle alors qu'elle varie entre 0,44 et 0,84 à l'échelle de la lactation. Les teneurs en acides gras qui dépendent davantage du régime alimentaire de la vache (Legarto *et al.*, 2014) sont moins héritables que les teneurs en protéines et en minéraux. Dans tous les cas, les valeurs d'héritabilité sont du même ordre de grandeur, voire plus élevées que les valeurs estimées à partir des mesures de référence (Vallas *et al.*, 2010, Cecchinato *et al.*, 2011, Bittante *et al.*, 2013, Cecchinato *et al.*, 2015, Poulsen *et al.*, 2015). De plus, les héritabilités ainsi que les coefficients de variation génétiques (CV_{gen}), estimés pour les paramètres fromagers sont tout à fait comparables à ceux estimés pour le TP (h^2 = 0,48 ; CV_{gen} = 4,4%) ou le TB (h^2 = 0,36 ; CV_{gen} = 6,2%) qui sont efficacement sélectionnés depuis de nombreuses années.

Q2 – Quels sont les liens génétiques entre les différents critères de fromageabilité, et entre fromageabilité et composition du lait ?

Conformément aux résultats de la littérature (Cecchinato *et al.*, 2011, Bittante *et al.*, 2013, Cecchinato *et al.*, 2015, Cecchinato et Bittante, 2016, Visentin *et al.*, 2017), les corrélations génétiques entre les rendements fromagers et les aptitudes à la coagulation sont toujours favorables. De bons rendements sont génétiquement associés à de bonnes aptitudes à la coagulation et une coagulation rapide est également associée à un caillé plus ferme. En revanche, contrairement à Vallas *et al.* (2010) et Cecchinato *et al.* (2011), nous ne mettons pas en évidence de lien génétique entre le seul critère d'acidification que nous avons pu prédire via

les spectres MIR, *i.e.* le pH du lait, et les autres paramètres fromagers (rendements et coagulation). Toutefois, ce résultat peut s'expliquer par la procédure de standardisation du pH qui a été appliquée avant l'emprésurage du lait et la mesure des rendements et des paramètres de coagulation dans le projet *From'MIR*. Enfin, les aptitudes à la coagulation mesurées pour différentes technologies fromagères (pâte pressée cuite et pâte molle) sont aussi favorablement corrélées d'un point de vue génétique. La sélection génétique d'un seul critère fromager (par exemple le rendement en extrait sec) aura donc un effet favorable sur l'ensemble des critères fromagers que nous avons étudiés à la fois pour les fromages à pâte pressée cuite et les fromages à pâte molle.

La plupart des études qui ont réalisé une analyse génétique des critères fromagers ont estimé leurs corrélations génétiques avec le TP (ou les caséines totales), le TB et plus rarement le lactose mais les résultats varient beaucoup en fonction de l'étude et du critère fromager considéré. De plus, très peu d'études se sont intéressées aux corrélations génétiques entre la composition fine et la fromageabilité du lait. Seuls deux articles (Bonfatti *et al.*, 2011a, Gustavsson *et al.*, 2014) rapportent des corrélations génétiques entre les paramètres de coagulation et les teneurs en protéines du lait. Et bien que des liens phénotypiques aient été observés entre la fromageabilité et la composition du lait en acides gras et en minéraux (Amenu et Deeth, 2007), les corrélations génétiques entre ces caractères n'ont, à notre connaissance, jamais été estimées.

Notre étude permet de mettre en évidence des corrélations génétiques très fortes entre la fromageabilité du lait et sa composition. Les rendements fromagers sont très fortement génétiquement liés au TB et dans une moindre mesure au TP. Les rendements fromagers dépendent des quantités de matière (protéique et grasse) présentes dans le volume de lait utilisé pour faire le fromage. Le TB étant plus variable (CV_{gen} = 6,2%) que le TP (CV_{gen} = 4,4%), les covariances génétiques et donc les corrélations génétiques sont plus fortes entre les rendements et le TB qu'entre les rendements et le TP. Lorsqu'on s'intéresse à la composition fine en acides gras, on observe que les rendements fromagers sont plus fortement liés aux acides gras saturés et en particulier, à l'acide palmitique (C16:0).

Les paramètres de coagulation sont en revanche plus fortement génétiquement liés au TP qu'au TB. La déstructuration des micelles de caséines étant le processus responsable de la formation d'un gel et donc de la coagulation du lait, il n'est pas étonnant de trouver un lien génétique

étroit entre la teneur en protéines du lait (80% de caséines) et les paramètres de coagulation. Exprimées en pourcentage de protéines dans le lait, les deux caséines les plus abondantes (αs1-CN et β-CN) sont les protéines les plus fortement corrélées aux paramètres de coagulation. Au contraire, lorsqu'on exprime les teneurs en protéines en pourcentage de protéines totales, le lien génétique le plus étroit avec les paramètres de coagulation est obtenu avec les deux autres caséines, notamment la κ-CN qui, parce qu'elle est localisée à la surface des micelles, joue un rôle particulier dans leur déstructuration.

Enfin, le calcium, le phosphore et le magnésium sont les minéraux qui présentent les corrélations génétiques les plus élevées avec les rendements et les paramètres de coagulation (de l'ordre de 0,40-0,50 en valeur absolue). Ces résultats sont également très cohérents avec les mécanismes impliqués dans la coagulation du lait puisqu'une grande partie du calcium, du phosphore et du magnésium est associée aux caséines dans les micelles. Et malgré les associations phénotypiques observées entre les critères fromagers et les autres minéraux (potassium et sodium), le citrate et le lactose (Glantz *et al.*, 2010, Sundekilde *et al.*, 2011, Bland *et al.*, 2015), on observe des corrélations génétiques très faibles entre ces constituants du lait et les paramètres fromagers prédits à partir des spectres MIR dans le projet *From'MIR*.

Dans tous les cas et conformément aux précédentes études, de plus fortes teneurs en acides gras, protéines et minéraux dans le lait sont génétiquement associées à de meilleurs rendements fromagers et à de meilleures aptitudes à la coagulation du lait.

Q3 - Le déterminisme génétique de la composition et de la fromageabilité du lait varie-t-il au cours de la lactation ?

L'héritabilité des caractères fromagers varie au cours de la lactation. Elle est minimale en début de lactation puis augmente rapidement pour atteindre sa valeur maximale en milieu de lactation. Ces résultats sont sans doute liés à la forte variabilité de la composition du lait au cours des premières semaines de lactation et donc à la forte variance résiduelle. Nous observons d'ailleurs le même profil de variation pour la composition du lait et notamment pour le TP et le TB, ce qui est conforme aux résultats obtenus par Druet *et al.* (2005). Peu de résultats sont par contre publiés à ce jour pour les caractères fromagers. A notre connaissance, une seule étude a appliqué un modèle de régression aléatoire sur deux paramètres de coagulation (RCT et a30) prédits par spectrométrie MIR en race Holstein (Pretto *et al.*, 2014). Les résultats de cette étude sont tout

à fait cohérents avec ceux que nous obtenons dans le projet *From'MIR*. Par ailleurs, pour chacun des caractères que nous avons analysés, nous montrons que le déterminisme génétique reste essentiellement le même tout au long de la lactation, comme le montre le premier vecteur propre de la matrice des corrélations entre stades (plus de 82% en moyenne sur tous les caractères). Ce résultat nous autorise, pour des analyses génétiques, à raisonner à l'échelle de la lactation ou avec un modèle sur contrôles élémentaires à répétabilité.

Q4 - Les performances de composition et de fromageabilité du lait mesurées en première lactation sont-elles représentatives de la performance de la vache sur l'ensemble de sa carrière ?

Pour tous les caractères de composition et de fromageabilité du lait, les corrélations génétiques sont très élevées entre les trois premières lactations. Ce résultat, cohérent avec ceux obtenus précédemment pour ce type de caractères (Druet *et al.*, 2005, Pretto *et al.*, 2014), suggère que les gènes impliqués dans le déterminisme de ces caractères sont les mêmes pour les vaches primipares ou multipares. La composition du lait et sa fromageabilité mesurées en première lactation sont donc représentatives du potentiel génétique de la vache sur l'ensemble de sa carrière (*a minima* les trois premières lactations). Pour des analyses génétiques, on peut donc sans aucun risque se contenter des performances mesurées en première lactation ou analyser l'ensemble des données avec un modèle à répétabilité.

Q5 - Existe-t-il des antagonismes génétiques entre la fromageabilité du lait et les caractères sélectionnés ?

Enfin, nous montrons qu'il n'existe aucun antagonisme génétique entre les paramètres fromagers et les caractères actuellement pris en compte dans l'index de synthèse racial en race Montbéliarde (ISU), en particulier le TP, le TB, la matière grasse, la matière protéique, les cellules somatiques, la fertilité ou la vitesse de traite. Toutefois, le lait des vaches les plus productives présente de moins bonnes aptitudes fromagères, ce qui est probablement lié à l'antagonisme génétique entre quantité de lait et taux et donc à un effet de dilution des matières protéique et grasse.

Chapitre 4

Gènes et variants candidats

4. Gènes et variants candidats

Dans les deux projets *PhénoFinlait* et *From'MIR*, une partie des vaches avec spectres MIR a été génotypée, essentiellement pour la puce 50K. Les génotypages de 8010 vaches (2967, 2737 et 2306 en races Montbéliarde, Normande et Holstein, respectivement) ont été réalisés entre 2010 et 2013 dans le cadre du projet *PhénoFinlait* (7530 au total pour la puce 50K et 480 pour la puce LD). Le projet *From'MIR* plus récent (2015-2018) a pu bénéficier des génotypages de 19 862 vaches effectués pour la sélection génomique par les éleveurs adhérents à l'entreprise Umotest (6505 pour la puce 50K et 13 357 pour la puce EuroG10K).

Dans ce chapitre, nous avons mis à profit ces données de génotypage pour identifier les gènes et les variants candidats ainsi que les réseaux de gènes qui affectent la composition et la fromageabilité du lait.

§4.1. Dans un premier temps, nous avons estimé les fréquences des variants des gènes des protéines du lait dans les trois principales races bovines laitières Montbéliarde, Normande et Holstein, travail réalisé dans le cadre d'un stage de deuxième année de Licence que j'ai encadré entre juin et août 2017 (Rahman, 2017).

§4.2. Nous avons ensuite recherché des QTL pour la composition protéique dans les races Montbéliarde, Normande et Holstein : analyses LDLA à partir des génotypes 50K (***Article 3***).

§4.3. A partir des données analysées au §4.2, nous avons utilisé les données du projet 1000 génomes pour imputer les séquences complètes des vaches et réaliser des GWAS intra-race et multi-race sur ces séquences imputées (***Article 4***).

§4.4. Les variants candidats détectés pour la composition protéique ont été ajoutés à la partie recherche de la puce EuroG10K et nous avons pu confirmer l'effet de la plupart de ces variants sur la composition et les aptitudes fromagères du lait prédits *via* les spectres MIR dans la population Montbéliarde du projet *From'MIR* (***Article 5***).

§4.5. Enfin, la dernière partie de ce chapitre est consacrée aux analyses GWAS et à l'analyse de réseaux de gènes sur les caractères de composition (protéines, acides gras, minéraux, citrate et lactose) et de fromageabilité du lait en race Montbéliarde (***Article 6***).

4.1. Variants génétiques des lactoprotéines en races Montbéliarde, Normande et Holstein

Bien avant la disponibilité des marqueurs moléculaires sur l'ADN, le polymorphisme génétique des principales protéines du lait de vache (lactoprotéines) a été identifié directement sur les protéines par électrophorèse (voir la revue de Grosclaude (1988)). Cette technique permet de distinguer les formes ou variants d'une même protéine qui diffèrent par leur charge ou leur taille et qui peuvent correspondre à des variations de la séquence des acides aminés et donc des gènes qui codent pour ces protéines. Les variants des lactoprotéines ont très tôt suscité l'intérêt car ils peuvent avoir des effets majeurs sur la production, la composition et la fromageabilité du lait. Aujourd'hui, on connait les gènes et les variations dans ces gènes qui induisent les variants des protéines du lait de vache (Martin *et al.*, 2002). Les gènes des caséines sont groupés dans une région de 250 kb sur le chromosome 6 (entre 87,1 et 87,4 Mb). On trouve dans l'ordre, les gènes des caséines αs1 (*CSN1S1*), β (*CSN2*), αs2 (*CSN1S2*) et κ (*CSN3*). Les gènes qui codent pour les protéines sériques α-lactalbumine et β-lactoglobuline sont respectivement situés sur les chromosomes 5 (*LALBA*) à 31,3 Mb et 11 (*PAEP*) à 103,3 Mb. Peu de variants existent dans les gènes *LALBA* et *CSN1S2* dans les races bovines et en particulier dans les races laitières françaises alors que les quatre autres gènes (*CSN1S1*, *CSN2*, *CSN3* et *PAEP*) présentent de nombreux polymorphismes.

4.1.1 Caractérisation des variants protéiques à partir des génotypes aux SNP

En raison de leurs effets avérés ou potentiels sur les caractères laitiers, plusieurs SNP localisés dans les gènes *CSN1S1*, *CSN2*, *CSN3* et *PAEP* sont présents sur la partie recherche de la puce EuroG10K (Boichard *et al.*, 2018). Treize de ces SNP présents sur la puce sont non synonymes et induisent un changement d'un acide aminé de la protéine. Ils nous permettent de caractériser les principaux variants protéiques (3 à 5) de chaque gène (***Tableau 4.1***) sur de grandes populations.

Tableau 4.1. *Gènes des lactoprotéines avec SNP non synonymes et polymorphes présents sur la puce EuroG10K et noms des variants protéiques associés*

Gène	Nombre de SNP	Noms des variants protéiques
CSN1S1	2	B, C, D
CSN2	4	A1, A2, A3, B, I
CSN3	2	A, B, E
PAEP	5	A, B, C, D, W

Les caractéristiques des 13 SNP sont en ***Annexe 3***. La ***Figure 4.1*** représente les arbres qui nous permettent de caractériser les variants protéiques à partir des génotypes aux SNP. Un variant protéique est parfois le résultat de la combinaison de plusieurs mutations proches dans un gène (changement de plusieurs acides aminés).

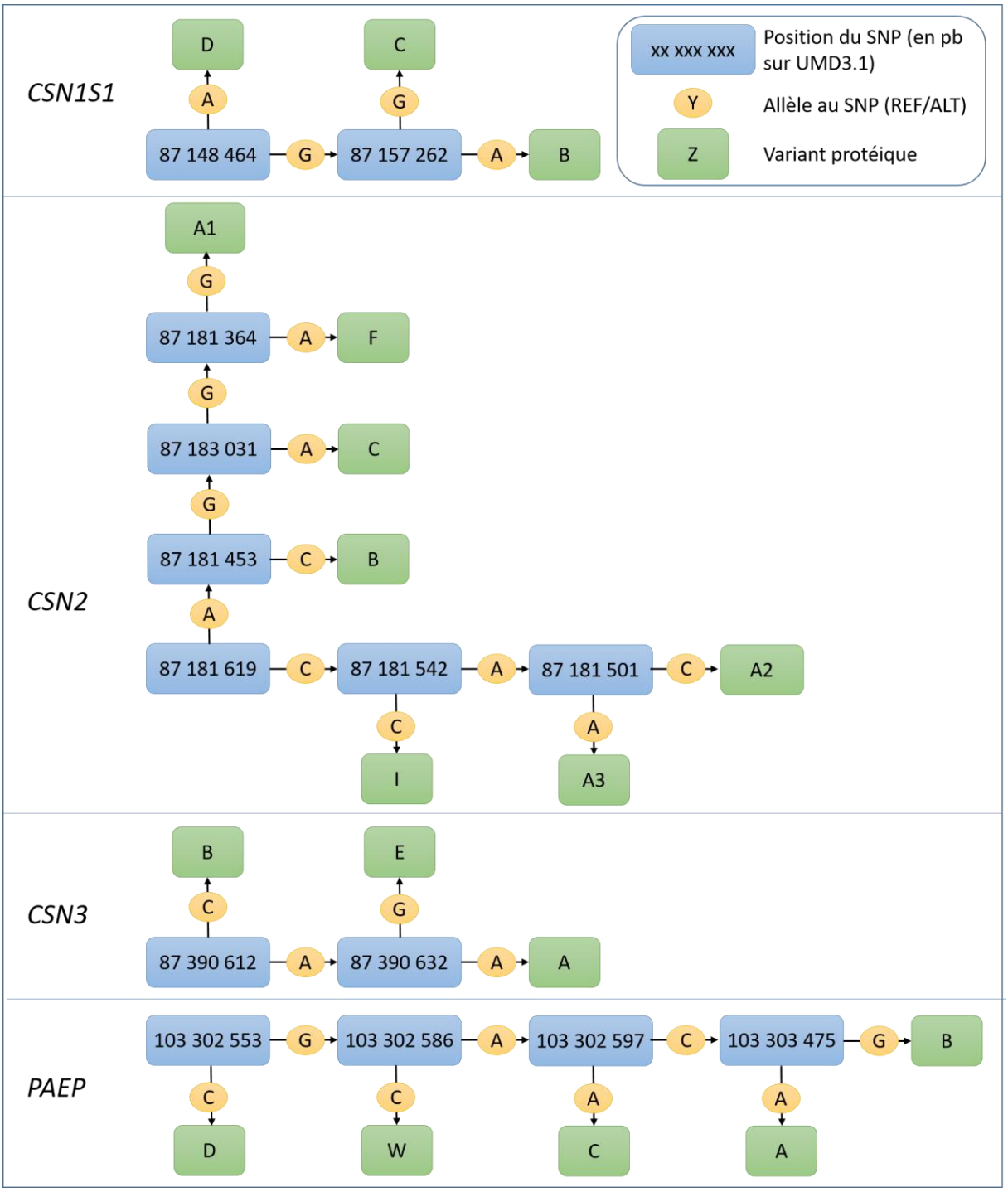


Figure 4.1. *Arbres de décision pour la caractérisation des variants protéiques dans les quatre gènes CSN1S1, CSN2, CSN3 et PAEP (pour le gène CSN2, avec les SNP disponibles, il n'a pas été possible de distinguer les variants A1, C et F).*

4.1.2. Inventaire et fréquences des variants protéiques dans les trois races

Les génotypes aux 13 SNP, disponibles pour un grand nombre d'animaux de toutes les races bovines génotypées pour la sélection génomique, dont les races laitières Montbéliarde (n = 132 387), Normande (n = 39 303) et Holstein (n = 180 711), permettent d'estimer la fréquence des variants protéiques dans les populations actuelles (***Figure 4.2***). A noter qu'avec les génotypages disponibles pour *CSN2*, il n'a pas été possible de distinguer les variants A1, C et F. De même, le polymorphisme responsable du variant D de *CSN3* n'est pas encore sur la puce.

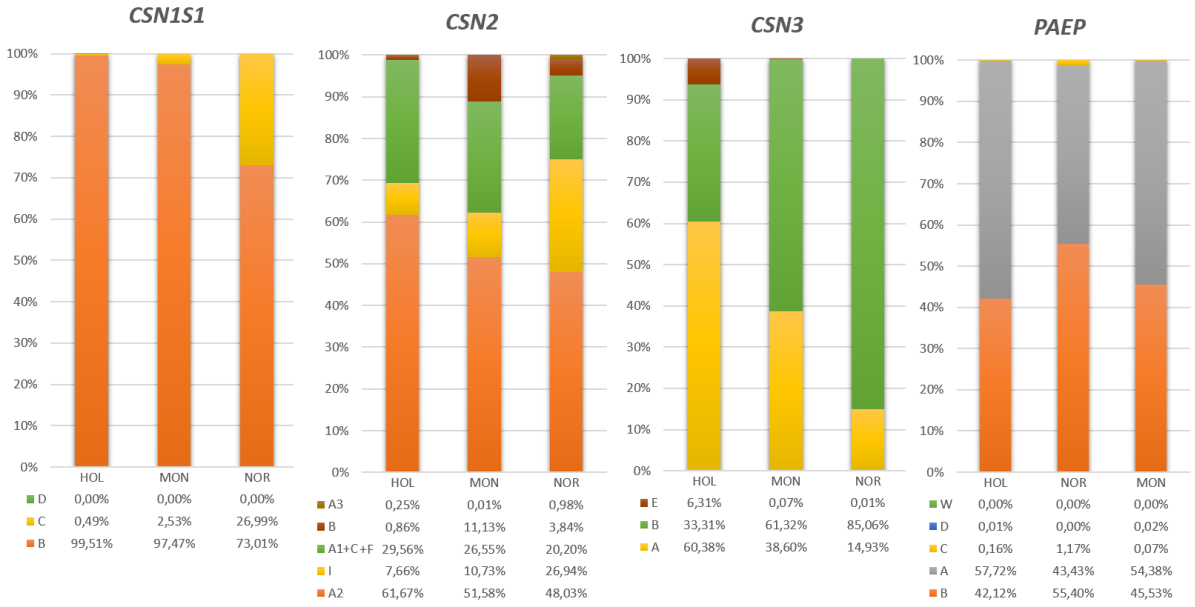


Figure 4.2. *Fréquence des variants protéiques dans les trois races Holstein (HOL), Montbéliarde (MON) et Normande (NOR)*

Dans le gène *CSN1S1*, deux variants seulement sont observés sur les trois connus et le variant B est largement majoritaire, surtout dans les races Holstein (99,5%) et Montbéliarde (97,5%). En race Normande, on trouve un deuxième variant C, en proportion assez importante (27%).

Dans le gène *CSN2*, cinq variants sont caractérisés. Le plus fréquent est le variant A2 dans les trois races (48 à 62%), puis le variant A1 (+ C et F) en races Holstein (30%) et Montbéliarde (27%) et le variant I (27%) en race Normande. Le variant B est trouvé surtout en race Montbéliarde (11%) et dans une moindre mesure en race Normande (4%). Enfin, un cinquième variant, A3, est également présent mais sa fréquence est inférieure à 1% dans les trois races.

Dans le gène *CSN3*, il y a deux variants majoritaires : A et B. Le variant A est le plus fréquent en race Holstein (60%) tandis que le variant B est trouvé majoritairement en races Montbéliarde (61%) et Normande (85%). Un troisième variant E est présent surtout en race Holstein (6%).

Enfin, dans le gène *PAEP*, deux variants A et B coexistent avec des fréquences relativement équilibrées dans les trois races. Un troisième variant C est présent mais très peu représenté dans les trois races (fréquence maximale = 1,2% en race Normande).

4.1.3. Evolution des fréquences des variants protéiques dans les trois races

Les tests relatifs aux variants des protéines du lait sont présents sur la partie recherche des puces EuroG10k utilisées à partir de 2013. Ils ne sont donc disponibles que pour des animaux jeunes. Cependant, pour estimer l'évolution des fréquences des variants au cours du temps, il est possible d'imputer les génotypes de ces SNP pour les animaux plus vieux, génotypés à partir d'une autre puce, LD ou 50K par exemple. Ce type d'imputation, dite *reverse* est la stratégie utilisée pour étudier l'effet des variants candidats de la partie recherche de la puce EuroG10K. Elle permet en effet d'imputer les génotypes des variants pour les animaux les plus vieux qui ont des performances. Cette stratégie, décrite en détail dans l'***Article 5*** (§4.5), nous a permis d'imputer les génotypes des 13 SNP pour l'ensemble des taureaux des trois races Montbéliarde (n=3224), Normande (n=2822) et Holstein (n=8843).

L'évolution des fréquences, calculées par année de naissance des taureaux (1995 – 2015), nous permet d'évaluer l'impact de la sélection au cours du temps (***Figure 4.3***). Globalement, les fréquences des variants de la caséine αs1 varient peu au fil des ans. Pour la caséine β, la fréquence du variant I, très faible dans les années 1990, augmente régulièrement jusqu'à aujourd'hui pour atteindre 17% en race Montbéliarde et 36% en race Normande alors qu'elle est relativement stable en race Holstein. Les fréquences des variants A et B de la caséine κ évoluent très peu en race Normande (B majoritaire) tandis que dans les deux autres races, la fréquence du variant B, plus faible au départ, a tendance à augmenter régulièrement au détriment de la fréquence du variant A. On observe sensiblement la même tendance pour les fréquences des variants A et B de la β lactoglobuline : elles sont stables en race Normande alors que le variant A, légèrement minoritaire dans les années 1990, est aujourd'hui le plus fréquent dans les deux autres races.

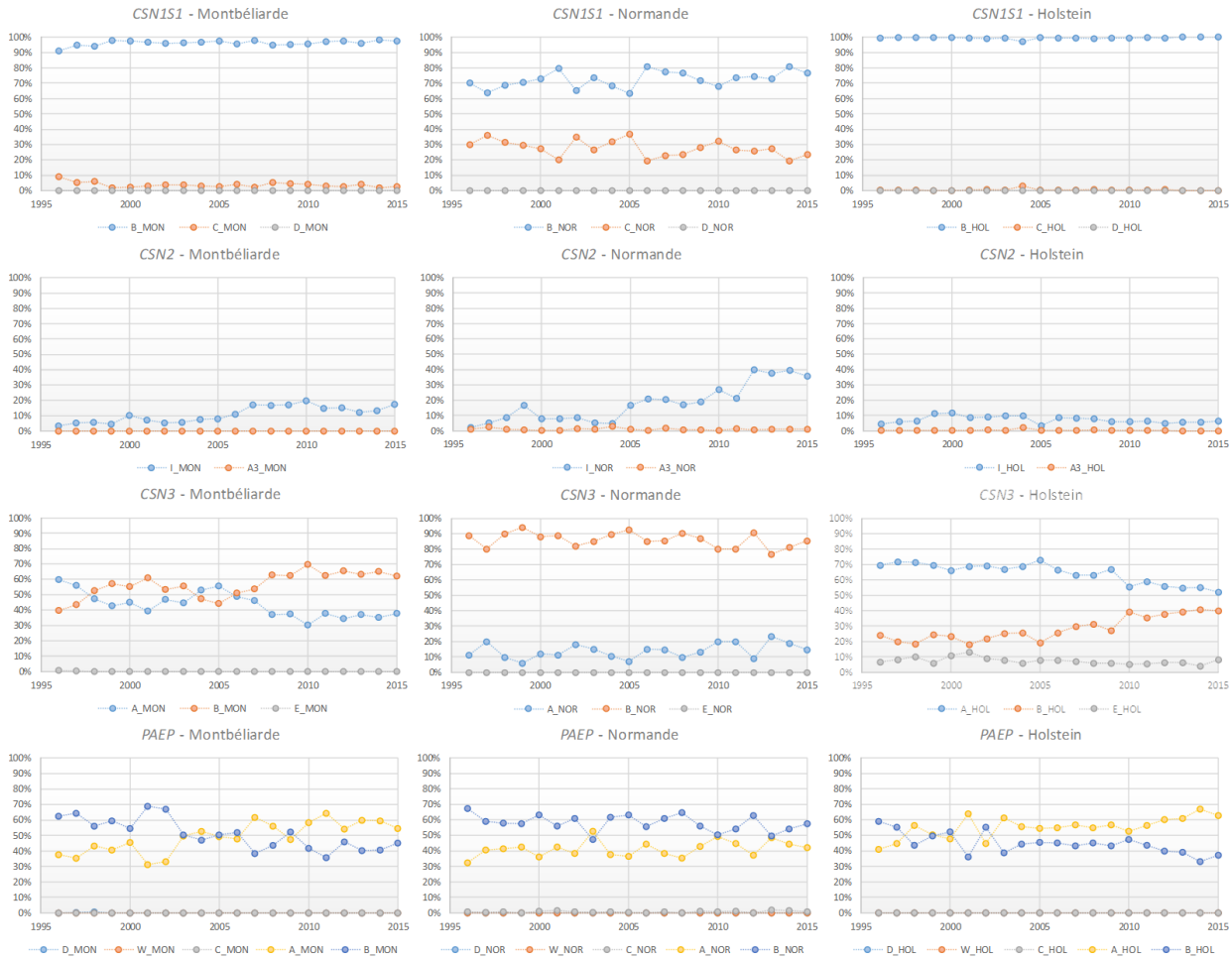


Figure 4.3. *Evolution des fréquences des variants protéiques par année de naissance des taureaux de race Montbéliarde (MON), Normande (NOR) et Holstein (HOL)*

4.1.4. Effets des variants protéiques sur les caractères de production laitière

Pour vérifier si l'évolution des fréquences des variants pouvait s'expliquer par la sélection sur les caractères de production laitière, nous avons testé l'effet des génotypes aux variants protéiques sur les quantités de lait (LAIT), de matière grasse (MG) et de matière protéique (MP), ainsi que sur les taux protéique (TP) et butyreux (TB) dans les trois races. Les performances analysées étaient les DYD (*Daughter Yield Deviation*) des taureaux, *i.e.* la moyenne des performances de leurs filles corrigées pour les effets fixés de milieu ainsi que pour le niveau génétique moyen de leurs mères. Le modèle, testé via la procédure GLM du logiciel SAS, comprenait l'effet des génotypes aux variants protéiques de chaque gène, l'effet de la mutation *K232A* du gène *DGAT1* connue pour ses effets très forts sur la composition et la production laitière (Grisart *et al.*, 2002) ainsi que l'effet du père pour prendre en compte les structures familiales de la population.

On a ainsi pu comparer les effets de certains variants entre eux dans certaines races et mis en évidence des effets significatifs. Les principaux résultats, synthétisés dans le ***Tableau 4.2***, montrent que les variants dont la fréquence augmente ont des effets positifs sur le LAIT (race Holstein) ou sur la MP et le TP (toutes races), ce qui est relativement cohérent avec la sélection pratiquée dans les trois grandes races laitières (voir formules des index économiques laitiers §1.6.4).

Tableau 4.2. *Effets comparés de certains variants protéiques sur les caractères de production laitière*

Gène	Variants	Races	LAIT	MG	MP	TB	TP
CSNS1	C vs B	Normande	ns*	ns	+	ns	+
CSN2	I vs A1, A2	3 races	ns	ns	+	+	+
CSN3	B vs A	Holstein / Normande	ns	ns	+ / ns	ns	+ / +
CSN3	E vs A, B	Holstein	+	ns	ns	-	-
PAEP	A vs B	3 races	ns	-	+	-	+

* non significatif ($P < 0.05$)

Pour les caséines, nous confirmons donc les effets favorables sur la teneur totale en protéines dans le lait du variant C de la caséine αs1 (Grosclaude, 1988), du variant I de la caséine β (Visker *et al.*, 2011, Vallas *et al.*, 2012) et du variant B de la caséine κ (Heck *et al.*, 2009). Les polymorphismes de chacune de ces caséines ont pour effet d'augmenter leur propre taux dans le lait, ce qui entraine également des effets favorables sur la coagulation du lait lors de sa transformation en fromage (Caroli *et al.*, 2009). En revanche, le variant E de la caséine κ qui dans notre étude a un effet positif sur la quantité de lait et des effets négatifs sur le TB et le TP en race Holstein (fréquence = 6%), a été associé à des effets défavorables sur les critères de coagulation du lait (Caroli *et al.*, 2009).

Le variant A de la β-lactoglobuline a un effet positif sur le TP et la MP et négatif sur le TB et la MG. Plusieurs études ont montré que ce variant est associé à des teneurs élevées en protéines sériques du lait (β-lactoglobuline et α-lactalbumine) et à des teneurs faibles en caséines, de sorte que globalement le TP peut ne pas être affecté (Lunden *et al.*, 1997, Bobe *et al.*, 1999, Heck *et al.*, 2009). Dans les trois races que nous avons étudiées, l'effet positif du variant A sur le TP et la MP est probablement responsable de l'augmentation très nette de sa fréquence au cours des 20 dernières années en race Normande (+9,6%), Montbéliarde (+17,3%) et Holstein (+22%). En revanche, malgré l'effet favorable sur les protéines totales du lait, le variant A,

parce qu'il affecte négativement le taux de caséines, a des effets défavorables sur la fromageabilité du lait et notamment sur les paramètres de coagulation (Caroli *et al.*, 2009).

Ces résultats permettent d'actualiser les fréquences des différents variants protéiques dans les trois principales races laitières françaises. Si les fréquences calculées sur les taureaux nés en 1995 sont tout à fait cohérentes avec celles publiées à la fin des années 80 (Grosclaude, 1988), nous montrons un état des lieux actuel sensiblement différent avec notamment l'émergence de trois nouveaux variants qui semblent avoir été sélectionnés indirectement depuis le milieu des années 2000 : le variant C de la caséine αs1, essentiellement en race Normande ; le variant I de la caséine β, présent dans les trois races et particulièrement fréquent en race Normande et enfin, le variant E de la caséine κ, présent surtout en race Holstein.

Cette étude met également en évidence l'augmentation de la fréquence de certains variants qui pourraient avoir des effets préjudiciables à la transformation du lait en fromage (variant E de la caséine κ et variant A de la β-lactoglobuline), y compris dans les races Normande et Montbéliarde dont le lait est utilisé pour produire de grandes quantités de fromages, notamment des fromages AOP.

4.2. Détection de QTL pour la composition protéique (puce 50K)

Identification of QTL and candidate mutations affecting major milk proteins in three French dairy cattle breeds.

M.P. Sanchez[1*], A. Govignon-Gion[†*], M. Ferrand[†], M. Gelé[†], D. Pourchet[‡], Y. Amigues[§], S. Fritz[*‖], M. Boussaha[*], A. Capitan[*‖], D. Rocha[*], G. Miranda[*], P. Martin[*], M. Brochard[†], D. Boichard[*]

[*] GABI, INRA, AgroParisTech, Université Paris Saclay, F-78350 Jouy-en-Josas, France

[†] Institut de l'Elevage, F-75012 Paris, France

[‡] ECEL Doubs - Territoire de Belfort, F-25640 Roulans, France

[§] LABOGENA DNA, F-78350 Jouy en Josas, France

[‖] Allice, F-75012 Paris, France

Journal of Dairy Science 2016. 99:8203–8215.

http://dx.doi.org/10.3168/jds.2016-11437

J. Dairy Sci. 99:8203–8215
http://dx.doi.org/10.3168/jds.2016-11437
© American Dairy Science Association®, 2016.

Whole-genome scan to detect quantitative trait loci associated with milk protein composition in 3 French dairy cattle breeds

M. P. Sanchez,*[1] A. Govignon-Gion,*† M. Ferrand,† M. Gelé,† D. Pourchet,‡ Y. Amigues,§ S. Fritz,*# M. Boussaha,* A. Capitan,*# D. Rocha,* G. Miranda,* P. Martin,* M. Brochard,† and D. Boichard*
*GABI, INRA, AgroParisTech, Université Paris Saclay, F-78350 Jouy-en-Josas, France
†Institut de l'Elevage, F-75012 Paris, France
‡Entreprise de Conseil en Elevage Doubs, Territoire de Belfort, F-25640 Roulans, France
§Labogena DNA, F-78350 Jouy en Josas, France
#Allice, F-75012 Paris, France

ABSTRACT

In the context of the PhénoFinLait project, a genome-wide analysis was performed to detect quantitative trait loci (QTL) that affect milk protein composition estimated using mid-infrared spectrometry in the Montbéliarde (MO), Normande (NO), and Holstein (HO) French dairy cattle breeds. The 6 main milk proteins (α-lactalbumin, β-lactoglobulin, and α_{S1}-, α_{S2}-, β-, and κ-caseins) expressed as grams per 100 g of milk (% of milk) or as grams per 100 g of protein (% of protein) were estimated in 848,068 test-day milk samples from 156,660 cows. Genotyping was performed for 2,773 MO, 2,673 NO, and 2,208 HO cows using the Illumina BovineSNP50 BeadChip (Illumina Inc., San Diego, CA). Individual test-day records were adjusted for environmental effects and then averaged per cow to define the phenotypes analyzed. Quantitative trait loci detection was performed within each breed using a linkage disequilibrium and linkage analysis approach. A total of 39 genomic regions distributed on 20 of the 29 *Bos taurus* autosomes (BTA) were significantly associated with milk protein composition at a genome-wide level of significance in at least 1 of the 3 breeds. The 9 most significant QTL were located on BTA2 (133 Mbp), BTA6 (38, 47, and 87 Mbp), BTA11 (103 Mbp), BTA14 (1.8 Mbp), BTA20 (32 and 58 Mbp), and BTA29 (8 Mbp). The BTA6 (87 Mbp), BTA11, and BTA20 (58 Mbp) QTL were found in all 3 breeds, and they had highly significant effects on κ-casein, β-lactoglobulin, and α-lactalbumin, expressed as a percentage of protein, respectively. Each of these QTL explained between 13% (BTA14) and 51% (BTA11) of the genetic variance of the trait. Many other QTL regions were also identified in at least one breed. They were located on 14 additional chromosomes (1, 3, 4, 5, 7, 15, 17, 19, 21, 22, 24, 25, 26, and 27), and they explained 2 to 8% of the genetic variance of 1 or more protein composition traits. Concordance analyses, performed between QTL status and sequence-derived polymorphisms from 13 bulls, revealed previously known causal polymorphisms in *LGB* (BTA11) and *GHR* (BTA20 at 32 Mbp) and excluded some other previously described mutations. These results constitute a first step in identifying causal mutations and using routinely collected mid-infrared predictions in future genomic selection programs to improve bovine milk protein composition.
Key words: dairy cattle, mid-infrared spectrometry, protein composition, quantitative trait loci

INTRODUCTION

Bovine milk contains 3 to 4% protein, which consists of about 80% caseins and 20% whey proteins. The relative fractions of protein play a key role in determining the functional properties of milk, such as clotting and cheese yield (Wedholm et al., 2006). The protein composition of bovine milk varies with different environmental factors (herd, season, stage of lactation, diet, and so forth), but it is mostly determined by genetic factors (Schopen et al., 2009; Bonfatti et al., 2011a; Gebreyesus et al., 2016).

Accurate genetic analyses of milk protein composition require large-scale studies. To date, these have been hampered by the complicated nature and cost of individual protein measurements in milk. Reference methods such as capillary zone electrophoresis are time consuming and expensive and have therefore only been applied to small or moderate numbers of milk samples. The largest QTL study, performed by Schopen et al. (2011), used more than 1,700 individual protein measurements obtained by liquid chromatography. More recently, mid-infrared (**MIR**) spectrometry has been shown to be useful for predicting milk protein composition (Bonfatti et al., 2011b; Ferrand et al., 2012), and it offers an alternative method for large-scale analyses.

Received May 10, 2016.
Accepted June 16, 2016.
[1]Corresponding author: marie-pierre.sanchez@jouy.inra.fr

Combined with high-throughput genotyping technologies, identifying the genomic regions responsible for genetic variation (QTL) in individual protein contents and accounting for these traits in genomic selection programs are now possible. PhénoFinLait is a major project that was initiated in 2008 to study milk in cattle, sheep, and goat dairy species (Gelé et al., 2014). One of its objectives was to dissect the genetic architecture of individual milk protein composition. In cattle, MIR predictive equations were derived from 450 reference samples analyzed using reverse-phase liquid chromatography. They were applied to the Montbéliarde (**MO**), Normande (**NO**), and Holstein (**HO**) French dairy breeds (Ferrand et al., 2012). The contents of the 6 major bovine milk proteins (α_{S1}-, α_{S2}-, β-, and κ-CN and α-LA and β-LG whey proteins) could be predicted with satisfactory accuracy. An initial study to estimate genetic parameters of these traits revealed relatively high heritability coefficients (Brochard et al., 2013) that were close to the values calculated from the protein contents estimated using a reference method (Schopen et al., 2009). These results suggested that the predictive accuracy of MIR protein content determinations was sufficiently accurate for genetic investigations.

To date, only one whole-genome association study using high-density SNP genotyping has been performed on milk protein composition in dairy cattle (Schopen et al., 2011). Those authors analyzed milk protein composition using capillary zone electrophoresis on 1,713 test-day samples from Dutch Holstein-Friesian cows. The present study reports the results of a whole-genome scan carried out on the lactation records of 2,773 MO, 2,673 NO, and 2,208 HO cows to identify QTL affecting individual protein composition predicted from MIR spectra. In addition, to validate effects of 10 known candidate mutations (located in 5 distinct genomic regions), concordance analyses between bull genotypes and QTL were carried out using data from the 1,000 bull genome project (Daetwyler et al., 2014).

MATERIALS AND METHODS

Animals, Milk Samples, and Phenotypic Data

In total, MIR spectra were collected on 848,068 milk samples from 156,660 cows collected between November 2009 and August 2012. These cows belonged to the 3 main French dairy breeds (MO, NO, and HO) and were selected according to their number of milk records. They were also distributed across 1,043 herds covering a broad range of geographical locations (16 small regions) and production systems (grass or maize silage, high or low input, conventional or organic, and so forth). In addition, to ensure broad genetic diversity, a list of AI bulls was set and herds were selected to maximize the number of daughters (hereinafter referred to as PhénoFinLait or **PFL** cows) born from these bulls and experiencing their first or second calving during the winter of 2009–2010. The MIR spectra were recovered from 5, 3, and 5 milk analysis laboratories for MO, NO, and HO cows, respectively, and all other information was retrieved from the National Information System. The whole milk protein content (**PC**) and milk protein composition were predicted using MIR spectra and equations previously derived by Ferrand et al. (2012). Individual protein contents were predicted for the 6 main milk proteins: α-LA and β-LG whey proteins and α_{S1}-, α_{S2}-, β-, and κ-CN and expressed as grams per 100 g milk (% of milk) or as grams per 100 g protein (% of protein). A total of 13 traits were analyzed: PC and the 6 proteins expressed as a percentage of milk and as a percentage of protein.

For each trait, a single phenotype per cow and across lactations was used in the QTL analysis. This phenotype was defined as the average per cow of test-day data, adjusted for all nongenetic effects. These nongenetic effects were estimated with the following breed-specific single trait model:

$$\mathbf{y} = \mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{Z}\mathbf{p} + \mathbf{e}, \qquad [1]$$

where μ is the overall mean, $\mathbf{y}$ is the vector of test-day observations, $\mathbf{a} \sim N\left(0, \mathbf{A}\sigma_a^2\right)$ is the vector of random genetic effects, $\mathbf{p} \sim N\left(0, \mathbf{I}\sigma_p^2\right)$ is the vector of random permanent environmental effects, and $\mathbf{e} \sim N\left(0, \mathbf{I}\sigma_e^2\right)$ is the vector of random residual effects, $\mathbf{X}$ and $\mathbf{Z}$ are incidence matrices, $\mathbf{A}$ is the relationship matrix among individuals, and $\mathbf{I}$ is the identity matrix. Variances σ_a^2, σ_p^2, and σ_e^2 are the additive genetic, permanent environmental, and residual variances, respectively. The vector $\boldsymbol{\beta}$ included the fixed effects of herd × test-day, parity × stage of lactation (63 levels), year × month of calving (34 levels), and spectrometer × test-month (102 levels). This model was applied to the first 3 lactations with at least 3 test-day records per lactation during the study period. In total, data from 344,542, 73,347, and 89,730 test-day records were analyzed, corresponding to 54,676 MO, 15,550 NO, and 17,983 HO cows, respectively (Table 1). All estimates were obtained using Genekit software (Ducrocq, 1998).

Genotyping

A total of 8,010 PFL cows (2,967 MO, 2,737 NO, and 2,306 HO) were genotyped with the Illumina BovineSNP50 BeadChip (Illumina Inc., San Diego, CA)

or, for a small group of them (6%), with the Illumina BovineLD BeadChip and subsequently imputed to Illumina BovineSNP50 genotypes. These cows were selected from the initial list of PFL cows by maximizing the number of test-days with MIR information while maintaining a good balance across breeds and sires. Quality control was applied to SNP mapped to the 29 autosomes (UMD3.1 assembly; Zimin et al., 2009). First, only samples with a call rate >95% were retained. Second, markers with a call frequency <90%, a minor allele frequency <5%, or genotype frequencies deviating from the Hardy–Weinberg equilibrium ($P <$ 1.E-4) were removed. After filtering, imputation and phasing were performed using DagPhase (Druet and Georges, 2010) as described by Boichard et al. (2012). Overall, 36,912, 37,363, and 39,683 SNP for 2,773 MO, 2,673 NO, and 2,208 HO cows, respectively, were used for QTL detection analyses (Table 1).

Table 1. Numbers of mid-infrared (MIR) spectra collected and cows in lactations 1 to 3

Breed	MIR spectra	Cows	Genotyped cows
Montbéliarde	344,542	54,676	2,773
Normande	73,347	15,550	2,673
Holstein	89,730	17,983	2,208
Total	507,619	88,209	7,654

QTL Detection Analyses

Because of the strong family structure in the design with planned half-sib families, we applied an approach maximizing detection power by accounting for linkage information within a family. This approach was an alternative to a simple association study in which the SNP association statistics have to be corrected for family structure (e.g., by using a genomic control). Therefore, after phase reconstruction, a linkage disequilibrium and linkage analysis approach combining the identity-by-descent method described by Meuwissen and Goddard (2000) and its extension to include linkage information (Meuwissen et al., 2002) was applied to detect QTL (Druet et al., 2008). This approach is a variance component mapping method that accounts for the transmission of haplotypes across generations and information from linkage disequilibrium between haplotypes. The method used to compute identity-by-descent probabilities and haplotype clustering was described by Druet et al. (2008). Each haplotype was defined by 6 adjacent markers. Then, a model including both polygenic and haplotype cluster effects was applied:

$$\mathbf{y}' = \mu + \mathbf{Z}_u\mathbf{u} + \mathbf{Z}_h\mathbf{h} + \mathbf{e}, \qquad [2]$$

where $\mathbf{y}'$ is the vector of preadjusted records derived from Equation [1] and averaged per cow; μ is the overall mean; $\mathbf{u} \sim N\left(0, \mathbf{A}\sigma_u^2\right)$ is the vector of random polygenic effects with $\mathbf{A}$ the matrix of additive genetic relationships among individuals, identical to the $\mathbf{A}$ matrix of Equation [1], and σ_u^2 the polygenic variance; $\mathbf{h} \sim N\left(0, \mathbf{H}\sigma_h^2\right)$ is the vector of random QTL effects corresponding to haplotype clusters with $\mathbf{H}$ the identity-by-descent matrix between the haplotype clusters and σ_h^2 the haplotype cluster variance; $\mathbf{e} \sim N\left(0, \mathbf{I}\sigma_e^2\right)$ is the vector of random residual effects with $\mathbf{I}$ the identity matrix and σ_e^2 the residual variance; $\mathbf{Z}_u$ and $\mathbf{Z}_h$ are incidence matrices relating phenotypes to corresponding polygenic and haplotype cluster effects, respectively.

The BLUPF90 software (Misztal et al., 2002) modified by Druet et al. (2008) was used to incorporate relationship matrices among QTL allele effects at each putative QTL position. Variance components $\left(\hat{\sigma}_u^2,\ \hat{\sigma}_h^2,\ \hat{\sigma}_e^2\right)$ were estimated maximizing likelihoods using an average information (AI)-REML approach. For each QTL, the proportion of total genetic variance due to the QTL was estimated:

$$\frac{\hat{\sigma}_h^2}{\hat{\sigma}_u^2 + \hat{\sigma}_h^2}.$$

The presence or absence of a QTL at a given position was tested by comparing Equation [2] to a polygenic model (i.e., Equation [2] without the term $\mathbf{Z}_h\mathbf{h}$) with the likelihood-ratio test statistic (**LRT**):

$$\text{LRT} = -2\ln\frac{L\left(H_0\right)}{L\left(H_1\right)},$$

where $L(H_0)$ and $L(H_1)$ were the maximum likelihood values when parameters were equal to their REML estimated values under the polygenic model with no QTL fitted (H_0) and under the alternative model with QTL (H_1), respectively. The distribution of the LRT had previously been described as in between the 1-df and the 2-df chi-squared distributions (Grignola et al., 1996).

Bonferroni correction was applied to the thresholds to account for the multiple testing of our analyses. Because adjacent haplotypes present high linkage disequilibrium, only nonoverlapping haplotypes were considered to define the number of independent tests. The haplotype definition with 6 markers resulted in 6,613 tests being considered for test correction. The 5%

genome-wide threshold of significance therefore corresponded to a nominal P-value of 7.6E-06 (0.05/6,613). A highly conservative approach was adopted by considering a 2-df chi-squared distribution for the null hypothesis, leading to a threshold of 23.6. An even more stringent threshold was also considered to highlight the most significant QTL (LRT $>$ 50; nominal P-value $<$ 1.4E-11), corresponding to the 9.3E-08 genome-wide threshold of significance.

Concordance Analyses Between QTL and Candidate Mutations

For highly significant QTL regions, and taking advantage of the half-sib structure, a concordance analysis (van den Berg et al., 2014) was performed in the regions of interest by comparing the sequence polymorphisms of sires with their QTL status. The following procedure was applied for each QTL region:

- Haplotype definition: In each breed, haplotypes of 15 SNP from the Bovine SNP50 BeadChip were formed considering the 14 SNP surrounding the SNP with the highest LRT value (SNPmax), that is, 7 SNP on the left side and 7 SNP on the right side. In the same QTL region, the 15 SNP haplotypes might differ between breeds because SNPmax was not always in exactly the same position. For each region, each sire was characterized by its 2 haplotypes.
- Determination of bull QTL status: A total of 187 bulls from the 3 breeds had daughters with phenotype and genotype data. Whole-genome sequences were available for 19 of them (Boussaha et al., 2015). The QTL status of the bulls was determined by testing the contrast between haplotype effects transmitted by a sire to his daughters. For the same bull, cows were distributed into 2 groups depending on the paternal haplotype they had received. The means of the 2 groups for the trait presenting the most significant effect in the QTL region were then compared using a t-test. The bull was considered to be heterozygous for the QTL if $P \leq 0.05$, homozygous if $P \geq 0.10$, and with an unknown genotype if $0.05 < P < 0.10$. This QTL status was determined for 13 bulls (7 MO, 2 NO, and 4 HO), based on 53 to 226 daughters per bull. The 6 other bulls had too few daughters (1–15) to properly estimate their status.
- Concordance analysis: Sequence-derived polymorphisms of these 13 bulls were extracted in candidate mutation regions from the corresponding VCF files (Boussaha et al., 2015). All 13 bulls had a high genotype quality score (30 or higher). Concordances were then determined between candidate mutations and bull QTL status in these 13 bulls. A polymorphism was considered concordant with a QTL if at least 90% of the bulls were either homozygous for both the polymorphism and the QTL or heterozygous for both the polymorphism and the QTL.

RESULTS

Accuracies, means, and standard deviations for the protein contents obtained from MIR predictions and expressed as a percentage in milk and a percentage in protein in the MO, NO, and HO breeds are reported in Table 2. The accuracy of the MIR predictions estimated by Ferrand et al. (2012) was greater for CN contents ($R^2 > 80\%$ and relative error $<8.4\%$) than for whey proteins ($59\% < R^2 < 74\%$; $11.7\% <$ relative error $< 14.4\%$). The means and standard deviations were very similar in the 3 breeds and quite similar to the values obtained using the reference analysis method.

Numerous significant genome-wide results were obtained. In total, 6,759, 3,981, and 3,942 tests were significant at the 7.6E-06 level, and among them 3,509, 1,910, and 1,629 exceeded the stringent level of 1.0E-11, in the MO, NO, and HO breeds, respectively.

Table 2. Milk protein composition: accuracy of MIR predictions and means ± standard deviations as a percentage of milk or as a percentage of proteins in the 3 breeds, Montbéliarde (MO), Normande (NO), and Holstein (HO)

	Accuracy[1]		As a percentage of milk			As a percentage of proteins		
Trait	R^2	RE	MO	NO	HO	MO	NO	HO
Protein content	1.00	0.73	3.4 ± 0.4	3.6 ± 0.4	3.3 ± 0.4	—	—	—
α-LA	0.59	14.4	0.14 ± 0.02	0.15 ± 0.02	0.14 ± 0.02	4.07 ± 0.28	4.16 ± 0.36	4.27 ± 0.42
β-LG	0.74	11.7	0.28 ± 0.05	0.28 ± 0.05	0.28 ± 0.05	8.25 ± 1.12	7.94 ± 1.03	8.46 ± 1.17
α_{S1}-CN	0.88	4.7	0.94 ± 0.10	0.99 ± 0.10	0.92 ± 0.11	27.8 ± 0.55	27.8 ± 0.68	27.9 ± 0.69
α_{S2}-CN	0.82	7.5	0.32 ± 0.04	0.35 ± 0.04	0.32 ± 0.04	9.53 ± 0.30	9.89 ± 0.33	9.69 ± 0.39
β-CN	0.92	3.7	1.24 ± 0.11	1.29 ± 0.11	1.20 ± 0.13	36.6 ± 0.88	36.2 ± 1.2	36.2 ± 1.2
κ-CN	0.80	8.4	0.33 ± 0.05	0.35 ± 0.05	0.31 ± 0.05	9.75 ± 0.60	9.87 ± 0.48	9.43 ± 0.58

[1]Accuracy of MIR predictions (R^2 and relative error) estimated by Ferrand et al. (2012).

Because of the high degree of linkage disequilibrium between neighboring haplotypes, significant results located within the same 4-Mbp interval were grouped into a single QTL region, regardless of the breeds or traits under study. Thirty-nine genomic regions were defined (Table 3; Figure 1), distributed over 20 of the 29 BTA. Several breeds or traits (potentially 3 × 13 = 39 breed × trait combinations) could therefore be affected by the same QTL region. By region, the number of significant breed by trait combinations ranged from 1 (for 17 QTL regions) to 29 (for the *6c* QTL region). These regions were generally tightly defined, including those comprising the largest number of significant results: the *20a*, *11c*, *14a*, and *6c* QTL regions presenting 10, 19, 21, and 29 significant results were located within intervals of about 500, 200, 400, and 1,300 kbp, respectively. Nine QTL regions mapped on BTA2, BTA6 (3 distinct regions), BTA11, BTA14, BTA20 (2 distinct regions), and BTA29 had very highly significant effects (LRT > 50) with respect to protein composition.

The effects of the QTL regions on protein composition (percentage of milk and percentage of proteins), expressed as a proportion of total genetic variance (**%GV**), are reported in Tables 4, 5, and 6 for the MO,

Table 3. List of the 39 genomic regions significantly [likelihood-ratio test statistic (LRT) >23.6] associated with milk protein composition (as a percentage of milk or as a percentage of proteins) in at least 1 of the 3 breeds (Montbéliarde, Normande, or Holstein)

Region[1]	Start (Mbp)	End (Mbp)	No. of significant tests[2]	LRT max	*P*-value
1	144.3	144.3	1	23.6	7.5E-06
2a	6.2	6.2	1	31.9	1.2E-07
2b	75.9	75.9	1	30.5	2.4E-07
2c	**131.8**	**134.1**	**5**	**54.7**	**1.3E-12**
3a	15	15.4	3	34.3	3.6E-08
3b	80.9	80.9	4	34.5	3.2E-08
4a	48.9	48.9	2	45.4	1.4E-10
4b	77.8	80.4	2	27.7	9.7E-07
5a	95	98.6	2	27.2	1.2E-06
5b	105.3	105.3	1	26.5	1.8E-06
5c	115.6	115.6	1	30.3	2.6E-07
6a	**37.7**	**37.7**	**1**	**102.5**	**5.5E-23**
6b	**46.6**	**46.7**	**2**	**86.2**	**1.9E-19**
6c	**86.9**	**88.2**	**29**	**309.8**	**5.3E-68**
7	46	46	1	29.8	3.4E-07
11a	12.7	12.7	2	26.5	1.8E-06
11b	63.1	63.1	1	24.1	5.8E-06
11c	**103.2**	**103.4**	**19**	**1,865.7**	**<1.0E-308**
14a	**1.7**	**2.1**	**21**	**184.7**	**7.8E-41**
15a	36.4	36.4	1	31.5	1.4E-07
15b	41.6	41.6	1	27	1.4E-06
15c	46.2	46.2	1	23.7	7.1E-06
15d	50.6	53.7	2	43.7	3.2E-10
17a	16.4	16.4	1	24.1	5.8E-06
17b	55.5	57	1	24.3	5.3E-06
19	61.1	61.1	1	40.1	2.0E-09
20a	**31.6**	**32.1**	**10**	**50.7**	**9.8E-12**
20b	36.2	36.3	1	38.7	3.9E-09
20c	**58**	**59.2**	**8**	**404.3**	**1.6E-88**
21	40.6	43.1	2	25.6	2.8E-06
22a	47.8	49.4	2	37.9	5.9E-09
22b	54.2	55.2	2	28.9	5.3E-07
24	58.3	58.3	2	30.7	2.2E-07
25a	27.1	30.7	1	24.3	5.3E-06
25b	58.3	58.3	1	25.6	2.8E-06
26	19.3	22.3	4	29.8	3.4E-07
27	36.1	36.2	2	30.8	2.1E-07
29a	**7.3**	**9.1**	**2**	**65.7**	**5.4E-15**
29b	43.8	44.3	4	27.1	1.3E-06

[1]Regions were defined by grouping SNP located within 4-Mb intervals.

[2]Number of breed × trait combinations with significant genome-wide results (LRT > 23.6; nominal $P < 7.6E-06$; Bonferroni corrected $P < 0.05$); in bold, regions with LRT > 50 (nominal $P < 1.4E-11$; Bonferroni corrected $P < 9.3E-08$).

NO, and HO breeds, respectively. In total, 26, 20, and 17 distinct QTL regions, corresponding to 54, 71, and 54 region × trait combinations, were significant at the genome-wide level in the MO, NO, and HO breeds, respectively. The number of QTL acting on PC differed as a function of the breed: 3 QTL in MO (*6c*, *20b*, and *25b*), 6 QTL in NO (*3b*, *6c*, *11a*, *14a*, *20a*, and *26*) and 4 QTL in HO (*6c*, *14a*, *20a*, and *29b*), the *6c* QTL being the only one common to all 3 breeds. All these QTL explained, respectively, 19.6, 26.5, and 21.5%GV of PC. For protein composition traits, 6 regions were common to the 3 breeds (*2c*, *6c*, *11c*, *14a*, *20c*, and *29a*) and generally affected the same traits, except for the *14a* QTL region that had effects on 10 different traits in the NO and HO breeds but only affected one CN content in MO. The 3 most significant regions, *6c*, *11c*, and *20c*, displayed high LRT values in all 3 breeds (Figure 2) and mainly affected κ-CN (11, 16, and 20%GV in MO, NO, and HO, respectively), β-LG (46, 41, and 51%GV), and α-LA (32, 25, and 26%GV) contents, expressed as a percentage of protein, respectively. The *11c* QTL had the strongest effects, with huge LRT values ranging from 1,563 to 1,866. Some QTL regions were found in only 1 breed. Breed-specific regions generally had weaker effects than shared regions, except for *6a* QTL, which explained about 31%GV of α_{S1}-CN (as a percentage of protein) in HO and had no significant effect in MO and NO cows.

Compared with proteins expressed as a percentage of milk, the number of significant results found for proteins expressed as a percentage of protein was higher in MO (33 vs. 18 region × trait combinations) and more similar in NO (30 vs. 35 region × trait combinations) and HO (27 vs. 23 region × trait combinations) cows. On average, the total genetic variance explained by all the QTL detected was higher for proteins expressed as a percentage of protein than for proteins expressed as a percentage of milk in MO (39 vs. 26%GV), NO (38 vs. 31%GV), and HO (43 vs. 26%GV) cows.

For each of the 6 milk protein fractions included in this study, up to 9 QTL regions were identified (2–7 as a percentage of milk and 2–8 as a percentage of protein). Depending on the protein and the unit of expression, the percentage of genetic variance explained by all the QTL detected ranged from 16 to 51% in MO, 26 to 43% in NO, and 19 to 55% in HO animals. As a percentage of protein, β-LG had the highest proportion of total genetic variance explained by the QTL (51, 43, and 55% in the MO, NO, and HO breeds, respectively), and it had the smallest number of QTL detected (2 in each breed). It is worth noting that for this trait, the 2 QTL detected co-localized (*6c* and *11c*) in the 3 breeds. The protein with the smallest proportion of genetic variance explained by the QTL was β-CN expressed as a percentage of milk (16, 26, and 19%GV in the MO, NO, and HO breeds).

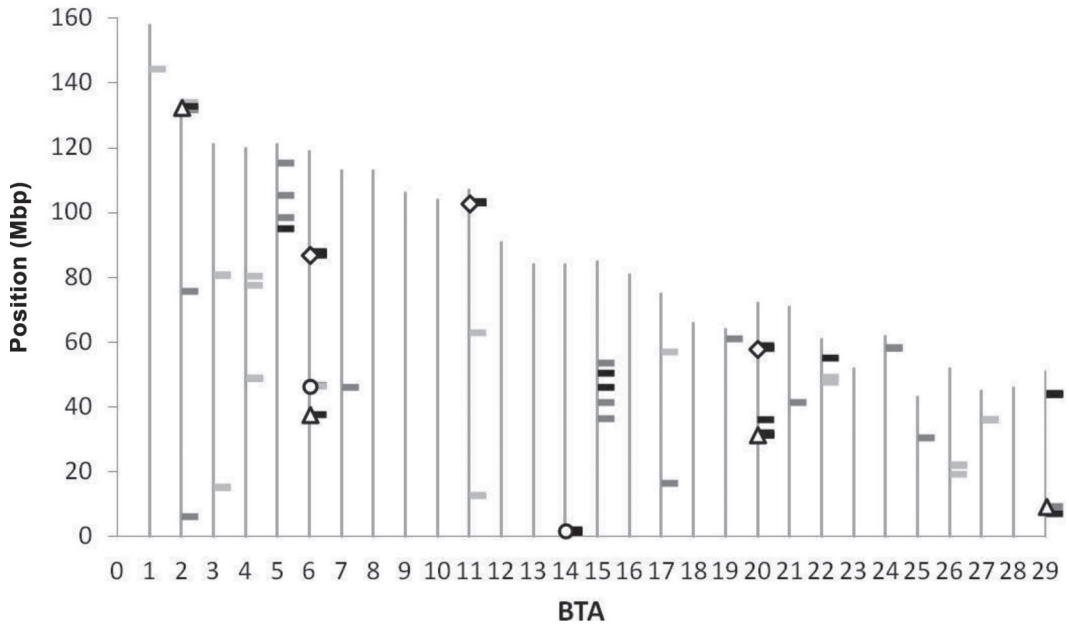


Figure 1. Genomic locations of QTL detected with likelihood-ratio test statistic (LRT) >23.6 in Holstein (black dash), Montbéliarde (dark gray dash), and Normande (light gray dash) breeds. QTL with LRT >50 found in 3 breeds (◊), 2 breeds (○), or 1 breed (Δ).

Table 4. Fraction (%) of genetic variance explained by each QTL region for protein content (PC) and protein composition as a percentage of milk and as a percentage of proteins (Montbéliarde breed)

	As a percentage of milk							As a percentage of protein						
Region	PC	α-LA	β-LG	α_{S1}-CN	α_{S2}-CN	β-CN	κ-CN	α-LA	β-LG	α_{S1}-CN	α_{S2}-CN	β-CN	κ-CN	No. of traits
2a										3.9				1
2b		4.5												1
2c											3.7	3.7		2
4a										4.4				1
4b												4.3		1
5a											3.5			1
5b												4.4		1
5c										3				1
6b								3		6.6				2
6c	13.4	10.2		12.7	14.6	12.6	16.9		4.8		8.1	7.6	11.4	10
7													5.5	1
11c			35.1					3.5	46.3	6.8		8.4	13.8	6
14a											3			1
15a													2.7	1
15b											4.2			1
15d										3.2		3.7		2
17a		3.1												1
19										3.6				1
20b	2.8			2.8	3.1									3
20c		13.2	4.5					31.8						3
21				3.6	3.7						3.9	4		4
24				3.8	3.6	2.9	3.1							4
25a								2.7						1
25b	3.4													1
27													4.4	1
29a										4.9			3.2	2
Total	19.6	31	39.6	22.9	25	15.5	20	41	51.1	36.4	26.4	36.1	41	

Table 5. Fraction (%) of genetic variance explained by each QTL region for protein content (PC) and protein composition as a percentage of milk and as a percentage of proteins (Normande breed)

	As a percentage of milk							As a percentage of protein						
Region	PC	α-LA	β-LG	α_{S1}-CN	α_{S2}-CN	β-CN	κ-CN	α-LA	β-LG	α_{S1}-CN	α_{S2}-CN	β-CN	κ-CN	No. of traits
1								2.6						1
2c											7.3		2.5	2
3a		4.7	1.7		4.6						5.3			4
3b	3.1			3.1		3.1	2.7							4
4a										5.7				1
4b				2.7	3.3		4				4.2	4		5
6b										10				1
6c	6.8	7.9		6.7	6.7	6.6	10.2		2.4		5.6	5.5	16	10
11a	3.5			3.4		3.7								3
11b											3.7			1
11c			26.2				2.9	3.6	41	6.9		4.4	11.5	7
14a	5.9	11		6.8	5.2	5.2	6.9	5.3		12.7		5.5	6	10
17b												6.9		1
20a	4.1		2.7	4.3	4.1	4	3.8							6
20c		9.3						25			5.5			3
22a		2.9					3					2.7	3.9	4
22b								3.7						1
26	3.1		2.1	3.2	2.8	3.5								5
27													3	1
29a										3.4				1
Total	26.5	35.8	32.7	30.2	26.7	26.1	33.5	40.2	43.4	38.7	31.6	29	42.9	

Table 6. Fraction (%) of genetic variance explained by each QTL region for protein content (PC) and protein composition as a percentage of milk and as a percentage of proteins (Holstein breed)

	As a percentage of milk							As a percentage of protein						
Region	PC	α-LA	β-LG	α_{S1}-CN	α_{S2}-CN	β-CN	κ-CN	α-LA	β-LG	α_{S1}-CN	α_{S2}-CN	β-CN	κ-CN	No. of traits
2c											7.6	5		2
5b											4.8			1
6a										30.7				1
6c	4.3	4.8		3.5	5.2	3.5	9.5		4		8	6.6	20.3	10
11c			32.9					3.7	51.3	3.3	4.1	8.4	11.8	7
14a	7.7	8.5	3	8.8	6.2	7.3	8.3				7.8	4.1	5.7	10
15c							3.2							1
15d													3.5	1
17b											5.8		3.6	2
19										4.5				1
20a	6.3			6.7	7.1		7.9					6.4	6.6	6
20b						4.8								1
20c		7.6	2.5					25.7						3
22b								5.8						1
25a										2.6				1
29a										6.4				1
29b	3.2	3.8	3	3.4		3								5
Total	21.5	24.7	41.4	22.4	18.5	18.6	28.9	35.2	55.3	47.5	38.1	30.5	51.5	

DISCUSSION

In this study, we report the results of a genome-wide scan of milk protein composition predicted from MIR spectra. It was conducted using 7,654 genotyped cows from the 3 main French dairy breeds. Numerous QTL regions (39 distinct regions in total: 17–26 per breed) were detected for protein composition, out of which 9 showed highly significant effects on protein composition, expressed as a proportion of total protein (percentage of protein) or milk (percentage of milk). In addition to genetic parameter results (Brochard et al., 2013), these QTL detection results confirmed that MIR predictions were accurate enough for genetic investigations. Indeed, repeated test-day records (6 on average per cow) compensated for the moderate MIR prediction accuracy of some proteins. The QTL detected explained a higher proportion of genetic variance for proteins expressed as a percentage of protein than as a percentage of milk. For instance, the *11c* QTL region had significant effects on only 1 trait as a percentage of milk (β-LG) but affected almost all proteins as a percentage of protein. In addition, some QTL regions presented significant results for traits expressed either as a percentage of protein or as a percentage of milk. Most previous studies in the literature on milk protein composition were performed on traits expressed as a percentage of protein, particularly because protein composition as a percentage of protein is more directly linked to traits of interest, such as the ratio of CN to total proteins affecting cheese yield (Wedholm et al., 2006). However, correlations among proteins are artificially increased and tend to be negative when they are expressed as a proportion of total proteins because their sum is fixed at 100% (Brochard et al., 2013). Analyses of proteins as a percentage of milk should therefore enable us to better understand the underlying biological mechanisms because no artificial correlations exist among these traits.

The number of QTL detected, and the magnitude of their effects, were in agreement with the moderate to high heritability estimates (h^2) obtained in the 3 breeds: $0.27 < h^2 < 0.86$ (Brochard et al., 2013). Moreover, these results were in agreement with the QTL results previously reported by Schopen et al. (2011) on protein contents estimated by capillary zone electrophoresis in 1,713 Dutch Holstein-Friesian cows. The authors highlighted several QTL throughout the genome, which included 3 with major effects located on BTA6, BTA11, and BTA14 explaining more than 10% of the genetic variance of the trait and 2 other QTL located on BTA5 and BTA29 with more moderate effects and explaining 5 to 10% of the genetic variance of the trait. All those QTL regions, except the one on BTA5, were also found in this study and corresponded

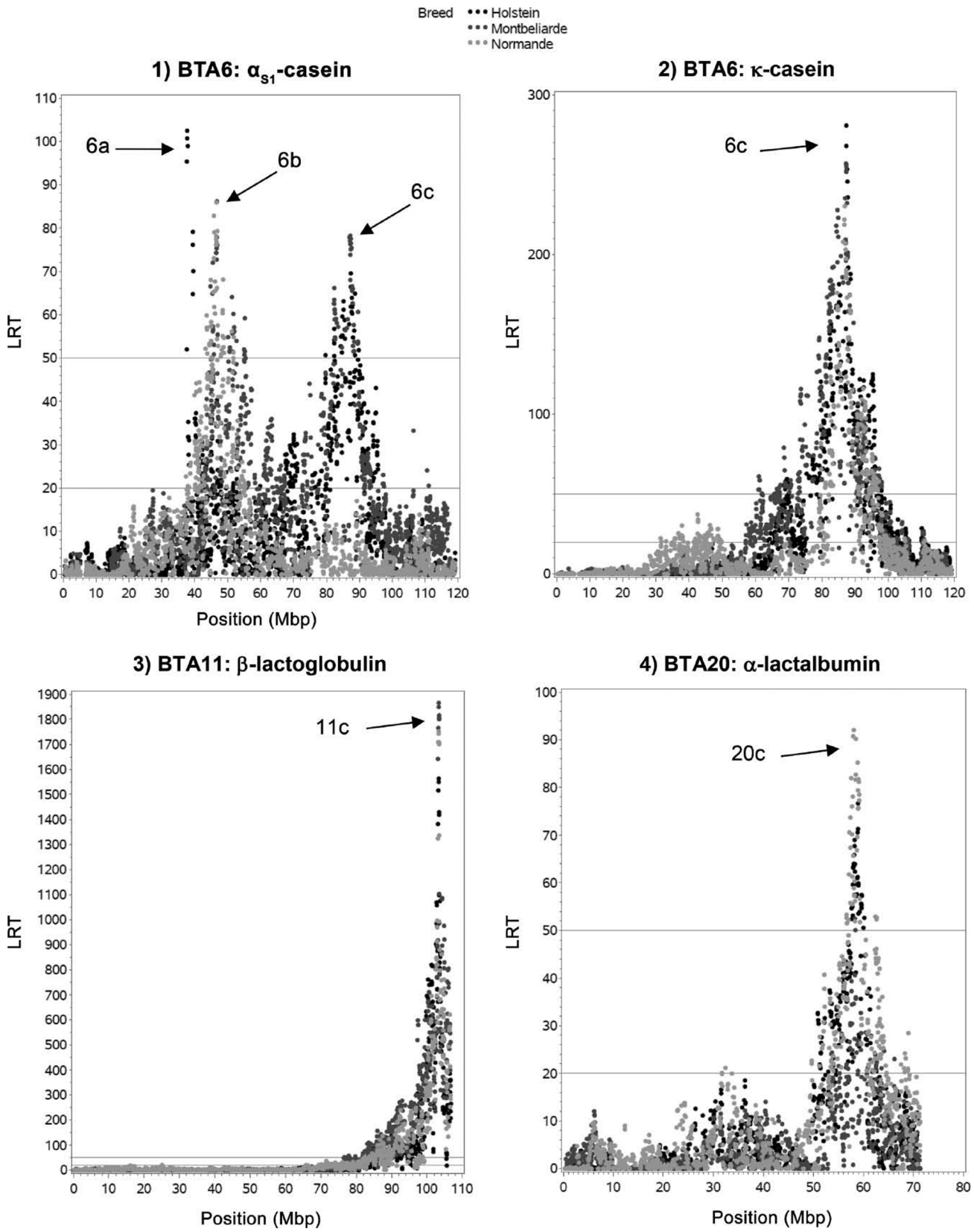


Figure 2. Likelihood ratio test (LRT) values plotted against SNP positions on BTA6 (1 and 2), BTA11 (3), and BTA20 (4) for protein contents expressed as a percentage of protein in Holstein, Montbéliarde, and Normande breeds.

to the *6c*, *11c*, *14a*, and *29b* QTL, respectively. Another study tested the association between 53 SNP located within or close to milk protein genes (BTA5, BTA6, and BTA11) and the milk protein profile measured by reverse-phase HPLC in purebred Holstein and (Holstein × Jersey) × Holstein backcross cows (Huang et al., 2012). Those authors found SNP that significantly explained a high percentage of the phenotypic variance of milk protein composition on BTA6 and BTA11 but not on BTA5, which is consistent with results found in the present study. In addition, 6 other QTL (*2c*, *6a*, *6b*, *20a*, *20c*, and *29a*) with effects ranging from 5 to 32% of the genetic variance of the trait were detected in our study. Schopen et al. (2011) also reported the *6b* and *20a* QTL as having weaker effects (about 3%), but they did not detect the *2c*, *6a*, *20c*, and *29a* QTL. This was not surprising with respect to *2c* and *29a* because these 2 QTL were only found in the NO population. However, it was more unexpected for *6a* and *20c*, which exhibited highly significant effects in HO cows and explained 31 and 26% of the genetic variance of α_{S1}-CN and α-LA, respectively. The *20c* QTL located at 58 to 59 Mbp was highly significant and explained 32, 25, and 26% of the genetic variance of the α-LA content in the MO, NO, and HO breeds, respectively. Microsatellite marker–based associations have been previously reported between PC and this region (Ashwell et al., 2004; Bagnato et al., 2008) and more recently with milk yield using SNP from the Illumina BovineHD BeadChip (Kemper et al., 2015), but the current report is the first time this region has been associated with milk protein composition.

The estimates of variance components were quite high. Because they were determined one by one in a one-QTL model, we cannot exclude the possibility that they were somewhat overestimated. Nevertheless, individual milk protein fractions are intermediate phenotypes compared, for example, with protein content, and QTL effects can be expected to be more marked.

Among the numerous QTL detected during our study, several were of particular interest because they had strong effects on milk protein composition, they were found in several breeds, or both factors were present. Identifying the causal mutations responsible for these effects could therefore be valuable for improving these traits through genomic selection. In each of the *6a*, *6c*, *11c*, *14a*, and *20a* QTL regions, candidate mutations had previously been associated with protein content in the literature. For these QTL regions, we tested whether the mutations previously described in the literature might be responsible for the effects observed in the PFL populations. To achieve this, we took advantage of the genome sequence of some PFL bulls from the 1,000 bull genomes project to perform concordance analyses between the QTL statuses and candidate mutation genotypes of the bulls.

6a QTL and the ABCG2 Gene

The *6a* QTL, located at around 37.7 Mbp, explained almost one-third of the genetic variance of α_{S1}-CN in HO cows. The Y581S mutation in the *ABCG2* gene, located about 300 kbp downstream of the *6a* QTL region, had previously been proposed as the causal mutation for milk yield and composition in Holstein cows. All the PFL bulls sequenced were homozygous A/A at this locus, whereas 6 were homozygous and 6 were heterozygous for the *6a* QTL. The Y581S mutation in the *ABCG2* gene was not responsible for the effects observed in our study. Nevertheless, the most significant SNP in this region was located at 37,653,391 bp; that is, between the *HERC3* (45 kbp downstream) and *PIGY* (24 kbp upstream) genes. A nearby SNP located at 37,631,640 bp on the BovineSNP50 BeadChip had previously been found to be associated with the milk protein percentage in Chinese Holstein cows (Jiang et al., 2010) and this was confirmed subsequently in the same population using a multiple SNP approach (Fang et al., 2014). This SNP was not tested during our study because it failed the quality control filters in the PFL HO population.

6b QTL and CN Genes

The locations of the most significant SNP in the *6c* QTL region ranged from 86.9 to 88.2 Mbp, depending on the trait and the breed. Major effects were found for all the analyzed traits expressed as a percentage of milk or as a percentage of protein in the 3 breeds and were particularly high for CN contents. The *6c* region includes genes encoding for caseins that are tightly linked in a 250-kb cluster (*CSN1S1*, *CSN2*, *CSN1S2*, and *CSN3*). The genotypes for 5 known polymorphisms in *CSN1S1* (87,157,262), *CSN2* (87,181,501 and 87,181,619), and *CSN3* (87,390,612 and 87,390,632) were tested in the PFL bulls (7 were heterozygous and 5 were homozygous for the QTL) and none of them was concordant with the QTL statuses. Nevertheless, the effects of these genes cannot be discarded because of the presence of different haplotypes formed by several polymorphisms in strong linkage disequilibrium in this region (Martin et al., 2002). Concordance analysis is known to be sensitive to multiple causal variants, and this hypothesis cannot be excluded in this case.

11c QTL and the LGB Gene

In the *11c* QTL region, a single SNP located at position 103,289,035 displayed the most significant effects in the MO, NO, and HO breeds during QTL detection analyses (41 to 51% of the genetic variance of β-LG content). It was in full linkage disequilibrium with the 2 mutations described by Ganai et al. (2009) as the causal genetic polymorphisms of β-LG protein variants A and B in the *LGB* gene (103,303,475 and 103,304,757 bp) in all the MO, NO, and HO bulls tested. Genotypes for the QTL and causal genetic polymorphisms were therefore perfectly concordant for all the bulls tested (7 were heterozygous and 3 were homozygous for the QTL).

14a QTL and the DGAT1 Gene

The K232A *DGAT1* polymorphism on BTA14 (Grisart et al., 2002) was investigated to explain the effects of the *14a* QTL region. Only 1 (HO) of the 13 tested bulls was found to be heterozygous for the QTL in the centromeric region of BTA14. However, reaching any conclusion was not possible because of the very poor sequence quality in this region for several of the bulls tested and particularly for the HO bull heterozygous for the QTL. Nevertheless, the very low frequency of K232A in MO and NO animals allowed us to hypothesize that it could not explain the QTL in these 2 breeds.

20b QTL and the GHR Gene

The *20b* QTL was located between 31.6 and 32.1 Mbp and had effects in both the NO and HO breeds. In NO cows, these effects were moderate (3–4% of genetic variance) and only found with respect to the protein composition expressed as a percentage of milk, whereas in HO cows the effects concerned both protein composition as a percentage of milk (6–8% of genetic variance) and some CN contents as a percentage of protein (β- and κ-CN). In both breeds, the effects on PC were significant (4 and 6% of genetic variance in the NO and HO breeds, respectively). In this QTL region, the F279Y mutation in the *GHR* gene at position 31,909,478 had previously been found to affect PC (Blott et al., 2003). Two PFL bulls were heterozygous and 9 were homozygous for the QTL, and genotypes at this site were concordant with their QTL status, suggesting that the mutation in the *GHR* gene might explain the effects detected in our study.

CONCLUSIONS

The identification of numerous QTL with marked effects on protein composition predicted from MIR spectra could help selection of these traits from routinely collected MIR spectra. The accuracy of MIR predictions remained moderate, but it was compensated for by the multiplicity of test-day measures and the strong effect of the QTL. The PFL population, consisting of 2,208 to 2,773 phenotyped and genotyped cows depending on the breed, represents a first reference population for genomic selection. In the near future, MIR spectra combined with large-scale cow genotyping will enable the rapid development of such reference populations. Therefore, if a clear breeding objective were defined, genomic selection could easily be implemented to target these traits. Moreover, through the use of whole-genome sequences during our study, 2 known mutations (*LGB* and *GHR*) were confirmed to be strong candidates responsible for the QTL effects, and another one (*ABCG2*) was discarded. For all the other regions, association studies applied to cow whole-genome sequences imputed from BovineSNP50 BeadChip genotypes (via BovineHD BeadChip genotypes) are expected to identify new candidate mutations for milk protein composition.

ACKNOWLEDGMENTS

The PhénoFinLait project has received financial support from Agence Nationale de le Recherche (ANR-08-GANI-034 Lactoscan), APIS-GENE, Compte d′Affectation Spéciale Développement Agricole et Rural, Centre National Interprofessionnel de l′Economie Laitière, FranceAgriMer, France Génétique Elevage, and the French Ministry of Agriculture. The authors thank the breeders who participated in PhénoFinLait; colleagues from Institut de l'Elevage and INRA who designed and coordinated farm samples and data collection; the partners of the project, laboratories, manufacturers, and DHI organizations that provided data; and the members of the PhénoFinLait scientific committee who advised and managed this work. The whole genome sequences of bulls were funded by the ANR CartoSeq project (ANR-10-GENM-0018) and APIS-GENE.

REFERENCES

Ashwell, M. S., D. W. Heyen, T. S. Sonstegard, C. P. Van Tassell, Y. Da, P. M. VanRaden, M. Ron, J. I. Weller, and H. A. Lewin. 2004. Detection of quantitative trait loci affecting milk production, health, and reproductive traits in Holstein cattle. J. Dairy Sci. 87:468–475.

Bagnato, A., F. Schiavini, A. Rossoni, C. Maltecca, M. Dolezal, I. Medugorac, J. Sölkner, V. Russo, L. Fontanesi, A. Friedmann, M. Soller, and E. Lipkin. 2008. Quantitative trait loci affecting milk yield and protein percentage in a three-country Brown Swiss population. J. Dairy Sci. 91:767–783.

Blott, S., J. J. Kim, S. Moisio, A. Schmidt-Küntzel, A. Cornet, P. Berzi, N. Cambisano, C. Ford, B. Grisart, D. Johnson, L. Karim, P. Simon, R. Snell, R. Spelman, J. Wong, J. Vilkki, M. Georges, F. Farnir, and W. Coppieters. 2003. Molecular dissection of a quantitative trait locus: A phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. Genetics 163:253–266.

Boichard, D., F. Guillaume, A. Baur, P. Croiseau, M. Rossignol, M. Boscher, T. Druet, L. Genestout, J. Colleau, L. Journaux, V. Ducrocq, and S. Fritz. 2012. Genomic selection in French dairy cattle. Anim. Prod. Sci. 52:115–120.

Bonfatti, V., A. Cecchinato, L. Gallo, A. Blasco, and P. Carnier. 2011a. Genetic analysis of detailed milk protein composition and coagulation properties in Simmental cattle. J. Dairy Sci. 94:5183–5193.

Bonfatti, V., G. Di Martino, and P. Carnier. 2011b. Effectiveness of mid-infrared spectroscopy for the prediction of detailed protein composition and contents of protein genetic variants of individual milk of Simmental cows. J. Dairy Sci. 94:5776–5785.

Boussaha, M., D. Esquerre, J. Barbieri, A. Djari, A. Pinton, R. Letaief, G. Salin, F. Escudie, A. Roulet, S. Fritz, F. Samson, C. Grohs, M. Bernard, C. Klopp, D. Boichard, and D. Rocha. 2015. Genome-wide study of structural variants in bovine Holstein, Montbeliarde and Normande dairy breeds. PLoS ONE 10:e0135931.

Brochard, M., M. P. Sanchez, A. Govignon-Gion, M. Ferrand, M. Gelé, D. Pourchet, G. Miranda, P. Martin, and D. Boichard. 2013. Paramètres génétiques pour la composition protéique du lait dans 3 races bovines. Page 158 in Rencontres autour des Recherches sur les Ruminants Vol. 20, Paris, France. INRA, Paris, France.

Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. Van Binsbergen, R. F. Brøndum, X. Liao, A. Djari, S. Rodriguez, C. Grohs, D. Esquerré, O. Bouchez, M. N. Rossignol, C. Klopp, D. Rocha, S. Fritz, A. Eggen, P. J. Bowman, D. Coote, A. J. Chamberlain, C. P. VanTassell, I. Hulsegge, M. E. Goddard, B. Guldbrandtsen, M. S. Lund, R. F. Veerkamp, D. Boichard, R. Fries, and B. J. Hayes. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat. Genet. 46:858–865. http://dx.doi.org/10.1038/ng.3034.

Druet, T., S. Fritz, M. Boussaha, S. Ben-Jemaa, F. Guillaume, D. Derbala, D. Zelenika, D. Lechner, C. Charon, D. Boichard, I. G. Gut, A. Eggen, and M. Gautier. 2008. Fine mapping of quantitative trait loci affecting female fertility in dairy cattle on BTA03 using a dense single-nucleotide polymorphism map. Genetics 178:2227–2235.

Druet, T., and M. Georges. 2010. A hidden Markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. Genetics 184:789–798.

Ducrocq, V. 1998. Genekit, BLUP software. June 2011 version ed. INRA GABI, Jouy-en-Josas, France.

Fang, M., W. Fu, D. Jiang, Q. Zhang, D. Sun, X. Ding, and J. Liu. 2014. A multiple-SNP approach for genome-wide association study of milk production traits in Chinese Holstein cattle. PLoS ONE 9:e99544.

Ferrand, M., G. Miranda, S. Guisnel, H. Larroque, O. Leray, F. Lahalle, M. Brochard, and P. Martin. 2012. Determination of protein composition in milk by mid-infrared spectrometry. Pages 41–45 in Proc. International Strategies and New Developments in Milk Analysis. VI ICAR Reference Laboratory Network Meeting, Cork, Ireland. ICAR, Rome, Italy.

Ganai, N. A., H. Bovenhuis, J. A. van Arendonk, and M. H. Visker. 2009. Novel polymorphisms in the bovine beta-lactoglobulin gene and their effects on beta-lactoglobulin protein concentration in milk. Anim. Genet. 40:127–133.

Gebreyesus, G., M. S. Lund, L. Janss, N. A. Poulsen, L. B. Larsen, H. Bovenhuis, and A. J. Buitenhuis. 2016. Short communication: Multi-trait estimation of genetic parameters for milk protein composition in the Danish Holstein. J. Dairy Sci. 99:2863–2866.

Gelé, M., S. Minery, J. M. Astruc, P. Brunschwig, M. Ferrand, G. Lagriffoul, H. Larroque, J. Legarto, P. Martin, G. Miranda, I. Palhière, P. Trossat, and M. Brochard. 2014. Phénotypage et génotypage à grande échelle de la composition fine des laits dans les filières bovine, ovine et caprine. Prod. Anim. 27:255–268.

Grignola, F. E., I. Hoeschele, Q. Zhang, and G. Thaller. 1996. Mapping quantitative trait loci in outcross populations via residual maximum likelihood. II. A simulation study. Genet. Sel. Evol. 28:491–504.

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and R. Snell. 2002. Positional candidate cloning of a QTL in dairy cattle: Identification of a missense mutation in the bovine *DGAT1* gene with major effect on milk yield and composition. Genome Res. 12:222–231.

Huang, W., F. Peñagaricano, K. R. Ahmad, J. A. Lucey, K. A. Weigel, and H. Khatib. 2012. Association between milk protein gene variants and protein composition traits in dairy cattle. J. Dairy Sci. 95:440–449.

Jiang, L., J. Liu, D. Sun, P. Ma, X. Ding, Y. Yu, and Q. Zhang. 2010. Genome wide association studies for milk production traits in Chinese Holstein population. PLoS ONE 5:e13661.

Kemper, K. E., C. M. Reich, P. J. Bowman, C. J. Vander Jagt, A. J. Chamberlain, B. A. Mason, B. J. Hayes, and M. E. Goddard. 2015. Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy of across-breed genomic predictions. Genet. Sel. Evol. 47:29.

Martin, P., M. Szymanowska, L. Zwierzchowski, and C. Leroux. 2002. The impact of genetic polymorphisms on the protein composition of ruminant milks. Reprod. Nutr. Dev. 42:433–459.

Meuwissen, T. H., A. Karlsen, S. Lien, I. Olsaker, and M. E. Goddard. 2002. Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. Genetics 161:373–379.

Meuwissen, T. H., and M. E. Goddard. 2000. Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. Genetics 155:421–430.

Misztal, I., T. Tsuruta, B. Strabel, B. Auvray, and T. Druet. 2002. BLUPF90 and related programs (BGF90). Pages 21–22 in Proc. 7th World Congress on Genetics Applied to Livestock Production. Montpellier, France. Editions Quae, Versailles, France.

Schopen, G. C., J. M. Heck, H. Bovenhuis, M. H. Visker, H. J. van Valenberg, and J. A. van Arendonk. 2009. Genetic parameters for major milk proteins in Dutch Holstein-Friesians. J. Dairy Sci. 92:1182–1191.

Schopen, G. C., M. H. Visker, P. D. Koks, E. Mullaart, J. A. van Arendonk, and H. Bovenhuis. 2011. Whole-genome association study for milk protein composition in dairy cattle. J. Dairy Sci. 94:3148–3158.

van den Berg, I., S. Fritz, S. Rodriguez, D. Rocha, M. Boussaha, M. S. Lund, and D. Boichard. 2014. Concordance analysis for QTL detection in dairy cattle: A case study of leg morphology. Genet. Sel. Evol. 46:31.

Wedholm, A., L. B. Larsen, H. Lindmark-Månsson, A. H. Karlsson, and A. Andrén. 2006. Effect of protein composition on the cheese-making properties of milk from individual dairy cows. J. Dairy Sci. 89:3296–3305.

Zimin, A., A. Delcher, L. Florea, D. Kelley, M. Schatz, D. Puiu, F. Hanrahan, G. Pertea, C. Van Tassell, T. Sonstegard, G. Marcais, M. Roberts, P. Subramanian, J. Yorke, and S. Salzberg. 2009. A whole-genome assembly of the domestic cow, *Bos taurus*. Genome Biol. 10:R42.

4.3. GWAS pour la composition protéique à l'échelle de la séquence

Within-breed and multi-breed GWAS on imputed whole genome sequence variants reveal candidate mutations affecting milk protein composition in dairy cattle.

Marie-Pierre Sanchez[1*], Armelle Govignon-Gion[2,1], Pascal Croiseau[1], Sébastien Fritz[1,3], Chris Hozé[1,3], Guy Miranda[1], Patrice Martin[1], Anne Barbat-Leterrier[1], Rabia Letaïef[1], Dominique Rocha[1], Mickaël Brochard[2], Mekki Boussaha[1], Didier Boichard[1]

[1]GABI, INRA, AgroParisTech, Université Paris Saclay, F-78350 Jouy-en-Josas, France

[2]Institut de l'Elevage, F-75012 Paris, France

[3]Allice, F-75012 Paris, France

Genetics Selection Evolution 2017. 49:68.

http://dx.doi.org/10.1186/s12711-017-0344-z

Sanchez *et al. Genet Sel Evol (2017) 49:68*
DOI 10.1186/s12711-017-0344-z

Genetics Selection Evolution

RESEARCH ARTICLE Open Access

Within-breed and multi-breed GWAS on imputed whole-genome sequence variants reveal candidate mutations affecting milk protein composition in dairy cattle

Marie-Pierre Sanchez[1*], Armelle Govignon-Gion[1,2], Pascal Croiseau[1], Sébastien Fritz[1,3], Chris Hozé[1,3], Guy Miranda[1], Patrice Martin[1], Anne Barbat-Leterrier[1], Rabia Letaïef[1], Dominique Rocha[1], Mickaël Brochard[2], Mekki Boussaha[1] and Didier Boichard[1]

Abstract

Background: Genome-wide association studies (GWAS) were performed at the sequence level to identify candidate mutations that affect the expression of six major milk proteins in Montbéliarde (MON), Normande (NOR), and Holstein (HOL) dairy cattle. Whey protein (α-lactalbumin and β-lactoglobulin) and casein (αs1, αs2, β, and κ) contents were estimated by mid-infrared (MIR) spectrometry, with medium to high accuracy ($0.59 \leq R^2 \leq 0.92$), for 848,068 test-day milk samples from 156,660 cows in the first three lactations. Milk composition was evaluated as average test-day measurements adjusted for environmental effects. Next, we genotyped a subset of 8080 cows (2967 MON, 2737 NOR, and 2306 HOL) with the BovineSNP50 Beadchip. For each breed, genotypes were first imputed to high-density (HD) using HD single nucleotide polymorphisms (SNPs) genotypes of 522 MON, 546 NOR, and 776 HOL bulls. The resulting HD SNP genotypes were subsequently imputed to the sequence level using 27 million high-quality sequence variants selected from Run4 of the 1000 Bull Genomes consortium (1147 bulls). Within-breed, multi-breed, and conditional GWAS were performed.

Results: Thirty-four distinct genomic regions were identified. Three regions on chromosomes 6, 11, and 20 had very significant effects on milk composition and were shared across the three breeds. Other significant effects, which partially overlapped across breeds, were found on almost all the autosomes. Multi-breed analyses provided a larger number of significant genomic regions with smaller confidence intervals than within-breed analyses. Combinations of within-breed, multi-breed, and conditional analyses led to the identification of putative causative variants in several candidate genes that presented significant protein–protein interactions enrichment, including those with previously described effects on milk composition (*SLC37A1*, *MGST1*, *ABCG2*, *CSN1S1*, *CSN2*, *CSN1S2*, *CSN3*, *PAEP*, *DGAT1*, *AGPAT6*) and those with effects reported for the first time here (*ALPL*, *ANKH*, *PICALM*).

Conclusions: GWAS applied to fine-scale phenotypes, multiple breeds, and whole-genome sequences seems to be effective to identify candidate gene variants. However, although we identified functional links between some candidate genes and milk phenotypes, the causality between candidate variants and milk protein composition remains to be demonstrated. Nevertheless, the identification of potential causative mutations that underlie milk protein composition may have immediate applications for improvements in cheese-making.

*Correspondence: marie-pierre.sanchez@inra.fr
[1] GABI, INRA, AgroParisTech, Université Paris Saclay, 78350 Jouy-en-Josas, France
Full list of author information is available at the end of the article

BioMed Central

© The Author(s) 2017. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Background

In cattle, milk protein composition is mostly influenced by genetic factors [1–4] and is of interest because it determines cheese-making properties [5]. Bovine milk protein composition can be predicted at a large scale by analyzing mid-infrared (MIR) spectra, which is routinely performed [6, 7]. Combined with cow genotyping, this technique may open avenues to investigate the genomic regions that influence milk protein composition. In a previous genome-wide association study (GWAS) based on the bovine 50 K single nucleotide polymorphism (SNP) array, we highlighted numerous genomic regions with very significant effects on milk protein composition in the three main breeds of French dairy cattle: Holstein (HOL), Montbéliarde (MON), and Normande (NOR) [8]. However, because the 50 K SNP array contains only a small fraction of the total number of genomic variants, we were not able to directly pinpoint candidate mutations.

In Run4 of the 1000 bull genome reference population, a database containing more than 56 million SNPs and small insertions/deletions (InDel) was constructed by analyzing whole-genome sequences (WGS) from 1147 bulls representing 27 different breeds, including 288 HOL, 28 MON and 24 NOR bulls. These data can then be used to impute WGS from experimentally or routinely obtained 50 K SNP genotypes [9]. In this way, imputed WGS can be obtained for a large number of animals and in particular, those with phenotypes.

Since WGS contain almost all the genomic variants, they should contain the causal mutations for a given trait and, thus they provide a much higher GWAS resolution. However, due to the long-range linkage disequilibrium that exists within dairy cattle breeds, the resolution of within-breed GWAS is often limited. For causal mutations that are shared among breeds, a multi-breed model can be used to refine regions that harbour quantitative trait loci (QTL). This approach takes advantage of the historical recombination events that have occurred in each breed, resulting in linkage disequilibrium over shorter distances and better resolution [10].

Here, we report the results of a GWAS at the sequence level for six major milk protein contents, namely α-lactalbumin and β-lactoglobulin and αs1, αs2, β, and κ caseins from HOL, MON, and NOR cows. The results of within-breed, multi-breed, and conditional analyses, that fit the most significant variant in addition to other tested variants, are examined together in order to pinpoint potential candidate variants in each genomic region.

Methods

Animals, phenotypes, and genotypes

For this study, we did not perform any animal experiment, thus no ethical approval was required. Details on the animals and milk analyses are in Sanchez et al. [8]. Briefly, MIR spectra were obtained for 848,068 milk samples from 156,660 cows of the three main French dairy breeds: Montbéliarde (MON), Normande (NOR), and Holstein (HOL). These spectra were used to predict milk protein content (PC) and milk protein composition with the equations derived as described by Ferrand et al. [7]. More details about the method and the calibration population used are in Sanchez et al. [4]. The contents of the six main milk proteins (α_{s1}-CN, α_{s2}-CN, β-CN, κ-CN, α-LA, and β-LG) were predicted in g/100 g protein. Total casein content and total whey protein content were also analyzed (Σ-CN and Σ-WP, respectively). In order to adjust phenotypes for non-genetic effects, a within-breed mixed model was applied to test-day data using the GENEKIT software [11]. This single-trait repeatability model included genetic, permanent environmental, and residual random effects, as well as herd × test-day, parity × stage of lactation, year × month of calving, and spectrometer × test month fixed effects. We applied this model to data from the first two lactations that included at least three test-day records across lactations during the study period. Then, test-day data were corrected for all non-genetic effects included in the model and averaged per cow. Thus, for each trait and each cow, a single phenotype was defined and subsequently used in GWAS analyses. In total, 293,780, 58,594, and 72,973 test-day records were analyzed, which corresponded to 44,959 MON, 12,428 NOR, and 14,530 HOL cows, i.e. an average of 6.5, 4.7, and 5.0 test-day records per cow, respectively.

Among these cows, 8010 were genotyped with the Illumina BovineSNP50 BeadChip (Illumina Inc., San Diego). We applied the following quality control filters: the individual call rate had to be higher than 95%, the SNP call rate higher than 90%, the minor allele frequency (MAF) higher than 5%, and genotype frequencies had to be in Hardy–Weinberg equilibrium with $P > 10^{-4}$. The final dataset included between 37,332 and 41,028 SNPs (Table 1), depending on the within-breed or multi-breed population considered, for 7907 cows (3032 MON, 2659 NOR, and 2216 HOL) with phenotypes.

Imputation to whole-genome sequences

The 50 K SNP genotypes of the 7907 cows were imputed to whole-genome sequence (WGS) using FImpute software, which accurately and quickly analyzes large datasets [12]. A two-step approach was applied in order to improve the accuracy of results: from 50 to 777 K high-density (HD) SNPs, and then, from imputed HD SNPs to WGS [13]. All imputations were performed separately for each breed using either a breed-specific (from 50 K to HD SNPs) or a multi-breed (from HD SNPs to WGS)

Table 1 Features of the Montbéliarde (MON), Normande (NOR), Holstein (HOL), and multi-breed populations

Number of	MON	NOR	HOL	Multi-breed
Phenotyped cows	44,959	12,428	14,530	71,917
Total test-day records	293,780	58,594	72,973	425,347
Test-day records per cow	6.5	4.7	5	5.9
Genotyped cows	3032	2659	2216	7907
Polymorphic 50 K SNPs	37,332	37,690	39,158	41,028
Polymorphic HD SNPs	548,185	549,359	553,712	586,749
Polymorphic sequence variants	15,957,336	14,809,860	15,116,501	18,366,748
Sequence variants (MAF $\geq$ 2%)	11,755,172	11,445,432	11,592,432	13,534,013

reference panel depending on the targeted density [14]. In each MON, NOR and HOL breed, imputations to the HD SNP level were performed using a within-breed reference population that included respectively 522 MON, 546 NOR, and 776 HOL bulls that had been genotyped with the Illumina BovineHD BeadChip (Illumina Inc., San Diego, CA). Around 550,000 SNPs were retained in each breed after removing SNPs that failed in the quality control filters, as described above for the 50 K (Table 1). WGS variants were imputed from HD SNP genotypes using WGS variants of the 1147 bulls from Run4 of the 1000 Bull Genomes consortium; these bulls represent 27 cattle breeds (see Additional file 1: Table S1), with 288 HOL, 28 MON, and 24 NOR individuals [9]. The protocol used was defined in the "1000 bull genomes" consortium [9]. Whole-genomes of all individuals were used for 2 × 100 bp paired-end sequencing using Illumina sequencing-by-synthesis technology and sequence reads were further filtered for quality and subsequently aligned to the UMD3.1 reference sequence, as previously described [9, 15]. Small genomic variations (SNPs and InDel) were detected using SAMtools 0.0.18 [16]. Raw variants were further filtered to produce 27,754,235 autosomal variants [15]. Filtered variants were subsequently annotated with the Ensembl variant effect predictor (VEP) pipeline v81 [17] and effect of the amino acid changes was predicted using the SIFT tool [18].

Precision of imputation from HD to sequence was assessed by comparing imputed genotypes with those obtained by re-genotyping a subset of the same cows with a custom chip. This additional information was not used in the imputation process. Two datasets were available: (1) a group of 168 Holstein cows that were genotyped with the first version (V1) of the EuroG10k Illumina chip, with 721 additional markers; and (2) a group of 2142 Montbéliarde cows that were genotyped with the fourth version (V4) of the same EuroG10k chip containing 3082 additional SNPs. Only SNPs with good technical quality (call rate > 95%, validation of the clusters by visual inspection, within-breed allelic frequency not significantly different across chip versions) were used. Imputation accuracy was measured by the squared correlation between true and imputed genotypes and by the genotypic and allelic concordance rate.

In order to remove SNPs with the lowest accuracies of imputation, only variants with a MAF higher than 0.02 were retained for further association analyses. Thus, about 11 million variants were included in each within-breed analysis and around 13 million were included in multi-breed analyses (Table 1).

Whole-genome sequence association analyses

We performed single-trait association analyses between all the polymorphic variants and the nine measured milk protein composition traits: PC, α-LA, β-LG, α_{s1}-CN, α_{s2}-CN, β-CN, κ-CN, Σ-CN, and Σ-WP (Table 2).

All association analyses were performed using the *mlma* option of the GCTA software, which applies a mixed linear model that includes the candidate variant [19]:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{x}b + \mathbf{u} + \mathbf{e}, \quad (1)$$

where $\mathbf{y}$ is the vector of pre-adjusted phenotypes, averaged per cow; μ is the overall mean; b is the additive fixed effect of the candidate variant to be tested for association; $\mathbf{x}$ is the vector of imputed genotypes coded as 0, 1, or 2 (number of copies of the second allele); $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G}\sigma_u^2)$ is the vector of random polygenic effects, with $\mathbf{G}$ the genomic relationship matrix (GRM), calculated by using the HD SNP genotypes [20], and σ_u^2 the polygenic variance, estimated based on the null model ($\mathbf{y} = \mathbf{1}\mu + \mathbf{u} + \mathbf{e}$) and then fixed while testing for the association between each variant and the trait; and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$ is the vector of random residual effects, with $\mathbf{I}$ the identity matrix and σ_e^2 the residual variance. Within-breed, the number of test-day records did not differ very much across cows, thus, the residual variance was assumed to be constant across cows.

For multi-breed association analyses, Model (2) was applied by adding a fixed breed effect $\mathbf{v}$ to Model (1), with $\mathbf{W}$ as the incidence matrix relating phenotypes to breed effect (three levels), and $\mathbf{x}$, $\mathbf{b}$, $\mathbf{u}$, and $\mathbf{e}$ as defined previously:

$$\mathbf{y} = \mathbf{W}\mathbf{v} + \mathbf{x}b + \mathbf{u} + \mathbf{e}. \quad (2)$$

The Bonferroni correction was applied to the thresholds in order to account for multiple testing. A very stringent

Table 2 MIR predictions for milk protein composition in Montbéliarde (MON), Normande (NOR), and Holstein (HOL) cows

Trait		Accuracy[a]		Means ± standard deviations[b]		
		R^2	RE	MON	NOR	HOL
PC	Protein content	1.00	0.73	3.4 ± 0.4	3.6 ± 0.4	3.3 ± 0.4
α-LA	α-lactalbumin	0.59	14.4	4.07 ± 0.28	4.16 ± 0.36	4.27 ± 0.42
β-LG	β-lactoglobulin	0.74	11.7	8.25 ± 1.12	7.94 ± 1.03	8.46 ± 1.17
α_{s1}-CN	α_{s1}-casein	0.88	4.7	27.8 ± 0.55	27.8 ± 0.68	27.9 ± 0.69
α_{s2}-CN	α_{s2}-casein	0.82	7.5	9.53 ± 0.30	9.89 ± 0.33	9.69 ± 0.39
β-CN	β-casein	0.92	3.7	36.6 ± 0.88	36.2 ± 1.2	36.2 ± 1.2
κ-CN	κ-casein	0.80	8.4	9.75 ± 0.60	9.87 ± 0.48	9.43 ± 0.58
Σ-CN	Sum of caseins	0.97	2.7	83.7 ± 0.94	83.7 ± 1.5	83.1 ± 1.4
Σ-WP	Sum of whey proteins	0.73	8.9	12.6 ± 1.1	11.9 ± 1.2	12.6 ± 1.3

[a] Accuracy of MIR predictions (R^2 = coefficient of determination and RE = relative error) estimated by Ferrand et al. [7] for protein composition expressed as g/100 g milk

[b] g/100 g milk for protein content (PC) and g/100 g protein for other traits

correction was used, which considered all 13 million tests as independent. Therefore, the 5% genome-wide threshold of significance corresponded to a nominal *P* value of 3.7×10^{-9} ($-\log_{10}(P) = 8.4$). QTL regions were identified by grouping significant results that were located within the same 2 million base-pair (Mbp) interval in a single genomic region, regardless of the breeds or traits under study. QTL regions were determined by considering positions of variants included in the upper third of the peak. For a given trait in a given breed, when two consecutive QTL regions had overlapping confidence intervals, or when the distance between the limits of the confidence intervals was less than 1 Mbp, only the confidence interval that presented the most significant results was retained.

Conditional association analyses

In the most significant QTL regions, conditional analyses were carried out using the *cojo* option of GCTA [21] in order to conclude if multiple significant variants in a genomic region were due to LD with the same causal mutation or to the presence of multiple causal mutations. Association analyses were performed by including in the model the most significant variant or the putative causal mutation as a fixed effect and by testing all variants that were not in strong LD with the conditional variant ($r^2 < 0.9$).

Annotation and protein interactions

Sequence-derived polymorphisms were extracted for candidate mutation regions from the corresponding VCF files [22]. All variants with a $-\log_{10}(P)$ higher than 8.4 and located within confidence intervals were annotated. To avoid missing important genes, confidence intervals were extended by 100 kb on each side.

In addition, functional protein–protein interactions (PPI) encoded by candidate genes were investigated, as well as gene ontology (GO) enrichment, using the STRING Genomics 10.0 database of protein–protein interaction (PPI) networks [23]. This database provides (1) known PPI from curated databases or experiments and (2) PPI predicted on the basis of gene neighborhood, gene fusions, gene co-occurrence, text mining in literature, co-expression, or protein homology. A global PPI network was constructed which retained only interactions with a high level of confidence (score > 0.4).

Results

The results of imputation accuracy at the sequence level for SNPs used in the GWAS analyses (MAF ≥ 2%) are in Table 3. Squared correlations between imputed and true genotypes in the validation set reached 76 and 84%, in Montbeliarde and Holstein breeds, respectively. This table also presents the overall results of concordance rate. Figure 1 shows the imputation precision according to MAF in the two breeds.

Among the 13 million tested variants, 71,755 had genome-wide significant effects ($-\log_{10}(P) \geq 8.4$) in at least one within-or multi-breed analysis and for at least one milk protein composition trait.

Among these, 29,722, 27,787, and 30,988 were found in within-breed MON, NOR, and HOL analyses, respectively. Some of these variants had significant effects in multiple breeds: 7343 in both MON and NOR, 8055 in NOR and HOL, 8068 in HOL and MON, and 3080 in all three breeds (Table 4; Fig. 2a).

For each trait, the number of significantly associated variants was relatively consistent between breeds. It was lower (from 193 to 2394) for α_{s2}-CN, β-CN, α_{s1}-CN, and PC; higher (from 8716 to 19,952) for β-LG, κ-WP, and

Table 3 Accuracies of imputations on whole-genome sequences in Holstein (HOL) and Montbéliarde (MON) breeds

Breed	HOL	MON
Number of cows	168	2142
EuroG10k chip version	V1	V4
Number of markers in the custom part	721	3082
Number of markers after quality control and MAF ≥ 0.02	221	1108
R^2 (%)	83.7	76.1
Genotypic concordance rate (%)	93.7	89.7
Allelic concordance rate (%)	96.5	94.0

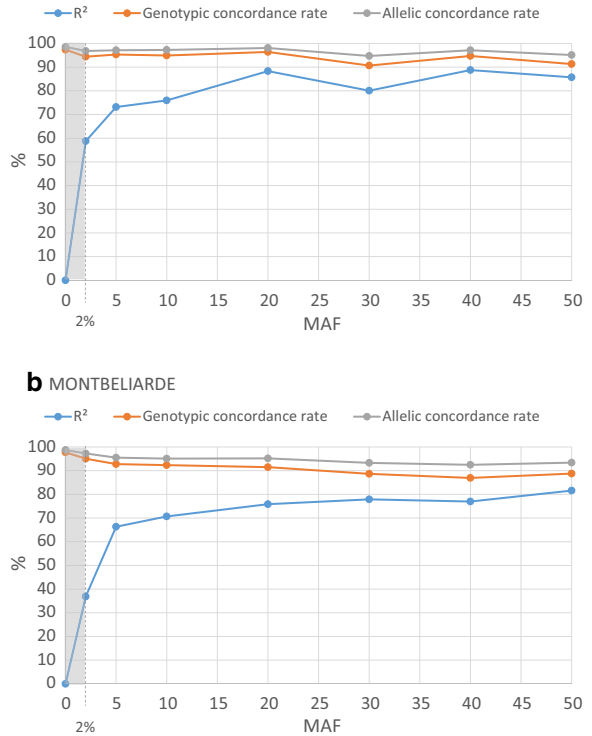


Fig. 1 Precision of imputations at the whole-genome sequence level in **a** Holstein and **b** Montbéliarde breeds, according to MAF. The 2% limit corresponds to the MAF threshold for markers used in GWAS

Σ-CN; and intermediate (from 4110 to 8248) for α-LA and κ-CN. Among these variants, 0 (PC) to 2266 (β-LG) were shared among the three breeds. Multi-breed analyses were more powerful, and detected a larger number of distinct variants with significant effects (34,248) than any of the within-breed analyses. However, the number of variants detected per trait was larger in one of the within-breed analyses than in the multi-breed analysis for PC, α-LA, β-LG, Σ-CN, and Σ-WP (Table 4), probably because of the long-range within-breed LD.

QTL regions were defined by merging the overlapping QTL regions obtained for the different traits and breeds and by grouping the corresponding significant results. Confidence intervals of these regions were defined as described in the Methods section. Thus, 34 QTL regions with significant effects on one or several milk protein composition traits were identified in within-breed and/or multi-breed analyses (see Additional file 2: Table S2). Three of these, located on chromosomes 6, 11, and 14, had significant pleiotropic effects on almost all protein composition traits analyzed (see Additional file 3: Table S3), while most (31 QTL) generally affected only one trait (see Additional file 4: Table S4).

The 34 QTL were distributed on 17 of the 29 bovine autosomes, with one to seven QTL per chromosome. Almost all of them (31) were detected in multi-breed analyses while 11, 8, and 11 QTL regions were found in MON, NOR, and HOL within-breed analyses, respectively. Four QTL regions, located on *Bos taurus* chromosome BTA6 (two regions at 45.8–46.9 Mbp and 85.2–87.4 Mbp), BTA11 (103.3 Mbp), and BTA20 (58.3–58.4 Mbp), were detected in three breeds. One additional region on BTA29 at about 9.6 Mbp was common to MON and HOL, and another region on BTA14 at 1.7–1.8 Mbp was common to HOL and NOR (Fig. 2b). The six QTL shared between two or three breeds had the most significant effects, along with one QTL detected only in the NOR breed on BTA2, at 131.8 Mbp ($-\log_{10}(P) \geq 20$; P value $< 10^{-13}$ after Bonferroni correction).

Multi-breed analyses led to the detection of a larger number of QTL regions than within-breed analyses: 14 of the 31 QTL detected in multi-breed analyses were not found in within-breed analyses. For the 17 QTL regions found in both within-and multi-breed analyses, the $-\log_{10}(P)$ value of the most significant (top) variant was almost always higher in multi-than in within-breed analyses; this was true even for most of the regions that had significant effects in only one within-breed analysis. For these QTL, the mean $-\log_{10}(P)$ value of the most significant (top) variant was 64 in multi-breed analyses versus 49, 46, and 42 in MON, NOR and HOL within-breed analyses, respectively. In addition, the QTL confidence intervals generated by multi-breed analyses contained a smaller number of variants than those produced by within-breed analyses. For the 17 QTL regions, an average of 134 variants (2–374) were found in multi-breed analyses versus 189 (39–335), 287 (61–872), and 308 (9–1236) in MON, NOR, and HOL within-breed analyses, respectively. However, in some QTL regions, specifically those located on BTA2 (131.8 Mbp), 6 (38 Mbp), and 19 (61 Mbp), the number of significant variants was smaller in within-breed analyses than in the multi-breed analysis.

Table 4 Number of variants with genome-wide significant effects ($-\log_{10}(P) > 8.4$) for milk composition traits in within- and multi-breed analyses

Trait	Within-breed analyses				Multi-breed analyses
	MON[a]	NOR[a]	HOL[a]	Shared among three breeds	
PC	1905	1201	2394	0	2350
α-LA	4590	6490	8248	213	7224
β-LG	19,952	16,048	15,517	2266	18,612
α_{s1}-CN	2232	708	629	182	2280
α_{s2}-CN	866	193	636	1	1947
β-CN	665	734	524	96	1652
κ-CN	4110	5878	6532	553	7012
Σ-CN	13,920	8716	11,833	961	12,698
Σ-WP	16,583	13,126	15,327	1916	16,546
Total number of distinct variants	29,722	27,787	30,988	3080	34,248

[a] Montbéliarde (MON), Normande (NOR), and Holstein (HOL) cows

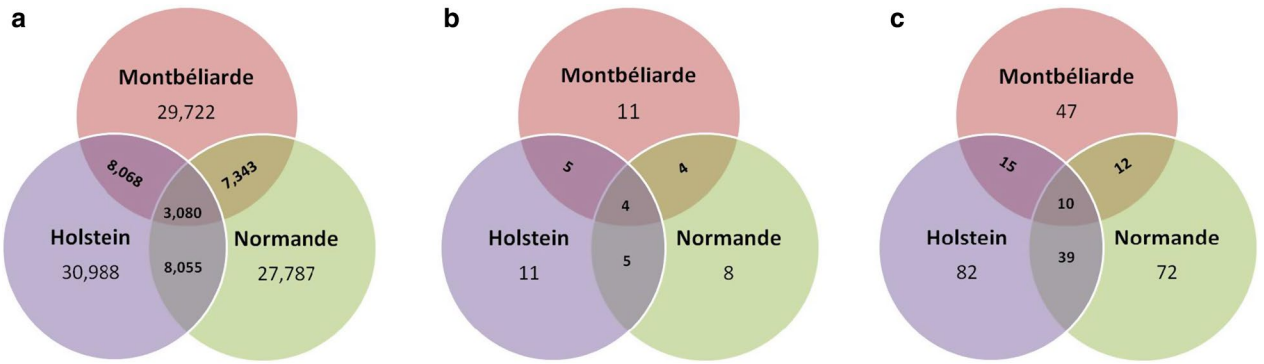


Fig. 2 Number of overlapping **a** variants with genome-wide significant effects ($-\log_{10}(P) \geq 8.4$), **b** QTL regions, and **c** genes containing variants with genome-wide significant effects among Montbéliarde, Normande, and Holstein breeds

Manhattan plots of three of the most significant QTL regions are in Fig. 3 for the three densities of markers (50 K SNP, HD SNP, or sequence). In each of these regions, several peaks are detected with the WGS data, whereas with the 50 K SNP density and in some cases with the HD SNP density, only one peak was observed.

All variants included within confidence intervals (+100 kb on each side) were functionally annotated (Table 5) and (see Additional file 5: Table S5). The percentage of variants that were located within genes ranged from 60.5% in HOL to 73.4% in NOR within-breed analyses, and it was intermediate in multi-breed analyses (65.8%). The vast majority of the genic variants were located within introns and in upstream or downstream regions. A total of 25, 82, 72, and 56 missense variants were found in MON, NOR, HOL, and multi-breed analyses, respectively; among these, we detected the previously reported missense mutations in the *PAEP* (103,303,475 bp) and *DGAT1* (1,802,266 bp) genes.

In 29 QTL regions, annotation led to the identification of candidate genes for milk protein composition. In total, 47, 72, and 82 candidate genes were identified in MON, NOR, and HOL within-breed analyses (109 in multi-breed analyses). Some of these were shared across breeds: 12 were found in both MON and NOR, 15 in MON and HOL, 39 in HOL and NOR, among which 10 were common to the three breeds (Fig. 2c). However, within a given region, the top variant was always different among the different breeds. The top variant was located in a gene in 21 of these regions, while in the remaining eight regions, the top variant was intergenic. However, these eight regions contained other variants located within confidence intervals that were annotated in genes, and of these, the most significant one was denoted the top genic variant. Genic variants with the most significant results were located within intron regions for 15 QTL and mainly upstream or downstream regulatory regions for 14 QTL. In total, 22 genes were identified as

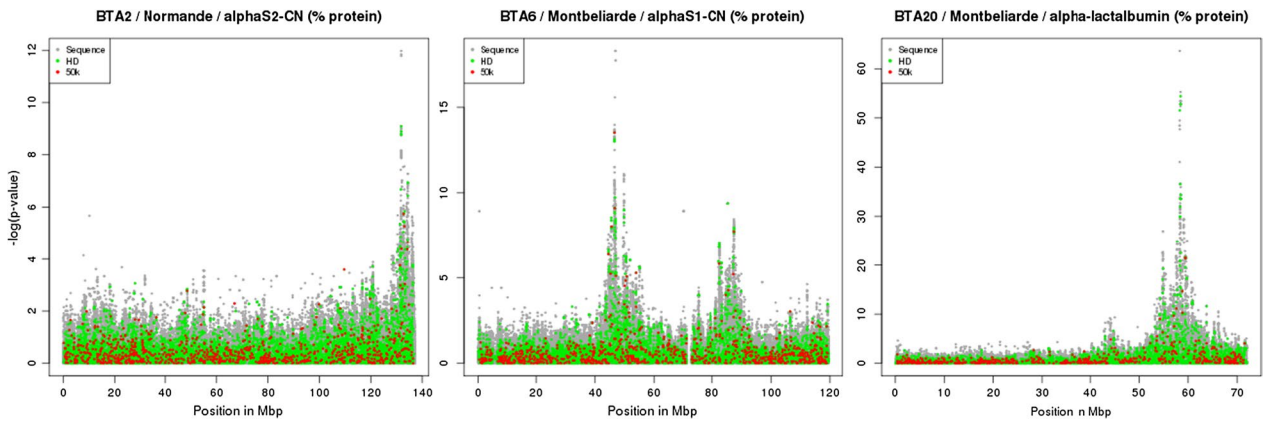


Fig. 3 $-\log_{10}(P)$ plotted against the position of variants on BTA2, 6, and 20 using three SNP densities: 50 K, HD, and whole-genome sequence SNPs

Table 5 Functional annotations of variants included within confidence intervals (±100 kb) of the 34 QTL in the three within-and multi-breed analyses

Functional annotation	Within-breed analyses			Multi-breed analyses
	MON[a]	NOR[a]	HOL[a]	
Intergenic	1514	1465	2676	1971
Intronic	1079	1804	1937	1737
3′ UTR	11	14	69	35
5′ UTR	14	27	16	18
Downstream	710	988	1276	1159
Inframe insertion	0	0	1	0
Missense	25	82	72	56
Splice acceptor	0	0	3	0
Synonymous	30	114	118	91
Upstream	509	1009	612	685
% genic	61.1	73.4	60.5	65.8
% genic non intronic	33.4	40.6	32.0	35.5

[a] Montbéliarde (MON), Normande (NOR), and Holstein (HOL) cows

the best candidates to explain the majority of the variability of milk protein composition in MON, NOR, and HOL cows. They were located on BTA1 (*SLC37A1*), BTA2 (*ALPL*), BTA5 (*MGST1*), BTA6 (*ABCG2*, *MEPE*, *PKD2*, *HERC3*, *SEPSECS*, *SEL1L3*, *DHX15*, *CSN1S1*, *CSN2*, *CSN1S2*, and *CSN3*), BTA11 (*PAEP*), BTA14 (*DGAT1*, *RECQL4*, *MROH1*, and *BOP1*), BTA20 (*ANKH*), BTA27 (*AGPAT6*), and BTA29 (*PICALM*).

Protein–protein interactions (PPI), as well as GO enrichment, were investigated for the 22 most plausible candidate genes of our study. Network proteins encoded by these genes had significantly more interactions than expected (10 edges identified; PPI enrichment P value $= 3.4 \times 10^{-9}$; Fig. 4), while GO terms for 12 biological processes, seven cellular components, and one molecular function were significantly (FDR < 0.05) enriched with two to nine of these genes for milk protein composition (Table 6).

Discussion

In this paper, we report the results of a whole-genome sequence scan for milk protein composition predicted from MIR spectra. We conducted within-and multi-breed analyses using imputed WGS of 7907 cows from three French dairy breeds. This approach led to the detection of 34 distinct regions that affect the protein composition of milk. The use of imputed WGS enabled us to confirm 22 of the 39 QTL that were previously detected from 50 K SNP genotypes [8] and to identify 12 novel QTL. In addition to genetic parameter results [4] and QTL detection results with the 50 K chip [8], these results confirm that MIR predictions are sufficiently accurate for genetic investigations. Repeated test-day records compensated for the moderate MIR prediction accuracy of some proteins.

Seventeen QTL that had been detected with 50 K SNP genotype data were not found with imputed WGS, possibly because different methods were used in the two studies (linkage disequilibrium and linkage analysis in the 50 K SNP study versus GWAS in the current imputed-WGS study) and also possibly because of the more stringent significance thresholds applied here. For GWAS on WGS data, the very stringent threshold that we used (with Bonferroni correction considering all variants as independent) probably reduced the detection power but minimized the number of false positive QTL.

Instead, the better resolution of the WGS data, combined with the power of the multi-breed GWAS approach, led to the detection of 12 QTL that were not previously found in the 50 K SNP study. To evaluate the

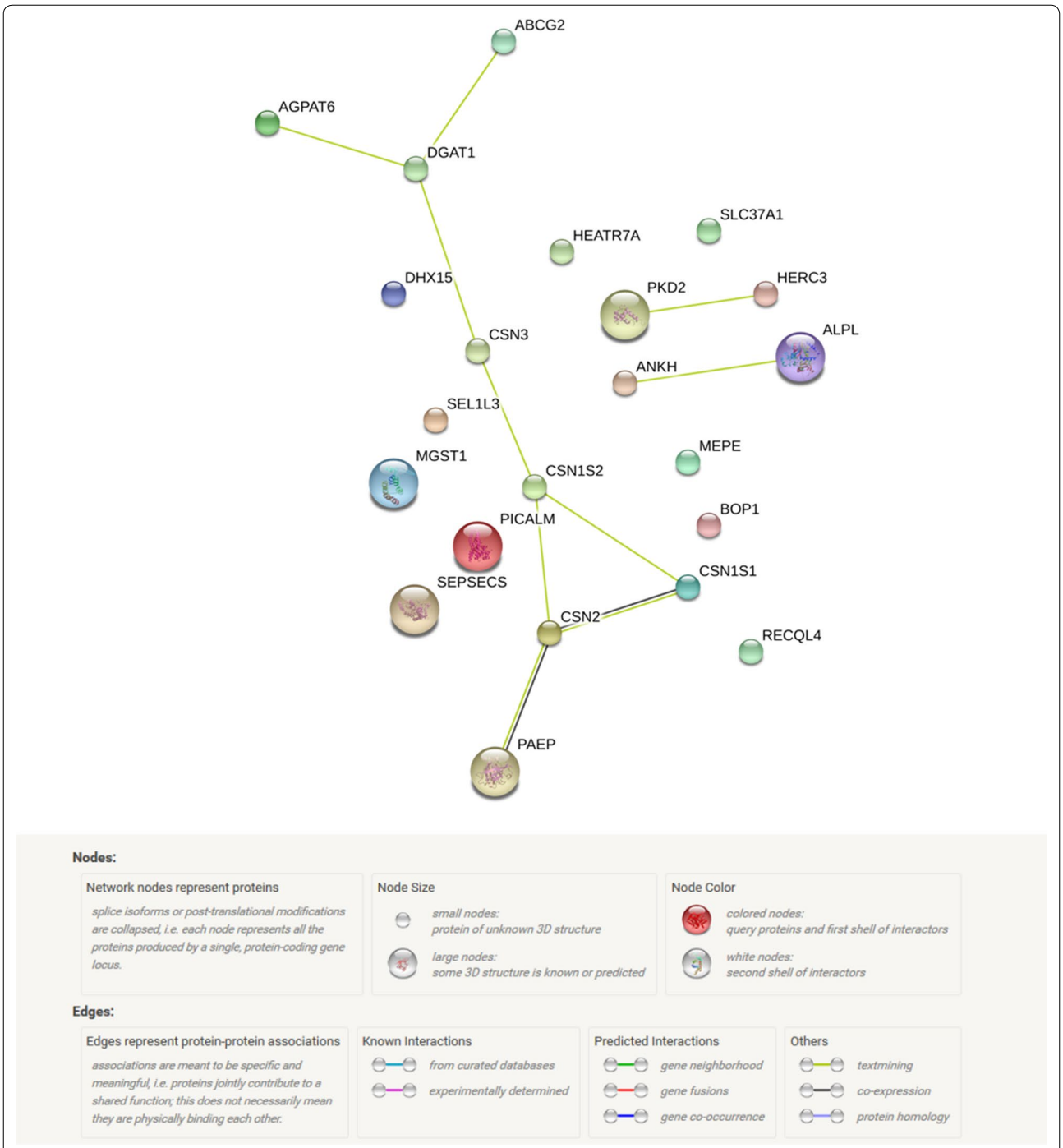


Fig. 4 Protein network of the 22 most probable candidate genes detected, according to STRING v10.0 action view

impact of marker density on GWAS results, we extracted 50 K and HD GWAS results from the WGS results. In several genomic regions, for example the regions on BTA2, 6, and 20 (Fig. 3), the increased resolution of the WGS data clearly makes it possible to identify two or more peaks whereas analysis of the 50 K SNP data detected only one peak.

Furthermore, the WGS resolution enables the use of a multi-breed approach, which is expected to better estimate the effects of rare variants and to reduce

Table 6 Gene Ontology (GO) functional enrichment with false discovery rate (FDR) < 0.05

	Pathway ID	Pathway description	Gene count	FDR	Genes
Biological process	GO.1903494	Response to dehydroepiandrosterone	4	1.73e−08	*CSN1S1, CSN1S2, CSN2, CSN3*
	GO.1903496	Response to 11-deoxycorticosterone	4	1.73e−08	*CSN1S1, CSN1S2, CSN2, CSN3*
	GO.0032570	Response to progesterone	4	1.81e−07	*CSN1S1, CSN1S2, CSN2, CSN3*
	GO.0097305	Response to alcohol	5	3.69e−07	*ALPL, CSN1S1, CSN1S2, CSN2, CSN3*
	GO.0032355	Response to estradiol	4	2.34e−06	*CSN1S1, CSN1S2, CSN2, CSN3*
	GO.1901700	Response to oxygen-containing compound	6	9.04e−05	*ALPL, CSN1S1, CSN1S2, CSN2, CSN3, PKD2*
	GO.0014070	Response to organic cyclic compound	5	0.000176	*ALPL, CSN1S1, CSN1S2, CSN2, CSN3*
	GO.0033993	Response to lipid	5	0.000181	*ALPL, CSN1S1, CSN1S2, CSN2, CSN3*
	GO.0009719	Response to endogenous stimulus	5	0.00205	*CSN1S1, CSN1S2, CSN2, CSN3, PKD2*
	GO.0048732	Gland development	3	0.0281	*CSN2, CSN3, PKD2*
	GO.0060416	Response to growth hormone	2	0.0281	*CSN1S1, CSN1S2*
	GO.0007595	Lactation	2	0.0298	*CSN2, CSN3*
Cellular component	GO.0005796	Golgi lumen	4	1.97e−08	*CSN1S1, CSN1S2, CSN2, CSN3*
	GO.0012505	Endomembrane system	8	0.00253	*AGPAT6, CSN1S1, CSN1S2, CSN2, CSN3, DGAT1, MGST1, PKD2*
	GO.0005576	Extracellular region	7	0.0372	*ALPL, CSN1S1, CSN1S2, CSN2, CSN3, PAEP, PKD2*
	GO.0005789	Endoplasmic reticulum membrane	4	0.0372	*AGPAT6, DGAT1, MGST1, PKD2*
	GO.0042175	Nuclear outer membrane-endoplasmic reticulum membrane network	4	0.0372	*AGPAT6, DGAT1, MGST1, PKD2*
	GO.0044444	Cytoplasmic part	9	0.0372	*ABCG2, AGPAT6, CSN1S1, CSN1S2, CSN2, CSN3, DGAT1, MGST1, PKD2*
	GO.0044446	Intracellular organelle part	9	0.0372	*ABCG2, AGPAT6, CSN1S1, CSN1S2, CSN2, CSN3, DGAT1, MGST1, PKD2*
Molecular function	GO.0035375	Zymogen binding	2	0.0177	*CSN1S2, CSN3*

LD between neighboring variants. With the multi-breed analysis, we detected 14 QTL that were not detected in any of our within-breed analyses (see, for example, regions of the *MGST1* and *AGPAT6* genes described below). For QTL that were detected in both within-and multi-breed analyses, the multi-breed approach provided smaller confidence intervals of the QTL than within-breed analyses. The three French breeds used in our study are not strongly related. Based on 50 K SNP data, Gautier et al. [24] reported a partitioning of the genetic diversity of cattle into distinct groups of breeds with high geographical consistency. The three breeds were classified into three distinct groups: from Eastern France and Alps for MON, from Northern European for HOL and from the Channel Islands and Northwestern France for NOR. Thus, our results illustrate the extent to which a multi-breed approach can complement and enhance the information gained from within-breed analyses even if breeds pooled in multi-breed analyses have different genetic origins.

In a previous study [25], the imputation from 50 K to HD SNP densities was found to be very accurate in all three breeds with the number of HD genotypes used here (>500) in calibration. For the second imputation step, from HD SNPs to WGS, we used the Run 4 reference population of the 1000 Bull Genomes consortium, which contained 1147 bulls, of which 288 were HOL, 28 were MON, and 24 were NOR. Due to the larger number of sequenced HOL bulls compared to the two other breeds, imputation is more accurate with HOL data than with MON data. In NOR, we anticipate that imputation accuracy is close to that obtained in MON due to similar population structures and similar numbers of whole-genome sequences for major ancestors in both breeds. These results are in agreement with or are better than those already published in cattle. Daetwyler et al. [9] showed that the use of the 1000 Bull Genome multi-breed population (Run 2, 234 bulls) led to a similar imputation accuracy among data obtained from Holstein–Friesian, Fleckvieh, and Jersey cattle (near 80% of correlation) in spite of differences in the number of bulls in the reference population (129 Holstein–Friesian, 43 Fleckvieh, and 15 Jersey). Among the *PhénoFinlait* cows genotyped with the 50 K SNP Beadchip and then imputed to WGS, 1077 MON, 238 NOR, and 498 HOL originated from nine MON, five NOR and eight HOL bulls with WGS available

in the Run4 reference population, i.e. 36, 9 and 22% of the PhénoFinlait cows, respectively. As expected, imputation accuracy dropped for variants with a low MAF. In order to limit the impact of imputation errors on the GWAS results, variants with a MAF lower than 2% were discarded from the analyses and almost all the genetic variants proposed as candidate variants in this study have moderate to high MAF.

Combining within-breed, multi-breed, and conditional GWAS analyses with functional annotations appears to be a good strategy for the differentiation of shared and breed-specific QTL. This approach also enables the direct identification of candidate genes with a very small number of candidate variants, or even in some cases, one unique variant which appears to be the best candidate to explain the observed effects.

On average, depending on the breed, between 60 and 73% of the QTL variants that we detected in the GWAS were located in genes; this is about twice as high as the percentage of genic variants at the whole-genome scale (35%; [15]). The most significant variants were located in 49 distinct genes, of which 22 were of particular interest, either because they were found in more than one breed or associated with several traits, or because they were previously described as influencing milk composition. These 22 genes, which are located in 11 distinct genomic regions and present significant protein–protein enrichment, are the most plausible candidates to explain a large part of the variation in milk protein composition among MON, NOR, and HOL cows. In four genomic regions (on BTA1, 2, 11, and 27), we identified one unique candidate variant (or a few candidate variants in LD) shared by all three breeds (in the *SLC37A1*, *ALPL*, *PAEP*, and *AGPAT6* genes, respectively). In three other genes, we suggest the presence of a breed-specific candidate variant (*MGST1* on BTA5 and *PICALM* on BTA29) or several candidate causative variants (*ANKH* on BTA20). Finally, four regions, including the *DGAT1* region on BTA14 and three regions on BTA6 (*ABCG2* region, a region at about 46 Mbp, and the casein gene cluster), were more complex, because they contained several candidate genes, each with several candidate variants. Eight of these candidate genes (*SLC37A1*, *MGST1*, *CSN1S1*, *CSN2*, *CSN1S2*, *CSN3*, *PAEP*, and *ANKH*) are known to be overexpressed in the mammary gland compared to other 17 tissues [26] and between two and nine of them are associated with one of the 20 GO terms in our study. The next sections describe these regions in more detail.

SLC37A1 (BTA1) and αs1-CN/ α-LA

The *SLC37A1* (*solute carrier family 37, member A1*) gene, which encodes a glucose-6-phosphate transporter that is involved in the homeostasis of blood glucose, is highly expressed in the mammary gland [27]. It could be a good candidate gene to explain the effects of the QTL identified on BTA1 on αs1-CN in both MON and multi-breed analyses and on the α-LA phenotype in the multi-breed analysis. In total, 138 distinct variants of this gene were located within the confidence intervals of the QTL, of which 133 were intronic, two were synonymous, and three were downstream (see Additional file 6: Figure S1a). For the αs1-CN/MON, αs1-CN/multi, and α-LA/multi results, the 80, 81, and 74 most significant variants in the peaks, respectively, were in intronic regions of *SLC37A1*. One downstream variant was detected for αs1-CN in the MON analysis, which ranked 104th among the significant variants, while multi-breed analyses revealed three downstream variants that ranked 81st, 87th, and 103rd. All intronic variants that are located at the top of the peaks are in strong LD but only one variant (indel), located at 144,397,274 bp, was common to all three TOP10 lists; it was 1st in the αs1-CN/MON ranking, 9th in the αs1-CN/multi-breed ranking, and 4th in the α-LA/multi-breed ranking. The top1 intronic variant detected in the αs1-CN/multi-breeds analysis, at 144,398,814 bp, ranked 75th in the αs1-CN/MON peak and 76th in the α-LA/multi-breed peak.

Two previous studies described the effects of *SLC37A1* gene variants on milk production traits. In an analysis of HD SNP genotypes, Kemper et al. [27] described six variants that are located between 144.325 and 144.525 Mbp in this region; the variant with the most significant effect was located in an intronic region of the gene (144,414,936 bp). In our study, this variant was included within the confidence interval of the QTL detected by the multi-breed analysis ($-\log_{10}(P) = 10.2$), but it ranked 101st. Two other intronic variants in strong LD in the *SLC37A1* gene, at 144,367,474 and 144,377,960 bp, were previously proposed as the best candidate mutations for changes in phosphorus concentration and milk production traits [28]. However, in our study in spite of relatively high MAF values (from 0.30 to 0.41 depending on the breed), these variants had a $-\log_{10}(P)$ value lower than 6 for all analyzed traits. In another study of targeted QTL regions after imputation to WGS level, the variant with the most significant effects was located at 144,381,564 bp [29]. This variant is close to the candidate variant identified in our analysis, but it can be excluded as the causal variant in our populations since it is monomorphic in the MON, NOR, and HOL individuals analyzed here.

The conditional analyses that we performed included the two best candidate variants as well as the candidate variant described by Kemper et al. [27]. These revealed that including the variant located at 144,398,814 bp in the model completely removed the original signal while with each of the two other variants, a less significant peak

persisted (see Additional file 6: Figure S1a). This variant, which has contrasting effects on αs1-CN and α-LA phenotypes, but with a more marked effect on the former, therefore constitutes the most probable candidate variant for the effects detected in our study.

ALPL (BTA2) and αs2-CN

The QTL identified on BTA2 at 131.8 Mbp had significant effects on several traits (αs2-CN, β-CN, and κ-CN). In particular, although the αs2-CN-associated peaks were detected in all within-and multi-breed analyses, even if in the MON and HOL analyses, the maximal $-\log_{10}(P)$ values did not reach the stringent threshold of 8.4 that we applied in this study (7 and 6.9, respectively; see Additional file 6: Figure S1b). In all analyses, the most significant variants were located in intronic regions of the *ALPL* (*alkaline phosphatase*) gene, which encodes a member of the alkaline phosphatase family of proteins. The most significant variant differed among the three within-breed analyses: it was located at 131,806,882 bp in NOR, 131,850,456 bp in MON, and 131,808,301 bp in HOL sequences. Instead, the top-ranked variant in the peak detected in the multi-breed GWAS was located at 131,806,882 bp. All three single-breed conditional analyses that included each of these variants as fixed effects lacked peaks (see Additional file 6: Figure S1b). These results suggest that all three intronic variants are in strong LD in the three breeds and that the causal mutation could be shared among breeds. Among all the variants at the top of the peaks, the intronic variant at 131,806,882 bp appears to be the most probable candidate variant in the *ALPL* gene for the observed effects on αs2-CN content; it ranked 1st, 6th, 26th, and 1st in the NOR, MON, HOL, and multi-breed peaks, respectively.

MGST1 (BTA5) and milk protein content (PC)

One region on BTA5 that contains 63 variants affected PC in the multi-breed analysis. The MON and NOR within-breed analyses revealed no peaks ($-\log_{10}(P) < 6$), whereas the HOL analysis detected a single peak with a $-\log_{10}(P) = 8$, which was close to the significance threshold of 8.4. Only one gene, *MGST1* (*microsomal glutathione S-transferase 1*), was present within the confidence interval obtained in the multi-breed analysis. Fifty-one variants were located in intronic (29), exonic (1 synonymous), 5′-UTR (2), or regulatory (19 in the upstream region) regions of the gene. The variant with the most significant effects was located at 93,950,211 bp in the upstream region and its $-\log_{10}(P)$ value was 9.3, versus a value of 8.0 for the variants that ranked 2nd (93,950,116 bp and 93,950,288 bp), which were located, respectively, in the 5′-UTR and upstream regions of the gene. The MAF value for these variants was low in the MON population (0.006; <MAF threshold of 0.02) and ranged from 0.08 to 0.12 in NOR, from 0.37 to 0.42 in HOL, and from 0.19 to 0.22 in the multi-breed population. Thus, the fact that peaks were detected only in HOL (close to significance) and multi-breed (significant) analyses could be due to the relatively low MAF for these variants in MON and NOR. The most significant variants in our study are located near a variant that was reported by Raven et al. [29] to be responsible for changes in fat percentage in Holstein cows (at 93,951,731 bp (upstream) and ranked 23rd in our study) and also near variants previously linked to fat yield by Iso-Touru et al. [30] and Van den Berg et al. [31] (93,945,694 and 93,945,738 bp, respectively; both were intronic variants and were not significant here). Conditional analyses including each of the six variants as a fixed effect showed that all variants except those reported by Iso-Touru et al. [30] and Van den Berg et al. [31] explained the effects observed in our study (see Additional file 6: Figure S1c). Thus, the effects observed on fat content by Raven et al. [29] and on protein content in our study could be explained by the same causative variant. Recently, Littlejohn et al. [32] confirmed that *MGST1* has causative pleiotropic effects on milk composition (percentage and yield of fat, protein, and lactose). These authors failed to identify causative variants in the gene but they pointed to a cluster of 17 variants that were grouped in a 10-kbp segment of the *MGST1* gene (93,944,937–93,954,751). Only one of these 17 variants is located in the confidence interval of the QTL that we detected and this is an intronic variant (93,949,810 bp) that ranked 7th in the peak in spite of having a higher MAF (0.32) than the most significant variants (MAF = 0.19–0.22). Thus, our study highlights three new candidate mutations in the *MGST1* gene, which are located very close to each other, in the 5′-UTR region (93,950,116 bp) or in the upstream region (93,950,211 and 93,950,288 bp) of the gene.

ABCG2, *MEPE*, *PKD2*, and *HERC3* (BTA6) and αs1-CN

Several QTL were found on BTA6. The first one, detected in HOL and multi-breed analyses, was located in the 37.6–38.4 Mbp region, which contains the Y581S polymorphism of the *ABCG2* gene (38,027,010 bp) that was described by Cohen-Zinder et al. [33] as a causative mutation for changes in milk yield and composition. This missense variant had MAF values of 0.0029 and 0.0018 in HOL and multi-breed populations, respectively, and therefore did not pass the MAF filter in both analyses. In spite of a low MAF, the Y581S polymorphism had a highly significant effect on the αs1-CN phenotype in both HOL and multi-breed analyses, with $-\log_{10}(P)$ values of 31 and 21, respectively; these values were higher than those of the top variant in the peaks after filtering

for MAF (20 and 15, respectively). However, among the sires of the HOL cows, six bulls were previously found to be heterozygous for the QTL detected in this region, but homozygous for this mutation [8]. Thus, we suggest that other mutations could be responsible for the QTL that affects milk protein composition.

In the HOL analysis, nine variants with MAF ranging from 0.022 to 0.041 were located within the confidence interval of the QTL. The most significant variants were located in intronic regions of the *ABCG2* gene, at 38,015,146 and 38,020,110 bp. Other variants, which are located in three other genes, i.e. *MEPE* (one downstream), *PKD2* (one intronic), and *HERC3* (two intronic), also had highly significant effects on αs1-CN. Due to the relatively low MAF of the candidate variants located in this region, these results require further analyses, including a larger number of animals and more accurate imputation or direct genotyping.

SEL1L3, *SEPSECS*, and *DHX15* (BTA6) and αs1-CN

In all within-breed and multi-breed analyses, the αs1-CN phenotype was affected by another region of BTA6 at 45.8–46.9 Mbp. However, the most likely candidate genes differed among breeds. In MON, the nine variants with the most significant effects were located in intronic regions of the *SEL1L3* gene (max. at 46,874,151 bp). In NOR, the top 116 variants in the peak were intergenic, while the genic variant with the most significant effects was located in an intron of the *SEPSECS* gene (46,277,697 bp). In HOL, the most significant genic variants (*DHX15*) ranked 16th in the peak (45,639,181 and 45,640,564 bp). Finally, among the top 80 variants detected by the multi-breed analysis, only one was genic, which was located in an intron of the *SEL1L3* gene (46,874,514 bp, ranked 3rd in the MON analysis). There is insufficient concordance among these results to propose a single set of candidate variants.

Pleiotropic effects of the casein gene region (BTA6)

On BTA6, we found a QTL that affected both the overall protein content of milk and the content of all four individual caseins in all three breeds. Variants with the most significant effects were located in an 840-kb interval that contains the 250-kb casein gene cluster (87,062,878–87,903,002 bp); other variants with effects on αs1-CN and β-CN in MON were located at 85.2 Mbp. In all within- and multi-breed analyses, the most significant effects were detected for the κ-CN phenotype, followed by αs1, αs2, or β-CN depending on the breed. In each analysis, the variant with the most significant effects on κ-CN was located within or in the immediate vicinity of the *CSN3* gene, which encodes the κ casein: at 87,376,747 bp (upstream) in NOR, 87,392,592 bp (5′-UTR) in MON and multi-breed, and 87,394,293 bp (downstream) in HOL. Each of these variants, as well as the κ casein A/B variant (87,390,576 bp, missense), was therefore included as a fixed effect in the conditional analyses. The results were breed-specific: in MON, the κ-CN-associated peak disappeared after fixing the upstream, missense, or 5′-UTR variant; in HOL, the peak disappeared after fixing the upstream, 3′-UTR, or downstream variant; but in NOR, the peak remained with the inclusion in the model of any of the four variants. Thus, none of the four candidate variants succeeded in explaining all the effects observed on κ-CN in the three breeds.

Instead, the peaks associated with the αs2-CN and β-CN phenotypes in NOR and the PC and αs2-CN phenotypes in MON could be explained by two distinct groups of six SNPs in complete LD, which were respectively located in the *CSN2* gene (three downstream and three intronic) and in the upstream region of the *ODAM* (odontogenic ameloblast-associated) gene (between the *CSN1S2* and *CSN3* genes).

Finally, the A1/B and A2 variants of *CSN2*, which ranked 147th and 86th, respectively, for their effects on PC and αs2-CN in NOR, were responsible for the αs2-CN phenotype in NOR but not for any other effect on the other traits or in the other breeds.

These results illustrate the complexity that is inherent with the analysis of the casein gene cluster, which contains the four genes *CSN1S1-CSN2-CSN1S2-CSN3* (encoding, respectively, αs1, β, αs2, and κ caseins). The polymorphisms of the amino-acid sequences of caseins are well known, and the effects on milk composition and cheese-making abilities have been well described (reviewed in Grosclaude et al. [34] and Caroli et al. [35]). Nevertheless, the effects of known polymorphisms are not always consistent between studies, likely because variations in the content of individual caseins are caused by several linked polymorphisms in the casein genes. Thus, it is likely that the most significant variants highlighted in our study are those that better explain haplotype effects. A multi-marker approach could facilitate efforts to distinguish the effects of all the causal polymorphisms located in this region.

Pleiotropic effects of the *PAEP* gene region (BTA11)

The most significant effects on protein composition were found for variants that are located on BTA11. Contents of each individual protein in milk, with the exception of αs2-CN, were affected by this region in all three breeds. Effects were most significant for β-LG and, to a lesser extent, for κ-CN in all within-and multi-breed analyses. All of the most significant variants were located in or close to the *PAEP* (*progestagen-associated endometrial protein*) gene, also named *LGB* gene, which encodes the

β-LG protein. The β-LG protein variants A and B, which are common in most cattle breeds, are associated with different β-LG levels in milk [34]. They differ by two amino-acid substitutions, caused by two missense mutations at 103,303,475 and 103,304,757 bp [36]. Interestingly, although these two variants had highly significant effects on β-LG in our study, they did not rank high in the peaks. In the MON and NOR analyses, both mutations were in complete LD and ranked 85th and 213rd, respectively, while in HOL, the two mutations ranked 48th and 109th, respectively (116th and 120th in multi-breed analysis). As suggested by Ganai et al. [36], differences in β-LG content may be caused by different levels of expression of the A and B alleles rather than by the direct effect of amino-acid substitutions. Among the top 30 variants in the within-and multi-breed analyses, only one, located at 103,298,431 bp in the upstream region of the *PAEP* gene, was shared by the four analyses. Moreover, this variant is one of the most significant in each analysis, ranking 6th, 4th, 1st, and 3rd, respectively, in the MON, NOR, HOL, and multi-breed analyses. The inclusion in conditional analyses of one of the causal missense variants or the most probable upstream variant identified in our study led to similar results in MON and HOL but not in NOR (see Additional file 6: Figure S1d). A peak remained in the conditional NOR analysis when missense mutations were fixed, but disappeared with the inclusion of the upstream variant at 103,298,431 bp. Thus, these results indicate that the missense mutations that cause the A and B variant protein polymorphisms do not explain all the variation associated with this region. Another variant, which is located in a regulatory region of *PAEP*, is more or less linked to the missense variants depending on the breed and appears to be a good candidate to explain different levels of expression of β-LG protein variants.

Pleiotropic effects of the *DGAT1* gene region (BTA14)

Very significant effects on different protein composition traits were associated with the region of the *DGAT1* gene in NOR and HOL but not in MON. This region affected PC and κ-CN in both NOR and HOL; αs1-CN, β-CN, and α-LA only in NOR, and αs2-CN only in HOL. Moreover, individual proteins with the lowest *P* value were κ-CN in NOR and αs2-CN in HOL. The A allele of the *DGAT1* K232A polymorphism, which decreases fat and protein percentages as well as fat yield, and increases milk and protein yields [37], was present at a frequency of 9.4% in NOR, 15.8% in HOL, and only 0.6% in MON. However, our study confirmed that this causative variant was not the most significant for all traits analyzed. It ranked 18th to 72th among variants in the NOR analysis, depending on the trait, and outside the confidence interval for all traits in HOL. These results suggest, first, that not all variations observed in this region are associated with the K232A polymorphism and, second, that other specific causative mutations could explain the effects detected in NOR and HOL.

A large number of genes are annotated in the 1-Mbp region between 1.5 and 2.5 Mbp on BTA14 and, depending on the trait and the breed, between 66 and 494 variants located within the confidence intervals of this QTL are located in 17 to 30 of those genes. Among the top 50 variants for all traits, six were missense variants, of which two were found in NOR (*DGAT1* and *BOP1*) and four in HOL (three in *RECQL4* and one in *MROH1*). In this region, no variant remained significant in the conditional analyses for NOR when the *DGAT1* (K232A) or *BOP1* (1,842,678 bp) variants were included, and for HOL when the variants in *RECQL4* (one of the three variants in complete LD: 1,617,841, 1,618,978 and 1,619,555 bp) or *MROH1* (1,878,165 bp) were included. In contrast, a less significant peak persisted when the *DGAT1* or *BOP1* variant was included in the HOL analyses and when the *RECQL4* or *MROH1* variant was included in the NOR analyses (see Additional file 6: Figure S1e). Among the six missense variants, only the *RECQL4* variant at 1,617,841 bp has a predicted deleterious effect, with a SIFT score of zero. Therefore, in addition to the *DGAT1* K232A polymorphism previously identified as having effects on milk composition, we report additional candidate missense mutations in *BOP1*, *MROH1*, and *RECQL4* genes, which could be partly responsible for the effects associated with the centromeric end of BTA14.

ANKH (BTA20) and α-LA

The GWAS on WGS data detected a QTL with very significant effects on the α-LA phenotype in all three within-breed analyses and in the multi-breed analysis; this confirmed our previous report based on a GWAS using 50 K SNP data [8]. Confidence intervals of the QTL included between one and four genes depending on the within- or multi-breed analysis, and *ANKH* was the only gene to be highlighted in all four analyses. *ANKH* encodes an inorganic pyrophosphate transport regulator that helps to prevent the deposition of minerals (calcium and phosphorous) in bones and α-LA exhibits a high affinity to metal ions, calcium in particular. In addition, *ANKH* is highly expressed in mammary tissue in Holstein and Jersey cows [27] and we observed a significant interaction between *ANKH* and *ALPL* (candidate gene on BTA2 for effects on αs2-CN), which suggests a functional link between these two genes (Fig. 4). Thus, *ANKH* constitutes a good functional candidate for effects on α-LA in HOL, MON, and NOR. However, none of the top 50 variants in this QTL were shared among the

three breeds. In each breed, the most significant variant was located either in intronic regions of the *ANKH* gene (at 58,422,697 bp in NOR and at 58,450,656 bp in multi-breed analyses) or in an intergenic region. In MON and HOL, for which the most significant variants were intergenic, *ANKH* intronic variants ranked 2nd (at 58,446,560 bp) and 13th (at 58,491,204 bp), respectively. After fixing the most significant variant from each within-breed analysis, a peak remained in all conditional analyses (see Additional file 6: Figure S1f), which suggests that several causative mutations in the *ANKH* gene could be responsible for the variation of the amount of α-LA in milk. The most significant variants could be those that are most tightly linked to the causative mutations in each breed, which could explain why they were breed-specific.

AGPAT6 (BTA27) and κ-CN

The multi-breed analysis detected a QTL for κ-CN content located at about 36.2 Mbp on BTA27, while in within-breed analyses, peaks were present in MON and NOR but they did not reach significance ($-\log_{10}(P) < 8.4$), and no peak was observed in HOL (see Additional file 6: Figure S1g). In the multi-breed analysis, the four most significant variants were located in an intergenic region but the variants that ranked 5th to 17th were located in the *AGPAT6* gene, which was previously described as a functional gene for milk fat content with pleiotropic effects on other milk components, in particular protein content [38]. The five most significant variants in the gene were in complete LD and located in the upstream region (at 36,209,319, 36,211,252, 36,211,258, and 36,211,708 bp) or in the 5′-UTR region (at 36,212,352 bp) of the *AGPAT6* gene. For the five linked variants, MAF were equal to 0.46 in MON, 0.47 in NOR, and 0.39 in HOL (0.44 in multi-breed population). When the κ-CN phenotype was conditioned on the effect of any of these mutations, the association signals completely disappeared in the MON, NOR, and multi-breed analyses (see Additional file 6: Figure S1g). The four variants located in the upstream region were previously identified as candidate causal polymorphisms in both Holstein and Fleckvieh cows by Daetwyler et al. [9]. These authors pointed to the polymorphism at 36,211,252 bp as the most plausible causative mutation because it presented a high probability of being within a transcription binding site. In addition, Littlejohn et al. [38] described strong associations between milk composition traits (fat, protein, and lactose) and 10 variants in the *AGPAT6* gene. Three of these 10 variants were among the most significant variants in our study, located at 36,209,319, 36,211,708, and 36,212,352 bp. Thus, we identified five putative causative variants in the *AGPAT6* gene for milk protein composition; of these, the variant at 36,212,352 bp appears to be the most plausible causative mutation because it is located in the 5′-UTR region of the *AGPAT6* gene. However, the lack of a significant effect in the HOL analyses, in spite of the high MAF of the candidate variants, probably reflects additional effects yet to be explained.

PICALM (BTA29) and αs1-CN

The αs1-CN phenotype was influenced by a genomic region that is located at about 9.5 Mbp on BTA29. Significant associations were found in MON, HOL, and multi-breed analyses, and a peak close to significance was found in NOR ($-\log_{10}(P) = 7.9$) (see Additional file 6: Figure S1h). In the MON and HOL analyses, the most significant variants were intergenic and, likewise, in the multi-breed analysis, all nine variants located within the confidence interval were intergenic. The most significant non-intergenic variants were located in the *PICALM* gene in MON and HOL. Two intronic variants ranked 11th in the peak detected in MON (9,651,065 and 9,656,439 bp) and one variant that ranked 10th in the HOL analysis, is located in the upstream region of the gene (9,611,304 bp). When conditional GWAS analyses were performed, the inclusion of the intronic variants removed the peak in MON but not in HOL analyses, and conversely, inclusion of the upstream variant removed the peak in HOL but not in MON analyses. In NOR, the peak in question persisted when either intronic or upstream variants were fixed (see Additional file 6: Figure S1h). These results suggest that either the causative variant is different between breeds or that several linked causative variants explain the significant effects observed in this region. The *PICALM* gene encodes a phosphatidylinositol-binding clathrin assembly protein, and polymorphisms in this gene are associated with the risk of Alzheimer's disease [39] in humans. However, to date, no link was reported between polymorphisms in this gene and bovine milk composition

Conclusions

Our study provides evidence that a GWAS-based approach applied to fine-scale phenotypes, whole-genome sequences, and multiple breeds provides enough resolution to identify candidate genes and directly pinpoint a limited number of candidate variants in most of these genes. Several variants, some shared among breeds, were identified as plausible candidate mutations for changes in milk protein composition in the three main French dairy cattle breeds. They were located both in genes that had previously been found to affect milk composition (*SLC37A1*, *MGST1*, *ABCG2*, *CSN1S1*, *CSN2*, *CSN1S2*, *CSN3*, *PAEP*, *DGAT1*, *AGPAT6*) and in genes for which no such relationship was known (*ALPL*,

ANKH, *PICALM*). In the future, functional analyses will enable the establishment of causative links between these candidate variants and milk protein phenotypes. However, even before such studies are completed, our results offer the opportunity to improve cheese-making properties through the identification of genetic variants associated with changes in milk composition. Direct consequences of these results on practical selection are not obvious and depend on potential premiums on protein composition and on incentives proposed by the milk processing industry. Nevertheless, it would be desirable to favour caseins against whey proteins at least for milk collected for cheese production. Such an option could be implemented by including variants that affect individual proteins in genomic evaluation models.

Additional files

Additional file 1: Table S1. The 1000 bull genome population (RUN4). (Daetwyler HD, personal communication).

Additional file 2: Table S2. Number of variants included within confidence intervals for each QTL region and trait, regardless of breed.

Additional file 3: Table S3. Description of the pleiotropic QTL regions detected in within-breed (MON, NOR, or HOL) or multi-breed (Multi) analyses.

Additional file 4: Table S4. Description of other significant QTL regions detected in within-breed (MON, NOR, or HOL) or multi-breed (Multi) analyses.

Additional file 5: Table S5. Functional annotations of variants included within confidence intervals (± 100 kb) of the 34 QTL for each trait in the three within-breed Montbéliarde (MON), Normande (NOR), and Holstein (HOL) or in multi-breed analyses.

Additional file 6: Figure S1. $-\log_{10}(P)$ plotted against the position of variants detected by GWAS (in *grey*) and conditional GWAS (GWAS_COJO; in *blue*) **a** On BTA1, **b** BTA2, **c** BTA5, **d** BTA11, **e** BTA14, **f** BTA20, **g** BTA27 and **h** BTA29

Authors' contributions

MPS analyzed the data and wrote the manuscript. DB, SF, and MBr designed the *PhénoFinlait* project. GM and PM provided reference analyses for milk protein composition. AG, PC, CH, AB, and MBo provided support in computing. RL and DR contributed to the estimation of imputation accuracy. All authors read and approved the final manuscript.

Author details

[1] GABI, INRA, AgroParisTech, Université Paris Saclay, 78350 Jouy-en-Josas, France. [2] Institut de l'Elevage, 75012 Paris, France. [3] Allice, 75012 Paris, France.

Availability of data and material

GWAS results obtained during the current study are available from the corresponding author on reasonable request.

Acknowledgements

The authors gratefully acknowledge the breeders who participated in the *PhénoFinlait* program; colleagues from the Institut de l'Elevage and INRA who designed and coordinated the farm sampling program and data collection; the partners of the program, laboratories, manufacturers, and DHI organizations who provided data; Marion Ferrand who developed the MIR prediction equations; and the members of the *PhénoFinlait* scientific committee who advised and managed this work. The authors would also like to thank the contribution of the 1000 Bull Genomes consortium.

Competing interests

The authors declare that they have no competing interests.

Funding

The *PhénoFinlait* program received financial support from ANR (ANR-08-GANI-034 Lactoscan), APIS-GENE, CASDAR, CNIEL, FranceAgriMer, France Génétique Elevage, and the French Ministry of Agriculture. The *Cartoseq* project was funded by ANR (ANR10-GENM-0018) and APIS-GENE.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 7 April 2017 Accepted: 30 August 2017
Published online: 18 September 2017

References

1. Schopen GC, Heck JM, Bovenhuis H, Visker MH, van Valenberg HJ, van Arendonk JA. Genetic parameters for major milk proteins in Dutch Holstein-Friesians. J Dairy Sci. 2009;92:1182–91.
2. Bonfatti V, Cecchinato A, Gallo L, Blasco A, Carnier P. Genetic analysis of detailed milk protein composition and coagulation properties in Simmental cattle. J Dairy Sci. 2011;94:5183–93.
3. Gebreyesus G, Lund MS, Janss L, Poulsen NA, Larsen LB, Bovenhuis H, et al. Short communication: multi-trait estimation of genetic parameters for milk protein composition in the Danish Holstein. J Dairy Sci. 2016;99:2863–6.
4. Sanchez MP, Ferrand M, Gelé M, Pourchet D, Miranda G, Martin P, et al. Short communication: genetic parameters for milk protein composition predicted using mid-infrared spectroscopy in the French Montbéliarde, Normande, and Holstein dairy cattle breeds. J Dairy Sci. 2017;100:6371–5.
5. Wedholm A, Larsen LB, Lindmark-Månsson H, Karlsson AH, Andrén A. Effect of protein composition on the cheese-making properties of milk from individual dairy cows. J Dairy Sci. 2006;89:3296–305.
6. Bonfatti V, Di Martino G, Carnier P. Effectiveness of mid-infrared spectroscopy for the prediction of detailed protein composition and contents of protein genetic variants of individual milk of Simmental cows. J Dairy Sci. 2011;94:5776–85.
7. Ferrand M, Miranda G, Guisnel S, Larroque H, Leray O, Lahalle F, et al. Determination of protein composition in milk by mid-infrared spectrometry. In Proceedings of the international strategies and new developments in milk analysis VI ICAR Reference Laboratory Network Meeting: 28 May 2012; Cork. 2013;16:41–5.
8. Sanchez MP, Govignon-Gion A, Ferrand M, Gele M, Pourchet D, Amigues Y, et al. Whole-genome scan to detect quantitative trait loci associated with milk protein composition in 3 French dairy cattle breeds. J Dairy Sci. 2016;99:8203–15.
9. Daetwyler HD, Capitan A, Pausch H, Stothard P, Van Binsbergen R, Brøndum RF, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet. 2014;46:858–67.
10. Raven L, Cocks B, Hayes B. Multibreed genome wide association can improve precision of mapping causative variants underlying milk production in dairy cattle. BMC Genomics. 2014;15:62.
11. Ducrocq V. Genekit, BLUP software. Version June 2011.
12. Sargolzaei M, Chesnais JP, Schenkel FS. A new approach for efficient genotype imputation using information from relatives. BMC Genomics. 2014;15:478.
13. van Binsbergen R, Bink MC, Calus MP, van Eeuwijk FA, Hayes BJ, Hulsegge I, et al. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genet Sel Evol. 2014;46:41.
14. Bouwman AC, Veerkamp RF. Consequences of splitting whole-genome sequencing effort over multiple breeds on imputation accuracy. BMC Genet. 2014;15:105.

15. Boussaha M, Michot P, Letaief R, Hoze C, Fritz S, Grohs C, et al. Construction of a large collection of small genome variations in French dairy and beef breeds using whole-genome sequences. Genet Sel Evol. 2016;48:87.
16. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
17. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP effect predictor. Bioinformatics. 2010;26:2069–70.
18. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4:1073–82.
19. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88:76–82.
20. Fu WX, Liu Y, Lu X, Niu XY, Ding XD, Liu JF, et al. A genome-wide association study identifies two novel promising candidate genes affecting *Escherichia coli* F4ab/F4ac susceptibility in swine. PLoS One. 2012;7:e32127.
21. Yang j, Ferreira T, Morris AP, Medland SE, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication and Meta-analysis (DIAGRAM) Consortium, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet. 2012;44:369–75.
22. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl variant effect predictor. Genome Biol. 2016;17:122.
23. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucl Acids Res. 2015;43:D447–52.
24. Gautier M, Laloe D, Moazami-Goudarzi K. Insights into the genetic history of French cattle from dense SNP data on 47 worldwide breeds. PLoS One. 2010;5:e13038.
25. Hoze C, Fouilloux MN, Venot E, Guillaume F, Dassonneville R, Fritz S, et al. High-density marker imputation accuracy in sixteen French cattle breeds. Genet Sel Evol. 2013;45:33.
26. Chamberlain AJ, Vander Jagt CJ, Hayes BJ, Khansefid M, Marett LC, Millen CA, et al. Extensive variation between tissues in allele specific expression in an outbred mammal. BMC Genomics. 2015;16:993.
27. Kemper KE, Reich CM, Bowman PJ, vander Jagt CJ, Chamberlain AJ, Mason BA, et al. Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy of across-breed genomic predictions. Genet Sel Evol. 2015;47:29.
28. Kemper KE, Littlejohn MD, Lopdell T, Hayes BJ, Bennett LE, Williams RP, et al. Leveraging genetically simple traits to identify small-effect variants for complex phenotypes. BMC Genomics. 2016;17:858.
29. Raven LA, Cocks BG, Kemper KE, Chamberlain AJ, Vander Jagt CJ, Goddard ME, et al. Targeted imputation of sequence variants and gene expression profiling identifies twelve candidate genes associated with lactation volume, composition and calving interval in dairy cattle. Mamm Genome. 2016;27:81–97.
30. Iso-Touru T, Sahana G, Guldbrandtsen B, Lund MS, Vilkki J. Genome-wide association analysis of milk yield traits in Nordic red cattle using imputed whole genome sequence variants. BMC Genet. 2016;17:55.
31. van den Berg I, Boichard D, Lund MS. Comparing power and precision of within-breed and multibreed genome-wide association studies of production traits using whole-genome sequence data for 5 French and Danish dairy cattle breeds. J Dairy Sci. 2016;99:8932–45.
32. Littlejohn MD, Tiplady K, Fink TA, Lehnert K, Lopdell T, Johnson T, et al. Sequence-based association analysis reveals an *MGST1* eQTL with pleiotropic effects on bovine milk composition. Sci Rep. 2016;6:25376.
33. Cohen-Zinder M, Seroussi E, Larkin DM, Loor JJ, Everts-van der Wind A, Lee JH, et al. Identification of a missense mutation in the bovine *ABCG2* gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. Genome Res. 2005;15:936–44.
34. Grosclaude F. Le polymorphisme génétique des principales lactoprotéines bovines. INRA Prod Anim. 1988;1:5–17.
35. Caroli AM, Chessa S, Erhardt GJ. Invited review: milk protein polymorphisms in cattle: effect on animal breeding and human nutrition. J Dairy Sci. 2009;92:5335–52.
36. Ganai NA, Bovenhuis H, van Arendonk JA, Visker MH. Novel polymorphisms in the bovine *beta-lactoglobulin* gene and their effects on beta-lactoglobulin protein concentration in milk. Anim Genet. 2009;40:127–33.
37. Grisart B, Coppieters W, Farnir F, Karim L, Ford C, Berzi P, et al. Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine *DGAT1* gene with major effect on milk yield and composition. Genome Res. 2002;12:222–31.
38. Littlejohn MD, Tiplady K, Lopdell T, Law TA, Scott A, Harland C, et al. Expression variants of the lipogenic *AGPAT6* gene affect diverse milk composition phenotypes in *Bos taurus*. PLoS One. 2014;9:e85757.
39. Harold D, Abraham R, Hollingworth P, Sims R, Gerrish A, Hamshere ML, et al. Genome-wide association study identifies variants at *CLU* and *PICALM* associated with Alzheimer's disease. Nat Genet. 2009;41:1088–93.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit



4.4. Confirmation des mutations candidates

Short Communication: Confirmation of candidate causative variants on milk composition and cheese-making properties in Montbéliarde cows.

M. P. Sanchez[1*], M. Ferrand[†], M. Gelé[†], D. Pourchet[‡], G. Miranda[*], P. Martin[*], M. Brochard[†], D. Boichard[*]

[*] GABI, INRA, AgroParisTech, Université Paris Saclay, F-78350 Jouy-en-Josas, France

[†] Institut de l'Elevage, F-75012 Paris, France

[‡] ECEL Doubs - Territoire de Belfort, F-25640 Roulans, France

Journal of Dairy Science 2018. 101:10076–10081

https://doi.org/10.3168/jds.2018-14986

J. Dairy Sci. 101:10076–10081
https://doi.org/10.3168/jds.2018-14986
© American Dairy Science Association®, 2018.

Short communication: Confirmation of candidate causative variants on milk composition and cheesemaking properties in Montbéliarde cows

M. P. Sanchez,*[1] V. Wolf,† M. El Jabri,‡ E. Beuvier,§ O. Rolet-Répécaud,§ Y. Gaüzère,# S. Minéry,† M. Brochard,‖ A. Michenet,*¶ S. Taussat,*¶ A. Barbat-Leterrier,* A. Delacroix-Buchet,* C. Laithier,† S. Fritz,*¶ and D. Boichard*

*GABI, INRA, AgroParisTech, Université Paris Saclay, F-78350 Jouy-en-Josas, France
†Conseil Elevage 25-90, F-25640 Roulans, France
‡Institut de l'Elevage, F-75012 Paris, France
§URTAL, INRA, F-39800 Poligny, France
#Ecole Nationale d'Industrie Laitière et des Biotechnologies, F-39801 Poligny, France
‖Umotest, F-01250 Ceyzériat, France
¶Allice, F-75012 Paris, France

ABSTRACT

In a previous study, we identified candidate causative variants located in 24 functional candidate genes for milk protein and fatty acid composition in Montbéliarde, Normande, and Holstein cows. We designed these variants on the custom part of the EuroG10K BeadChip (Illumina Inc., San Diego, CA), which is routinely used for genomic selection analyses in French dairy cattle. To validate the effects of these candidate variants on milk composition and to estimate their effects on cheesemaking properties, a genome-wide association study was performed on milk protein, fatty acid and mineral composition, as well as on 9 cheesemaking traits (3 laboratory cheese yields, 5 coagulation traits, and milk pH). All the traits were predicted from midinfrared spectra in the Montbéliarde cow population of the Franche-Comté region. A total of 194 candidate variants located in 24 genes and 17 genomic regions were imputed on 19,862 cows with phenotypes and genotyped with either the BovineSNP50 (Illumina Inc.) or the EuroG10K BeadChip. We then tested the effect of each SNP in a mixed linear model including random polygenic effects estimated with a genomic relationship matrix. We confirm here the effects of candidate causative variants located in 17 functional candidate genes on both cheesemaking properties and milk composition traits. In each candidate gene, we identified the most plausible causative variant: 4 are missense in the *ALPL*, *SLC26A4*, *CSN3*, and *SCD* genes, 7 are located in 5′UTR (*AGPAT6*), 3′ untranslated region (*GPT*), or upstream (*CSN1S1*, *CSN1S2*, *PAEP*, *DGAT1*, and *PICALM*) regions, and 6 are located in introns of the *SLC37A1*, *MGST1*, *CSN2*, *BRI3BP*, *FASN*, and *ANKH* genes.

Key words: Montbéliarde, cheesemaking properties, milk composition, candidate variants

Short Communication

Cheesemaking properties (**CMP**) are strongly related to bovine milk composition (Wedholm et al., 2006). In the PhénoFinlait project, we predicted protein and fatty acid milk composition using mid-infrared (**MIR**) spectrometry in the French Holstein, Montbéliarde, and Normande dairy cattle breeds (Ferrand-Calmels et al., 2014; Sanchez et al., 2017a). Genome-wide association study (**GWAS**) on whole-genome sequences imputed using data of run 4 of the 1,000 bull genome project (Daetwyler et al., 2014) led to the identification of candidate causative mutations in 24 candidate genes (Boichard et al., 2014; Sanchez et al., 2017b). To validate these mutations in an independent population, we designed them on the custom part of the EuroG10K Beadchip (Illumina Inc., San Diego, CA). The independent population, consisting of Montbéliarde cows, originated from the FROM'MIR project that studied CMP and milk composition (proteins, fatty acids, and minerals) from MIR spectra. In the present study, we tested effects of candidate causative variants evidenced in the PhénoFinlait project on both CMP and milk composition traits predicted from MIR spectra in 19,862 FROM'MIR Montbéliarde cows.

A total of 416 milk samples (246 individual cows, 100 herds, and 70 dairy vats) were used to develop prediction equations for CMP. The sampling was designed to maximize genetic diversity as well as diversity in milk composition. All the samples, collected in Protected Designation of Origin and Protected Geographi-

Received April 27, 2018.
Accepted July 20, 2018.
[1] Corresponding author: marie-pierre.sanchez@inra.fr

cal Indication cheese area of Franche-Comté (eastern France), were aliquoted and analyzed within 24 h by MIR spectroscopy using MilkoScan FT6000 (Foss, Hillerød, Denmark) and by CMP reference methods for soft cheese (**SC**) and pressed cooked cheese (**PCC**) parameters. Three laboratory cheese yield, 13 milk rennet coagulation [Formoptic; adapted from a mechanical Formagraph (Foss Electric, Hillerod, Denmark) by Chr. Hansen (Horsholm, Denmark) and ENILBio (Poligny, France)], and 8 milk acidification by lactic acid bacteria parameters were measured. Equations of predictions were developed for all the 24 CMP traits using partial least squares regression. Accuracies of prediction (R^2) ranged from 0.04 to 0.89 according to traits (El Jabri et al., 2017; Laithier et al., 2017; Sanchez et al., 2018). Only 9 CMP traits with medium to high prediction accuracies ($0.54 \leq R^2 \leq 0.89$) were retained for genetic analyses (See definitions of the traits and their elementary statistics in Table 1). The 3 following laboratory curd yields (**CY**): CY_{FRESH} = 100 × (g of curd/g of milk), CY_{DM} = 100 × (1 − g of DM whey/g of DM milk), and $CY_{FAT\text{-}PROT}$ = (protein content + fat content) × (g of milk/g of curd) were measured in duplicate using the method of Hurtaud et al. (1993, 1995) that applies centrifugation (2,700 × *g*, 15 min, 32°C); laboratory CY are therefore expected to be higher than actual CY obtained in dairy industry. Five coagulation traits measured with a Formoptic device adapted from a mechanical Formagraph (Foss) by Chr. Hansen and ENILBio assessing curd firming and rennet coagulation time (**K10** and **RCT**, respectively) or curd firmness [at RCT and at 2 times RCT (**a** and **a2**, respectively)]; K10/RCT were measured for both SC and PCC, whereas a2 was measured for only the SC. Finally, we retained milk pH for PCC. Full details on reference measurements, MIR prediction equations as well as cheesemaking traits can be found in Sanchez et al. (2018).

Moreover, 15 milk composition traits (Table 1) were predicted from MIR spectra using equations of prediction previously developed in PhénoFinlait (Ferrand et al., 2012; Ferrand-Calmels et al., 2014; Sanchez et al., 2017a) and Optimir (Gengler et al., 2016) projects. These included 6 proteins (α-LA and β-LG whey proteins and α_{S1}-CN, α_{S2}-CN, β-CN and κ-CN), 4 fatty acids (total SFA, UFA, MUFA, and PUFA), and 5 min-

Table 1. Midinfrared predictions of cheesemaking properties and milk composition (n = 1,442,371), mean, SD, and CV, and performances of prediction equations, R^2 and root mean square error (RMSE), in a validation set

Trait		Mean	SD	CV	R^2_{val}	$RMSE_{val}$
Cheesemaking properties[1]						
CY_{FRESH}	100 × (g of curd/g of milk), %	38.2	7.2	18.9	0.86	2.78
CY_{DM}	100 × (g of DM of curd/g of DM of milk), %	67.1	4.8	7.2	0.89	1.67
$CY_{FAT\text{-}PROT}$	(g of milk fat + g of milk protein)/kg of curd, g/kg	186.1	21.8	11.7	0.54	14.50
a_{PCC}	Curd firmness at rennet coagulation time (RCT), firm index (FI)	18.8	2.5	13.3	0.72	1.44
$K10/RCT_{PCC}$	Curd organization index standardized for RCT	0.40	0.09	24.7	0.62	0.06
a_{SC}	Curd firmness at RCT, FI	19.1	2.7	13.9	0.73	1.52
$a2_{SC}$	Curd firmness at 2 × RCT, FI	23.0	2.1	9.1	0.64	1.37
$K10/RCT_{SC}$	Curd organization index standardized for RCT	0.36	0.10	27.9	0.62	0.071
pH_{0_PCC}	Started value of pH	6.5	0.07	1.1	0.65	0.035
Milk protein composition, g/100 g of protein						
α-LA	α-lactalbumin	4.0	0.35	8.8	0.59	0.22
β-LG	β-lactoglobulin	12.3	1.6	13.2	0.74	0.82
α_{S1}-CN	α_{S1}-casein	32.3	0.27	0.8	0.88	0.094
α_{S2}-CN	α_{S2}-casein	9.7	0.34	3.5	0.82	0.14
β-CN	β-casein	29.8	1.1	3.7	0.92	0.31
κ-CN	κ-casein	8.7	0.43	5.0	0.8	0.19
Milk fatty acid composition, g/100 g fat						
SFA	Saturated fatty acids	70.0	7.1	10.1	0.995	0.37
MUFA	Monounsaturated fatty acids	26.9	5.4	20.1	0.97	0.95
UFA	Unsaturated fatty acids	30.5	5.2	17.1	0.98	0.70
PUFA	Polyunsaturated fatty acids	3.3	1.0	30.3	0.76	0.044
Milk mineral composition, mg/kg of milk						
Na	Sodium	340	46.2	13.6	0.44	34.6
Ca	Calcium	1,162	95.8	8.3	0.82	40.6
P	Phosphorous	1,008	81.0	8.0	0.75	40.5
Mg	Magnesium	100	7.4	7.3	0.77	3.55
K	Potassium	1,476	106.9	7.2	0.68	60.5

[1]For pressed cooked cheese (PCC) and soft cheese (SC). CY = curd yield.

erals (Na, Ca, P, Mg, and K). Equations of prediction were applied on about 6 million MIR spectra collected from 330,000 Montbéliarde cows in the Franche-Comté region. Data from cows with at least 3 test-day records during the first lactation (1,442,371 test-day records from 189,817 cows) were adjusted for nongenetic effects using a mixed model. Herd × test-day × spectrometer and stage of lactation were included in this model as fixed effects, whereas animal genetic and permanent environment effects were assumed random. Data adjusted for fixed effects were then averaged per cow. A subset of 19,862 FROM'MIR cows were genotyped for the BovineSNP50 BeadChip (6,505 cows; Illumina Inc.) or for the customized low-density EuroG10K BeadChip (13,357 cows mainly for versions 1 to 5) for genomic selection purpose. All genotypes were imputed at 50K SNP density to the custom part SNP of version 7 of the EuroG10K BeadChip with FImpute software (Sargolzaei et al., 2014) using 177,736 cows genotyped for the BovineSNP50 or EuroG10K (versions 1 to 7) BeadChips. Mean squared correlations (R^2) between imputed and true genotypes reached 91.6% in a validation set for variants with minor allele frequency (**MAF**) ≥1%. Single-trait association analyses were performed between all the polymorphic variants with MAF ≥1% (45,120 SNP) and the 24 traits (9 CMP and 15 milk composition traits). A mixed linear model was applied with the GCTA software (Yang et al., 2011), including a mean, the additive fixed effect of the candidate variant, and the random polygenic effects of animals estimated with the genomic relationship matrix calculated from the genotypes of the BovineSNP50 BeadChip. The SNP effect was considered significant if its $-\log_{10}(P)$ value estimated assuming a Student distribution was higher than 6 (5% threshold after Bonferroni correction, 0.05/45,120).

Significant effects were found on at least 1 of the 24 traits analyzed for 162 of the 194 candidate variants (MAF ≥1%) previously selected in the PhénoFinlait project for milk protein or fatty acid composition (Boichard et al., 2014; Sanchez et al., 2017b). We confirmed effects on both CMP and milk composition traits (proteins, fatty acids, and minerals) for 13 of the 14 candidate regions. We found the well-known regions of caseins [BTA6], *PAEP* (BTA11), or *DGAT1* (BTA14) genes as well as other regions on BTA1, 2, 4, 5, 17, 19, 20, 26, 27, and 29. In these regions, candidate variants of the EuroG10K custom chip were systematically more significant than the SNP of the BovineSNP50 BeadChip. We targeted 24 genes (1 to 5 per region) found to be the best candidates in the PhénoFinlait project (Table 2). For each gene, 1 to 33 candidate variants were present in the custom part of the EuroG10K BeadChip (i.e., 245 in total). Among them, 162 with MAF higher than 1% had significant effects on at least 1 CMP or milk composition trait and 110 were ranked among the 10 most significant variants (top 10) for at least 1 of the traits analyzed. Effects of variants located in *BOP1* (7), *MROH1* (13), and *CYPB11* (1) genes on BTA14 were not tested because they had too low MAF (<0.01). Four other candidate genes could be excluded because all their polymorphisms had no significant effects (*ABCG2* and *DHX37*) or because significant variants were not located in the top 10 of the peak (*GPSM1* and *RECQL4*).

Seventeen candidate genes, each containing between 1 to 27 candidate variants, ranked in the top 10 of peaks, were kept for further investigation. Ranks of each variant were then examined in peaks for all traits to find the best candidate causative variant in each gene. In most of the genes, 1 variant, reported in Table 2, was ranked at the top of the peak for several traits analyzed in this study. Four of these variants were missense in *ALPL*, *SLC26A4*, *CSN3*, and *SCD* genes. In the *CSN3* gene, the candidate variant, located at 87,390,612 bp, is the missense variant responsible for the κ-CN A/B polymorphism. In the GWAS peaks, it was ranked first and second for 2 cheese yield traits ($CY_{FAT\text{-}PROT}$ and $K10/RCT_{PCC}$, respectively) and first to eighteenth for protein contents (Table 3). In *SCD* on BTA26, we observed a missense variant at 21,144,708 bp that was previously found with effects on milk fatty acid composition (Li et al., 2016). In our study, it was the variant with the most significant effects for CY_{FRESH}, CY_{DM}, and α_{S1}- and α_{S2}-CN contents. We found 7 other candidate variants located in 5′UTR (*AGPAT6*), 3′UTR (*GPT*), or upstream (*CSN1S1*, *CSN1S2*, *PAEP*, *DGAT1*, and *PICALM*) regions. Finally, 6 variants were located in introns of the *SLC37A1*, *MGST1*, *CSN2*, *BRI3BP*, *FASN*, and *ANKH* genes. Surprisingly, polymorphisms previously found as causal variants in *PAEP* (Ganai et al., 2009) and *DGAT1* (Grisart et al., 2002) genes were not the most significant for any traits in our study. In each of these genes, we identified an upstream variant that was the best candidate. The *K232A* mutation in the *DGAT1* gene was excluded from the final results because it had an MAF lower than 1% in the Montbéliarde cows (0.9%). Nevertheless, effects associated with this mutation presented $-\log(P)$ values higher than those of the candidate variant located in the upstream region of the *DGAT1* gene (e.g., 143 against 125 for fat content). Thus, we cannot exclude that the mutation described by Grisart et al. (2002) is the causative variant for the effects observed in our study. In contrast, despite a high MAF in Montbéliarde cows for the 2 causal variants identified by Ganai et

Table 2. Best candidate variants (VAR) in 24 candidate genes (R^2 = imputation accuracy estimated in a validation population)

BTA	Gene	Total VAR[1]	VAR MAF[2] >1%	Sign. VAR[3]	Top 10 VAR[4]	Best candidate VAR (bp)	Functional annotation	MAF	R^2
1	*SLC37A1*	14	13	12	10	144,398,764	Intronic	0.45	98.4
2	*ALPL*	8	8	7	7	131,812,821	Missense	0.38	95.8
4	*SLC26A4*	4	4	2	2	48,990,317	Missense	0.28	99.1
5	*MGST1*	26	22	19	7	93,945,738	Intronic	0.07	98.9
6	*ABCG2*	5	3	0	0	—	—	—	—
6	*CSN1S1*	11	9	9	4	87,141,456	Upstream	0.30	98.3
6	*CSN1S2*	10	7	7	4	87,261,372	Upstream	0.30	98.3
6	*CSN2*	14	13	10	8	87,187,426	Intronic	0.20	93.7
6	*CSN3*	21	18	17	10	87,390,612	Missense	0.40	95.6
11	*GPSM1*	3	3	3	0	—	—	—	—
11	*PAEP*	33	31	31	27	103,298,431	Upstream	0.45	1
14	*RECQL4*	3	3	3	0	—	—	—	—
14	*GPT*	2	2	2	2	1,623,927	3′UTR[5]	0.48	96.3
14	*DGAT1*	9	1	1	1	1,795,176	Upstream	0.48	97.7
14	*BOP1*	7	0	0	0	—	—	—	—
14	*MROH1*	13	0	0	0	—	—	—	—
14	*CYP11B1*	1	0	0	0	—	—	—	—
17	*BRI3BP*	12	12	9	9	53,072,959	Intronic	0.06	98.5
17	*DHX37*	6	6	0	0	—	—	—	—
19	*FASN*	10	10	8	5	51,386,735	Intronic	0.37	94.7
20	*ANKH*	20	17	11	3	58,427,343	Intronic	0.07	98.8
26	*SCD*	6	6	5	5	21,144,708	Missense	0.46	96.2
27	*AGPAT6*	4	4	4	4	36,212,352	5′UTR	0.50	96.9
29	*PICALM*	3	2	2	2	9,611,304	Upstream	0.22	95.5

[1]Number of total variants in the gene in the custom part of the EuroG10K BeadChip (Illumina Inc., San Diego, CA).

[2]Number of variants with minor allele frequency (MAF) $\geq$0.01.

[3]Number of variants with MAF $\geq$0.01 and –log(*P*-value) $\geq$6.

[4]Number of variants ranked in the top 10 for at least 1 trait.

[5]UTR = untranslated region.

al. (2009) in the *PAEP* gene (45% for both variants located at 103,303,475 and 103,304,757 bp), they were never ranked at the top of the peak. Moreover, on average for all the traits we analyzed, the ranks of the 2 *PAEP* causal variant mutations were respectively 27.6 and 17.6, whereas it was 9.7 for the candidate mutation located at 103,298,431 bp in the upstream region of the *PAEP* gene. This latter variant could be located in a regulatory region of the *PAEP* gene and could modulate its expression.

Analyses of both CMP and milk composition traits show that variants with significant effects on CMP were also significant on milk protein, fatty acids, or mineral composition. This result confirms the genetic links, previously described via genetic correlations, between CMP and milk composition and in particular with protein composition (Wedholm et al., 2006). Moreover, considering all results together can help to establish the functional link existing between these traits and candidate genes. For example, the best candidate variant in the *SLC37A1* gene, which encodes a glucose-6-phosphate transporter, had significant effects on 3 CMP traits [$8 \leq -\log(P) \leq 30$] and 5 milk composition traits [$13 \leq -\log(P) \leq 167$] with the most significant effects obtained on phosphorous. Similarly, the best candidate variant in the *ANKH* gene, encoding an inorganic pyrophosphate transport regulator that helps to prevent the deposition of Ca and P in bones, had significant effects on 5 CMP [$7 \leq -\log(P) \leq 46$] and 9 milk composition traits [$30 \leq -\log(P) \leq 175$], with the most significant effects found for α-LA that exhibit a high affinity to Ca. These 2 examples illustrate the interest to consider fine-scale phenotypes in complement to complex phenotypes, such as CMP.

We confirmed the effects of 13 genomic regions, previously identified on milk composition, on CMP and milk protein, fatty acids, and mineral composition predicted from MIR spectra in an independent population of 19,862 Montbéliarde cows. We showed that simultaneously analyzing fine-scale phenotypes and traits of interest can facilitate the identification of functional candidate genes. We reported candidate causative variants in 17 genes that have functional links with traits studied. To explore other genomic regions and to find other candidate variants, a GWAS will be performed on whole-genome sequence variants after imputations with the run 6 of the 1,000 bull genome project. Pleiotropic GWAS results will then be exploited to search for a set of interacting genes co-associated with CMP and milk composition.

Table 3. Ranks of the 17 best candidate variants in the significant peaks identified for each milk cheesemaking and composition trait defined in Table 1

BTA	Gene	Location of the best candidate variant (bp)	Cheesemaking properties								Protein composition						Fatty acid composition				Mineral composition				
			CY_{FRESH}	CY_{DM}	$CY_{FAT\text{-}PROT}$	K10/ RCT_{PCC}	a_{PCC}	K10/ RCT_{SC}	a_{SC}	$a2_{SC}$	pH_{IPCC}	α-LA	β-LG	α_{S1} CN	α_{S2} CN	β-CN	κ-CN	SFA	MUFA	UFA	PUFA	Na	Ca	P	Mg	K
1	*SLC37A1*	144,398,764			5			3			1			4								4		2	1	3
2	*ALPL*	131,812,821					2								5		4					2				
4	*SLC26A4*	48,990,317								1	1			1	1		1						1			
5	*MGST1*	93,945,738	4	3	5	18				22				5				1	1	1	1					
6	*CSN1S1*	87,141,456	32	26	47	37	51	42	26	39			19	28	5		26					9	8		40	46
6	*CSN1S2*	87,261,372	2	3	28	3	13	4	5	3			58	50	66	50	38								7	
6	*CSN2*	87,187,426	51	50	48	38	53	48	28	45			20	18	4		19						4		41	52
6	*CSN3*	87,390,612	31	28	1	11		2	50	13		5	2	1	18	3	2									
11	*PAEP*	103,298,431	1	1	7	5	21	5	3	2		1	12	23	2	18	2	10						9		7
14	*GPT*	1,623,927																7	5	5	6					7
14	*DGAT1*	1,795,176	76	55	46	77		77		77		85	45	34			68	1	1	1	1		68	77		11
17	*BRI3BP*	53,072,959				2		2		2					3	4	1									1
19	*FASN*	51,386,735																1	1	1						
20	*ANKH*	58,427,343	4		2		1		1		1	1		1	2		1						1		1	2
26	*SCD*	21,144,708	1	1										1	1			12	11	12	15					
27	*AGPAT6*	36,212,352	1	1	4											1		1	1	1	2				4	
29	*PICALM*	9,611,304												2	2							2	2			

ACKNOWLEDGMENTS

This study was funded by the French Ministry of Agriculture, Agro-food and Forest, the French Dairy Interbranch Organization (CNIEL; Paris, France), the Regional Union of Protected Designation cheeses of Franche-Comté (URFAC; Paris, France) and the Regional Council of Bourgogne- Franche-Comté (Poligny, France), under the project FROM'MIR (Besançon, France).

REFERENCES

Boichard, D., A. Govignon-Gion, H. Larroque, C. Maroteau, I. Palhiere, G. Tosser-Klop, R. Rupp, M. P. Sanchez, and M. Brochard. 2014. Genetic determinism of milk composition in fatty acids and proteins in ruminants, and selection potential. INRA Prod. Anim. 27:283–298.

Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. Van Binsbergen, R. F. Brøndum, X. Liao, A. Djari, S. Rodriguez, C. Grohs, D. Esquerré, O. Bouchez, M. N. Rossignol, C. Klopp, D. Rocha, S. Fritz, A. Eggen, P. J. Bowman, D. Coote, A. J. Chamberlain, C. P. VanTassell, I. Hulsegge, M. E. Goddard, B. Guldbrandtsen, M. S. Lund, R. F. Veerkamp, D. Boichard, R. Fries, and B. J. Hayes. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat. Genet. 46:858–865. https://doi.org/10.1038/ng.3034.

El Jabri, M., M. P. Sanchez, C. Laithier, E. Doutart, V. Wolf, D. Pourchet, P. Grosperrin, E. Beuvier, O. Rolet-Répécaud, Y. Gaüzère, O. Belysheva, A. Delacroix-Buchet, and D. Boichard. 2017. Bayesian regression models and variable selection methods before PLS regression. Application to the prediction of milk cheese-making properties using infared spectral data. Pages 15–17 in Chimiometrie XVIII. Paris, France. AgroParisTech, Paris, France.

Ferrand, M., G. Miranda, S. Guisnel, H. Larroque, O. Leray, F. Lahalle, M. Brochard, and P. Martin. 2012. Determination of protein composition in milk by mid-infrared spectrometry. Pages 41–45 in Proc. International Strategies and New Developments in Milk Analysis. VI ICAR Reference Laboratory Network Meeting, Cork, Ireland. ICAR, Rome, Italy.

Ferrand-Calmels, M., I. Palhiere, M. Brochard, O. Leray, J. Astruc, M. Aurel, S. Barbey, F. Bouvier, P. Brunschwig, H. Caillatt, M. Douguet, F. Faucon-Lahalle, M. Gele, G. Thomas, J. Trommenschlager, and H. Larroque. 2014. Prediction of fatty acid profiles in cow, ewe, and goat milk by mid-infrared spectrometry. J. Dairy Sci. 97:17–35. https://doi.org/10.3168/jds.2013-6648.

Ganai, N. A., H. Bovenhuis, J. A. van Arendonk, and M. H. Visker. 2009. Novel polymorphisms in the bovine beta-lactoglobulin gene and their effects on beta-lactoglobulin protein concentration in milk. Anim. Genet. 40:127–133. https://doi.org/10.1111/j.1365-2052.2008.01806.x.

Gengler, N., H. Soyeurt, F. Dehareng, C. Bastin, F. Colinet, H. Hammami, M. Vanrobays, A. Laine, S. Vanderick, C. Grelet, A. Vanlierde, E. Froidmont, and P. Dardenne. 2016. Capitalizing on fine milk composition for breeding and management of dairy cows. J. Dairy Sci. 99:4071–4079. https://doi.org/10.3168/jds.2015-10140.

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and R. Snell. 2002. Positional candidate cloning of a QTL in dairy cattle: Identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res. 12:222–231. https://doi.org/10.1101/gr.224202.

Hurtaud, C., H. Rulquin, M. Delaite, and R. Verite. 1995. Prediction of cheese yielding efficiency of individual milk of dairy cows—Correlation with coagulation parameters and laboratory curd yield. Ann. Zootech. 44:385–398.

Hurtaud, C., H. Rulquin, and R. Verite. 1993. Effect of infused volatile fatty acids and caseinate on milk composition and coagulation in

dairy cows. J. Dairy Sci. 76:3011–3020. https://doi.org/10.3168/jds.S0022-0302(93)77640-7.

Laithier, C., V. Wolf, M. El Jabri, P. Trossat, S. Gavoye, D. Pourchet, P. Grosperrin, E. Beuvier, O. Rolet-Répécaud, Y. Gauzère, O. Belysheva, E. Notz, and A. Delacroix-Buchet. 2017. Prediction of cheesemaking properties of Montbeliarde milks used for PDO/PGI cheeses production in Franche-Comté by mid-infrared spectrometry. Pages 15–19 in 12th International Meeting on Mountain Cheese, June 20–22, 2017. Padova, Italy. Padova University Press, Padova, Italy.

Li, C., D. Sun, S. Zhang, L. Liu, M. Alim, and Q. Zhang. 2016. A post-GWAS confirming the SCD gene associated with milk medium- and long-chain unsaturated fatty acids in Chinese Holstein population. Anim. Genet. 47:483–490. https://doi.org/10.1111/age.12432.

Sanchez, M. P., M. El Jabri, S. Minéry, V. Wolf, E. Beuvier, C. Laithier, A. Delacroix-Buchet, M. Brochard, and D. Boichard. 2018. Genetic parameters for cheese-making properties and milk composition predicted from mid-infrared spectra in a large dataset of Montbéliarde cows. J. Dairy Sci. https://doi.org/10.3168/jds.2018-14878.

Sanchez, M. P., M. Ferrand, M. Gele, D. Pourchet, G. Miranda, P. Martin, M. Brochard, and D. Boichard. 2017a. Short communication: Genetic parameters for milk protein composition predicted using mid-infrared spectroscopy in the French Montbéliarde, Normande, and Holstein dairy cattle breeds. J. Dairy Sci. 100:6371–6375. https://doi.org/10.3168/jds.2017-12663.

Sanchez, M. P., A. Govignon-Gion, P. Croiseau, S. Fritz, C. Hozé, G. Miranda, P. Martin, A. Barbat-Leterrier, R. Letaïef, D. Rocha, M. Brochard, M. Boussaha, and D. Boichard. 2017b. Within-breed and multi-breed GWAS on imputed whole-genome sequence variants reveal candidate mutations affecting milk protein composition in dairy cattle. Genet. Sel. Evol. 49:68. https://doi.org/10.1186/s12711-017-0344-z.

Sargolzaei, M., J. Chesnais, and F. Schenkel. 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15:478. https://doi.org/10.1186/1471-2164-15-478.

Wedholm, A., L. B. Larsen, H. Lindmark-Månsson, A. H. Karlsson, and A. Andrén. 2006. Effect of protein composition on the cheese-making properties of milk from individual dairy cows. J. Dairy Sci. 89:3296–3305. https://doi.org/10.3168/jds.S0022-0302(06)72366-9.

Yang, J., S. Lee, M. Goddard, and P. Visscher. 2011. GCTA: A Tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 88:76–82. https://doi.org/10.1016/j.ajhg.2010.11.011.

4.5. GWAS et réseaux de gènes pour la composition et la fromageabilité du lait

Sequence-based GWAS, network and pathway analyses reveal genes co-regulated and co-associated with milk cheese-making properties and milk composition in Montbéliarde cows.

Marie-Pierre Sanchez[1*], Yuliaxis Ramayo-Caldas[1], Valérie Wolf[2], Cécile Laithier[3], Mohammed El Jabri[3], Alexis Michenet[1,4], Mekki Boussaha[1], Sébastien Taussat[1,4], Sébastien Fritz[1,4], Agnès Delacroix-Buchet[1], Mickaël Brochard[5], Didier Boichard[1]

[1] GABI, INRA, AgroParisTech, Université Paris Saclay, F-78350 Jouy-en-Josas, France

[2] Conseil Elevage 25-90, F-25640 Roulans, France

[3] Institut de l'Elevage, F-75012 Paris, France

[4]Allice, F-75012 Paris, France

[5] Umotest, F-01250 Ceyzériat, France

Genetics Selection Evolution 2019. 51:34.

https://doi.org/10.1186/s12711-019-0473-7

Sanchez *et al. Genet Sel Evol* (2019) 51:34
https://doi.org/10.1186/s12711-019-0473-7

GSE Genetics Selection Evolution

RESEARCH ARTICLE Open Access

Sequence-based GWAS, network and pathway analyses reveal genes co-associated with milk cheese-making properties and milk composition in Montbéliarde cows

Marie-Pierre Sanchez[1*], Yuliaxis Ramayo-Caldas[1], Valérie Wolf[2], Cécile Laithier[3], Mohammed El Jabri[3], Alexis Michenet[1,4], Mekki Boussaha[1], Sébastien Taussat[1,4], Sébastien Fritz[1,4], Agnès Delacroix-Buchet[1], Mickaël Brochard[5] and Didier Boichard[1]

Abstract

Background: Milk quality in dairy cattle is routinely assessed via analysis of mid-infrared (MIR) spectra; this approach can also be used to predict the milk's cheese-making properties (CMP) and composition. When this method of high-throughput phenotyping is combined with efficient imputations of whole-genome sequence data from cows' genotyping data, it provides a unique and powerful framework with which to carry out genomic analyses. The goal of this study was to use this approach to identify genes and gene networks associated with milk CMP and composition in the Montbéliarde breed.

Results: Milk cheese yields, coagulation traits, milk pH and contents of proteins, fatty acids, minerals, citrate, and lactose were predicted from MIR spectra. Thirty-six phenotypes from primiparous Montbéliarde cows (1,442,371 test-day records from 189,817 cows) were adjusted for non-genetic effects and averaged per cow. 50 K genotypes, which were available for a subset of 19,586 cows, were imputed at the sequence level using Run6 of the 1000 Bull Genomes Project (comprising 2333 animals). The individual effects of 8.5 million variants were evaluated in a genome-wide association study (GWAS) which led to the detection of 59 QTL regions, most of which had highly significant effects on CMP and milk composition. The results of the GWAS were further subjected to an association weight matrix and the partial correlation and information theory approach and we identified a set of 736 co-associated genes. Among these, the well-known caseins, *PAEP* and *DGAT1*, together with dozens of other genes such as *SLC37A1*, *ALPL*, *MGST1*, *SEL1L3*, *GPT*, *BRI3BP*, *SCD*, *GPAT4*, *FASN*, and *ANKH*, explained from 12 to 30% of the phenotypic variance of CMP traits. We were further able to identify metabolic pathways (e.g., phosphate and phospholipid metabolism and inorganic anion transport) and key regulator genes, such as *PPARA*, *ASXL3,* and *bta-mir-200c* that are functionally linked to milk composition.

Conclusions: By using an approach that integrated GWAS with network and pathway analyses at the whole-genome sequence level, we propose candidate variants that explain a substantial proportion of the phenotypic variance of CMP traits and could thus be included in genomic evaluation models to improve milk CMP in Montbéliarde cows.

*Correspondence: marie-pierre.sanchez@inra.fr
[1] GABI, INRA, AgroParisTech, Université Paris Saclay, 78350 Jouy-en-Josas, France
Full list of author information is available at the end of the article

BMC

© The Author(s) 2019. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Background

About 40% of the bovine milk produced worldwide is processed into cheese; because of this, the cheese-making properties (CMP) of bovine milk are economically important for the dairy industry. Direct measurement of CMP is costly and time-consuming, and cannot be obtained on a very large scale. However, mid-infrared (MIR) spectrometry, which is already widely employed to predict milk composition, has been shown to provide indirect measures of CMP that are sufficiently reliable to be used in genetic analyses [1]. Indeed, because of their strong dependence on milk composition traits [2], milk CMP, especially cheese yields and coagulation properties, can be routinely assessed at low cost from MIR spectra [3]. The information obtained from high-throughput MIR spectra can then be combined with genotypic data from cows that are generated for the purpose of genomic selection to provide a unique resource for large-scale genomic analyses of CMP aimed at identifying the genes involved in the genetic determinism of these traits.

Genomic regions containing quantitative trait loci (QTL) that affect traits of interest, such as CMP, can be identified by genome-wide association studies (GWAS). By combining the results of genotyping for genomic selection with reference data from the 1000 Bull Genomes Project, it becomes possible to carry out GWAS on imputed whole-genome sequences (WGS) that should contain the causative mutations for traits of interest [4]. However, even if these analyses are carried out at the sequence level, GWAS alone is generally not sufficient to identify causative genes, let alone causative variants for complex and polygenic traits. Indeed, due to the long-range linkage disequilibrium (LD) in dairy cattle, many variants with almost identical P-values that are potentially located in more than one gene or in intergenic regions are generally found in a QTL region, which complicates identification of the causative mutations. Moreover, complex traits are typically influenced by many genomic regions, most of which explain only a small proportion of the phenotypic variance and are thus difficult to detect by GWAS. Finally, GWAS performed on a single trait and single marker cannot take either the pleiotropic effects of variants or the interactions between them into account. Thus, a GWAS-based approach is a good starting point for identifying QTL regions but needs to be supplemented by additional analyses to capture a larger proportion of the genetic variance and to understand in depth the genetic architecture of complex traits, such as CMP. In the last decade, methods have been developed that build on GWAS results by using gene network analysis to highlight co-associated genes for a set of correlated traits [5, 6]. Once the gene network is built, it is then possible to carry out in silico functional analyses, based on databases from bovine or other organisms' genomes, to identify key regulators that modulate gene expression or to highlight the enrichment of gene-sets linked to certain metabolic pathways. Gene network approaches have been applied to milk CMP [7], fatty acid composition [8, 9], and protein composition [10, 11] but, to date, there has been no joint analysis of CMP and milk composition in spite of the close relationship between the two groups of traits. Moreover, all previous studies examined only a limited number of cows (164 to 1100 cows) and genomic variants (50 K or HD SNP chips).

The goal of the FROM'MIR project is to analyze CMP and milk composition traits predicted from MIR spectra in the Montbéliarde dairy breed from the Franche-Comté region, which boasts the highest production of protected designation of origin (PDO) cheeses in France. Nine CMP traits (three measures of cheese yield, five coagulation traits, and one acidification trait) and 27 milk composition traits (protein, fatty acid, mineral, citrate, and lactose contents) were predicted with a relatively high degree of accuracy from more than 6.6 million MIR spectra of milk samples collected from 410,622 cows. Of these cows, 19,586 were genotyped with a SNP chip. A prior study revealed medium-to-high heritabilities for CMP traits as well as high genetic correlations among CMP traits and between CMP and some milk composition traits [3].

The objectives of the current study were first, to fine-map QTL for CMP and milk composition traits via GWAS of imputed WGS from 19,586 cows, and second, to explore the GWAS results using association weight matrices (AWM) [5] and partial correlation and information theory (PCIT) [6] analyses, in order to identify gene networks and metabolic and regulatory pathways that are associated with milk cheese-making and composition traits.

Methods

Animals, MIR spectra, and 50 K genotypes

For this study, we did not perform any experiments on animals; thus, no ethical approval was required. Details of the animals, milk analyses, and prediction equations were described in a prior study by Sanchez et al. [3]. Briefly, prediction equations were developed for nine CMP traits from 416 milk samples for which both reference measurements for those CMP traits and MIR spectra were taken. The CMP traits, described in Table 1, included three laboratory cheese yields (CY_{FRESH}, CY_{DM}, and $CY_{FAT\text{-}PROT}$), five coagulation traits for pressed cooked cheese (PCC) and soft cheese (SC) ($K10/RCT_{PCC}$, $K10/RCT_{SC}$, a_{PCC}, a_{SC}, and $a2_{SC}$), and milk pH after adding starter for PCC (pH_{0_PCC}). The accuracies of MIR predictions, assessed by

Table 1 Means, standard deviations (SD) for cheese-making properties and milk composition traits in the genotyped population (N = 19,586), and accuracy of MIR predictions equations (R^2_{val})

Trait	Description and unit	Mean	SD	R^2_{val}
Cheese-making properties[a]				
CY_{FRESH}	100 × (g curd/g milk), in %	37.7	4.95	0.82
CY_{DM}	100 × (g DM curd/g DM milk), in %	66.8	3.31	0.89
$CY_{FAT\text{-}PROT}$	(g milk fat + g milk protein)/kg curd, in g kg^{-1}	189.7	14.3	0.54
a_{PCC}	Curd firmness at rennet coagulation time (RCT), in firm index (FI)	18.8	1.72	0.76
$K10/RCT_{PCC}$	Curd organization index standardized for RCT	0.37	0.06	0.68
a_{SC}	Curd firmness at RCT, in FI	18.9	1.80	0.76
$a2_{SC}$	Curd firmness at 2 times RCT, in FI	22.8	1.41	0.69
$K10/RCT_{SC}$	Curd organization index standardized for RCT	0.37	0.07	0.72
pH_{0_PCC}	Initial value of pH	6.52	0.04	0.62
Protein composition				
PC	Protein content, in g/100 g milk	3.36	0.20	1.00
α-LA	α-lactalbumin, in g/100 g protein	4.01	0.20	0.59
β-LG	β-lactoglobulin, in g/100 g protein	12.4	1.09	0.74
αs1-CN	αs1-casein, in g/100 g protein	32.2	0.18	0.88
αs2-CN	αs2-casein, in g/100 g protein	9.73	0.19	0.82
β-CN	β-casein, in g/100 g protein	29.7	0.68	0.92
κ-CN	κ-casein, in g/100 g protein	8.74	0.24	0.80
ΣCN	Total caseins, in g/100 g protein	80.8	0.74	0.98
ΣWP	Total whey proteins, in g/100 g protein	16.9	1.15	0.54
Fatty acid composition				
FC	Fat content, in g/100 g milk	3.73	0.32	1.00
SFA	Saturated fatty acids, in g/100 g fat	70.6	3.05	1.00
MUFA	Mono-unsaturated fatty acids, in g/100 g fat	26.5	2.68	0.97
UFA	Unsaturated fatty acids, in g/100 g fat	30.0	2.93	0.98
PUFA	Poly-unsaturated fatty acids, in g/100 g fat	3.33	0.39	0.76
Σ C4-C10	Sum of C4 to C10 fatty acids, in g/100 g fat	11.6	0.71	0.95
Σ C4-C12	Sum of C4 to C12 fatty acids, in g/100 g fat	14.2	0.93	0.95
C14:0	Myristic acid, in g/100 g fat	11.1	1.05	0.94
C16:0	Palmitic acid, in g/100 g fat	28.8	2.53	0.94
C18:0	Stearic acid, in g/100 g fat	10.5	1.42	0.84
C18:1	Oleic acid, in g/100 g fat	23.2	2.59	0.96
Minerals				
Ca	Calcium, in mg/kg milk	1165	69.6	0.82
P	Phosphorous, in mg/kg milk	1014	62.5	0.75
Mg	Magnesium, in mg/kg milk	100.9	5.5	0.77
K	Potassium, in mg/kg milk	1496	69.3	0.68
Na	Sodium, in mg/kg milk	338.3	29.1	0.44
Other compounds				
Lactose	Lactose, in g/kg milk	49.3	1.4	0.92
Citrate	Citrate, in g/kg milk	0.83	0.11	0.90

[a] For pressed cooked cheese (PCC) and soft cheese (SC)

the coefficient of determination (R^2), varied between 0.54 and 0.89 depending on the CMP trait (Table 1). Milk composition was also predicted using equations that were developed in previous projects ($0.44 < R^2 < 1$; Table 1). Milk proteins and fatty acids were predicted with equations that were developed in the PhénoFinlait project [12–14], whereas for minerals and citrate content we used equations that were generated by the

Optimir project [15]. Lactose was predicted by a Foss equation.

Prediction equations were applied to the original dataset, which comprised 6,670,769 milk samples originating from 410,622 Montbéliarde cows. Data from cows with at least three test-day records during the first lactation (1,442,371 test-day records from 189,817 cows) were adjusted for non-genetic effects in a mixed model with the Genekit software [16]. Herd × test-day × spectrometer, age at calving, and stage of lactation were included in this model as fixed effects, while animal genetic and permanent environmental effects were assumed to be random. Test-day data adjusted for fixed effects were then averaged over a lactation for each cow. A subset of 19,586 cows for which MIR spectra were available had also been genotyped for the purpose of genomic selection by using the BovineSNP50 (50 K, 6505 cows) or the EuroG10 K BeadChip (Illumina Inc., San Diego, 13,081 cows). Means and standard deviations of the traits for this subset are in Table 1. Using FImpute software [17], all genotypes were imputed to the 50 K-SNP level. A total of 43,801 autosomal SNPs were retained after quality control filters were applied. These filters were taken directly from the French national evaluation system [18]: individual call rate higher than 95%, SNP call rate higher than 90%, minor allele frequency (MAF) higher than 1% in at least one major French dairy cattle breed, and genotype frequencies in Hardy–Weinberg equilibrium with $P > 10^{-4}$.

Imputation to whole-genome sequences

The 50 K SNP genotypes of the 19,586 cows were then imputed to whole-genome sequences (WGS). A two-step approach was applied in order to improve the accuracy of imputed genotypes of the WGS variants [19]: from 50 to 777 K high-density (HD) SNPs using FImpute software [17], and then, from imputed HD SNPs to WGS, using Minimac software [20]. In spite of a longer computing time, Minimac was preferred over FImpute to impute on WGS because it infers allele dosages in addition to the best-guess genotypes. Compared to the best-guess genotypes, allele dosages are expected to be more correlated to true genotypes [21] and to lead to a better targeting of causative mutations in GWAS analyses [22]. Imputations from 50 K to the HD SNP level were performed using a within-breed reference set of 522 Montbéliard bulls that were genotyped with the Illumina BovineHD BeadChip (Illumina Inc., San Diego, CA) [23]. WGS variants were imputed from HD SNP genotypes using WGS variants of 2333 *Bos taurus* animals, from the 6th run of the 1000 Bull Genomes Project [21, 24]. These animals represent 51 cattle breeds and include 54 Montbéliard individuals, most of them being major ancestor bulls with a high cumulated contribution to the breed (80.6%). We applied the protocol defined by the "1000 Bull Genomes" consortium [4, 25]: (1) short reads were filtered for quality and aligned to the UMD3.1 reference sequence [4, 26], and small genomic variations (SNPs and indels) were detected using SAMtools 0.0.18 [27]; (2) raw variants were filtered to produce 26,738,438 autosomal variants as described in Boussaha et al. [26]; and (3) filtered variants were annotated with the Ensembl variant effect predictor (VEP) pipeline v81 [28] and effects of amino-acid changes were predicted using the SIFT tool [29].

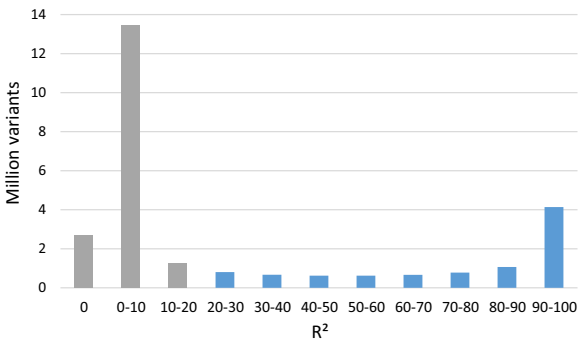


Fig. 1 Distribution of imputation accuracies (coefficient of determination, R^2) for 26.7 million sequence variants imputed with Minimac. Variants in blue, with $R^2 > 0.20$, were retained for GWAS analyses

The precision of imputation from HD SNP to sequence was assessed using the coefficient of determination (R^2) calculated with Minimac software [20]. In order to remove variants with low imputation accuracies, only variants with an R^2 higher than 20% and a MAF higher than 1% were retained for further association analyses, i.e. 8,551,748 variants with a mean R^2 of 76% (Fig. 1).

Whole-genome sequence association analyses

We performed single-trait association analyses between all 8,551,748 variants and the 36 CMP and milk composition traits described in Table 1. All association analyses were performed using the *mlma* option of the GCTA software (version 1.24), which applies a mixed linear model that includes the variant to be tested [30]:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{x}b + \mathbf{u} + \mathbf{e}, \quad (1)$$

where $\mathbf{y}$ is the vector of pre-adjusted phenotypes, averaged per cow; μ is the overall mean; b is the additive fixed effect of the variant to be tested for association; $\mathbf{x}$ is the vector of predicted allele dosages, varying between 0 and 2; $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G}\sigma_u^2)$ is the vector of random polygenic effects, with $\mathbf{G}$ the genomic relationship matrix (GRM), calculated using the HD SNP genotypes [31], and σ_u^2 is the polygenic variance, estimated based on the null model ($\mathbf{y} = \mathbf{1}\mu + \mathbf{u} + \mathbf{e}$) and then fixed while testing for the association between each variant and the trait of interest;

and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma^2_{\mathbf{e}})$ is the vector of random residual effects, with $\mathbf{I}$ the identity matrix and σ^2_e the residual variance.

The Bonferroni correction was applied to the thresholds in order to account for multiple testing. We used a very stringent correction, which considered all 8.5 million tests as independent. Therefore, the 5% genome-wide threshold of significance corresponded to a nominal P value of 5.8×10^{-9} ($-\log_{10}(P) = 8.2$). When a given trait was significantly affected by multiple variants, the variants that were located less than 1 Mbp apart were grouped in the same QTL region. The bounds of QTL regions were then determined by considering the positions of variants that were included in the upper third of the peak. For each trait, the percentage of phenotypic variance explained by each QTL was calculated as follows: $\%\sigma^2_P = 100\left(\frac{2p(1-p)\alpha^2}{\sigma^2_P}\right)$, with σ^2_P the phenotypic variance of the trait, and p and α are the frequency and the estimated allelic substitution effect, respectively, of the variant with the most significant effect in the QTL region.

Co-associated gene network analysis

Co-associated genes were detected from the GWAS results using the AWM approach [5, 6]. We first constructed two $n \times m$ matrices with variants row-wise ($n = 8{,}551{,}748$) and traits column-wise ($m = 36$). The first matrix contained variants' z-score standardized additive effects, whereas the second one contained the P-values associated with those effects. Among the CMP traits, CY_{DM} was selected as the key phenotype because it has the highest economic importance to the cheese-making process. The AWM was constructed following the procedure described in Ramayo-Caldas et al. [32]. SNPs were included in the analysis if their P-value for CY_{DM} was less than or equal to 0.001. Due to the large number of traits analyzed, we calculated correlation coefficients between SNP additive effects for different traits and then selected the set of traits correlated with CY_{DM} ($|r| \geq 0.25$). Next, we explored the dependency among traits and we estimated that on average, six other phenotypes were associated with these SNPs at the same P-value ($P \leq 0.001$). Other variants with significant effects on at least six traits were finally included in the analysis. Based on VEP annotation [33], we then selected only the SNPs that were located within or close to (within 10 kb of) genes. Among these, we retained only one variant per gene, i.e. the SNP that was associated with the largest number of traits or, in case of a tie, the variant for which the sum of P-values for the traits was the lowest.

Subsequently, to identify significant gene–gene interactions, partial correlations were computed using the PCIT algorithm developed by Reverter and Chan [34]; the algorithm was implemented in an R package designed for this purpose [35]. We visualized the gene network with Cytoscape 3.6.1 [36], with each node representing a gene and each edge representing a significant interaction. The centrality parameters of each node were assessed using the CentiScaPe 2.2 plug-in for Cytoscape [37]. For each node, we calculated the number of adjacent genes (degree parameter) and the relative node contribution (eigenvector parameter), with the latter value being higher (or lower) if the gene was connected to highly (or poorly) connected genes.

Identification of key regulators

Potential key regulators of the gene network were identified using two approaches. First, we used the iRegulon 1.3 plug-in for Cytoscape [38] to identify transcription factors (TF) in silico; this method was based on human datasets but included orthologous regions of ten other vertebrate genomes, including *Bos taurus*. Two types of data were used to identify regulatory regions that were shared by the genes identified in the network: (1) TF binding site motifs in the cis-regulatory regions, and (2) thousands of ChIP-Seq (chromatin immunoprecipitation followed by high-throughput sequencing) datasets from the ENCODE project [39] corresponding to targets of known TF. More details are in Janky et al. [38]. We then applied an information loss-less approach [6] that explored the connectivity of all regulators in the network, including TF, miRNA, and lnRNA. As recommended by Reverter and Fortes [6], we tested trios of TF genes to find the minimal set of TF genes with maximal coverage of the network.

Gene-set enrichment analysis

Next, we searched in the gene network for enrichment in gene ontology (GO) terms and pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG), using the ClueGO 2.5.1 plug-in for Cytoscape [40]. In order to avoid selecting GO terms that were too general (too many genes) or too specific (too few genes), we selected the 4th to 8th levels of the GO hierarchy. A gene set was considered to be enriched if the P-value associated with the hypergeometric test was lower than 0.05, after application of the Benjamini–Hochberg correction for multiple testing. GO terms and KEGG pathways were subsequently clustered in functional groups if the kappa statistic was higher than 0.4.

Results

GWAS analyses

GWAS that was carried out on 8,551,748 imputed WGS variants for the 36 CMP and milk composition traits revealed 236,332 significant variant × trait combinations ($-\log_{10}(P) > 8.2$), corresponding to 79,803 different

variants. Due to the high maximal $-\log_{10}(P)$ value for a large number of genomic regions (up to 560 for one of the QTL detected on chromosome 11), the number of variants with significant effects ($-\log_{10}(P) > 8.2$) was sometimes very large in a given region. Thus, to best target candidate variants, we selected only the variants that were located in the upper third of the peaks, as described in the Methods section. In doing so, we defined 59 QTL regions, which contained 6757 distinct variants (Table 2). In each of the QTL regions, we designated "candidate variants" as the variants that were located within the confidence intervals of the QTL and the "best candidate variant" (described in Table 2) as the variant within a gene (or its upstream/downstream regions) with the most significant effects.

These QTL regions varied in size (from 9.2 kbp to 8.9 Mbp) and contained from 6 to 401 variants; they were distributed on all *Bos taurus* autosomes (BTA) with the exception of BTA8 and 23 (Fig. 2 and [see Additional file 1: Figure S1]). In almost all the QTL regions (56), we identified variants that were located in one or more candidate genes. Around 60% (i.e. 4312 of 7393) of the variants detected in the QTL regions were located within or in the upstream/downstream region of 264 genes [see Additional file 2: Table S1]. Fifty-one of these variants were predicted to be responsible for an amino-acid change in the protein, whereas most of them (2972) were located in introns (Table 3).

We found the most significant effects around 103.3 Mbp on BTA11 ($-\log_{10}(P) = 560$), 144.4 Mbp on BTA1 ($-\log_{10}(P) = 210$), 58.4 Mbp on BTA20 ($-\log_{10}(P) = 177$), 1.6 Mbp on BTA14 ($-\log_{10}(P) = 123$), and 46.9 Mbp on BTA6 ($-\log_{10}(P) = 120$). In each of these five QTL, we identified variants that were located in candidate genes, which were, respectively, *PAEP*, *SLC37A1*, *ANKH*, *GPT*, and *SEL1L3*. All the variants were located in introns of the genes, with the exception of the best candidate variant of the *GPT* gene, which was found in the upstream region. Four other QTL had more moderate but nevertheless strong effects ($-\log_{10}(P)$ between 60 and 83), on BTA5 (118 Mbp), BTA6 (87.4 Mbp), BTA17 (53.1 Mbp), and BTA27 (36.2 Mbp), with the best candidate variants located in *GRAMD4* (upstream region), *CSN3* (downstream region), *BRI3BP* (upstream region), and *GPAT4* (3'UTR region), respectively. We also found candidate variants ($-\log_{10}(P)$ between 25 and 50) in 11 other candidate genes, on BTA2 (*ALPL*), BTA4 (*CBLL1*), BTA5 (*MGST1*), BTA7 (*FSTL4*), BTA12 (*ABCC4*), BTA19 (*FASN*), BTA22 (*FAM19A4* and *KLF15*), BTA25 (*FAM57B*), BTA26 (*SCD*), and BTA29 (*EED*). Finally, many other variants were identified in various genomic regions that had more moderate but significant effects after application of the Bonferroni correction ($-\log_{10}(P) > 8.2$); most of these were located in genes. All the QTL regions are described in detail in Table 2.

On average, each QTL had significant effects on about six traits. Only 13 QTL affected a single trait, while the other 46 QTL had pleiotropic effects on two to 26 traits. The QTL that affected the largest number of traits was located at about 1.6 Mbp on BTA14. For most traits, including FC, the variant with the strongest effect was not the well-known K232A polymorphism in the *DGAT1* gene [see Additional file 3: Table S2]. More than half of the QTL (33), and in particular those with the most significant effects, had effects on CMP traits. Almost all of the QTL with significant effects on CMP traits presented significant pleiotropic effects on milk composition traits, as well. In contrast, the remaining 26 QTL affected milk composition (protein, fatty acid, mineral, citrate, or lactose content) but not CMP. Among traits, we observed large differences in both the number of QTL detected (ranging from 6 to 19) and in the total percentage of phenotypic variance (ranging from 4.7 to 62.4%) that was explained by the detected QTLs, and simply estimated by the sum of percentages per QTL. Overall, the larger the number of detected QTL for a given trait, the lower the percentage of phenotypic variance that was explained by each. For example, in our study, the most polygenic trait, a_{SC}, was influenced by 19 QTL, each of which explained only 0.2 to 1.9% of the phenotypic variance. In contrast, we detected only six QTL for ΣWP but the QTL with the most important effect explained 56% of the phenotypic variance of this trait. As expected, the most heritable traits were those that presented the highest values of the total phenotypic variance explained by the QTL. The trait for which the largest amount of total phenotypic variance was explained by the QTL was β-LG (62%), which was also the most heritable trait analyzed in our study. For CMP traits, which are moderately heritable, from 12% (curd firmness) to 30% (curd firming time) of the phenotypic variability was explained by the QTL (i.e. from 27 to 65% of the genetic variance). Cheese yields presented intermediate results, as the detected QTL captured about 20% of their phenotypic variance, i.e. about 50% of their genetic variance. For CMP traits, the QTL that contributed the most were those detected in the regions of the *PAEP*, *casein*, and *DGAT1* genes. However, other QTL regions on BTA5, 6, 16, 20, and 22 also generated noteworthy contributions. For protein composition traits, the highest-contributing QTL region was the *PAEP* gene region (up to 59% for β-LG). The region of the *casein* genes had a more moderate contribution (0.7–5.6%, depending on the trait), while the lesser-known QTL detected on BTA20 (at about 58 Mbp) explained 18, 9, and 7% of

Table 2 The 59 QTL regions identified by GWAS, the most likely candidate variant, and number of traits affected by the QTL

QTL region						Best candidate variant[a]										Number of traits[d]			
N	BTA	From (bp)	To (bp)	# variants	#genes	bp	Variant ID	Rank	R^2	MAF	Effect	SE	$-\log_{10}P$	Gene	Functional annotation	Trait most strongly affected	CMP	Proteins	Fatty acids	Minerals
1	1	144,389,419	144,398,814	6	1	144,395,375	rs136069703	1	0.97	0.38	29.5	0.951	210.4	SLC37A1	Intron	P	3	1	1	4
2	2	5,718,384	6,567,522	111	5	5,865,070	rs136830094	1	0.91	0.22	−0.02	0.003	15.6	INPP1	Downstream	Lactose	0	1	0	2
3	2	131,808,301	131,888,417	88	1	131,816,616	rs133677653	1	0.91	0.49	−0.05	0.003	41.4	ALPL	Intron	κ-CN	1	3	0	1
4	3	7,442,755	8,149,715	56	5	7,935,102	rs383033753	7	0.50	0.14	−0.65	0.096	10.7	FCGR2B	Intron	CY_{DM}	2	0	1	0
5	3	15,514,034	15,928,379	401	7	15,525,599	rs110073735	1	0.99	0.03	−2.78	0.410	10.9	EFNA1	Intron	Lactose	1	0	0	3
6	3	34,235,208	34,355,357	35	4	34,327,146	rs210558120	1	0.34	0.43	0.26	0.035	12.9	KIAA1324	Intron	a_{SC}	2	0	0	0
7	4	49,033,707	49,153,995	39	2	49,033,707	rs380575157	1	0.47	0.20	0.02	0.001	30.7	CBLL1	Upstream	$pH0_{-PCC}$	2	3	0	2
8	4	75,743,094	79,803,738	59	5	77,825,429	rs385069094	1	0.57	0.01	−2.34	0.360	10.1	GCK	Intron	Mg	0	1	0	3
9	4	92,588,016	92,966,245	227	7	92,624,543	rs379514460	1	0.34	0.03	−29.0	4.226	11.2	FSCN3	Upstream	Ca	0	0	0	1
10	5	29,947,476	31,423,430	315	25	29,947,476	rs442522314	2	0.63	0.09	−2.26	0.383	8.4	COX14	Downstream	Lactose	0	0	0	1
11	5	93,892,583	93,945,738	9	1	93,943,700	rs210744452	1	0.77	0.05	0.11	0.010	27.7	MGST1	Intron	FC	2	0	7	0
12	5	117,126,900	119,221,867	51	3	117,972,265	rs525880746	1	0.79	0.04	42.4	2.186	83.0	GRAMD4	Upstream	Ca	7	4	1	6
13	6	37,857,989	38,326,250	268	8	38,326,250	rs382477515	1	0.37	0.42	−0.10	0.017	9.5	IBSP	Upstream	β-CN	0	1	0	0
14	6	46,555,489	47,082,793	89	3	46,876,802	rs110408618	1	1.00	0.17	27.8	1.189	120.5	SEL1L3	Intron	K	4	5	1	4
15	6	87,125,482	87,961,577	143	7	87,392,899	rs382350292	3	0.53	0.23	0.03	0.002	76.4	CSN3	Downstream	$K10/R_{SC}$	8	8	0	1
16	6	108,967,622	109,145,154	91	5	109,026,822	rs208213463	1	0.93	0.47	−8.42	0.904	19.9	GAK	Intron	Ca	2	2	0	2
17	7	958,741	1,146,962	359	4	975,515	rs383246531	1	0.93	0.37	−1.56	0.180	17.4	47695[b]	Intron	Lactose	0	0	0	2
18	7	41,576,519	46,646,915	399	23	46,452,656	rs209828204	6	0.87	0.07	1.36	0.127	26.0	FSTL4	Downstream	a_{PCC}	2	2	0	1
19	9	102,731,669	103,130,410	207	3	102,885,120	rs134445867	15	0.90	0.03	−0.14	0.023	9.3	MPC1	Intron	C4_C10	0	0	1	0
20	10	1,984,741	2,326,212	86	1	2,096,282	rs385793060	67	0.99	0.35	−2.04	0.380	7.1	47622[b]	Upstream	Na	0	0	0	1
21	10	48,359,318	50,266,445	213	1	49,459,919	rs109896326	1	1.00	0.14	0.30	0.046	9.8	RORA	Intron	UNSAT	0	0	4	0
22	10	99,714,774	99,843,583	20	0	99,714,774	rs440530756	1	0.22	0.02	−0.18	0.028	9.8	–	Intergenic	PUNSAT	0	0	2	0
23	11	9,023,594	9,684,851	11	3	9,684,851	rs384459785	8	0.62	0.07	−2.82	0.385	12.6	POLE4	Intron	$CY_{FAT-PROT}$	3	0	3	0
24	11	14,152,677	15,493,112	132	6	14,284,886	rs384594145	48	0.90	0.10	−0.01	0.001	18.7	XDH	Upstream	pH_{0-PCC}	1	0	0	1
25	11	86,907,041	86,916,295	9	1	86,912,990	rs481567394	1	0.21	0.05	0.02	0.003	16.7	ATP6V1C2	Intron	pH_{0-PCC}	1	0	0	0
26	11	103,273,963	103,322,890	214	4	103,301,982	rs109907194	48	0.89	0.46	−1.09	0.017	559.7	PAEP	Intron	ΣWP	8	7	2	3
27	12	68,616,690	77,578,414	337	7	70,162,028	rs721489054	22	0.83	0.17	0.22	0.019	29.0	ABCC4	Intron	C14:0	0	0	7	2
28	13	20,094,707	22,437,171	291	4	21,053,894	rs378591536	7	0.34	0.07	−0.24	0.034	12.1	23216[b]	Intron	C14:0	0	0	1	0
29	13	45,394,264	48,611,254	75	4	46,734,011	rs379821485	18	0.76	0.15	−0.01	0.001	10.6	RF0026	Downstream	pH_{0-PCC}	1	0	1	0
30	13	52,289,279	55,114,121	80	9	54,938,610	rs110422533	2	1.00	0.39	−6.60	1.007	10.3	GID8	Synonymous	K	0	0	2	1

Table 2 (continued)

QTL region						Best candidate variant[a]										Number of traits[d]			
N	BTA	From (bp)	To (bp)	# variants	#genes	bp	Variant ID	Rank	R^2	MAF	Effect	SE	$-\log_{10}P$	Gene	Functional annotation	Trait most strongly affected	CMP	Proteins	Fatty acids	Minerals
31	13	64,648,620	64,870,118	19	3	64,812,464	rs43717459	1	0.33	0.10	−0.24	0.027	18.9	ACSS2	Intron	C4–C10	0	0	1	0
32	13	79,225,285	79,326,265	6	1	79,326,265	rs109862148	3	0.38	0.07	−0.04	0.007	8.9	FAM65C	Downstream	αs2-CN	0	1	0	0
33	14	1,622,956	1,881,116	27	9	1,629,753	rs109035586	1	0.36	0.33	0.17	0.007	122.7	GPT	Upstream	FC	6	6	10	4
34	15	39,854,757	40,283,508	140	2	39,885,845	rs134953698	2	0.79	0.18	−0.04	0.005	10.4	ARNTL	Intron	PUNSAT	0	0	1	0
35	15	52,993,384	54,612,060	341	10	53,943,342	rs381948106	1	0.47	0.04	17.5	2.747	9.7	RAB6A	Intron	P	0	0	2	2
36	16	1,607,369	3,050,452	10	0	1,609,129	rs42450079	1	0.54	0.28	−0.04	0.005	17.0	–	Intergenic	κ-CN	2	6	0	0
37	16	60,539,357	63,891,341	178	3	60,646,127	rs137615589	72	0.75	0.14	0.02	0.004	9.8	38238[b]	Upstream	a_{SC}	2	1	0	0
38	16	67,700,538	67,811,269	167	2	67,758,163	rs42465711	1	1.00	0.4	0.31	0.044	11.9	SWT1	Intron	CYDM	2	1	2	1
39	16	70,170,086	71,493,899	59	1	71,432,479	rs109766366	5	0.79	0.43	−0.23	0.032	12.3	PROX1	Intron	UNSAT	0	0	4	2
40	17	29,348,215	30,211,790	203	9	29,938,428	rs207509104	21	0.55	0.36	0.10	0.015	10.3	LARP1B	Synonymous	C4–C10	0	0	1	0
41	17	52,753,338	53,240,467	283	6	53,072,959	rs448501071	5	1.00	0.06	−0.32	0.019	62.4	BRI3BP	Intron	C4–C10	2	3	2	1
42	18	10,566,605	11,091,131	68	5	11,002,789	rs41867427	2	0.61	0.06	14.3	2.071	11.3	CRISPLD2	Intron	a_{SC}	2	0	0	1
43	19	51,304,834	51,538,272	67	2	51,383,847	rs136067046	1	0.83	0.32	0.20	0.013	49.4	FASN	Upstream	C14:0	0	0	9	0
44	19	55,229,384	57,240,571	267	10	57,151,350	rs42848485	1	0.63	0.46	9.07	1.079	16.4	FADS6	Intron	Ca	2	1	1	1
45	19	60,407,923	62,177,206	142	0	61,135,270	rs41923848	1	0.91	0.13	−2.28	0.210	26.6	–	Intergenic	Lactose	0	3	0	3
46	20	58,245,970	58,457,768	87	1	58,446,058	rs137085630	22	0.99	0.06	−1.01	0.036	176.5	ANKH	Intron	Citrate	5	4	2	5
47	21	40,120,343	44,138,058	63	2	41,638,428	rs137153434	25	0.21	0.12	0.07	0.009	12.7	G2E3	Upstream	PC	3	1	0	1
48	22	32,877,755	33,466,544	46	2	32,877,755	rs208141216	10	0.30	0.08	0.08	0.007	25.7	FAM19A4	Intron	PC	8	1	1	2
49	22	55,186,094	55,273,619	72	1	55,254,221	rs43597796	1	0.93	0.49	0.00	0.001	12.6	ATP2B2	Intron	pH_{0-PCC}	1	1	0	0
50	22	61,257,725	61,312,492	18	1	61,284,069	rs109001472	1	1.00	0.47	−0.09	0.008	28.3	KLF15	Intron	C4–C10	0	0	2	0
51	24	50,420,365	50,550,020	114	2	50,465,348	rs383068825	4	0.53	0.34	−10.7	1.282	16.0	SKA1	Downstream	K	1	1	0	2
52	24	58,744,952	58,825,217	23	3	58,817,202	rs208779762	1	0.93	0.43	−6.57	0.807	15.4	LMAN1	Upstream	P	1	1	0	1
53	25	2,994,081	3,261,509	28	3	3,241,838	rs137696417	16	0.61	0.18	−0.03	0.007	6.5	ADCY9	Downstream	κ-CN	0	1	0	0
54	25	25,642,563	29,605,418	82	11	26,498,356	rs137150057	1	1.00	0.34	−0.11	0.020	11.3	FAM57B	5′UTR	C18:0	1	1	6	0
55	26	20,727,700	21,427,109	61	4	21149234	rs136334180	11	1.00	0.31	0.39	0.035	27.6	SCD	Upstream	UNSAT	0	0	11	0
56	26	32,773,808	33,925,908	95	2	33,233,277	rs385554497	36	0.27	0.02	1.55	0.270	8.1	ACSL5	intron	a_{PCC}	1	0	0	0
57	27	36,165,492	36,235,730	11	1	36,212,352	rs208675276	1	0.54	0.41	0.65	0.040	60.4	GPAT4[c]	5′ UTR	C16:0	3	0	7	1
58	28	6,008,464	7,038,810	141	3	6,027,037	rs382911338	1	0.37	0.18	−0.01	0.001	17.1	PCNX2	Intron	pH_{0-PCC}	1	2	0	1
59	29	9,253,006	9,622,389	145	3	9,343,362	rs133715120	6	0.61	0.33	−2.76	0.264	24.9	EED	Intron	Lactose	0	3	1	4

[a] When a gene was present in the confidence interval of the QTL, the best candidate variant was the genic variant with the most significant effects (intergenic variants were discarded)

[b] 47695, 47622, 23216, and 38238 for ENSBTAG00000047695, ENSBTAG00000047622, ENSBTAG00000023216, and ENSBTAG00000038238, respectively

[c] Also named *AGPAT6*

[d] Number of milk cheese-making (CMP), protein, fatty acid, and mineral composition traits with significant effects

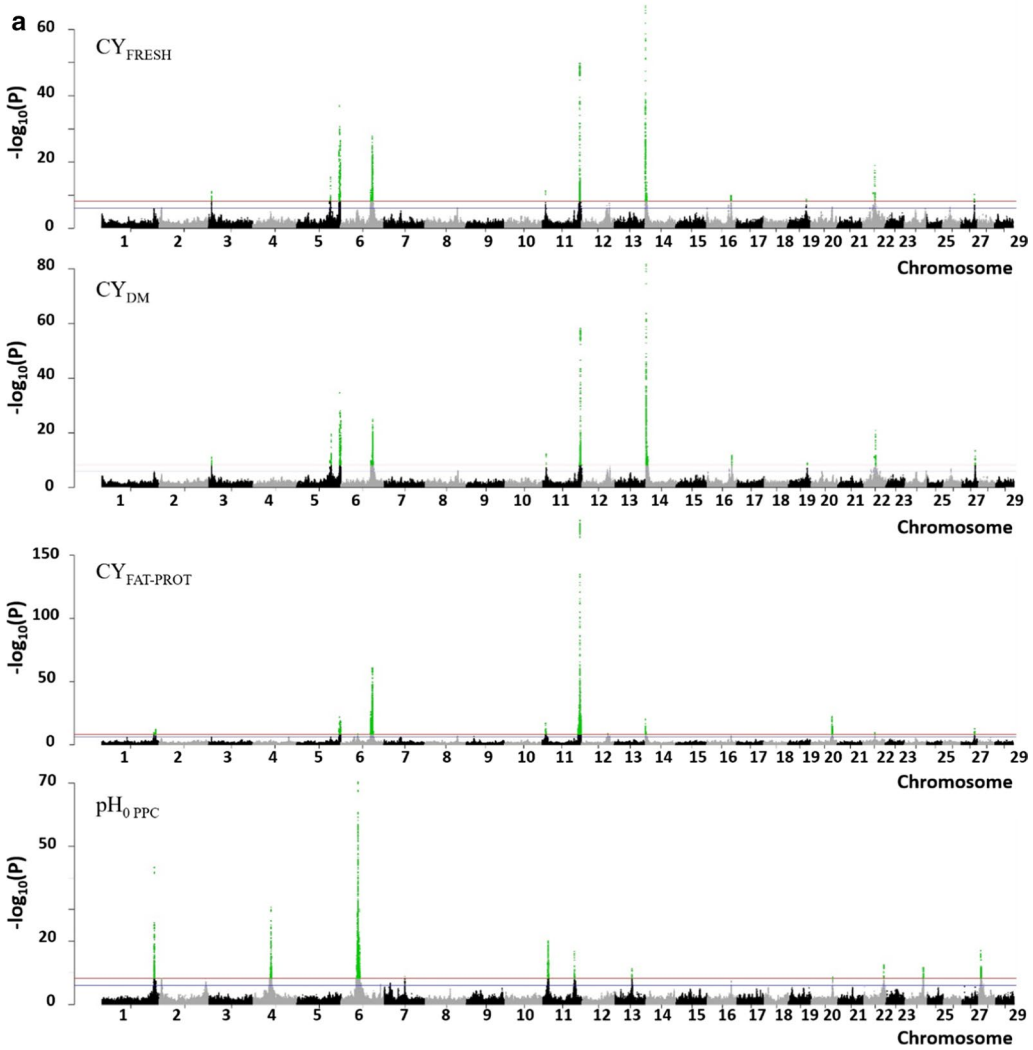


Fig. 2 $-\log_{10}$(P) values plotted against the position of variants on *Bos taurus* autosomes for cheese-making traits. **a** Cheese yields (CY) and $pH_{0\ PPC}$, **b** coagulation traits

the phenotypic variance of α-LA, αs1-CN, and κ-CN, respectively. For fatty acid content, the QTL that we detected explained a much smaller part of the phenotypic variability. The top-contributing QTL were the *DGAT1* gene region on BTA14 (12% for FC), *FASN* on BTA19 (1.5% for C14:0), *GPAT4* on BTA27 (3.2% for C16:0), and *SCD* on BTA26 (2% for C18:1). In contrast to fatty acids but similarly to proteins, a relatively large part of the phenotypic variance in mineral content was explained by QTL that were located in the region of the *SLC37A1* gene (3, 5, and 10% for Mg, K, and P, respectively) and the *ANKH* gene (20% for Mg). Two other regions influenced mineral content to a lesser extent: those at 117 Mbp on BTA5 (*GRAMD4*) and at 46 Mbp on BTA6 (*SEL1L3*).

Gene network

Using the AWM procedure, we reduced the set of 8.5 million variants tested in the GWAS to a set of 38,858 variants that had the most significant effects (P-value ≤ 0.001) on the key phenotype (CY_{DM}). Seven CMP (CY_{FRESH}, $CY_{FAT\text{-}PROT}$, and the five coagulation traits) and eight milk composition traits (PC, FC, UNSAT, PUNSAT, C18:1, Ca, Mg, and P) were correlated with CY_{DM} ($r \geq 0.25$). On average, each of the 38,858 variants had significant effects (P-value ≤ 0.001) on six of the correlated traits. We also retained 2322 additional variants that had significant effects on at least six of the correlated phenotypes. Thus, the final dataset included 41,180 variants, which had significant effects on CY_{DM} or on at least six correlated traits. Of

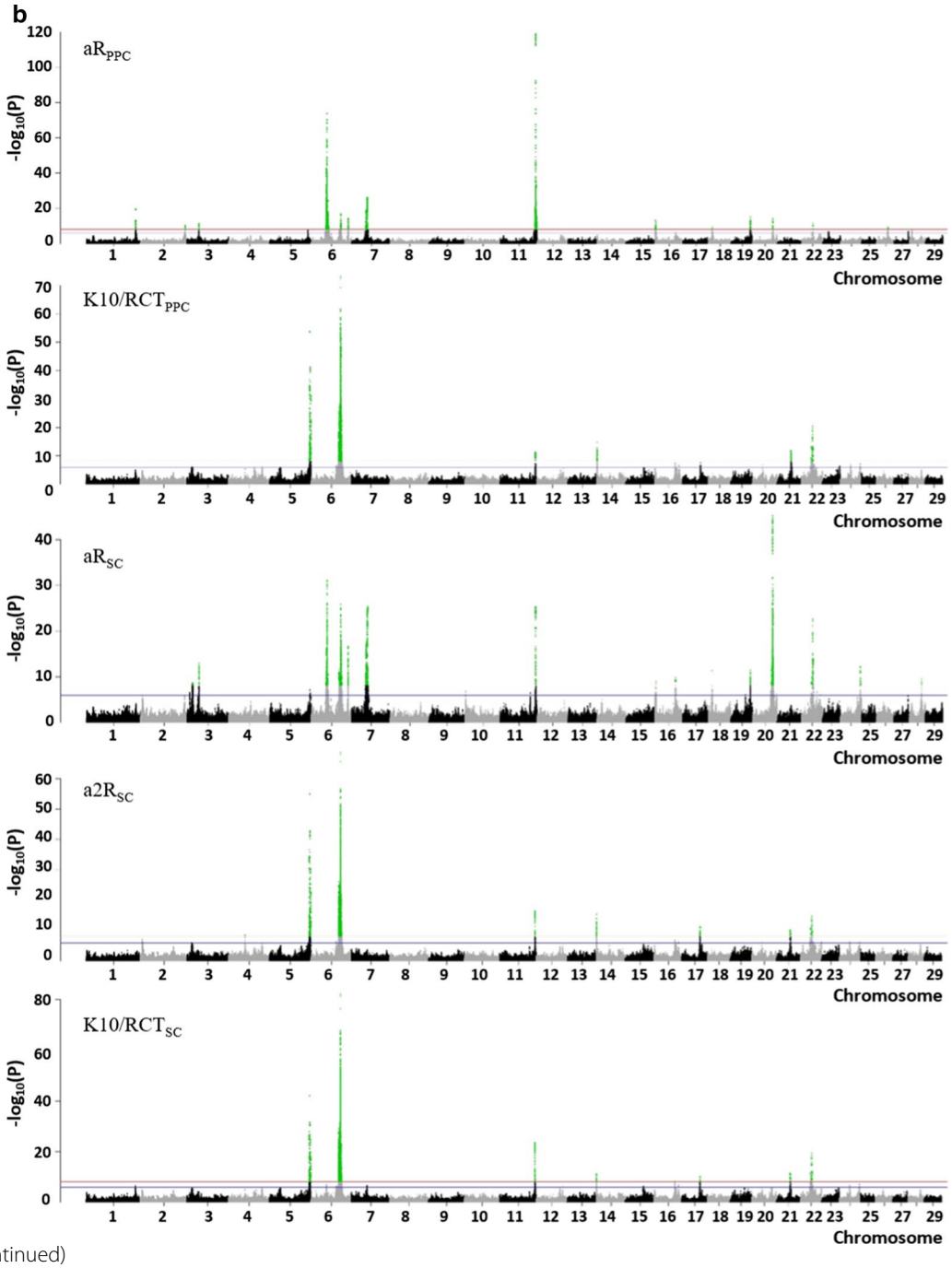


Fig. 2 (continued)

these 41,180 variants, 15,330 were located in 736 genes (±10 kb); the PCIT approach subsequently revealed 59,168 significant interactions among these genes. Thus, by merging the AWM and the PCIT approaches, the GWAS results on milk CMP and composition traits could be interpreted as a gene network of 736 nodes and 59,168 edges. The list of the 736 genes selected by AWM is in [see Additional file 4: Table S3].

For most of the traits, correlation coefficients from the z-score additive effects of the 736 variants retained by the AWM procedure were close to the correlation coefficients obtained from pedigree for the 16 phenotypes (Table 4). This suggested that the additive effects of the variants retained in the AWM analysis explained a large and representative part of the genetic relationships among the traits.

Table 3 Functional annotations of variants included in the 59 QTL regions

Functional annotation	Number of variants	%
Intergenic	3081	41.7
Intronic	2972	40.2
Upstream	604	8.2
Downstream	584	7.9
3′ UTR	26	0.35
5′ UTR	10	0.14
Synonymous	65	0.88
Missense	51	0.69
Total	7393	100

Among the 736 genes, 86 were located within QTL regions that had been highlighted by the GWAS analysis with a most-stringent threshold; these included the best candidate genes for 25 QTL. The remaining 650 genes were unique to the AWM analysis and had not been detected by GWAS. In contrast, 178 genes located within the confidence intervals of QTL detected with GWAS were not found in AWM analyses.

For each node of the gene network, we calculated the number of adjacent genes and the relative node contribution. Figure 3 lists the values of these parameters for the nodes of the gene network that were also best candidate genes in the GWAS analyses. This revealed genes that were highly connected with other genes in the network (*SWT1*, *GPT*, *MGST1*, *FCGR2B*, *CSN3*, *G2E3*, and *GRAMD4*), genes that were moderately connected (*RAB6A*, *FAM19A4*, *INPP1*, *CBLL1*, *ANKH*, *LMAN1*, *ARNTL*, *SLC37A1*, and *EED*), and genes that were poorly connected (*PAEP*, *FASN*, *GPAT4*, *SEL1L3*, *KIAA1324*, and *PROX1*).

In silico functional analyses

Key regulators in the gene network were identified in silico using two approaches. From the analyses of binding site motifs and ChIP-Seq datasets, first we identified eight TF that presented a significant normalized enrichment score (NES). Each of these TF targeted from 136 to 261 genes in the gene network (Table 5), and all eight together targeted more than half of the network genes (416). Using an information loss-less approach, we then identified among the 736 genes the trios of regulators (TF, miRNA, and lnRNA) that had the best coverage of the whole gene network, i.e. trios that demonstrated the largest number of interactions with genes of the network with the least amount of overlap. With this second approach, we found 61 regulators, each with two to 276 significant interactions with genes of the network. The trios that covered the largest number of genes were *ASXL3—HIC2—RNF2* and *HIC2—ZPFM2—bta-mir-200c*. These two trios interacted with the majority of the genes of the network, i.e. 529 and 528 genes, respectively.

Genes of the network were found to be enriched in five KEGG pathways and 115 GO terms (corrected *P*-value between 2.10^{-17} and 2.10^{-4}), which clustered into 44 functional groups (Fig. 4 and [see Additional file 5: Table S4]). The largest group comprised 15 GO terms; it contained 31 genes of the gene network and was related to the metabolic processes associated with potassium transport. The next three groups, with 28 GO terms and one KEGG pathway all related to phosphate and phospholipid metabolism, contained 66 genes of the network. Among these, there were many of the genes that had been highlighted by GWAS as having the most significant effects on milk CMP and composition traits: *CSN1S1*, *DGAT1*, *FASN*, *GPAT4*, *INPP1*, *PPARA*, *PROX1*, and *SCD*. Other groups, (for details [see Additional file 5: Table S4]), had a functional relationship with milk composition through endopeptidase activity (16 genes, including *CSN2* and *GRAMD4*), protein glycosylation (19 genes), carboxylic acid biosynthesis (24 genes including *FASN*, *PAEP*, *PPARA*, *PROX1*, and *SCD*), inorganic anion transport (10 genes including *ANKH* and *SLC37A1*), and Ca- (11 genes) and phospholipase- (9 genes) signaling pathways.

Discussion

GWAS and gene network analyses are complementary

The GWAS approach used here—performed on whole-genome sequences from a large number of animals for complex cheese-making traits as well as fine-scale milk composition traits—led to the identification of 59 QTL regions. In order to limit the detection of false positives, we retained only the QTL that still demonstrated significant effects after applying the Bonferroni correction ($P\text{-value} < 5.8 \times 10^{-9}$) and therefore those that presented the strongest effects overall. The downside of this approach was that all the QTL in our analysis explained, on average, less than 50% of the total genetic variation of each trait (i.e. 20% of the phenotypic variance), and this value was probably overestimated. Indeed, when the true effect is small or when the P-value threshold is very low, the detection power is limited and a significant effect may be overestimated, leading to an overestimation of SNP variance. Some QTL were identified with very good resolution (narrow peaks), such as the 12 QTL for which only one candidate gene was identified within the confidence interval. Other QTL regions were larger and more gene-rich (up to 25 genes within the confidence interval), and identification of the best candidate gene was not straightforward. To address these two shortcomings—specifically, to capture the missing genetic

Table 4 Genomic correlations calculated using additive effects of the 736 SNPs selected by the AWM (above the diagonal) and genetic correlations estimated from pedigree or taken from Sanchez et al. [3] (below the diagonal)

	CY_{FRESH}	CY_{DM}	$CY_{FAT\text{-}PROT}$	$K10/RCT_{PCC}$	a_{PCC}	$K10/RCT_{SC}$	a_{SC}	$a2_{SC}$	PC	FC	C18:1	UFA	PUFA	Ca	Mg	P
CY_{FRESH}		1.00	−0.87	−0.84	0.25	−0.84	0.52	0.85	0.84	0.96	−0.82	−0.78	−0.80	0.70	0.52	0.83
CY_{DM}	0.97		−0.88	−0.85	0.27	−0.85	0.55	0.86	0.85	0.95	−0.80	−0.76	−0.78	0.71	0.55	0.84
$CY_{FAT\text{-}PROT}$	−0.82	−0.84		0.79	−0.61	0.84	−0.73	−0.79	−0.72	−0.72	0.57	0.51	0.58	−0.55	−0.54	−0.75
$K10/RCT_{PCC}$	−0.72	−0.73	0.65		−0.31	0.99	−0.69	−0.99	−0.98	−0.74	0.57	0.52	0.44	−0.80	−0.75	−0.86
a_{PCC}	0.77	0.78	−0.65	−0.76		−0.41	0.79	0.29	0.20	0.01	0.08	0.14	0.02	−0.12	0.29	0.35
$K10/RCT_{SC}$	−0.74	−0.76	0.71	0.80	−0.78		−0.75	−0.98	−0.95	−0.71	0.53	0.47	0.41	−0.75	−0.76	−0.87
a_{SC}	0.76	0.78	−0.67	−0.73	0.76	−0.77		0.63	0.64	0.32	−0.19	−0.14	−0.15	0.36	0.72	0.62
$a2_{SC}$	0.72	0.75	−0.64	−0.72	0.77	−0.77	0.74		0.97	0.75	−0.59	−0.53	−0.45	0.79	0.68	0.87
PC	0.74	0.75	−0.52	−0.80	0.94	−0.81	0.91	0.89		0.77	−0.59	−0.55	−0.46	0.84	0.76	0.86
FC	0.91	0.87	−0.57	−0.53	0.55	−0.47	0.51	0.48	0.60		−0.89	−0.87	−0.87	0.69	0.41	0.75
C18:1	−0.38	−0.45	0.22	0.20	−0.22	0.13	−0.18	−0.17	−0.23	−0.57		0.99	0.93	−0.52	−0.19	−0.56
UFA	−0.34	−0.40	0.17	0.13	−0.14	0.05	−0.10	−0.09	−0.21	−0.55	0.47		0.93	−0.50	−0.16	−0.51
PUFA	−0.47	−0.42	0.30	0.00	−0.02	−0.04	0.01	0.04	−0.29	−0.59	0.71	0.74		−0.42	−0.13	−0.50
Ca	0.41	0.40	−0.25	−0.46	0.45	−0.37	0.39	0.42	0.41	0.26	−0.30	−0.30	−0.35		0.76	0.71
Mg	0.54	0.58	−0.44	−0.58	0.59	−0.58	0.54	0.54	0.40	0.20	−0.25	−0.25	−0.05	0.60		0.70
P	0.40	0.41	−0.31	−0.54	0.50	−0.53	0.42	0.40	0.40	0.29	−0.31	−0.30	−0.39	0.34	0.58	

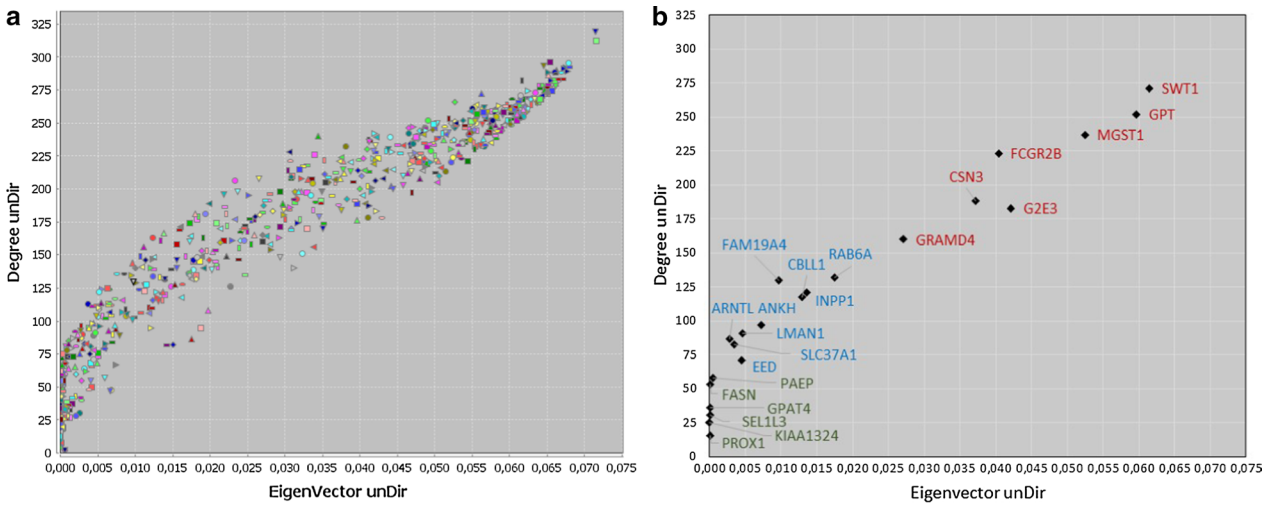


Fig. 3 Centiscape scatter plot view: number of adjacent genes (degree) and the relative node contribution (eigenvector) **a** for the 736 genes of the gene network and **b** for the best candidate genes

Table 5 Transcription factors (TFs) identified as key regulators of milk cheese-making and composition traits from both binding-site motifs and ChIP-Seq datasets, which presented significant normalized enrichment scores (NES)

TF	NES	Number of binding site motifs	Number of ChIP-Seq datasets	Number of target genes	Chromosome	Gene start (bp)	Gene end (bp)	Gene description
HSPA1L	4.90	5	1	261	23	27,334,344	27,338,328	Heat shock 70 kDa protein 1-like
SMAD5	4.63	4	2	253	7	49,155,483	49,217,780	SMAD family member 5
HNF1B	4.56	3	5	242	19	14,287,673	14,349,579	HNF1 homeobox B
SMAP2	4.30	7	1	236	3	106,311,859	106,358,978	Small ArfGAP2
TFAP2A	4.29	3	1	233	23	45,480,546	45,499,034	Transcription factor AP-2 alpha
BCL11A	4.25	5	1	195	11	43,071,977	43,174,031	B Cell CLL/Lymphoma 11A
SMAD3	4.02	3	1	170	10	13,958,174	13,980,371	SMAD family member 3
RXRA	3.49	2	1	136	11	105,990,344	106,015,000	Retinoid X receptor alpha

variability and to better identify functional candidate genes within QTL regions—we carried out additional analyses, which complemented our GWAS results. The AWM-PCIT approach enabled us to identify a more comprehensive gene network of 736 genes from lower significant GWAS results (P-value < 0.001) by taking co-associations between traits into account. When we used the additive effects of variants that were located in these genes to calculate correlations between traits, the values obtained were similar to the genetic correlations we calculated from pedigree [3], suggesting that the gene network adequately explained the genetic relationships between traits. Finally, in silico functional analyses of genes of the network helped us to identify metabolic pathways and key regulators with functional links to milk cheese-making and composition traits. This last step, in addition to establishing functional links between the gene network and the analyzed traits, enabled us to identify candidate genes in some QTL regions. Therefore, by combining the results obtained through these different approaches, we are able to propose candidate genes for the main QTL regions, and for each, the best candidate for the causative variant, or at least, a variant in high LD with the causative variant.

Functional candidate genes

As expected, we confirmed the strong effects of the cluster of casein genes and the *PAEP* gene regions on protein composition as well as milk CMP. The QTL detected in the casein genes region explained up to 20% σ^2_p of the

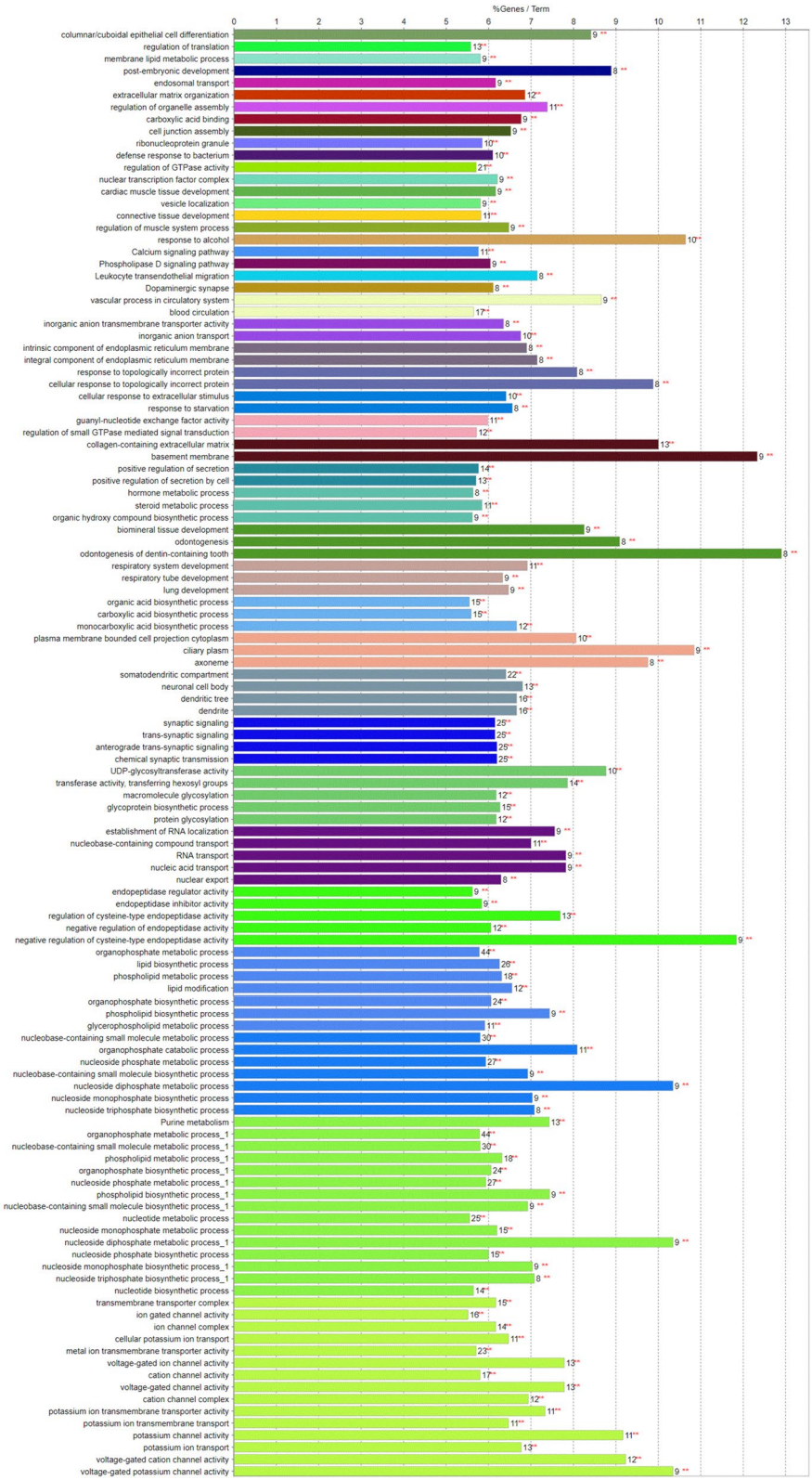


Fig. 4 Description of the five KEGG pathways and 105 GO terms that were significantly enriched among genes of the network and which clustered in 44 functional groups

curd-firming time while the *PAEP* gene region explained up to 8.5% σ^2_P of cheese yields. The best candidate gene variants, i.e. variants with the most significant effects on traits, were located in the downstream region of the *CSN3* gene, which encodes κ-CN (at 87,392,899 bp on BTA6), and in an intronic region of the *PAEP* gene, which encodes β-LG (at 103,301,982 pb on BTA11). The missense variants that are respectively responsible for the κ-CN [41] and β-LG [42] A/B polymorphisms had much weaker effects: they were ranked 100th and 56th, respectively, among the variants. The region of the *DGAT1* gene on BTA14 had also large effects on milk composition (12% σ^2_P for FC) and on CMP (6.4% σ^2_P for CY_{DM}). In spite of its low MAF in Montbéliarde cows (0.015), the *K232A DGAT1* mutation [43] was the top-ranked variant for traits that were linked with some protein and phosphorous contents, and coagulation traits (1st for PC, α-LA, β-CN, and P; 2nd for $a2_{SC}$, $K10/RCT_{SC}$, and $K10/RCT_{PCC}$) and it was one of the 736 variants retained by the AWM. However, in this gene-rich region, the *GPT* gene, which we found to be highly connected, i.e. presenting significant gene–gene interactions with many other genes of the AWM gene network, appeared to be also a good candidate for FC, CY_{DM}, CY_{FRESH}, and fatty acid composition. The best candidate variant, located in the upstream region of *GPT* (*glutamic-pyruvic transaminase*) at 1,629,753 bp (rs109035586), was ranked 1st for 12 traits, including FC, cheese yields, fatty acid composition, αS1-CN, and CITRATE. Interestingly, two polymorphisms in the *GPT* gene, including a missense variant that is located very close to the best candidate variant detected in our study (1,629,600 bp), were also recently found to be associated with fat percentage in a concordance analysis carried out on imputed whole-genome sequences of Holstein bulls [44]. This variant was also highly significant in our study but was ranked 44th among variants with significant effects on FC.

In addition to the three well-known QTL regions described above, we also found evidence that other genomic regions have highly significant effects on the traits analyzed; specifically, our analysis highlighted the *SLC37A1*, *ALPL*, *MGST1*, *SEL1L3*, *FASN*, *ANKH*, *BRI3BP*, *SCD*, and *GPAT4* genes, which we had also previously detected in a sequence-based GWAS on milk protein and fatty acid composition [45, 46]. We confirm here their effects on milk composition and note their effects on CMP. As previously found, the *MGST1*, *FASN*, *SCD*, and *GPAT4* genes mainly affected fatty acids whereas the *SLC37A1*, *ALPL*, *SEL1L3*, *BRI3BP*, and *ANKH* genes had effects mainly on proteins and minerals. As a consequence, and in accordance with genetic correlations that we had previously estimated from this dataset [3], the former set of genes exclusively influenced cheese yields whereas the latter set had greater effects on coagulation traits. Strong effects of *ALPL*, *ANKH,* and *SEL1L3,* which we had previously identified for protein composition [45], were confirmed for milk composition and CMP. In each of these regions, the current analysis reduced the size of the confidence intervals of the QTL and, in six of them, only one gene was found that encoded a known protein (*SLC37A1*, *ALPL*, *MGST1*, *SEL1L3*, *ANKH*, and *GPAT4*).

On BTA17, we found two QTL regions associated with de novo milk fatty acid synthesis, i.e. synthesis within the mammary epithelial cells of fatty acids C4:0 to C10:0. The first was within the *LARP1B* (*La ribonucleoprotein domain family member 1B*) gene, for which the best candidate was a synonymous variant located at 29,938,428 bp. This result corroborates the discovery of Duchemin et al. [47], who identified *LARP1B* as a causative gene for de novo synthesis of milk fatty acids through the imputation of sequence variants in this region. These authors noted a splice-region variant at 29,940,555 bp, which was close to the variant that we detected here. However, in spite of its high MAF (0.40), we excluded this variant because it was not significant in our study (P-value $= 10^{-4}$ vs. 5.10^{-11} for the variant located at 29,938,428 bp). This region had limited effects in our study and affected only short FA traits. Instead, further along the same chromosome, we identified another region with much more significant effects on de novo fatty acid synthesis that also affected CMP and protein and mineral composition. The best candidate gene for this region was *BRI3BP* (*BRI3 binding protein*), with the most significant variant located at 53,072,959 bp in an intron of *BRI3BP*. This variant had been previously highlighted for its effects on FA composition in an independent population [48] and, in another study, we recently confirmed its effects on both CMP and milk composition traits [46]. Thus, it is a serious candidate for the causative variant behind the strong effects that we observed in the region. Although the *BRI3BP* gene was not an obvious functional candidate, it has been also described as affecting de novo fatty acid synthesis in a recent GWAS performed on imputed sequence variants in this region [49]. The most significant variant found by the authors of this study was also intronic (53,078,216 bp) but that particular variant was ranked 31st among variants with significant effects on C4–C10.

Finally, we identified other candidate genes that contained variants with non-negligible effects on milk composition and CMP traits. Among these, both GWAS and AWM analyses highlighted *FCGR2B*, *KIAA1324*, *CBLL1*, *GRAMD4*, *ARNTL*, *RAB6A*, *ENSBTAG00000038238*, *SWT1*, *G2E3*, *FAM19A4*, *LMAN1*, and *EED*. The *FCGRB2*, *KIAA1324*, *G2E3*, *LMAN1*, and *EED* genes have been previously identified as candidate genes for milk yield or milk composition [50–54], whereas the

functional link between the other genes and bovine milk composition and cheese-making traits remains to be discovered.

Co-association gene network

The *SLC37A1* (*solute carrier family 37 member 1, a phosphorous antiporter*) and *ANKH* (*inorganic pyrophosphate transport regulator*) genes, which encode transmembrane proteins involved in ion transport, both play a role in the inorganic anion transport that was revealed by the GO analysis. Thus, these genes are good candidates for having an effect on CMP and milk composition, with the strongest effects obtained for phosphorous (about 11% σ^2_P) and citrate (about 32% σ^2_P) contents, respectively. For each of these genes, we propose here an intronic candidate variant, located at 58,446,058 bp for *ANKH* and at 144,395,375 bp for *SLC37A1*. Very close to but distinct from those identified in previous studies [45, 53, 55], this variant is more significant in spite of a slightly lower imputation accuracy.

A set of genes, including those detected previously (*DGAT1*, *FASN*, *GPAT4*, *CSN1S1*, *PAEP*, and *SCD*) and those noted here for the first time (*INPP1*, *PPARA*, *PROX1*), appeared to play a role in phosphate and phospholipid metabolism as well as in the biosynthesis of carboxylic acids, which are fatty acid precursors. *PROX1* (*prospero homeobox 1*) and *PPARA* (*peroxisome proliferator activated receptor alpha*) encode transcription factors; the former interacted with only 16 genes while the latter interacted with 128 genes within the network, including with *FASN*, *SCD*, *GPAT4*, and *DGAT1*. *PPARA* belongs to a superfamily of hormone receptors (*PPAR*) that regulate the transcription of genes involved in different lipid metabolism pathways [56]. *FASN* (*fatty acid synthase*) and *SCD* (*stearoyl-coenzyme A desaturase 1)* encode key enzymes in de novo fatty acid synthesis and fatty acid desaturation, respectively, and *GPAT4* (*glycerol-3-phosphate acyltransferase 4*) is paralogous to *DGAT1* (*diacylglycerol O-acyltransferase 1*), with the two genes occupying adjacent nodes of the mammary triglyceride synthesis chain [57]. In addition to their effects on protein composition, the *PAEP* and *CSN1S1* genes, which encode milk β-LG and αs1-CN proteins, respectively, are also associated with genes involved in fatty acid metabolism. These results suggest a close link between milk fatty acid and protein metabolism. In goats, variants that are responsible for a decrease in *CSN1S1* gene expression were also associated with a decrease in fat content, probably due to disruption of the structure and secretion of fat globules [58]. A similar relationship was pointed out in cattle by Knutsen et al. [49], who found a major effect of the *PAEP* gene region on the C4:0 content of bovine milk, and Pausch et al. [53], who identified strong pleiotropic effects of variants located in the *CSN1S1* gene on fat and protein content. In addition, a strong association between PAEP and omega-3 fatty acids was observed by Boichard et al. [48]. All of these genes, which contain the top-ranked variants for, in particular, cheese yields and fatty acid composition, thus represent good candidates. Alone, they explained the largest part of the phenotypic variance captured in the present study for CY_{DM} and FC, i.e. around 16% out of 20%.

In addition to the PPARA TF, we highlight here other genes for putative regulators as well, such as *ASXL3* (*additional sex combs like 3, transcriptional regulator*) and *bta-mir-200c*, which interact with many genes of the network (276 and 240, respectively). Both are good candidates for key regulators in the network, as the protein encoded by *ASXL3* has been shown to negatively regulate lipogenesis and *bta-mir-200c* miRNA has been found to be highly expressed in the mammary gland [59–61] and present in milk whey [62]. Interestingly, all of the regulators that we identified in our study were different from the TF found in previous studies that applied similar approaches to study milk proteins [10] or fatty acids [9]. Unlike these studies, we analyzed here milk protein, fatty acid, and mineral composition as well as cheese-making traits all together, which might explain the identification of different regulatory pathways. However, in spite of this, some of the significantly enriched GO terms or KEGG pathways that we highlight here were concordant with those previously reported for CMP traits (Ca signaling pathway) [7], milk protein content (potassium ion transport) [10], or fatty acid content (hormone and steroid metabolic processes) [9].

Causative variants

The approach that we used, which combines GWAS and post-GWAS analyses, was successful both in confirming previously reported candidate genes and in identifying new candidates that appear to be functionally linked to the analyzed traits. This was possible because our analyses were based on a large sample size, sequence-level genotypes, and detailed phenotypes for milk components in addition to complex CMP traits. However, for most of these genes, the top-ranked variant identified here was different both from what we had found before in an analysis of milk protein and fatty acid composition and from what had been detected in previous studies. Since the first GWAS on WGS imputed from the 1000 Bull Genomes reference population, in 2014 [4], to date published GWAS based on this approach have generally converged towards the same candidate genes but rarely towards the same best candidate variants in these genes. Using data from humans, Faye et al. [63] showed that when the causal variant is less accurately genotyped or imputed than

one of its highly correlated neighboring variants, the neighboring variant can capture the association better than the causal variant. However, in our study, the HD SNP, imputed more accurately than sequence variants, were rarely the top variants of the peaks, with the noticeable exception in the *SCD* gene. For *SLC37A1*, the peak variant was more significant than variants already proposed in other studies and slightly better imputed. Nevertheless, we can anticipate that by accumulating bovine sequence data from different breeds and different populations, future runs of the 1000 Bull Genome Project will lead to better identification of causative variants by GWAS. More specifically, the expansion of the bovine sequence database should increase the accuracy of imputed genotypes and thus the probability of identifying the right variant. In addition, if GWAS analyses can be carried out in different breeds, meta-analyses should lead to a better resolution due to the linkage disequilibrium at shorter distances between breeds than within breed, and thus to a better discrimination of causal variants.

Conclusions

By combining GWAS and AWM approaches at the whole-genome sequence level on milk cheese-making and composition traits predicted from MIR spectra, this study highlights candidate genes with major effects that are functionally related to milk composition. For most of these, we are able to propose some candidate variants that are likely to be either causative or in linkage disequilibrium with causative variants. In addition to providing a better understanding of the metabolic pathways involved in the genetic determinism of cheese-making traits, this study should make it possible to select a set of variants that explain a large part of the genetic variability of cheese-making traits. The increase in the number of cows for which both genotypes and phenotypes are available allows better detection of variants which could be included in genomic prediction to more accurately select animals with high genetic merit for CMP and finally improve the efficiency of the cheese-making process, which is of vital economic importance in the dairy industry.

Additional files

Additional file 1: Figure S1. $-\log_{10}(P)$ plotted against the position of variants on *Bos Taurus* autosomes for milk composition. Manhattan Plot obtained from GWAS for milk composition traits.

Additional file 2: Table S1. Description of the 264 genes located in confidence intervals of the QTL detected by GWAS for milk cheese-making and composition traits. Name and position of candidate genes identified by GWAS.

Additional file 3: Table S2. Percentage of the phenotypic variance of milk CMP and composition traits explained by each QTL. Individual effects of QTL on all CMP and composition traits, expressed as a percentage of the phenotypic variance of the trait.

Additional file 4: Table S3. Description of the 736 genes selected by the gene network analysis for milk cheese-making and composition traits. Name and position of candidate genes identified by AWM.

Additional file 5: Table S4. Gene ontology (GO) terms and KEGG pathways for genes selected by AWM. Name, description and list of genes of the gene ontology terms and KEGG pathways identified in the gene network analysis.

Acknowledgements
The authors gratefully acknowledge the breeders who participated in the *FROM'MIR* project; colleagues from the Conseil-Elevage 25-90 and INRA who coordinated farm sampling and data collection, and performed cheese-making reference measurements; the members of the scientific committee for advising on and managing this work; and the contribution of the 1000 Bull Genomes consortium.

Authors' contributions
MPS performed GWAS, network, and pathway analyses and wrote the manuscript. YRC and AM provided support in developing computing programs, ST performed imputation analyses, and MBo managed sequence analyses of the 1000 Bull Genomes Project. MEJ developed the MIR prediction equations for milk cheese-making properties. VW, CL, SF, ADB, MBr, and DB designed and managed the *FROM'MIR* project. All authors read and approved the final manuscript.

Funding
This study was funded by the French Ministry of Agriculture, Agro-food, and Forest, the French Dairy Interbranch Organization (CNIEL), the Regional Union of Protected Designation Cheeses of Franche-Comté (URFAC), and the Regional Council of Bourgogne-Franche-Comté, under the project *FROM'MIR*.

Availability of data and material
The data (genotypes and phenotypes) that enabled the findings of this study were made available by UMOTEST, CEL25-90, and HSCEL. However, restrictions apply to the availability of these data: they were used under license for the current study, and are not publicly available.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] GABI, INRA, AgroParisTech, Université Paris Saclay, 78350 Jouy-en-Josas, France. [2] Conseil Elevage 25-90, 25640 Roulans, France. [3] Institut de l'Elevage, 75012 Paris, France. [4] Allice, 75012 Paris, France. [5] Umotest, 01250 Ceyzériat, France.

Received: 20 December 2018 Accepted: 7 June 2019
Published online: 01 July 2019

References

1. De Marchi M, Toffanin V, Cassandro M, Penasa M. Invited review: mid-infrared spectroscopy as phenotyping tool for milk traits. J Dairy Sci. 2014;97:1171–86.

2. Wedholm A, Larsen LB, Lindmark-Månsson H, Karlsson AH, Andrén A. Effect of protein composition on the cheese-making properties of milk from individual dairy cows. J Dairy Sci. 2006;89:3296–305.
3. Sanchez MP, El Jabri M, Minéry S, Wolf V, Beuvier E, Laithier C, et al. Genetic parameters for cheese-making properties and milk composition predicted from mid-infrared spectra in a large dataset of Montbéliarde cows. J Dairy Sci. 2018;101:10048–61.
4. Daetwyler HD, Capitan A, Pausch H, Stothard P, Van Binsbergen R, Brøndum RF, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet. 2014;46:858–67.
5. Fortes MRS, Reverter A, Zhang Y, Collis E, Nagaraj SH, Jonsson NN, et al. Association weight matrix for the genetic dissection of puberty in beef cattle. Proc Natl Acad Sci USA. 2010;107:13642–7.
6. Reverter A, Fortes MB. Breeding and genetics symposium: building single nucleotide polymorphism-derived gene regulatory networks: towards functional genomewide association studies. J Anim Sci. 2013;91:530–6.
7. Dadousis C, Pegolo S, Rosa GJM, Gianola D, Bittante G, Cecchinato A. Pathway-based genome-wide association analysis of milk coagulation properties, curd firmness, cheese yield, and curd nutrient recovery in dairy cattle. J Dairy Sci. 2017;100:1223–31.
8. Buitenhuis B, Janss LL, Poulsen NA, Larsen LB, Larsen MK, Sorensen P. Genome-wide association and biological pathway analysis for milk-fat composition in Danish Holstein and Danish Jersey cattle. BMC Genom. 2014;15:1112.
9. Pegolo S, Dadousis C, Mach N, Ramayo-Caldas Y, Mele M, Conte G, et al. SNP co-association and network analyses identify E2F3, KDM5A and BACH2 as key regulators of the bovine milk fatty acid profile. Sci Rep. 2017;7:17317.
10. Pegolo S, Mach N, Ramayo-Caldas Y, Schiavon S, Bittante G, Cecchinato A. Integration of GWAS, pathway and network analyses reveals novel mechanistic insights into the synthesis of milk proteins in dairy cows. Sci Rep. 2018;8:566.
11. Gambra R, Penagaricano F, Kropp J, Khateeb K, Weigel KA, Lucey J, et al. Genomic architecture of bovine kappa-casein and beta-lactoglobulin. J Dairy Sci. 2013;96:5333–43.
12. Ferrand M, Miranda G, Guisnel S, Larroque H, Leray O, Lahalle F, et al. Determination of protein composition in milk by mid-infrared spectrometry. In Proceedings of the VI ICAR reference laboratory network meeting: 28 May 2012; Cork; 2012
13. Ferrand-Calmels M, Palhiere I, Brochard M, Leray O, Astruc JM, Aurel MR, et al. Prediction of fatty acid profiles in cow, ewe, and goat milk by mid-infrared spectrometry. J Dairy Sci. 2014;97:17–35.
14. Sanchez MP, Ferrand M, Gele M, Pourchet D, Miranda G, Martin P, et al. Short communication: genetic parameters for milk protein composition predicted using mid-infrared spectroscopy in the French Montbeliarde, Normande, and Holstein dairy cattle breeds. J Dairy Sci. 2017;100:6371–5.
15. Gengler N, Soyeurt H, Dehareng F, Bastin C, Colinet F, Hammami H, et al. Capitalizing on fine milk composition for breeding and management of dairy cows. J Dairy Sci. 2016;99:4071–9.
16. Ducrocq V. Genekit, BLUP software. June 2011 version. Jouy-en-Josas: INRA GABI; 1998.
17. Sargolzaei M, Chesnais JP, Schenkel FS. A new approach for efficient genotype imputation using information from relatives. BMC Genom. 2014;15:478.
18. Boichard D, Guillaume F, Baur A, Croiseau P, Rossignol M, Boscher MY, et al. Genomic selection in French dairy cattle. Anim Prod Sci. 2012;52:115–20.
19. van Binsbergen R, Bink MC, Calus MP, van Eeuwijk FA, Hayes BJ, Hulsegge I, et al. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genet Sel Evol. 2014;46:41.
20. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet. 2012;44:955–9.
21. Brondum RF, Guldbrandtsen B, Sahana G, Lund MS, Su G. Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. BMC Genom. 2014;15:728.
22. Pausch H, MacLeod I, Fries R, Emmerling R, Bowman PJ, Daetwyler HD, et al. Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal variant detection in cattle. Genet Sel Evol. 2017;49:24.
23. Hoze C, Fouilloux MN, Venot E, Guillaume F, Dassonneville R, Fritz S, et al. High-density marker imputation accuracy in sixteen French cattle breeds. Genet Sel Evol. 2013;45:33.
24. Bouwman AC, Veerkamp RF. Consequences of splitting whole-genome sequencing effort over multiple breeds on imputation accuracy. BMC Genet. 2014;15:105.
25. Bouwman AC, Daetwyler HD, Chamberlain AJ, Ponce CH, Sargolzaei M, Schenkel FS, et al. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. Nat Genet. 2018;50:362–7.
26. Boussaha M, Michot P, Letaief R, Hoze C, Fritz S, Grohs C, et al. Construction of a large collection of small genome variations in French dairy and beef breeds using whole-genome sequences. Genet Sel Evol. 2016;48:87.
27. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
28. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 2010;26:2069–70.
29. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4:1073–82.
30. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88:76–82.
31. Fu WX, Liu Y, Lu X, Niu XY, Ding XD, Liu JF, et al. A genome-wide association study identifies two novel promising candidate genes affecting *Escherichia coli* F4ab/F4ac susceptibility in swine. PLoS One. 2012;7:e32127.
32. Ramayo-Caldas Y, Renand G, Ballester M, Saintilan R, Rocha D. Multi-breed and multi-trait co-association analysis of meat tenderness and other meat quality traits in three French beef cattle breeds. Genet Sel Evol. 2016;48:37.
33. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl variant effect predictor. Genome Biol. 2016;17:122.
34. Reverter A, Chan EK. Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. Bioinformatics. 2008;24:2491–7.
35. Watson-Haigh NS, Kadarmideen HN, Reverter A. PCIT: an R package for weighted gene co-expression networks based on partial correlation and information theory approaches. Bioinformatics. 2010;26:411–3.
36. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504.
37. Scardoni G, Petterlini M, Laudanna C. Analyzing biological network parameters with CentiScaPe. Bioinformatics. 2009;25:2857–9.
38. Janky R, Verfaillie A, Imrichova H, Van de Sande B, Standaert L, Christiaens V, et al. iRegulon: from a gene list to a gene regulatory network using large motif and track collections. PLoS Comput Biol. 2014;10:e1003731.
39. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, et al. Architecture of the human regulatory network derived from ENCODE data. Nature. 2012;489:91–100.
40. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, et al. ClueGO: a Cytoscape plug-into decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009;25:1091–3.
41. Grosclaude F, Mahé MF, Mercier JC, Ribadeau-Dumas B. Localisation des substitutions d'acides aminés différenciant les varianst A et B de la caséine kappa bovine. Ann Genet Sel Anim. 1972;4:515–21.
42. Ganai NA, Bovenhuis H, van Arendonk JA, Visker MH. Novel polymorphisms in the bovine *beta-lactoglobulin* gene and their effects on beta-lactoglobulin protein concentration in milk. Anim Genet. 2009;40:127–33.
43. Grisart B, Coppieters W, Farnir F, Karim L, Ford C, Berzi P, et al. Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine *DGAT1* gene with major effect on milk yield and composition. Genome Res. 2002;12:222–31.
44. Weller JI, Bickhart DM, Wiggans GR, Tooker ME, O'Connell JR, Jiang J, et al. Determination of quantitative trait nucleotides by concordance analysis between quantitative trait loci and marker genotypes of US Holsteins. J Dairy Sci. 2018;101:9089–107.
45. Sanchez MP, Govignon-Gion A, Croiseau P, Fritz S, Hozé C, Miranda G, et al. Within-breed and multi-breed GWAS on imputed whole-genome

sequence variants reveal candidate mutations affecting milk protein composition in dairy cattle. Genet Sel Evol. 2017;49:68.
46. Sanchez MP, Wolf V, El Jabri M, Beuvier E, Rolet-Répécaud O, Gaüzère Y, et al. Short communication: confirmation of candidate causative variants on milk composition and cheesemaking properties in Montbéliarde cows. J Dairy Sci. 2018;101:10076–81.
47. Duchemin SI, Bovenhuis H, Megens HJ, Van Arendonk JAM, Visker MHPW. Fine-mapping of BTA17 using imputed sequences for associations with de novo synthesized fatty acids in bovine milk. J Dairy Sci. 2017;100:9125–35.
48. Boichard D, Govignon-Gion A, Larroque H, Maroteau C, Palhiere I, Tosser-Klop G, et al. Genetic determinism of milk composition in fatty acids and proteins in ruminants, and selection potential. Prod Anim. 2014;27:283–98.
49. Knutsen TM, Olsen HG, Tafintseva V, Svendsen M, Kohler A, Kent MP, et al. Unravelling genetic variation underlying de novo-synthesis of bovine milk fatty acids. Sci Rep. 2018;8:2179.
50. Kemper KE, Reich CM, Bowman PJ, vander Jagt CJ, Chamberlain AJ, Mason BA, et al. Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy of across-breed genomic predictions. Genet Sel Evol. 2015;47:29.
51. Jiang J, Gao Y, Hou Y, Li W, Zhang S, Zhang Q, et al. Whole-genome resequencing of Holstein bulls for indel discovery and identification of genes associated with milk composition traits in dairy cattle. PLoS One. 2016;11:e0168946.
52. Li C, Sun D, Zhang S, Wang S, Wu X, Zhang Q, et al. Genome wide association study identifies 20 novel promising genes associated with milk fatty acid traits in Chinese Holstein. PLoS One. 2014;9:e96186.
53. Pausch H, Emmerling R, Gredler-Grandl B, Fries R, Daetwyler HD, Goddard ME. Meta-analysis of sequence-based association studies across three cattle breeds reveals 25 QTL for fat and protein percentages in milk at nucleotide resolution. BMC Genom. 2017;18:853.
54. Lopdell TJ, Tiplady K, Struchalin M, Johnson TJJ, Keehan M, Sherlock R, et al. DNA and RNA-sequence based GWAS highlights membrane-transport genes as key modulators of milk lactose content. BMC Genom. 2017;18:968.
55. Kemper KE, Littlejohn MD, Lopdell T, Hayes BJ, Bennett LE, Williams RP, et al. Leveraging genetically simple traits to identify small-effect variants for complex phenotypes. BMC Genom. 2016;17:858.
56. Schoonjans K, Staels B, Auwerx J. The peroxisome proliferator activated receptors (PPARs) and their effects on lipid metabolism and adipocyte differentiation. Biochim Biophys Acta. 1996;1302:93–109.
57. Coleman RA, Lee DP. Enzymes of triacylglycerol synthesis and their regulation. Prog Lipid Res. 2004;43:134–76.
58. Martin P, Leroux C. Caprine gene specifying alpha(s1)-casein: a highly suspicious factor with both multiple and unexpected effects. Prod Anim. 2000;13:125–32.
59. Li R, Dudemaine PL, Zhao X, Lei C, Ibeagha-Awemu EM. Comparative analysis of the miRNome of bovine milk fat, whey and cells. PLoS One. 2016;11:e0154129.
60. Li Z, Liu H, Jin X, Lo L, Liu J. Expression profiles of microRNAs from lactating and non-lactating bovine mammary glands and identification of miRNA related to lactation. BMC Genom. 2012;13:731.
61. Le Guillou S, Marthey S, Laloe D, Laubier J, Mobuchon L, Leroux C, et al. Characterisation and comparison of lactating mouse and bovine mammary gland miRNomes. PLoS One. 2014;9:e91938.
62. Chen X, Gao C, Li H, Huang L, Sun Q, Dong Y, et al. Identification and characterization of microRNAs in raw milk during different periods of lactation, commercial fluid, and powdered milk products. Cell Res. 2010;20:1128–37.
63. Faye LL, Machiela MJ, Kraft P, Bull SB, Sun L. Re-ranking sequencing variants in the post-GWAS era for accurate causal variant identification. PLoS Genet. 2013;9:e1003609.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC

4.6. Bilan du chapitre 4

Grâce aux dispositifs des projets *PhénoFinlait* et *From'MIR*, puissants par leur grand nombre de vaches à la fois phénotypées et génotypées, nous détectons de nombreux gènes, dont certains ont des effets forts sur les caractères de composition et de fromageabilité du lait. Parmi les gènes qui ont les effets les plus importants, nous confirmons les effets des gènes des lactoprotéines (caséines et *PAEP*) et de *DGAT1*, déjà connus. Nous montrons par ailleurs l'évolution de la fréquence de certains variants dans les gènes des lactoprotéines dans les races Montbéliarde, Normande et Holstein. En outre, nous mettons en évidence les effets d'autres gènes, dont les effets sur ces caractères sont peu ou pas connus (*SLC37A1*, *ALPL*, *MGST1*, *SEL1L3*, *GPT*, *SCD*, *GPAT4*, *FASN*, *ANKH*…).

De plus, grâce aux données du projet 1000 génomes bovins et à la puce EuroG10K, nous proposons des variants candidats dans chacun des gènes et confirmons l'effet de certains d'entre eux, soit parce qu'ils sont causaux, soit parce qu'ils sont très proches des variants causaux. Ces variants candidats expliquent une part substantielle de la variabilité phénotypique des caractères, jusqu'à 30% pour les critères fromagers (paramètres de coagulation) et jusqu'à 60% pour la composition du lait (taux des protéines sériques).

Enfin, une approche « réseaux de gènes », réalisée sur l'ensemble des caractères a permis de mettre en évidence des gènes pléiotropes, co-associés aux caractères de composition et de fromageabilité du lait, ainsi que les voies métaboliques (*e.g.* phosphate et phospholipides, transport des anions inorganiques…) et les régulateurs associés à ces gènes. Parmi eux, on trouve *PPARA*, *ASXL3* et *bta-mir-200c* qui sont fonctionnellement liés à la composition du lait.

Chapitre 5

Vers une sélection génomique des aptitudes fromagères

5. Vers une sélection génomique des aptitudes fromagères

Les caractères de composition et de fromageabilité du lait étant héritables et sous l'influence de gènes avec des effets parfois très forts, ils doivent pouvoir être sélectionnés assez facilement. Par ailleurs, avec environ 20 000 vaches de race Montbéliarde qui sont à la fois phénotypées et génotypées, le projet *From'MIR* offre une première population de référence de taille conséquente pour envisager une sélection génomique.

Dans ce chapitre, nous présentons une étude d'estimation de la précision d'une évaluation génomique sur les caractères fromagers prédits par spectrométrie MIR en race Montbéliarde. Nous évaluons ensuite le progrès génétique réalisé sur ces caractères au cours de ces 13 dernières années. Pour finir, nous simulons différents scénarios de sélection sur la fromageabilité et pour chacun de ces scénarios, nous calculons les réponses à la sélection attendues pour les caractères fromagers mais aussi pour les caractères actuellement sélectionnés en race Montbéliarde.

5.1. Estimation de la précision d'une évaluation génomique

Compte tenu du jeu de données *From'MIR* (grand nombre de vaches avec phénotypes, dont une partie génotypée), nous avons choisi d'évaluer la précision d'une évaluation génomique sur une population de validation à partir d'une méthode en une étape (SS-GBLUP) qui combine l'information apportée par le pedigree et les marqueurs moléculaires.

5.1.1. Matériel et méthodes

5.1.1.1. Données et caractères analysés

Un certain nombre de filtres ont été appliqués au jeu de données *From'MIR* qui contient au départ les performances de 4,8 millions de contrôles de 311 613 vaches (voir §2.2.5.3). Nous avons sélectionné les vaches dont l'âge au premier vêlage était compris entre 22 et 44 mois qui avaient une première lactation avec au moins 3 contrôles et des lactations complètes ou commencées au moins 90 jours avant la date de fin de récolte des données. Enfin, nous avons gardé les contrôles jusqu'à 305 jours de lactation et pour chaque caractère, les contrôles avec des prédictions comprises dans l'intervalle de ± 3 écart-types autour de la moyenne.

Au final, nous avions donc à disposition 2 869 353 contrôles de 191 532 vaches parmi lesquelles 19 564 étaient génotypées. Le pedigree a été constitué en remontant quatre

générations. Les caractères choisis sont les neuf critères fromagers les mieux prédits par la spectrométrie MIR, *i.e.* les trois rendements fromagers (CY_{FRESH}, CY_{DM} et $CY_{FAT\text{-}PROT}$), cinq paramètres de coagulation en pâte pressée cuite (a_{PCC} et $K10/RCT_{PCC}$) et pâte molle (a_{SC}, $a2_{SC}$ et $K10/RCT_{SC}$) et le pH du lait (pH_{0_PCC}) ainsi que les taux de caséines (Σ CN) et de calcium (Ca) dans le lait.

5.1.1.2. Constitution des populations d'apprentissage et de validation

A partir des 19 564 vaches avec phénotypes et génotypes, nous avons constitué une population d'apprentissage (**APP**) et une population de validation (**VAL**) aussi indépendantes que possible (***Figure 5.1***).

a) Les premières lactations (L1) des vaches génotypées ont dans un premier temps été triées par mois de vêlage entre janvier 2012 et mars 2017 (derniers contrôles en juin 2017). Les vaches primipares lors de la dernière campagne de vêlage (à partir d'août 2016) ont servi à constituer la population VAL, leurs performances ont donc été exclues des données utilisées pour l'évaluation génomique. La population APP a par conséquent été définie par les vaches primipares des campagnes précédentes. Pour mimer au mieux une sélection génomique, nous avons par ailleurs éliminé de la population APP les contrôles contemporains aux contrôles de la population VAL, *i.e.* ceux réalisés après août 2016. De plus, pour avoir dans chacune des populations APP et VAL un phénotype relativement précis, nous avons gardé les vaches avec au moins trois et cinq contrôles, respectivement. Cela a eu pour conséquence d'éliminer de la population APP les vaches ayant vêlé pour la première fois en juin et juillet 2016. Après ce premier tri, nous disposions de 13 272 vaches dans APP et 4032 vaches dans VAL, soit 17 304 vaches au total.

b) Dans un deuxième temps, des filtres complémentaires ont été appliqués en fonction de la structure familiale des deux populations. Les vaches sélectionnées à l'étape a) étaient filles de 642 taureaux différents. Parmi ces taureaux, 180 avaient des filles à la fois dans la population APP et dans la population VAL. De plus, les filles de ces 180 taureaux représentaient à elles seules 75% des vaches sélectionnées précédemment. Pour éviter d'avoir des vaches très apparentées (demi-sœurs) dans les populations APP et VAL, pouvant conduire à surestimer la précision de l'évaluation génomique, nous avons sélectionné des familles différentes dans les deux populations. Lorsqu'au moins 40% des filles d'un des 180 taureaux étaient dans la population APP (140 taureaux), nous avons gardé les filles dans cette population et éliminé les

demi-sœurs de la population VAL. A l'inverse, pour les 40 autres taureaux qui avaient moins de 40% de leurs filles dans la population APP, nous avons gardé les filles de la population VAL et éliminé les demi-sœurs qui se trouvaient dans la population APP.

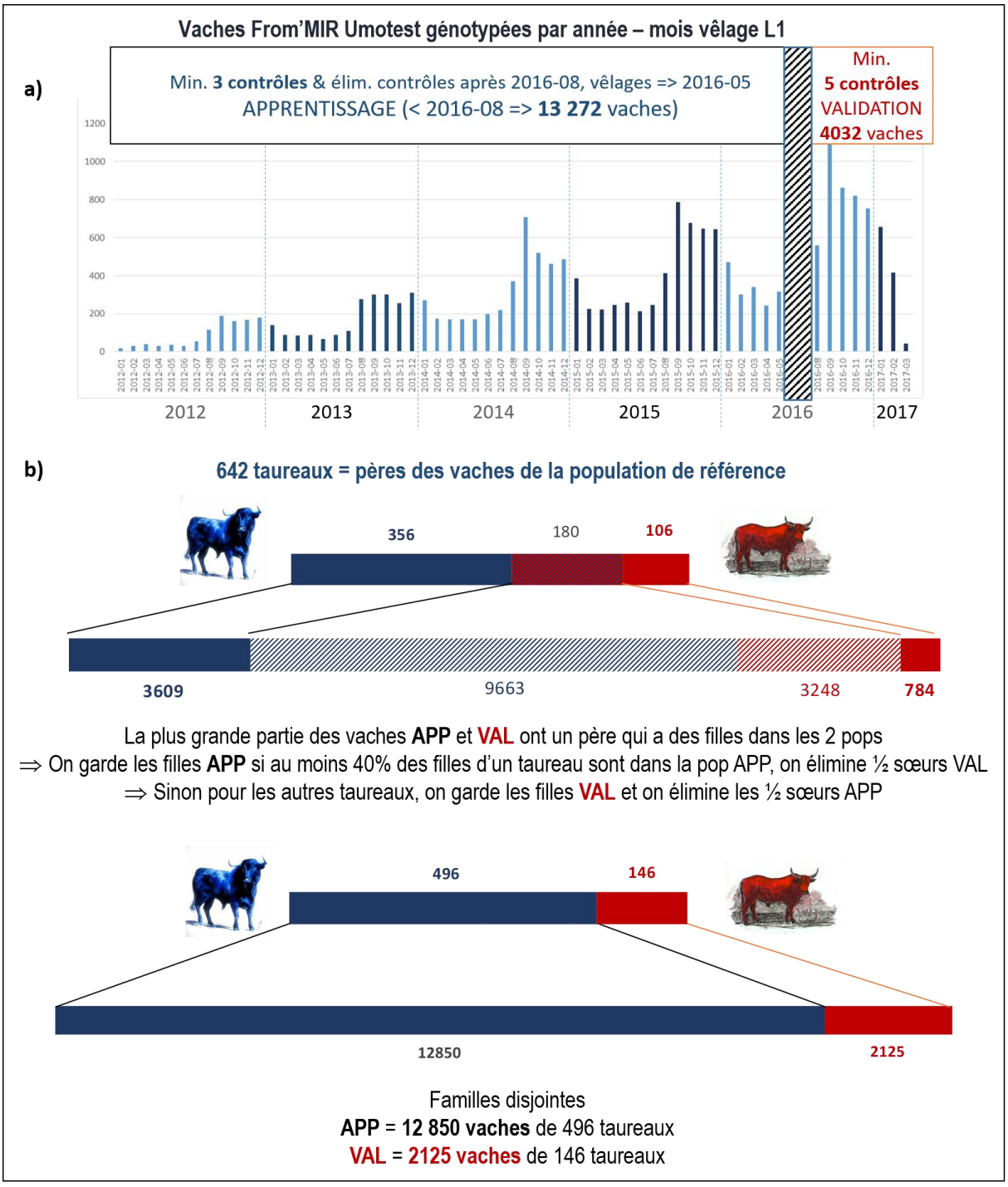


Figure 5.1. *Constitution des populations d'apprentissage (APP) et de validation (VAL) à partir des 19 564 vaches From'MIR Umotest phénotypées et génotypées*

En raison du génotypage relativement récent des vaches, nous avons observé une plus forte proportion de vaches génotypées dans la dernière campagne de vêlage. Nous avons donc pu nous permettre d'être un peu plus sévères sur le nombre de contrôles par vache et la proportion de filles par taureau pour sélectionner les vaches de la population VAL utilisées pour calculer la précision de l'évaluation génomique. Un plus grand nombre de contrôles dans la population de validation conduit à un phénotype plus précis. En conséquence, la corrélation entre prédiction génomique et phénotype de validation est plus élevée et l'estimation du CD est plus précise.

Les populations APP et VAL ont ainsi été créées avec des vaches ayant réalisé leur première lactation dans des campagnes différentes et issues de familles disjointes. *In fine*, la population APP se composait de **12 850** vaches génotypées et 147 069 vaches non génotypées, issues de 496 taureaux, et la population VAL contenait **2125** vaches génotypées, filles de 146 autres taureaux.

5.1.1.3. Méthode et logiciel d'évaluation génomique utilisés

La méthode d'évaluation génomique utilisée est la méthode dite **HSSGBLUP** (Hybrid-SS-GBLUP) décrite par Fernando *et al.* (2016) et implémentée dans le logiciel HSSGBLUP à l'INRA dans l'équipe Génétique et Génomique Bovine (Tribout, comm. pers.). C'est une méthode d'évaluation *Single Step* qui, comparativement à une méthode en deux étapes (évaluation polygénique puis génomique), combine les informations apportées par les phénotypes, le pedigree et les marqueurs pour prédire des valeurs génomiques moins biaisées et plus précises pour l'ensemble des animaux du pedigree (génotypés et non génotypés). Par rapport au modèle classique du *Single Step* (Aguilar *et al.*, 2010), la méthode HSSGBLUP présente en outre l'avantage d'être particulièrement adaptée aux populations qui ont un grand nombre d'animaux génotypés, sans approximation numérique.

5.1.1.4. Modèles testés

La méthode HSSGBLUP a été appliquée à deux types de modèles : lactation (MOY) et contrôles élémentaires (CTL). Le modèle MOY considère la performance moyenne de la vache sur la lactation tandis que le modèle CTL prend en compte les performances de la vache à chacun de ses contrôles. Chaque type de modèle a été appliqué aux données des premières lactations (1) ou aux données des trois premières lactations (3). De sorte que quatre modèles différents, incluant un nombre de vaches équivalent (légères différences dues aux filtres

appliqués aux données) mais présentant un nombre d'observations différent, ont été testés : **MOY1** (une observation par vache), **MOY3** (jusqu'à trois observations par vache, près de 2 en moyenne), **CTL1** (7 observations en moyenne par vache sur la première lactation) et **CTL3** (13 observations en moyenne sur les trois premières lactations). Les effets et les effectifs de chaque modèle sont décrits dans le ***Tableau 5.1***. Les variances génétique additive (σ_a^2), résiduelle (σ_e^2) et de l'environnement permanent (σ_p^2) utilisées dans le programme HSSGBLUP ont été estimées par le logiciel WOMBAT (Meyer, 2007) avec les 4 modèles (chapitre 3).

Tableau 5.1. *Description des modèles utilisés pour l'évaluation génomique*

	MOY1	MOY3	CTL1	CTL3
Modèle	y = Xb + Za + e	y = Xb + Za + Zp + e	y = Xb + Za + Zp + e	y = Xb + Za + Zp + e
Effets aléatoires (variances σ^2 estimées par WOMBAT)	Animal σ^2_a Résiduel σ^2_e	Animal σ^2_a Env. permanent σ^2_p Résiduel σ^2_e	Animal σ^2_a Env. permanent σ^2_p Résiduel σ^2_e	Animal σ^2_a Env. permanent σ^2_p Résiduel σ^2_e
Effets fixés **b**	Troupeau x Campagne Mois x An vêlage Age au 1er vêlage Nombre de contrôles	Troupeau x Campagne Mois x An vêlage Age au vêlage Nombre de contrôles x rang lactation	Troupeau x Date contrôle Mois x An vêlage Age au 1er vêlage Stade de lactation	Troupeau x Date contrôle Mois x An vêlage Age au vêlage Stade de lactation x rang lactation
# Vaches avec phénotypes	155 479	159 919	155 961	155 961
# Phénotypes	155 479	287 800	1 115 000	2 070 000
Pedigree # Total animaux # Dont génotypés	 413 787 21 764	 422 857 21 874	 414 919 21 747	 423 348 21 874

5.1.1.5. Densités et pondérations des marqueurs

Grâce à l'imputation *reverse* réalisée sur les vaches *From'MIR* (Sanchez *et al.*, 2018b), nous avions à disposition les génotypes de la puce 50K augmentée des SNP de la partie recherche de la puce EuroG10K. Parmi les SNP de la partie recherche, figuraient certains variants candidats (prédictifs ou causaux) pour la composition et la fromageabilité du lait mis en évidence dans les projets *PhénoFinlait* et *From'MIR* (Boichard *et al.*, 2014, Sanchez *et al.*, 2016a, Sanchez *et al.*, 2017b, Sanchez *et al.*, 2018b, Sanchez *et al.*, 2019). Dix-huit de ces variants (SNP), représentant chacun une région QTL avec de gros effets sur la composition et / ou la fromageabilité du lait, ont ainsi été sélectionnés. Selon le caractère, 5 à 14 de ces SNP présentaient des effets significatifs dans les GWAS (***Tableau 5.2***). Pour chaque caractère, les parts de variance génétique associées à chacun de ces SNP ont été calculées à partir des effets estimés dans les analyses GWAS selon la formule suivante : $\%\ \sigma_a^2 = 100\ \left(\frac{2p(1-p)\alpha^2}{\sigma_a^2}\right)$ avec σ_a^2

la variance génétique additive du caractère, *p* et *(1-p)* les fréquences alléliques et α l'effet de substitution allélique du variant.

Plusieurs jeux de marqueurs et pondérations associées à ces marqueurs ont donc été testés.

1) Les quatre modèles MOY1, MOY3, CTL1 et CTL3 ont été testés avec les SNP de la puce 50K qui avaient une MAF (*Minor Allele Frequency*) supérieure à 1‰, soit 41 942 marqueurs. La variance des effets de ces SNP est supposée constante (Scénario **50K**).

2) Sur les quatre modèles, celui qui a conduit aux valeurs génomiques les plus précises a également été testé avec les 47 794 SNP (MAF > 1‰) de la puce 50K et de la partie recherche de la puce EuroG10K i) avec des variances des effets constantes pour chaque SNP (scénario **50K+**) ou ii) avec des variances spécifiques pour les effets de 5 à 14 SNP candidats sélectionnés selon le caractère. Ces QTL expliquent de 17,5 à 58,4% de la variance génétique du caractère. Les parts de variance génétique sont attribuées à chaque SNP à partir des effets estimés par les GWAS, présentés dans le ***Tableau 5.2***. Dans ce modèle, les autres SNP (47 780 à 47 789 selon le caractère) ont des variances d'effets constantes et se partagent le reste de la variance génétique, soit 41,6 à 82,5% de la variance génétique totale selon le caractère (scénario **QTL**).

Tableau 5.2. *SNP sélectionnés à partir des résultats GWAS et % de variance génétique associée à chaque caractère dans le scénario QTL*

QTL	BTA	CY_{FRESH}	CY_{DM}	$CY_{FAT-PROT}$	a_{PCC}	$K10/RCT_{PCC}$	a_{SC}	$a2_{SC}$	$K10/RCT_{SC}$	$pH0_{_PCC}$	ΣCN	Ca	# Traits	Position SNP (pb)	MAF	Annotation fonctionnelle
1	1	0,00	0,00	0,75	1,13	0,00	0,00	0,00	0,48	7,62	0,00	0,00	4	144 395 109	0,45	intron SLC37A1
2	2	0,55	0,00	0,00	0,00	0,00	0,94	0,58	0,00	1,20	0,00	0,51	5	5 802 738	0,21	intergénique
3	2	0,00	0,00	0,00	0,39	0,00	0,00	0,00	0,00	0,00	1,34	0,00	2	131 812 821	0,37	missense ALPL
4	5	1,32	1,78	1,14	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	3	93 945 738	0,12	intron MGST1
5	5	2,76	2,72	1,23	0,00	3,89	0,41	3,55	2,64	0,00	1,93	5,38	9	117 972 265	0,04	upstream GRAMD4
6	6	2,47	0,00	3,01	4,35	0,00	2,25	0,00	1,99	12,5	4,26	0,73	8	46 751 233	0,27	downstream SLC34A2
7	6	7,70	7,25	8,70	2,34	42,7	4,31	35,8	42,1	0,00	1,53	1,32	10	87 390 612	0,40	missense CSN3
8	7	0,00	0,00	1,48	3,05	0,00	2,90	0,00	1,04	0,00	2,20	3,12	6	46 318 618	0,08	upstream ENSBTAG00000004032
9	11	8,35	10,4	23,0	10,4	2,11	2,99	2,81	3,29	0,00	16,2	0,00	9	103 303 475	0,46	missense PAEP
10	14	12,7	16,4	2,82	0,00	3,44	0,00	3,22	2,28	0,00	2,47	4,21	8	1 629 753	0,33	upstream GPT
11	16	0,45	0,45	0,00	0,26	0,58	0,58	0,47	0,57	0,00	0,52	0,00	8	60 615 012	0,08	intergénique
12	17	0,00	0,00	0,00	0,00	1,90	0,00	1,18	1,04	0,00	0,58	0,00	4	53 072 959	0,06	upstream BRI3BP
13	19	1,39	1,67	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	2	51 383 847	0,32	upstream FASN
14	20	0,41	0,00	1,34	0,56	0,00	2,41	0,00	0,87	1,02	2,33	2,19	8	58 427 343	0,07	intron ANKH
15	22	1,93	2,25	0,73	0,36	2,00	0,90	1,27	1,65	0,00	0,78	0,00	9	32 827 786	0,26	intron FAM19A4
16	24	0,46	0,48	0,00	0,00	0,55	0,62	0,41	0,48	0,00	0,00	0,00	6	58 809 468	0,21	intron LMAN1
17	26	0,50	0,59	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	2	21 144 708	0,46	missense SCD
18	27	1,83	2,60	1,72	0,00	0,00	0,00	0,00	0,00	1,20	0,00	0,00	4	36 212 352	0,41	5'UTR GPAT4
% σ^2_a totale		42,9	46,5	45,9	22,8	57,1	18,3	49,3	58,4	23,5	34,2	17,5				
h^2		0,38	0,39	0,37	0,42	0,47	0,45	0,48	0,47	0,37	0,46	0,50				
# QTL		**14**	**11**	**11**	**9**	**8**	**10**	**9**	**12**	**5**	**11**	**7**				

5.1.1.6. Estimation de la précision des valeurs génomiques

Le programme HSSGBLUP fournit une valeur génétique estimée (**VGE**) pour chaque individu du pedigree. Les performances de la population de validation n'ayant pas été prises en compte pour estimer les VGE, nous pouvons estimer la précision à partir de la corrélation entre les VGE et les performances corrigées pour les effets de milieu dans la population VAL. Les effets utilisés pour corriger les performances de la population VAL ont été estimés à l'aide du logiciel GENEKIT (Ducrocq, 2011) selon les deux modèles différents MOY1 (moyennes L1) et CTL1 (contrôles élémentaires L1) décrits dans le ***Tableau 5.1***. Pour chaque caractère, quatre corrélations ont ensuite été calculées entre

- les performances moyennes corrigées des effets de milieu par le modèle MOY1 d'une part et les VGE estimées avec le modèle MOY1 ($\mathbf{r_{MOY1}}$) et le modèle MOY3 ($\mathbf{r_{MOY3}}$) d'autre part ;
- les moyennes des performances élémentaires corrigées des effets de milieu et de l'effet d'environnement permanent par le modèle CTL1 d'une part et les VGE estimées avec le modèle CTL1 ($\mathbf{r_{CTL1}}$) et le modèle CTL3 ($\mathbf{r_{CTL3}}$) d'autre part.

Les CD (coefficients de détermination) ont été calculés en divisant le carré de la corrélation par l'héritabilité du caractère utilisé pour la validation et dérivée du chapitre 3.

5.1.2. Résultats & discussion

5.1.2.1. Comparaison des modèles

Les quatre modèles MOY1, MOY3, CTL1 et CTL3 ont été testés avec les marqueurs de la puce 50K. La ***Figure 5.2*** présente les valeurs de CD obtenues dans la population de validation pour les quatre modèles et les 11 caractères analysés.

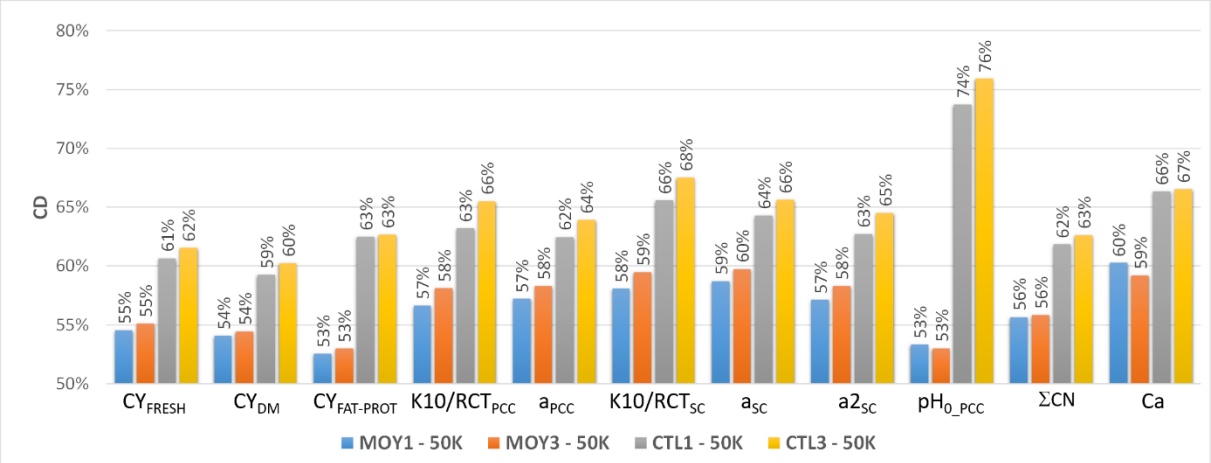


Figure 5.2. *CD estimés dans la population de validation pour les 4 modèles*

Quel que soit le modèle testé et le caractère analysé, les CD sont toujours supérieurs à 50%. Ce résultat est déjà important en lui-même puisque le seuil de 50% est le seuil de publication actuellement appliqué en race Montbéliarde pour les index génomiques des taureaux (Idele, 2018). La précision est légèrement plus élevée pour les paramètres de coagulation et de composition du lait mais dans l'ensemble, les CD sont relativement proches pour tous les caractères. Pour un type de modèle donné (MOY ou CTL), la prise en compte des performances des deuxième et troisième lactations en plus de la L1 améliore légèrement la précision (+1 point de CD en moyenne). Dans tous les cas, les modèles de type contrôles élémentaires (CTL1 et CTL3) donnent des valeurs génomiques nettement plus précises que les modèles de type lactation (MOY1 et MOY3) : + 8 points de CD en moyenne. Sur les 11 caractères, les valeurs de CD moyennes sont égales à 56, 57, 64 et 65% pour les modèles MOY1, MOY3, CTL1 et CTL3, respectivement.

5.1.2.2. Comparaison des densités et pondérations des SNP

Nous avons donc sélectionné le modèle CTL3 qui, avec les marqueurs de la puce 50K, donne les résultats les plus précis pour les 11 caractères et avons ré-estimé les valeurs génétiques en appliquant les scénarios 50K+ et QTL. La comparaison des CD pour les 11 caractères est présentée dans la ***Figure 5.3*** pour les trois scénarios 50K (*idem* Figure 5.2), 50K+ et QTL.

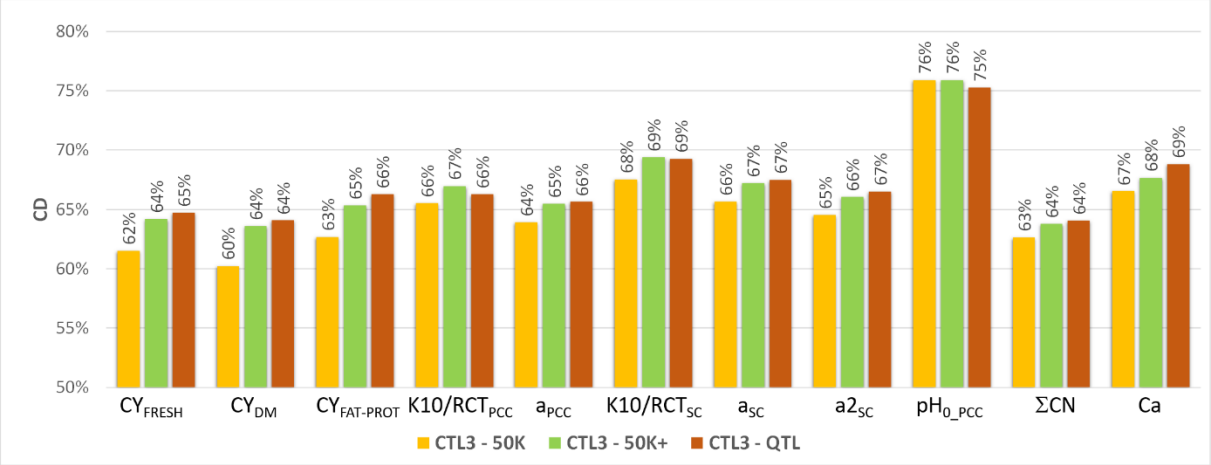


Figure 5.3. *CD estimés dans la population de validation pour le modèle CTL3 et les 3 scénarios*

Bien que nous ayons déjà atteint des bons niveaux de CD avec les SNP de la puce 50K, l'ajout des marqueurs de la partie recherche de la puce EuroG10K (scénario 50K+ : 5852 SNP supplémentaires par rapport au scénario 50K) améliore encore la précision des valeurs génomiques. En moyenne sur tous les caractères, le gain est d'environ +2 points de CD : +2,7

à +3,4 points pour les rendements fromagers et +1,5 à +1,9 points pour les paramètres de coagulation, les caséines et le calcium dans le lait. Le scénario QTL, qui attribue des pondérations plus élevées aux QTL, ne permet qu'un gain supplémentaire marginal de 0,25 point en moyenne et cette valeur cache des disparités selon les caractères (+0,2 à +1,1 point pour a_{PCC}, ΣCN, a_{SC}, $a2_{SC}$, CY_{DM}, CY_{FRESH}, $CY_{FAT\text{-}PROT}$ et Ca, respectivement et -0,1 à -0,7 point pour $K10/RCT_{SC}$, pH_{0_PCC} et $K10/RCT_{PCC}$, respectivement).

En revanche, les scénarios ne sont pas classés de la même façon lorsqu'on s'intéresse aux pentes de la droite de régression des performances corrigées sur les VGE (***Figure 5.4***). Pour tous les caractères et les trois scénarios, les pentes sont inférieures à 1, ce qui indique un biais dans l'estimation des VGE. Le scénario qui génère les VGE les moins biaisées est le scénario QTL, les deux autres scénarios sont équivalents avec toutefois des résultats un peu meilleurs pour le scénario 50K.

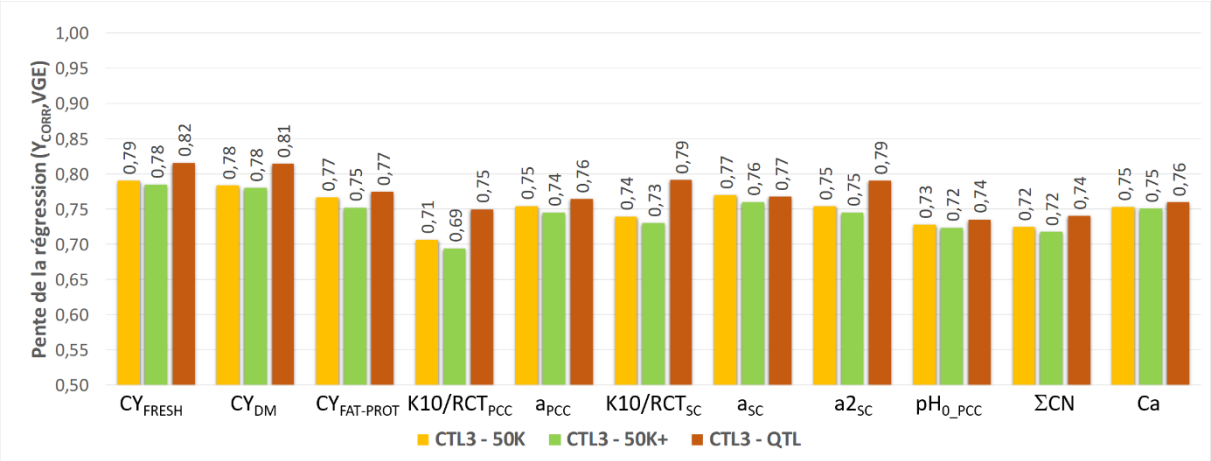


Figure 5.4. *Pentes de la régression des performances corrigées (Y_{CORR}) sur les valeurs génétiques (VGE) pour le modèle CTL3 et les 3 scénarios*

En plus des variants candidats identifiés dans les projets *PhénoFinlait* et *From'MIR*, la partie recherche de la puce EuroG10K est particulièrement enrichie en SNP dans les régions QTL détectés pour les caractères de production laitière notamment (Boichard *et al.*, 2018). Parmi ces régions QTL, on trouve par exemple les régions des gènes *DGAT1*, des caséines, *PAEP*… qui ont également des effets forts sur les caractères fromagers. Cela explique probablement le gain de précision obtenu avec le scénario 50K+ par rapport au scénario 50K. On constate d'ailleurs que le gain est le plus fort pour les rendements fromagers qui sont très corrélés génétiquement au TB (Sanchez *et al.*, 2018a) et donc particulièrement sous l'influence de la région du gène

DGAT1 (Sanchez *et al.*, 2019). Les estimations des effets des marqueurs dans cette région (entre 0 et 3 Mb sur le chromosome 14) sont représentées dans la ***Figure 5.5*** pour le rendement en extrait sec et les trois scénarios.

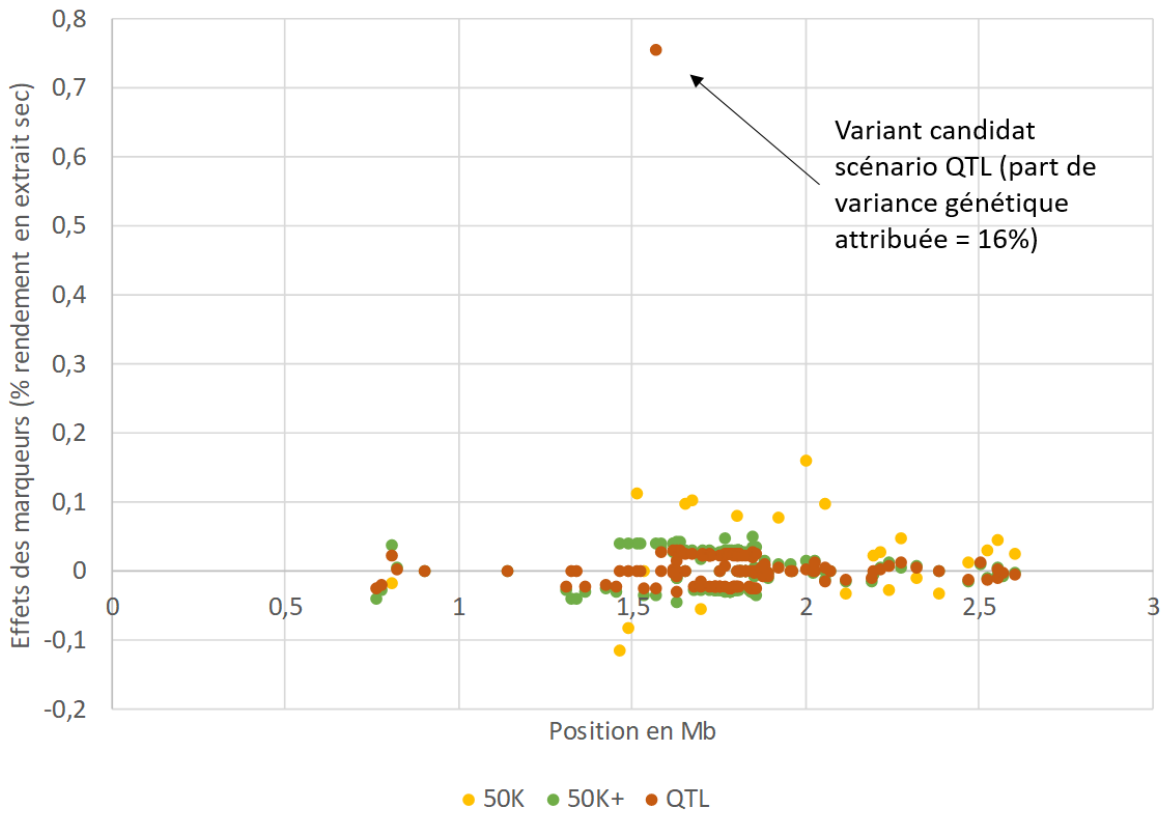


Figure 5.5. *Estimation des effets des SNP dans la région du gène DGAT1 pour le rendement en extrait sec CY_{DM} et les 3 scénarios 50K, 50K+ et QTL*

On remarque tout d'abord un fort enrichissement de la région en SNP sur la partie recherche de la puce EuroG10K (153 SNP dans le scénario 50+ contre 25 SNP dans le scénario 50K). Par ailleurs, exprimés en valeur absolue de l'unité du caractère (points de rendement en extrait sec), les effets moyens des marqueurs de la région estimés dans le scénario 50K sont plus élevés (0,052) que ceux estimés dans le scénario 50K+ (0,021) ou QTL (0,019). Dans le scenario 50+, l'effet de la région chromosomique se répartit sur un plus grand nombre de marqueurs, avec des effets individuels plus faibles. Comme attendu, dans le scénario QTL, le variant candidat auquel nous avons attribué 16% de la variance génétique du caractère a de loin les plus gros effets sur le caractère (0,75 point de rendement), les 152 autres SNP de la région ayant alors un effet faible. Indépendamment des poids associés aux SNP, le fait d'enrichir la région en marqueurs a pour effet de mieux capter les effets de la mutation causale et d'augmenter la

précision. Inclure la mutation causale avec un poids approprié donne un résultat assez similaire en termes de précision, l'effet fort de la mutation causale conduisant à la réduction des effets des marqueurs proches. Mais si le fait d'ajouter des pondérations plus fortes aux variants candidats entraine un gain de précision relativement faible, il permet en revanche d'estimer des valeurs génétiques moins biaisées.

Les modèles et scénarios comparés dans cette étude montrent que les valeurs génétiques les plus précises et les moins biaisées sont obtenues :
i) avec un modèle de type contrôles élémentaires appliqué aux trois premières lactations, qui utilise le maximum d'information et une modélisation plus précise du phénotype ;
ii) avec les génotypes aux marqueurs de la puce 50K et aux marqueurs de la partie recherche de la puce EuroG10K, option dans laquelle les marqueurs sont à la fois plus nombreux et relativement concentrés dans les régions QTL ;
iii) avec une pondération plus forte pour les variants causaux ou prédictifs.

5.2. Estimation du progrès génétique réalisé sur les caractères fromagers

Nous avons ensuite estimé le progrès génétique réalisé sur les critères fromagers dans la population de taureaux et de vaches de race Montbéliarde de l'entreprise Umotest.

5.2.1. Matériel et méthodes

5.2.1.1. Modèle et estimation des valeurs génétiques

Une nouvelle évaluation a été réalisée à l'aide du logiciel HSSGBLUP (§5.1.1.3) avec le modèle CTL3 (contrôles élémentaires des trois premières lactations ; §5.1.1.4) appliqué à l'ensemble des données (tous les phénotypes conservés). Nous avions ainsi 2,6 millions d'observations de 190 261 vaches, tout le reste étant par ailleurs égal au scénario CTL3-50K testé précédemment (423 348 animaux dans le pedigree, dont 21 874 génotypés pour 41 942 marqueurs). Les effets des marqueurs étant estimés directement par la méthode HSSGBLUP, cette analyse nous a fourni les formules de prédiction des valeurs génétiques à partir des génotypes aux marqueurs de la puce 50K.

5.2.1.2. Animaux et génotypages

La disponibilité des génotypages (puce 50K ou puce LD imputée 50K) pour un grand nombre d'animaux de race Montbéliarde de l'entreprise de sélection Umotest (21 171 mâles et 311 761 femelles au total en janvier 2019 ; ***Figure 5.6***) nous a permis d'estimer le progrès génétique réalisé sur les taureaux et les vaches pour les 11 caractères fromagers décrits dans le §5.1.1.1. Pour estimer le progrès génétique de ces caractères, nous avons appliqué les équations de prédiction des valeurs génétiques sur les génotypes des animaux contemporains des individus du pedigree *From'MIR*, *i.e.* les taureaux nés entre 2005 et 2018 et les vaches nées entre 2008 et 2018.

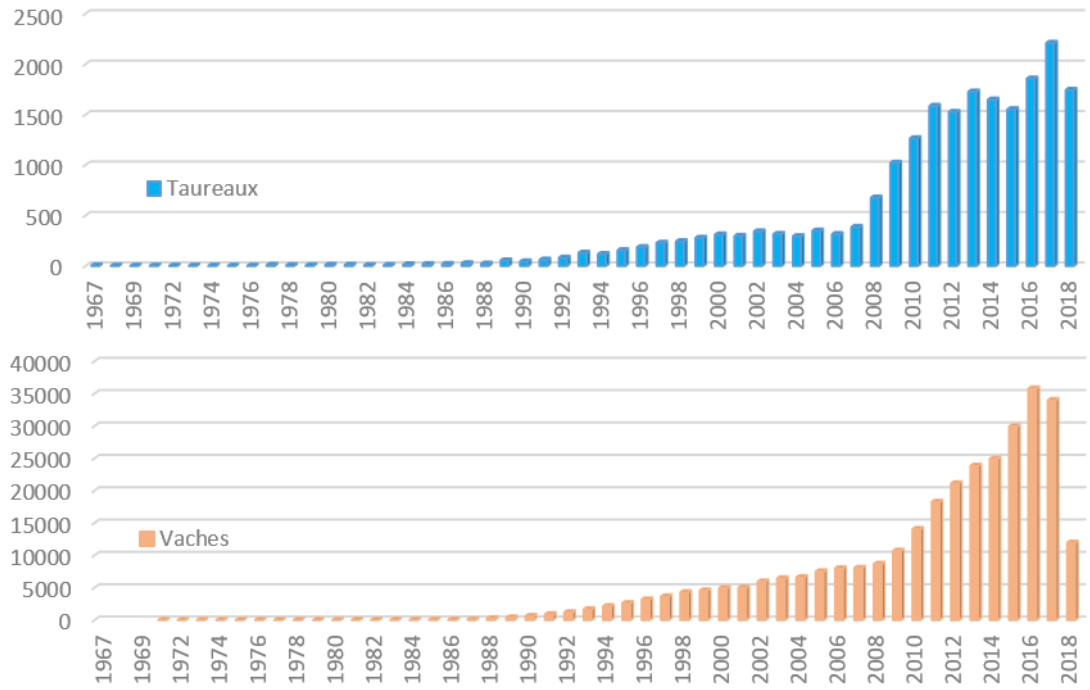


Figure 5.6. *Nombre de taureaux et de vaches de race Montbéliarde (Umotest) génotypés par année de naissance*

5.2.2. Résultats

Les courbes d'évolution génétique réalisée sur les mâles et les femelles sont dans la ***Figure 5.7*** pour les 11 caractères. On note une amélioration régulière des valeurs génétiques de tous les critères fromagers sur la période 2005-2018, *i.e.* en moyenne une augmentation des valeurs génétiques de tous les caractères sauf évidemment celles du rendement en MSU ($CY_{FAT\text{-}PROT}$) et des K10/RCT (inverse de la vitesse d'organisation) que l'on cherche à diminuer. Dans la zone AOP Comté, le niveau génétique moyen des taureaux et des vaches de race Montbéliarde

(animaux génotypés Umotest) pour les rendements fromagers et les paramètres de coagulation est donc aujourd'hui plus élevé qu'en 2005.

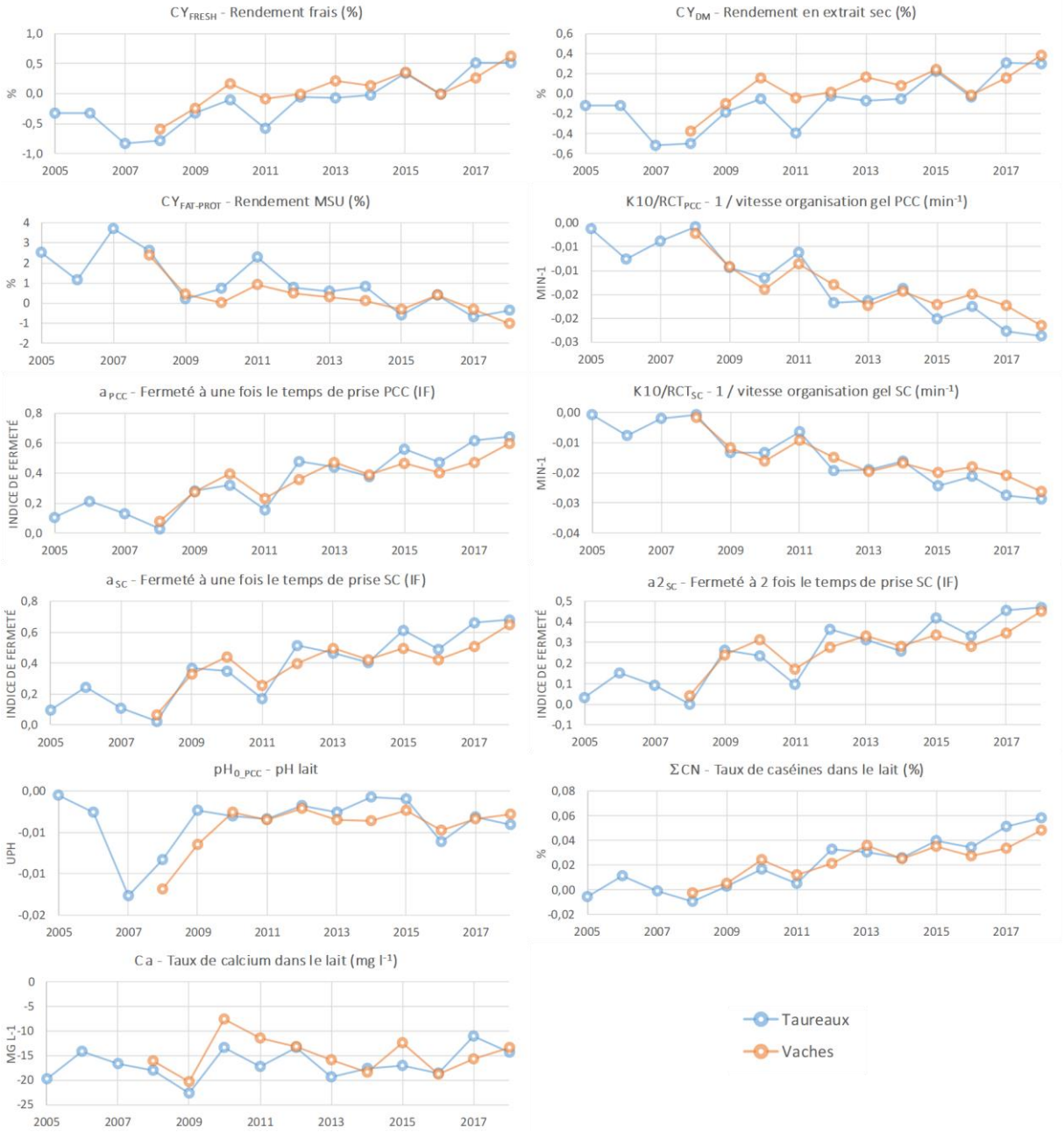


Figure 5.7. *Courbes de l'évolution génétique réalisée par année de naissance des taureaux nés entre 2005 et 2018 et des vaches nées entre 2008 et 2018 pour les 11 caractères fromagers*

L'écart entre les périodes extrêmes (2005-2018 pour les taureaux et 2008-2018 pour les vaches) a été exprimé en écart-type génétique (ET_G) du caractère (***Figure 5.8***). Les paramètres de coagulation et le taux de caséines dans le lait sont les caractères qui évoluent le plus fortement durant la période, entre 0,5 et 0,6 ET_G selon le caractère. L'évolution est un peu plus modérée

pour les rendements (entre 0,2 et 0,4 ET_G). Pour le pH du lait, les résultats ne sont pas cohérents entre les mâles et les femelles mais cette différence s'explique par la diminution observée entre 2005 et 2007 chez les taureaux (Figure 5.7). Quant au taux de calcium, son évolution est beaucoup plus lente que celle des autres caractères (+0,1 ET_G dans la population de taureaux).

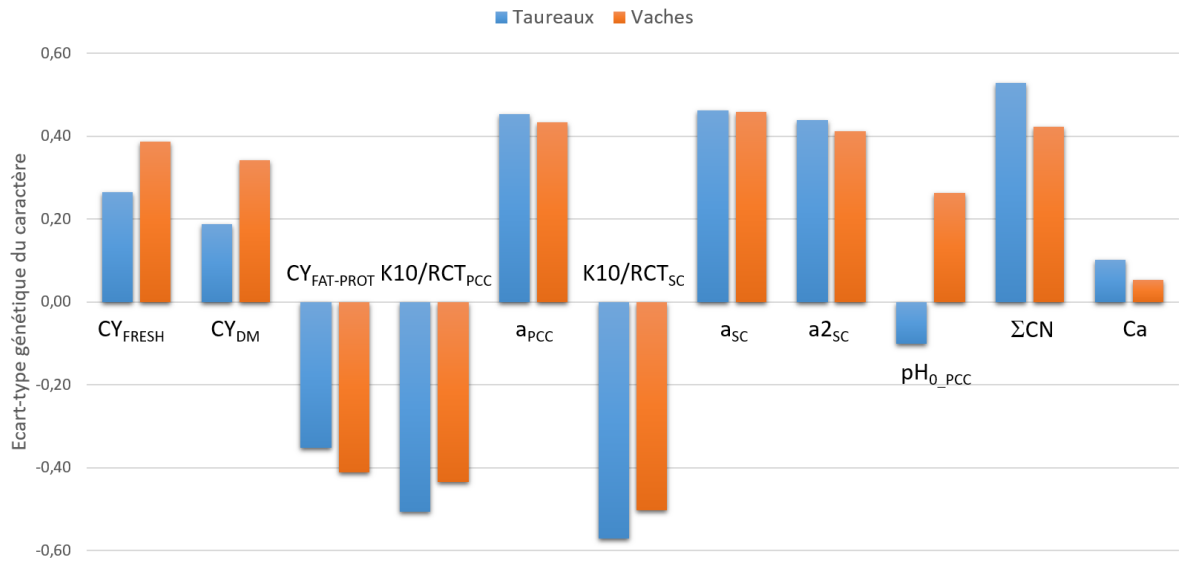


Figure 5.8. *Différences entre les valeurs génétiques estimées entre 2018 et 2005 pour les taureaux et entre 2018 et 2008 pour les vaches pour les 11 caractères fromagers (en écart-type génétique du caractère)*

Les résultats présentés ici sont cohérents avec la sélection pratiquée en race Montbéliarde. Comme vu au §1.6.4.1, l'index économique laitier (SYNTLAIT) donne un poids très important au TP et cet index représente à lui seul 45% de la pondération économique dans l'index de synthèse global (ISU) de la race. On observe d'ailleurs une réponse forte sur les caséines qui représentent environ 80% des protéines totales du lait. En sélectionnant sur le TP en race Montbéliarde, on a donc dans le même temps amélioré les caractères fromagers et en particulier les paramètres de coagulation qui sont les plus corrélés génétiquement au TP (Sanchez *et al.*, 2018a).

5.3. Etudes des potentialités de sélection des caractères fromagers

Suite à ces résultats, nous avons initié une étude pour évaluer l'intérêt de prendre en compte les caractères fromagers dans l'objectif de sélection. Ce travail, réalisé dans le cadre d'un stage de fin d'études que j'ai co-encadré entre janvier et août 2018 (Beaumont, 2018), comportait plusieurs étapes.

5.3.1. Concertation avec un groupe d'experts

Avant d'envisager de définir un nouvel objectif de sélection incluant les caractères fromagers, il était nécessaire de consulter les différents acteurs de la filière AOP Comté pour déterminer avec eux le choix des caractères à améliorer ainsi que le consentement à payer, *i.e.* le prix que les fromagers / affineurs sont prêts à payer aux éleveurs pour un lait plus fromageable. Ce dernier point est crucial pour réaliser toute étude économique. Le consentement à payer permet de répartir les bénéfices apportés par la sélection sur les caractères fromagers à tous les niveaux de la filière (du producteur de lait à l'affineur de fromage) pour que le système aboutisse au résultat escompté, une amélioration de la fromageabilité du lait.

Dans un premier temps, une concertation a été organisée avec 10 experts de la filière AOP Comté qui étaient des techniciens, fromagers ou affineurs choisis pour leur bonne connaissance de la transformation fromagère et du modèle économique de la filière. Le groupe d'experts a été réuni en région Franche-Comté puis certains d'entre eux ont été consultés par téléphone individuellement. Cette étape a pu bénéficier du soutien du service ASTRE (Approches Sociales et TRavail en Elevage) de l'Institut de l'Elevage pour tout ce qui concernait la préparation, l'organisation et l'animation des réunions et des interviews.

Le travail de hiérarchisation de l'importance des caractères fromagers demandé aux experts a permis de faire ressortir les caractères fromagers les plus importants pour la filière AOP Comté, à savoir le rendement en extrait sec, et la fermeté et la vitesse d'organisation du caillé en modèle pâte pressée cuite. Tous les experts interviewés par téléphone s'accordaient sur le fait qu'il était important de mettre en place une valorisation économique des caractères fromagers, et il leur paraissait possible notamment de dévaloriser un peu le TP au profit des caractères fromagers. Cependant, les experts n'ont pas été en mesure de définir un consentement à payer précis, invoquant les effets que pourrait avoir la sélection des caractères fromagers étudiés dans *From'MIR* (caractéristiques technologiques liées à la physico-chimie du lait) sur la qualité finale des produits, particulièrement importante en filière AOP.

5.3.2. Simulation de scénarios de sélection

Nous avons tiré parti de la consultation des acteurs de la filière pour simuler un nouvel objectif de sélection avec les trois caractères fromagers mis en avant par les experts. Nous voulions ainsi

évaluer dans quelle mesure il était possible d'accroître davantage le gain génétique sur les caractères fromagers et estimer l'impact relatif sur les autres caractères.

5.3.2.1. Scénarios de sélection

Trois scénarios ont été simulés.

1) Le premier scénario correspond à la situation actuelle, la sélection est réalisée sur l'index de synthèse global calculé ainsi :

ISU = 0,45 SYNTLAIT + 0,145 STMA + 0,18 REPRO + 0,05 LGF + 0,05 VTRA + 0,125 MO
avec

- l'index économique laitier SYNTLAIT = 1,050 (MP + 0,1 MG + 3 TP + 0,5 TB) qui combine les index des matières protéique (MP) et grasse (MP) ainsi que ceux des taux protéique (TP) et butyreux (TB) ;
- l'index de synthèse de santé de la mamelle STMA = 0,6 CEL + 0,4 MACL qui combine les index des comptages cellulaires (CEL) et des mammites cliniques (MACL) ;
- l'index de synthèse de la fertilité REPRO = 0,5 FER + 0,25 FERG + 0,25 IVIA1 qui combine les index fertilité (FER), fertilité génisse (FERG) et l'intervalle vêlage première insémination (IVIA1) ;
- l'index de la longévité fonctionnelle LGF ;
- l'index de la vitesse de traite VTRA ;
- l'index de synthèse de la morphologie MO = 0,4 Mamelle + 0,3 Corps + 0,15 Aplombs + 0,10 Bassin + 0,05 Aptitude Bouchère.

2) Le deuxième scénario ajoute l'index de la teneur en caséines dans le lait (CAS) à l'ISU actuel, nous l'avons défini avec les pondérations suivantes :

ISU-COMP = 0,7 ISU + 0,3 CAS

3) Le troisième scénario donne le même poids à l'ISU que précédemment et répartit les 30% restant à parts égales aux trois caractères fromagers priorisés par les experts : le rendement en extrait sec (CY_{DM}) et les deux paramètres de coagulation pour les fromages de type pâte pressée cuite (la fermeté du gel a_{PCC} et l'inverse de la vitesse d'organisation du gel $K10/RCT_{PCC}$) :

ISU-FROM = 0,7 ISU + 0,1 (CY_{DM} + a_{PCC} - $K10/RCT_{PCC}$)

Dans ces deux derniers scénarios, le poids accordé aux nouveaux caractères est particulièrement élevé, probablement nettement supérieur à ce qui peut être envisagé. Ces scénarios donnent donc une tendance extrême.

A noter que pour avoir des résultats directement comparables pour tous les caractères, tous les index élémentaires et les index de synthèse ont été exprimés en écart-type génétique.

5.3.2.2. Estimation des réponses à la sélection

Pour chacun des scénarios, nous avons simulé une sélection par troncature en utilisant les valeurs génétiques des taureaux calculées dans le paragraphe précédent pour les caractères fromagers (§5.2) et les index génomiques de l'évaluation officielle pour les autres caractères qui sont dans l'objectif de sélection. Pour avoir un effectif de taureaux suffisant par année de naissance, nous avons gardé les taureaux nés entre 2009 et 2017 (***Figure 5.6***) et mimé une stratégie de sélection comparable à celle pratiquée aujourd'hui par l'entreprise Umotest en race Montbéliarde. Chaque index de synthèse global ISU, ISU-COMP et ISU-FROM a été calculé pour tous les taureaux. Pour chaque scénario, les taureaux ont ensuite été triés sur cet index par année de naissance et les 80 meilleurs taureaux ont été choisis chaque année parmi l'ensemble des candidats disponibles (environ 1600 en moyenne). En plus de l'index de synthèse global, nous avions à disposition pour tous les taureaux les index génomiques officiels pour l'ensemble des caractères évalués en routine. Nous avons donc calculé les réponses à la sélection sur les 11 caractères fromagers, les index élémentaires MP, MG, TP, TB, CEL, MACL et LAIT ainsi que les index de synthèse ISU, STMA, REPRO, MO et VB. L'index élémentaire LAIT correspond au caractère quantité de lait, l'index de synthèse VB est l'index de valeur bouchère et les autres index sont décrits plus haut. Dans chaque scénario et pour chaque index, les réponses à la sélection ont été estimées annuellement en calculant la différentielle de sélection, *i.e.* la différence entre la moyenne des index des 80 meilleurs taureaux et la moyenne des taureaux candidats. La moyenne des réponses annuelles a ensuite été calculée sur les neuf années de la période (2009-2017). Les index étant exprimés en écart-type génétique, les réponses à la sélection sont également exprimées dans la même unité.

5.3.2.3. Résultats

Les réponses à la sélection sur tous les index sont présentées sur la ***Figure 5.9*** pour les trois scénarios.

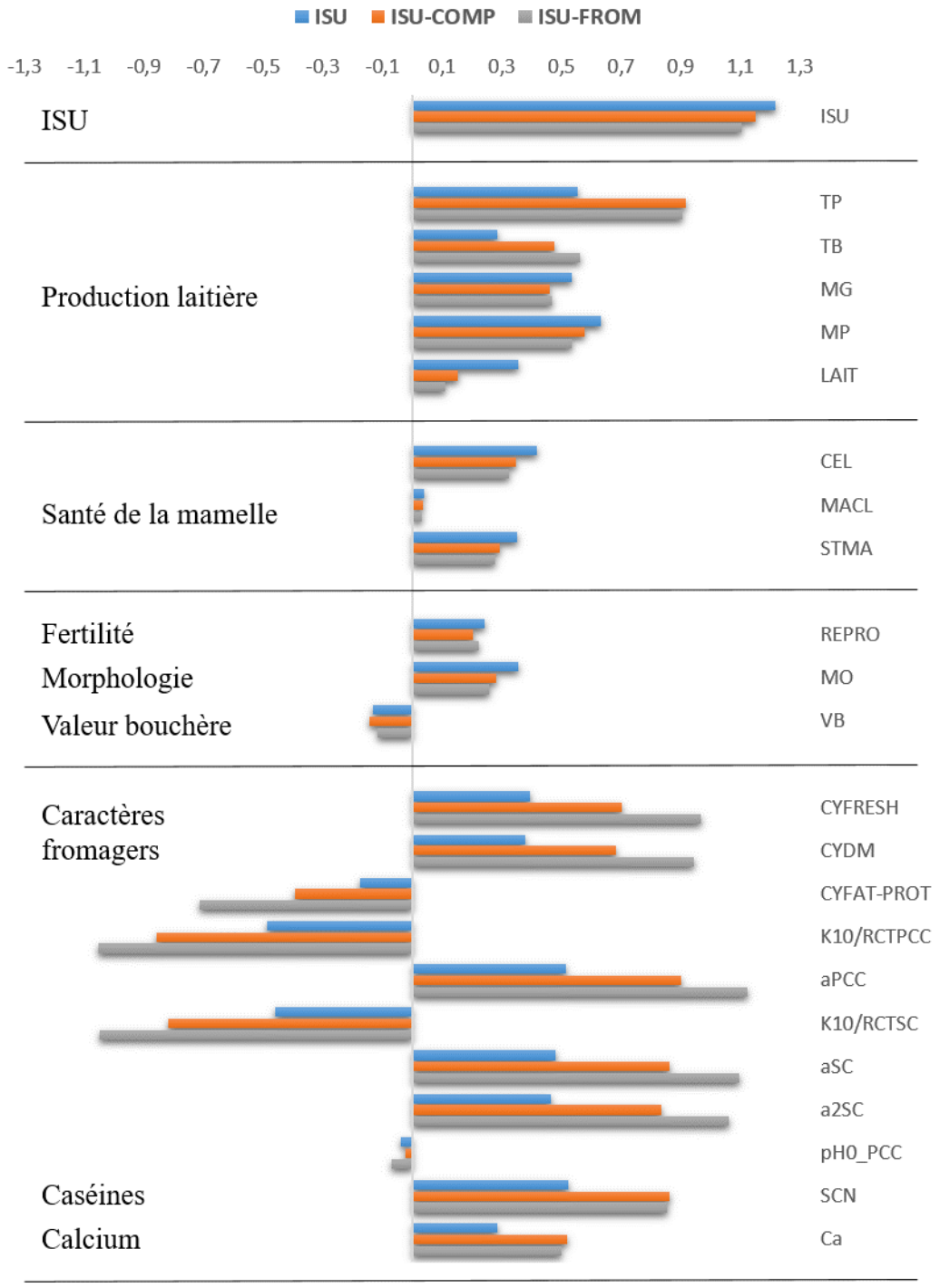


Figure 5.9. *Réponses à la sélection sur ISU, ISU-COMP et ISU-FROM estimées sur les index fromagers et les index des autres caractères (en écart-type génétique)*

Comme attendu, l'ajout de nouveaux critères dans l'objectif de sélection entraine des variations de réponse pour tous les index de façon plus ou moins marquée. Pour quasiment tous les caractères, le scénario ISU-COMP entraine des niveaux de réponse intermédiaires par rapport aux scénarios ISU et ISU-FROM.

Pour les caractères autres que les caractères fromagers, comparés à la situation actuelle (ISU), les deux scénarios alternatifs entrainent :

1) une réponse amoindrie pour les quantités de matières protéique et grasse et surtout de lait mais une meilleure réponse sur les taux (TP et TB) ;
2) une réponse un peu plus faible sur les caractères liés à la santé de la mamelle ;
3) une réponse quasi-équivalente pour les index de synthèse de fertilité et de valeur bouchère mais une réponse un peu plus faible sur la synthèse morphologie.

Toutefois, les résultats négatifs sont à relativiser dans la mesure où l'ajout des caséines ou des caractères fromagers dans l'objectif de sélection entraine une diminution du gain de l'ordre de 0,1 ET_G pour MP, MG, CEL, MACL, STMA et MO et de l'ordre de 0,2 ET_G pour le LAIT. Ces valeurs sont relativement limitées, compte tenu du poids élevé accordé au taux de caséines ou aux aptitudes fromagères dans les index alternatifs.

En revanche, le fait de sélectionner les taureaux sur l'ISU-COMP et *a fortiori* sur l'ISU-FROM entraine des niveaux de réponse beaucoup plus forts sur l'ensemble des caractères fromagers étudiés dans *From'MIR*, *i.e.* tous les rendements fromagers et tous les paramètres de coagulation mesurés sur les fromages de type pâte pressée cuite et pâte molle. Par exemple, pour le rendement en extrait sec, le gain est presque doublé avec le scénario ISU-COMP et presque triplé avec le scénario ISU-FROM par rapport au scénario ISU. On passe d'un gain génétique de 0,38 avec l'ISU à un gain de 0,68 et 0,95 ET_G pour les scénarios ISU-COMP et ISU-FROM, respectivement. Les gains sur les autres caractères fromagers sont du même ordre de grandeur. Les taux de caséines et de calcium dans le lait sont également nettement améliorés de manière à peu près équivalente avec les deux scénarios alternatifs. Pour le taux de calcium par exemple, on passe d'une différentielle de 0,28 ET_G avec la sélection sur l'ISU à un progrès de 0,50 et 0,52 ET_G si on sélectionne sur ISU-FROM ou sur ISU-COMP, respectivement.

Cette étude montre donc qu'en ajoutant les caractères fromagers dans l'objectif de sélection, il est possible d'obtenir un gain génétique nettement plus conséquent pour les caractères fromagers et sur certains caractères de composition du lait (TP, TB mais aussi caséines et calcium) qu'en sélectionnant sur l'ISU actuel et ce, tout en ne dégradant pas trop le progrès génétique des autres caractères.

5.4. Bilan du chapitre 5

La précision des valeurs génomiques de la fromageabilité du lait est tout à fait cohérente avec la précision théorique attendue eu égard aux valeurs d'héritabilité et à la taille de la population de référence utilisée (Goddard et Hayes, 2009). Avec ce niveau de précision, comparable aux précisions obtenues pour d'autres caractères actuellement sélectionnés (*e.g.* caractères de production laitière), il est tout à fait possible de mettre en place dès à présent une évaluation génomique sur les caractères fromagers. D'autant que, pour avoir une estimation rigoureuse de la précision, nous avons choisi de partitionner la population de référence en une population d'apprentissage et une population de validation aussi indépendantes que possible. Cette approche a eu pour effet de réduire fortement la taille de la population utilisée pour estimer les effets des marqueurs (12 850 vaches sur les 19 564 vaches de la population de référence). La prise en compte de l'ensemble des vaches phénotypées et génotypées devrait donc permettre d'accroître encore la précision de prédiction des valeurs génomiques. De plus, le nombre de vaches phénotypées et génotypées augmente rapidement.

Nous montrons également ici un progrès génétique des rendements fromagers et des paramètres de coagulation au cours des 13 dernières années. Du fait des corrélations génétiques fortes entre les taux protéique et butyreux d'une part et les caractères fromagers d'autre part (Sanchez *et al.*, 2018a), ces derniers ont probablement été sélectionnés indirectement avec les taux, et tout particulièrement avec le taux protéique qui a un poids économique relativement important dans l'index de synthèse racial (ISU) en race Montbéliarde.

Enfin, un nouvel objectif de sélection qui combinerait les caractères présents dans l'ISU actuel aux trois caractères fromagers identifiés comme étant les plus importants dans la filière AOP Comté permettrait d'améliorer plus rapidement l'ensemble des caractères fromagers (rendements, paramètres de coagulation pour les fromages à pâte pressée cuite mais aussi pour les fromages à pâte molle) et les caractères de composition du lait (TP, TB, taux de caséines et de calcium dans le lait) avec un impact limité sur le progrès génétique des autres caractères présents dans l'objectif, avec cependant un certain impact sur la quantité de lait.

Ces résultats encourageants ont conduit les partenaires du projet *From'Mir* à demander une évaluation génomique pilote dans la zone AOP Comté. Un projet, dans lequel sont également impliqués l'INRA et la société GenEval, aujourd'hui en charge des évaluations génétiques bovines, est donc en cours. Il bénéficiera de tous les résultats obtenus dans cette thèse : choix

des caractères et des modèles, paramètres génétiques, variants candidats de la composition et de la fromageabilité disponibles sur la puce EuroG10K, estimation des effets des variants (GWAS) pour leurs pondérations dans les modèles d'évaluation génomique… Ainsi, l'évaluation génomique pilote des aptitudes fromagères du lait des vaches Montbéliarde de la région Franche-Comté pourra être réalisée dès cet automne (2019) pour *in fine* réaliser des évaluations en routine en 2020.

Chapitre 6

Discussions, conclusions et perspectives

6. Discussions, conclusions et perspectives

Les résultats ayant déjà fait l'objet de discussions dans les articles qui figurent dans ce manuscrit, nous faisons dans le chapitre 6 un bref rappel des principales conclusions et nous les resituons dans le contexte actuel de la sélection en France mais également dans un contexte de recherche plus large.

6.1. Apports de *PhénoFinlait* et *From'MIR*

Les travaux réalisés dans cette thèse ont pu bénéficier des dispositifs puissants de deux projets de recherche dont l'objectif était d'étudier la composition et la fromageabilité du lait prédites à partir des spectres MIR produits en routine dans le cadre du contrôle laitier. Le premier projet, *PhénoFinlait*, s'est intéressé à la composition fine du lait des trois principales races bovines laitières françaises Holstein, Montbéliarde et Normande. Dans le cadre de cette thèse, nous avons étudié les teneurs en lactoprotéines prédites à partir de plus de 900 000 spectres MIR d'environ 160 000 vaches. Dans le second projet *From'MIR*, nous avions à disposition non seulement la composition fine du lait (protéines, acides gras, minéraux, citrate et lactose) mais aussi la fromageabilité du lait en lien avec sa physico-chimie (rendement et coagulation) pour plus de six millions de spectres MIR issus de plus de 400 000 vaches Montbéliarde de la région Franche-Comté (zone AOP Comté).

Nous avons dans un premier temps testé des analyses de régression bayésiennes pour prédire les aptitudes fromagères du lait à partir des spectres MIR et nous avons montré sur le jeu de données *From'MIR* que ces analyses aboutissaient à des prédictions moins précises que les analyses de régression PLS (*Partial Least Square*) classiquement utilisées (El Jabri *et al.*, 2019) et mises en œuvre ensuite pour la réalisation du projet.

Les jeux de données volumineux et originaux des deux projets nous ont ensuite permis d'estimer les paramètres génétiques de manière très précise avec différents modèles. Nous mettons ainsi en évidence :

- des valeurs d'héritabilités fortes pour les teneurs en lactoprotéines dans les trois races Montbéliarde, Normande et Holstein (Sanchez *et al.*, 2017a) ;
- des valeurs d'héritabilité modérées à fortes pour les caractères fromagers et de composition fine du lait (protéines, acides gras, minéraux, citrate et lactose) sur la population de vaches Montbéliarde de la zone AOP Comté (Sanchez *et al.*, 2018a) ;

- des corrélations génétiques favorables entre tous les caractères fromagers : rendements, vitesse et fermeté du caillé pour les fromages de type pâte pressée cuite (ex. Comté) ou pâte molle (ex. Camembert) (Sanchez *et al.*, 2018a) ;
- des corrélations génétiques fortes entre les caractères fromagers et la composition du lait et plus particulièrement entre les rendements fromagers et la teneur en matière grasse ainsi qu'entre les paramètres de coagulation et la teneur en protéines et en minéraux (calcium et phosphore) du lait (Sanchez *et al.*, 2018a) ;
- un déterminisme génétique relativement constant au cours de la lactation et entre lactations ;
- aucun antagonisme génétique entre les caractères fromagers et les caractères actuellement dans l'objectif de sélection en race Montbéliarde.

Une partie des vaches avec phénotypes ayant été génotypée (environ 2300 à 3000 vaches par race dans *PhénoFinlait* et 20 000 vaches dans *From'MIR*), il a été possible de rechercher les régions du génome affectant la composition et la fromageabilité du lait. Ces analyses, qui ont pu bénéficier des données de séquences du projet 1000 génomes bovins ainsi que de la puce du consortium *Eurogenomics* (EuroG10K), nous ont permis :
- de mettre en évidence des régions du génome, avec parfois des effets très forts sur les teneurs en protéines du lait et souvent partagées entre les trois races Montbéliarde, Normande et Holstein (Sanchez *et al.*, 2016a) ;
- de détecter des gènes et des variations (variants) dans ces gènes que nous avons pu ajouter sur la partie recherche de la puce EuroG10K (Sanchez *et al.*, 2017b), une méthode de choix pour confirmer leurs effets à grande échelle et les utiliser en sélection ;
- de confirmer les effets des variants détectés pour les teneurs en protéines dans le projet *PhénoFinlait* sur la population de vaches de race Montbéliarde du projet *From'MIR*, relativement peu dépendante des vaches du projet *PhénoFinlait* (Sanchez *et al.*, 2018b) ;
- d'identifier des régions du génome sur les caractères de fromageabilité du lait en race Montbéliarde dont certaines étaient communes à celles détectées pour la composition en protéines (Sanchez *et al.*, 2019) ;
- de mettre en évidence un réseau de 736 gènes co-associés aux caractères de composition et de fromageabilité du lait qui révèle des voies métaboliques et des gènes régulateurs fonctionnellement liés aux caractères étudiés (Sanchez *et al.*, 2019).

Encouragés par ces résultats, nous avons dans un dernier temps étudié les possibilités de sélectionner les caractères fromagers dans la population de vaches Montbéliarde de la zone AOP Comté :

- nous montrons qu'une évaluation en une étape (single step) qui estime les valeurs génétiques de l'ensemble des vaches (génotypées et / ou phénotypées) conduit à des résultats tout à fait satisfaisants, nous avons ainsi pu mesurer les gains liés à l'utilisation des données répétées, à l'utilisation d'un modèle sur contrôles élémentaires et à l'ajout de marqueurs, y compris de variants supposés causaux (en moyenne sur 11 caractères, CD = 67%) ;
- nous constatons par ailleurs une amélioration génétique des caractères fromagers durant ces 13 dernières années, probablement liée à la forte pondération économique attribuée au taux protéique dans l'objectif de sélection en race Montbéliarde ;
- en simulant différents scénarios de sélection, nous montrons que l'ajout des caractères fromagers dans l'objectif de sélection permettrait d'améliorer nettement le gain génétique de la fromageabilité du lait avec un impact limité sur le gain génétique des autres caractères actuellement sélectionnés en race Montbéliarde.

6.2. Qualité des fromages produits à partir d'un lait plus fromageable

Le projet *From'MIR* s'est focalisé sur les qualités technologiques du lait pour la fabrication de fromages. Cependant, la qualité des fromages finis est un critère essentiel, tout spécialement dans un contexte AOP. Elle se détermine après une période d'affinage et dépend des conditions de milieu, et tout spécialement de la flore microbienne. Cette qualité du produit ne peut pas être prédite simplement à partir de la composition du lait et il est difficile de produire des équations MIR de la qualité finale des fabrications individuelles. Or, un risque potentiel, redouté par la filière, est qu'une modification des caractéristiques de fromageabilité se traduise par une altération de la qualité du produit fini.

Une étude complémentaire a donc été réalisée dans *From'MIR* en conditions de mini-fabrications pour i) vérifier la cohérence entre les rendements fromagers mesurés en laboratoire et ceux mesurés en conditions réelles de fabrication et ii) évaluer l'effet d'une amélioration de ces caractères sur les qualités sensorielles ou organoleptiques des fromages (texture et goût), *via* un jury de dégustation. Des laits avec des taux protéiques équivalents et des rendements en extrait sec de laboratoire supérieurs à la médiane de la population ont été sélectionnés pour fabriquer deux types de fromages à pâte molle (type Camembert au lait pasteurisé, 2 L de lait

par fromage) et à pâte pressée cuite (type Comté au lait cru affiné sept mois, 120 L de lait par fromage). Les laits de troupeaux collectés au cours de 2 périodes (hiver et été), 3 jours distincts et dans 3 troupeaux différents (+ mélange des 3 troupeaux) ont été transformés en 24 fromages de chaque type (48 mini-fabrications au total).

Cette étude permet tout d'abord de valider la méthode de référence utilisée dans *From'MIR* pour mesurer les rendements de laboratoire. De plus, l'analyse sensorielle des fromages, réalisée par un jury de dégustation, ne révèle aucun défaut de goût pour les deux types de fromage et aucun défaut de texture sur les fromages à pâte molle alors que quelques défauts sont observés sur les fromages à pâte pressée cuite (Laithier, comm. pers.). Ces résultats sont bien évidemment à confirmer sur un plus grand effectif et en conditions de terrain. Ils soulèvent néanmoins la question de la relation entre les aptitudes technologiques du lait et les qualités organoleptiques du fromage.

Enfin, le projet *From'MIR* a étudié la fromageabilité en lien avec les propriétés physico-chimiques du lait. Or, la flore microbienne du lait cru (bactéries, levures, moisissures…) a un rôle crucial dans la transformation en fromage, en particulier sur ses qualités organoleptiques (Montel *et al.,* 2014). Le microbiote du lait impacte donc tout particulièrement les fromages fabriqués à partir de lait cru comme le Comté. Ce point, en particulier la relation hôte-microbiote en lien avec la qualité des fromages, n'a pas pu être adressé dans cette étude, ni dans cette thèse.

6.3. Les suites de *From'MIR*

Les résultats des analyses génétiques du projet *From'MIR* ont été diffusés à la communauté scientifique internationale au travers des articles scientifiques qui figurent dans ce manuscrit et de présentations en congrès internationaux (voir liste des publications en fin de manuscrit) mais aussi à tous les professionnels (éleveurs, fromagers, affineurs…). Ils ont ainsi été présentés lors de deux journées de restitution des résultats à l'échelle régionale (26 avril 2018, Dannemarie-sur-Crête, Doubs) et nationale (28 juin 2018, Paris) et une présentation a eu lieu lors des dernières journées Rencontres Recherche Ruminants (3R) (Gaudillière *et al.*, 2018). De plus, des Newsletters et des fiches techniques ont été régulièrement rédigées et diffusées au cours de l'avancement du projet et plusieurs articles sont parus dans la presse régionale ou technique (***Annexe 4***).

6.3.1. Au niveau régional

Les résultats très encourageants du projet de recherche *From'MIR*, ainsi que tous les efforts de vulgarisation et de diffusion qui lui ont été associés, ont permis de créer une dynamique assez forte au sein de la filière laitière franc-comtoise.

Ainsi aujourd'hui, une évaluation génomique pilote des caractères fromagers est en cours de mise en place. Elle est prévue à l'automne 2019 pour une évaluation en routine dès 2020. Parallèlement à cette évaluation, un observatoire régional de la fromageabilité du lait (*Observalait*) est en cours de création en Franche-Comté. Ses objectifs sont notamment d'entretenir les équations de prédiction avec l'apport des nouvelles données et d'exploiter les spectres MIR pour assurer un suivi en routine du lait à toutes les échelles (individuelle, troupeau et cuve fromagère) jusqu'au fromage affiné (qualité, goût…). A moyen terme, tous les acteurs de la filière (éleveurs, fromagers et affineurs) disposeront donc des outils de phénotypage, de génétique et des conclusions apportées par l'observatoire régional.

6.3.2. Au niveau national

D'autres régions en France sont également grandes productrices de fromages, et notamment de fromages AOP. Les équations développées pour *From'MIR* ne sont probablement pas directement applicables à d'autres populations. Les prédictions sont en effet affectées par les races et les systèmes de production. Toutefois, les résultats acquis suscitent déjà l'intérêt d'autres filières fromagères (autres races ou autres espèces) qui voudraient mettre en place un projet analogue (région AOP Camembert en race Normande, ovins laitiers…). Ils pourront bénéficier de l'expérience et des résultats de *From'MIR* avec éventuellement une possibilité à terme de mettre en commun les populations de calibration des différentes races bovines pour améliorer la précision des équations de prédiction de la fromageabilité.

6.3.3. Et au-delà

Une collaboration est en cours avec la Belgique qui a également développé des équations de fromageabilité dans le cadre du projet *ProFARMilk*. Appliquées aux mêmes spectres MIR, les équations des deux projets donnent des prédictions fortement corrélées, notamment pour les rendements de laboratoire. Les travaux se poursuivent et un des objectifs, à terme, pourrait être le renforcement de la robustesse des équations de prédiction.

6.4. Quel impact économique aurait un lait plus fromageable ?

Durant le stage de M. Beaumont, un dialogue conduit avec la filière fromagère de Franche Comté a permis de classer les critères par ordre d'importance mais n'a pas permis de déterminer un consentement à payer. En l'absence de ces valeurs, il n'a pas été possible de réaliser une simulation bio-économique détaillée des exploitations pour déterminer les poids économiques relatifs des caractères dans l'objectif de sélection.

Cependant, les suites du projet *From'MIR* montrent que tout est en marche pour à moyen terme inclure les caractères fromagers dans l'objectif de sélection en race Montbéliarde. Les premières actions seront vraisemblablement axées sur la mise à disposition des index mâles et femelles et de l'observation de l'impact des effets génétiques pour les différents acteurs de la filière. Une redéfinition éventuelle de l'objectif de sélection ne sera décidée que dans un second temps. Pour cela, il faudra une définition consensuelle à l'échelle de la filière des poids économiques des caractères et donc de la valorisation des aptitudes fromagères au niveau de l'éleveur. Les critères de coagulation n'affectent que le processus de fabrication, il revient aux fromagers d'indiquer s'ils souhaitent une évolution. Un gain de rendement affecterait davantage l'économie de la filière, en réduisant la quantité de lait nécessaire par kg de fromage et en modifiant la productivité. Avec la formule de paiement actuelle du lait, qui valorise les taux butyreux et protéique, une sélection sur le rendement fromager, qui induirait une hausse des taux (+1,1g de TP et +2,1g de TB pour +2,2% de rendement en extrait sec), se traduirait par une valorisation supérieure du prix du litre de lait (13,2€ / 1000L) et une répartition assez équilibrée de la plus-value entre l'éleveur et le fromager. Toutefois, ce raisonnement ne prend pas en compte les réponses sur les autres caractères, en particulier sur la quantité de lait. Dans le cadre franc-comtois avec un prix élevé du litre de lait par rapport à ses composants, le gain pour l'éleveur est beaucoup plus réduit et l'objectif de sélection devra être défini pour minimiser cette perte de quantité et satisfaire les besoins des différents acteurs.

6.5. Sélection sur la fromageabilité dans les autres pays

Malgré l'intérêt économique potentiel de l'amélioration des caractères fromagers, à notre connaissance, très peu de pays intègrent aujourd'hui les caractères fromagers dans leur objectif de sélection. Les Etats-Unis calculent un index dit fromager (*Dollars Cheese Merit Index*, CM$) mais qui, tout comme l'ISU Montbéliard actuel, donne simplement un poids plus important au taux protéique (VanRaden, 2004). En Belgique, le projet *ProFARMilk* (2011-2017) a étudié les

aptitudes technologiques du lait prédites par la spectrométrie MIR pour la transformation en fromage, en yaourt et en beurre (Colinet *et al.*, 2013). Ce projet a abouti à la mise en place d'un outil de suivi des aptitudes à la transformation du lait dans le cadre du contrôle laitier en Wallonie pour dans un premier temps, écarter les laits non coagulants et à terme, permettre la sélection de ces critères (AWE, 2018). L'Italie, qui transforme plus de 80% du lait qu'elle produit en fromage, utilise depuis de nombreuses années déjà des indicateurs des aptitudes fromagères (*e.g.* le taux de caséines ou le comportement lactodynamique du lait mesuré au Formagraph) dans la grille de paiement du lait dans la région AOP Parmigiano-Reggiano (Malacarne *et al.*, 2004). De plus, certains paramètres de coagulation prédits par la spectrométrie MIR sont enregistrés en routine depuis 2011 (De Marchi *et al.*, 2012, Pretto *et al.*, 2012). Un index (*Cheese Aptitude Index*) qui combine la vitesse (RCT) et la fermeté (a30) de coagulation avec des poids équivalents est publié depuis janvier 2012 pour les taureaux Holstein de la région Veneto (http://www.intermizoo.com/research/cheese-aptitude-index).

En France, sans que la sélection ne soit appliquée sur ces critères, nous avons observé des évolutions parfois assez fortes des fréquences de certains variants des lactoprotéines. On peut supposer qu'il s'agit de réponses indirectes à la sélection sur les taux. Il convient d'informer sur ces évolutions et de déterminer si elles sont favorables ou non. Ainsi, l'augmentation du variant A de la β-lactoglobuline augmente le taux de protéines sériques et ne devrait pas être favorable à la fromageabilité. L'effet de certains variants est controversé, comme par exemple le variant A2 de la caséine β, supposé mieux toléré chez une partie des consommateurs. Compte tenu de ces allégations santé (vraies ou supposées) et de la valeur ajoutée possible, des productions de niche se développent ou sont envisagées avec des génotypes particuliers.

6.6. Les spectres MIR du lait

6.6.1. Un outil de phénotypage fin pour la sélection

En France, et plus largement dans le monde, les laboratoires d'analyse du lait utilisent la spectrométrie MIR pour prédire les constituants majeurs du lait (protéines, matière grasse, lactose et urée) depuis les années 1970 (Biggs, 1978). Ces dernières années, plusieurs études ont montré le potentiel de cet outil pour prédire la composition fine du lait en acides gras (Soyeurt *et al.*, 2006), minéraux (Soyeurt *et al.*, 2009) et protéines (Ferrand *et al.*, 2012) et donc ses qualités nutritionnelles et technologiques (Dal Zotto *et al.*, 2008). Plusieurs revues

rapportent ainsi les différentes utilisations de la spectrométrie MIR dans le lait (De Marchi *et al.*, 2014, Gengler *et al.*, 2016, De Marchi *et al.*, 2018).

La composition du lait étant sous l'influence d'un grand nombre de facteurs, son évolution au cours de la lactation est le reflet de l'état général de la vache et notamment de son bien-être, de son statut sanitaire ou métabolique *via* la composition en acides gras (balance énergétique), en corps cétoniques (acétone et β-hydroxybutyrate, indicateurs d'acétonémie) ou en lactoferrine (glycoprotéine du lait impliquée dans les mécanismes immunitaires et inflammatoires) mais également de son impact sur l'environnement (efficacité alimentaire, émissions de méthane…).

Face au nouveau contexte économique, sociétal et environnemental qui oriente vers une production axée sur la qualité des produits, la robustesse des animaux et la réduction de leur impact sur l'environnement, la spectrométrie MIR apparait donc comme un outil de choix pour prédire ces nouveaux caractères à grande échelle et envisager de les inclure dans les objectifs de sélection (Boichard et Brochard, 2012).

De nombreux projets de recherche se sont ainsi développés ces dernières années, citons par exemple les projets européens *RobustMilk* (2008-2012) (Veerkamp *et al.*, 2013) et *GplusE* (2014-2018) (Grelet *et al.*, 2019). Ces études montrent qu'à partir d'un simple spectre MIR du lait, il est possible de prédire différents caractères liés à la physiologie, au métabolisme et à l'état sanitaire de la vache : ingestion et efficacité alimentaire (McParland *et al.*, 2014), acétonémie (Grelet *et al.*, 2016), fertilité (Bastin *et al.*, 2016), santé (Veerkamp *et al.*, 2013, Bastin *et al.*, 2016) et émissions de méthane (Vanlierde *et al.*, 2018). Et bien que les composants fins du lait soient des indicateurs, ces études montrent que ces caractères complexes sont prédits plus précisément à partir des spectres MIR directement.

6.6.2. La génétique du spectre MIR

Plutôt que de s'intéresser aux phénotypes prédits indirectement par les spectres MIR, certains auteurs ont étudié la génétique du spectre MIR (longueurs d'onde) directement. Dagnachew *et al.* (2013), qui ont imaginé cette approche dans l'espèce caprine, ont montré qu'elle pouvait conduire à des valeurs génétiques estimées (EBV) plus précises que celles estimées indirectement à partir des prédictions MIR. La ***Figure 6.1*** représente les différences entre les deux approches.

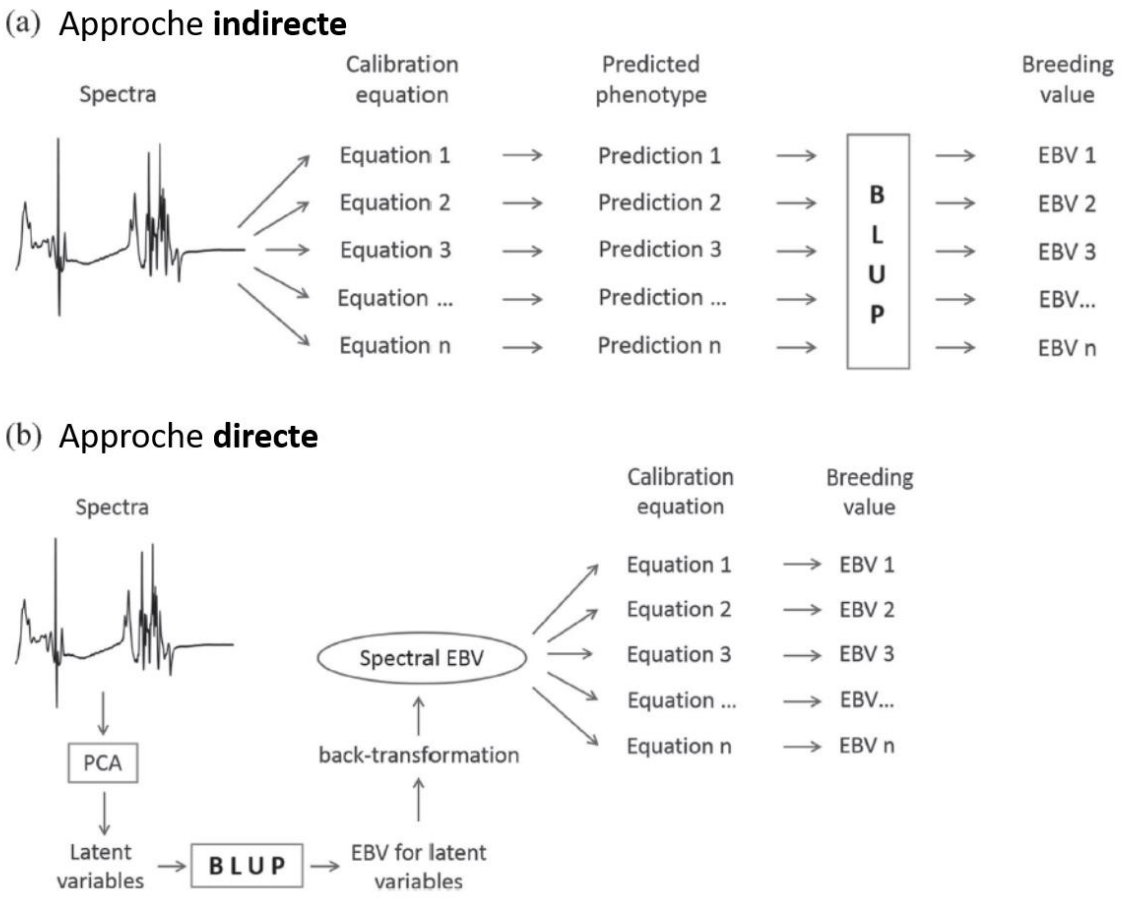


Figure 6.1. *Comparaison des approches indirecte (a) et directe (b) pour estimer les valeurs génétiques (EBV) à partir des spectres MIR, d'après Bonfatti et al. (2017b)*

Dans l'approche **indirecte** (a), qui est celle communément utilisée, les phénotypes sont prédits à partir des spectres MIR, *via* les équations de prédiction, et les EBV sont estimées sur ces prédictions. L'approche **directe** (b) estime les EBV des variables latentes du spectre (*i.e.* les variables identifiées par une analyse en composantes principales qui expliquent la quasi-totalité de la variabilité du spectre), généralement peu nombreuses (< 10). Les EBV des variables latentes sont ensuite rétro-transformées pour obtenir les EBV de toutes les longueurs d'onde du spectre. Enfin les équations de prédiction sont appliquées directement sur ces EBV pour estimer les valeurs génétiques des phénotypes.

Bonfatti *et al.* (2017b) ont ainsi comparé les valeurs génétiques estimées (EBV) à partir des longueurs d'onde directement à celles estimées à partir des prédictions MIR pour différents caractères technologiques (coagulation et rendements fromagers) et de composition (protéines, acides gras et calcium) du lait. Dans leur étude, l'approche directe ne donne des bons résultats

que pour les caractères très corrélés aux taux protéique et butyreux et ces résultats sont probablement liés à la méthode utilisée pour réduire la dimension des spectres.

Sous réserve de trouver des méthodes appropriées pour résumer au mieux toute l'information du spectre, Soyeurt *et al.* (2010) suggèrent que l'utilisation directe des spectres pourrait permettre une meilleure gestion de certains désordres métaboliques (acétonémie, mammites…) au niveau des troupeaux.

6.6.3. Remarques sur les prédictions MIR

Les prédictions MIR sont des prédictions indirectes d'un phénotype quelconque associé à une modification de composition du lait induisant des variations d'absorbance. Dans les situations les plus favorables, le phénotype d'intérêt est lui-même responsable des variations d'absorbance et les équations sont alors très robustes. Mais avec la multiplication des phénotypes prédits par MIR, cette situation devient plus rare et les prédictions sont de plus en plus souvent indirectes. Elles dépendent alors fortement de l'association entre le phénotype d'intérêt et la modification de composition du lait induisant la variation d'absorbance. Quand cette association est forte et systématique, les prédictions sont généralisables à d'autres échantillons obtenus dans des conditions différentes (autres races, autres systèmes de production…). Mais quand l'association n'est pas systématique, la prédiction peut devenir très imprécise ou très biaisée. Il est essentiel, lorsque les conditions varient, de valider l'équation de prédiction.

Les résultats ne sont pas présentés dans ce manuscrit car ils ne sont pas interprétables mais nous avons estimé les héritabilités de l'ensemble des 24 caractères fromagers prédits par la spectrométrie MIR, y compris ceux qui étaient très mal prédits ($0,10 < R^2 < 0,90$). On pourrait supposer que l'héritabilité baisse avec la qualité de prédiction, dans la mesure où l'on introduit une erreur croissante. Or, les estimations d'héritabilité de tous les caractères sont relativement proches et dépendent peu de la précision des équations. On peut donc conclure que les prédictions MIR correspondent à quelque chose d'héritable, même quand elles sont peu corrélées au caractère à prédire.

Les méthodes MIR sont donc très attractives par leur puissance élevée et leur coût réduit. Mais cela ne doit pas cacher leurs limites qu'il faut garder en tête. A chaque fois que c'est possible, il est préférable de valider les équations sur de nouveaux échantillons de référence. Par ailleurs, nos études ont également fourni des estimations de corrélations génétiques, des QTL et des

réseaux de gènes interprétables et logiques, en rapport avec la biologie des caractères. Ces résultats confortent notre analyse que ces prédictions sont fiables et utilisables à grande échelle.

6.7. Identification des mutations causales

Dans ce travail de thèse, nous mettons en évidence un certain nombre de gènes candidats et des variants candidats dans ces gènes avec des effets très significatifs qui expliquent une proportion relativement importante de la variabilité génétique de la composition fine et de la fromageabilité du lait. Identifier les effets de ces variants dans plusieurs populations ou plusieurs races permet de les confirmer et d'exclure le fait qu'ils soient des faux-positifs. Par ailleurs, l'approche réseau de gènes met en évidence des voies métaboliques et des gènes régulateurs cohérents avec la biologie des caractères analysés. Toutefois, nous ne savons pas aujourd'hui si les variants candidats sont causaux ou en déséquilibre de liaison avec les variants causaux. Or, l'identification des mutations causales, responsables des effets sur les caractères d'intérêt, représente un défi majeur à la fois pour comprendre les mécanismes impliqués dans le déterminisme de ces caractères mais aussi à des fins de sélection génomique pour accroître la précision et la persistance des prédictions génomiques ou pour la prise en compte d'effets non additifs et d'interactions épistatiques. Nous verrons dans ce paragraphe comment d'une part, une meilleure connaissance du génome bovin et de son fonctionnement et d'autre part, des analyses d'expression et / ou fonctionnelles peuvent nous permettre d'identifier ces mutations causales.

6.7.1. Vers une meilleure connaissance du génome bovin et de son fonctionnement

Les travaux réalisés dans cette thèse se sont appuyés sur l'assemblage de la séquence bovine UMD3.1, disponible depuis fin 2009. Or, depuis avril 2018, un nouvel assemblage bovin ARS-UCD1.2 est disponible. Les technologies de séquençage ayant beaucoup évolué en près de neuf ans, la nouvelle version est plus complète et précise. L'ordre des variants sur le génome étant plus juste, on peut supposer que l'imputation sur la séquence sera plus précise. Les données de séquences du projet 1000 génomes bovins sont donc en cours de réalignement avec le nouvel assemblage et le run7 sera disponible au printemps 2019. Par rapport au run précédent, le run7 contiendra donc des séquences de meilleure qualité pour un nombre accru de bovins, environ 4100 du genre *Bos* (espèces *Taurus* et *Indicus*) de plus de 100 races ou types génétiques (Schnabel *et al.*, 2019). Cette population de référence, nettement enrichie par rapport au run6

(2333 bovins de 70 races), permettra d'améliorer la qualité des imputations au niveau de la séquence et donc de faciliter la mise en évidence des variants causaux.

Généralement, parce qu'il nécessiterait une analyse spécifique, le chromosome X (hémizygote chez les mâles XY) n'est pas inclus dans les analyses pangénomiques. Ce chromosome représente pourtant 3 et 6% du génome des taureaux et des vaches, respectivement. En s'appuyant sur le nouvel assemblage, une étude très récente a identifié précisément une région pseudo-autosomale de 5,7 Mb sur le chromosome X qui pourrait donc être ajoutée aux autosomes dans les analyses d'imputation, de GWAS et de sélection génomique (Johnson *et al.*, 2019) comme pour les autosomes. Par ailleurs, les caractères fromagers n'étant exprimés que par la femelle qui porte deux chromosomes X, une analyse d'association standard serait également possible, après adaptation des procédures d'imputation. Une telle approche permettrait d'étudier une région du génome bovin, encore peu explorée jusqu'à aujourd'hui pour des raisons pratiques et peu scientifiques.

De plus, la plupart des variants étudiés jusqu'à présent sur le génome bovin sont des variants liés à la variation d'un seul nucléotide (SNP) ou des petites insertions/délétions. D'autres variants structuraux sont présents sur le génome : les longues insertions/délétions, les CNV (*Copy Number Variation*) qui sont des fragments d'ADN répétés, les translocations, les inversions… Ces variants sont moins nombreux que les SNP mais la probabilité qu'ils aient un effet biologique est plus forte. Leur identification sur le génome bovin, actuellement en cours (Boussaha *et al.*, 2016, Letaief *et al.*, 2017), reste complexe et difficile à systématiser, ce qui limite leur utilisation à grande échelle. Leur typage par puce nécessite beaucoup de mise au point mais constitue une voie prometteuse pour des analyses systématiques d'imputation, d'analyses GWAS et de prédiction génomique (Mesbah-Uddin *et al.*, 2018).

La mise en commun de données des séquences à l'échelle internationale, ainsi que les avancées technologiques en matière de séquençage du génome, nous apportent une connaissance de plus en plus précise et exhaustive des variations de la séquence du génome bovin. Toutefois, cette information ne permet pas à elle seule de décrypter le fonctionnement du génome. En effet, si il est relativement facile de prédire la conséquence d'une variation dans une région codante d'un gène, le logiciel SIFT (Kumar *et al.*, 2009) permet par exemple de prédire l'impact d'une mutation non synonyme responsable du changement d'un acide aminé de la protéine, il est beaucoup plus difficile de prédire l'effet des variations dans les autres régions du génome

(introns, régions en amont (*upstream*) ou en aval (*downstream*) des gènes, région intergénique…). On cite très souvent dans la littérature l'exemple du gène *DGAT1* qui code l'enzyme *Diacylglycérol Acyltransférase I* qui catalyse la dernière étape de synthèse des triglycérides dans le lait. Une double mutation dans un exon de ce gène entraine la substitution lysine / alanine en position 232 (*K232A*) de la protéine, ce qui a pour effet de modifier la vitesse maximale de l'enzyme (Grisart *et al.,* 2002). Cette mutation a permis d'expliquer les effets du QTL identifié au préalable dans la région de ce gène (terminaison centromérique du chromosome 14) pour le taux butyreux du lait (Coppieters *et al.*, 1998).

Toutefois, cet exemple d'identification d'une mutation causale, pourtant déjà ancien, reste une exception, en partie parce que la plupart des QTL identifiés ne sont pas associés à des variations dans les parties codantes des gènes. En effet, tout comme nous le montrons pour les caractères de composition et de fromageabilité du lait, les variants avec les effets les plus significatifs dans les analyses GWAS sont la plupart du temps localisés dans des régions non codantes du génome bovin et donc probablement dans des régions régulatrices. Sans une bonne connaissance de ces régions, il est difficile de cibler le variant causal parmi tous les variants en déséquilibre de liaison dans la région QTL. Il est possible, comme nous l'avons fait, de réaliser des analyses *in silico* à partir de bases de données d'annotation qui sont alimentées au fil du temps par de grands projets internationaux, tels le consortium ENCODE (*Encyclopedia of DNA Elements*) pour le génome humain (ENCODE Project Consortium, 2012) ou le consortium FAANG (*Functional Annotation of Animal Genomes*) pour les génomes des principales espèces d'élevage, dont le bovin (Giuffra *et al.*, 2019).

6.7.2. Analyses d'expression pangénomiques et analyses fonctionnelles ciblées

Un variant localisé dans une région régulatrice peut avoir un effet sur le niveau d'expression d'un gène (transcription en ARN puis traduction en protéine) et affecter un phénotype. La régulation peut se faire localement (*cis*-régulation) si le variant régulateur est situé à moins d'une mégabase du gène cible ou à distance (*trans*-régulation) si le variant régulateur est situé à plus de cinq mégabases du gène cible, voire sur un autre chromosome (Westra et Franke, 2014). Les méthodes de séquençage de l'ARN (*RNA-Seq*), qui permettent de mettre en évidence ces différences d'expression, peuvent être utilisées pour rechercher des QTL d'expression sur l'ensemble du génome. Chamberlain *et al.* (2018) définissent trois types de QTL d'expression comme étant des variants associés i) à l'expression totale d'un gène (eQTL pour *expression*

QTL), ii) à une expression différentielle des deux copies du gène en fonction de l'allèle porté (aseQTL pour *allele specific expression QTL*) et iii) à des modifications de l'épissage du transcrit et donc à des isoformes différents (sQTL pour *splicing QTL*). L'expression d'un gène étant spécifique d'un tissu, le choix du tissu à analyser se fait en fonction du caractère étudié. Pour les caractères de composition et de fromageabilité du lait, on rechercherait les QTL d'expression dans le lait ou dans les cellules de la glande mammaire.

Pour valider fonctionnellement un variant, on utilise une approche plus ciblée *via* la comparaison de lignées de cellules ou d'animaux d'une espèce modèle (*e.g.* souris) homozygotes pour les deux allèles du variant. La culture de cellules peut permettre d'identifier des modifications de certains mécanismes cellulaires (*e.g.* activité enzymatique, localisation intra-cellulaire d'une molécule, …) par comparaison des deux types de cellules transformées par mutagénèse dirigée. Si la fonction du gène candidat n'est pas connue, des techniques d'édition du génome, tel le système CRISPR/Cas9, peuvent être mises en œuvre sur des lignées d'animaux d'une espèce modèle. Elles peuvent permettent de manière efficace d'invalider un gène ou d'introduire une mutation spécifique et ainsi créer des lignées transgéniques homozygotes extrêmes pour les deux allèles du variant testé. Il est ensuite possible de faire différentes analyses à différents stades (phénotypes, tissus…) sur les animaux des deux lignées pour valider l'effet du variant.

Ces dernières approches sont relativement lourdes et elles peuvent être inefficaces si le phénotype ne s'exprime pas de la même manière dans les cellules ou dans l'espèce modèle. Et bien évidemment, pour un caractère quantitatif, il ne sera pas possible de valider tous les variants candidats de cette manière. On pourra par exemple réserver les validations fonctionnelles aux variants qui présentent les effets les plus forts sur le caractère d'intérêt.

Pour les caractères que nous avons étudiés, deux régions paraissent particulièrement intéressantes pour une exploration plus approfondie. Il s'agit des régions situées sur le chromosome 1 (gène *SLC37A1*) et sur le chromosome 20 (gène *ANKH*). Les deux régions ont des effets très forts sur la composition du lait dans les trois principales races bovines laitières et sur les aptitudes fromagères en race Montbéliarde. Par ailleurs, le gène *ANKH*, qui est le meilleur candidat pour le QTL que nous identifions sur le chromosome 20, a été très peu étudié jusqu'à présent. Les gènes *SLC37A1* (*solute carrier family 37, member A1*) et *ANKH* (*inorganic pyrophosphate transport regulator*) codent respectivement un transporteur du glucose-6-

phosphate et une protéine transmembranaire impliquée dans le transport des ions. Dans le réseau de gènes que nous avons identifié, une approche *Gene Ontology* (GO) nous a permis de mettre en évidence un enrichissement en gènes du terme de GO « transport des anions inorganiques » qui contient notamment les deux gènes *SLC37A1* et *ANKH*. Ces résultats suggèrent un lien fonctionnel entre les deux gènes. Les meilleurs variants candidats de ces gènes sont des variants introniques localisés à 144 385 375 pb dans le gène *SLC37A1* et 58 446 058 pb dans le gène *ANKH*. Ces deux variants, ainsi que tous les autres candidats issus des travaux de cette thèse, ont été ajoutés sur la partie recherche de la puce EuroG10K et ils vont donc être génotypés sur un grand nombre de bovins de différentes races. Sous réserve d'avoir accès aux phénotypes, leurs effets pourront donc être confirmés sur de nouvelles populations. Par ailleurs, il sera intéressant dans un premier temps d'utiliser les données du nouveau run du projet 1000 génomes bovins (run7 : nouvel assemblage bovin et meilleure qualité d'imputation) pour essayer d'identifier de meilleurs variants candidats fonctionnels dans chacun de ces gènes. Dans un second temps, l'effet des variants des gènes *SLC37A1* et *ANKH*, impliqués dans des mécanismes cellulaires, pourrait être validée sur des lignées cellulaires homozygotes.

Quelques remarques finales…

Au terme de ce manuscrit, je reviens sur le parcours un peu atypique qu'est le mien. Ingénieur d'études, j'ai démarré cette thèse à mi-carrière, un peu plus de trois ans après mon arrivée dans l'équipe génétique et génomique bovine (G2B) de l'unité GABI, après avoir travaillé dans l'équipe de génétique porcine de la même unité. A mon arrivée dans l'équipe G2B, on m'a confié le volet « protéines » du projet *PhénoFinlait* et quelques années plus tard, le projet *From'MIR* se mettait en place. Ayant toujours été attirée par une thèse, j'ai pensé que le cadre formel d'un doctorat me donnerait l'opportunité de réaliser l'analyse intégrée des données des deux projets et surtout de m'investir en profondeur dans un sujet de recherche. Cela restait toutefois un challenge dans la mesure où je gardais certaines de mes missions d'ingénieur et ne devais consacrer que 70% de mon temps à ma thèse. Je fais le bilan ici de ces trois années au moment très particulier où j'écris les dernières lignes de ce manuscrit et je peux dire aujourd'hui que mes objectifs sont atteints, peut-être même au-delà de mes espérances. C'est en grande partie parce que j'ai eu la chance de travailler sur des projets d'envergure avec des données originales. J'ai ainsi pu appliquer un large éventail d'approches statistiques, de génétique quantitative et de génomique sur des caractères originaux (composition et fromageabilité du lait) prédits par la spectrométrie MIR et peu étudiés jusqu'à présent dans l'équipe. J'ai de plus bénéficié d'un juste équilibre entre encadrement et autonomie qui a laissé de la place à l'initiative et dans lequel je me suis épanouie. Les résultats qui en découlent sont originaux, ils ont été valorisés dans des articles scientifiques et dans des communications à des congrès mais aussi par des applications directes sur le terrain. La valorisation est à finaliser (article en cours de revue à GSE, projet d'écriture d'un article pour INRA Productions Animales, présentation au congrès EAAP en août 2019 de l'étude de la précision des évaluations génomiques de la fromageabilité et projet d'écriture d'un article scientifique) mais l'ensemble des analyses est terminé. Ce travail a été enrichissant et très motivant pour la suite de ma carrière que j'aborderai avec plus de compétences, de connaissances, d'autonomie et de recul.

Bibliographie additionnelle

Bibliographie additionnelle *

* Références bibliographiques citées dans le corps du manuscrit (les références spécifiques des articles scientifiques publiés sont listées en fin de chaque article)

Aguilar, I., I. Misztal, D. Johnson, A. Legarra, S. Tsuruta, and T. Lawlor. 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. Journal of Dairy Science 93:743-752.

Alais, C. 1984. Science du lait: principes des techniques laitières. 4e ed.

Amenu, B. and H. Deeth. 2007. The impact of milk composition on cheddar cheese manufacture. Australian Journal of Dairy Technology 62:171-184.

AWE. 2018. La transformation fermière du lait. Page http://www.awenet.be/awe/UserFiles/file/asbl/lait/Folder%20transformation%20du%20lait%202018.pdf.

Bastin, C., L. Theron, A. Laine, and N. Gengler. 2016. On the role of mid-infrared predicted phenotypes in fertility and health dairy breeding programs. Journal of Dairy Science 99:4080-4094.

Baur, A., S. Fritz, J. Promp, O. Bulot, D. Boichard, V. Ducrocq, and P. Croiseau. 2014. Implementation of the French official genomic evaluation in Brown Swiss dairy cattle. in 10th World Congress of Genetics Applied to Livestock Production. Vancouver, Canada.

Beaumont, M. 2018. Etude des possibilités de sélection des vaches de race Montbéliarde sur de nouveaux caractères de fromageabilité du lait. Mémoire de fin d'études, UniLaSalle, Beauvais, France, janvier-août 2018.

Biggs, D. A. 1978. Instrumental infrared estimation of fat, protein, and lactose in milk - Collaborative study. Journal of the Association of Official Analytical Chemists 61:1015-1034.

Bittante, G., C. Cipolat-Gotet, and A. Cecchinato. 2013. Genetic parameters of different measures of cheese yield and milk nutrient recovery from an individual model cheese-manufacturing process. Journal of Dairy Science 96:7966-7979.

Bland, J., A. Grandison, and C. Fagan. 2015. Evaluation of milk compositional variables on coagulation properties using partial least squares. Journal of Dairy Research 82:8-14.

Bobe, G., D. Beitz, A. Freeman, and G. Lindberg. 1999. Effect of milk protein genotypes on milk protein composition and its genetic parameter estimates. Journal of Dairy Science 82:2797-2804.

Boichard, D., M. Boussaha, A. Capitan, D. Rocha, C. Hozé, M. P. Sanchez, T. Tribout, R. Letaief, P. Croiseau, C. Grohs, W. Li, C. Harland, C. Charlier, M. S. Lund, G. Sahana, M. Georges, S. Barbier, W. Coppieters, S. Fritz, and B. Guldbrandtsen. 2018. Experience from large scale use of the EuroGenomics custom SNP chip in cattle. in 11th WCGALP. Auckland, New Zealand.

Boichard, D. and M. Brochard. 2012. New phenotypes for new breeding goals in dairy cattle. Animal 6:544-550.

Boichard, D., H. Chung, R. Dassonneville, X. David, A. Eggen, S. Fritz, K. J. Gietzen, B. J. Hayes, C. T. Lawley, T. S. Sonstegard, C. P. Van Tassell, P. M. VanRaden, K. A. Viaud-Martinez, G. R. Wiggans, and B. L. Consortium. 2012a. Design of a bovine low-density SNP array optimized for imputation. PLos One 7:e34130.

Boichard, D., A. Govignon-Gion, H. Larroque, C. Maroteau, I. Palhiere, G. Tosser-Klopp, R. Rupp, M. P. Sanchez, M. Brochard, Y. Amigues, M. Y. Boscher, and H. Leveziel. 2014. Genetic determinism of milk composition in fatty acids and proteins in ruminants, and selection potential. INRA Productions Animales 27:283-298.

Boichard, D., F. Guillaume, A. Baur, P. Croiseau, M. Rossignol, M. Boscher, T. Druet, L. Genestout, J. Colleau, L. Journaux, V. Ducrocq, and S. Fritz. 2012b. Genomic selection in French dairy cattle. Animal Production Science 52:115-120.

Bonfatti, V., A. Cecchinato, L. Gallo, A. Blasco, and P. Carnier. 2011a. Genetic analysis of detailed milk protein composition and coagulation properties in Simmental cattle. Journal of Dairy Science 94:5183-5193.

Bonfatti, V., L. Degano, A. Menegoz, and P. Carnier. 2016. Short communication: Mid-infrared spectroscopy prediction of fine milk composition and technological properties in Italian Simmental. Journal of Dairy Science 99:8216-8221.

Bonfatti, V., G. Di Martino, and P. Carnier. 2011b. Effectiveness of mid-infrared spectroscopy for the prediction of detailed protein composition and contents of protein genetic variants of individual milk of Simmental cows. Journal of Dairy Science 94:5776-5785.

Bonfatti, V., F. Tiezzi, F. Miglior, and P. Carnier. 2017a. Comparison of Bayesian regression models and partial least squares regression for the development of infrared prediction equations. Journal of Dairy Science 100:7306-7319.

Bonfatti, V., D. Vicario, L. Degano, A. Lugo, and P. Carnier. 2017b. Comparison between direct and indirect methods for exploiting Fourier transform spectral information in estimation of breeding values for fine composition and technological properties of milk. Journal of Dairy Science 100:2057-2067.

Bonfatti, V., D. Vicario, A. Lugo, and R. Carnier. 2017c. Genetic parameters of measures and population-wide infrared predictions of 92 traits describing the fine composition and technological properties of milk in Italian Simmental cattle. Journal of Dairy Science 100:5526-5540.

Boussaha, M., P. Michot, R. Letaief, C. Hoze, S. Fritz, C. Grohs, D. Esquerre, A. Duchesne, R. Philippe, V. Blanquet, F. Phocas, S. Floriot, D. Rocha, C. Klopp, A. Capitan, and D. Boichard. 2016. Construction of a large collection of small genome variations in French dairy and beef breeds using whole-genome sequences. Genetics Selection Evolution 48:87.

Bouwman, A. C., H. D. Daetwyler, A. J. Chamberlain, C. H. Ponce, M. Sargolzaei, F. S. Schenkel, G. Sahana, A. Govignon-Gion, S. Boitard, M. Dolezal, H. Pausch, R. F. Brondum,

P. J. Bowman, B. Thomsen, B. Guldbrandtsen, M. S. Lund, B. Servin, D. J. Garrick, J. Reecy, J. Vilkki, A. Bagnato, M. Wang, J. L. Hoff, R. D. Schnabel, J. F. Taylor, A. A. E. Vinkhuyzen, F. Panitz, C. Bendixen, L. E. Holm, B. Gredler, C. Hoze, M. Boussaha, M. P. Sanchez, D. Rocha, A. Capitan, T. Tribout, A. Barbat, P. Croiseau, C. Drogemuller, V. Jagannathan, C. V. Jagt, J. J. Crowley, A. Bieber, D. C. Purfield, D. P. Berry, R. Emmerling, K. U. Gotz, M. Frischknecht, I. Russ, J. Solkner, C. P. Van Tassell, R. Fries, P. Stothard, R. F. Veerkamp, D. Boichard, M. E. Goddard, and B. J. Hayes. 2018. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. Nature Genetics 50:362-367.

Bovenhuis, H. and J. Weller. 1994. Mapping and analysis of dairy-cattle quantitative trait loci by maximum-likelihood methodology using milk protein genes as genetic-markers. Genetics 137:267-280.

Buitenhuis, B., N. Poulsen, L. Larsen, and J. Sehested. 2015. Estimation of genetic parameters and detection of quantitative trait loci for minerals in Danish Holstein and Danish Jersey milk. BMC Genetics 16:52.

Caroli, A., S. Chessa, and G. Erhardt. 2009. Invited review: Milk protein polymorphisms in cattle: Effect on animal breeding and human nutrition. Journal of Dairy Science 92:5335-5352.

Cecchinato, A., A. Albera, C. Cipolat-Gotet, A. Ferragina, and G. Bittante. 2015. Genetic parameters of cheese yield and curd nutrient recovery or whey loss traits predicted using Fourier-transform infrared spectroscopy of samples collected during milk recording on Holstein, Brown Swiss, and Simmental dairy cows. Journal of Dairy Science 98:4914-4927.

Cecchinato, A. and G. Bittante. 2016. Genetic and environmental relationships of different measures of individual cheese yield and curd nutrients recovery with coagulation properties of bovine milk. Journal of Dairy Science 99:1975-1989.

Cecchinato, A., M. De Marchi, L. Gallo, G. Bittante, and P. Carnier. 2009. Mid-infrared spectroscopy predictions as indicator traits in breeding programs for enhanced coagulation properties of milk. Journal of Dairy Science 92:5304-5313.

Cecchinato, A., M. Penasa, M. De Marchi, L. Gallo, G. Bittante, and P. Carnier. 2011. Genetic parameters of coagulation properties, milk yield, quality, and acidity estimated using coagulating and noncoagulating milk information in Brown Swiss and Holstein-Friesian cows. Journal of Dairy Science 94:4205-4213.

Chamberlain, A. J., B. J. Hayes, R. Xiang, C. J. Vander Jagt, C. M. Reich, I. M. MacLeod, C. P. Prowse-Wilkins, B. A. Mason, H. Daetwyler, and M. E. Goddard. 2018. Identification of regulatory variation in dairy cattle with RNA sequence data. in 11th WCGALP. Auckland, New Zealand.

Cipolat-Gotet, C., A. Cecchinato, M. De Marchi, M. Penasa, and G. Bittante. 2012. Comparison between mechanical and near-infrared methods for assessing coagulation properties of bovine milk. Journal of Dairy Science 95:6806-6819.

CNAOL. 2018. Chiffres clés 2017. Conseil National des Appellations d'Origine Laitières ed.

CNIEL. 2018. L'économie laitière en chiffres - Edition 2018. Centre National Interprofessionnel de l'Economie Laitière ed.

Colinet, F. G., T. Troch, O. Abbas, V. Baeten, F. Dehareng, E. Froidmont, H. Soyeurt, P. Dardenne, M. Sindic, and N. Gengler. 2013. Potentiel d'utilisation de la spectrométrie moyen infrarouge pour prédire le rendement fromager du lait et étudier sa variabilité génétique. in Rencontres Recherche Ruminants. Vol. 20, Paris, France.

Coppa, M., A. Ferlay, C. Leroux, M. Jestin, Y. Chilliard, B. Martin, and D. Andueza. 2010. Prediction of milk fatty acid composition by near infrared reflectance spectroscopy. International Dairy Journal 20:182-189.

Coppieters, W., J. Riquet, J. Arranz, P. Berzi, N. Cambisano, B. Grisart, L. Karim, F. Marcq, L. Moreau, C. Nezer, P. Simon, P. Vanmanshoven, D. Wagenaar, and M. Georges. 1998. A QTL with major effect on milk yield and composition maps to bovine Chromosome 14. Mammalian Genome 9:540-544.

Corrieu, G., H. Spinnler, Y. Jomier, and D. Picque. 1988. Automated system to follow up and control the acidification activity of lactic acid starters. French Patent FR 2:629-612.

Couvreur, S. and C. Hurtaud. 2007. Globule milk fat: Secretion, composition, function and variation factors. INRA Productions Animales 20:369-382.

Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. Van Binsbergen, R. F. Brøndum, X. Liao, A. Djari, S. Rodriguez, C. Grohs, D. Esquerré, O. Bouchez, M. N. Rossignol, C. Klopp, D. Rocha, S. Fritz, A. Eggen, P. J. Bowman, D. Coote, A. J. Chamberlain, C. P. VanTassell, I. Hulsegge, M. E. Goddard, B. Guldbrandtsen, M. S. Lund, R. F. Veerkamp, D. Boichard, R. Fries, and B. J. Hayes. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nature Genetics 46:858-867.

Dagnachew, B., T. Meuwissen, and T. Adnoy. 2013. Genetic components of milk Fourier-transform infrared spectra used to predict breeding values for milk composition and quality traits in dairy goats. Journal of Dairy Science 96:5933-5942.

Dal Zotto, R., M. De Marchi, A. Cecchinato, M. Penasa, M. Cassandro, P. Carnier, L. Gallo, and G. Bittante. 2008. Reproducibility and repeatability of measures of milk coagulation properties and predictive ability of mid-infrared reflectance spectroscopy. Journal of Dairy Science 91:4103-4112.

De Kruif, C. G. and C. Holt. 2003. Casein Micelle Structure, Functions and Interactions. Pages 233-276 in Advanced Dairy Chemistry—1 Proteins: Part A / Part B. P. F. Fox and P. L. H. McSweeney, ed. Springer US, Boston, MA.

De Marchi, M., V. Bonfatti, A. Cecchinato, G. Di Martino, and P. Carnier. 2009a. Prediction of protein composition of individual cow milk using mid-infrared spectroscopy. Italian Journal of Animal Science 8:399-401.

De Marchi, M., C. Fagan, C. O'Donnell, A. Cecchinato, R. Dal Zotto, M. Cassandro, M. Penasa, and G. Bittante. 2009b. Prediction of coagulation properties, titratable acidity, and pH of bovine milk using mid-infrared spectroscopy. Journal of Dairy Science 92:423-432.

De Marchi, M., M. Penasa, F. Tiezzi, V. Toffanin, and M. Cassandro. 2012. Prediction of milk coagulation properties by Fourier Transform Mid-Infrared Spectroscopy (FTMIR) for genetic purposes, herd management and dairy profitability. Pages 47-53 in International Strategies and New Developments in Milk Analysis. VI ICAR Reference Laboratory Network Meeting. Vol. ICAR Technical Series No. 16, Cork, Ireland.

De Marchi, M., M. Penasa, A. Zidi, and C. Manuelian. 2018. Invited review: Use of infrared technologies for the assessment of dairy products-Applications and perspectives. Journal of Dairy Science 101:10589-10604.

De Marchi, M., V. Toffanin, M. Cassandro, and M. Penasa. 2014. Invited review: Mid-infrared spectroscopy as phenotyping tool for milk traits. Journal of Dairy Science 97:1171-1186.

Druet, T., F. Jaffrezic, and V. Ducrocq. 2005. Estimation of genetic parameters for test day records of dairy traits in the first three lactations. Genetics Selection Evolution 37:257-271.

Ducrocq, V. 2011. Genekit, BLUP software. INRA, Jouy-en-Josas, France.

Eck, A. 1984. Le fromage. Technique et documentation (Lavoisier) ed.

El Jabri, M., M. P. Sanchez, C. Laithier, E. Doutart, V. Wolf, D. Pourchet, P. Grosperrin, E. Beuvier, O. Rolet-Répécaud, Y. Gaüzère, O. Belysheva, A. Delacroix-Buchet, and D. Boichard. 2017. Bayesian regression models and variable selection methods before PLS regression. Application to the prediction of milk cheese-making properties using infared spectral data. in Chimiometrie XVIII. Paris, France.

El Jabri, M., M. P. Sanchez, P. Trossat, C. Laithier, V. Wolf, P. Grosperrin, E. Beuvier, O. Rolet-Répécaud, S. Gavoye, Y. Gauzere, O. Belysheva, E. Notz, D. Boichard, and A. Delacroix-Buchet. 2019. Comparison of Bayesian and PLS regression methods for mid-infrared prediction of cheese-making properties in Montbéliarde cows. Sous presse dans Journal of Dairy Science. 102. https://doi.org/10.3168/jds.2019-16320

ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489:57-74.

Erbe, M., B. Hayes, L. Matukumalli, S. Goswami, P. Bowman, C. Reich, B. Mason, and M. Goddard. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. Journal of Dairy Science 95:4114-4129.

Fang, Z., M. Visker, G. Miranda, A. Delacroix-Buchet, H. Bovenhuis, and P. Martin. 2016. The relationships among bovine alpha(S)-casein phosphorylation isoforms suggest different phosphorylation pathways. Journal of Dairy Science 99:8168-8177.

Fernando, R., H. Cheng, B. Golden, and D. Garrick. 2016. Computational strategies for alternative single-step Bayesian regression models with large numbers of genotyped and non-genotyped animals. Genetics Selection Evolution 48:96.

Ferragina, A., C. Cipolat-Gotet, A. Cecchinato, and G. Bittante. 2013. The use of Fourier-transform infrared spectroscopy to predict cheese yield and nutrient recovery or whey loss traits from unprocessed bovine milk samples. Journal of Dairy Science 96:7980-7990.

Ferragina, A., G. de los Campos, A. Vazquez, A. Cecchinato, and G. Bittante. 2015. Bayesian regression models outperform partial least squares methods for predicting milk components and technological properties using infrared spectral data. Journal of Dairy Science 98:8133-8151.

Ferrand, M., G. Miranda, S. Guisnel, H. Larroque, O. Leray, F. Lahalle, M. Brochard, and P. Martin. 2012. Determination of protein composition in milk by mid-infrared spectrometry. Pages 41-45 in Proc. International Strategies and New Developments in Milk Analysis. VI ICAR Reference Laboratory Network Meeting, Cork, Ireland.

Ferrand-Calmels, M., I. Palhiere, M. Brochard, O. Leray, J. Astruc, M. Aurel, S. Barbey, F. Bouvier, P. Brunschwig, H. Caillatt, M. Douguet, F. Faucon-Lahalle, M. Gele, G. Thomas, J. Trommenschlager, and H. Larroque. 2014. Prediction of fatty acid profiles in cow, ewe, and goat milk by mid-infrared spectrometry. Journal of Dairy Science 97:17-35.

Fleming, A., F. Schenkel, J. Chen, F. Malchiodi, V. Bonfatti, R. Ali, B. Mallard, M. Corredig, and F. Miglior. 2017. Prediction of milk fatty acid content with mid-infrared spectroscopy in Canadian dairy cattle using differently distributed model development sets. Journal of Dairy Science 100:5073-5081.

Fortes, M., A. Reverter, Y. Zhang, E. Collis, S. Nagaraj, N. Jonsson, K. Prayaga, W. Barris, and R. Hawken. 2010. Association weight matrix for the genetic dissection of puberty in beef cattle. Proceedings of National Academy of Sciences USA 107:13642-13647.

FranceAgriMer. 2018. Données et bilans FranceAgriMer: les produits carnés et laitiers - Données statistiques 2017 - France / UE / Monde. FranceAgriMer ed.

Garnsworthy, P., S. Feng, A. Lock, and M. Royal. 2010. Short communication: Heritability of milk fatty acid composition and stearoyl-CoA desaturase indices in dairy cows. Journal of Dairy Science 93:1743-1748.

Gaudillière, N., M. Gelé, M. P. Sanchez, M. El Jabri, V. Wolf, D. Boichard, M. Brochard, A. Delacroix-Buchet, and C. Laithier. 2018. La fromageabilité du lait en race Montbéliarde dans les élevages AOP et IGP de Franche-Comté : variabilité et facteurs de variation. in Rencontres Recherche Ruminants. Vol. 24, Paris.

Gebreyesus, G., M. S. Lund, L. Janss, N. A. Poulsen, L. B. Larsen, H. Bovenhuis, and A. J. Buitenhuis. 2016. Short communication: Multi-trait estimation of genetic parameters for milk protein composition in the Danish Holstein. Journal of Dairy Science 99:2863-2866.

Gelé, M., S. Minery, J. M. Astruc, P. Brunschwig, M. Ferrand, G. Lagriffoul, H. Larroque, J. Legarto, P. Martin, G. Miranda, I. Palhière, P. Trossat, and M. Brochard. 2014. Phénotypage et génotypage à grande échelle de la composition fine des laits dans les filières bovine, ovine et caprine. INRA Productions Animales 27:255-268.

Gengler, N., H. Soyeurt, F. Dehareng, C. Bastin, F. Colinet, H. Hammami, M. Vanrobays, A. Laine, S. Vanderick, C. Grelet, A. Vanlierde, E. Froidmont, and P. Dardenne. 2016. Capitalizing on fine milk composition for breeding and management of dairy cows. Journal of Dairy Science 99:4071-4079.

Georges, M., D. Nielsen, M. Mackinnon, Mishra, A, R. Okimoto, A. Pasquino, L. Sargeant, Sorensen, A, M. Steele, X. Zhao, J. Womack, and I. Hoeschele. 1995. Mapping quantitative trait loci controlling milk-production in dairy-cattle by exploiting progeny testing. Genetics 139:907-920.

Giuffra, E., C. K. Tuggle, and F. Consortium. 2019. Functional Annotation of Animal Genomes (FAANG): Current Achievements and Roadmap. Annual Review of Animal Biosciences 7:65-88.

Glantz, M., T. Devold, G. Vegarud, H. Mansson, H. Stalhammar, and M. Paulsson. 2010. Importance of casein micelle size and milk composition for milk gelation. Journal of Dairy Science 93:1444-1451.

Goddard, M. and B. Hayes. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nature Review Genetics 10:381-391.

Grelet, C., C. Bastin, M. Gele, J. Daviere, M. Johan, A. Werner, R. Reding, J. Pierna, F. Colinet, P. Dardenne, N. Gengler, H. Soyeurt, and F. Dehareng. 2016. Development of Fourier transform mid-infrared calibrations to predict acetone, beta-hydroxybutyrate, and citrate contents in bovine milk through a European dairy network. Journal of Dairy Science 99:4816-4825.

Grelet, C., J. Pierna, P. Dardenne, V. Baeten, and F. Dehareng. 2015. Standardization of milk mid-infrared spectra from a European dairy network. Journal of Dairy Science 98:2150-2160.

Grelet, C., J. Pierna, P. Dardenne, H. Soyeurt, A. Vanlierde, F. Colinet, C. Bastin, N. Gengler, V. Baeten, and F. Dehareng. 2017. Standardization of milk mid-infrared spectrometers for the transfer and use of multiple models. Journal of Dairy Science 100:7910-7921.

Grelet, C., A. Vanlierde, M. Hostens, L. Foldager, M. Salavati, K. L. Ingvartsen, M. Crowe, M. T. Sorensen, E. Froidmont, C. P. Ferris, C. Marchitelli, F. Becker, T. Larsen, F. Carter, and F. Dehareng. 2019. Potential of milk mid-IR spectra to predict metabolic status of cows through blood components and an innovative clustering approach. Animal 13:649-658.

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and R. Snell. 2002. Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Research 12:222-231.

Grosclaude, F. 1988. Le polymorphisme génétique des principales lactoprotéines bovines. INRA Productions Animales. 1:5-17.

Gustavsson, F., M. Glantz, N. Poulsen, L. Wadso, H. Stalhammar, A. Andren, H. Mansson, L. Larsen, M. Paulsson, and W. Fikse. 2014. Genetic parameters for rennet- and acid-induced coagulation properties in milk from Swedish Red dairy cows. Journal of Dairy Science 97:5219-5229.

Habier, D., R. Fernando, K. Kizilkaya, and D. Garrick. 2011. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics 12:186.

Haile-Mariam, M. and J. Pryce. 2017. Genetic parameters for lactose and its correlation with other milk production traits and fitness traits in pasture-based production systems. Journal of Dairy Science 100:3754-3766.

Heck, J., A. Schennink, H. van Valenberg, H. Bovenhuis, M. Visker, J. van Arendonk, and A. van Hooijdonk. 2009. Effects of milk protein variants on the protein composition of bovine milk. Journal of Dairy Science 92:1192-1202.

Howie, B., C. Fuchsberger, M. Stephens, J. Marchini, and G. Abecasis. 2012. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nature Genetics 44:955.

Hurtaud, C., J. Peyraud, G. Michel, D. Berthelot, and L. Delaby. 2009. Winter feeding systems and dairy cow breed have an impact on milk composition and flavour of two Protected Designation of Origin French cheeses. Animal 3:1327-1338.

Hurtaud, C., H. Rulquin, M. Delaite, and R. Verite. 1995. Prediction of cheese yielding efficiency of individual milk of dairy cows - correlation with coagulation parameters and laboratory curd yield. Annales de Zootechnie 44:385-398.

Idele. 2018. Evaluation génétique des taureaux Monbéliards: production laitière, morphologie et caractères fonctionnels. Page 17 pages. Edition 18/35 Décembre 2018 ed. Idele, Geneval.

Idele and CNE. 2018. Les chiffres clés du GEB - bovins 2017 - Productions lait et viande. Institut de l'élevage et Confédération Nationale de l'Elevage ed.

Johnson, T., M. Keehan, C. Harland, T. Lopdell, R. J. Spelman, S. R. Davis, B. D. Rosen, T. P. L. Smith, and C. Couldrey. 2019. Short communication: Identification of the pseudoautosomal region in the Hereford bovine reference genome assembly ARS-UCD1.2. Journal of Dairy Science 102:3254-3258.

Karoui, R., A. Mouazen, E. Dufour, R. Schoonheydt, and J. de Baerdemaeker. 2006. Utilisation of front-face fluorescence spectroscopy for the determination of some selected chemical parameters in soft cheeses. Lait 86:155-169.

Kumar, P., S. Henikoff, and P. Ng. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature Protocols 4:1073-1082.

Legarra, A., A. Ricard, and L. Varona. 2018. GWAS by GBLUP: Single and Multimarker EMMAX and Bayes Factors, with an Example in Detection of a Major Gene for Horse Gait. G3-Genes Genomes Genetics 8:2301-2308.

Legarto, J., M. Gelé, A. Ferlay, C. Hurtaud, G. lagriffoul, I. Palhière, J. Peyraud, B. Rouillé, and P. Brunschwig. 2014. Effets des conduites d'élevage sur la production de lait, les taux butyreux et protéique et la composition en acides gras du lait de vache, chèvre et brebis évaluée par spectrométrie dans le moyen infrarouge. INRA Productions Animales 27:269-282.

Leonil, J., M. Michalski, and P. Martin. 2013. Supramolecular structures of milk: structure and nutritional impact of the casein micelle and the milk fat globule. INRA Productions Animales 26:129-143.

Letaief, R., E. Rebours, C. Grohs, C. Meersseman, S. Fritz, L. Trouilh, D. Esquerré, J. Barbieri, C. Klopp, R. Philippe, V. Blanquet, D. Boichard, D. Rocha, and M. Boussaha. 2017. Identification of copy number variation in French dairy and beef breeds using next-generation sequencing. Genetics Selection Evolution 49:77.

Lunden, A., M. Nilsson, and L. Janson. 1997. Marked effect of beta-lactoglobulin polymorphism on the ratio of casein to total protein in milk. Journal of Dairy Science 80:2996-3005.

Mahaut, M., R. Jeantet, and G. Brulé. 2000. Initiation à la technologie fromagère. Technique et documentation (Lavoisier) ed.

Malacarne, M., P. Formaggioni, P. Franceschi, and A. Summer. 2004. Seasonal variations of milk quality in Parmigiano-Reggiano cheese manufacture on a period of 10 years. Pages 63-77 in Scienza e Tecnica Lattiero Casearia. Vol. 55, Italy.

Martin, P., M. Szymanowska, L. Zwierzchowski, and C. Leroux. 2002. The impact of genetic polymorphisms on the protein composition of ruminant milks. Reproduction Nutrition Development 42:433-459.

McLaren, W., L. Gil, S. Hunt, H. Riat, G. Ritchie, A. Thormann, P. Flicek, and F. Cunningham. 2016. The Ensembl Variant Effect Predictor. Genome Biology 17:122.

McParland, S., E. Lewis, E. Kennedy, S. Moore, B. McCarthy, M. O'Donovan, S. Butler, J. Pryce, and D. Berry. 2014. Mid-infrared spectrometry of milk as a predictor of energy intake and efficiency in lactating dairy cows. Journal of Dairy Science 97:5863-5871.

Mesbah-Uddin, M., B. Guldbrandtsen, T. Iso-Touru, J. Vilkki, D. De Koning, D. Boichard, M. Lund, and G. Sahana. 2018. Genome-wide mapping of large deletions and their population-genetic properties in dairy cattle. DNA Research 25:49-59.

Meuwissen, T. H. and M. E. Goddard. 2000. Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. Genetics 155:421-430.

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829.

Meyer, K. 2007. WOMBAT - A tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). Journal of Zhejiang University Science B 8:815-821.

Miranda, G., N. Boumahrou, L. Bianchi, A. Pinard, B. Saadaoui, A. Guillot, C. Henry, C. Bevilacqua, C. Beauvallet, S. Bellier, C. Cebo, and P. Martin. 2011. Understanding milk protein complexity to produce accurate phenotypes. in 8th Int. Milk Genomics Consortium Symp., Melbourne, Australia.

Montel, M., S. Buchin, A. Mallet, C. Delbes-Paus, D. Vuitton, N. Desmasures, and F. Berthier. 2014. Traditional cheeses: Rich and diverse microbiota with associated benefits. International Journal of Food Microbiology 177:136-154.

Perez, P. and G. de los Campos. 2014. Genome-Wide Regression and Prediction with the BGLR Statistical Package. Genetics 198:483-495.

Poulsen, N. A., A. J. Buitenhuis, and L. B. Larsen. 2015. Phenotypic and genetic associations of milk traits with milk coagulation properties. Journal of Dairy Science 98:2079-2087.

Pretto, D., N. Lopez-Villalobos, M. Penasa, and M. Cassandro. 2012. Genetic response for milk production traits, somatic cell score, acidity and coagulation properties in Italian Holstein-Friesian population under current and alternative selection indices and breeding objectives. Livestock Science 150:59-66.

Pretto, D., M. Vallas, E. Parna, A. Tanavots, H. Kiiman, and T. Kaart. 2014. Short communication: Genetic correlation and heritability of milk coagulation traits within and across lactations in Holstein cows using multiple-lactation random regression animal models. Journal of Dairy Science 97:7980-7984.

Rahman, R. 2017. Etudes des variants des protéines du lait de vache. Mémoire de L2, Université de Versailles Saint-Quentin-en-Yvelines, Versailles, France, juin - août 2017.

Raven, L., B. Cocks, and B. Hayes. 2014. Multibreed genome wide association can improve precision of mapping causative variants underlying milk production in dairy cattle. BMC Genomics 15:62.

Rutten, M., H. Bovenhuis, K. Hettinga, H. van Valenberg, and J. van Arendonk. 2009. Predicting bovine milk fat composition using infrared spectroscopy based on milk samples collected in winter and summer. Journal of Dairy Science 92:6202-6209.

Salque, M., P. Bogucki, J. Pyzel, I. Sobkowiak-Tabaka, R. Grygiel, M. Szmyt, and R. Evershed. 2013. Earliest evidence for cheese making in the sixth millennium BC in northern Europe. Nature 493:522-525.

Sanchez, M. P., M. El Jabri, S. Minéry, V. Wolf, E. Beuvier, C. Laithier, A. Delacroix-Buchet, M. Brochard, and D. Boichard. 2018a. Genetic parameters for cheese-making properties and milk composition predicted from mid-infrared spectra in a large dataset of Montbéliarde cows. Journal of Dairy Science 101:10048-10061.

Sanchez, M. P., M. Ferrand, M. Gele, D. Pourchet, G. Miranda, P. Martin, M. Brochard, and D. Boichard. 2017a. Short communication: Genetic parameters for milk protein composition predicted using mid-infrared spectroscopy in the French Montbeliarde, Normande, and Holstein dairy cattle breeds. Journal of Dairy Science 100:6371-6375.

Sanchez, M. P., A. Govignon-Gion, P. Croiseau, S. Fritz, C. Hozé, G. Miranda, P. Martin, A. Barbat-Leterrier, R. Letaïef, D. Rocha, M. Brochard, M. Boussaha, and D. Boichard. 2017b. Within-breed and multi-breed GWAS on imputed whole-genome sequence variants reveal candidate mutations affecting milk protein composition in dairy cattle. Genetics Selection Evolution 49:68.

Sanchez, M. P., A. Govignon-Gion, M. Ferrand, M. Gele, D. Pourchet, Y. Amigues, S. Fritz, M. Boussaha, A. Capitan, D. Rocha, G. Miranda, P. Martin, M. Brochard, and D. Boichard. 2016a. Whole-genome scan to detect quantitative trait loci associated with milk protein composition in 3 French dairy cattle breeds. Journal of Dairy Science 99:8203-8215.

Sanchez, M. P., D. Jonas, A. Baur, V. Ducrocq, C. Hozé, R. Saintilan, F. Phocas, S. Fritz, D. Boichard, and P. Croiseau. 2016b. Implementation of genomic selection in three French regional dairy cattle breeds. in Proc. 67th European Association of Animal Production, Belfast, Ireland.

Sanchez, M. P., Y. Ramayo-Caldas, V. Wolf, C. Laithier, M. El Jabri, M. Boussaha, S. Taussat, S. Fritz, A. Delacroix-Buchet, M. Brochard, and D. Boichard. 2019. Sequence-based GWAS, network and pathway analyses reveal genes co-associated with milk cheese-making properties and milk composition in Montbéliarde cows. Genetics Selection Evolution. 51:34.

Sanchez, M. P., V. Wolf, M. El Jabri, E. Beuvier, O. Rolet-Répécaud, Y. Gaüzère, S. Minéry, M. Brochard, A. Michenet, S. Taussat, A. Barbat-Leterrier, A. Delacroix-Buchet, C. Laithier, S. Fritz, and D. Boichard. 2018b. Short communication: Confirmation of candidate causative variants on milk composition and cheesemaking properties in Montbéliarde cows. Journal of Dairy Science 101:10076-10081.

Sargolzaei, M., J. Chesnais, and F. Schenkel. 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15:478.

Schnabel, R., H. Daetwyler, and A. Chamberlain. 2019. Bovine haplotype reference consortium 1000 bulls and GTEx: international projects to advance bovine genomic research. in Plant and Animal Genome XXVII. San Diego, California, USA.

Schopen, G. C., J. M. Heck, H. Bovenhuis, M. H. Visker, H. J. van Valenberg, and J. A. van Arendonk. 2009. Genetic parameters for major milk proteins in Dutch Holstein-Friesians. Journal of Dairy Science 92:1182-1191.

Soyeurt, H., D. Bruwier, J. Romnee, N. Gengler, C. Bertozzi, D. Veselko, and P. Dardenne. 2009. Potential estimation of major mineral contents in cow milk using mid-infrared spectrometry. Journal of Dairy Science 92:2444-2454.

Soyeurt, H., P. Dardenne, F. Dehareng, G. Lognay, D. Veselko, M. Marlier, C. Bertozzi, P. Mayeres, and N. Gengler. 2006. Estimating fatty acid content in cow milk using mid-infrared spectrometry. Journal of Dairy Science 89:3690-3695.

Soyeurt, H., F. Dehareng, N. Gengler, S. McParland, E. Wall, D. Berry, M. Coffey, and P. Dardenne. 2011. Mid-infrared prediction of bovine milk fatty acids across multiple breeds, production systems, and countries. Journal of Dairy Science 94:1657-1667.

Soyeurt, H., I. Misztal, and N. Gengler. 2010. Genetic variability of milk components based on mid-infrared spectral data. Journal of Dairy Science 93:1722-1728.

Spelman, R., W. Coppieters, L. Karim, J. vanArendonk, and H. Bovenhuis. 1996. Quantitative trait loci analysis for five milk production traits on chromosome six in the Dutch Holstein-Friesian population. Genetics 144:1799-1807.

Sundekilde, U., P. Frederiksen, M. Clausen, L. Larsen, and H. Bertram. 2011. Relationship between the Metabolite Profile and Technological Properties of Bovine Milk from Two Dairy Breeds Elucidated by NMR-Based Metabolomics. Journal of Agricultural and Food Chemistry 59:7360-7367.

Toffanin, V., M. De Marchi, N. Lopez-Villalobos, and M. Cassandro. 2015a. Effectiveness of mid-infrared spectroscopy for prediction of the contents of calcium and phosphorus, and titratable acidity of milk and their relationship with milk quality and coagulation properties. International Dairy Journal 41:68-73.

Toffanin, V., M. Penasa, S. McParland, D. Berry, M. Cassandro, and M. De Marchi. 2015b. Genetic parameters for milk mineral content and acidity predicted by mid-infrared spectroscopy in Holstein-Friesian cows. Animal 9:775-780.

Vallas, M., H. Bovenhuis, T. Kaart, K. Parna, H. Kiiman, and E. Parna. 2010. Genetic parameters for milk coagulation properties in Estonian Holstein cows. Journal of Dairy Science 93:3789-3796.

Vallas, M., T. Kaart, S. Varv, K. Parna, I. Joudu, H. Viinalass, and E. Parna. 2012. Composite beta-kappa-casein genotypes and their effect on composition and coagulation of milk from Estonian Holstein cows. Journal of Dairy Science 95:6760-6769.

van den Berg, I., D. Boichard, and M. Lund. 2016. Comparing power and precision of within-breed and multibreed genome-wide association studies of production traits using whole-genome sequence data for 5 French and Danish dairy cattle breeds. Journal of Dairy Science 99:8932-8945.

van Hulzen, K. J. E., R. C. Sprong, R. van der Meer, and J. A. M. van Arendonk. 2009. Genetic and nongenetic variation in concentration of selenium, calcium, potassium, zinc, magnesium, and phosphorus in milk of Dutch Holstein-Friesian cows. Journal of Dairy Science 92:5754-5759.

Vanlierde, A., H. Soyeurt, N. Gengler, F. Colinet, E. Froidmont, M. Kreuzer, F. Grandl, M. Bell, P. Lund, D. Olijhoek, M. Eugene, C. Martin, B. Kuhla, and F. Dehareng. 2018. Short communication: Development of an equation for estimating methane emissions of dairy cows from milk Fourier transform mid-infrared spectra by using reference data obtained exclusively from respiration chambers. Journal of Dairy Science 101:7618-7624.

VanRaden, P. M. 2004. Invited review: Selection on net merit to improve lifetime profit. Journal of Dairy Science 87:3125-3131.

Veerkamp, R. F., L. Kaal, Y. De Haas, and J. D. Oldham. 2013. Breeding for robust cows that produce healthier milk: RobustMilk. Advances in Animal Biosciences 4:594-599.

Visentin, G., A. McDermott, S. McParland, D. Berry, O. Kenny, A. Brodkorb, M. Fenelon, and M. De Marchi. 2015. Prediction of bovine milk technological traits from mid-infrared spectroscopy analysis in dairy cows. Journal of Dairy Science 98:6620-6629.

Visentin, G., S. McParland, M. De Marchi, A. McDermott, M. Fenelon, M. Penasa, and D. Berry. 2017. Processing characteristics of dairy cow milk are moderately heritable. Journal of Dairy Science 100:6343-6355.

Visker, M., B. Dibbits, S. Kinders, H. van Valenberg, J. van Arendonk, and H. Bovenhuis. 2011. Association of bovine beta-casein protein variant I with milk production and milk protein composition. Animal Genetics 42:212-218.

Visscher, P., N. Wray, Q. Zhang, P. Sklar, M. McCarthy, M. Brown, and J. Yang. 2017. 10 Years of GWAS Discovery: Biology, Function, and Translation. American Journal of Human Genetics 101:5-22.

Walstra, P. 1999. Casein sub-micelles: do they exist ? International Dairy Journal 9:189-192.

Wang, Y., D. Veltkamp, and B. Kowalski. 1991. Multivariate Instrument Standardization. Analytical Chemistry 63:2750-2756.

Wedholm, A., L. B. Larsen, H. Lindmark-Månsson, A. H. Karlsson, and A. Andrén. 2006. Effect of protein composition on the cheese-making properties of milk from individual dairy cows. Journal of Dairy Science 89:3296-3305.

Westra, H. and L. Franke. 2014. From genome to function by studying eQTLs. Biochimica Et Biophysica Acta-Molecular Basis of Disease 1842:1896-1902.

Yang, J., T. Ferreira, A. Morris, S. Medland, P. Madden, A. Heath, N. Martin, G. Montgomery, M. Weedon, R. Loos, T. Frayling, M. McCarthy, J. Hirschhorn, M. Goddard, P. Visscher, G. I. A. Trai, D. G. R. Meta-A, G. I. A. Trai, and D. G. R. Meta-A. 2012. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nature Genetics 44:369-375.

Yang, J., S. Lee, M. Goddard, and P. Visscher. 2011. GCTA: a tool for genome-wide complex trait analysis. American Journal of Human Genetics 88:76-82.

Liste des publications 2016-2019

Liste des publications entre 2016 et 2019

Résultats du travail de thèse

Articles scientifiques

1. **Sanchez M-P.**, Govignon-Gion A., Ferrand M., Gelé M., Pourchet D., Amigues Y., Fritz S., Boussaha M., Capitan A., Rocha D., Miranda G., Martin P., Brochard M., Boichard D. **2016**. Identification of QTL and candidate mutations affecting major milk proteins in three French dairy cattle breeds. Journal of Dairy Science 99:8203–8215. http://dx.doi.org/10.3168/jds.2016-11437

2. **Sanchez M.-P.**, Ferrand M., Gelé M., Pourchet D., Miranda G., Martin P., Brochard M., Boichard D. **2017**. Short-communication: Genetic parameters for milk protein composition predicted using mid-infrared spectroscopy in the French Montbéliarde, Normande and Holstein dairy cattle breeds. Journal of Dairy Science 100:6371-6375. https://doi.org/10.3168/jds.2017-12663

3. **Sanchez M.-P.**, Govignon-Gion A., Croiseau P., Fritz S., Hozé C., Miranda G., Martin P., Barbat-Leterrier A., Brochard M., Boussaha M., Boichard D. **2017**. Within-breed and multi-breed GWAS on imputed whole genome sequence variants reveal candidate mutations affecting milk protein composition in dairy cattle. Genetics Selection Evolution 49:68. https://doi.org/10.1186/s12711-017-0344-z

4. **Sanchez M.-P.**, El Jabri M., Minéry S., Wolf V., Beuvier E., Laithier C., Delacroix-Buchet A., Brochard M. & Boichard. D. **2018**. Genetic parameters for cheese-making properties and milk composition predicted from mid-infrared spectrometry in a large dataset of Montbéliarde cows. Journal of Dairy Science 101:10048–10061. https://doi.org/10.3168/jds.2018-14878

5. **Sanchez M.-P.,** Wolf V., El Jabri M., Beuvier E., Rolet-Répécaud O., Gaüzère Y., Minéry S., Brochard M., Fritz S., Michenet A., Taussat S., Barbat-Leterrier A., Delacroix-Buchet A., Laithier C., Boichard. D. **2018**. Short Communication: Confirmation of candidate causative variants on milk composition and cheese-making properties in Montbéliarde cows. Journal of Dairy Science 101:10076–10081. https://doi.org/10.3168/jds.2018-14986

6. **Sanchez M.-P.**, Ramayo-Caldas Y., Wolf V., Laithier C., El Jabri M., Boussaha M., Taussat S., Fritz S., Delacroix-Buchet A., Brochard M., and Boichard D. **2019**. Sequence-based GWAS, network and pathway analyses reveal genes co-associated with milk cheese-making properties and milk composition in Montbéliarde cows. Genetics Selection Evolution 51:34. https://doi.org/10.1186/s12711-019-0473-7

7. El Jabri M., **Sanchez M-P.**, Trossat P., Laithier C., Wolf, V., Grosperrin P., Beuvier E., Rolet-Répécaud O., Gavoye S., Gaüzère Y., Belysheva O., Notz E., Boichard D., Delacroix-Buchet A. **2019**. Comparison of Bayesian and PLS regression methods for mid-infrared prediction of cheese-making properties in Montbéliarde cows. Sous presse dans Journal of Dairy Science 102. https://doi.org/10.3168/jds.2019-16320

8. **Sanchez M-P**., Wolf V., Laithier C., El Jabri M., Beuvier E., Rolet-Répécaud O., Gaudillière N., Minéry S., Ramayo-Caldas Y., Tribout T., Michenet A., Boussaha M., Taussat S., Fritz S., Delacroix-Buchet A., Grosperrin P., Brochard M. and Boichard D. 2019. Analyse génétique de la fromageabilité du lait de vache prédite par spectrométrie moyen infrarouge en race Montbéliarde. Soumis pour publication à INRA Productions Animales le 21/05/2019.

Communications à des congrès / séminaires

1. **Sanchez M-P.**, Govignon-Gion A., Barbat A., Gelé M., Fritz S., Miranda G., Martin P., Boussaha M., Brochard M., Croiseau P., Boichard D. **2016**. Using whole genome sequences to identify candidate mutations of milk fatty acids and proteins in dairy cattle. 24th Plant & animal Genome. San diego, CA, United States, January 9-13th, 2016.

2. Boichard D., **Sanchez M-P.**, Govignon-Gion A., Barbat A., Boussaha M., Tribout T., Lefebvre R., Saintilan R., Hozé C., Fritz S., Croiseau P. **2016**. Identification of candidate causal variants underlying QTL in dairy cattle through GWAS and Bayesian approach at the sequence level. 24th Plant & animal Genome. San diego, CA, United States, January 9-13th, 2016.

3. **Sanchez M-P. 2016.** Genetic analysis of bovine milk composition and cheese-making abilities predicted from MIR spectra. Séminaire des doctorants du département de génétique animale INRA, Castanet-Tolosan, 16-17 mars 2016.

4. El Jabri M., **Sanchez M-P.**, Laithier C., Doutart E., Wolf V., Pourchet D., Grosperrin P., Beuvier E., Rolet-Répécaud O., Gaüzère Y., Belysheva O., Delacroix-Buchet A., Boichard. D. **2017**. Bayesian regression models and variable selection methods before PLS regression. Application to the prediction of mill cheese-making properties using infrared spectral data. Chimiometrie XVIII, Paris, France, January 30th – February 1st, 2017.

5. **Sanchez M-P. 2017.** Genetic analysis of bovine milk composition and cheese-making abilities predicted from MIR spectra. Séminaire des doctorants du département de génétique animale INRA, Bruz, 9-10 mai 2017.

6. **Sanchez M-P.**, Wolf V., El Jabri M., Beuvier E., Rolet-Répécaud O., Gaüzère Y., Minéry S., Brochard M. Fritz S., Michenet A., Taussat S., Barbat-Leterrier A., Delacroix-Buchet A., Laithier C., Boichard. D. **2018**. Validation of candidate causative variants on milk composition and cheese-making properties in Montbéliarde cows. In proceedings of the 11th World Congress on Genetics Applied to Livestock Production: 11th WCGALP, 12-16 Feb. 2018, Auckland, New Zealand.

7. **Sanchez M-P.**, Wolf V., El Jabri M., Boussaha M., Taussat S., Laithier C., Delacroix-Buchet A., Brochard M., Boichard. D. **2018**. GWAS on whole genome sequences for cheese-making traits and milk composition in Montbéliarde cows. 69th EAAP meeting. Dubrovnik, Croatia, August 27 – 30, 2018.

8. **Sanchez M-P.,** Wolf V., El Jabri M., Boussaha M., Taussat S., Laithier C., Delacroix-Buchet A., Brochard M., Boichard. D. **2018**. GWAS on whole genome sequences for

cheese-making traits and milk composition in Montbéliarde cows. Journées scientifiques du département de génétique animale, INRA. Dienné, France, 2-4 octobre, 2018.

9. Gaudilliere N., Gelé M., **Sanchez M.-P.**, El Jabri M., Wolf V., Boichard D., Brochard M., Delacroix-Buchet A., Laithier C. **2018**. La fromageabilité du lait en race Montbéliarde dans les élevages AOP et IGP de Franche-Comté : variabilité et facteurs de variation. Rencontres Recherches Ruminants 2018, Paris, 7-8 décembre 2018.

10. El Jabri M., Rolet-Répécaud O., **Sanchez, M-P.**, Wolf, V., Beuvier E., Notz E., Gaudillière N., Laithier C., Doutart E., Boichard D., Delacroix-Buchet A. **2019**. Bayes and partial least square regression methods for infrared prediction of dairy vat milk casein micelle size. 12th EFITA, Rhodes, Greece, June 27-29, 2019.

11. **Sanchez M.-P.**, Tribout T., Wolf V., El Jabri M., Gaudillière N., Fritz S., Laithier C., Delacroix-Buchet A., Brochard M., Boichard D. **2019**. Towards a genomic evaluation of cheese-making traits including candidate SNP in Montbéliarde cows. Accepté à 70th EAAP meeting. Ghent, Belgium, August 26 – 30, 2019.

12. **Sanchez M.-P.**, Ramayo-Caldas Y., Wolf V., Laithier C., El Jabri M., Boussaha M., Taussat S., Fritz S., Delacroix-Buchet A., Brochard M., and Boichard D. **2019**. Sequence-based GWAS, network and pathway analyses reveal genes co-associated with milk cheese-making properties and milk composition in Montbéliarde cows. Soumis à 16th IMGC meeting. Aarhus, Denmark, November 12 – 14, 2019.

Encadrement de stages

1. Rahman, R. **2017**. Etudes des variants des protéines du lait de vache. Mémoire de L2, Université de Versailles Saint-Quentin-en-Yvelines, Versailles, France, juin - août 2017.

2. Beaumont, M. **2018**. Etude des possibilités de sélection des vaches de race Montbéliarde sur de nouveaux caractères de fromageabilité du lait. Mémoire de fin d'études, UniLaSalle, Beauvais, France, janvier-août 2018.

Autres résultats sur la période 2016-2019

Articles scientifiques

1. Jonas D., Ducrocq V., Fritz S., Baur A., **Sanchez M.-P.**, Croiseau P. **2017**. Genomic evaluation of regional dairy cattle breeds in single-breed and multibreed contexts. Journal of animal breeding and genetics 134:3-13.

2. Teissier M., **Sanchez M.-P.**, Boussaha M., Barbat-Leterrier A., Hozé C., Robert-Granié C., Croiseau P. **2017**. Use of meta-analyses and joint analyses to select variants in whole genome sequences for genomic evaluation: an application in milk production of French dairy cattle breeds. Journal of Dairy Science 101:3126-3139. https://doi.org/10.3168/jds.2017-13587

3. Bouwman A. C., Daetwyler H. D., Chamberlain A. J., Ponce C. H., Sargolzaei M., Schenkel F.S., Sahana G., Govignon-Gion A., Boitard S., Dolezal M., Pausch H., Brøndum R. F., Bowman P. J., Thomsen B., Guldbrandtsen B., Lund M. S., Servin B., Garrick D. J., Reecy J., Vilkki J., Bagnato A., Wang M., Hoff J. L., Schnabel R. D., Taylor J. F., Vinkhuyzen A. A. E., Panitz F., Bendixen C., Holm L. E., Gredler B., Hozé C., Boussaha M., **Sanchez M.-P.,** Rocha D., Capitan A., Tribout T., Barbat A., Croiseau P., Drögemüller C., Jagannathan V., Vander Jagt C., Crowley J.J., Bieber A., Purfield D. C., Berry D. P., Emmerling R., Götz K. U., Frischknecht M., Russ I, Sölkner J., Van Tassell C. P., Fries R., Stothard P., Veerkamp R. F., Boichard D., Goddard M. E. and Hayes B. J. **2018**. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. Nature Genetics 50:362-367. https://doi.org/10.1038/s41588-018-0056-5

Communications à des congrès / séminaires

13. **Sanchez M.P.**, Jonas D., Baur A., Ducrocq V., Hoze C., Saintilan R., Phocas F., Fritz S., Boichard D., Croiseau P. **2016**. Mise en place d'une évaluation génomique en races Abondance, Tarentaise et Vosgienne. Rencontres Recherches Ruminants 2016, Paris, 7-8 décembre 2016.

14. **Sanchez M-P.**, Guatteo R., Davergne A., Grohs C., Capitan A., Blanquefort P., Delafosse A., Joly A., Ngwa-Mbot D., Biet F., Fourichon C., Boichard D. **2016**. Whole genome association analysis of resistance / susceptibility to paratuberculosis in French Holstein and Normande cattle. 13th International Colloquium on Paratuberculosis, Nantes, France, 20-24 June 2016.

15. Marete A., Boussaha M., Rocha D., Fritz S., Michot P., Capitan A., **Sanchez M-P.**, Letaief R., Baur A., Sando Lund M., Boichard D. **2016**. Candidate variant confirmation in three French dairy breeds: GWAS exploits on big data. 5th ICQG, Madison, Wisconsin, USA, 12-17 June 2016.

16. **Sanchez M-P.**, Guatteo R., Davergne A., Grohs C., Capitan A., Blanquefort P., Delafosse A., Joly A., Fourichon C., Boichard D. **2016**. GWAS of resistance to paratuberculosis in French Holstein and Normande cattle. 67th EAAP meeting. Belfast, Ireland, August 29 – September 2, 2016.

17. **Sanchez M-P.**, Jonas D., Baur A., Ducrocq V., Hozé C., Saintilan R., Phocas P., Fritz S., Boichard D., Croiseau P. **2016**. Implementation of genomic selection in three French regional dairy cattle breeds. 67th EAAP meeting. Belfast, Ireland, August 29 – September 2, 2016.

18. Tribout T., Barbat M., Govignon-Gion A., Launay A., Lefebvre R., Barbat A., Boussaha M., Croiseau P., **Sanchez M-P.**, Fritz S. **2016**. Using whole genome sequences to identify QTL for udder health and morphology in French dairy cattle. 67th EAAP meeting. Belfast, Ireland, August 29 – September 2, 2016.

19. Venot E., Boichard D., Ducrocq V., Croiseau P., Fritz S., Baur A., Saintilan R., Tribout T., **Sanchez M-P.**, Boulesteix P., Astruc J.-M., Legarra A., Robert-Granie C., Carillier C., Palhiere I., Tortereau F., Rupp R., Larroque H., Loywick V., Mattalia S. **2016**. French genomic experience: genomics for all ruminant species. Presented at 40. Biennal session of ICAR, Puerto Varas, CHL (2016-10-24 - 2016-10-28).

20. Croiseau P., Tribout T., Boichard D., **Sanchez M.P.**, Fritz S. **2017**. Use of whole sequence GWAS to improve genomic evaluation in dairy cattle. 25th Plant & animal Genome. San Diego, CA, United States, January 14-18th, 2017.

21. Boichard. D., Boussaha M., Capitan A., Rocha D., Hozé C., **Sanchez M-P.**, Tribout T., Letaief R., Croiseau P., Grohs C., Li W., Harland C., Charlier C., Lund M.S. Sahana G., Georges M., Barbier S., Coppieters W., Fritz S., Guldbrandtsen B. **2018**. Experience from large scale use of the EuroGenomics custom SNP chip in cattle. In proceedings of the 11th World Congress on Genetics Applied to Livestock Production: 11th WCGALP, 12-16 Feb. 2018, Auckland, New Zealand.

22. Croiseau P., Tribout T., Boichard. D., **Sanchez M-P.**, Fritz S. **2018**. Use of whole sequence GWAS to improve genomic evaluation in dairy cattle. In proceedings of the 11th World Congress on Genetics Applied to Livestock Production: 11th WCGALP, 12-16 Feb. 2018, Auckland, New Zealand.

23. **Sanchez, M.P.,** Guatteo, R., Davergne, A., Grohs, C., Taussat, S., Blanquefort, P., Delafosse, A., Joly, A., Fourichon, C., Boichard, D. **2018**. Sequence-based association study of resistance to paratuberculosis in Holstein and Normande cattle. 69th EAAP meeting. Dubrovnik, Croatia, August 27 – 30, 2018.

24. Croiseau P., Hozé C., Fritz S., **Sanchez M-P.,** Tribout T. **2018**. Inclusion of candidate mutations in genomic evaluation for French dairy cattle breeds. 69th EAAP meeting. Dubrovnik, Croatia, August 27 – 30, 2018.

25. Bourdon C., Boussaha M., Rupp R., **Sanchez M-P.**, Tribout T., Bardou P., Aujean E., Tosser-Klopp G., Le Provost F. Recherche d'associations entre microARNs, variants génétiques et QTL laitiers chez les bovins, caprins et ovins. Rencontres Recherches Ruminants 2018, Paris, 7-8 décembre 2018.

Encadrement de stages

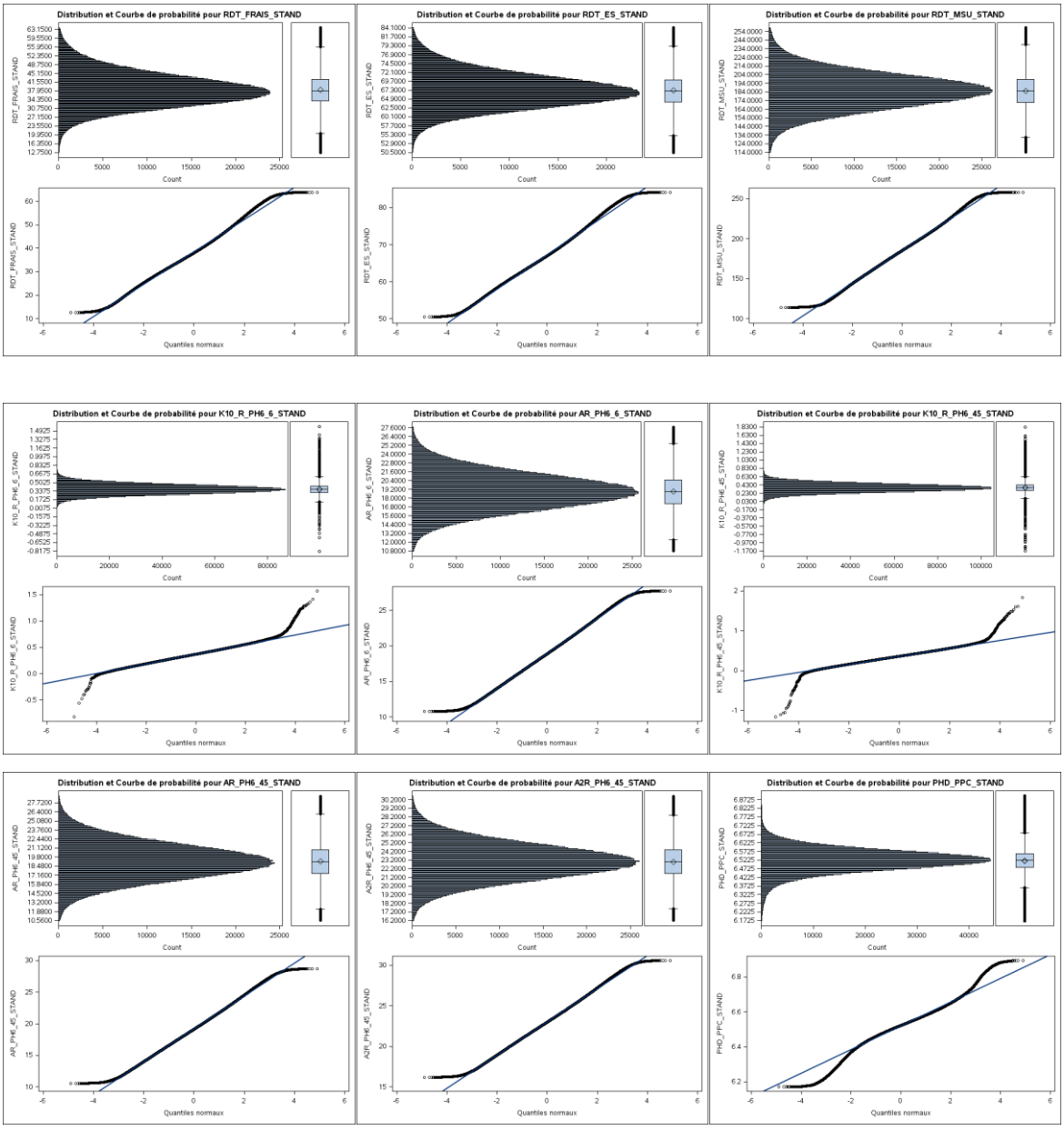
1. Saout, J. **2018**. Etude du déterminisme génétique de la résistance à la paratuberculose bovine dans les races Holstein et Normande. Mémoire de Master 2 BMC, Université de Rennes 1, UFR SVE, Rennes, France, janvier-juin 2018.

Annexes

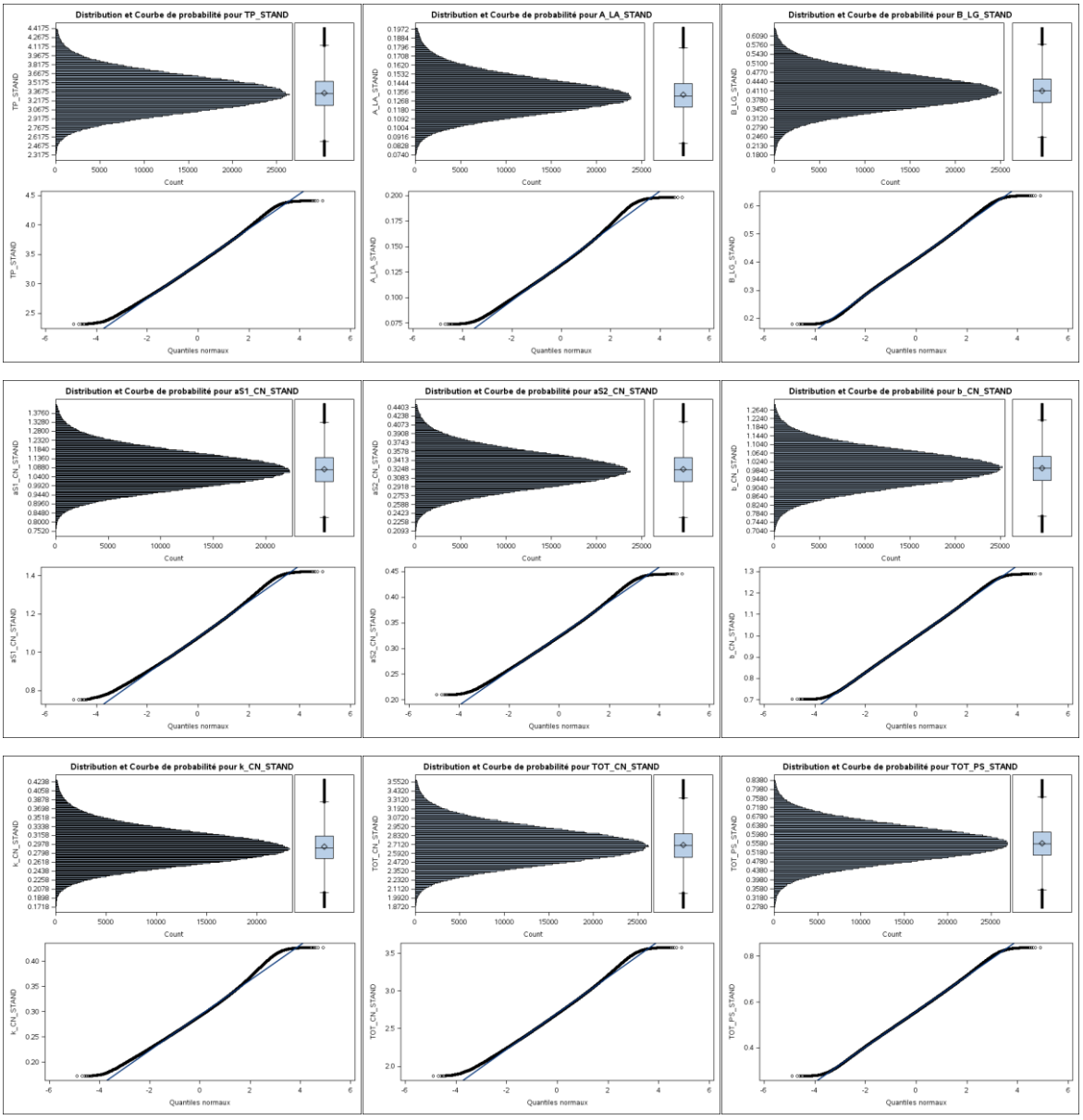
Annexes

Annexe 1 - Distributions des caractères prédits par spectrométrie MIR dans le projet *From'MIR*

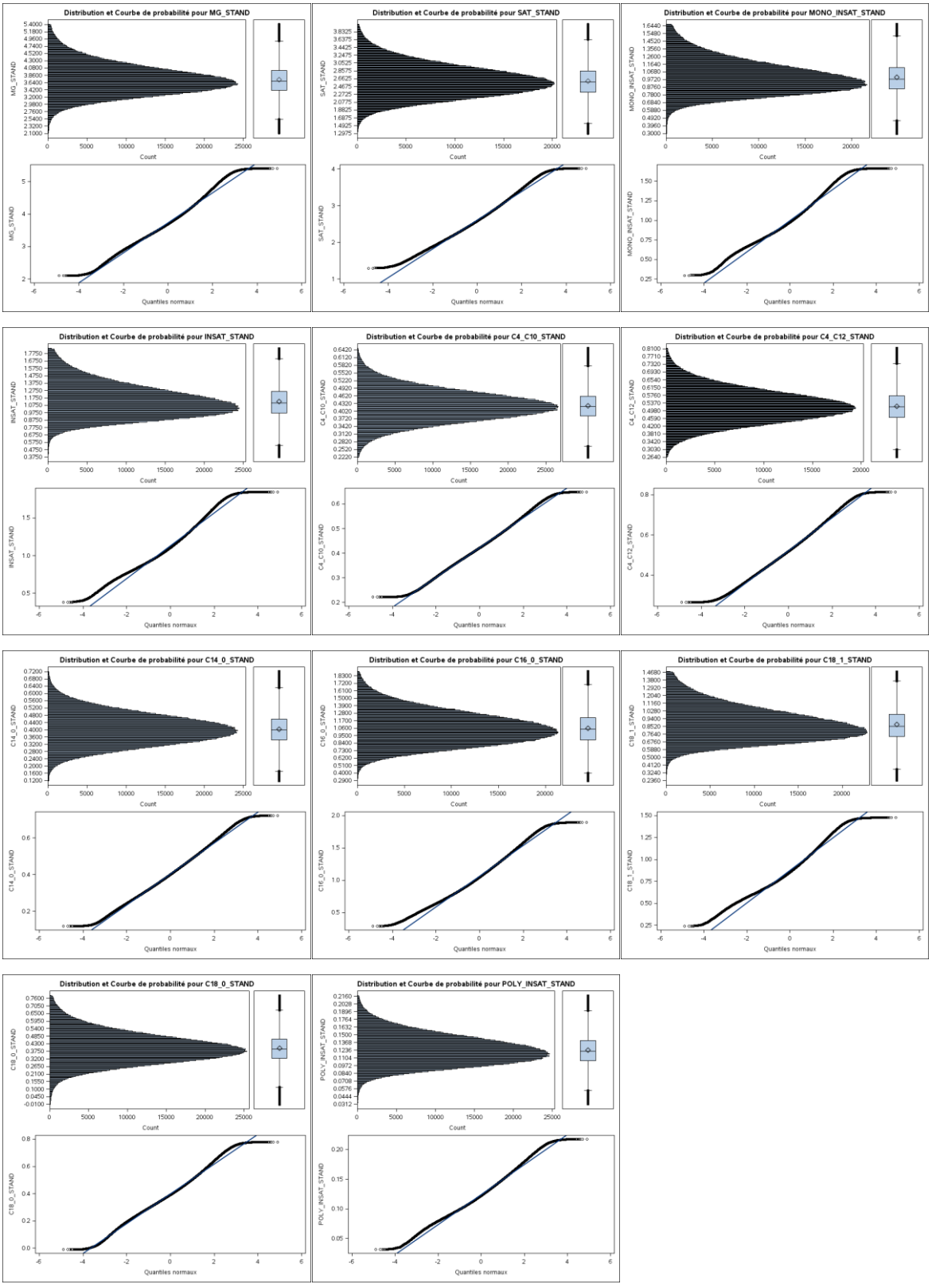
Critères fromagers



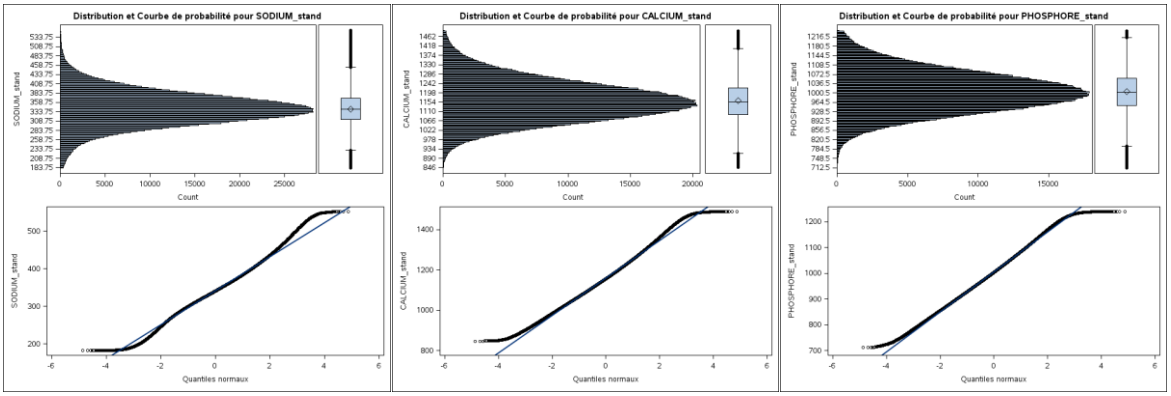
Composition protéique

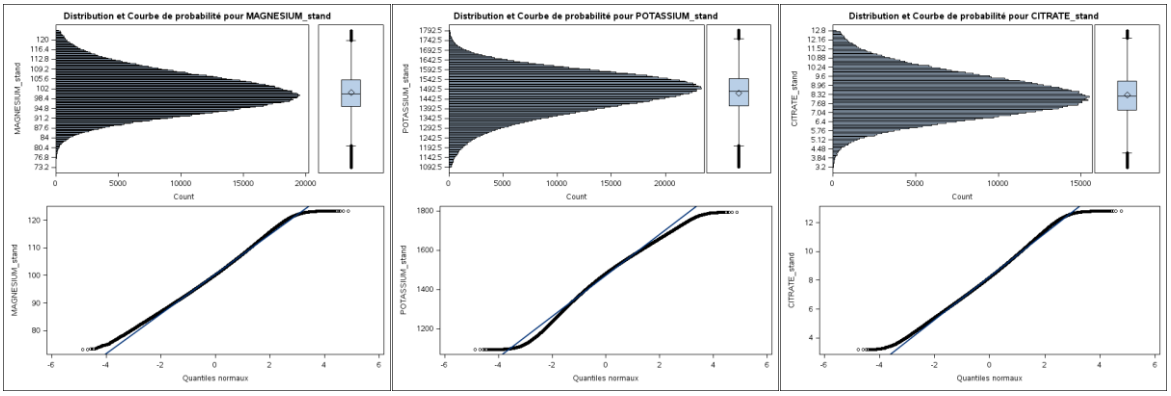


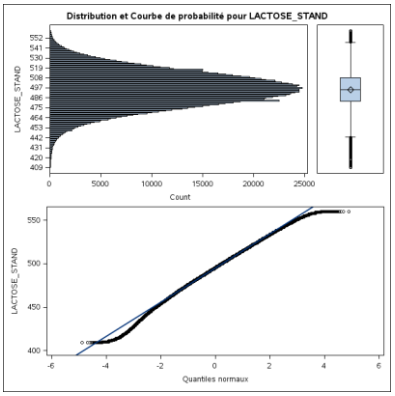
Composition en acides gras



Composition en minéraux, citrate et lactose

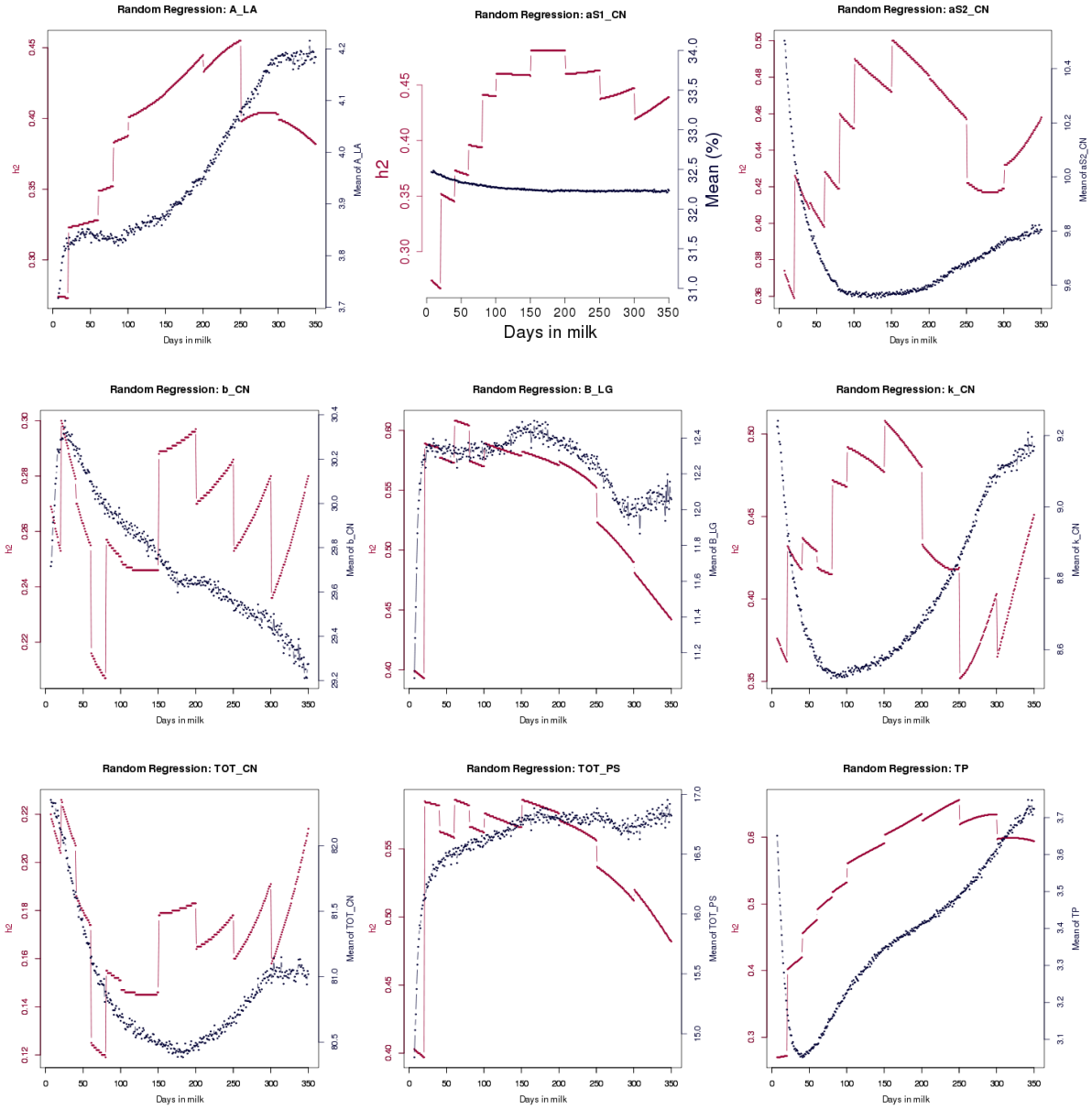




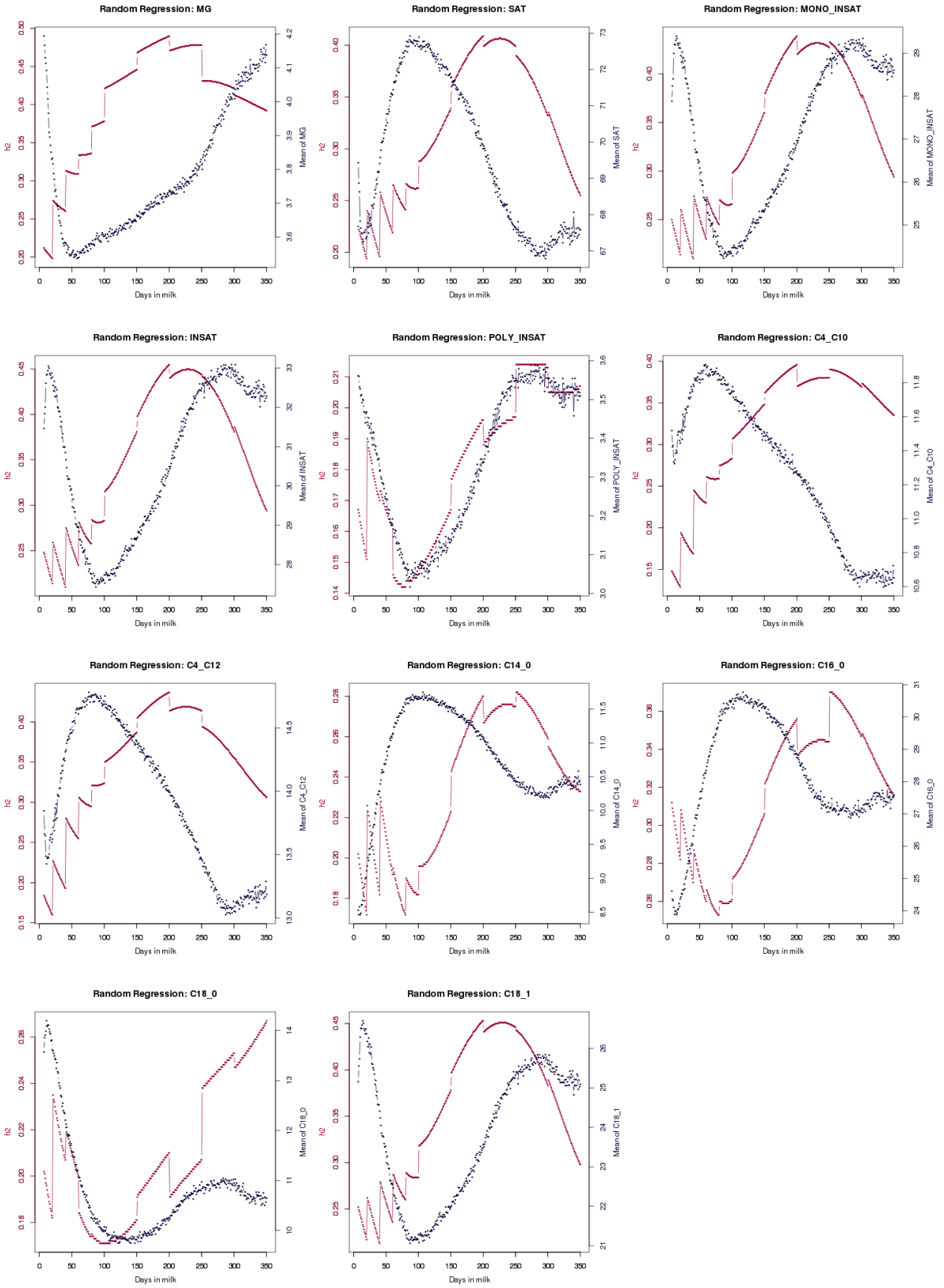


Annexe 2 - Courbes de lactation et héritabilités estimées par régression aléatoire

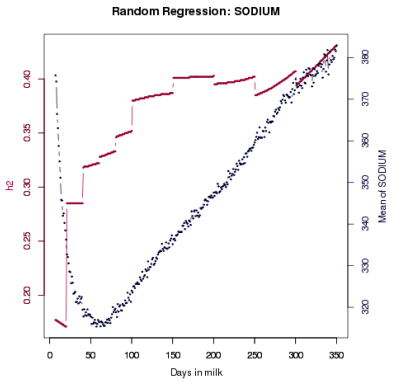
Composition protéique

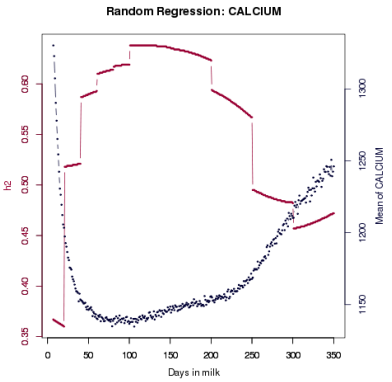


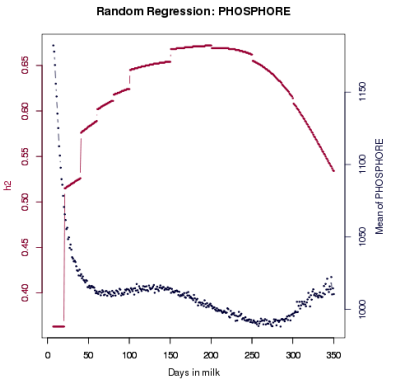
Composition en acides gras

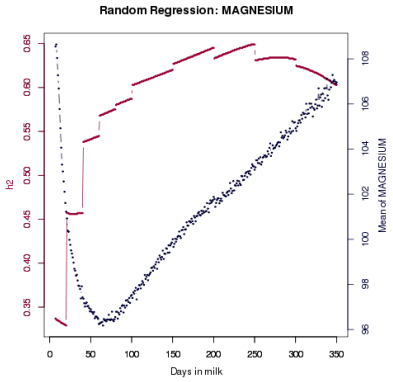


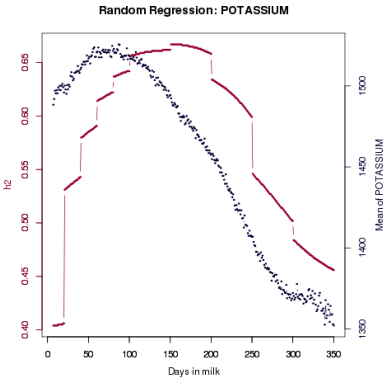
Composition en minéraux, citrate et lactose

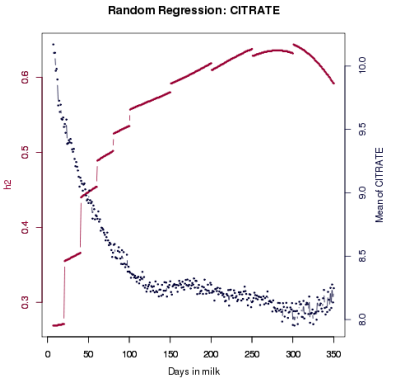


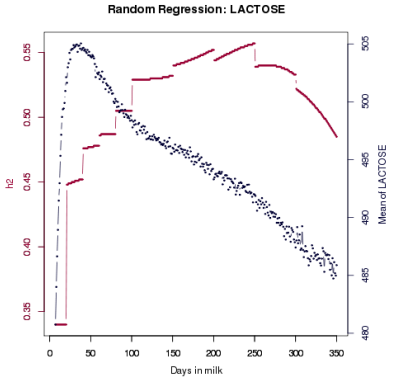












Annexe 3 – Variants génétiques des lactoprotéines

Liste des 15 mutations non synonymes (SNP) de la puce EuroG10K permettant de caractériser les principaux variants protéiques de 4 lactoprotéines, les lignes grisées correspondent aux SNP pour lesquels nous n'avons pas pu exploiter les génotypes dans cette thèse.

Gène	BTA	Position SNP (pb)	Allèles (REF/ALT)	Score SIFT*	Substitution	Variant Protéique REF	Variant Protéique ALT	TOP/BOTTOM**
CSN1S1	6	87148464	G/A	0,54	A68T	B	D	[A/G]
CSN1S1	6	87157262	A/G	0,22	E207G	B	C	[A/G]
CSN2	6	87181364	G/A	0,90	P167L	A2	F	[A/G]
CSN2	6	87181453	G/C	0,79	S137R	A2	B	[C/G]
CSN2	6	87181501	G/T	0,58	H121Q	A2	A3	[A/C]
CSN2	6	87181542	T/G	0,10	M108L	A2	I	[A/C]
CSN2	6	87181619	G/T	0,02	P82H	A2	A1, B, C, F, G	[A/C]
CSN2	6	87183031	C/T	0,18	E52K	A2	C	[A/G]
CSN3	6	87390612	A/C	0,02	D169A	A	B, B2	[A/C]
CSN3	6	87390632	A/G	0,03	S176G	A	E	[A/G]
LGB	11	103302553	G/C	0,03	E61Q	B	D	[C/G]
LGB	11	103302586	A/C	0,01	I72L	B	W	[A/C]
LGB	11	103302597	G/T	0,76	Q75H	B	C	[A/C]
LGB	11	103303475	G/A	0,62	G80D	B	A	[A/G]
LGB	11	103304757	C/T	0,76	A134V	B	A	[A/G]

* Pour une mutation non synonyme, le score SIFT, entre 0 et 1, prédit dans quelle mesure la fonction de la protéine pourrait être affectée par le changement d'acide aminé ; la mutation est considérée délétère si le score SIFT < 0,05, tolérée si le score SIFT > 0,05. ** Format TOP/BOTTOM Illumina avec en couleur, les SNP pour lesquels le format REF/ALT ne correspond pas au format TOP/BOTTOM.

Annexe 4 – Restitution et diffusion des résultats *From'MIR*

Programmes des journées de restitution

Régionale (Dannemarie-sur-Crête, 26/04/18)



Journée de restitution FROM'MIR :
développer des outils de prédiction et de conseils pour maîtriser la fromageabilité des laits.
PROGRAMME

10h00-10h30 • **Introduction** P. Grosperrin (Conseil Elevage 25-90), P. Duboz (CIGC), C. Laithier (Idele), V. Wolf (Conseil Elevage 25-90)

10h30-11h40 • **Du lait fromageable ?**
- **Qu'est-ce qu'un lait fromageable ?** A. Delacroix-Buchet (INRA Jouy-en-Josas)
- **Prédire la fromageabilité par le Moyen Infra-Rouge : mise au point des équations et validation sur minifabrications**
  M. El Jabri (Idele), S. Gavoye (Actalia) & C. Laithier (Idele)
- **Quid de l'effet échelle – Vache - Troupeau - Cuve de fromagerie - sur les liens entre composition physico-chimique fine et aptitude fromagère des laits ?**
  E. Beuvier (INRA Poligny) & S. Gavoye (Actalia)

Clôture de séance E. Notz (CTFC)

11h40-13h00 • **Quels sont les facteurs qui impactent la fromageabilité des laits ?**
- **À l'échelle individuelle ?**
  M. Gelé (Idele), M-P Sanchez (INRA Jouy-en-Josas) & M. Brochard (UMOTEST)
- **À l'échelle troupeaux ?** N. Gaudillière (Conseil Elevage 25-90)
- **À l'échelle des cuves de fromagerie ?**
  E. Notz (CTFC) & Y. Bouton (CIGC)

Clôture de séance Y. Bernard (Eva-Jura)

14h30-16h00 • **FROM'MIR sur le terrain**
- **Illustration : utilisation de FROM'MIR à la fromagerie de Loray**
  N. Gaudillière (Conseil Elevage 25-90) & E. Notz (CTFC)
- **Table ronde : Discussion sur les résultats de FROM'MIR**
  S. Fonquernie (Conseil Régional de Bourgogne Franche-Comté),
  B. Richard (Président de la Fruitière de Loray),
  C. Vermot Desroches (Président du CIGC),
  E. Fevrier (Président du syndicat du Mont d'Or),
  A. Alix (Conseil Elevage 25-90) & E. Chevalier (CIGC, Monts et Terroirs)

16h00-16h30 • **Les perspectives de FROM'MIR**
- **FROM'MIR demain ?** P. Grosperrin (Conseil Elevage 25-90)

Nationale (Paris, 28/06/18)



Journée de restitution FROM'MIR :
développer des outils de prédiction et de conseils pour maîtriser la fromageabilité des laits.
Programme prévisionnel

09h30-10h00 **Accueil café**

10h00-10h30 **Introduction** D. Pourchet, P. Grosperrin (Conseil Elevage 25-90), T. Geslain (CNIEL), C. Laithier (Idele), V. Wolf (Conseil Elevage 25-90)

Du lait fromageable ?

10h30-12h00
- **Qu'est-ce qu'un lait fromageable ?** A. Delacroix-Buchet (INRA Jouy-en-Josas)
- **Prédire la fromageabilité par le Moyen Infra-Rouge : mise au point des équations et validation sur minifabrications**
  - **Dispositif mis en place :** C. Laithier (Idele) & V. Wolf (CEL 25-90)
  - **Résultats obtenus :** M. El Jabri (Idele), P. Trossat (Actalia) & C. Laithier (Idele)
- **Quid de l'effet échelle – Vache - Troupeau - Cuve de fromagerie - sur les liens entre composition physico-chimique fine et aptitude fromagère des laits ?**
  E. Beuvier (INRA Poligny) & S. Gavoye (Actalia)

Quels sont les facteurs qui impactent la fromageabilité des laits ?

12h00-12h30 puis 14h00-15h00
- **À l'échelle individuelle ?**
  M. Gelé (Idele), M-P Sanchez (INRA Jouy-en-Josas) & M. Brochard (UMOTEST)
- **À l'échelle troupeaux ?** N. Gaudillière (Conseil Elevage 25-90)
- **À l'échelle des cuves de fromagerie ?**
  E. Notz (CTFC) & Y. Bouton (CIGC)

Quelle suites à FROM'MIR ?

15h00-15h45
- **Suite donnée à FROM'MIR en Franche Comté : construction d'un observatoire de la fromageabilité**
  E. Chevalier (CIGC)
- **Témoignages d'autres filières intéressées pour travailler sur la fromageabilité du lait**

Conclusions générales

15h45-16h30 P. Martin (INRA Jouy en Josas), M. Marguet (Idele) & T. Geslain (CNIEL)

Quelques exemples d'articles, fiche technique et Newsletter du projet From'MIR

L'éleveur laitier - n°272 – juillet-août 2018



CONSEIL ELEVAGE 25-90

Mir. Il est possible d'utiliser le spectre infrarouge Mir du lait pour estimer de façon fiable le rendement fromager et l'aptitude à la coagulation des laits individuels. Cela vaut aussi pour certains critères de fromageabilité des laits de troupeaux.

RECHERCHE

Il y a de l'avenir dans la fromageabilité des laits

Pilote. From'Mir, programme pilote sur la fromageabilité, ouvre un boulevard aux entreprises de sélection... mais pas seulement.

Le programme de recherche pilote sur la fromageabilité des laits From'Mir, orchestré autour des filières fromagères et d'élevage de Franche-Comté, est arrivé à terme. Le voile qu'il lève et les perspectives qu'il porte méritaient bien les 1,5 million d'euros investis et le soutien du Cniel, notamment. From'Mir montre qu'il est possible d'utiliser le spectre infrarouge Mir du lait pour estimer de façon fiable le rendement fromager et l'aptitude à la coagulation des laits individuels. Cela vaut aussi pour certains critères de fromageabilité des laits de troupeaux, moins à l'échelle des laits de cuve en fromagerie.

Pour aboutir à ces équations de prédiction Mir, le travail a porté sur le lait de 250 montbéliardes, 100 troupeaux et 70 cuves de fromagerie. Il a consisté aussi à vérifier sur des minifabrications la pertinence des prédictions. From'Mir confirme également l'intérêt d'aller au-delà du TP ou du TB/TP, prédicteurs classiques de la fromageabilité, pour analyser plus finement la composition du lait (taille des micelles de caséine, composition en minéraux).

Sélectionner sur ces critères aurait du sens

Ce travail de recherche ouvre aussi un boulevard à la génétique. L'héritabilité assez forte constatée de certains critères de fromageabilité prédictibles par le Mir (rendement, égouttage, coagulation) offre en effet la possibilité de les sélectionner de façon efficace. Cette sélection serait favorable aux taux et sans antagonisme fort ou neutre avec le potentiel lait ou les critères fonctionnels. Les producteurs de lait à comté, désormais plafonnés au niveau de leur productivité laitière par hectare, apprécieront. À défaut de produire plus de lait, ils pourront travailler à un lait plus fromageable... avec, au final, plus de kilos de comté. From'Mir a d'ores et déjà identifié des gènes expliquant la variabilité de la fromageabilité... ouvrant la voie à une indexation.

Effet positif des fourrages de qualité en quantité

Outre la sélection, d'autres leviers de pilotage à l'échelle du troupeau sont à portée de main des éleveurs pour améliorer la fromageabilité. From'Mir souligne l'effet très significatif de la qualité des fourrages et du niveau d'ingestion des animaux sur l'égouttage (rendement extrait sec) et la coagulation. Sans surprise, la quantité de concentrés apportée joue positivement. En revanche, les effets de la composition du concentré sont beaucoup plus ténus. Au pâturage, où nombre de facteurs ont été ciblés, un seul apparaît flagrant : l'accès au point d'eau. Il joue aussi sur le rendement extrait sec.

JEAN-MICHEL VOCORET

La Terre de chez nous – 29 juin 2018

Restitution des travaux de FROM'MIR

La fromageabilité en un coup d'œil

Valérie Wolf travaille au pôle de recherche et développement de Conseil Elevage 25/90, plus particulièrement sur FROM'MIR, un projet inédit au niveau français dont l'objectif est d'estimer la fromageabilité du lait à travers une simple analyse.

FROM'MIR est un programme qui a été lancé en 2014. Ce projet est porté par Conseil élevage 25/90, en lien avec l'Institut de l'Elevage et de nombreux autres partenaires techniques et financiers.

Du lait à la fromageabilité

FROM'MIR, à quoi cela va-t-il servir ? Valérie Wolf qui travaille au pôle recherche et développement de Conseil élevage 25/90 l'explique : *« On est dans une région à vocation fromagère. On a des analyses de lait. On sait qu'à partir de celles-ci, on peut estimer le taux butyrique, le taux protéique. Ce sont les analyses Moyen-Infra Rouge du lait (MIR). Ces analyses sont déjà utilisées dans le cadre du conseil en élevage et c'est le même principe qui est utilisé pour le paiement du lait à la qualité. Or, on savait qu'on pouvait aller plus loin, avoir accès aux protéines, aux différents acides gras qui composent ce lait, on voulait aller encore plus loin et estimer la fromageabilité ».* La fromageabilité, c'est le potentiel qu'a un lait pour être transformé en fromage. Les équipes du programme se sont notamment penchées sur deux critères :

- Le rendement en frais et en extrait sec : on peut assimiler le rendement frais à un égouttage physique, c'est la quantité de caillé obtenue à partir d'un échantillon de lait. Quant au rendement en extrait sec il permet de calculer ce qui est retenu de l'extrait sec du lait dans l'extrait sec du caillé.
- L'aptitude à la coagulation du lait, c'est-à-dire son aptitude à coaguler sous l'action de la présure et à donner un gel plus ou moins ferme.

A partir d'analyses de lait de références, les équipes de FROM'MIR ont donc obtenu des équations permettant de passer de l'analyse MIR du lait à une estimation de la fromageabilité.

L'utilisation de la spectroscopie MIR est très intéressante pour plusieurs raisons : c'est une analyse connue et de routine, qui se fait déjà en grand nombre à un coût modeste. De plus elle produit un spectre, une donnée brute qui peut être stockée telle quelle et être analysée à nouveau si les pistes de recherche et développement évoluent. Pour Valérie Wolf, *« c'est une des grandes plus-values de ce programme, avoir permis de créer une grande bibliothèque de spectres MIR, qui à l'aide des équations développées dans FROM'MIR, permettra d'étudier la fromageabilité du lait à grande échelle ! »*

A l'heure du bilan

Une des grandes originalités de FROM'MIR est de travailler à trois échelles : de la vache, du troupeau et de la cuve du fromager. A l'heure du bilan, on peut constater que les équations obtenues sont fiables à deux niveaux. Ainsi *« pour une vache, on sait très bien estimer le rendement et l'aptitude à la coagulation du lait »* se félicite Valérie Wolf. De même *« à l'échelle des laits de troupeaux, c'est un peu moins bon, mais cela fonctionne bien pour certains paramètres »*. Mais à l'échelle des laits de cuve, il faudra encore du travail pour à améliorer les résultats. Cela vient du fait que pour construire un bon modèle, il faut de la variabilité dans les échantillons comparés et à l'échelle des fromageries la variabilité entre les laits est réduite.

Cette bibliothèque de données va donc permettre d'étudier les facteurs d'influence de la fromageabilité. Ce travail a déjà commencé en identifiant les impacts de l'alimentation mais également de la génétique. Ainsi il a été mis en évidence au niveau des laits de troupeaux que la qualité et la quantité du fourrage ingérée ont un impact très fort sur la fromageabilité du lait. Ainsi *« FROM'MIR a clairement mis en évidence que les vaches rationnées montraient une détérioration des TP et TB et donc de la fromageabilité »* détaille Valérie Wolf. L'utilisation de concentrés a aussi un impact, mais pas aussi évident qu'on aurait pu le croire. Les acteurs du projet ont ainsi pu constater que *« l'augmentation des concentrés a un impact modeste sur la fromageabilité qui s'explique par son effet favorable sur le TP »*. Le volet génétique a également été étudié. Celui-ci a un impact important et *« il a pu être clairement démontré que la fromageabilité est un critère héritable qui se transmet de génération en génération. Il y a donc un intérêt à travailler dessus pour la race montbéliarde »*, l'étude n'ayant concerné que les laits issus de cette race.

A présent, la demande des professionnels encadrant le projet est de se laisser le temps pour s'approprier ces nouveaux indicateurs, de voir comment ils évolueront dans le temps. *« Mais il y a une envie à terme d'aller plus loin que la cuve et d'aller jusqu'au fromage, afin de voir s'il existe une filiation entre fromageabilité des troupeaux et qualité sensorielle des fromages »* s'enthousiasme Valérie Wolf.

Morgane Branger



Beaucoup d'acteurs ont pu assister au séminaire de restitution des résultats du programme. Ainsi, celui-ci a été soutenu financièrement par de nombreux acteurs : ministère de l'Agriculture, de l'Agro-alimentaire et de la Forêt, Centre National Interprofessionnel de l'Economie Laitière, Union Régionale des Fromages d'Appellation d'origine Comtois et en région Bourgogne Franche-Comté (© Conseil élevage 25-90).

Faire le lien entre fromageabilité et pratiques

La fromageabilité représente le revenu des éleveurs. L'améliorer, c'est améliorer leur paye. D'où l'importance de l'étude réalisée sur le terrain, à la coopérative de Loray, afin de relier pratiques agricoles, pratiques du fromager et fromageabilité.

Bertrand Richard est le président de la coop de Loray. Sa coopérative a accueilli en début d'année les techniciens de conseil élevage 25/90 venus expérimenter sur le terrain FROM'MIR.

La Terre de Chez Nous (TCN.) : Pouvez-vous nous décrire en quelques mots votre coopérative ? Et comment vous avez été démarché pour prendre part à ce programme ?

Bertrand Richard, président de la coopérative de Loray (B. R.) : On travaille à peu près 2,8 millions de litres de lait à l'année. Sur la coop, nous sommes dix exploitants, deux fromagers et un apprenti. Le directeur du Contrôle laitier m'a contacté car ils recherchaient une petite coopérative, où tout le monde était au Contrôle laitier. Il m'a sollicité pour savoir si on voulait participer au programme FROM'MIR. On m'a expliqué le programme, j'ai donné mon accord de principe avant de solliciter mes producteurs. Ils sont revenus présenter le programme à tous les producteurs de la coop et on a fixé une date pour cette journée d'étude sur le terrain.

TCN : Comment s'est déroulée cette journée d'étude ?

B. R. : Sur une journée au mois de janvier, les équipes du Conseil élevage sont venues chez chaque producteur pour analyser le lait, à la traite du soir et du matin, pour suivre celui-ci le lendemain en cuve et voir ce qui allait se passer au niveau de la fromagerie. Les fromages issus de ces laits-là ont été identifiés pour être ensuite suivis à l'affinage.

En parallèle de ces prélèvements de lait, on nous a demandé de préparer un échantillon du fourrage distribué la veille et les jours suivants. Chaque exploitation a eu un audit technique pour comparer leur fonctionnement.

TCN : Depuis, Conseil élevage est revenu vers vous avec les premiers résultats de leurs études. Qu'est-ce que cela a apporté à votre coopérative ?

B. R. : On a eu une restitution pour les producteurs et travailleurs de la coop. On a fait le point sur la fromageabilité de nos laits en lien avec nos pratiques agricoles, mais aussi en lien avec les habitudes des fromagers. Ça a permis à chacun de découvrir les techniques de l'autre et leur impact sur le fromage. Grâce à ça, on s'est dit qu'on allait se mettre autour d'une même table, pour échanger d'avantage et mieux se comprendre entre agriculteurs et fromagers. Avec notre petite structure, on a déjà un fromager qui participe beaucoup à la vie de la coop. On est assez proches, mais là je pense que ça nous a encore rapprochés car il a mieux compris les techniques agricoles. D'un point de vue technique, on a vu que pour la fromageabilité il y a des écarts entre tous les laits produits sur nos fermes. J'ai retenu que là où il y avait les meilleurs fourrages, il y avait les meilleurs laits, avec les meilleurs taux, qui influençaient ensuite sur la qualité des fromages et les rendements.

TCN : Que retirez-vous de cette expérience ?

B. R. : Je suis content de l'avoir fait. C'était une démarche lourde, à un moment pas forcément opportun pour nous. Mais je pense qu'il faut qu'on continue de développer ces recherches pour continuer à développer la qualité de nos fromages. Il y a des choses à faire. Mais attention, là c'est mon avis personnel, il faudra veiller à ne pas aller trop loin et à ne pas uniformiser trop nos systèmes pour obtenir une meilleure fromageabilité. Car la grande richesse de notre produit, c'est sa diversité.

M. B.



Demain, pourra-t-on prévoir la qualité des fromages à partir d'un simple échantillon de lait ? © Conseil élevage 25-90.

Réussir lait – 12 juin 2018

Réussir lait **12 juin 2018** à 11h00 | Par Annick Conté partager :

Lait Fromage Bovins Lait Transformation Élevage Bovin

From'Mir : estimer et piloter la fromageabilité des laits

Conduit en Franche-Comté pendant quatre ans, le programme From'Mir montre tout l'intérêt de s'intéresser à la composition fine du lait.

Abonnez-vous Imprimer Envoyer



Le programme s'appuie sur le spectre Mir, utilisé en routine pour le paiement du lait, pour estimer la fromageabilité des laits individuels et de troupeaux. - © Jérôme Chabanne

Quand on s'intéresse à la fromageabilité des laits, il ne faut pas se limiter aux taux butyreux et protéique. C'est ce que démontre le programme From'Mir qui se termine en juin: le spectre Mir (spectrométrie moyen infrarouge), utilisé en routine pour le paiement du lait et véritable reflet de la composition physicochimique du lait, permet d'estimer le rendement fromager et l'aptitude à la coagulation des laits individuels. Il en est de même pour certains paramètres de fromageabilité à l'échelle des laits de troupeaux. From'Mir confirme aussi l'impact d'éléments plus fins comme la taille des miscelles de caséine et l'importance de la composition en minéraux notamment le calcium.

Mise en évidence de facteurs génétiques

Autre apport du programme: il met en évidence des gènes (codant pour différents types de caséines) expliquant la variabilité de la fromageabilité, ceci en analysant finement le génôme de 20 000 Montbéliardes. Et il montre que les critères de fromageabilité du lait peuvent être sélectionnés efficacement: une sélection sur ces critères serait favorable aux TB et TP et sans antagonisme avec d'autres critères en sélection aujourd'hui. Ce qui permet d'envisager des index « fromageabilité » pour les taureaux et une sélection des vaches. From'Mir ouvre donc de nouvelles pistes de pilotage de la fromageabilité en élevage. Mais la génétique n'est pas le seul levier de pilotage : la qualité des fourrages, le niveau d'ingestion, la prévention des troubles métaboliques sont utilisables.

« *From'Mir a permis de créer une dynamique de travail sur la fromageabilité en Franche-Comté ainsi qu'un partenariat efficace entre des entreprises de l'amont et de l'aval (585 éleveurs et 52 fromageries), des organismes de conseil et de recherche (plus de 70 techniciens et scientifiques)*[(1)], souligne Cécile Laithier, chef de projet, Idele. Pour demain, L'objectif est de créer un observatoire de la fromageabilité en Franche-Comté. Il permettra d'exploiter plus d'un million de spectres Mir collectés en routine, ainsi que d'autres données sur les laits de troupeaux, des cuves jusqu'aux fromages affinés. Cette expérience francomtoise acquise dans From'Mir sera valorisable par d'autres filières. Une journée nationale de restitution est prévue le 28 juin.

La revue laitière française – n°782 – Juin 2018

essentiel

FROM'MIR

Des leviers pour estimer et maîtriser la **fromageabilité** des laits



Né d'une réflexion générale,le projet From'Mir sur la fromageabilité des laits s'appuie sur l'outil d'analyse par infrarouge, employée dans le paiement du lait. Une méthode d'analyse rapide, précise, robuste et peu onéreuse qui peut faire parler l'échantillon de lait au-delà des seuls taux butyrique et protéique.

Le programme de recherche From'Mir, conduit en Franche-Comté pendant quatre ans, a révélé l'intérêt d'étudier la composition fine des laits et les paramètres influant la fromageabilité. Un travail impactant pour l'ensemble des filières fromagères.

Estimer la fromageabilité des laits de vache de Franche-Comté, identifier et comprendre les facteurs qui l'influencent pour mieux la maîtriser, tout en valorisant les nombreuses analyses réalisées sur les laits collectés : telles sont les finalités du programme de recherche From'Mir qui, initié il y a quatre ans par les filières AOP locales, s'achève ce mois de juin 2018 en offrant des résultats très prometteurs. Considéré comme pilote pour l'ensemble des filières laitières françaises, et labellisé Vitagora, ce programme a mobilisé, de la part des organisations professionnelles (Cniel, Union régionale des fromages d'appellation d'origine comtois…) et des institutions (ministère de l'Agriculture, de l'Agroalimentaire et de la Forêt, région Bourgogne Franche-Comté…), pas moins de 1,5 million d'euros, auxquels se sont ajoutés des financements des entreprises.

Alors que la qualité des laits, et en particulier leur aptitude à être transformés en fromage, est communément évaluée à partir des taux protéiques et butyreux, le programme From'Mir propose d'aller plus loin.

DE NOUVELLES ANALYSES DE ROUTINE ?

Basé sur la technique d'analyses MIR (spectrométrie moyen infrarouge), il confirme l'intérêt *« de préciser et de faire parler davantage les échantillons de lait »*, en d'autres mots, d'étudier plus finement la composition physicochimique du lait quand on s'intéresse à sa fromageabilité. Ce qui est une première en France !

Les chercheurs ne partaient pas de rien. La somme de connaissances accumulées par les analyses MIR et par des programmes de recherche spécifiques était en effet déjà abyssale. Utilisée en routine, depuis de nombreuses années, pour le paiement du lait et le conseil en élevage, la méthode MIR avait déjà fourni une multitude de données brutes, baptisées « spectres MIR ». Soit de véritables empreintes très dépendantes de la composition physicochimique du lait (protéines, acides gras, minéraux). Aussi, plusieurs programmes de recherche antérieurs avaient apporté leurs lots d'informations fines, comme les programmes PhénoFinLait (protéines) et OptiMir (minéraux). Aujourd'hui, sept millions de spectres sont disponibles dans la seule base de données du Conseil élevage de Franche-Comté ! Mais jusqu'ici, seuls les laits individuels avaient été étudiés, et pas du tout les laits de mélange (troupeaux, cuves de fromageries). Et c'est sur ces laits de mélanges, en plus des laits individuels, que From'Mir a proposé d'investiguer.

« From'Mir est remarquable par son approche de la fromageabilité. Il propose, pour l'évaluer, de s'affranchir de l'exclusivité faite aux taux protéiques et butyreux. Et il crée des liens triangulaires solides entre les critères de fromageabilité, une composition très fine du lait et les différents facteurs influençant cette composition », résume Eric Notz, du CFTC, le Centre technique des fromages comtois.

UN TERRAIN D'ÉTUDES TRÈS CADRÉ

La fromageabilité étant une notion complexe, deux modèles fromagers, pâtes molles et pâtes pressées cuites, ont été étudiés selon trois critères liés à la physicochimie du lait : le rendement fromager (poids du caillé/poids du lait avant coagulation en frais ou en extrait sec mesuré en laboratoire), l'aptitude à la coagulation enzymatique (mesures des temps de prise, de la fermeté du gel et de la vitesse d'organisation du gel) et l'aptitude à l'acidification par les bactéries lactiques. Ces trois critères ont été estimés directement à partir de spectres MIR, sur tous les laits de la chaîne de fabrication : des laits individuels de vaches montbéliardes (250 vaches), des laits de troupeaux (100 troupeaux parmi les 2 000 du contrôle laitier de la zone AOP/IGP de Franche-Comté) et des laits de cuves de fromagerie (70 cuves dans 55 fromageries de Franche-Comté). Les équations d'estimation de ces critères à travers les analyses MIR ont été établies par comparaison avec des mesures réalisées parallèlement en laboratoire (analyses de référence). Les estimations obtenues ont, en outre, été validées en conditions réelles de fabrication à l'aide de minifabrications fromagères.

DES RÉSULTATS PROMETTEURS SUR LES LAITS INDIVIDUELS

Qu'il s'agisse de pâtes molles ou de pâtes pressées cuites, les spectres MIR étudiés se sont montrés très utiles sur bien des points. Par exemple, pour estimer le rendement fromager et l'aptitude à la coagulation des laits individuels. Ils ont aussi confirmé que la fermeté du gel évaluée sur le modèle pâte pressée cuite n'est pas due uniquement aux taux protéiques et butyreux (35 % de la fermeté expliquée par le TP et le TB), mais qu'elle a aussi un lien étroit avec la taille des micelles de caséines et le taux de caséines kappa. Autre exemple, en pâtes pressées cuites, il était clair que les temps de prise du gel s'allongeaient quand les taux de caséines bêta et kappa étaient bas. Par contre, à l'échelle des laits de troupeaux, en pâtes molles comme en pâtes pressées cuites, les liens avec la fermeté du gel de caillé sont plus faibles. Et ceci est encore moins vrai sur les laits de cuve.

LA GÉNÉTIQUE, UN FACTEUR TRÈS INFLUENT

Autre résultat : la mise en évidence de facteurs génétiques influençant clairement la fromageabilité des laits. À partir de l'analyse fine du génome de 20 000 vaches montbéliardes, le programme a, en effet, souligné l'existence de gènes expliquant la variabilité du comportement des laits en fromagerie (gènes codant pour les différents variants des différents types de caséines : kappa (A,B,E), beta, alpha s1 et alpha s2 et la beta lactoglobuline). La création d'une indexation génétique des critères de fromageabilité par animal est dès lors envisageable. Et, en jouant sur la stratégie de réforme et/ou de sélection des taureaux, il deviendra tout à fait possible d'améliorer progressivement la génétique des vaches montbéliardes de génération en génération sur des critères de fromageabilité ciblés. D'autant plus que, selon From'Mir, les critères de fromageabilité peuvent être sélectionnés efficacement par la génétique, sans antagonisme fort avec les autres critères en sélection aujourd'hui (niveau de production laitière, taux de matière grasse et de matière protéique) ! *« Le volet génétique de From'Mir profitera à toutes les filières fromagères françaises qui pourront trouver de la valorisation à travers le profilage de leurs races laitières »*, souligne André Alix, président de Conseil élevage 25-90.

LA CONDUITE DES TROUPEAUX AUSSI

D'autres leviers sont également efficaces au niveau du pilotage des troupeaux, les plus ••• *des laits mis en œuvre, et qu'elles s'additionnent dans la cuve du fromager pour apporter le potentiel de diversité de goût de nos AOP. »*

Quoi qu'il en soit, ce travail réaffirme l'importance du dialogue entre tous les maillons de la chaîne laitière. D'ores et déjà, les estimations de la fromageabilité, obtenues par le programme à l'échelle des vaches et des troupeaux, peuvent servir de base de travail pour le fromager, pour mieux comprendre ce qui se passe dans sa cuve et anticiper des adaptations technologiques. De la même façon, face à des évolutions constatées de la fromageabilité lors de ses fabrications, le fromager peut interroger les spectres MIR de son lait de mélange et de celui des éleveurs collectés sur les données de fromageabilité de leur lait. Une démarche de progrès peut alors se déclencher sur cette base. La méthode d'analyse MIR permet ainsi de créer une bibliothèque d'archive de la fromageabilité chimique potentielle des laits. Cette bibliothèque peut ensuite être analysée pour mieux comprendre les liens existants entre la qualité sensorielle des fromages et la qualité fromagère du lait transformé.

Et demain ? Quelle suite donner à From'Mir ? *« L'aventure doit se poursuivre*, affirme Philippe Grosperrin, directeur de Conseil élevage 25-90. *Nous allons proposer une suite au programme à nos financeurs actuels et aux différents partenaires. » « La Franche-Comté a eu une très grande chance d'être une région pilote pour ce type de recherche, il faut que nous poursuivions »*, insiste Eric Chevalier, directeur chez Monts & Terroirs (Sodiaal). Selon lui, *« il faudrait étendre la démarche dans les fruitières, chez les affineurs… »*.

UN OBSERVATOIRE RÉGIONAL DE LA FROMAGEABILITÉ EN FRANCHE-COMTÉ

L'objectif de la filière laitière francomtoise est de créer un observatoire régional de la fromageabilité en Franche-Comté qui exploiterait les spectres MIR ainsi que d'autres données sur les laits de troupeaux de cuves jusqu'aux fromages affinés (qualité des fromages, goût…). *« Le fromager devrait pouvoir aller consulter, quand il le souhaite, les spectres MIR disponibles dans des bases de données toujours accessibles pour pouvoir modifier ses pratiques en temps quasi réel*, remarque Philippe Grosperrin. *Et après avoir identifié les profils de fromageabilité du lait adaptés pour la qualité fromagère recherchée, le fromager n'attendra plus des mois d'affinage pour ajuster en continu ses paramètres technologiques. »*

Ces bases de données pourraient aussi servir de base à de futurs travaux de recherche. *« From'Mir concerne toutes les régions de France*, insiste en conclusion Claude Vermot Desroches, président du Comité interprofessionnel de gestion du comté (CIGC). *Le progrès est considérable car le fromager pourra réagir beaucoup plus vite, les référentiels des AOP pourront être enrichis… Mais le travail sur la richesse microbienne sera essentiel, surtout dans nos filières au lait cru, avec lequel l'équilibre entre population microbienne et composition chimique est fragile. »*

SABINE CARANTINO

C'EST QUOI LA FROMAGEABILITÉ ?

La fromageabilité est l'aptitude du lait à se transformer en fromage sous l'action d'un agent coagulant et des levains lactiques, en minimisant les pertes lors de l'égouttage et en maximisant la rétention dans le caillé des éléments de lait contribuant à l'élaboration de la qualité sensorielle finale des fromages. Cette aptitude s'évalue de manière différente selon la technologie fromagère concernée. La fromageabilité, notamment pour les fromages de terroir au lait cru, est une notion complexe. Elle induit, par exemple, qu'il y ait un bon équilibre entre la composition chimique et la composition microbiologique du lait.

CHIFFRES CLÉS

- 585 éleveurs et 52 fromageries ont participé au projet
- plus d'un millier d'échantillons prélevés
- 48 minifabrications réalisées
- plus de 70 techniciens et scientifiques mobilisés dont 1 doctorant et 5 stagiaires

ILS ONT DIT…



Claude Vermot Desroches, président du CIGC

« Attention à ce que tout ce travail sur la génétique ne fasse pas perdre de la diversité à nos AOP. Il nous faudra pour cela avoir une approche très fine de la sélection. Mais vive From'Mir ! Un tel travail d'approfondissement des connaissances physicochimiques et génétiques, et d'amélioration des pratiques a forcément un gros impact sur les filières traditionnelles. Il leur permet de mieux mettre en avant la valeur de leurs métiers et de leurs produits. »



Cécile Laithier, chef de projet From'Mir à l'Idele

« Les équations d'estimation établies par From'Mir sont de très bons outils de prédiction de la fromageabilité des laits individuels. Elles sont un peu moins efficaces sur les laits de troupeaux et restent à améliorer sur les laits de cuves de fromageries. Des travaux sont encore nécessaires. Pour ces laits, de nouvelles pistes seront explorées à partir de l'utilisation des laits individuels. »

À SAVOIR

Les partenaires financiers du projet From'Mir :
Cniel, ministère de l'Agriculture, région Bourgogne Franche-Comté, Union régionale des fromages d'appellation d'origine comtois (comté, morbier, mont d'or, bleu de Gex).

Les partenaires techniques :
Actalia, Allice, Conseil élevage 25-90, CIGC, CRA Bourgogne et Franche-Comté, CTFC, Enilbio, EVA Jura, Haute-Saône Conseil élevage, Inra de Poligny et de Jouy en Josas, Institut de l'élevage, Umotest.

ILS ONT DIT…



Philippe Grosperrin, directeur de Conseil élevage 25-90

« From'Mir s'est révélé être un dispositif puissant pour mener à bien un travail génétique. Il a révélé des héritabilités variables, de modérées à fortes, pour les critères de fromageabilité. »



Eric Chevalier, directeur chez Monts & Terroirs (Sodiaal)

« Nous sommes dans l'univers du lait cru. Glaner de précieuses informations sur la composition fine du lait et les vulgariser en langage fromager est un plus majeur dans le travail quotidien de nos fromagers. »

RLF782-ESSE-LEAD-FROMAGEABILITE.indd 4 14/05/2018 10:31
?-ESSE-LEAD-FROMAGEABILITE.indd 5 14/05/2018 10:31

Quels facteurs impactent La fromageabilité du lait ?

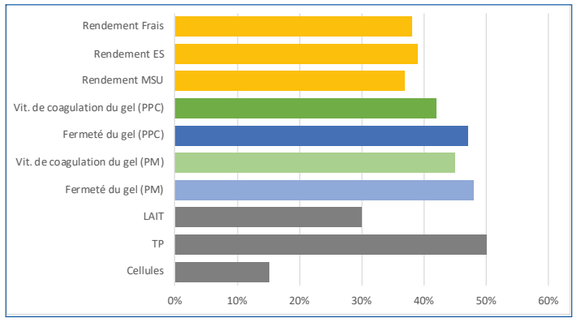
Le programme FROM'MIR permet, pour la première fois en France, d'élaborer des outils d'estimation de la fromageabilité du lait en routine à partir des analyses Moyen Infrarouge du lait de vache. Ce programme a été conduit en race Montbéliarde au sein des filières fromagères AOP et IGP de Franche-Comté.



Un facteur génétique fort
Le potentiel fromager de la Montbéliarde décrypté

Des caractères héritables

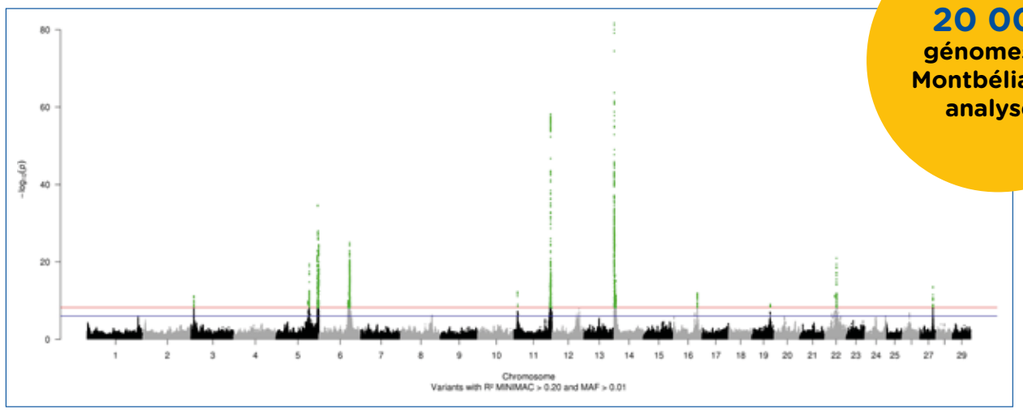
L'héritabilité représente la part des différences expliquée par la génétique et transmissible à la descendance.
Avec des valeurs comprises entre 37 et 48%, les critères fromagers (rendements et aptitude à la coagulation) sont sélectionnables.
Par ailleurs ils sont fortement et favorablement corrélés entre eux et avec les taux protéiques et butyreux notamment. .



De nombreuses régions du génome impliquées

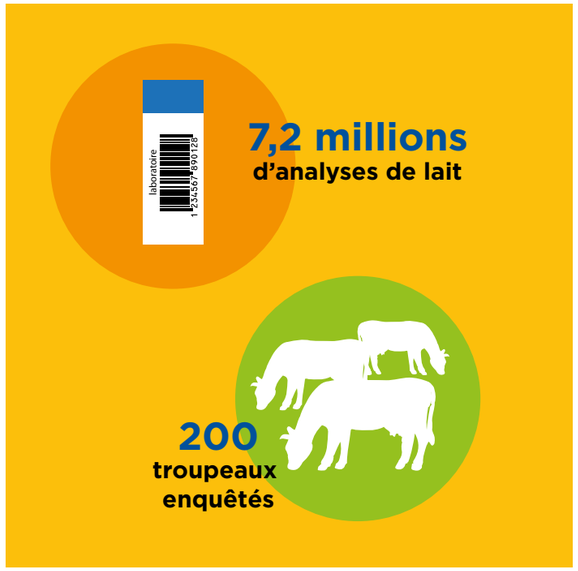
L'analyse de plus de 27M de « lettres » de la séquence ADN de 20 000 femelles a permis de confirmer (caséines…) et d'identifier de nouvelles zones du génome bovin impliquées dans la fromageabilité du lait.

GWAS on whole sequence : RDT_ES



Exemple : Détection (pics verts) sur les chromosomes 1 à 29, des zones du génome ayant un effet très significatif sur le rendement de laboratoire en extrait sec.

20 000
génomes de Montbéliardes analysés



Des 1ers index génomiques en Montbéliarde

Neuf caractères fromagers au total ont pu être indexés. Ils sont en cours d'étude afin d'analyser l'impact économique potentiel de ces nouveaux critères et les réponses à la sélection. Cela permettra d'éclairer sur l'usage possible de ces nouveaux index.
Les 1ers résultats révèlent que la sélection actuelle est plutôt favorable grâce à la place importante accordée au TP.

Des leviers en élevage
Maîtriser la fromageabilité du lait, c'est possible !



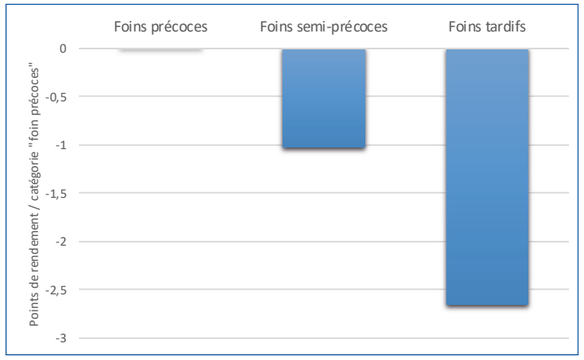
Du fourrage à volonté !

FROM'MIR a permis de démontrer qu'une restriction en fourrages en hiver avait un effet négatif sur l'aptitude à la coagulation du lait. Limiter l'ingestion chez la vache laitière a un impact négatif sur le taux protéique.

Du fourrage de qualité !

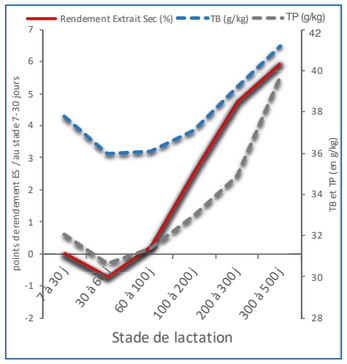
Dans les systèmes herbagers de Franche-Comté où le foin constitue le principal fourrage en hiver, il existe un lien fort entre le stade de récolte et le rendement fromager.
Les foins récoltés précocement ont de meilleures valeurs alimentaires.

Rendements ES (%) en fonction de la qualité du foin



La répartition des vêlages : un facteur à prendre en compte

Evolution du rendement en Extrait Sec (%) au cours de la laction



Les rendements et l'aptitude à la coagulation atteignent leurs valeurs minimales aux alentours du pic de lactation puis augmentent jusqu'à la fin de la lactation. Cet effet sera notable dans les élevages où les vêlages ne sont pas étalés tout au long de l'année.

La fromageabilité mesurée par FROM'MIR :

Aptitude à la coagulation enzymatique : capacité d'un lait à se transformer en caillé après addition de présure.
Deux paramètres mesurés : la vitesse d'organisation et la fermeté du gel.
Ces paramètres sont mesurés selon deux modalités d'emprésurage, l'une de type pâte molle (PM) et l'autre de type pâte pressée cuite (PPC).

Rendement fromager de laboratoire frais : Rapport (en %) de la masse de caillé -obtenu après emprésurage, découpage du caillé et centrifugation- à la masse de lait mis en œuvre. Le rendement de laboratoire extrait sec correspond au rapport (en %) de la masse d'extrait sec du caillé à la masse d'extrait sec du lait mise en œuvre. Le rendement dit « MSU » (comme matière sèche utile) correspond à la quantité de matière sèche utile (matière grasse et protéique) nécessaire pour obtenir 1 kg de caillé.









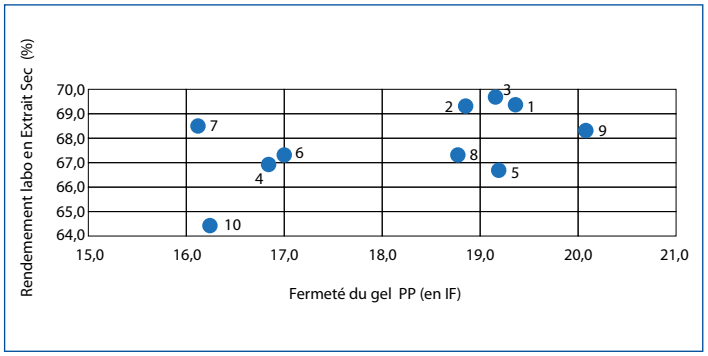
La fromageabilité est l'affaire de tous

Cas de la coopérative de Loray

Les outils mis au point dans FROM'MIR ont été testés pour la première fois en conditions réelles dans une fruitière du Doubs qui collecte et transforme en Comté le lait de 11 producteurs.

La diversité des résultats s'explique par la diversité des pratiques et conditions d'élevage

Fermeté du gel et rendement laboratoire des laits de tank des producteurs

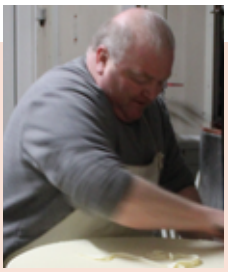


Chaque élevage a fait l'objet d'une visite pour enregistrer les pratiques des éleveurs, la qualité des fourrages, la composition des aliments et les conditions de logement. Les situations sont très variées dans les 11 élevages de la coopérative et expliquent en grande partie la diversité des résultats obtenus quant à la fromageabilité des laits



Bertrand Richard,
Président de la coopérative de Loray

« On n'aurait jamais imaginé qu'il y avait autant de diversité au sein des élevages de la coopérative »



Cyriaque Abram,
fromager à la coopérative de Loray

« Jusqu'à maintenant je n'avais pas vraiment conscience de tout ce qui pouvait se passer en amont du tank et qui pouvait influencer mon travail à la fromagerie »



FROM'MIR, la suite...

L'appropriation complète des travaux de FROM'MIR nécessite de mettre en place des outils pour suivre et analyser en continu les évolutions de la fromageabilité et les relier aux observations de terrain et à toutes les données d'ores et déjà existantes au sein de l'ensemble des filières. L'objectif des membres du consortium FROM'MIR est donc de mettre en place un observatoire qui permettra de faire communiquer les bases de données et les valoriser. Cet observatoire consolidera également les résultats acquis dans le programme FROM'MIR.

Les méthodologies utilisées ainsi que l'expérience acquise au cours du programme FROM'MIR sont une base de réflexion pour la construction et la conduite d'autres études pour d'autres races, d'autres types de transformation fromagère ou différents systèmes d'élevages.

Les résultats ont été obtenus dans le cadre du programme FROM'MIR avec le soutien financier du ministère de l'Agriculture, de l'Agro-alimentaire et de la Forêt, du Centre National Interprofessionnel de l'Economie Laitière (CNIEL), de l'Union Régionale des Fromages d'Appellation d'origine Comtois (URFAC) et de la région Bourgogne Franche-Comté.

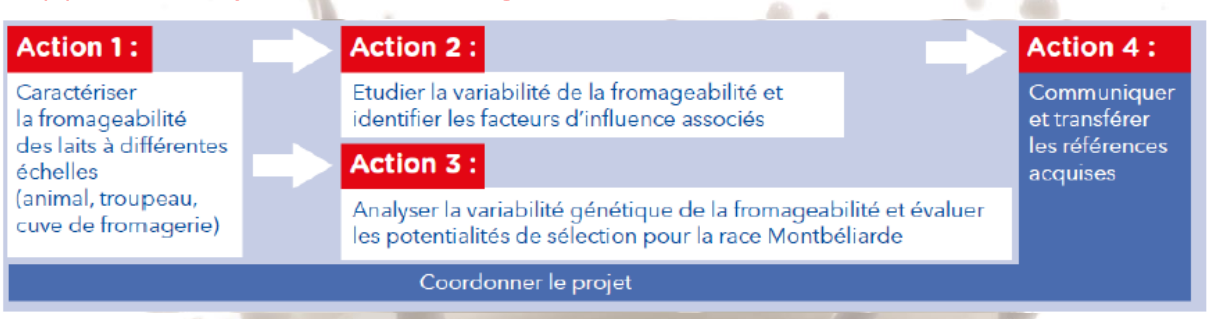
Rédacteurs : BROCHARD Mickaël (Umotest), SANCHEZ Marie-Pierre (INRA), MINERY Stéphanie (Idele), GELE Marine (Idele), WOLF Valérie (Conseil Elevage 25-90), LAITHIER Cécile (Idele), BOICHARD Didier (INRA), GROSPERRIN Philippe. (Conseil Elevage 25-90), DELACROIX-BUCHET Agnès (INRA) et GAUDILLIERE Nicolas (Conseil Elevage 25-90)

Crédits photo : Conseil Élevage 25-90, Thierry Petit

Newsletter FROM'MIR n° 5

La newsletter n° 5 a pour objectif de présenter les facteurs de variation de la fromageabilité à l'échelle individuelle, du troupeau et de la cuve de fromagerie. Un zoom sur l'UMT eBIS, UMOTEST et les Entreprises de Conseil en élevage sera réalisé.

Rappel des objectifs et de l'organisation de FROM'MIR

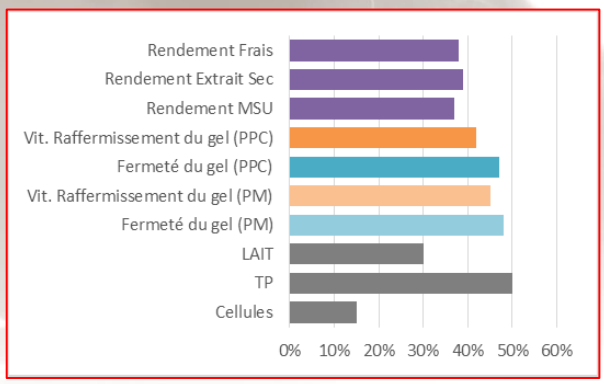


Quels facteurs impactent la fromageabilité du lait ?

La génétique : un levier d'action important à l'échelle individuelle

Pour un caractère donné, l'héritabilité, qui peut varier entre 0 et 100%, mesure la part de la variabilité de ce caractère qui est génétique et donc transmissible à la descendance. Plus cette valeur sera forte, plus il sera possible d'améliorer génétiquement ce caractère par sélection.

Les valeurs d'héritabilité varient de 37 à 48% pour les paramètres fromagers. Elles sont du même ordre de grandeur que celles des taux protéique et butyreux, ce qui montre qu'il est possible d'améliorer la fromageabilité des laits par la voie génétique.



Les corrélations génétiques sont fortes et toujours favorables entre les différents critères fromagers. La sélection sur l'un des critères fromagers aura donc un effet améliorateur sur tous les autres critères fromagers et la composition du lait.

Près de 20 000 vaches disposant de résultats de contrôles de performances avec spectres MIR enregistrés avaient été génotypées (puce EuroG10K ou puce 50k) via Umotest pour le calcul des index génomiques des caractères classiques. En combinant ces données à celles du projet « 1 000 génomes bovins », nous avons pu reconstituer la séquence du génome complet des 20 000 vaches, soit plus de 27 millions de variations du génome. L'analyse conjointe de ces données du génome et des phénotypes a permis d'identifier les régions du génome et les voies métaboliques impliquées dans le déterminisme des critères fromagers.

En plus des régions déjà connues pour leurs effets sur la composition du lait (gènes des caséines, gène PAEP qui code la β-lactoglobuline et gène DGAT1 impliqué dans la synthèse des acides gras), nous avons mis en évidence un grand nombre de régions du génome avec des effets forts sur les critères fromagers (Figure 1). Une partie de ces régions correspond à des gènes impliqués dans le métabolisme des acides gras, des protéines et des minéraux, et d'autres sont encore à élucider.

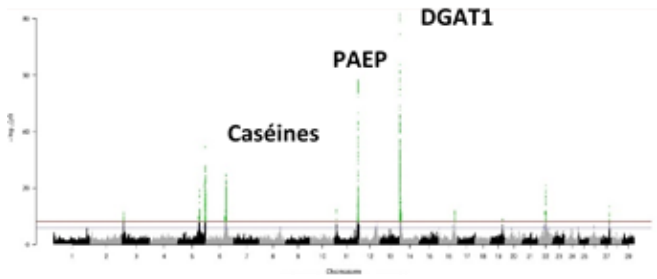


Figure 1. Exemple de résultat obtenu par analyse de la séquence complète des vaches. Chaque pic vert représente une région du génome avec des effets très significatifs sur le rendement en extrait sec.

Une évaluation génomique des critères fromagers est possible grâce à la quantité de données disponibles et à l'effectif important de vaches typées. Un pilote a été développé. Des études complémentaires sont en cours pour estimer la précision de ces évaluations, l'importance économique de ces critères fromagers par rapport aux autres critères en sélection et les réponses à la sélection pour l'ensemble des caractères évalués.

Le stade de lactation influence fortement la fromageabilité du lait

A l'échelle individuelle, le rendement et l'aptitude à la coagulation suivent des évolutions cohérentes avec celles des taux au cours de la lactation : les rendements, la vitesse d'organisation et la fermeté du gel atteignent leurs valeurs minimales aux alentours du pic de lactation puis augmentent jusqu'à la fin de la lactation (Figure 2). Ces effets deviennent forts à partir de 200 jours de lactation. A l'échelle des troupeaux, le stade de lactation est un facteur de variation de la fromageabilité qui aura un impact notable seulement dans les élevages où les vêlages ne sont pas étalés tout au long de l'année.

Comparées aux multipares, les primipares présentent des laits de plus faible fromageabilité, en relation sans doute avec un taux plus élevé de protéines solubles.

Compte tenu du lien fort entre les paramètres de fromageabilité et les taux, le rapport TB/TP est un facteur qui influence fortement la fromageabilité. A l'échelle des vaches, un rapport TB/TP supérieur à 1,5 révèle fréquemment un fort déficit énergétique en début de lactation. Il est également associé à un rendement, une vitesse d'organisation et une fermeté du gel moins élevés.

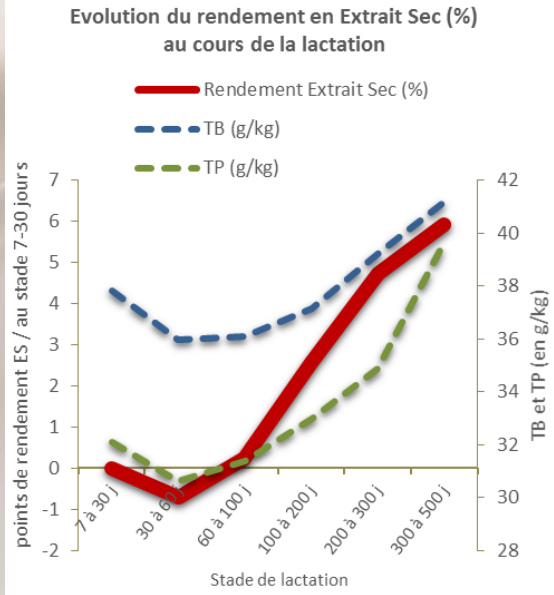
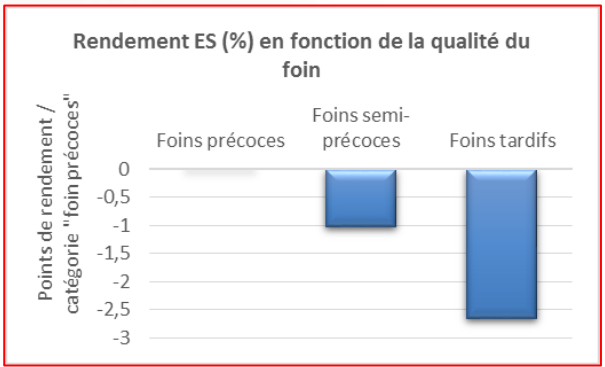


Figure 2. Evolution du rendement en Extrait Sec (%) au cours de la lactation

La qualité de la ration de base et son niveau d'ingestion

L'alimentation a un effet sur la fromageabilité qui s'explique principalement par son influence sur les taux.

La qualité de la ration de base a un impact majeur sur les paramètres de fromageabilité. Les rendements, la vitesse d'organisation et la fermeté du gel sont plus élevés lorsque le foin qui compose la ration du troupeau a été récolté précocement. Ces foins précoces ont de meilleures valeurs alimentaires et sont plus ingestibles que les foins tardifs.

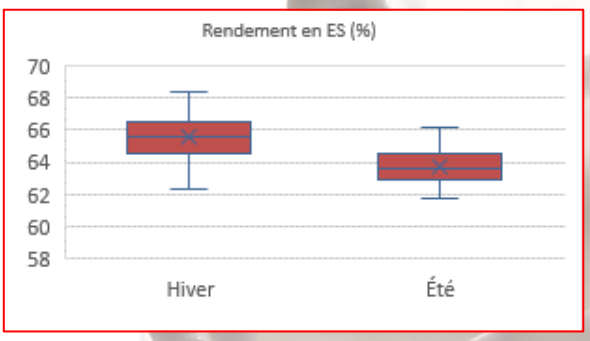


Les facteurs qui influencent le niveau d'ingestion ont un impact sur la fromageabilité du lait. Une restriction en fourrages en hiver est un facteur qui pénalise fortement le rendement. Au pâturage, un accès à l'eau limitant a un impact négatif sur la vitesse d'organisation et la fermeté du gel. Dans une moindre mesure, l'augmentation de la quantité de concentrés distribuée a un effet positif mais modeste sur la vitesse d'organisation et la fermeté du gel qui s'explique par son effet favorable sur le TP.



Le confort du bâtiment, la place disponible à l'auge et pour le couchage, la gestion des périodes de transition et les pratiques de distribution de la ration sont d'autres facteurs qui peuvent influencer le niveau d'ingestion. Ils peuvent contribuer à la maîtrise de la fromageabilité du lait.

La fromageabilité à l'échelle de la cuve de fromagerie



La saison a un impact important sur la fromageabilité des laits de cuve, les rendements sont plus élevés en hiver en lien avec des taux (TB et TP) plus élevés également. La vitesse d'organisation du gel (PM) est plus élevée en hiver. D'autres facteurs ont été testés tels que la zone géographique, l'historique thermique du lait, la taille de la fromagerie et l'homogénéité des volumes de lait livré de chaque élevage. Ils ne sont pas ressortis comme significatifs.

S'intéresser aux échelles individuelle et troupeau pour mieux caractériser la fromageabilité de la cuve ?

FROM'MIR démontre l'importance de considérer la fromageabilité à toutes les échelles. A l'échelle de la cuve, le fromager pourra s'appuyer sur les résultats de fromageabilité des laits individuels et troupeaux des éleveurs livrant à la coopérative pour mieux comprendre la variabilité de la fromageabilité dans sa cuve, en lien avec la qualité finale des fromages fabriqués.

FROM'MIR sur le terrain : cas concret à la coopérative de Loray

Les équations établies dans le cadre du programme FROM'MIR ont été utilisées pour la première fois en conditions réelles pour évaluer la fromageabilité des laits utilisés lors d'une journée de fabrication en janvier 2018 à la coopérative de Loray dans le Doubs.

La coopérative de Loray (25) en chiffres :

- 11 producteurs
- 2,8 millions de litres de lait collectés
- Transformation en AOP Comté
- Affinage : ETS Rivoire-Jacquemin

Cet essai a permis de démontrer la plus-value de ces nouveaux outils pour travailler collectivement sur la maîtrise de la fromageabilité au sein d'un atelier de fabrication. L'estimation directe de la fromageabilité des laits permet d'engager un travail technique plus efficace que l'analyse des taux (TB et TP) qui ne sont que des prédicteurs indirects de la fromageabilité. Fromagers et producteurs ont engagé un dialogue autour d'un langage commun matérialisé par les paramètres estimés grâce à FROM'MIR.

Bertrand RICHARD
Président de la coopérative de Loray :
« On ne se serait jamais imaginé qu'il y avait autant de diversité au sein des élevages de la coopérative »



Cyriaque ABRAM
Fromager à la coopérative de Loray :
« Jusqu'à maintenant je n'avais pas vraiment conscience de tout ce qui pouvait se passer en amont du tank et qui pouvait influencer mon travail à la fromagerie »



Par ailleurs, cette expérimentation a permis d'illustrer la forte variabilité des pratiques et des résultats des producteurs au sein d'une coopérative composée de seulement 11 élevages. Cette diversité fait la force et la typicité des produits AOP. Les fromages produits au cours de cette journée de fabrication seront suivis au cours de leur affinage et leurs caractéristiques sensorielles seront analysées avant leur commercialisation.

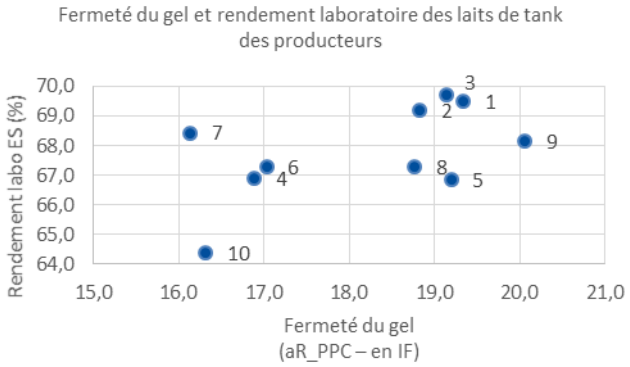
Fromageabilité des laits des vaches qui composent la cuve : **une variabilité importante**

	Moyenne* (N=348 VL)	Ecart-type (CV)
Lait (kg/VL)	23,9	6,3 (26%)
TB (g/kg de lait)	38,6	4,8 (12%)
TP (g/kg de lait)	33,1	3,4 (10%)
Type de données	***Estimations MIR***	
Rendement Extrait Sec (%)	65,9	4,8 (7%)
Fermeté du gel (aR_PPC en IF)	18,4	2,4 (13%)

* Moyennes pondérées de productions laitières individuelles

Chaque élevage a fait l'objet d'une visite qui a permis d'apprécier les pratiques des éleveurs, la qualité des fourrages, la composition des aliments et les conditions de logement. Bien que la coopérative soit de taille modeste, les situations dans les élevages sont très variées et expliquent en grande partie la diversité des résultats obtenus quant à la fromageabilité des laits à l'échelle des élevages.

A l'échelle des troupeaux : la diversité des résultats s'explique par la diversité des pratiques des éleveurs et des conditions d'élevage

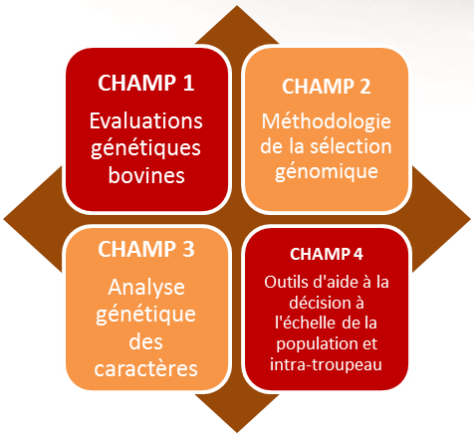


Zoom sur les activités des partenaires

Zoom sur l'UMT eBIS

L'UMT eBIS, labellisée en 2017 pour 5 ans, associe 3 partenaires : l'INRA-GABI, l'institut de l'élevage et Allice au sein de l'équipe Génétique et Génomique Bovine (G2B) à Jouy en Josas (78). Elle fait suite à l'UMT 3G et rassemble les mêmes partenaires pour mettre en œuvre la sélection génomique dans toutes les populations, une technologie qui a complètement renouvelé la sélection bovine. Le nouveau projet prolonge le précédent mais intègre les nouveaux enjeux de la sélection associés aux changements globaux, aux évolutions technologiques, mais aussi aux évolutions réglementaires.

4 champs dans le projet de l'UMT Ebis



33 agents, 23 ETP

Direction : Sophie MATTALIA, Didier BOICHARD et Sébastien FRITZ

Les travaux de l'UMT eBIS dans FROM'MIR



L'UMT réalise les analyses de l'action 3 du projet FROM'MIR « Analyse de la variabilité génétique de la fromageabilité et évaluation des potentialités de sélection pour la race Montbéliarde » dont les résultats sont détaillés dans cette Newsletter.

Zoom sur Umotest

Umotest, est une union de coopératives agricoles (insémination) qui diffuse ses taureaux Montbéliards auprès de 15 000 éleveurs français et dont l'activité de ses coopératives adhérentes représente 80% de l'insémination française dans cette race, des zones AOP de l'Est, aux plaines océaniques bretonnes en passant par les zones herbagères auvergnates. Sa mission première est de créer des solutions adaptées aux éleveurs de Montbéliarde. Elle s'appuie pour cela sur un programme de sélection partagé entre toutes les coopératives. Umotest est un lieu de concertation et de décisions pour conduire le programme de sélection au quotidien mais également pour anticiper les besoins des éleveurs et imaginer les services, outils, indicateurs qui leur seront utiles. Dans ce cadre-là, Umotest est très actif en matière de R&D partenariale. Elle a été à l'initiative ou motrice de plusieurs innovations récentes (génomique, sexage de la semence…). L'innovation est dans ses gènes !

> http://www.umotest.com/presentation.html

Rôle d'Umotest dans FROM'MIR

Umotest avec les Conseils en Elevage du Doubs-Territoire de Belfort, de Haute-Saône et l'Institut de l'Elevage a été à l'initiative d'un projet sur la fromageabilité. Umotest a accompagné et soutenu le Conseil Elevage 25-90 pour monter et conduire le projet FROM'MIR.
Umotest s'est fortement impliquée dans FROM'MIR, par la participation de MM. Hervé Bole (éleveur administrateur) et Tristan Gaiffe (directeur général) au pilotage et de Mickaël Brochard (responsable R&D) dans le suivi et la réalisation du programme. Umotest a également mis à disposition du projet l'intégralité des génotypes de femelles (et mâles) dont elle disposait.



Le rôle de Mickaël Brochard a consisté à animer et participer, avec l'Institut de l'Elevage, l'INRA et Allice (UMTeBIS), aux travaux du 3ème volet de FROM'MIR dédié à l'analyse des facteurs génétiques, dont les principaux résultats sont présentés dans cette newsletter. Il a également contribué au comité de suivi global du projet, au groupe communication ainsi qu'au volet 1 du projet concernant le développement des équations de prédictions de la fromageabilité par le MIR pour transférer autant que possible l'expérience de projets de recherche antérieurs.

Zoom sur les Entreprises de Conseil en élevage de Franche-Comté

Conseil Elevage 25-90, Haute-Saône Conseil Elevage et Eva Jura (anciennement Jura Conseil Elevage) réalisent le **contrôle de performances** en élevage. Cette mission consiste à prélever un échantillon de lait par vache traite à fréquence régulière dans les élevages. Cela répond à deux objectifs :

- Un objectif génétique et collectif : collecter des phénotypes nécessaires pour conduire les programmes génétiques et établir les index.
- Un objectif économique et individuel : élaborer un suivi et le pilotage de l'atelier lait.

Le cœur de métier de ces trois entreprises, c'est aussi le c**onseil en élevage. Les conseillers techniques** valorisent les données issues du contrôle de performance et apportent une expertise dans les élevages laitiers de la région pour améliorer la rentabilité des exploitations laitières. Le conseil technique porte notamment sur l'alimentation, les fourrages, la qualité du lait, la reproduction, les prévisions laitières, l'élevage des génisses.







Résolument tournées vers l'avenir, les entreprises de conseil en élevage apportent constamment de nouveaux services aux éleveurs et investissent de façon importante dans la R&D. Les données que ces entreprises collectent sur le terrain sont valorisées pour développer des outils de conseils toujours plus pointus pour apporter une vraie plus-value à leurs éleveurs adhérents, tels que le constat de gestation, CetoMIR/CETODETECT (permet d'identifier le déficit énergétique en début de lactation) et PROFIL'AGE, un outil de pilotage des troupeaux basés sur l'interprétation du profil en acides gras du lait.

	CONSEIL ÉLEVAGE 25-90	HAUTE-SAÔNE CONSEIL ÉLEVAGE	eva JURA au cœur de l'élevage
Agents de traites	88	28	35
Conseillers techniques & cadres	46	15	19
Nombre d'élevages adhérents	1592	500	770
Nombre d'analyse de laits / an	800 000	300 000	415 000

Le rôle des Entreprises de Conseil en Elevage dans FROM'MIR

Conseil Elevage 25-90 est l'organisme chef de file du programme et a assuré la co-animation avec l'institut de l'élevage.

Les entreprises de conseil en élevage ont collecté l'ensemble des échantillons de laits individuels et de troupeaux qui ont servi à la bonne réalisation du programme. Soucieux de garantir la qualité des échantillons prélevés, des efforts importants ont été mis en place par les trois ECEL sur la logistique de prélèvement pour garantir la conformité des échantillons. L'ensemble des échantillons ont été prélevés dans plus de 650 élevages sur la région, et une cinquantaine de conseillers ont contribué à ces prélèvements mais aussi aux enquêtes réalisées en élevages. L'action visant à mettre en évidence les facteurs de variation de la fromageabilité a été encadrée par Conseil Elevage 25-90.

Zoom sur le laboratoire d'analyse de Conseil Elevage 25-90

Conseil Elevage 25-90 possède son laboratoire pour l'analyse par spectrométrie Moyen InfraRouge des échantillons de lait collectés dans la cadre du contrôle de performance. Cette mission est assurée par le LDA 39 dans le Jura et le LIAL de Rioz en Haute-Saône.

Le laboratoire est un outil déterminant dans les projets de R&D à Conseil Elevage 25-90. L'innovation se conduisant rarement seule, les partenariats sont nombreux dans le cadre de la fédération nationale France Conseil Elevage ou encore du GIEE européen EMR (www.milkrecording.eu). Conseil Elevage 25-90 a analysé l'ensemble des échantillons prélevés dans le cadre de FROM'MIR par spectrométrie Moyen InfraRouge dans son laboratoire d'analyse. Les spectres MIR du lait, obtenus par cette méthode, ont été valorisés pour la mise au point des équations.

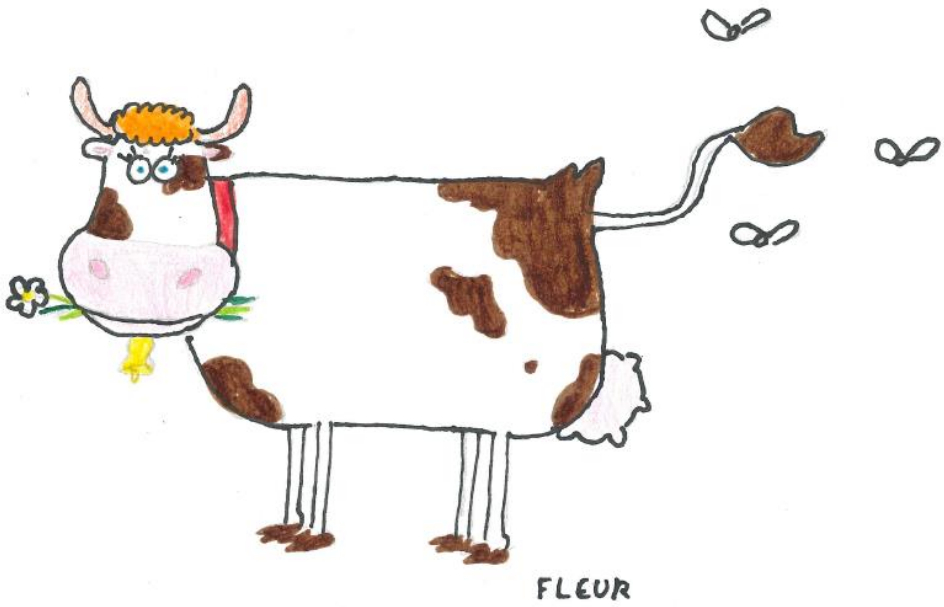


Contacts : Cécile LAITHIER — Institut de l'Elevage — cecile.laithier@idele.fr
Valérie WOLF — CEL 25-90 — valerie.wolf@cel2590.fr

Crédit photos : CEL 25-90, INRA, © Emilie AUJÉ
Mise en page : Isabelle GUIGUE, Institut de l'Elevage

PARTENAIRES TECHNIQUES ET FINANCIERS :




FLEUR

université PARIS-SACLAY

ÉCOLE DOCTORALE
Agriculture, alimentation, biologie, environnement, santé (ABIES)

Titre : Analyse génétique de la composition protéique & des aptitudes fromagères du lait de vache prédites à partir des spectres moyen infrarouge

Mots clés : *Bovins laitiers - Composition du lait – Aptitudes fromagères – Variants causaux – Sélection génomique*

Résumé : Les aptitudes du lait à la transformation en fromage sont étroitement liées à sa composition, notamment en protéines. Ces caractères, difficiles à mesurer directement, ont été prédits à partir des spectres moyen infrarouge (MIR) du lait pour la composition en protéines dans les 3 races bovines Montbéliarde, Normande et Holstein (projet *PhénoFinlait*) et pour 9 aptitudes fromagères et la composition fine du lait en race Montbéliarde (projet *From'MIR*). La méthode *Partial Least Squares* (PLS) a fourni des prédictions MIR plus précises que les méthodes bayésiennes testées.

Une analyse génétique a été réalisée pour ces caractères prédits à partir de plus de six millions de spectres MIR de plus de 400 000 vaches.

Les caractères fromagers et de composition du lait sont modérément à fortement héritables. Les corrélations génétiques entre caractères fromagers (rendements et coagulation) et avec la composition du lait (protéines, acides gras et minéraux) sont élevées et favorables.

Les génotypes de 28 000 vaches ont été imputés jusqu'à la séquence complète grâce aux données du projet 1000 génomes bovins.

Des analyses d'association (GWAS) révèlent de nombreux gènes et variants avec des effets forts sur la fromageabilité et la composition du lait. Un réseau de 736 gènes, par ailleurs associé à ces caractères, permet d'identifier des voies métaboliques et des gènes régulateurs fonctionnellement liés à ces caractères.

Un prototype d'évaluation génomique a été mis en place en race Montbéliarde. Un modèle de type contrôles élémentaires, incluant les variants détectés par les GWAS et présumés causaux, donne les estimations des valeurs génomiques les plus précises. La simulation d'une sélection incluant les caractères fromagers montre qu'il est possible d'améliorer la fromageabilité du lait avec un impact limité sur le gain génétique des autres caractères sélectionnés.

Les travaux présentés dans cette thèse ont abouti 1) à la détection de gènes (dont certains jamais décrits auparavant) et de variants candidats pour la composition et la fromageabilité du lait et 2) à la mise en place d'une évaluation génomique de la fromageabilité du lait en race Montbéliarde dans la zone AOP Comté.

Title : Genetic analysis of bovine milk protein composition and cheese-making traits predicted from mid-infrared spectra

Keywords : *Dairy cattle – Milk composition – Cheese-making properties – Causal variants – Genomic selection*

Abstract: The ability of milk to be processed into cheese is closely linked to its composition, in particular in proteins. These traits, which are difficult to measure directly, were predicted from milk mid-infrared (MIR) spectra for protein composition in 3 cattle breeds Montbéliarde, Normande and Holstein (*PhénoFinlait* project) and for 9 milk cheese-making properties (CMP) and composition traits in Montbéliarde cows (*From'MIR* project). The Partial Least Squares method provided more accurate predictions than the Bayesian methods tested.

A genetic analysis was performed on these traits, predicted from more than six million MIR spectra of more than 400,000 cows.

Milk CMP and composition traits are moderately to highly heritable. Genetic correlations between CMP (cheese yields and coagulation) and with milk composition (proteins, fatty acids and minerals) are high and favorable.

The genotypes of 28,000 cows were imputed to whole genome sequences using the 1000 bovine genome reference population.

Genome wide association studies (GWAS) reveal many genes and variants in these genes with strong effects on CMP and milk composition. A network of 736 genes, associated with these traits, enable the identification of metabolic pathways and regulatory genes functionally linked to these traits.

A pilot genomic evaluation was set up in Montbéliarde cows. A test-day model, including variants detected by GWAS, provides the most accurate genomic value estimates. Simulation of a selection shows that it is possible to improve the cheesability of milk with a limited impact on the genetic gain of the traits that currently make up the breeding objective.

The work presented in this thesis led to 1) the detection of genes (some of which have never been described before) and candidate variants for milk CMP and composition traits and 2) the implementation of a genomic evaluation of CMP predicted from MIR spectra in Montbéliarde cows of the Comté PDO area.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France