Optimization of association genetics and genomic selection strategies for populations of different diversity levels : Application in maize (Zea mays L.)

Renaud Rincent

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

AgroParisTech
INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

Ecole Doctorale ABIES
Agriculture Agriculture Alimentation Food Biologie Biology Environnement Environment Santé Health

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

L'Institut des Sciences et Industries du Vivant et de l'Environnement

(AgroParisTech)

Spécialité : Sciences de la vie et santé

présentée et soutenue publiquement par

Renaud RINCENT

le 11 Avril 2014

Optimization of association genetics and genomic selection strategies for populations of different diversity levels. Application in maize (Zea mays L.).

Directeur de thèse : **Alain Charcosset**
Co-encadrement de la thèse : **Laurence Moreau, Milena Ouzunova, Pascal Flament, Pierre Dubreuil**

Travaux réalisés dans le cadre d'une convention CIFRE, avec l'aide de l'ANRT

Jury

M. John HICKEY, Docteur, Roslin Institute, Edinburgh	Rapporteur
M. Gilles CHARMET, Directeur de Recherches, INRA de Clermont-Ferrand	Rapporteur
M. David CAUSEUR, Professeur, Mathématiques appliquées, Agrocampus Ouest	Examinateur
M. Etienne VERRIER, Professeur, AgroParisTech	Examinateur
M. Pascal FLAMENT, Head of research support functions, Limagrain	Co-encadrant
Mme Laurence MOREAU, Chargée de Recherches, INRA du Moulon	Co-encadrante
M. Alain CHARCOSSET, Directeur de Recherches, INRA du Moulon	Directeur de Thèse

AgroParisTech
UMR de Génétique Végétale
INRA/Univ. Paris Sud/CNRS
Ferme du Moulon, 91190 Gif-sur-Yvette, France

Remerciements

- Pour ces 3 ans fermes du Moulon -

Mes remerciements vont d'abord à Alain et Laurence qui m'ont fait confiance et soutenu au cours de ces trois années. La richesse de nos échanges et de vos engagements m'ont été très précieux. C'est en bonne partie grâce à vous que j'ai pu découvrir, profiter et m'intégrer peu à peu au monde passionnant de la recherche : j'ai maintenant le doigt pris dans l'engrenage de la machine infernale, bravo vous pouvez être fier de vous ! J'espère pouvoir continuer à travailler avec vous sur de nouvelles idées... Merci également aux autres membres de GQMS de m'avoir fait confiance pour donner quelques cours aux étudiants (Julie), pour les échanges qu'on a pu avoir (Tristan, Cyril, Stéphane, Delphine). Merci Tristan (mon prof de TD de stat à Grignon !) pour tes supers cours et démos en direct, ta pédagogie, ton grand tableau et nos échanges. Merci Cyril d'avoir toujours pris le temps de m'expliquer des choses au bureau ou au champs, et pour ton initiation aux dégustations œnologiques, et merci de m'avoir fait découvrir que j'étais marié à un Nez... Merci André de m'avoir recommandé le stage chez KWS avec Dietrich, qui m'a lancé sur la voie de la génétique. Merci pour votre passion communicative, et de prendre à chaque fois plaisir à discuter avec nous, les étudiants.

Pour leur accompagnement, le temps qu'ils ont pris à m'accueillir et à m'exposer leurs problématiques, je remercie mes collègues de KWS, en particulier Milena et Thomas (vielen Dank), mes collègues de Limagrain (Pascal, Zivan, Simon, Sébastien...) et de Biogemma (Pierre et Sébastien), ainsi que Sylvie Guillaume pour son efficacité et sa gentillesse. Merci à Pascal, Milena et Sébastien d'avoir accepté de financer ma thèse. Un grand merci à tous ceux qui ont travaillé sur les expérimentations du réseau INRA (Cyril et Jacques en particulier) et de nos partenaires Allemands et Espagnols pour le très gros travail de terrain sans lequel ma thèse n'aurait pas été possible. Egalement merci à tous ceux qui ont réalisé les analyses de laboratoire (Delphine, Valérie, Fabrice...).

Je remercie John Hickey, Gilles Charmet, David Causeur, Etienne Verrier et Pascal Flament d'avoir accepté de faire partie de mon jury de thèse.

Je remercie également toute la République démocratique du Moulon pour sa très bonne ambiance collective, son ouverture d'esprit et sa richesse humaine et scientifique. L'ambiance change en permanence avec les arrivées et départs mensuels, l'évolution de la faune endémique, mais elle reste toujours très bonne. Je remercie d'abord l'équipe de choc de jeunes présents à mon arrivée: Nico, Marion, Amandine, Fabio, Pierre M, Sophie, Mathieu, Bub,

Ashwin, Yannick, Charlotte, Paulina, Stéphanie, Pierre R, Yves, Beatriz, Abdul, ceux arrivés plus tard Mariangela, Héloïse, Margot, Telma, Aude, Sandra, les Sara, Margaux-Alison, Christophe, Adrien, Jean-Tristan, Cyril. Merci à mes collègues de bureau Mariangela (bout de choco ?) et Héloïse (bout de choco ? Ah non c'est vrai...), pour l'architecture d'intérieur, la déco et la bonne ambiance. Merci Sophie pour nos discussions et ta précieuse aide, Nicolas et Amandine pour vos conseils. Merci à Christine, Jérôme, Maud, Karin, Isabelle, Dominique, Domé, Matthieu, Philippe pour nos discussions. Merci Valérie pour ta grande cuisine et ta gentillesse. Philippe, Marlène, Xavier, Fabrice, pour les rigolades.

Je remercie les collègues de Jouy (de m'avoir laisser harceler Bananier), et notamment Hervé et Estelle pour nos échanges. Merci Estelle pour tes conseils, ton soutien et ta bonne humeur.

Je remercie également toutes les personnes que j'ai croisées au cours de ma thèse et avec qui j'ai pu avoir de riches échanges : l'équipe d'Orléans (Vincent et Leopoldo en particulier), de Toulouse (Andres et Bertrand), Jose Crossa, Catherine Giauffret, Christopher Sauvage, Denis Laloë et tous les autres.

Merci aux copains pour leur amitié forte et riche, et les grands moments passés ensemble, merci Laurent, Baptiste, Anne, Anaïs, Alexis, Gildas et Schérou, Armand, David et Chloé.

Je remercie tout particulièrement ma famille, sans qui tout cela n'aurait pas été possible. Vous avez toujours été présents et m'avez donné de la force dans les moments difficiles.

Je remercie Sarah du fond du cœur pour son soutien, sa gentillesse et son courage.

"Nous n'avons que faire d'aller trier des miracles et des difficultés étrangères : il me semble que parmi les choses que nous voyons ordinairement, il y a des étrangetés si incompréhensibles, qu'elles surpassent toute la difficulté des miracles. Quel monstre est-ce, que cette goutte de semence, de quoi nous sommes produits, porte en soi les impressions, non de la forme corporelle seulement, mais des pensements et des inclinations de nos pères. Cette goutte d'eau, où loge ce nombre infini de formes : et comme portent-elles ces ressemblances, d'un progrès si téméraire et si déréglé, que l'arrière-fils répondra à son bisaïeul, le neveu à l'oncle."

Essais de **M. de Montaigne** (1580). Livre II, chap. 37.

A mes parents

ABSTRACT

Major progresses have been achieved in genotyping technologies, which makes it easier to decipher the relationship between genotype and phenotype. This contributed to the understanding of the genetic architecture of traits (Genome Wide Association Studies, GWAS), and to better predictions of genetic value to improve breeding efficiency (Genomic Selection, GS). The objective of this thesis was to define efficient ways of leading these approaches. We first derived analytically the power from classical GWAS mixed model and showed that it was lower for markers with a small minimum allele frequency, a strong differentiation among population subgroups and that are strongly correlated with markers used for estimating the kinship matrix K. We considered therefore two alternative estimators of K. Simulations showed that these were as efficient as classical estimators to control false positive and provided more power. We confirmed these results on true datasets collected on two maize panels, and could increase by up to 40% the number of detected associations. These panels, genotyped with a 50k SNP-array and phenotyped for flowering and biomass traits, were used to characterize the diversity of Dent and Flint groups and detect QTLs. In GS, studies highlighted the importance of relationship between the calibration set (CS) and the predicted set on the accuracy of predictions. Considering low present genotyping cost, we proposed a sampling algorithm of the CS based on the G-BLUP model, which resulted in higher accuracies than other sampling strategies for all the traits considered. It could reach the same accuracy than a randomly sampled CS with half of the phenotyping effort.

Key words: maize, genomic selection, association mapping, power, accuracy, biomass.

RESUME

D’importants progrès ont été réalisés dans les domaines du génotypage et du séquençage, ce qui permet de mieux comprendre la relation génotype/phénotype. Il est possible d'analyser l’architecture génétique des caractères (génétique d’association, GA), ou de prédire la valeur génétique des candidats à la sélection (sélection génomique, SG). L’objectif de cette thèse était de développer des outils pour mener ces stratégies de manière optimale. Nous avons d’abord dérivé analytiquement la puissance du modèle mixte de GA, et montré que la puissance était plus faible pour les marqueurs présentant une faible diversité, une forte différentiation entre sous groupes et une forte corrélation avec les marqueurs utilisés pour estimer l’apparentement (K). Nous avons donc considéré deux estimateurs alternatifs de K. Des simulations ont montré qu'ils sont aussi efficaces que la méthode classique pour contrôler

les faux positifs et augmentent la puissance. Ces résultats ont été confirmés sur les panels corné et denté du programme Cornfed, avec une augmentation de 40% du nombre de SNP détectés. Ces panels, génotypés avec une puce 50k SNP et phénotypés pour leur précocité et leur biomasse ont permis de décrire la diversité de ces groupes et de détecter des QTL. En SG, des études ont montré l'importance de la composition du jeu de calibration sur la fiabilité des prédictions. Nous avons proposé un algorithme d'échantillonnage dérivé de la théorie du G-BLUP permettant de maximiser la fiabilité des prédictions. Par rapport à un échantillon aléatoire, il permettrait de diminuer de moitié l'effort de phénotypage pour atteindre une même fiabilité de prédiction sur les panels Cornfed.

Mots clés : maïs, sélection génomique, génétique d'association, puissance, fiabilité, biomasse.

Table of contents

General introduction

Plant breeding appeared 9 000 to 12 000 years ago, when humans became sedentary and developed agriculture. The first plants that were cultivated for a given species accumulated alleles which facilitated the cultivation, harvest and/or use of harvested products. Note that these favorable alleles may have long existed in wild populations or appeared simultaneously through mutations. This transition from wild reproduction to cultivation occurred independently for many species in several regions of the world and is referred to as domestication. After the first steps of domestication, the process was continued by farmers to increase the value of plants for previous criteria. Both during domestication and later steps, seeds from the plants with the best agronomical characteristics were selected for the sowing of the next season. Divergence between domesticated individuals and their wild ancestors increased with time and could result in huge phenotypic variability. This is for example the case of maize (*Zea mays* ssp. *mays*), which became very different from the teosinte subspecies (ssps. *parviglumis* and *mexicana*) from which it was domesticated in Mesoamerica starting around 9000 years ago (Beadle 1939; Matsuoka *et al*. 2002; Doebley 2004). The selection of individuals of higher phenotypic value, called selective breeding, generated plants improved in terms of utility for humans (e.g. yield, composition, precocity), instead of maximizing fitness only as would natural selection do. Selective breeding was used for millennia, until the 20th century for maize. One main limit of this approach is that it is based on the phenotype of single plants in a particular environment. As this phenotype is the result of both genotypic and environmental factors, it does not reflect directly the genetic potential, i.e. the Genetic Value (GV). This could be conceptualized only in the early 1900s after the founding work of precursory scientists.

Gregor Mendel, considered as the founder of genetics, first understood and described the inheritance of traits influenced by few genes (qualitative traits) by studying the segregation of color and shape in peas (Mendel 1866). His work was synthesized into the famous laws of inheritance: the law of segregation and the law of independent assortment. Approximately at the same time Francis Galton developed statistical approaches (1869, 1879) to study quantitative traits (continuous traits, for example human height), laying the foundation of the biometrical school. Mendel's theory was criticized at this time, in particular because it could not explain how continuous traits are inherited, and was thought to be contradictory to the approach of Francis Galton. R. A. Fisher later proved (1918) that Mendel's laws could be extended to continuous traits by showing that the combined effect of many genes and the environment could give rise to continuous phenotypic variations. It is also in the early 20th

century that W. Johannsen introduced the notions of genotype and phenotype in his famous experiments on variability between and within pure lines of beans (1903). These first developments of quantitative genetics allowed the mathematical formalization of the relationship between genotype and phenotype, the phenotype being seen as a realization of a genotype in an environment. This gave birth to many concepts of applied statistics used in numerous and various fields. Evolutionary theories also developed since the mid 1800s with the concept of natural selection (see Darwin's seminal book "On the Origin of species", 1859). This concept together with gradual evolution, and Mendelian genetics were synthesized in the so called "modern evolutionary synthesis" (Huxley 1942), the most accepted paradigm in evolutionary biology, which establishes that variation has to be heritable to undergo natural selection. We now know that these variations submitted to natural selection can have different origins including genetic and epigenetic factors.

In animal and plant breeding, statistical models could then be developed to predict and compare the gain of different selection strategies (Falconer and Mackay 1996; Gallais 1990), and as a result optimize these. Genetic progress was formally decomposed into four components: genetic variability, selection intensity, generation interval, and the accuracy of the estimations of GVs. The replicated observation of a genotype in different environments, or the observations of related individuals (first statistically modeled by Henderson, 1963) allowed the distinction between the effect of the genotype (GV), the effect of the environment (micro and macro environment), and the potential interaction between the genotypic and environmental effects. The selection strategies based on GV estimates have been extensively and efficiently used in breeding. In plant breeding, the possibility to generate numerous individuals with the same genotype, through cloning or most often the production of inbred lines, allows the evaluation of the genotype in field trial networks. This was the most common approach used for phenotypic evaluation in plants until recently. In maize, which is mostly allogamous, inbred lines have poor performance because of inbreeding and are thus crossed to produce hybrids, taking advantage of heterosis (Shull 1908). Hybrid breeding in maize contributed to a huge increase in productivity, with average grain yields increasing from 1.5 to 8 t/ha between 1935 and 2000 in the USA (Troyer 2005). One limitation of these strategies mainly based on phenotypic data is that they are conducted without knowing the genes underlying the variation of the phenotypic trait (number, positions, and effects) and thus without knowing the favorable alleles that could be combined to produce an improved genotype.

The question is then, how to identify favorable alleles ? A first answer was obtained, again on peas, by Sax (1923), who identified an association between the size (quantitative trait) and the color (qualitative trait) of seeds. His experiment thus revealed that a local mutation (responsible for the seed color) was associated with a quantitative trait (the size of the seeds). The color can be seen here as a phenotypic marker: it directly reveals the genotype at a locus (implied in seed color), which was associated with the genotype at a locus influencing a quantitative trait (a so-called Quantitative Trait Locus or QTL, associated here with the size) through physical linkage. The law of independent assortment states that the alleles at a locus segregate independently from the alleles at another locus during meioses if they are located on different chromosomes. If not, the two loci are physically linked and are separated only if a crossover occurs between them. The probability that a recombination occurs during meioses defines the concept of genetic distance (expressed in centiMorgan, cM). As a consequence, two linked genes are more or less correlated (in Linkage Disequilibrium, LD), depending on the genetic distance that separates them. Correlation between two linked loci implies that a marker can capture (at least partially) the effect of nearby QTL(s). Phenotypic markers are however often of poor interest, because they are rare and often dominant. The development of molecular markers in the 1960s made it possible to carry out the first QTL detection experiments with 10-30 polymorphic markers within a given population. The first molecular markers were protein variants (isozymes) identified by electrophoresis. These variations have the advantage of being codominant but they are not very polymorphic and not numerous enough to cover the entire genome. In the 1980s, new approaches appeared, enabling to detect polymorphism at the DNA level, revealing polymorphism in the presence or absence of restriction sites (Restriction Fragment Length Polymorphism, RFLP), in the length of the amplified fragments (Amplified Fragment Length Polymorphism, AFLP) or in the number of copies of microsatellites (Single Sequence Repeat, SSR). This permitted the development of an increasing number of markers and the first genomewide QTL mapping approaches really started in 1988 with the seminal paper of Paterson *et al.*. The progress made in DNA sequencing later allowed the identification of numerous polymorphisms at the level of single nucleotides (called SNP). These SNPs rapidly became the most commonly used markers, because they can be automatically analyzed with SNP-arrays providing cheap, numerous and codominant markers. The fact that SNPs are generally biallelic, and thus less informative than SSRs, is counterbalanced by the fact that thousands to millions of SNPs are now available for many species. High throughput SNP-arrays have been developed and are extensively used in

human, animal and plant genetics. In maize, a 50,000 SNP-array was developed (GANAL *et al.* 2011) following the sequencing of B73 (ZHOU *et al.* 2009; WEI *et al.* 2009a; WEI *et al.* 2009b), the first maize inbred line sequenced, and the resequencing of numerous inbred lines. Technological progress in sequencing makes it now possible to genotype individuals directly by sequencing portions of their genomes. Several Genotyping By Sequencing (GBS) strategies are now available (ELSHIRE *et al.* 2011).

These tools, combined with phenotypic data, offer different ways of detecting QTLs. In linkage-based QTL detection, individuals with contrasted phenotypes are crossed to produce a segregating population. In this kind of populations linkage between markers and QTLs makes it possible to detect associations between phenotypic variability and marker polymorphism. Major QTLs were detected with this approach, and the underlying gene was sometimes identified after analyzing numerous recombinant individuals in the genomic region of interest (HUANG *et al.* 1997; SALVI and TUBEROSA 2005; GIULIANI *et al.* 2005; DUCROCQ *et al.* 2009). However the low diversity of the material used as parents (a significant proportion of QTLs are monomorphic), and the low resolution of the detection (often confined to a range of 10 to 30 cM, FLINT-GARCIA *et al.* 2003; ZHU *et al.* 2008) are important limits to this approach and makes it difficult to identify the underlying genetic factor(s).These difficulties can be circumvented to some extent by increasing the number of parents and the size of the total population (YU *et al.* 2008; CAVANAGH *et al.* 2008; BARDOL *et al.* 2013).
Also, the fast increase of available molecular markers allowed to work on more diverse materials with no or limited relatedness. The approach known as Genome Wide Association Study (GWAS) consists of combining genotypic and phenotypic information of diversity panel in a statistical model to detect marker-trait associations. Such panels have accumulated numerous historical recombination events between highly diverse ancestral haplotypes. It results in a lower LD extent than in segregating populations, and as a consequence a much higher resolution (RAFALSKI and MORGANTE 2004). However, contrary to linkage mapping populations, LD in association mapping panels is not only due to genetic linkage, but can also be caused by population structure, relatedness, drift and selection (JANNINK and WALSH 2002; FLINT-GARCIA *et al.* 2003). The contribution of these factors relative to linkage can be evaluated statistically (MANGIN *et al.* 2012) and proved for instance to be substantial in grapevine and maize (MANGIN *et al.* 2012; BOUCHET *et al.* 2013). This component of LD due to population structure and relatedness can generate false positives and has thus to be taken into account in association mapping models (EWENS and SPIELMAN 1995; THORNSBERRY *et*

*al.*2001). Once these effects are correctly modeled, only marker-trait associations due to linkage should be detected. Population structure (Q matrix) and kinship (K matrix) are unknown but they can be estimated using molecular markers (PRITCHARD *et al.* 2000; PRICE *et al.* 2006; VANRADEN 2008; ALEXANDER *et al.* 2009; ASTLE and BALDING 2009). Major genes were identified with GWAS in human, animal and plant genetics (OZAKI *et al.* 2002; BELÓ *et al.* 2007; JONES *et al.* 2008). However, one of the main drawback of these structure and relatedness corrections is that it also reduces the number of detectable true positives, particularly if the trait is correlated to the population structure (LARSSON *et al.* 2013). For this reason, it is of highest importance to estimate Q and K in an efficient way to maximize detection power and control false positive rate efficiently (YU *et al.* 2006).

Once QTLs have been detected, markers can be used in breeding programs to follow the favorable alleles in a cross to select improved individuals. This marker-assisted selection (MAS) has typically been efficiently used to introgress resistance alleles in elite material (SANZ-ALFEREZ *et al.* 1995; THABUIS *et al.* 2004; RANDHAWA *et al.* 2009; RIAR *et al.* 2012). This is more difficult when the trait is influenced by many genes, which is often the case in quantitative traits. In that case only the main QTLs are detected, and as a result only a fraction of the total genetic variability is explained. In addition to this, it becomes difficult to pyramid all the favorable alleles in one individual (SERVIN 2004) when the number of QTLs is high (HOSPITAL and CHARCOSSET 1997). In such cases, LANDE and THOMPSON (1990) proposed to select individuals based on an estimation of their genetic value obtained by summing the effect of markers significantly associated to QTLs and possibly combine this information with the phenotype to manage undetected QTL. Comparison of different MAS strategies revealed that the main interest of marker-based selection was its efficiency to reduce generation interval (HOSPITAL *et al.* 1997). One limit of this approach is that the selection of individuals based on their QTL-based predictions often result in the fast fixation in the first generations of favorable alleles at the biggest QTLs but not at the others (HOSPITAL *et al.* 1997; MOREAU *et al.* 2004). Moreover, the marker-QTL associations tend to decrease along generations due to the accumulation of recombination events, which reduces the efficiency of MAS. Finally, the effect of the detected QTLs is often overestimated because only significant associations are considered, and these detected associations are likely to be biased upward (Beavis 1998). Correlatively, the use of a significance threshold implies that the identified QTLs capture only a fraction of the genetic variance of quantitative traits, even if a sufficient coverage is used. This phenomenon was first described in human genetics and defined as the "missing

heritability" (Maher 2008). We now know that considerable population size is required to get sufficient power for the detection of small to intermediate QTLs (VISSCHER 2008), which is expensive and not always possible. This is an important problem in the deciphering of genetic architecture because most of the quantitative traits of interest are influenced by many genes of small effect (oil content or flowering time in maize were found to be influenced by more than 50 QTLs, LAURIE *et al.* 2004; BUCKLER *et al.* 2009).

When the number of QTLs is that high, it becomes interesting to estimate all the marker effects simultaneously to circumvent the limitations of QTL detection. In that case, the objective is to predict as accurately as possible the GVs of individuals candidate to selection, including possibly unphenotyped individuals. This was first proposed by WHITTAKER *et al.* (2000) and further formalized and extended to situations where the number of markers is much higher than the number of observations by MEUWISSEN *et al.* (2001), who called this approach genomic selection (GS). GS can be applied as follows: in a first step the genotypes and phenotypes of reference individuals (the calibration set) are combined to calibrate the chosen statistical model (RR-BLUP, RA-BLUP, BayesA, BayesB or others, see HESLOT *et al.*(2012) for a review). In a second step, the calibrated model is used to predict the genotyped selection candidates, which can then be selected without being phenotyped. These individuals can (i) belong to the same generation as the calibration set, making it possible to increase selection intensity, or (ii) belong to a next generation of yet unphenotyped individuals, making it possible to conduct new cycles of selection more rapidly. GS is expected to be more efficient than post-QTL MAS, because a more important part of the genetic variance is captured, reducing the amount of missing heritability (YANG *et al.* 2010). MEUWISSEN *et al.* (2001) proposed prediction models based on the mixed model or the bayesian frameworks, which combine the information brought by the observations and prior knowledge on the trait architecture (for example obtained from QTL detections). In the mixed model with all available markers included as random effects (Ridge Regression Best Linear Unbiased Prediction, or RR-BLUP), we suppose that the traits is influenced by a large number of genes having small and independent effects (infinitesimal model). This assumption seems reasonable for many quantitative traits and the predictions obtained with RR-BLUP are often as accurate as more complex models, such as Bayesian models, neural networks, or machine learning (HESLOT *et al.* 2012; RESENDE *et al.* 2012). However, prior assumptions on the proportion of causal SNPs can sometimes extent the validity of the model to more genetically distant individuals (HABIER *et al.* 2007). Interestingly, some studies revealed that the

prediction accuracies was not only due to LD between markers and QTLs but also to the efficiency of the markers to capture relatedness between individuals (HABIER *et al.* 2007). Molecular markers can indeed be used to estimate kinship between individuals (LOISELLE *et al*. 1995; RITLAND 1996; VANRADEN 2008; ASTLE and BALDING 2009) and the resulting realized relationship matrix can be more informative than pedigree because it takes Mendelian sampling into account (and pedigree information is not always available, and sometimes of poor quality). It was proven that a traditional BLUP model with pedigree matrix replaced by realized kinship was equivalent to RR-BLUP in some conditions presented by HABIER *et al.*(2007), GODDARD (2009) and HAYES *et al.* (2009b). This mixed model (called Realized Additive BLUP or RA-BLUP, ZHONG *et al.* 2009) is close to the classical model used in GWAS to control false positives (YU *et al.* 2006). GS has been successfully implemented in dairy cattle and is expected to double genetic progress thanks to the replacement of progeny testing by genomic predictions, and could potentially diminish inbreeding at the same time (HAYES *et al.* 2009a). In plant breeding, simulations (ZHONG *et al.* 2009; JANNINK 2010; HESLOT *et al.* 2012) and fields experiments (CROSSA *et al.* 2010; ALBRECHT *et al.* 2011; ZHAO *et al.* 2011; HOFHEINZ *et al.* 2012; WINDHAUSEN *et al.* 2012; BARDOL *et al.*, in review) gave encouraging results in populations with variable levels of diversity. BERNARDO and YU (2007) showed for instance using simulations, that GS provided 18 to 43% more genetic gain per cycle than traditional marker assisted recurrent selection in biparental populations. CROSSA *et al.* (2010) confirmed the potential interest of GS in more diverse material. Theoretical and experimental results revealed few critical aspects, which have imperatively to be considered when designing GS procedures including marker density, statistical model, phenotypic evaluation, and genetic distance between and within the calibration set and the predicted individuals. All these factors influence the accuracy of the predictions and as a result the genetic progress. Because the predictive ability of a model relies on the kinship between individuals and the LD between QTLs and markers, it is quite clear that relatedness between the calibration set and the prediction set, and the accordance of LD phase in both sets can affect accuracies. Some studies revealed indeed that prediction accuracy could be considerably reduced in case of low relatedness between both sets (HABIER *et al.* 2007, 2010; LY *et al.* 2013; RIEDELSHEIMER *et al.* 2013). It is therefore of the highest importance to define the calibration set in an efficient way.

Molecular markers are therefore of considerable interest in genetics to either detect loci of interest and/or improve selection efficiency. Because markers can capture QTL effects thanks

to LD, they can be used to detect QTLs (for example in GWAS) or to predict GVs (GS). GWAS and GS are based on close statistical models, but in GWAS the objective is to detect QTLs, whereas in GS the objective is to predict GVs. The efficiency of different GWAS and GS strategies can be estimated and possibly optimized by estimating their detection power (for GWAS) or their prediction accuracy (for GS). The main objective of this thesis was to optimize the use of available molecular information to maximize QTL detection power in GWAS and prediction accuracy in GS. For this, we proposed new approaches that can be used at critical steps of GWAS and GS, namely the estimation of a relevant kinship matrix to maximize power and control false positive rate efficiently in GWAS, and optimize the composition of the calibration set in GS to maximize prediction accuracy of selection candidates. These approaches were evaluated and compared to existing procedures using simulations based on existing genotypes and using true experimental data. These experimental data were obtained within the European "Cornfed" project, which was developed to characterize the variation of biomass related traits in maize in view of increasing the efficiency of breeding programs targeting this trait. This project includes in particular a Dent (CF-Dent) and a Flint (CF-Flint) panels, expanding a previous panel comprising less representatives of these groups and also including tropical materials (CK-panel, CAMUS-KULANDAIVELU *et al*., 2006). Flint and Dent represent complementary heterotic groups to create hybrid varieties adapted to Northern European environmental conditions. The two Cornfed panels, each composed of 300 lines, were genotyped with the 50,000 SNP-array and phenotyped in a Western European trial network for traits related to flowering time and biomass productivity.

The first chapter of this thesis is dedicated to the analytical study of power in GWAS in panels presenting different levels of diversity. It highlights the parameters influencing power and proposes new kinship estimators to maximize power. The efficiency of these estimators are evaluated with simulations based on the CF-Dent, CF-Flint and CK-panel (CAMUS-KULANDAIVELU et al. 2006) genotypes. In the second chapter, we used molecular (50k SNP-array) to analyze diversity and Linkage Disequilibrium (LD) in the CF-Dent and CF-Flint panels. Phenotypic variation for flowering time and biomass production was analyzed based on 10 to 11 Western European trials. Chapter 2 also presents GWAS results using models derived in chapter 1, illustrating the interest of approaches evaluated in chapter 1 through simulations. The third and last chapter is devoted to the optimization of the calibration set in GS. We proposed an algorithm for this, and validated its ability in the CF-Dent and CF-Flint

panels. These three chapters are presented as scientific articles, chapters 1 and 3 were published in Genetics, and chapter 2 is organized in view of submission to Theor. Appl. Genet.. The chapters were not ordered chronologically with respect to work realized during the PhD, but in a way that, for both GWAS and GS approaches, methodological aspects are presented first, and then followed by application on true phenotypes. GWAS was presented first and GS second, because we characterized the Cornfed panels in terms of diversity, Linkage Disequilibrium and detection power in a same study. It also seemed interesting to us to present first insights in the genetic determinism of traits to facilitate the interpretation of GS results. Finally, limits and perspectives of the present work with respect to genetic analyses and breeding applications are discussed in a last section.

Chapter 1

Recovering power in association mapping panels with variable levels of linkage disequilibrium

R. Rincent,[1,2,3,4] L. Moreau,[1] H. Monod,[5] E. Kuhn,[5] A.E. Melchinger,[6] R. A. Malvar,[7] J. Moreno-Gonzalez,[8]S. Nicolas,[1] D. Madur,[1]V. Combes,[1]F. Dumas,[1]T. Altmann,[9]D. Brunel,[10]M. Ouzunova[3], P. Flament[4], P. Dubreuil[2], A. Charcosset[1,12], T. Mary-Huard[1,11]

[1]UMR de Génétique Végétale, INRA – Université Paris-Sud – CNRS, 91190 Gif-sur-Yvette, France,
[2]BIOGEMMA, Genetics and Genomics in Cereals, 63720 Chappes, France,
[3]KWS Saat AG, 37555 Einbeck, Germany,
[4]Limagrain, site d'ULICE, BP173, 63204 Riom Cedex, France,
[5]INRA, Unité de Mathématique et Informatique Appliquées, UR 341, 78352 Jouy-en-Josas, France.
[6]Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70599, Stuttgart, Germany,
[7]Misión Biológica de Galicia, Spanish National Research Council,36080 Pontevedra, Spain,
[8]Centro de Investigaciones Agrarias de Mabegondo, 15080 La Coruna, Spain
[9]Max-Planck Institute for Molecular Plant Physiology, 14476 Potsdam-Golm, Germany,
Leibniz-Institute of Plant Genetics and Crop Plant Research (IPK), 06466 Gatersleben, Germany,
[10]INRA, UR 1279 Etude du Polymorphisme des Génomes Végétaux, CEA Institut de Génomique, Centre National de Génotypage, CP5724, 91057 Evry, France.
[11] INRA/AgroParisTech, UMR 518, 75231, Paris, France

Short running title: Recovering power in association mapping

Keywords: Association mapping, power, kinship, linkage disequilibrium, *Zea mays* L.

[12]**Corresponding author:** Alain Charcosset
UMR de Génétique Végétale
INRA - Univ Paris-Sud - CNRS - AgroParisTech
Ferme du Moulon,
F-91190, Gif-sur-Yvette, France
Tel: +33 1 69 33 23 35
Fax: +33 1 69 33 23 40
E-mail: alain.charcosset@moulon.inra.fr

ABSTRACT

Association mapping has permitted the discovery of major QTLs in many species. It can be applied to existing populations and, as a consequence, it is generally necessary to take into account structure and relatedness among individuals in the statistical model to control false positives. We studied analytically power in association studies by computing non-centrality parameter of the tests and its relationship with parameters characterizing diversity (genetic differentiation between groups and allele frequencies) and kinship between individuals. Investigation of three different maize diversity panels genotyped with the 50k SNPs array highlighted contrasted average power among panels and revealed gaps of power of classical mixed models in regions with high Linkage Disequilibrium (LD). These gaps could be related to the fact that markers are used for both testing association and estimating relatedness. We thus considered two alternative approaches to estimate the kinship matrix to recover power in regions of high LD. In the first one, we estimated the kinship with all the markers located on other chromosomes than the tested SNP. In the second one, correlation between markers was taken into account to weight the contribution of each marker to the kinship. Simulations revealed that these two approaches were efficient to control false positives and more powerful than classical models.

INTRODUCTION

Quantitative traits are determined by the polymorphism of many genes or genomic regions with small effects, i.e. Quantitative Trait Loci (QTL). Understanding the genetic architecture of such traits, which supposes the identification of these causal loci, is now facilitated by a dramatic increase in the number of molecular markers available. This makes it possible to conduct genome-wide association studies (GWAS), in which phenotypes and genotypes of individuals in highly diverse panels are used to detect QTLs (LYNCH and WALSH, 1998). Such panels have accumulated numerous historical recombinations, leading to a low extent of linkage disequilibrium (LD). Compared to linkage mapping, more markers are therefore needed to capture causal signals but with a much higher mapping resolution (RAFALSKI and MORGANTE 2004). Major genes were identified by this approach in human, animal and plant genetics (OZAKI *et al.* 2002; BELÓ *et al.* 2007; JONES *et al.* 2008). However, contrary to linkage mapping populations, LD in association mapping panels is not only due to genetic linkage, but can also be caused by population structure, relatedness, drift and selection (JANNINK and WALSH 2002; FLINT-GARCIA *et al.* 2003). The contribution of these factors relative to linkage can be evaluated statistically (MANGIN *et al.*2012) and proved for instance to be substantial in grapevine and maize (MANGIN *et al.* 2012; BOUCHET *et al.* 2013). This component of LD due to population structure and relatedness can generate false positives and has thus to be taken into account in association mapping models to control false positives (EWENS and SPIELMAN 1995; THORNSBERRY *et al.* 2001). Once these effects are correctly modeled, only marker-trait associations due to linkage should be detected.

Population structure can be estimated with softwares such as STRUCTURE (PRITCHARD *et al.* 2000) and ADMIXTURE (ALEXANDER *et al.* 2009), or by Principal Component Analysis on the genotypic data (PRICE *et al.* 2006). These methods permit the estimation of a structure matrix (***Q***) attributing the admixture coefficient of each individual in each group. Relatedness (***K*** matrix) can be estimated in different ways including Identity By State (IBS), or estimators of Identity By Descent (IBD) considering marker allelic frequencies (VANRADEN 2008; ASTLE and BALDING 2009). YU *et al.* (2006) proposed a mixed model approach (***Q***+***K***) to detect QTL in the context of association mapping. This model has the advantage of controlling false positive rate by including a fixed structure effect (through ***Q***) and/or a random polygenic effect (through ***K***). It was used in many association mapping studies and permitted the detection of QTLs in humans, animals and plants (ZHAO *et al.* 2007a; HUANG *et al.* 2010; KANG *et al.* 2010a; PRICE *et al.* 2010; ZHANG *et*

*al.*2010; BOUCHET *et al.* 2013; ROMAY *et al.* 2013). However, one of the main drawbacks of these structure and relatedness corrections is that it also reduces the number of detectable true positives, particularly if the trait is correlated to the population structure (LARSSON *et al.* 2013).Also, including the tested SNP in the computation of ***K*** is expected to decrease power at this SNP (LISTGARTEN *et al.* 2012). In order to increase the power of GWAS, some authors therefore proposed to use only a subset of SNPs as covariates or to estimate genetic similarity (LISTGARTEN *et al.* 2012; BERNARDO 2013). SPEED *et al.*(2012) proposed to weight the contribution of the SNPs in the kinship estimation to increase the accuracy of heritability estimates.

It is particularly important to evaluate the power of panels and statistical approaches to discover QTLs. Power may be analytically investigated using the non-centrality parameter of the test statistics. This strategy has first been applied in linkage mapping, where several authors showed how power is influenced by the size of the population, heritability, the effect captured by the marker and the allelic frequencies (SOLLER *et al.* 1976; KNAPP and BRIDGES 1990; REBAI and GOFFINET 1993; CHARCOSSET and GALLAIS 1996). Such analytical approach has also been applied in association studies in human and animal genetics (SHAM *et al.* 2000; WANG 2008; PURCELL *et al.* 2003; TEYSSÈDRE *et al.* 2012). Alternatively, the estimation of power has also been addressed through simulation studies (see for instance STICH and MELCHINGER 2009; ERBE *et al.* 2010; MACLEOD *et al.* 2010; BRADBURY *et al.* 2011; YU *et al.* 2006; ZHAO *et al.* 2007b). We can retain from these studies that power of association mapping diminishes with structure and relatedness in addition to the parameters identified in linkage analysis, and that the way of estimating ***K*** has an effect on power (STICH *et al.* 2008). To our knowledge no study was conducted to compare analytically the power along the genome in different association mapping designs.

In this study we derived analytically the power at each marker for the classical mixed model involving relatedness between individuals (YU *et al.*, 2006). This analytical expression of power makes it possible to study the effect of different parameters on local power along the genome. We first used it to compare three diversity panels with different diversity patterns. We highlighted a loss of power due to the use of the genotypic information both to test marker effect and to estimate ***K***, and that this was particularly strong in regions of high LD. We therefore evaluated two alternative estimation strategies of the kinship matrix to increase power in GWAS. In the first one, we used an estimated ***K*** matrix specific to each chromosome: only the markers that are physically unlinked to the tested SNP are used to estimate ***K***. In the second one, we weighted the contribution of each marker in the estimation of ***K*** by taking into account intra-chromosomic LD. We compared in

simulations based on true genotypes of maize inbreds the efficiency of the different strategies to detect QTLs and to control false positives.

MATERIALS AND METHODS

Statistical models for association mapping and power evaluation

Mixed models are now routinely used to control type I error in GWAS (YU *et al.* 2006). Relatedness among individuals is taken into account by considering that the random polygenic effects are not independent, with a covariance matrix determined by kinship (***K***, with as many rows and columns as individuals: N). As ***K*** includes information on both population structure and relatedness, it is in general not useful to consider admixture information as fixed effects covariates (ASTLE and BALDING 2009). We therefore considered the following statistical model (denoted by $\mathcal{M}_{\mathrm{K}}$):

$$\boldsymbol{Y} = \mathbf{1}\mu + \boldsymbol{X_l}\beta_l + \boldsymbol{U} + \boldsymbol{E}\ ,$$

$$= \boldsymbol{X\beta} + \boldsymbol{U} + \boldsymbol{E}\ \text{, with } \boldsymbol{X} = [\mathbf{1}\boldsymbol{X_l}] \text{ and } \boldsymbol{\beta^T} = (\mu, \beta_l)$$

where $\boldsymbol{Y}$ is the vector of N phenotypes, μ is the intercept, $\mathbf{1}$is a vector of N 1, $\boldsymbol{X_l}$ is the vector of N genotypes at the tested locus (0 and 1 corresponding to homozygotes and 0.5 to heterozygotes), β_l is the additive effect of locus *l* to be estimated, $\boldsymbol{U} \rightsquigarrow N(0, \boldsymbol{K}\sigma^2_{gl})$ is the vector of random polygenic effects, σ^2_{gl} being the residual polygenic variance, $\boldsymbol{E} \rightsquigarrow N(0, \boldsymbol{I}\sigma^2_e)$ is the vector of remaining residual effects with variance σ^2_e, $\boldsymbol{I}$ is an identity matrix of size equal to the number of individuals (N), *U* and *E* are independent.

Locus effects in this mixed model can be tested using Wald statistics (WALD 1943). In the general case, a given linear combination of fixed effects $\boldsymbol{L^T\beta} = 0$ (H0 hypothesis) can be tested against $\boldsymbol{L^T\beta} \neq 0$ (the alternative hypothesis H1) using:

$$\boldsymbol{W} = \left(\boldsymbol{L^T}\widehat{\boldsymbol{\beta}}\right)^{\boldsymbol{T}} \left[\boldsymbol{L^T}\left(\boldsymbol{X^T}\left(\boldsymbol{K}\hat{\sigma}^2_{gl} + \boldsymbol{I}\hat{\sigma}^2_e\right)^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{L}\right]^{-1}\left(\boldsymbol{L^T}\widehat{\boldsymbol{\beta}}\right),$$

where $\widehat{\boldsymbol{\beta}}$ is a vector of fixed effect estimates, $\boldsymbol{L}$ is a linear combination, $\hat{\sigma}^2_{gl}$ and $\hat{\sigma}^2_e$ are the REML estimates of σ^2_{gl} and σ^2_e.

In GWAS we test the particular linear combination: $\boldsymbol{L}^T\boldsymbol{\beta} = \beta_l = 0$against $\boldsymbol{L}^T\boldsymbol{\beta} = \beta_l \neq 0$, with $\boldsymbol{L} = \binom{0}{1}$if the only fixed effects are the intercept and the marker additive effect. Note that the approach could be extended to more complex effects such as dominance by adding extra term(s) in fixed effects. When the variances are known, W follows a χ^2 distribution: $\chi^2(\nu_1;\ NCP = \lambda)$ where $\nu_1 = rank(\boldsymbol{X_l}) = 1$ and λ is the non-centrality parameter (NCP). The non-centrality parameter is equal to:

$$\lambda = \beta_l \left[\boldsymbol{L^T}\left(\boldsymbol{X^T}\left(\boldsymbol{K}\sigma^2_{gl} + \boldsymbol{I}\sigma^2_e\right)^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{L}\right]^{-1}\beta_l.$$

Under H0, $= 0$; whereas under H1, λ is positive. Power can thus be determined as the probability $P(\chi^2_{[ddl=\nu_1\,;\,NCP=\lambda]} > \chi^2_{crit})$, λ being the NCP and $\chi^2_{crit} = \chi^2_{[ddl=\nu_1\,;\,NCP=0\,;\,1-\alpha]}$ the value of the central χ^2 (1-α) quantile, where α corresponds to the chosen type I error level. The power of the test increases as the NCP increases. λdepends on the QTL effect β_l (the magnitude of departure from H0), the marker genotypes and the variance and covariance components. Hence in addition to the number of individuals, power can be influenced by the marker genotypes, the marker effect (β_l), the heritability (through $\sigma^2_{gl}$and σ^2_e) and the relatedness between individuals (***K***).

Analytical evaluation of the impact of panel characteristics on power

When genotypic data are available in a given association mapping panel, it is possible to evaluate analytically power at each marker thanks to the above formula. Consider a panel where N individuals were genotyped at M markers (SNPs). The potential power at a given marker can be investigated by setting a QTL effect β_l, a background genetic variance σ^2_{gl} and a residual variance σ^2_e to reach a given heritability h^2.Power at a given marker can then be related to parameters characterizing the marker in the panel of interest. It is first expected to depend on allele frequencies, that can be characterized by the Minor Allele Frequency (MAF). Also, according to the analytical expression of the NCP, power at a marker in $\mathcal{M}_K$ can be influenced by its correlation with the kinship that reflects both the structure of the panel and the relationships between individuals. It is thus interesting to relate power at a given marker to its Nei's index of differentiation (*Fst*) among genetic groups (NEI, 1973) and to its correlation with the kinship matrix. Let us denote by $\boldsymbol{K\_M_l}$ the kinship matrix evaluated from the considered marker *l* only. To define how power at a given marker is affected by its correlation to ***K***, one can calculate the correlation between $\boldsymbol{K\_M_l}$ and ***K*** at each marker. This correlation between local and global kinship is further referred to as *CorK*. These statistics (*Fst*, MAF, *CorK* and analytical power) can be calculated for each marker in any association mapping panel.

In this article, we applied this strategy to three maize panels (see below). We represented the relationship between MAF, *Fst*, *CorK* and local power with the two following approaches. In the first one, analytical power was represented as level plots considering MAF and *Fst* as x and y-axes, with the R function level.plot. The same procedure was applied to MAF and *CorK*. In the second approach, cubic smoothing splines were adjusted along the genome to the *Fst*, *CorK* and power for the markers with a MAF above 0.4, using the R function smooth.spline (HASTIE and TIBSHIRANI 1990).

Kinship estimation

In practice the kinship matrix ***K*** is unknown and has to be estimated. One classically used estimator was proposed by ASTLE and BALDING(2009) and is defined as: $K\_Freq_{i,j} = \frac{1}{L}\sum_{l=1}^{L} \frac{(G_{i,l}-p_l)(G_{j,l}-p_l)}{\sigma_l^2}$, where $G_{i,l}$ and $G_{j,l}$ are the genotypes of individuals i and j at marler l ($G_{i,l}$= 0 or 1 for homozygotes, 0.5 for heterozygotes), p_l is the frequency of the allele coded 1, σ_l^2 is the variance of $G_{i,l}$, respectively. One problem that might arise from this formula and other classical estimators as the Identity by State, or the formula of VanRaden (2008), is that LD between SNPs is not taken into account. As a result more weight is given in the kinship estimation to the regions of the genome that carry several markers in strong LD and power may be lower in these regions.

We therefore considered two alternative approaches to limit this effect. In the first one, the kinship matrix (***K_Chr***) was estimated with all the markers other than those located on the same chromosome as the marker being tested. If the markers located on the other chromosomes are sufficient to reliably estimate relatedness, this method is expected to reasonably control the risk of detecting false positives and avoids considering in the kinship matrix markers linked with the tested marker: $K\_Chr_{i,j,c} = \frac{1}{L_{-c}}\sum_{l \notin c} \frac{(G_{i,l}-p_l)(G_{j,l}-p_l)}{\sigma_l^2}$, where c is the considered chromosome, L_{-c} is the number of markers not located on chromosome c.

In the second approach we used all the markers as estimators of relatedness but we weighted the contribution of each marker. The kinship estimator $K\_Freq_{i,j}$ can be understood as follows: each marker l yields an estimator $\hat{k}_{ijl} = \frac{(G_{i,l}-p_l)(G_{j,l}-p_l)}{\sigma_l^2}$ of the true kinship coefficient k_{ij} between individuals i and j, that are then averaged over all markers to obtain$K_Freq_{i,j} = \frac{1}{L}\sum_l \hat{k}_{ijl}$. This average would be optimal if all estimators had the same variance, and were independent. In practice

none of these conditions is satisfied: the error variance of each estimator depends on the MAF of the marker, and LD between markers generates correlations between markers. As a consequence, estimators with poor precision (high error variance) will have the same weight as estimators with high precision. Moreover, *m* highly correlated estimators will accumulate a weight of *m*/L without providing *m* independent information, i.e. too much weight is attributed to highly correlated estimators. Alternatively, one may look for the weighted combination$K_LD_{i,j} = \sum_l \omega_l \hat{k}_{ijl}$, that is the best linear combination of coefficient $\hat{k}_{ijl}$, $l = 1, \dots, L$ to estimate k_{ij} without bias. Define $\mathbb{E}_{ij}(\hat{k}_{ijl})$ and $\mathbb{V}_{ij}(\hat{k}_{ijl})$ as the mean and variance of estimator $\hat{k}_{ijl}$ over all couples of individuals (i,j) having the same kinship k_{ij}. Note Δ the covariance matrix between estimators $\hat{k}_{ijl}$, i.e. $\Delta_{ll'} = \mathbb{C}ov_{ij}(\hat{k}_{ijl}, \hat{k}_{ijl'})$, $\Omega = (\omega_1, \dots, \omega_L)^T$the vector of weights, and $K_{ij} = \left(\hat{k}_{ij1}, \dots, \hat{k}_{ijL}\right)^T$the vector of marker estimators.Then $K\_LD_{i,j}$ satisfies:

$$
\begin{aligned}
& min\ \mathbb{V}_{ij}\left(K\_LD_{i,j}\right) \text{under constraint } \mathbb{E}_{ij}(K\_LD_{i,j}) = k_{ij} \\
\Leftrightarrow \quad & \min_{\Omega}\ \mathbb{V}_{ij}(\Omega^T K_{ij}) \text{under constraint } \mathbb{E}_{ij}(\Omega^T K_{ij}) = k_{ij} \\
\Leftrightarrow \quad & \min_{\Omega}\ \Omega^T \Delta \Omega \text{under constraint} \Omega^T \mathbb{E}_{ij}(K_{ij}) = k_{ij}
\end{aligned}
$$

In this formulation the optimal weights may be negative, we added extra constraints to ensure the positivity of the weights, leading to the following optimization program:

$$\min_{\Omega}\ \Omega^T \Delta \Omega \text{ under constraint } \Omega^T \mathbb{E}_{ij}(K_{ij}) = k_{ij} \text{ and } \omega_l \geq 0, \text{ for all } l. \qquad (1)$$

In practice, obtaining the optimal weights requires (i) the knowledge of matrix $\boldsymbol{\Delta}$ and (ii) to solve the optimization problem (1). The exact expression of matrix $\boldsymbol{\Delta}$ is unknown, but one can estimate this matrix from the panel data using the classical moment estimator:

$$\widehat{\mathbb{C}ov}_{ij}\left(\hat{k}_{ijl}, \hat{k}_{ijl'}\right) = \frac{n(n-1)}{2} \sum_i \sum_{j>i} (\hat{k}_{ijl} - \widehat{\mathbb{E}}_{ij}\left(\hat{k}_{ijl}\right))(\hat{k}_{ijl'} - \widehat{\mathbb{E}}_{ij}\left(\hat{k}_{ijl'}\right)).$$

The resulting estimated matrix is then plugged into the optimization program (1). Then to solve the optimization program, one should note that (1) is a quadratic problem with linear constraints, and therefore can be solved using classical optimization techniques (in this article we used the R package solve.QP that implements the dual method of GOLDFARB and IDNANI, 1983).

The main limitation of this strategy lies in step (i): when estimating the covariance, one actually replaces the expectation over all couples having the same kinship k_{ij} by an averaging over all couples in the panel - assuming implicitly that they all have the same kinship. Even if the kinship

differs between couples, this weighting increases the contribution of markers with a high diversity (leading to a high precision) and not highly correlated with other markers. It therefore corrects the two drawbacks of the naive averaged estimator mentioned earlier.

Let us denote the statistical model for association mapping described above by $\mathcal{M}_{K\_Freq}$, $\mathcal{M}_{K\_Chr}$ and $\mathcal{M}_{K\_LD}$ with ***K*** estimated as ***K_Freq, K_Chr*** and ***K_LD***, respectively.

Simulation based evaluation of the impact of the estimation of *K* on false positive control and power

The closed form expression of the non-centrality parameter already revealed that kinship affects power. Comparing the impact of different kinship estimators on power implies to evaluate their ability to guarantee the expected nominal control of false positives under different hypotheses on trait genetic determinism. To this end, we simulated traits influenced by L biallelic QTLs (SNPs).In a first step, QTLs were sampled randomly among the SNPs located on all the chromosomes except one. The chromosome without QTL (further referred to as "H0-chromosome") was used to estimate the false positive rate. All the H0-markers (the markers on the H0-chromosome) were tested with the above mentioned statistical models for each run of simulation. The efficiency of the different estimations of ***K*** to control false positives was evaluated by comparing expected and observed quantiles of H0-Pvalues and histograms of H0-Pvalues. In a second step we applied the same procedure, but now sampling the QTLs among the M SNPs (on all chromosomes). A QTL was declared detected when the Pvalue of the corresponding SNP in the genetic model was below the significance threshold. Power of a given model was computed as the number of QTL which were detected. We also applied a less restrictive definition of QTL detection, considering that a QTL could be detected by SNPs located near it. To do so, another analysis was conducted in which markers within a given genetic distance of a QTL were considered H1-markers and the others H0-markers. The realized false discovery rate (FDR) was defined as the proportion of H0-markers among the markers declared significant. Power of QTL detection was estimated by considering that a QTL was detected when at least one of the corresponding H1-markers had a significant Pvalue. This general method will be exemplified with parameters specific to three maize panels described below.

Genetic material and genotyping data

The above mentioned power analyses (analytical evaluation of power and simulation based evaluation of alternative methods) were applied to three diversity panels of maize. The first panel (called C-K) was described in CAMUS-KULANDAIVELU *et al.* (2006). It is composed of 375 inbred lines covering American and European diversity. It includes Tropical, Dent and Flint lines. The second and third panels are the Dent and Flint panels of the "Cornfed" project (CF-Dent and CF-Flint), described in RINCENT *et al.* (2012). They include lines of the C-K panel and lines derived from recent breeding schemes. Both are composed of 300 lines. These panels were genotyped with the 50k SNPs array described in GANAL *et al.* (2011), as presented in BOUCHET *et al*. (2013) and RINCENT *et al.* (2012). Individuals which had marker missing rate and/or heterozygosity higher than 0.1 and 0.05, respectively, were eliminated. Markers, which had missing rate and/or average heterozygosity higher than 0.2 and 0.15, respectively, were eliminated. In each panel, few individuals were highly related. One individual was removed for pairs identical for more than 98% of the loci. In total 315, 277 and 267 individuals and 44487, 45434, and 44255 markers passed the genotyping filter criteria for the C-K, CF-Dent and CF-Flint designs, respectively. Missing genotypes (below 2% in both panels) were imputed with the software BEAGLE (BROWNING and BROWNING 2009). Panels were all adjusted to 267 individuals in order to compare power for a same population size. Individuals removed were chosen at random. To avoid the ascertainment bias noted by GANAL *et al.* (2011), we only used the markers that were developed by comparing the sequences of nested association mapping founder lines (PANZEA SNPs, GORE *et al.* 2009) in the estimation of admixture and relationship coefficients (29996, 30119 and 29132 markers passed the filter criteria for the C-K, CF-Dent and CF-Flint lines respectively).

Admixture in the CF-Dent and CF-Flint panels was investigated using the SNP data with the software ADMIXTURE (ALEXANDER *et al.* 2009), with a number of groups equal to four, determined according to the cross-validation procedure presented in ADMIXTURE. For the C-K panel we used the admixture in five groups estimated by CAMUS-KULANDAIVELU *et al.* (2006) using 55 SSRs chosen for their broad genome coverage and reproducibility. We estimated the differentiation index among genetic groups (*Fst*, NEI, 1973) at each marker using the R package r-hierfstat (GOUDET 2005).

Finally, the relationship between LD and power along the genome can be empirically investigated using two different measures of LD. Raw LD can be estimated as the squared correlation between

allelic doses at two loci (r^2). Linkage related LD (denoted by r^2K) can be estimated using the algorithm proposed by MANGIN *et al.* (2012), which corrects r^2 by ***K_Freq***. LD within these panels (r^2), possibly corrected by ***K_Freq*** (r^2K), was estimated within a sample of 4000 markers regularly spaced on the physical map.

Specific parameterization

For analytical investigation of power in the three maize panels, the total additive genetic variance σ_g^2 was set to 1000, β_l was set to 17.9, which corresponds to a QTL explaining 8% of the total genetic variance if it had a minor allele frequency (MAF) of 0.5, σ_e^2 was chosen to get an heritability of 0.8. Under these hypotheses, analytical power was investigated for an α type I risk equal to $1{,}25\ 10^{-6}$ which led to a risk of 0.05 with a Bonferroni correction on 40 000 tests. We also considered less stringent threshold corresponding to Bonferroni corrections on 4 000 and 400 tests, although the number of tests was always the same. Power under these hypotheses was calculated in R 3.0.0 (R development Core Team, 2013) for each marker.

To estimate kinship with the different formulas presented above, we considered that all individuals were inbred and we estimated σ_l^2as $p_l(1-p_l)$.For comparing the different methods for kinship estimation, we simulated traits influenced by 50 or 100 biallelic QTLs (QTL effects follow a geometric series as in LANDE and THOMPSON (1990), with parameter *a* set to 0.96 and 0.98 when 50 or 100 QTLs were simulated, respectively). Sign of allelic effect at a given locus was assigned randomly. Genotypic values of the individuals were calculated as the sum of the allelic effects at these QTLs. Phenotypes were obtained by adding a residual noise following a normal distribution with mean 0 and variance equal to: $\sigma_g^2\left({}^{1}/_{h^2}-1\right)$, where the heritability h^2 is set to 0.8.We performed 100 runs of simulations for each scenario using the R 3.0.0 software (R development Core Team, 2013).Each chromosome was used ten times as the H0-chromosome.For all simulations, the statistical tests were made with EMMAX (KANG *et al.* 2010b) to reduce computational time, and then with ASREML-R (GILMOUR *et al*. 2006) on the markers which had a Pvalue below 0.001 with EMMAX. For Pvalues above 0.001, Pvalues obtained with EMMAX and ASREML-R were very close and highly correlated. As investigations of the two criteria for QTL detection (causal factor only or window around it) led to very comparable results with respect to the main focus of our study, results considering a window around causal factor are therefore presented as supplementary information (Table S1).

RESULTS

Diversity and Linkage Disequilibrium in maize panels

Diversity and Linkage Disequilibrium (LD) were investigated within the different panels to provide elements on their ability to detect QTL (ie. their power) along the genome. On average, the Minor Alelle Frequency (MAF) was lower in the CF-Flint than in the other panels. Differentiation among genetic groups (*Fst*) was higher for CF-Dent (0.15) than for C-K (0.11) and CF-Flint (0.08) (Table 1). The raw LD (r^2) and its correction by Kinship (r^2K) were variable between and within panels (Figure 1). LD was on average higher in the dent panel. Within each panel, it was higher for centromeric than for telomeric regions. High r^2 values were observed between physically linked markers but also unlinked markers. This last situation occurred mainly between centromeric regions (Figure 1A, chromosomes 5, 7, and 8 and Figure 1B, chromosome 7). Inter-chromosomic LD was reduced to a large extent when considering r^2K rather than r^2. Taking into account covariance between individuals (r^2K) also reduced intra-chromosomic LD, in particular between distant blocks with high LD (Figure 1B chromosome 10). Considering r^2K instead of r^2 globally had the strongest impact in the CF-Dent panel.

Relationship between MAF, *Fst, CorK* and power

Above described parametrization of QTL effects was used to investigate the influence of MAF, *Fst*, and the correlation between local and global covariance matrices (estimated as *CorK_Freq*) on power in the three maize panels. Level plots (Figure 2) showed that the MAF, the *Fst*, and *CorK_Freq* had important effects on power, with very similar graphs in all the panels. The highest power was achieved when MAF was high and *Fst* or *CorK_Freq* was low. When the MAF was below 0.1, power was close to 0 even if the marker had a low *Fst* or low *CorK_Freq*. Some regions of the level plots were not covered by the available markers (regions in white on Figure 2), in particular there was no marker with a *CorK_Freq* below 0.03.Note that the graphs obtained using ***K_Chr*** (or the ***IBS***) were similar to those obtained with ***K_Freq*** and led to the same general conclusions (results not shown).

The parameters related to power (MAF, *Fst*, *CorK_Freq*) varied between panels (Table 1, see above). As a consequence from above described relationships, the mean analytical power of statistical model $\mathcal{M}_{\text{K\_Freq}}$ varied between the three panels (Table1), and was higher in the C-K panel (11.3%) than in the CF-Dent and CF-Flint panels (below 9.0%).

Variation of analytical power and *CorK* along chromosomes

Power scans (analytical power at each marker plotted against its physical position) of model $\mathcal{M}_{K\_Freq}$ revealed an extreme variability along the genome in the three panels (Figure 3). In all panels, power at a given location ranged from zero to a maximal value, which depended on the position according to a V-shaped curve (Figures 3, 4). This maximal value was the lowest near centromeres and the highest near telomeres. This global trend was particularly strong in the CF-Dent panel and less pronounced in the C-K panel, for which the maximum power was stable for larger segments. The V-shaped curve also had different local trends for the different chromosomes for a given panel. For instance in the CF-Flint panel, depletion in power in centromeric region was longer for chromosome 7 than for chromosome 6 (Figure 3C).

Table 1: Average and standard deviation of analytical power and of the parameters related to power. Analytical power of model $\mathcal{M}_{K\_Freq}$ was estimated in each panel (reduced to a size of 267 individuals), assuming an heritability of 0.8, a marker effect that would explain 8% of the background genetic variance if it had a Minor Allele Frequency (MAF) of 0.5, and a type I risk of 0.05 with a Bonferroni correction on 40000 tests.

	Power ($\mathcal{M}_{K\_Freq}$)		*CorK_Freq*[a]		*CorK_Chr*[b]		MAF[c]		*Fst*[d]	
	Average	SD	Average	SD	Average	SD	Average	SD	Average	SD
C-K	0.113	0.090	0.087	0.032	0.083	0.029	0.269	0.132	0.112	0.116
CF-Dent	0.090	0.081	0.103	0.033	0.093	0.030	0.260	0.139	0.146	0.118
CF-Flint	0.088	0.086	0.094	0.032	0.088	0.030	0.240	0.147	0.083	0.076

[a]Correlation between the kinship matrix estimated with a single marker (***K_Freq_M**$_l$*) and the kinship matrix estimated with all the PANZEA markers (***K_Freq***). [b]Correlation between the kinship matrix estimated with a single marker (***K_Freq_M**$_l$*) and the kinship matrix estimated with all the PANZEA markers but those located on the same chromosome.[c] Minor Allele Frequency. [d] Nei's differentiation index among genetic groups.

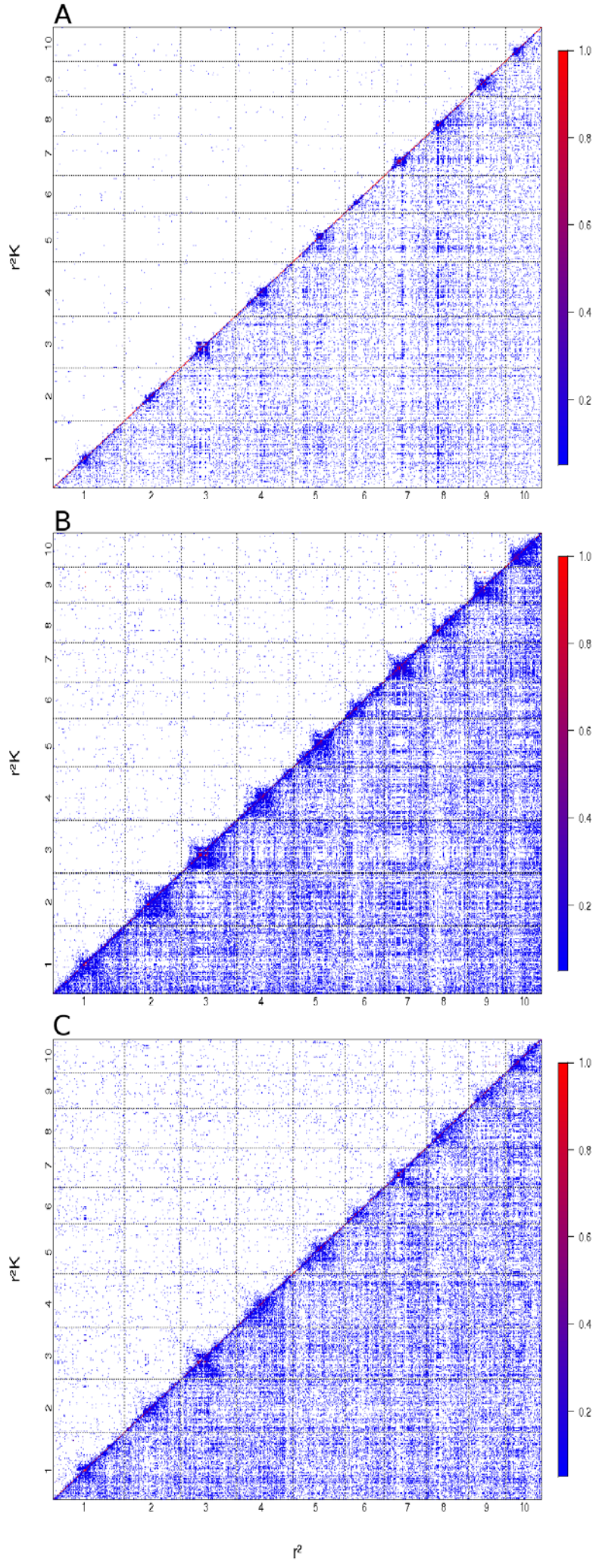

Figure1: Linkage Disequilibrium in the C-K (A), CF-Dent (B) and CF-Flint (C) panels estimated with 4000 markers sampled according to their physical position. Raw squared correlations (r^2) are represented below the diagonal, and r^2 corrected by kinship (r^2K) estimated as ***K_Freq*** are presented above the diagonal. Cells corresponding to LD below 0.05 are in white. Markers were ordered according to their physical position.

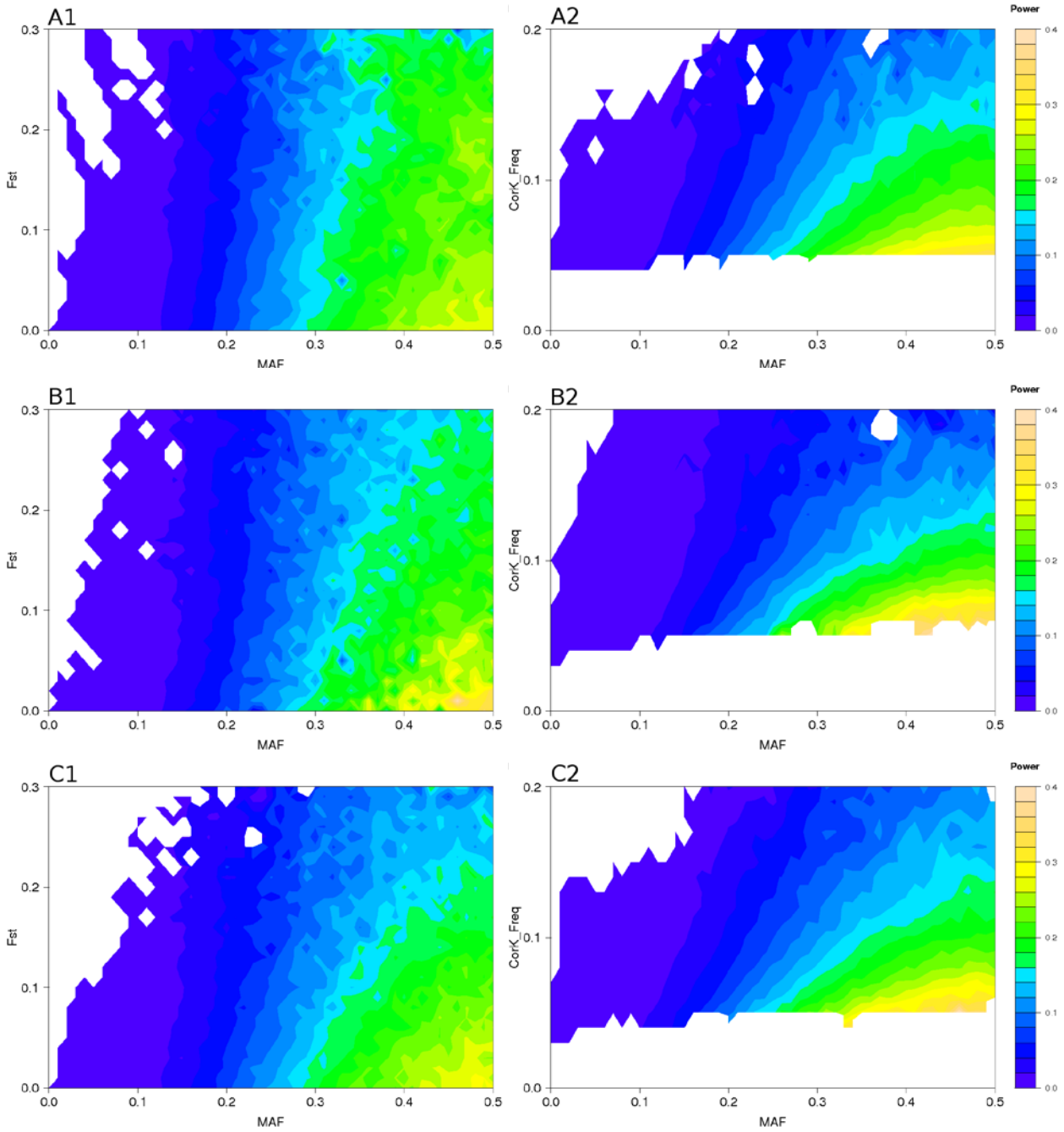

Figure 2: Level plots of power of model $\mathcal{M}_{K\_Freq}$ in the C-K (A), CF-Dent (B), and CF-Flint (C) panels. Each color corresponds to a range of power described by the right hand side scale. x axis corresponds to the MAF. y axis is the *Fst* (A1, B1, C1) or the correlation between the kinship matrix estimated with the considered marker only and the kinship matrix estimated with all the PANZEA markers (A2, B2, C2).

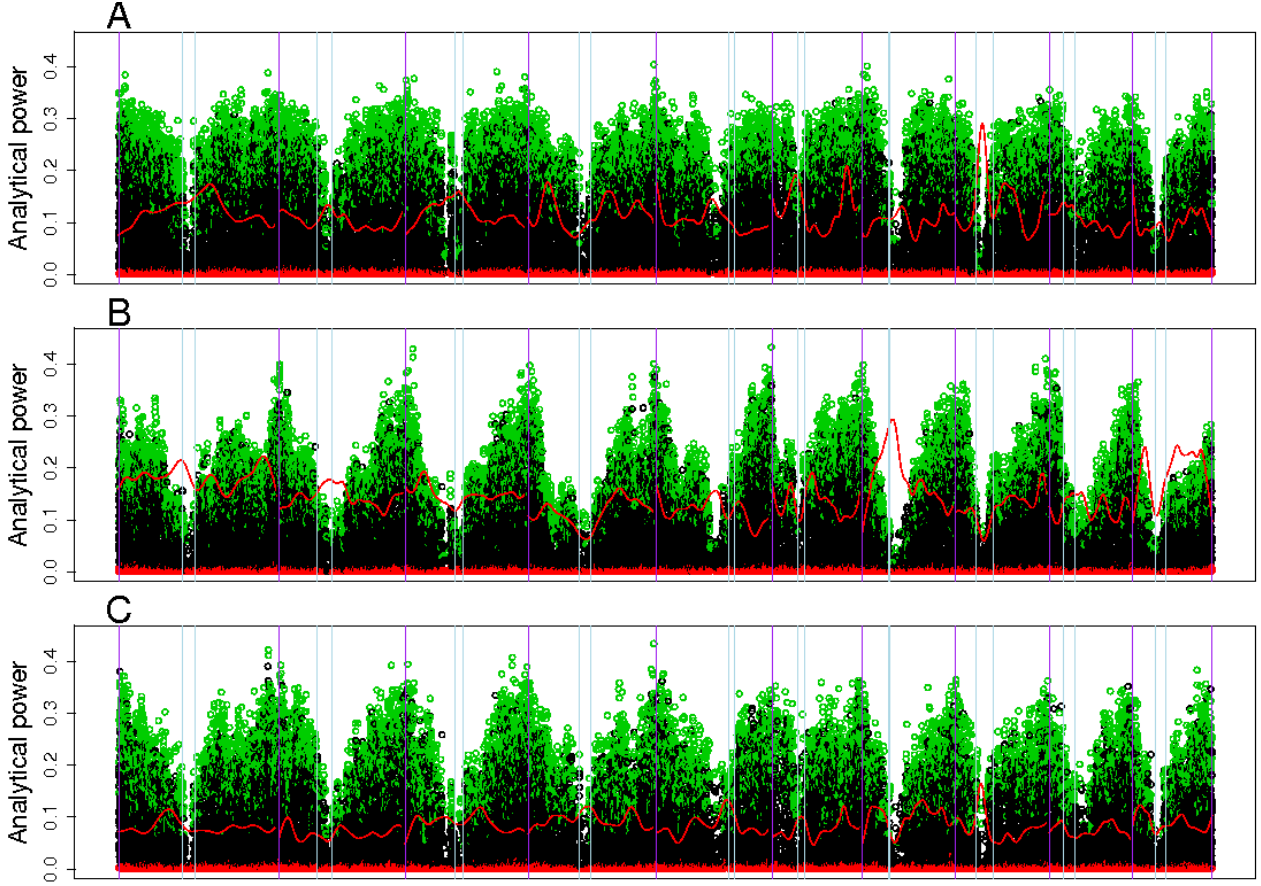

Figure 3: Power scan of statistical model $\mathcal{M}_{K\_Freq}$ in the C-K (A), CF-Dent (B) and CF-Flint (C) panels. Power at each marker is plotted against its physical position. Markers with a MAF above 0.4 and below 0.1 are represented by green and red dots, respectively. Red curve displays local *Fst*. Purple and light blue vertical lines indicate the chromosome and the centromere limits, respectively.

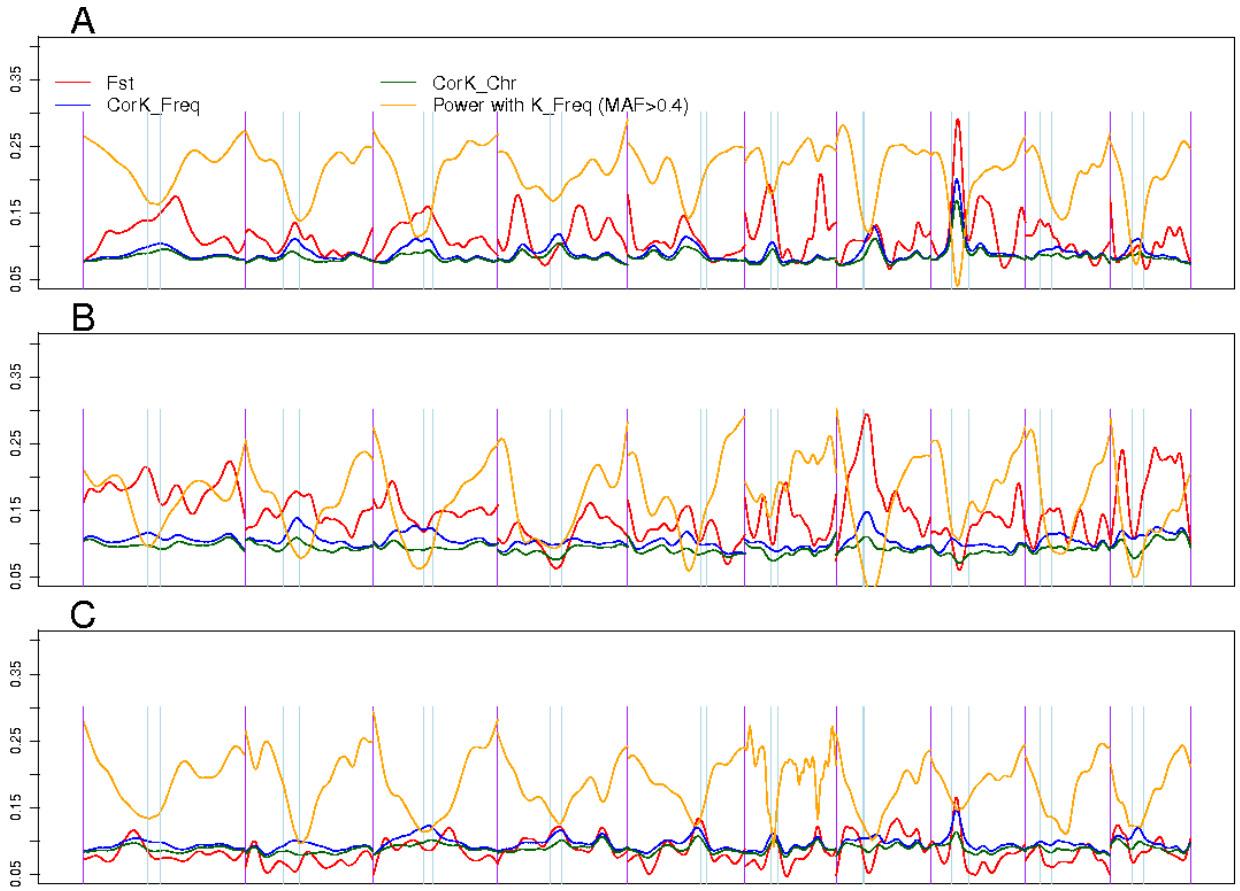

Figure 4: Scan of parameters related to power along the genome in C-K (A), CF-Dent (B), and CF-Flint (C) panels. *Fst* is Nei's index of differentiation among genetic groups. *CorK_Freq* is the correlation between the kinship matrix estimated with the considered marker only (***K_Freq_M$_l$***) and the kinship matrix estimated with all the PANZEA markers (***K_Freq***). *CorK_Chr* is the correlation between the kinship matrix estimated with the considered marker only (***K_Freq_M$_l$***) and the kinship matrix estimated with all the PANZEA markers but those located on the same chromosome than the tested SNP (***K_Chr***). For each parameter a smoothing spline was used along the genome. The orange curve was adjusted to the analytical power at markers with a MAF above 0.4.

Power of model $\mathcal{M}_{K\_Freq}$ was in accordance with trends of *CorK_Freq* along the genome. Correlation between the covariance matrix at the marker and the global covariance matrix (***K_Freq*** and ***K_Chr***) was significantly lower for ***K_Chr*** than for ***K_Freq***, and particularly in the pericentromeric regions (Figure4). We observed that peaks of *Fst* corresponded generally to peaks of both correlations (*CorK_Freq* and *CorK_Chr*) (Figure4B, chromosome 7, and Figures 4A and 4C chromosome 8). Conversely, pericentromeric regions with low *Fst* corresponded to a peak of *CorK_Freq* and a drop of *CorK_Chr* (Figure4B, chromosomes 8 and 10, and Figure 4C chromosome 7). *CorK_Freq*, *CorK_Chr* and the difference between these two parameters were higher in the CF-Dent panel than in the two others.

Simulation based assessment of kinship estimation on false positive control and power

Simulating different genetic models using the genotypes of the three panels allowed the comparison of the efficiency of the three statistical models to control false positives and to detect QTLs. The efficiency to control false positives depended on the genetic model (number of QTLs), the panel, and the estimation procedure for K (Table 2). The distribution of the Pvalues under H0 revealed that $\mathcal{M}_{K\_Freq}$ was conservative (Figure 5A) whereas the alternative models $\mathcal{M}_{K\_Chr}$ and $\mathcal{M}_{K\_LD}$ gave distributions closer to the expected one (Figures 5B and 5C). The observed Pvalue quantiles were closer to the expected Pvalue quantiles with $\mathcal{M}_{K\_Chr}$ and $\mathcal{M}_{K\_LD}$ than with $\mathcal{M}_{K\_Freq}$ (Table 2). $\mathcal{M}_{K\_Freq}$ resulted in fewer small Pvalues than expected under H0, for example in the CF-Dent panel we observed only half of the Pvalues that were expected to be below 0.001. Observed Pvalue quantiles with $\mathcal{M}_{K\_Chr}$ and $\mathcal{M}_{K\_LD}$ were very close to the expected Pvalue quantiles, although also most of the time below it.

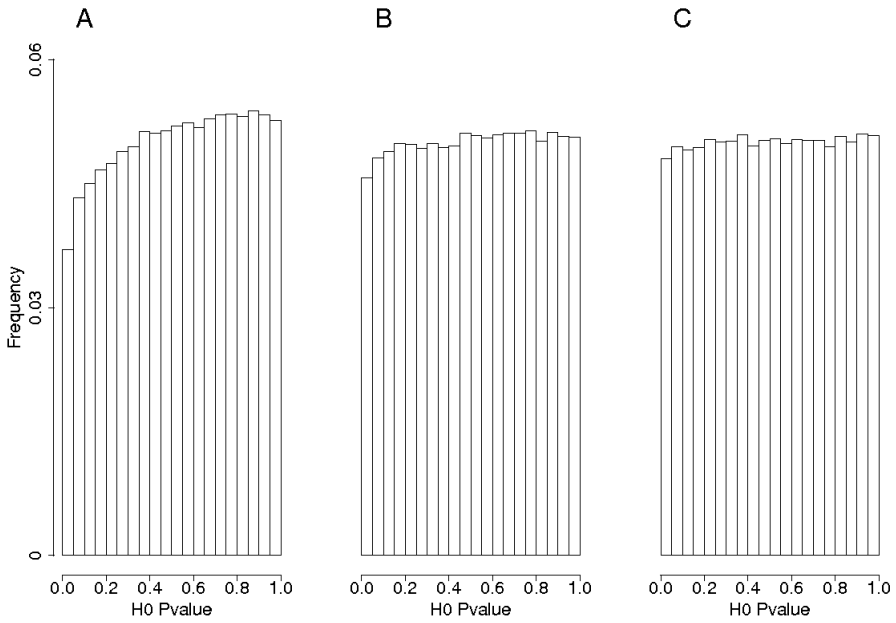

Figure 5: Histograms of Pvalues of the markers on the H0-chromosome using $\mathcal{M}_{K\_Freq}$(A), $\mathcal{M}_{K\_Chr}$(B), and $\mathcal{M}_{K\_LD}$(C). This was obtained when simulating 100 QTLs in the CF-Dent panel.

Table 2: Quantiles of the Pvalues under H0 in each panel with the three statistical models and considering two different genetic models (50 or 100 QTLs). We estimated the average and the standard deviation of the 0.001 and 0.01 quantiles over the 100 runs of simulation.

			1 ‰ quantile		1% quantile	
Panel	Nb QTLs	Approach	Average(‰)	SD(‰)	Average(%)	SD(%)
C-K	50	M_{K_Freq}	0.8	0.7	0.81	0.28
	50	M_{K_Chr}	1.0	0.9	0.98	0.35
	50	M_{K_LD}	0.9	0.9	0.94	0.32
	100	M_{K_Freq}	0.8	0.6	0.89	0.25
	100	M_{K_Chr}	1.1	0.8	1.08	0.34
	100	M_{K_LD}	1.0	0.8	1.07	0.34
CF-Dent	50	M_{K_Freq}	0.5	0.5	0.61	0.23
	50	M_{K_Chr}	0.8	1.0	0.94	0.53
	50	M_{K_LD}	0.7	0.8	0.85	0.34
	100	M_{K_Freq}	0.4	0.5	0.63	0.30
	100	M_{K_Chr}	0.8	0.9	1.02	0.65
	100	M_{K_LD}	0.9	1.6	0.94	0.42
CF-Flint	50	M_{K_Freq}	0.6	0.7	0.74	0.25
	50	M_{K_Chr}	0.9	0.9	0.99	0.39
	50	M_{K_LD}	1.2	1.1	1.09	0.47
	100	M_{K_Freq}	0.5	0.5	0.73	0.25
	100	M_{K_Chr}	0.7	0.6	0.92	0.39
	100	M_{K_LD}	1.0	0.7	1.06	0.39

The second step of the simulations revealed the ability of the different statistical models to detect QTLs in the different panels. With the usual Bonferroni correction, only few QTLs were detected (Table 3). In each scenario $\mathcal{M}_{K\_Chr}$ and $\mathcal{M}_{K\_LD}$ were more powerful than $\mathcal{M}_{K\_Freq}$. For example, they respectively permitted the detection of 2.1, 1.3 and 1.2 QTL (SNP considered as QTL) on average in the CF-Dent panel when 50 QTLs were segregating. The difference of power (proportion of SNP considered as QTL detected) between the different models was more important for less stringent significance threshold. The difference of power between $\mathcal{M}_{K\_Chr}$ and $\mathcal{M}_{K\_Freq}$ was the highest in the CF-Dent panel. More QTLs were found in the scenario with 50 QTLs than in the scenario with 100 QTLs. This was expected, QTLs having a lower effect on the trait in the 100 than in the 50-QTLs scenario.

Table 3: Number of QTLs detected with the three statistical models in each panel at different thresholds assuming different genetic models (50 or 100 QTLs). We computed the average and the standard deviation of the number of QTLs detected in the 100 runs of simulation.

			T[a]		10*T		100*T	
Panel	Nb QTLs	Approach	Average	SD	Average	SD	Average	SD
C-K	50	$M_{K\ Freq}$	1.4	1.0	2.5	1.2	4.2	1.6
	50	$M_{K\ Chr}$	1.7	1.1	3.2	1.5	4.9	1.7
	50	$M_{K\ LD}$	1.6	1.1	2.6	1.3	4.3	1.7
	100	$M_{K\ Freq}$	0.3	0.5	0.9	0.8	2.1	1.2
	100	$M_{K\ Chr}$	0.5	0.7	1.3	1.0	2.8	1.5
	100	$M_{K\ LD}$	0.4	0.6	1.1	0.9	2.3	1.4
CF-Dent	50	$M_{K\ Freq}$	1.2	1.0	2.2	1.3	3.6	1.3
	50	$M_{K\ Chr}$	2.1	1.4	3.4	1.5	5.3	1.6
	50	$M_{K\ LD}$	1.3	1.1	2.5	1.3	4.1	1.4
	100	$M_{K\ Freq}$	0.3	0.6	0.9	0.9	2.0	1.4
	100	$M_{K\ Chr}$	0.8	1.0	1.7	1.3	3.4	1.7
	100	$M_{K\ LD}$	0.5	0.7	1.0	1.1	2.4	1.4
CF-Flint	50	$M_{K\ Freq}$	1.4	1.0	2.4	1.1	3.7	1.2
	50	$M_{K\ Chr}$	1.8	1.2	3.0	1.0	4.5	1.3
	50	$M_{K\ LD}$	1.4	0.9	2.4	1.1	4.0	1.3
	100	$M_{K\ Freq}$	0.3	0.6	0.8	0.9	1.9	1.1
	100	$M_{K\ Chr}$	0.6	0.8	1.4	1.2	2.8	1.4
	100	$M_{K\ LD}$	0.4	0.7	1.0	1.1	2.1	1.3

[a] Significance threshold T was set considering a type I risk of 5% with a Bonferroni correction assuming 40 000 tests.

DISCUSSION AND CONCLUSIONS

Analytical investigation of potential power along the genome with usual model ($\mathcal{M}_{K\_Freq}$)

Power is a key parameter in association mapping, because it indicates how likely the discovery of a QTL is. We presented a general method based on non centrality parameter to derive analytically theoretical power at each marker locus in a given panel of individuals. It was applied to three different association mapping panels. While being adjusted to the same population size, these different panels had different average power. They also displayed different local patterns of power along the genome.

Power could be related to three parameters characterizing each marker: its MAF, its differentiation index among genetic groups (*Fst*), and the correlation between its individual kinship matrix with that estimated with all the markers (*CorK_Freq* when ***K_Freq*** is considered). Power at a marker with a low MAF is limited, even if this marker is orthogonal to structure and kinship (Figures 2, 3). This effect was highlighted already for linkage studies (SOLLER *et al.* 1976 and CHARCOSSET and GALLAIS 1996) and GWAS (LONSDALE *et al.* 2013) and can be explained by the fact that when one of the two alleles is rare, the marker cannot contribute much to the genetic variation. The correlation between kinship at single markers and the global kinship had a strong negative effect on power (Figure 2). The *Fst* among admixture group also had an important effect on local power (Figures 2, 3). This confirmed that admixture is reflected by the kinship matrix, because differentiated regions had a low power although we used a model with relatedness but no admixture ($\mathcal{M}_{K\_Freq}$). The level plots showing analytical power at different MAF and *CorK_Freq* were very similar in the three panels (Figure 2 A2, B2, C2), but those showing power at different MAF and *Fst* differed (Figure 2 A1, B1 and C1). This suggests that group differentiation has different relative contribution to local kinship variation in the different panels. At a given pair of MAF and *Fst* value, power was lower in the CF-Dent and CF-Flint panels than in the C-K panel, whereas five groups were used in this panel instead of four in the two others. The C-K panel is composed of highly diverse groups (Tropical, Dent and Flint lines) and so the admixture matrix captured ancestral population structure but only a small part of kinship. On the opposite, the CF-Dent and CF-Flint panels are composed of less heterogeneous material and so the admixture matrix captured more relatedness. Finally, shape of the level plots (Figure 2) also suggested that the effect of the different parameters affecting power were not additive. For example *Fst* and *CorK_Freq* had a stronger effect on power for markers with

higher MAF, and MAF had a stronger effect on power for less differentiated markers. These results show that controlling false positives using the ***K_Freq*** model also implies reducing power at differentiated markers (LARSSON *et al.* 2013). It is interesting to note that no marker had a *CorK_Freq* below 0.03 (Figure 2). To investigate the maximum power that could be reached theoretically, we generated for each panel a vector of zeros and ones simulating a marker genotype and applied a simple exchange algorithm until analytical power reached a maximum. These virtual markers (one for each panel) had analytical power much higher (above 0.8) than the maximal analytical power of the existing SNPs (below 0.44 in each panel). They had a MAF of 0.5 and a *CorK_Freq* value below 0.017. This difference illustrates that the maximum power is strongly constrained by the evolution process that led to the panels.

Both *Fst* and *CorK_Freq* appeared highly variable along the genome in each panel. High differentiation (*Fst*) was observed in particular in pericentromeric regions (Figures 3A and 3C, chromosome 8 and Figure 3B, chromosome 7). Pericentromeric regions are known to be more structured than telomeric regions (CARNEIRO *et al.* 2009; FRANCHINI *et al.* 2010) because of lower recombination rates. *CorK_Freq* was also higher in regions of high LD (mostly pericentromeric regions, see Figures 1 and 4). Beyond the effect of group differentiation, markers in regions of high LD are indeed correlated to many other SNPs that all contribute to the estimation of ***K_Freq***. These LD and *Fst* features led to the observed V-shape analytical power curve along the chromosome, particularly in the CF-Dent panel in which LD was more extended (Figures 1, 3). This is in good agreement with published manhattan plots of GWAS results which showed a reduced number of low Pvalues in the centromeric regions (BOUCHET *et al*. 2013; LARSSON *et al.* 2013). In our three panels, we observed that this problem also arose with other classical estimators of relatedness (results not shown) such as the IBS estimator or the first estimator provided on page 4416 in VANRADEN (2008).

As MAF, *Fst*, LD extent, and consequently *CorK_Freq* were different in the three panels (Table 1), average power was highly variable among the three panels (adjusted for the same population size). Among the three diversity panels, the C-K panel appeared to be the most powerful on average due to its higher MAF, lesser LD extent and its lower relatedness. It should be noted that this analytical study assumed that the variance components were known. It was therefore necessary to confirm these results with simulations.

Simulation based comparison of type I risk and power of statistical models associated with different estimations of *K*

Removing the markers on the same chromosome than the tested one ($\mathcal{M}_{K\_Chr}$) permitted to decrease the correlation between the kinship at the tested SNP and the global covariance (*CorK_Chr* in Figure 4). *CorK_Chr* remained nevertheless high in structured regions (high *Fst*), i.e. regions with important differentiation between genetic groups (Figures 4A and 4C, chromosome 8), which suggests that ***K_Chr*** was efficient to estimate covariance between individuals.

To evaluate models involving different kinship estimators for their ability (i) to control false positives at nominal levels and (ii) to detect QTLs, we conducted simulations based on the genotypes of the diversity panels. Using all the markers to estimate kinship matrix ($\mathcal{M}_{K\_Freq}$) led to an over-correction of the H0-Pvalues (Table 2, Figure 5). This was particularly the case in the panel with the highest level of LD (CF-Dent). Under H0, the Pvalue distributions of the two alternative models were much closer to the expected distribution, and revealed that these approaches were also efficient to control false positives (Figure 5). Results obtained with $\mathcal{M}_{K\_Chr}$ showed that molecular information carried by nine of the ten chromosomes was sufficient to reliably estimate covariance between individuals to control for false positives.

Knowing that the three estimations of the kinship matrix (***K_Freq, K_Chr*** and ***K_LD***) were efficient to control false positives, we could compare their power in a second step of simulations. QTLs were sampled from the ten chromosomes, and power of $\mathcal{M}_{K\_Freq}$, $\mathcal{M}_{K\_Chr}$ and $\mathcal{M}_{K_LD}$at different threshold was evaluated at the SNPs/QTLs. The alternative models were more powerful than the usual model $\mathcal{M}_{K\_Freq}$ (Table 3). In particular estimating the covariance matrix using the markers on the non tested chromosome ($\mathcal{M}_{K\_Chr}$) resulted in higher power in each scenario in each panel. As expected the gain of power was higher in the panel with more extended LD (CF-Dent). The gain of power was lower with $\mathcal{M}_{K\_LD}$, but we suppose that this approach could be improved by taking into account gene density along the genome, or a priori information on genetic architecture, and by using a better estimate of the covariance between the marker based estimators when computing optimal marker weights. Note that further research on the ***K_LD*** estimator should also consider its scalability when applied to very high dimensional datasets.

To check the stability of these results, when considering that a QTL could be detected by SNPs located near it, we used another simulation approach, in which SNPs within a genetic window around the QTL positions were considered as H1-markers and the others as H0-markers. The results (Table S1) confirmed that at a given realized FDR, the alternative models and in particular $\mathcal{M}_{K\_Chr}$ were more powerful than the traditional model ($\mathcal{M}_{K\_Freq}$). Considering that true discoveries were within 5 cM of the QTLs, $\mathcal{M}_{K\_Freq}$ had a power to detect QTLs of 11%, $\mathcal{M}_{K\_Chr}$ of 26% and $\mathcal{M}_{K\_LD}$ of 19% at a realized FDR of 10%, when 100 QTLs were simulated in the CF-Dent panel.

In conclusion, the derivation of analytical power permitted to highlight which parameters are linked to power in Association Mapping. In particular the kinship between individuals (***K***) clearly influenced the Non Centrality Parameter. Analytical power scan in three diversity panels also confirmed that the way of estimating ***K*** can affect power. In particular, usual model ($\mathcal{M}_{K\_Freq}$) has a low power in regions of high LD. We proposed two alternative approaches to recover this gap of power, and we could show with simulations based on true genotypes that they were more powerful at given type I risks than usual models.

Acknowledgments

This research was jointly supported as “Cornfed project” by the French National Agency for Research (ANR), the German Federal Ministry of Education and Research (BMBF) and the Spanish ministry of Science and Innovation (MICINN, research project EUI2008-3635). R. Rincent is jointly funded by Limagrain, Biogemma, KWS and the French ANRt. L. Moreau, T. Mary-Huard and A. Charcosset conducted this research in the framework of Amaizing Investissement d'Avenir program. The authors thank the reviewers and the editor for their comments which improved the manuscript.

LITERATURE

ALEXANDER D. H., J. NOVEMBRE, and K. LANGE, 2009 Fast model-based estimation of ancestry in unrelated individuals. Genome Res. **19**: 1655–1664.

ASTLE W. andD. J.BALDING, 2009 Population Structure and Cryptic Relatedness in Genetic Association Studies.Stat. Sci. **24**: 451–471.

BELO A., P. ZHENG, S. LUCK, B.SHEN, D. J.MEYER *et al.,* 2007 Whole genome scan detects an allelic variant of fad2 associated with increased oleic acid levels in maize. Mol. Genet. Genomics **279**: 1–10.

BERNARDO R., 2013 Genomewide Markers for Controlling Background Variation in Association Mapping. The Plant Genome **6**: 1-9.

BOUCHET S., B. SERVIN, P. BERTIN, D. MADUR, V. COMBES *et al.,* 2013 Adaptation of Maize to Temperate Climates: Mid-Density Genome-Wide Association Genetics and Diversity Patterns Reveal Key Genomic Regions, with a Major Contribution of the Vgt2 (ZCN8) Locus. Plos One **8**: e71377.

BRADBURY P., T. PARKER, M. T. HAMBLIN and J.-L.JANNINK, 2011 Assessment of Power and False Discovery Rate in Genome-Wide Association Studies using the BarleyCAP Germplasm. Crop Sci.**51**: 52.

BROWNING B. L. and S. R.BROWNING, 2009 A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. Am. J. Hum. Genet.**84**: 210–223.

CAMUS-KULANDAIVELU L., J.-B.VEYRIERAS,D.MADUR, V. COMBES, M. FOURMANN *et al.,* 2006 Maize Adaptation to Temperate Climate: Relationship Between Population Structure and Polymorphism in the Dwarf8 Gene. Genetics **172**: 2449–2463.

CARNEIRO M., N. FERRAND, and M. W. NACHMAN, 2009 Recombination and Speciation: Loci Near Centromeres Are More Differentiated Than Loci Near Telomeres Between Subspecies of the European Rabbit (Oryctolagus cuniculus). Genetics **181**: 593–606.

CHARCOSSET A. and A.GALLAIS, 1996 Estimation of the contribution of quantitative trait loci (QTL) to the variance of a quantitative trait by means of genetic markers. Theoret. Appl.

Genet.**93**: 1193–1201.

ERBE M., F. YTOURNEL, E. C. G. PIMENTEL, A. R. SHARIFI AND H. SIMIANER, 2010 Comparison of three whole genome association mapping approaches in selected populations. Zuchtungskunde **82**: 77-97

EWENS W. and R.SPIELMAN, 1995 The Transmission Disequilibrium Test - History, Subdivision and Admixture. Am. J. Hum. Genet.**57**: 455–464.

FLINT-GARCIA, S. A., J. M. THORNSBERRY and E. S. BUCKLER, 2003 Structure of linkage disequilibrium in plants.Annu. Rev. Plant Biol.**54:** 357–374.

FRANCHINI P., P. COLANGELO, E. SOLANO, E. CAPANNA, E. VERHEYEN *et al.*, 2010 Reduced Gene Flow at Pericentromeric Loci in a Hybrid Zone Involving Chromosomal Races of the House Mouse Mus Musculus Domesticus. Evolution **64**: 2020–2032.

GANAL M. W., G. DURSTEWITZ, A. POLLEY, A. BÉRARD, E. S. BUCKLER *et al.*, 2011 A Large Maize (Zea mays L.) SNP Genotyping Array: Development and Germplasm Genotyping, and Genetic Mapping to Compare with the B73 Reference Genome (L Lukens, Ed.). PLoS ONE **6**: e28334.

GILMOUR A.R., B.GOGEL, B.R.CULLIS, and R.THOMPSON, 2009 ASREML user guide release 3.0.VSN International Ltd., Hemel Hempstead, UK

GOLDFARBD. and A. IDNANI, 1983 A numerically stable dual method for solving strictly convex quadratic programs. Math.Prog.**27**: 1–33.

GORE M. A., J.-M. CHIA, R. J. ELSHIRE, Q. SUN, E. S. ERSOZ *et al.*, 2009 A First-Generation Haplotype Map of Maize. Science **326**: 1115–1117.

GOUDET J., 2005 hierfstat, a package for r to compute and test hierarchical F-statistics. Mol. Ecol. Notes **5**: 184–186.

HASTIE, T. J. and R. J. TIBSHIRANI, 1990 Generalized Additive Models. Chapman and Hal.

HUANG X., X. WEI, T. SANG, Q. ZHAO, Q. FENG *et al.*, 2010 Genome-wide association studies of 14 agronomic traits in rice landraces. Nat. Genet.**42**: 961–967.

JANNINK, J. L., and B. WALSH, 2002 Association mapping in plant populations, pp. 59–68 in

Quantitative Genetics, Genomics and Plant Breeding, edited by M. S. Kang. CAB International, New York.

JONES P., K. CHASE, A. MARTIN, P. DAVERN, E. A.OSTRANDER *et al.*, 2008 Single-Nucleotide-Polymorphism-Based Association Mapping of Dog Stereotypes. Genetics **179**: 1033–1044.

KANG H. M., J. H. SUL, S. K. SERVICE, N. A. ZAITLEN, S.E.A. KONG *et al.*, 2010a Variance component model to account for sample structure in genome-wide association studies. Nat. Genet.**42**: 348–354.

KANG H. M., J. H. SUL, S. K. SERVICE, N. A. ZAITLEN, S. KONG *et al.*, 2010b Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. **42**: 348–354.

KNAPP S. J., and W. C.BRIDGES, 1990 Using molecular markers to estimate quantitative trait locus parameters: power and genetic variances for unreplicated and replicated progeny. Genetics **126**: 769–777.

LANDE, R., and R. THOMPSON, 1990 Efficiency of marker assisted selection in the improvement of quantitative traits. Genetics **124**: 743–756.

LARSSON S. J., A. E. LIPKA, and E. S. BUCKLER, 2013 Lessons from Dwarf8 on the Strengths and Weaknesses of Structured Association Mapping (JK Pritchard, Ed.). PLoS Genet.**9**: e1003246.

LISTGARTEN J., C. LIPPERT, C. M. KADIE, R. I. DAVIDSON, E. ESKIN *et al.*, 2012 Improved Linear Mixed Models for Genome-Wide Association Studies.Nat. Meth. **9**: 525–526.

LONSDALE J., J. THOMAS, M. SALVATORE, R. PHILLIPS, E. LO *et al.*, 2013 The Genotype-Tissue Expression (GTEx) project. Nat. Genet.**45**: 580–585.

LYNCH M., and B. WALSH, 1998 Genetics and Analysis of Quantitative Traits. Sinauer Associates, Sunderland, MA.

MACLEOD I. M., B. J. HAYES, K. W. SAVIN, A. J. CHAMBERLAIN, H. C. MCPARTLAN, M. E. GODDARD 2010 Power of a genome scan to detect and locate quantitative trait loci in cattle using dense single nucleotide polymorphisms. J. Anim. Breed.Genet.**127**:133-142.

MANGIN B., A. SIBERCHICOT, S. NICOLAS, A. DOLIGEZ, P. THIS *et al.*, 2012 Novel measures of

linkage disequilibrium that correct the bias due to population structure and relatedness. Heredity **108**: 285–291.

NEI M., 1973 Analysis of gene diversity in subdivided populations. PNAS **70**: 3321–3323.

OZAKI K., Y. OHNISHI, A. IIDA, A. SEKINE, R. YAMADA *et al.*, 2002 Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. Nat. Genet.**32**: 650–654.

PRICE A. L., N. J. PATTERSON, R. M. PLENGE, M. E. WEINBLATT, N.A. SHADICK *et al.*, 2006 Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet.**38**: 904–909.

PRICE, A. L., N. A. ZAITLEN, D. REICH, and N. PATTERSON, 2010 New approaches to population stratification in genome-wide association studies. Nat. Rev. Genet. **11**: 459–463.

PRITCHARD J. K., M.STEPHENS, and P. DONNELLY, 2000 Inference of population structure using multilocus genotype data. Genetics **155**: 945.

PURCELL S., S. S. CHERNY, and P. C. SHAM, 2003 Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics **19**: 149–150.

R development Core Team 2006 R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

RAFALSKI A. and M.MORGANTE, 2004 Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. Trends Genet.**20**: 103–111.

REBAI A. and B.GOFFINET, 1993 Power of tests for QTL detection using replicated progenies derived from a diallel cross. Theor. Appl. Genet.**86**: 1014–1022.

RINCENT R., D. LALOE, S. NICOLAS, T. ALTMANN, D. BRUNEL *et al.*, 2012 Maximizing the Reliability of Genomic Selection by Optimizing the Calibration Set of Reference Individuals: Comparison of Methods in Two Diverse Groups of Maize Inbreds (Zea mays L.). Genetics **192**: 715–728.

ROMAY M. C., M. J. MILLARD, J. C. GLAUBITZ, J. A. PEIFFER, K. L. SWARTS *et al.*, 2013 Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biology **14**:R55

SHAM P. C., S. S. CHERNY, S. PURCELL and J.K.HEWITT, 2000 Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data.Am. J. Hum. Genet **66**: 1616.

SOLLER M., A. GENIZI and T. BRODY, 1976 On the power of experimental designs for the detection of linkage between marker lociand quantitative loci in crosses between inbred lines. Theor. Appl. Genet.**47**:35–39.

SPEED D., HEMANI G., JOHNSON M. R., BALDING D. J., 2012 Improved Heritability Estimation from Genome-wide SNPs. Am. J. Hum. Genet.**91**: 1011–1021.

STICH B. and A. MELCHINGER, 2009 Comparison of mixed-model approaches for association mapping in rapeseed, potato, sugar beet, maize, and Arabidopsis. BMC Genomics **10**: 94.

STICH B., J. MOHRING, H.-P. PIEPHO, M.HECKENBERGER, E. S. BUCKLER *et al.*, 2008 Comparison of Mixed-Model Approaches for Association Mapping. Genetics **178**: 1745–1754.

TEYSSÈDRE S., J.-M.ELSEN, and A.RICARD, 2012 Statistical distributions of test statistics used for quantitative trait association mapping in structured populations.Genet. Sel. Evol.**44**: 32.

THORNSBERRY J. M., M. M. GOODMAN, J. DOEBLEY, S. KRESOVICH, D. NIELSEN *et al.*, 2001 Dwarf8 polymorphisms associate with variation in flowering time. Nat. Genet. **28**: 286–289.

VANRADEN P., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci.**91**: 4414–4423.

WALD A., 1943 Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large. Trans. Amer. Math. Soc.**54**: 426-482.

WANG K., 2008 An Analytic Study of the Power of Popular Quantitative-Trait-Locus Mapping Methods. Behav.Genet.**38**: 554–559.

YU J., G. PRESSOIR, W. H. BRIGGS, I. VROH BI, M. YAMASAKI *et al.*, 2006 A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. **38**: 203–208.

ZHANG Z., ERSOZ E., LAI C.-Q., TODHUNTER R. J., TIWARI H. K., GORE M. A., BRADBURY P. J., YU J., ARNETT D. K., ORDOVAS J. M., BUCKLER E. S., 2010 Mixed linear model approach adapted for genome-wide association studies. Nat. Genet. **42**: 355–360.

ZHAO K., M. J. ARANZANA, S. KIM, C. LISTER, C. SHINDO *et al.*, 2007a An Arabidopsis Example of Association Mapping in Structured Samples. PLoS Genet.**3**: e4.

ZHAO H. H., R. L. FERNANDO andJ.C.M. DEKKERS, 2007b Power and Precision of Alternate Methods for Linkage Disequilibrium Mapping of Quantitative Trait Loci. Genetics **175**: 1975–1986.

Chapter 2

Dent and Flint maize diversity panels reveal important genetic potential for increasing biomass production

R. Rincent,[1,2,3,4], S. Nicolas,[1] S. Bouchet,[1]T. Altmann,[6] D. Brunel,[7] P. Revilla,[8] V.M. Rodríguez,[8] J. Moreno-Gonzalez,[9] A. E. Melchinger,[10] E. Bauer,[11] C-C. Schoen,[11] N. Meyer,[2] C. Giauffret,[12] C. Bauland,[1] P. Jamin,[1] J. Laborde,[14] P. Flament,[4] L. Moreau[1], A. Charcosset,[1,14]

[1]UMR de Génétique Végétale, INRA – Université Paris-Sud – CNRS, 91190 Gif-sur-Yvette, France,
[2]BIOGEMMA, Genetics and Genomics in Cereals, 63720 Chappes, France,
[3]KWS Saat AG, Grimsehlstr 31, 37555 Einbeck, Germany,
[4]Limagrain, site d'ULICE, av G. Gershwin, BP173, 63204 Riom Cedex, France,
[5]AgroParisTech, Laboratoire de Génétique Animale et Biologie Intégrative, Domaine de Vilvert, 78352 Jouy-en-Josas, France,
[6]Max-Planck Institute for Molecular Plant Physiology, 14476 Potsdam-Golm, Germany, Leibniz-Institute of Plant Genetics and Crop Plant Research (IPK), 06466 Gatersleben, Germany,
[7]INRA, UR 1279 Etude du Polymorphisme des Génomes Végétaux, CEA Institut de Génomique, Centre National de Génotypage, 2, rue Gaston Crémieux, CP5724, 91057 Evry, France,[1]
[8]Misión Biológica de Galicia, Spanish National Research Council (CSIC).Apartado 28, 36080 Pontevedra, Spain,
[9]Centro de Investigaciones Agrarias de Mabegondo. Apartado 10, 15080 La Coruna, Spain
[10]Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, Fruwirthstr.21, 70599, Stuttgart, Germany,
[11]Department of Plant Breeding, Technische Universität München, 85354 Freising, Germany,
[12]INRA/Université des Sciences et Technologies de Lille, UMR1281, Stress Abiotiques et Différenciation des Végétaux Cultivés, Estrées-Mons, B.P. 136, 80203 Péronne Cedex, France,
[13]INRA Stn Expt Mais, 40590 St Martin De Hinx, France,

Short running title: Genetics of biomass in Dent and Flint panels
Keywords: biomass, association mapping, flowering time, power, *Zea mays* L.

[14]**Corresponding author:** Alain Charcosset
UMR de Génétique Végétale
INRA - Univ Paris-Sud - CNRS - AgroParisTech
Ferme du Moulon,
F-91190, Gif-sur-Yvette, France
Tel: +33 1 69 33 23 35
Fax: +33 1 69 33 23 40
E-mail: alain.charcosset@moulon.inra.fr

ABSTRACT

The high whole plant biomass productivity of maize makes it a potential source of energy in animal feeding and biofuel production. The variability and the genetic determinism of traits related to biomass are poorly known. We analyzed two highly diverse panels of Dent and Flint lines representing diverse complementary heterotic groups for Northern Europe. They were genotyped with the 50k SNP-array and phenotyped as hybrids (crossed to a tester of the complementary pool) in a western European field trial network for traits related to flowering time, plant height and biomass. The molecular information revealed major trends in recent breeding, conducting to different levels of structure and relatedness in the Dent and Flint panels. This study revealed important potential genetic progress for biomass production, even at constant precocity. Association mapping was run by combining genotypes and phenotypes in a mixed model with a random polygenic effect. This permitted the detection of significant associations, confirming height and flowering time QTLs found in literature. Biomass yield QTLs were detected in both panels but were unstable across the environments. Alternative kinship estimator only based on markers unlinked to the tested SNP increased the number of significant associations by around 40% with a satisfying control of the false positive rate.

INTRODUCTION

Maize is together with wheat and rice one of the three main sources of nutritional energy for humans and is extensively being used in animal feeding, either as grain or whole plant forage. The high efficiency of its C4 metabolism also makes it a resource for biofuel production, as attested by the recent development of BioGas in Germany (HERRMANN and RATH 2012; RATH *et al.* 2013). This worldwide importance is due to the adaptation to various climatic conditions that maize developed following its domestication approximately 8700 years ago in the regions of Mexico (MATSUOKA *et al.* 2002; REBOURG *et al.* 2003). It now has the broadest cultivated range of all crops, from the South of Chile to Canada, from altitudes near sea level to highlands above 3000 m (TENAILLON and CHARCOSSET 2011). In Europe, maize cultivation was adopted on a broad scale rapidly after the discovery of America (REBOURG *et al.* 2003) and a dramatic evolution of varieties occurred with the development of hybrids following World War 2. Dent lines from Northern American origin proved at that time highly complementary to flint lines from European origins to combine productivity and environmental adaptation features for maize cultivation in Northern Europe. These flint x dent hybrid varieties have proven extremely successful for both grain and silage production and the reciprocal selection of the two groups increased their differentiation and complementarity. However, their potential for biomass production remains poorly documented and it is therefore of high interest to investigate the variability of this trait and the underlying genetic determinism within these two groups.

Panels of highly diverse materials have proven most useful to investigate the organization of diversity available for breeding at phenotypic and genotypic levels. They also can lead to the discovery of genes of interest thanks to increasing availability of molecular markers, which now makes it possible to get dense molecular polymorphism information on the whole genome. Genotypic and phenotypic information can indeed be combined to detect QTLs contributing to the variability of traits of interest in genome-wide association studies (GWAS).This strategy was successfully used in many species and resulted in the identification of major genes (OZAKI *et al.* 2002; BELÓ *et al.* 2007; JONES *et al.* 2008).Highly diverse panels have accumulated numerous historical recombination events, leading to a low extent of linkage disequilibrium (LD), which is favorable to finely map QTLs. However, LD in association mapping panels is not only due to genetic linkage, but can also be caused by population structure, relatedness, drift and selection (JANNINK and WALSH 2003; FLINT-GARCIA *et al.* 2003). The contribution of these factors relative to linkage can be evaluated statistically (MANGIN *et al.*2012) and proved for instance to be substantial in grapevine and maize (MANGIN *et al.* 2012; BOUCHET *et al.* 2013). This component of LD due to

population structure and relatedness can generate false positives and has thus to be taken into account in association mapping models to control false positives (EWENS and SPIELMAN 1995; THORNSBERRY *et al.* 2001). Once these effects are correctly modeled, only marker-trait associations due to linkage should be detected. Efficient softwares were developed to infer population structure using genotypic data (PRITCHARD *et al.* 2000; ALEXANDER *et al.* 2009), and several estimators of relatedness between individuals are available (VANRADEN 2008; ASTLE and BALDING 2009; RINCENT *et al*. in press). The estimated admixture (Q) and kinship (K) matrices can be introduced in the GWAS statistical model to control false positive efficiently (YU *et al.* 2006).

The objectives of the present work were (i) to investigate diversity in European and American Dent and Flint inbred lines, (ii) evaluate variability of traits related to biomass and flowering time and (iii) detect QTLs for these traits. For this, original Dent and Flint panels were assembled within the European Cornfed project (RINCENT *et al.* 2012), which objective was to characterize the variation of biomass related traits in maize in view of increasing the efficiency of breeding programs targeting this trait. Flint and Dent represent complementary heterotic groups to create hybrid varieties adapted to Northern European environmental conditions. These panels include first cycle lines derived from landraces representing the materials from which these groups were created, and more recent lines created by public institutes and breeding companies, to cover most diversity available in European material. All lines were genotyped with a 50k SNPs array (GANAL *et al.* 2011) and phenotyped per se and as hybrids with a tester line representative of the opposite group in a field trial network composed of 9 to 11 Western European trials.

MATERIALS AND METHODS

Genetic material and genotyping data

A previous panel (further referred to as "C-K panel") of 375 lines representing a broad diversity of European and American materials was successfully used in association genetics in previous studies (CAMUS-KULANDAIVELU *et al*., 2006, DUCROCQ *et al*., 2008, BOUCHET *et al*., 2013). Within the “CornFed” project we developed two new specific Dent and Flint panels (CF-Dent and CF-Flint) aiming at analyzing more precisely the two genetic groups of interest for maize hybrid breeding in Northern Europe, as briefly described in a methodological context by RINCENT *et al.*(2012). Both panels are composed of 300 lines aiming at best representing the diversity of these groups and different generations of genetic materials. These include the first inbred lines created from Open Pollinated Varieties (OPVs), further referred to as first cycle lines, and more recent lines developed

by public institutes or, in the case of the dent panel, private companies. The dent panel (CF-Dent, see list in table S2) includes 124 lines from the C-K panel (CAMUS-KULANDAIVELU *et al.* 2006) determined as belonging to the "Corn Belt Dent" and "Stiff Stalk" groups with an admixture coefficient above 0.5, 58 from the University of Hohenheim, 25 from CSIC, 12 from CIAM, 58 from the ex-PVP (ex Plant Variety Protection) lines (Mikel 2006; Nelson *et al*. 2008), and 23 recent lines from INRA. Similarly the Flint panel (CF-Flint, see list in table S3) includes 118 lines of the C-K panel determined as belonging to the European Flint and Northern Flint groups with an admixture coefficient above 0.5. These were complemented by lines derived from breeding programs of the following institutes:70 from the University of Hohenheim (RIEDELSHEIMER *et al*. 2012), 56 from the Misión Biológica de Galicia and the Estación Experimental de Aula Dei (CSIC), 23 from the Centro Investigacións Agrarias de Mabegondo (CIAM), 23 from the Eidgenössische Technische Hochschule Zürich (ETHZ) and 10 recent lines from the Institut National de la Recherche Agronomique (INRA).Four lines (FP1, C105, F816 and EM1027) attributed by STRUCTURE to both Dent and Flint groups with probabilities close to 0.5 in CAMUS-KULANDAIVELU *et al.* (2006) were assigned to both CF-Dent and CF-Flint panels.

These panels were genotyped with the 50k SNPs array described in GANAL *et al.* (2011), as presented in RINCENT *et al.* (2012). Individuals which had marker missing rate and/or heterozygosity higher than 0.1 and 0.05, respectively, were eliminated. Markers which had missing rate and/or average heterozygosity higher than 0.2 and 0.15, respectively, were eliminated. In each panel, few individuals were highly related. One individual was removed for pairs identical for more than 98% of the loci. Three Dent lines and nine Flint lines were eliminated for this reason. Missing genotypes (below 2% in both panels) were imputed with the software BEAGLE (BROWNING and BROWNING 2009). In total 276 and 259 phenotyped individuals passed the genotyping filters for the CF-Dent and CF-Flint panels, respectively (tables S2 and S3).The filtered markers with a Minor Allele Frequency (MAF) above 0.05 were tested for association (42214 and 39076 markers for the CF-Dent and CF-Flint panels, respectively).

Diversity analysis

To avoid the ascertainment bias noted by GANAL *et al.* (2011), we only used the markers that were developed by comparing the sequences of nested association mapping founder lines (PANZEA SNPs, GORE *et al.* 2009) in the estimation of admixture and kinship coefficients. In total 29418 and 28513markers which had a MAF above 0.01were considered for the diversity analysis in the CF-Dent and CF-Flint lines respectively. Genotypic data of each panel were organized as G matrices

with *N* rows and *L* columns, *N* and *L* being the panel size and number of SNP loci respectively. Genotype of individual *i* at marker *k* ($G_{i,k}$) was coded as 1, 0.5 or 0 for homozygote for an arbitrarily chosen allele, heterozygote and the other homozygote, respectively.

Kinship was estimated following ASTLE AND BALDING (2009) as:

$K\_Freq_{i,j} = \frac{1}{L}\sum_{l=1}^{L} \frac{(G_{i,l}-p_l)(G_{j,l}-p_l)}{p_l(1-p_l)}$, where p_l is the frequency of the allele coded 1 of PANZEA marker l in the panel of interest. Note that contrary to the Identity By State (IBS, the proportion of shared alleles) estimation, this formula gives a higher weight to loci with a low diversity. Also, similarity is higher if two individuals share rare alleles than common alleles.

Admixture was estimated in the CF-Dent and CF-Flint panels using the software ADMIXTURE (ALEXANDER *et al.* 2009) with a number of groups varying from 2 to 8. This software is based on the same statistical model as STRUCTURE (PRITCHARD *et al.* 2000; FALUSH *et al.* 2003) but uses a fast numerical optimization algorithm, which permits to considerably reduce computational time. The groups identified by the software were interpreted using the available pedigree information. Differentiation among genetic groups (*Fst*, NEI 1973) was estimated at each locus using r-hierfstat (GOUDET 2005) for each number of groups Q (from 2 to 8), using the individuals attributed to one subgroup with a probability above 0.7 (these individuals are then considered as representative of the corresponding subgroup). Diversity(Expected heterozygosity, He) was also estimated at each marker as $2p_l(1-p_l)$. A Principal Coordinates Analysis (PCoA) was performed on the genetic distance matrices (GOWER 1966), estimated as $\mathbf{1}_{N,N} - K\_Freq$, where $\mathbf{1}_{N,N}$ is a matrix of ones of the same size as K_Freq. We also represented each panel by a network, in which two individuals were linked when their relationship coefficient was above 0.2, unlinked otherwise. For this, the genomic relationship matrix was transformed in a matrix of booleans indicating if the coefficients were above 0.2 or not. These networks were drawn with a Fruchterman and Reingold's force-directed algorithm (FRUCHTERMAN and REINGOLD, 1991) with the package « network » in R 3.0.0 (R development Core Team, 2013).

Linkage Disequilibrium (LD)

To estimate the minimum number of markers needed to cover the genome, we estimated intra-chromosomic LD using all the markers. LD was first estimated as the squared correlation between the allelic doses at two markers (denoted by r^2) located on the same chromosome (HILL and ROBERTSON 1968). As kinship has to be taken into account in the GWAS model to control false positives, we need to take it into account to estimate the number of markers required to cover the genome. For this reason, the approach of MANGIN *et al.* (2012) was used to correct for kinship and

estimate the part of LD only due to linkage (r^2K). To visualize the local variation of LD, r^2K was averaged along the genome using a sliding window of 4 Mbp. This was represented on a graph together with marker diversity *(He)* and differentiation (*Fst*) after adjusting cubic smoothing splines along the genome using the R function smooth.spline (HASTIE and TIBSHIRANI, 1990).

Genetic distances between loci were taken from the map of GANAL *et al.* (2011) based on the cross F2xF252. Unmapped markers were positioned according to the local ratio between physical and genetic distances. The variation of LD with the genetic distances on each chromosome was adjusted to the model of HILL and WEIR (1988), using only the pair of markers separated by less than 4 cM. We estimated the LD decay for each chromosome as the abscissa of the intersection between the fitted curve and the horizontal line $y = 0.1$. Knowing the length of each chromosome (in cM) we could estimate the minimum number of markers required on each chromosome to get an average r^2 or r^2K of 0.1 between each pair of adjacent markers.

Phenotypic data

The Flint and Dent lines were respectively crossed to a Dent (F353) and a Flint (UH007) tester to produce hybrid progenies for phenotypic evaluation. These two lines were representative of advanced materials within their respective group. The two hybrid panels were evaluated for flowering and biomass production related traits. Two separate experiments were conducted for the Dent and Flint hybrids, with five locations in 2010, and respectively 6 and 5 locations in 2011. Within each panel, the hybrids were divided into two groups of precocity and each group was evaluated in a different block. A small number of randomly chosen entries was replicated within block (18 entries) and across blocks (18 entries) to estimate experimental error and block effects. Male and Female flowering time, plant height (PLHT), dry matter content at harvest (DMC) and dry matter yield (DMY) were registered for each plot. Male and female flowering time were converted into growing degree days in base 6°C, using the mean daily air temperature measured at each location (these measures were respectively denoted by Tass_GDD6, Silk_GDD6). The Anthesis to Silking Interval (ASI_GDD6) was obtained by subtracting Tass_GDD6 from Silk_GDD6. DMC and DMY were observed at only nine of the ten trials for the Flint panel. DMC and DMY were corrected by flowering precocity (DMCcorr and DMYcorr) by regressing the raw data on Silk_GDD6 for each block for DMC or for each trial for DMY.

$$DMC_{ijkl} = \mu + \alpha_k \times Silk\_GDD6_{ijkl} + E_{ijkl} \quad \text{and} \quad DMCcorr_{ijkl} = \hat{E}_{ijkl}$$

$$DMY_{ijkl} = \mu + \alpha_j \times Silk\_GDD6_{ijk\,l} + E_{ijkl} \quad \text{and} \quad DMYcorr_{ijkl} = \hat{E}_{ijkl}$$

with *i*, *j*, *k* and *l* the indices indicating respectively the genotype, the trial, the block and the repetition in the block, μ is the intercept, α_k and α_j are the trial specific regression coefficients on silking for DMC and DMY, respectively.

Outlier plots with obviously extreme phenotypes were excluded from the study (less than 2.5% of the observations were removed in both panels).Least-squares means of genotypes over the global network were calculated with the GLM procedure (SAS Institute, 2008) by adjusting for block and trial effects. DMY adjusted means were not corrected by block effects. Such a correction would indeed rely on the performances of the genotypes common to the two blocks, which are likely to be affected by competition effects (early genotypes being penalized in the "late" block and late genotypes favored in the "early block"). Considering the important difference of residual variance among trials, we took heteroscedasticity into account by estimating a residual variance for each trial. For all the traits except DMY and DMYcorr, weighted least squares were computed using the following model:

$Y_{ijkl} = \mu + G_i + T_j + T(B)_{jk} + E_{ijkl}$, with $E_{ijkl} \sim N(0,\sigma_j^2)$, and for DMY and DMYcorr with the model: $Y_{ijkl} = \mu + G_i + T_j + E_{ijkl}$, with $E_{ijkl} \sim N(0,\sigma_j^2)$, where Y_{ijkl} is the phenotype of the repetition *l* of genotype *i* in block k of trial *j*, μ is the global mean, G_i is the fixed genotype effect of individual *i*, T_j is the effect of trial *j*, and $T(B)_{jk}$ is the effect of block *k* within trial *j*.

Trait heritability was estimated at the level of the experimental design. For traits other than DMY and DMYcorr, variance components of heritability were estimated in two steps. In a first step, genotypes were considered as fixed effect in order to get block effect estimates based only on the between block repetitions.

$$Y_{ijkl} = \mu + G_i + T_j + B_{k(j)} + \mathbf{E}_{ijk}$$

In a second step, phenotypes were corrected by block effects and were analyzed considering genotype and genotype x trial effects as random:

$$Y_{ijkl} - \widehat{B_{k(j)}} = \mu + \mathbf{G}_i + T_j + \mathbf{GxT}_{ij} + \mathbf{E}_{ijk},$$

where $\mathbf{GxT}_{ij}$ is the random interaction effect between genotype *i* an trial *j*.

For DMY and DMYcorr, variance components of heritability were estimated in one step only to prevent confounding block effects with competition between early and late lines:

$$Y_{ijkl} = \mu + \mathbf{G}_i + T_j + \mathbf{GxT}_{ij} + \mathbf{E}_{ijk}$$

Heritabilities were then estimated as: $h^2 = \frac{\sigma_g^2}{\sigma_g^2 + {\sigma_e^2}/{r} + \sigma_{gxt}^2/L}$, where σ_g^2, σ_e^2 and σ_{gxt}^2 are the variance estimates of the random effects $\mathbf{G}_i$, $\mathbf{E}_{ijk}$ and $\mathbf{GxT}_{ij}$, respectively. L is the mean number of environments, and r is the average number of repetitions. We also computed adjusted means and

heritabilities for each trial by simplifying accordingly above described statistical models.
The lines were also evaluated per se for Tass_GDD6, Silk_GDD6 (Dent and Flint lines) and PLHT (only the Flint lines). The Dent and Flint lines were evaluated in Saint-Martin de Hinx and Gif-sur-Yvette (France), respectively. Per se least-squares genotype means were calculated with the GLM procedure by adjusting for block effect. Variances of the per se experiment were estimated with the same mixed model used to estimate heritabilities at the trial level in the hybrid experiments.

Phenotypic characterization of the genetic groups within each panel

Genetic groups defined by admixture were compared within each panel for their phenotypic performance by estimating the genetic average of each group (denoted by μ_q) using the following model:
$Y_i = \sum_{q=1}^{Q} \mu_q F_{i,q} + \mathbf{E}_i$, where Y_i is the adjusted mean of individual i, $F_{i,q}$ is the admixture coefficient of individual i in group q, N_Q is the number of groups. N_Q=8 was considered for both panels based on the results of admixture.

Statistical model for association mapping

Mixed models are classically used to detect QTLs while controlling false positive rate in GWAS (YU *et al*. 2006). Relatedness among individuals is taken into account by considering that the random polygenic effects are not independent, with a covariance matrix determined by K. A fixed structure effect (associated to a structure matrix Q) can also be included if the dataset is highly structured. Comparison of Pvalues obtained with different (Q+K) models revealed that K was sufficient to control both structure and relatedness (fig. S1).
We tested each SNP with a MAF above 0.05 (42214 and 39076 SNPs in the CF-Dent and CF-Flint panels, respectively) in the following model: $Y = X\beta + U + E$, where Y is the vector of phenotypes (adjusted means of the per se performances, or of the hybrid performances at one trial or in the whole trial network),X includes a vector of 1 and the genotypes at the tested marker (coded as 0, 0.5 or 1 as mentioned above), β includes the intercept and the additive effect of the tested marker (β_l), defined as the difference between the two homozygous genotypes, $U \rightsquigarrow N(0, K.\sigma_{gl}^2)$ is the vector of random polygenic effects,K being the kinship estimate andσ_{gl}^2 the residual polygenic variance, $E \rightsquigarrow N(0, I.\sigma_e^2)$ is the vector of remaining residual effects with variance σ_e^2, I is an identity matrix of size equal to the number of individuals (N), U and E are independent. We used two different estimates of K in the model: K_Freq as presented above, and K_Chr (RINCENT *et al.* 2014) which is

computed only with the markers physically unlinked to the tested SNP: $K\_Chr_{i,j,c} = \frac{1}{L_{-c}} \sum_{l \notin c} \frac{(G_{i,l}-p_l)(G_{j,l}-p_l)}{p_l(1-p_l)}$, where c is the considered chromosome, L_{-c} is the number of markers not located on chromosome c. This second estimator was developed to take into account the fact that including markers in high LD with the tested SNP in the kinship estimation decreased power (LISTGARTEN 2012, RINCENT *et al.* 2014). Each marker was tested for association with the different traits using a Wald test (Wald 1943) in ASReml-R (GILMOUR *et al*. 2006). The scripts were written in R 3.0.0 (R development Core Team 2013). The statistical significance threshold was set to $0.05/M_{eff}$, which corresponds to a Bonferroni correction on M_{eff} tests, M_{eff} being the number of independent tests estimated as in LI and JI (2005). This procedure evaluated 3638 and 3527 independent tests in the CF-Dent and CF-Flint panels respectively, which led to a $-\log_{10}$(Pvalue) threshold of 4.9 in both panels. Significant SNPs separated by less than 100 kb were considered as a single QTL for the interpretation of the results.

RESULTS

Diversity and structure analysis

The histograms of the Minor Allele Frequencies (MAF) of the polymorphic PANZEA markers showed a slight deficit in rare alleles in the CF-Dent panel and a slight excess in the CF-Flint panel, compared to a uniform distribution (Fig. 1). This trend was consistent with the higher proportion of monomorphic PANZEA markers observed for the most typical lines (admixture above 0.95 at N_Q=8) of the Flint group than for those of the Dent group (18% and 15% respectively). MAF was on average slightly higher in the CF-Dent (0.25) than in the CF-Flint panel (0.24), which resulted in a lower index of diversity (Nei, 1978) in the Flint than in the Dent panel (0.36 and 0.37, respectively). Locus diversity *He* was variable along the genome (Fig. 2), with generally lower values in centromeric regions.

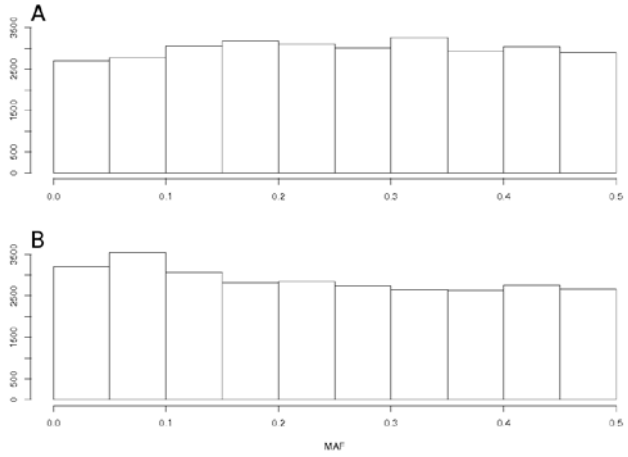

Figure 1: Histograms of the Minor Allele Frequencies of the polymorphic PANZEA markers in the CF-Dent (A) and CF-Flint (B) panels.

The cross-validation criterion proposed by ADMIXTURE suggested the presence of at least 4 main groups in both panels, and the criterion always improved with the number of groups (results not shown). For an expected number of genetic groups comprised between 2 and 8, all the subgroups identified by ADMIXTURE were interpretable in terms of pedigree and/or geographical origins. The genetic groups were composed of lines sharing a common recent ancestor (ex. F252), or a common ancestral origin (ex. Northern Flint). We noted that groups at level N_Q could generally be related to groups at level N_Q+1 by the subdivision of one subgroup into two (see Fig. S2 for an empirical synthesis). For a same number of groups, the differentiation among groups was higher in the CF-Dent than in the CF-Flint panel (Table 1). The *Fst* over the genome increased with the number of groups in both panels, but it reached a plateau at 7 in the CF-Flint panel. When considering four groups, *Fst* was variable along the genome (Fig. 2), in particular peaks of *Fst* were clearly visible in the CF-Dent panel (Chromosomes 7 and 10) and in the CF-Flint panel (Chromosome 8).

Table 1: Differenciation index among the genetic groups (*Fst*) estimated with hierfstat, for different number of groups varying from 2 to 8. Lines were attributed to a given group if their admixture was above a threshold of 0.7)

		N_Q=2	N_Q=3	N_Q=4	N_Q=5	N_Q=6	N_Q=7	N_Q=8
Fst	CF-Dent	0.07	0.15	0.24	0.26	0.30	0.32	0.35
	CF-Flint	0.07	0.13	0.15	0.18	0.21	0.23	0.22

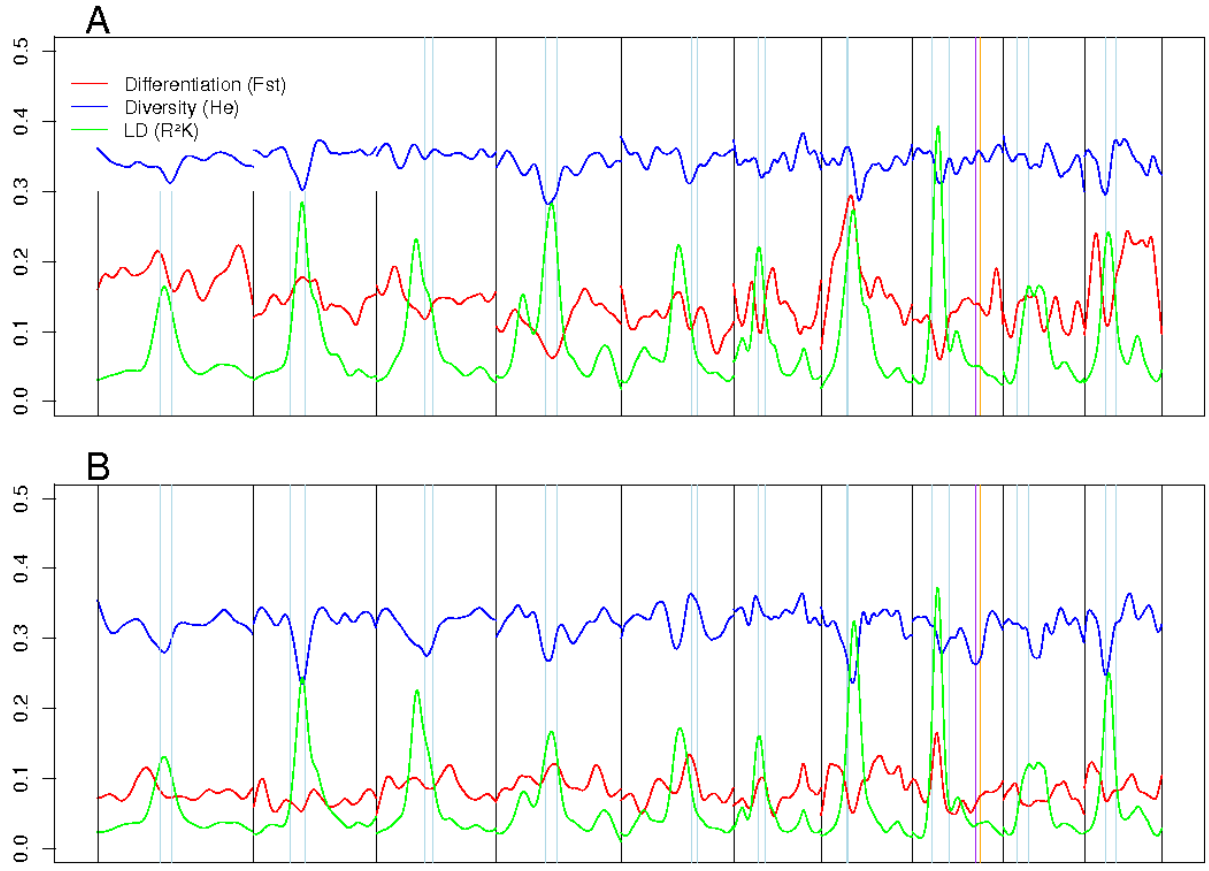

Figure 2: Differentiation among groups (*Fst*, estimated at Q=4), diversity *(He)* and Linkage Disequilibrium along the genome (physical distance in bp) in the CF-Dent (A) and CF-Flint (B) panels. For each parameter a cubic smoothing spline was adjusted along the genome. Centromere limits, *Vgt1* and *Vgt2* are located by blue, pink and purple lines, respectively.

The two first axes of the PCoA explained 16.1% and 15.7% of the variability in the CF-Dent and CF-Flint panels, respectively (Fig. 3). The different groups identified by ADMIXTURE were clearly identifiable on the PCoA graphs. The first axis separated the Iodent from the non Iodent lines in the CF-Dent panel, and the Northern Flint from the other Flint lines in the CF-Flint panel. Note that extreme positions along the axes were observed for the well known key founders of these groups (eg. Ph207 for Iodent, B73 for Stiff Stalk, Mo17 for Lancaster, D105 for Northern Flint).

Network representations of the CF-Dent and the CF-Flint panels revealed clusters of related individuals and isolated lines (Fig. 3). The shape of the network was different in the two panels: the Dent panel was composed of isolated lines and few clusters of related individuals. The network of the Flint panel also revealed clusters of related individuals but was much looser than the network of the CF-Dent panel. Groups identified with ADMIXTURE at N_Q=4 were in good agreement with the network visualization. In each panel, one of the four groups (called "Others" in Fig. 3) was composed of more heterogeneous material including many 1st cycle lines, and appeared fragmented in the network.

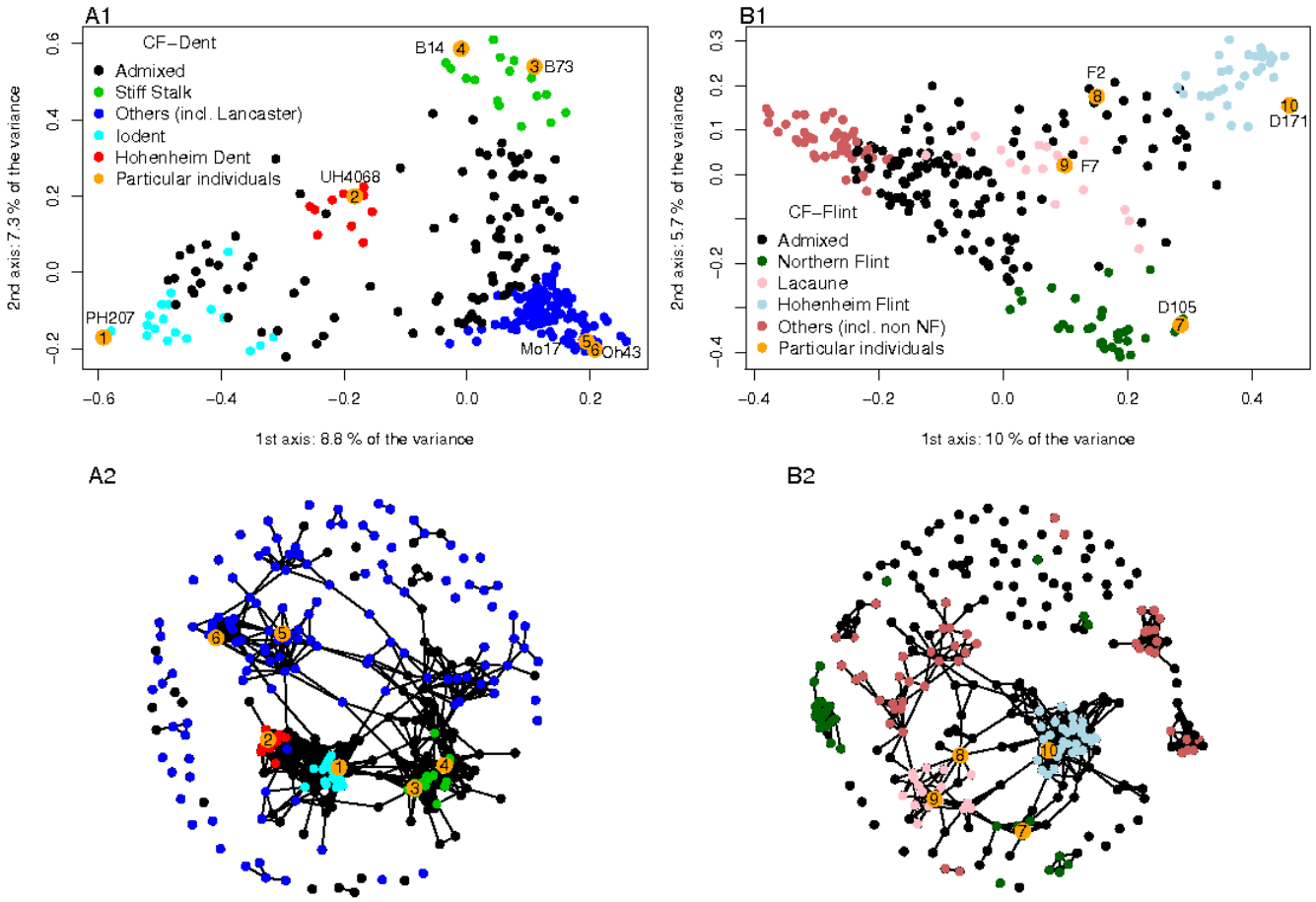

Figure 3: PCoA (A1 and B1) and network (A2 and B2) representations of the CF-Dent (A1 and A2) and CF-Flint (B1 and B2) panels. Both representation are based on the covariance matrix K_Freq. The most representative individuals of each subgroup at N_Q=4 (admixture above 0.7) were colored. Few key individuals are indicated in each panel (and numbered from 1 to 6 in CF-Dent and from 7 to10 in CF-Flint). In the network representation, individuals are linked if there covariance is above 0.2, unlinked otherwise. In these networks, distances are not informative.

Linkage disequilibrium

The LD was on average more extended in the CF-Dent than in the CF-Flint panel (0.21 and 0.12cM to reach an r^2 of 0.1 on average over all chromosomes, respectively, see Table 2). Inter-chromosomal LD was observed in both panels (fig. S3), particularly between centromeric regions. When considering physical distances, LD extent was highly variable between chromosomes and along chromosomes (Fig. 2), being more extended in centromeric regions. Taking relatedness into account substantially reduced the extent of LD in both panels, particularly in the CF-Dent panel (Table 2), and considerably reduced inter-chromosomic LD (fig. S3). For intra-chromosomic LD, the decrease observed when considering relatedness was particularly strong for chromosomes 3 and 8 in both panels, and chromosomes 4 and 7 in the CF-Dent panel only (Table 2). The chromosomes 3, 4 and 8 in the CF-Dent panel had a more extended LD (r^2K) than the others (Table 2). In the CF-Flint panel, all the chromosomes displayed similar r^2K except chromosome 8 for which LD was more extended (0.14 cM to reach a r^2K of 0.1 for chromosome 8, only 0.09 to 0.10 cM for the other chromosomes). Knowing the length of the chromosomes (in cM), these statistics allowed the estimation of the minimum number of markers required to cover the genome (assuming evenly spaced markers on the genetic map): more markers are needed in the CF-Flint (24387) than in the CF-Dent panel (19000) to get a r^2K of 0.1 between evenly spaced adjacent markers (Table 2).

Table 2: Extent of Linkage Disequilibrium and number of markers needed to reach an average r^2 or r^2K of 0.1 for each chromosome.

	CF-Dent				CF-Flint			
	r^2		r^2K		r^2		r^2K	
Chrom.	r^2 extent (cM)[a]	N markers[b]	r^2K extent (cM)[a]	N. markers[b]	r^2 extent (cM)[a]	N. markers[b]	r^2K extent (cM)[a]	N markers[b]
1	0.12	2740	0.09	3605	0.09	3636	0.09	3841
2	0.10	2461	0.09	2715	0.09	2702	0.09	2838
3	0.32	786	0.16	1572	0.16	1441	0.10	2578
4	0.27	853	0.18	1299	0.10	2209	0.09	2497
5	0.20	1179	0.13	1749	0.10	2390	0.09	2526
6	0.20	968	0.14	1332	0.10	1924	0.09	2037
7	0.28	741	0.11	1877	0.10	2086	0.09	2299
8	0.25	937	0.18	1349	0.23	1008	0.14	1736
9	0.19	993	0.11	1725	0.10	1924	0.09	2113
10	0.19	892	0.10	1778	0.10	1761	0.09	1923
Total		12552		19000		21081		24387

The genetic position of the markers was derived from the genetic map LHRE (Ganal et al. (2011). [a] genetic distance (in cM) to reach r^2 or r^2K equal to 0.1, after fitting Hill and Weir model. r^2 and r^2K calculated with the R package LDcorSV. [b] Number of markers required to reach an average r^2 of 0.1 between adjacent markers.

Table 3: Variances in the hybrid experimental design. The different traits are male (Tass_GDD6), female flowering time (Silk_GDD6), Anthesis To Silking Interval (ASI_GDD6) expressed in growing degree day in base 6°C, plant height (PLHT, cm), Dry Matter Content (DMC, %), Dry Matter Content (DMY, t/ha). DMCcorr and DMYcorr are the DMC and DMY corrected by Silk_GDD6.

		Tass_GDD6	Silk_GDD6	ASI_GDD6	PLHT	DMC	DMCcorr	DMY	DMYcorr
CF-Dent panel									
Trial network	Genot. variance	1322.3	1515.0	68.9	133.1	10.3	2.5	2.0	1.5
	TrialxG variance	295.7	436.9	87.0	51.7	3.6	2.6	1.9	1.6
	Resid. variance	324.5	375.0	218.7	129.7	5.1	5.1	3.3	3.5
	Nb of trial	11	11	11	10	11	11	11	11
	Heritability	0.96	0.96	0.73	0.89	0.93	0.80	0.82	0.78
Heritability per trial									
Trial	Mons 2010	0.77	0.84	0.58	0.62	0.81	0.42	0.54	0.57
	Pontevedra 2010	0.89	0.90	0.50	0.63	0.58	0.39	0.19	0.06
	Coruna 2010	0.81	0.77	0.31	.	0.71	0.47	0.65	0.51
	Roggestein 2010	0.88	0.92	0.69	0.86	0.89	0.52	0.58	0.58
	Einbeck 2010	0.92	0.92	0.43	0.80	0.89	0.82	0.73	0.64
	Mons 2011 late	0.92	0.90	0.43	0.47	0.86	0.46	0.79	0.77
	Moulon 2011	0.88	0.79	0.27	0.88	0.63	0.48	0.55	0.49
	Mons 2011 early	0.74	0.66	0.29	0.64	0.69	0.35	0.67	0.59
	Pontevedra 2011	0.80	0.87	0.31	0.33	0.63	0.32	0.26	0.10
	Coruna 2011	0.76	0.77	0.31	0.52	0.78	0.68	0.58	0.44
	Pocking 2011	0.83	0.82	0.38	0.41	0.84	0.65	0.60	0.58
CF-Flint panel									
Trial network	Genot. variance	1623.5	1558.6	53.5	193.9	6.8	2.2	1.9	1.4
	TrialxG variance	181.2	143.5	74.5	116.0	4.5	3.3	1.8	1.5
	Residual variance	345.4	363.7	218.9	196.3	6.5	6.6	3.8	4.2
	Nb of trial	10	10	9	9	9	9	9	9
	Heritability	0.97	0.97	0.65	0.86	0.86	0.69	0.76	0.71
Heritability per trial									
Trial	Mons 2010	0.69	0.78	0.24	0.65	.	.	.	.
	Pontevedra 2010	0.88	0.84	0.34	0.73	0.60	0.47	0.38	0.29
	Coruna 2010	0.90	0.84	0.47	.	0.47	0.41	0.45	0.31
	Roggestein 2010	0.92	0.90	0.74	0.71	0.75	0.66	0.41	0.41
	Einbeck 2010	0.90	0.94	0.36	0.78	0.82	0.61	0.56	0.47
	Moulon 2011	0.83	0.87	0.43	0.50	0.47	0.38	0.74	0.71
	Ploudaniel 2011	0.88	0.78	.	0.87	0.70	0.62	0.61	0.42
	Pontevedra 2011	0.91	0.88	0.58	0.51	0.35	0.20	0.35	0.32
	Coruna 2011	0.82	0.73	0.55	0.59	0.74	0.50	0.56	0.39
	Pocking 2011	0.82	0.77	0.00	0.85	0.78	0.66	0.62	0.62

Table 4: Variances in the per se experimental design (Dent panel, one trial), and correlation between the per se and the hybrid adjusted means (Corr Hyb/PerSe).

	CF-Dent			CF-Flint			
	Tass_GDD6	Silk_GDD6	ASI_GDD6	Tass_GDD6	Silk_GDD6	ASI_GDD6	PLHT
Genot. Variance	7382	8361	433	10440	8666	653	728.3
Residual variance	604	429	223	1186	461	1118	39.4
heritability	0.93	0.96	0.68	0.91	0.96	0.40	0.96
Corr Hyb/PerSe	0.85	0.87	0.43	0.68	0.77	0.22	0.58

Phenotypic variation

We observed a high variability for all the traits in both panels and in both hybrid and per se evaluations (Tables 3 and S1), with for instance least-squares means of DMY of the hybrids over the trial network varying between 11 and 20 t/ha in both panels. High heritabilities were observed at the trial network level (over 0.73 and 0.65 in the CF-Dent and CF-Flint panels, respectively). For most of the traits, heritability was higher in the CF-Dent than in the CF-Flint panel. This was related to higher residual variances in the CF-Flint panel. Tass_GDD6 and Silk_GDD6 were the most heritable traits (0.96 and 0.97 in the CF-Dent and CF-Flint panels respectively). ASI_GDD6 and yield traits (DMC, DMCcorr, DMY and DMYcorr) were the less heritable traits. The lowest heritability was 0.65 for ASI_GDD6 in the CF-Flint panel. The heritabilities of the per se evaluations were close to the heritabilities of the hybrid trial network (Table 4), although inbred lines were evaluated at only one trial. This was due to much higher genetic variances in the per se evaluation (up to 6.4 times higher). The correlation between the hybrid and the per se adjusted means were quite high for Tass_GDD6 and Silk_GDD6 (between 0.68 and 0.87), but lower for ASI_GDD6 (between 0.22 and 0.43). These correlations were higher in the CF-Dent than in the CF-Flint panels for the three traits (Table 4).

Phenotypic characterization of the genetic groups within each panel

For hybrid performances, we observed differences between the genetic groups identified within the two panels (Adjusted R^2 were between 0.11 and 0.47 in CF-Dent and between 0.05 and 0.41 in CF-Flint when considering 8 groups, Table 5). In the CF-Dent panel, the lines related to UH_4068 or to F252 displayed the earliest flowering time and the highest DMC and DMCcorr (Table 5). The Lancaster and Stiff Stalk groups displayed the latest flowering time and were also the most productive (DMY of up to 17.6 t/ha). In the CF-Flint panel, the Lacaune (Fv7 related), the Northern Flints, and the Hohenheim Flints displayed the earliest flowering time and the the highest DMC and DMCcorr. Groups from southern Europe (related to CIAM Aranga and descent from Italian Open Pollinated Varieties (OPV) or from other non Northern Flints introductions into Europe) displayed the latest flowering. The lines related to CIAM Aranga, to UHF047 or to Fv7 (Lacaune) were the most productive when crossed to the Dent tester (DMY of up to 16.6 t/ha). Despite the negative correlation between flowering precocity and productivity in both panels (results not shown), we could observe different levels of productivity for a same precocity in some cases. For example lines related to B73 and those related to 0h43 both displayed late flowering but the first group was more productive. In the Flint panel, the group "CIAM Aranga and EC18 related" was by far the most

productive, although earlier than other groups. We observed that the three Flint groups which had the highest contribution in first cycle lines, namely Italian OPVs, Pyrenean and NF, were the less productive, with DMY below 15 t/ha (Table 5). A similar trend was found for dents, with most first cycle lines grouped in the "Minnesota13" group, which displayed the lowest value for DMYcorr. We also noted substantial variation within genetic groups (see for example Iodent and Italian OPVs in tables S2 and S3, respectively), consistent with the limited proportion of variance explained by admixture for all the traits. Within a given group, the most typical lines (admixture above 0.98) could differ by up to 5 t/ha (e.g. non admixed individuals of the "UH_F047 family" group ranged from 12.7 to 17.7 t/ha, table S3). A formal analysis of genetic gain over breeding generations could not be conducted due to the complexity of the pedigrees but some interesting trends could be noted. For instance within the Ph207 group, most lines derived from Ph207 founder appear superior to it in terms of performance (table S2).

Table 5: Characterization of the different genetic groups at Q=8 in the CF-Dent and CF-Flint panels. The group means of each trait were obtained by regressing the adjusted means on the admixture coefficients. The different traits are male (Tass_GDD6), female flowering time (Silk_GDD6), Anthesis To Silking Interval (ASI_GDD6) expressed in growing degree day in base 6°C, plant height (PLHT, cm), Dry Matter Content (DMC, %), Dry Matter Yield (DMY, t/ha). DMCcorr and DMYcorr are the DMC and DMY corrected by Silk_GDD6.

	Genetic groups	Frequence	Tass_ GDD6	Silk_ GDD6	ASI_ GDD6	DMC	DMY	PLHT	DMCcorr	DMYcorr
CF-Dent	Stiff Stalk (B73 type)	0.07	906	916	10	31.9	17.5	267	-1.2	1.4
	Lancaster (MO17 type)	0.09	940	963	21	30.1	17.6	273	-1.0	1.0
	UH_4068 family (mostly Iodent at K=3)	0.09	854	866	12	38.2	16.1	253	2.0	0.5
	Iodent (Ph207 type)	0.15	870	887	18	36.1	16.1	252	1.0	0.3
	Stiff Stalk (B14 type)	0.12	913	927	13	33.0	17.3	263	0.2	1.1
	Minnesota13 (Wf9, A3 type)	0.27	890	916	25	32.6	15.0	253	-0.9	-1.1
	Lancaster (OH43 type)	0.09	903	920	18	31.4	16.1	254	-1.5	-0.1
	F252 family	0.11	840	853	15	39.1	14.7	241	2.1	-0.8
	Adj. R2 [a]		0.33	0.33	0.13	0.47	0.23	0.20	0.34	0.23
CF-Flint	Hohenheim Flint (D171 type, from composite)	0.13	848	876	21	33.7	14.7	247	1.1	0.1
	UH_F047 family	0.10	867	893	20	32.3	15.2	257	0.3	0.3
	Lacaune (Fv7 type)	0.11	843	874	24	33.5	15.4	239	0.8	0.7
	CIAM Aranga and EC18 related	0.06	899	931	22	30.6	16.6	258	0.0	1.3
	Descent from italian OPVs (numerous 1st cycles)	0.09	912	942	21	29.9	14.9	257	-0.3	-0.5
	Descent from non NF introductions in Europe (Spanish and others)	0.17	952	984	21	28.2	15.8	270	-0.4	-0.2
	Pyrenean (Numerous 1st cycle)	0.16	876	908	26	30.7	14.7	250	-0.8	-0.3
	NF (numerous 1st cycles)	0.18	855	891	27	32.2	14.5	253	0.0	-0.4
	Adj. R2 [a]		0.38	0.41	0.05	0.27	0.06	0.12	0.06	0.07

[a] Adjusted R^2 of the regression on the admixture at N_Q=8.

Association mapping results

The complete lists of significant SNPs are presented in tables S4 and S5, and the most significant associations (-log(Pvalue) above 5) are summarized in tables 7 and 8. The highest -log(Pvalue) were 9.98 on Chromosome 8 in CF-Dent and 6.71 on chromosome 1 in the CF-Flint panel, corresponding both to associations with flowering trait (Tass_GDD6 or Silk_GDD6).

Regarding the two statistical methods which were used, both kinship estimators (K_Freq and K_Chr) appeared efficient to control false positive rate, as revealed by QQ-plots (Fig. S1). At the chosen Bonferroni threshold, the kinship estimator K_Chr permitted the discovery of more SNPs than K_Freq for all the traits in both panel, at the trial or at the network level except for DMYcorr in the CF-Flint panel (Table 6). K_Chr permitted the discovery of 62 additional SNPs in the CF-Dent panel, and 15 in the CF-Flint panel (11 and 7 at the network level, corresponding to an increase of 41% and 39% in the CF-Dent and CF-Flint panels, respectively). Only 1 and 3 SNPs were identified with K_Freq but not with K_Chr in the CF-Dent and CF-Flint panels, respectively.

Table 6: Statistics on the significant SNPs and QTLs in the CF-Dent and CF-Flint panels evaluated on tester.

		Estimation of K	Tass_GDD6	Silk_GDD6	ASI_GDD6	PLHT	DMC	DMCcorr	DMY	DMYcorr	sum
CF-Dent	Asso_network[a]	K_Freq	12	8	0	2	3	0	1	1	27
	Asso_network	K_Chr	16	10	0	4	5	1	1	1	38
	Asso_trials[b]	K_Freq	35	27	22	12	14	5	48	33	196
	Asso_trials	K_Chr	45	39	26	14	23	8	56	47	258
	Asso_per_trial[c]	K_Freq	6.18	4.09	2	1.09	1.45	0.45	4.36	3.09	
	Asso_per_trial	K_Chr	7.91	5.82	2.36	1.36	2.27	0.73	5.18	4.45	
	Prop_Asso_specific[d]	K_Freq	0.69	0.74	1	1	0.93	1	1	0.97	4
	Prop_Asso_specific	K_Chr	0.71	0.77	1	0.93	0.96	1	0.98	0.96	2
	QTLs[e]	K_Freq and K_Chr	29	24	22	8	21	9	33	27	173
CF-Flint	Asso_network[a]	K_Freq	1	2	1	4	1	0	4	5	18
	Asso_network	K_Chr	2	3	2	8	1	0	4	5	25
	Asso_trials[b]	K_Freq	16	17	12	15	8	5	12	16	101
	Asso_trials	K_Chr	18	19	14	19	13	6	12	15	116
	Asso_per_trial[c]	K_Freq	2.3	2.3	1.2	1.6	0.8	0.5	1.2	1.6	
	Asso_per_trial	K_Chr	2.9	2.5	1.4	2.1	1.3	0.6	1.2	1.5	
	Prop_Asso_specific[d]	K_Freq	0.69	0.88	1	0.93	1	1	1	1	5
	Prop_Asso_specific	K_Chr	0.72	0.89	1	0.89	1	1	1	1	5
	QTLs[e]	Kfreq and K_Chr	14	16	15	18	11	6	13	15	108

[a] Number of significant SNPs when considering the trial network adjusted means, [b] Number of significant associations when considering the trial specific adjusted means, [c]Mean number of significant associations per trial, [d]Proportion of significant associations specific to one trial, [e] Number of regions (QTLs) detected.

Comparing the two panels, the total number of SNPs (Table 6) significant in at least one environment or at the network level with one of the two methods (K_Freq or K_Chr) was more than

two times higher in the CF-Dent (258 SNPs) than in the CF-Flint panel (116 SNPs). This difference was less pronounced when considering regions (QTLs) instead of SNPs (173 and 108 QTLs identified in the CF-Dent and CF-Flint panels, respectively). The only exception to this global trend was PLHT, for which more QTLs were discovered CF-Flint panel than in the CF-Dent.

Considering traits, more SNPs were discovered for DMY, DMYcorr, Tass_GDD6 and Silk_GDD6 than for DMC, DMCcorr, and ASI_GDD6 (Table 6). However, most of the DMY and DMYcorr SNPs (96% to 100%) were declared significant in only one environment, whereas some Tass_GDD6 and Silk_GDD6 QTLs were stable across most of the environments (Fig. 4, chromosomes 2, 3, 4, 7 and 8; and fig. 5 chromosome 1). The proportion of SNP significant in only one environment was higher in the CF-Flint than in the CF-Dent panel, with the exception of PLHT. At the network level, more SNPs were declared significant for Tass_GDD6 and Silk_GDD6 in the CF-Dent panel, and for DMY and DMYcorr in the CF-Flint panel. Note that some of the significant SNPs were associated with more than one trait (Tables S4 and S5). These pleiotropic effects particularly concerned the following couples of traits: Tass_GDD6 and Silk_GDD6, Tass_GDD6 (or Silk_GDD6) and DMC (or DMY, or PLHT), PLHT and DMY (or DMC, or Tass_GDD6, or Silk_GDD6). In the CF-Flint panel, one SNP was associated with Tass_GDD6, Silk_GDD6, PLHT and DMC.

When testing associations with per se adjusted means, QTLs of Tass_GDD6 and Silk_GDD6 were discovered but only one QTL of ASI_GDD6 (in the CF-Dent panel) and no QTL of PLHT. And again, more QTLs were found in the CF-Dent panel (25) than in the CF-Flint panel (14). Most of these QTLs were located on chromosomes 3 and 8 in the CF-Dent panel and on the chromosomes 1, 3 and 9 in the CF-Flint panel. Five QTLs of Tass_GDD6 and four of Silk_GDD6 were found associated with both hybrid and per se performances (including *Zcn8,* see discussion) in the CF-Dent panel. In the CF-Flint panel, only one QTL of Tass_GDD6 and one QTL of Silk_GDD6 were found in both hybrid and per se evaluations.

Table 7: Most significant associations in the CF-Dent panel at the network level.

Trait	Chr	Pos	MAF	-log_K_Freq[a]	-log_K_Chr[b]	effect[c]	Closest gene	Gene descr.
Tass_GDD6	8	123506141	0.27	8.81	9.98	11.65	GRMZM2G179264	ZCN8 protein
Tass_GDD6	2	178262299	0.22	7.07	7.22	11.68	GRMZM2G098828	ATP binding
Tass_GDD6	4	233828118	0.47	5.78	6.29	8.36	GRMZM2G064023	Citrate synthase activity
Tass_GDD6	8	115446396	0.14	5.37	5.93	11.58	GRMZM2G111396	Unknown
Tass_GDD6	7	122130497	0.05	5.71	5.79	17.30	GRMZM2G075348	Uncharacterized
Tass_GDD6	8	118188472	0.39	5.36	5.76	7.80	GRMZM2G047842	Uncharacterized
Tass_GDD6	8	126077120	0.37	4.56	5.74	6.88	GRMZM2G380515	Zinc ion binding
Tass_GDD6	3	150832948	0.48	5.11	5.23	-8.86	GRMZM2G082387	Transcription factor
Tass_GDD6	8	126287026	0.36	3.88	5.00	6.41	GRMZM2G118834	Uncharacterized
Silk_GDD6	8	123506141	0.27	7.83	8.86	11.97	GRMZM2G179264	ZCN8 protein
Silk_GDD6	2	178262299	0.22	7.42	7.57	12.97	GRMZM2G098828	ATP binding
Silk_GDD6	8	115446396	0.14	5.42	5.90	12.59	GRMZM2G111396	Unknown
Silk_GDD6	7	122130497	0.05	5.22	5.31	17.80	GRMZM2G075348	Uncharacterized
PLHT	2	186447969	0.14	4.59	5.19	4.28	GRMZM2G381059	Protein binding
PLHT	2	178262299	0.22	4.88	5.10	4.11	GRMZM2G098828	ATP binding
DMYcorr	5	190732112	0.19	6.00	6.07	0.49	GRMZM2G031952	Cytoskeleton
DMY	5	190732112	0.19	6.54	6.70	0.56	GRMZM2G031952	Cytoskeleton
DMC	3	150832948	0.48	5.26	5.41	0.74	GRMZM2G082387	Transcription factor
DMC	10	31219126	0.11	4.95	5.35	-1.05	AC189796.3	Unknown

[a] -log(Pvalue) with K_Freq, [b] -log(Pvalue) with K_Chr, [c] effect at the network level.

Table 8: Most significant associations in the CF-Flint panel at the network level.

Trait	Chr	Pos	MAF	-log_K_Freq[a]	-log_K_Chr[b]	effect[c]	Closest gene	Gene descr.
Tass_GDD6	1	53414468	0.24	5.37	5.75	12.14	GRMZM2G031001	DNA binding
Silk_GDD6	1	53414468	0.24	6.15	6.72	12.41	GRMZM2G031001	DNA binding
Silk_GDD6	1	300441295	0.36	5.10	5.44	-8.52	GRMZM2G377487	Unknown
PLHT	8	101237704	0.14	5.96	6.12	6.43	GRMZM2G055667	Fatty acid biosynthetic process
PLHT	1	154077833	0.17	5.31	6.10	6.03	GRMZM2G056039	Heat shock protein
PLHT	9	119310870	0.13	5.88	5.78	-7.25	GRMZM2G098179	Response to freezing
PLHT	1	153344342	0.25	4.62	5.64	5.11	GRMZM2G422631	Cell wal modification
PLHT	1	53414468	0.24	4.63	5.05	5.30	GRMZM2G031001	DNA binding
PLHT	8	84808001	0.06	4.83	5.03	8.79	GRMZM2G128809	RNA binding
DMYcorr	1	17966974	0.23	5.60	5.76	-0.52	GRMZM2G059102	Transcription factor
DMYcorr	1	154077833	0.17	5.20	5.45	0.55	GRMZM2G056039	Heat shock protein
DMY	1	154077833	0.17	6.00	6.42	0.65	GRMZM2G056039	Heat shock protein
DMY	1	153344342	0.25	4.91	5.39	0.53	GRMZM2G422631	Cell wal modification
DMC	4	152972399	0.17	5.30	5.33	-0.97	GRMZM2G406313	Cortical cell delineating
ASI_GDD6	7	32478358	0.08	5.40	5.68	-4.56	GRMZM2G472146	Signaling pathway
ASI_GDD6	7	99894530	0.24	4.79	5.09	-2.77	GRMZM2G166692	Unknown

[a] log(Pvalue) with K_Freq, [b] log(Pvalue) with K_Chr, [c] effect at the network level. [a]

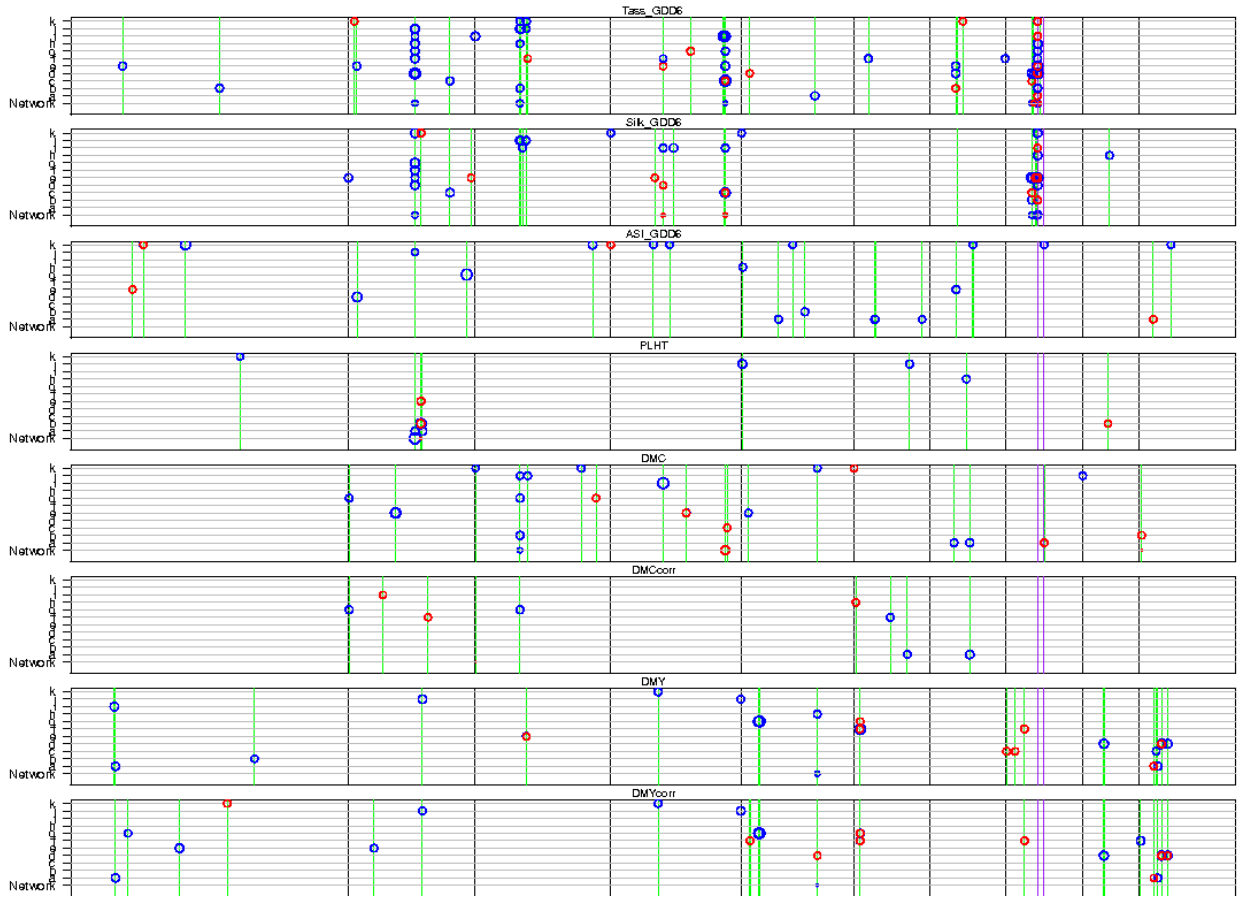

Figure 4: Significant SNPs identified in the CF-Dent panel for the different traits in the different environments and in the global adjusted means. Circle diameters is proportional to the -log10(Pvalue), and the red color indicates the additional significant SNPs when using K_Chr as covariance matrix (the markers physically linked to the tested SNP are not used to estimate kinship). Chromosomes are separated by black lines, Vgt1 and Vgt2 are indicated by purple lines. The trials are: a: Mons 2010, b: Pontevedra 2010, c: Coruna 2010, d: Roggestein 2010, e: Einbeck 2010, f: Mons 2011, g: Moulon 2011, h: Mons Precoce 2011, i: Pontevedra 2011, j: Coruna 2011, k: Pocking 2011.

Figure 5: Significant SNPs identified in the CF-Flint panel for the different traits in the different environments and in the global adjusted means. Circle diameters is proportional to the -log10(Pvalue), and the red color indicates the additional significant SNPs when using K_Chr as covariance matrix (the markers physically linked to the tested SNP are not used to estimate kinship). Chromosomes are separated by black lines, Vgt1 and Vgt2 are indicated by purple lines. The trials are: a: Mons 2010, b: Pontevedra 2010, c: Coruna 2010, d: Roggestein 2010, e: Einbeck 2010, f: Moulon 2011, g: Ploudaniel 2011, h: Pontevedra 2011, i: Coruna 2011, j: Pocking 2011.

DISCUSSION AND CONCLUSION

Genetic Diversity organization

The proportion of polymorphic PANZEA-markers was high in both panels (85% and 82% in the CF-Dent and CF-Flint panels, respectively). There was a high genetic diversity in the CF-Dent and CF-Flint panels (0.37 and 0.36, respectively), in the upper range of those reported in diversity studies based on SNPs (HAMBLIN *et al.* 2007, LU *et al.* 2009, TRUNTZLER *et al.* 2012, VAN INGHELANDT *et al.* 2010, YANG *et al.* 2010, BOUCHET *et al.* 2013). The slightly higher diversity in the Dent panel and the higher proportion of monomorphic markers in Flints are consistent with the observations of BOUCHET *et al.* (2013), who hypothesized that this could be the consequence of the severe bottleneck encountered by Flint material when diverging from tropical germplasm. As in our study (Fig. 1), BOUCHET *et al.* observed more rare alleles in the Flints and interpreted it as the possible effect of population expansion following bottleneck. The grouping made by ADMIXTURE based on the molecular information revealed the complex structure of both panels. From N_Q=2 to N_Q=8, all the identified groups could be interpreted using the pedigree information and/or known assignation to heterotic groups (Fig. S2). The groups identified in the CF-Flint panel appear to be related to the ancient history of this material. In particular, the double introduction of maize into Europe (in Southern Europe by Columbus in 1493, and in Northern Europe before 1539, REBOURG *et al.*, 2003) is still clearly visible in our results (Southern OPVs vs. Northern Flint, respectively, Fig. 3). The CF-Dent panel does not show such ancient historical patterns, consistent with the fact that this group originated from Corn-Belt dent open pollinated varieties which displayed limited population structure (CAMUS-KULANDAIVELU *et al.* 2006). Admixture groups observed in our study appear to be the result of the diverse breeding strategies which have been applied since the early development of hybrid maize in the US. The network and PCoA visualizations revealed that the material available relates to a large extent to a limited number of key lines, in particular in the CF-Dent panel (Fig. 3). Each key line and the material derived from it generated structure groups which were also clearly visible in the network and PCoA visualizations. This clustering around key lines (B73, Mo17, and PH207) corresponds to the three main dent groups (Stiff Stalk, Lancaster and Iodent, respectively) and was also shown by ROMAY *et al.* (2013) using Genotyping By Sequencing data. The fact that the Flint panel was less structured by modern breeding than the Dent panel is consistent with the fact that it was submitted to less breeding cycles. Hybrids involving Flint parents are indeed recent (1960s) compared to the first Dent hybrids (1930s) developed in the USA in the early 20th century. The different history of the panels also resulted in higher differentiation of

the groups in the CF-Dent than in the CF-Flint panel (this was true for N_Q=2 to N_Q=8, table 1). Although efforts made to assemble materials from different institutes, it appeared that some heterotic groups or families were common to these institutes. There are however some noticeable exceptions like CIAM-Aranga and Hohenheim Flints which appear specific from the institutes which created the corresponding lines.

Relatedness between individuals greatly influenced LD between pairs of markers (in particular between unlinked markers, fig. S3) in both panels but particularly in CF-Dent. When taking kinship into account, LD remained higher in the CF-Dent panel. As observed in previous studies (VAN INGHELANDT *et al.* 2010, BOUCHET *et al.* 2013), LD decreased with the physical (and genetic) distance. LD decay was variable between chromosomes, with an higher extend on chromosome 3, 4 and 8 in the CF-Dent panel, and chromosome 8 in the CF-Flint panel. This is in accordance with ROMAY *et al.* (2013) and KHOBRAGADE *et al.* (pers. com.), who identified particularly long haplotypes on chromosome 4, in regions including important domestication genes. Other important genes related to flowering time (*Vgt1* and *Vgt2*) are located on chromosome 8 (CHARDON *et al.* 2005; SALVI *et al.* 2007, 2009; VEYRIERAS *et al.* 2007; DUCROCQ *et al.* 2008; VAN INGHELANDT *et al.* 2012; BOUCHET *et al.* 2013). The slight drop of diversity in the region of *Vgt*1 and *Vgt*2 in the Flint panel (Fig. 2) may be due to the fixation of the early alleles during adaptation to short growing seasons. The higher LD extent in the CF-Dent panel resulted in a reduced number of SNPs required for a minimum coverage of the genome (19000 markers in comparison to 24387 markers in the CF-Flint). The number of SNPs available in GWAS in the panels (42214 and 39076 in the CF-Dent and CF-Flint panels respectively) makes it possible to conduct a first genome-wide analysis. However these available markers are not evenly spaced along the genetic map, and a LD of 0.1 between adjacent pairs of SNP is insufficient to detect QTLs of small to intermediate effect in our panels. In the CF-Flint panel, fewer markers were available for GWAS, whereas more markers were needed to cover the genome than in the CF-Dent panel. This could lead to a lower power in the CF-Flint panel in some regions of the genome. In both panels, we expect that a substantial gain in power could be obtained by increasing the number of markers (by combining GBS, sequencing and imputation for example).

Also, one of the main limitations in the dissection of quantitative traits is the size of the population under study, which affects GWAS power and the reliability of genomic predictions. For this reason, the panel size should be as large as possible. But we showed in this study, that at some point the sampling of additional individuals often results in relatedness (possibly high, fig. 3), which may decrease GWAS marginal gain of power. This highlights the importance of screening collections of landraces and of first cycle lines, which can probably be used to increase panel size and diversity

without increasing too much relatedness, and as a result increase the potential of the panels for the QTL detections, and for the inference of evolutionary events.

Trait variation within and among genetic groups

All the traits in both panels showed high genetic variability, which resulted in high heritabilities at the trial network levels. Male and female flowering time (Tass_GDD6 and Silk_GDD6) were the most heritable traits (above 0.96 at the network level), and Anthesis To Silking Interval (ASI_GDD6), which is highly sensitive to environmental stresses, was the less heritable trait (0.73 in CF-Dent and 0.65 in CF-Flint). Heritabilities at the trial-network level were in the range that is expected for the observed traits. Trial heritabilities of yield traits (DMC, DMCcorr, DMY and DMYcorr) were highly variable between trials, probably because of the different environmental conditions and of the different culture managements. The hybrid heritabilities were slightly higher for the Dent than for the Flint except for flowering time. This is mostly due to higher residual variances in the Flint panel, partly explained by plant lodging in some of the trials.

The heritabilities in the per se evaluation are close to the heritabilities in the hybrid evaluation for Tass_GDD6 and Silk_GDD6. This is due to a genetic variance 5.5 to 6.4 times higher, and a residual variance only 1.2 to 3.4 times higher than in the hybrid experimental design (tables 3 and 4). This difference of genetic variability between per se and hybrid evaluation is higher than what is expected under an additive model (in that case per se genetic variability should be four times higher than the hybrid genetic variability). This suggests the existence of a substantial amount of non additive genetic effects. The range of correlations between per se and hybrid adjusted means revealed the importance to evaluate biomass production potential of the lines in hybrid progenies and not per se only.

The high genetic diversity and phenotypic variability of these two panels is encouraging for the development of more productive biomass maize. Comparison of group materials revealed by population structure analysis showed a significant effect on all traits (Table 5). It highlighted groups with original characteristics like the "CIAM Aranga and EC18 related" group in the Flints, or the Stiff Stalk lines (particularly those related to B73) in the Dents, which displayed a high productivity relative to their earliness (Table 5). High variances nevertheless exist within genetic groups. Although a formal analysis was not possible due to the complexity of pedigrees, we observed some groups for which recent materials were more productive than that of founder ancestral lines (e.g. Ph207 and derivatives in Dents). This reveals that both Flint and dent groups have undergone genetic progress (Tables S2 and S3). However, substantial variability remains in the more recent

lines (e.g. group "CIAM Aranga" in Table S2), which is encouraging for further breeding. High heritability and variability within groups observed in this analysis is encouraging to run GWAS.

Association mapping results

The distribution of the P-values (QQ-plot, fig. S1) illustrates that a random polygenic effect was required to control false positive rate efficiently, and that both K_Freq and K_Chr were efficient for this (distribution near diagonal for P-values above 0.01). However the use of K_Chr instead of K_Freq substantially increased the number of significant SNPs (increase of around 40% in both panels). This confirms the importance of removing markers in LD with the tested marker from the kinship estimation. As expected from simulations in RINCENT *et al.* 2014, the gain of power was less important in the CF-Flint than in the CF-Dent panel.

QTLs were identified for all the traits in both panels (Tables 6, 7, 8, S3 and S4, Fig. 4 and 5). Globally, more QTLs were discovered in the CF-Dent than in the CF-Flint panel in the hybrid evaluation (173 and 108 QTLs respectively) and in the per se evaluation (25 and 14 QTLs, respectively). This is consistent with the higher MAF, number of markers and LD extent in the CF-Dent panel (see above).

As expected based on knowledge of trait complexity and consequences on power, more QTLs were found for Tass_GDD6 and Silk_GDD6 than for more complex traits (ASI_GDD6 or DMC), and these flowering QTLs were more stable across environments. In particular four regions of the genome in the CF-Dent panel (Fig. 4, chromosomes 2, 3, 4 and 8) and one region of the genome in the CF-Flint panel (Fig. 5, chromosome 1) were associated with flowering time in most of the environments. Polymorphism in the vicinity of ZcN8 gene appeared as the most significant in both hybrid and per se evaluations in the CF-Dent panel.It corresponds tothe *Vgt*2 QTL found in numerous studies (CHARDON *et al.* 2005; SALVI *et al.* 2007, 2009; VEYRIERAS *et al.* 2007; DUCROCQ *et al.* 2008; VANINGHELANDT *et al.* 2012; BOUCHET *et al.* 2013; ROMAY *et al*. 2013). Note that it was not significant in the flint panel, neither in hybrid nor per se evaluations, consistent with the quasi fixation of the early allele in Flint reported by BOUCHET *et al.* (2013). None of four other regions for flowering time appeared as strongly significant in BOUCHET *et al.* (2013). Also, the strong association with days to silking corresponding to gene ZmCCT (in ROMAY *et al.* 2013) on Chromosome 10 was not detected in our study, probably because the late allele at this locus (DUCROCQ *et al.*, 2009) is underrepresented in our panels, or marker density was too low in this region for capturing this effect.

PLHT was an exception to the global trend, as more QTLs were found in the CF-Flint than in the CF-Dent panel (18 and 8 QTLs, respectively), probably because it is the only trait (with

Tass_GDD6 to some extent) which had a much higher genetic variance in the CF-Flint than in the CF-Dent panel (table 3). We found common associations with the study of PEIFFER *et al.* (2014) in particular in the CF-Flint panel (e.g. the QTL close to position 249 Mb on chromosome 1 near the gene brassinosteroid-deficient dwarf1, PETTEM, 1956). Interestingly, most of the PLHT QTLs are not associated with flowering traits, as also found by PEIFFER *et al.* (2014).As both flowering time and plant height are increasingly documented in the literature and less subject than yield to GxE interactions, a formal meta-analysis of our study and literature investigations would be highly beneficial to go beyond these preliminary trends.

For DMY or DMYcorr, many significant associations were discovered but they were highly instable between environments (more than 96% of these SNPs were significant in only one environment). The genetic determinism of these traits is more difficult to investigate because of interactions with the environment and/or because they are highly integrative. We noted that some associations for DMY were common to flowering time, suggesting a pleiotropic effect of the corresponding QTL. Note however that the QTL observed at network level for DMYcorr and DMY at position 154077833 on chromosome 1 (Table 8) in the Flint and position 190732112 on chromosome 5 in the Dents do not belong to this category and therefore would be particularly interesting to select for biomass yield without modifying flowering time. Finally DMC, DMCcorr and ASI displayed the fewer number of detected QTL, highlighting that they are most likely affected by numerous factors of small effects and strong environmental effects.

Most of the significant SNPs identified with the hybrid adjusted means were different from those identified with the per se adjusted means. This could be due to interactions between alleles (dominance and possibly epistasis), which was also shown by the genetic variance higher than expected in the per se evaluations. This was more pronounced in the CF-Flint than in the CF-Dent panel. The proportion of SNPs significant in only one environment was also higher in the CF-Flint panel. We can hypothesize from the comparison between both panels, that the CF-Flint panel is probably submitted to more gene*gene and gene*environment interactions.

Conclusions:

We could illustrate, using genotypic and phenotypic information, that Dent and Flint groups have a different history and that this has strong consequences on diversity, variability, LD extent, which in turn influence detection power. The combination of phenotypic and genotypic data permitted the identification of flowering time and biomass related QTLs in both panels. This study would

probably be strongly enriched by increasing the number of markers, population size with original individuals and by using statistical models which takes interactions into account. Although further analyses are required, the identified biomass QTLs are potentially of considerable interest, because they could be introgressed in elite material to increase productivity.

Acknowledgments

We are very grateful to whom made possible the gathering of their inbred lines to our panels. In particular: Candice Gardner from United States Department of Agriculture North Central Regional Plant Introduction Station of Ames, USA ; Natalia de Leon from University of Wisconsin, USA, Geert Kleijer from Agroscope Changins-Wädenswil of Nyon, Switzerland, Wolfgang Schipprack from Universität Hohenheim of Eckartsweier, Germany, Rita Redaelli from Unita Di Ricerca per la Maiscoltura of Bergamo, Italy, Amando Ordás from Misión Biológica de Galicia of Pontevedra, Spain, Ángel Álvarez from Estacion Experimental de Aula Dei of Zaragoza, Spain, José Ignacio Ruiz de Galarreta from Centro Neiker de Arkaute of Vitoria, Spain, Laura Campo from Centro de Investigación Agraria Mabegondo of La Coruna, Spain and Jacques Laborde and colleagues from Institut National de la Rercherche Agronomique of Saint Martin de Hinx, France.

This research was jointly supported as “Cornfed project” by the French National Agency for Research (ANR), the German Federal Ministry of Education and Research (BMBF) and the Spanish ministry of Science and Innovation (MICINN). R. Rincent is jointly funded by Limagrain, Biogemma, KWS and the French ANRt. L. Moreau, S. Nicolas and A. Charcosset conducted this research in the framework of Amaizing Investissement d'Avenir program.

LITERATURE

ALEXANDER D. H., NOVEMBRE J., LANGE K., 2009 Fast model-based estimation of ancestry in unrelated individuals. Genome Res. **19**: 1655–1664.

AMIN N., C. M. VAN DUIJN and Y.S. AULCHENKO, 2007 A Genomic Background Based Method for Association Analysis in Related Individuals (P Heutink, Ed.). PLoS ONE **2**: e1274.

ASTLE W., BALDING D. J., 2009 Population Structure and Cryptic Relatedness in Genetic Association Studies. Stat. Sci. **24**: 451–471.

BEADLE G. W., 1939 Teosinte and the origin of maize. J. Hered. **30**: 245-247

BELÓ A., ZHENG P., LUCK S., SHEN B., MEYER D. J., LI B., TINGEY S., RAFALSKI A., 2007 Whole genome scan detects an allelic variant of fad2 associated with increased oleic acid levels in maize. Mol. Genet. Genomics **279**: 1–10.

BOUCHET S., SERVIN B., BERTIN P., MADUR D., COMBES V., DUMAS F., BRUNEL D., LABORDE J., CHARCOSSET A., NICOLAS S., 2013 Adaptation of Maize to Temperate Climates: Mid-Density Genome-Wide Association Genetics and Diversity Patterns Reveal Key Genomic Regions, with a Major Contribution of the Vgt2 (ZCN8) Locus. Plos One **8**: e71377.

BROWNING B. L., BROWNING S. R., 2009 A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. Am. J. Hum. Genet.**84**: 210–223.

CAMUS-KULANDAIVELU L., J.-B. VEYRIERAS, D.MADUR, V. COMBES, M. FOURMANN, *et al.*, 2006 Maize Adaptation to Temperate Climate: Relationship Between Population Structure and Polymorphism in the Dwarf8 Gene. Genetics **172**: 2449–2463.

CHARDON F., HOURCADE D., COMBES V., CHARCOSSET A., 2005 Mapping of a spontaneous mutation for early flowering time in maize highlights contrasting allelic series at two-linked QTL on chromosome 8. Theor. Appl. Genet. **112**: 1–11.

DOEBLEY J., 2004 The genetics of maize evolution. Annu. Rev. Genet. **38**: 37–59.

DUBREUIL P., DUFOUR P., KREJCI E., CAUSSE M., DE VIENNE D., GALLAIS A., CHARCOSSET A. (1996) Organization of RFLP diversity amonginbred lines of Maize representing the most

significant heteroticgroups. Crop Sci **36**: 790–799.

DUCROCQ S., MADUR D., VEYRIERAS J.-B., CAMUS-KULANDAIVELU L., KLOIBER-MAITZ M., PRESTERL T., OUZUNOVA M., MANICACCI D., CHARCOSSET A., 2008 Key Impact of Vgt1 on Flowering Time Adaptation in Maize: Evidence From Association Mapping and Ecogeographical Information. Genetics **178**: 2433–2437.

EWENS W., SPIELMAN R., 1995 The Transmission Disequilibrium Test - History, Subdivision and Admixture. Am. J. Hum. Genet.**57**: 455–464.

FALUSH D., STEPHENS M., PRITCHARD J. K., 2003 Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics **164**: 1567–1587.

FLINT-GARCIA, S.A., J.M. THORNSBERRY and E.S. BUCKLER2003. Structure of linkage disequilibrium in plants.Annu. Rev. Plant Biol. **54**:357–374.

FRUCHTERMAN T.M.J. and E.M. REINGOLD, 1991 Graph drawing by force-directed placement.Software Pract.Exper.**21**: 1129–1164.

GANAL M. W., DURSTEWITZ G., POLLEY A., BÉRARD A., BUCKLER E. S., CHARCOSSET A., CLARKE J. D., GRANER E.-M., HANSEN M., JOETS J., PASLIER M.-C. LE, MCMULLEN M. D., MONTALENT P., ROSE M., SCHÖN C.-C., SUN Q., WALTER H., MARTIN O. C., FALQUE M., 2011 A Large Maize (Zea mays L.) SNP Genotyping Array: Development and Germplasm Genotyping, and Genetic Mapping to Compare with the B73 Reference Genome (L Lukens, Ed.). Plos One **6**: e28334.

GILMOUR A.R., B.GOGEL, B.R.CULLIS, andR.THOMPSON,2009 ASREML user guide release 3.0.VSN International Ltd., Hemel Hempstead, UK

GORE M. A., CHIA J.-M., ELSHIRE R. J., SUN Q., ERSOZ E. S., HURWITZ B. L., PEIFFER J. A., MCMULLEN M. D., GRILLS G. S., ROSS-IBARRA J., WARE D. H., BUCKLER E. S., 2009 A First-Generation Haplotype Map of Maize. Science **326**: 1115–1117.

GOUDET J., 2005 hierfstat, a package for r to compute and test hierarchical F-statistics. Mol. Ecol. Notes **5**: 184–186.

GOWER J.C., 1966 Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika**53**: 325–338.

HAMBLIN M., WARBURTON M., BUCKLER E. 2007 Empirical comparison of simple sequence repeats and single nucleotide polymorphisms in assessment of maize diversity and relatedness. PLoS ONE 2.

HASTIE, T. J. and R. J. TIBSHIRANI, 1990 Generalized Additive Models.Chapman and Hal.

HERRMANN A. and J. RATH 2012.Biogas production from maize: current state, challenges and prospects. 1. Methane yield potential. BioEnergy Res. **5**(4):1027-1042.

HILL W. G. and B. S. WEIR, 1988 Variances and covariances of squared linkage disequilibria in finite populations. Theor Popul Biol**33**: 54–78.

INGHELANDT D. VAN, MELCHINGER A.E., LEBRETON C., STICH B. 2010 Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. Theor Appl Genet 120: 1289–1299.

INGHELANDT D. VAN, MELCHINGER A. E., MARTINANT J.-P., STICH B., 2012 Genome-wide association mapping of flowering time and northern corn leaf blight (Setosphaeria turcica) resistance in a vast commercial maize germplasm set. Bmc Plant Biol. **12**: 56.

JANNINK, J.L. and B. WALSH 2003 Association mapping in plant populations. p. 59–68. In Kang, M.S. (ed.) Quantitative genetics, genomics and plant breeding.CAB Int., New York, NY.

JONES P., CHASE K., MARTIN A., DAVERN P., OSTRANDER E. A., LARK K. G., 2008 Single-Nucleotide-Polymorphism-Based Association Mapping of Dog Stereotypes. Genetics **179**: 1033–1044.

LISTGARTEN J., C. LIPPERT, C. M. KADIE, R. I. DAVIDSON, E. ESKIN *et al.,* 2012 Improved Linear Mixed Models for Genome-Wide Association Studies. Nat. Meth. **9**: 525–526.

LI J., JI L., 2005 Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity **95**: 221–227.

LU Y, YAN J, GUIMARAES CT, TABA S, HAO Z, *et al*. 2009 Molecular characterization of global maize breeding germplasm based on genome-wide single nucleotide polymorphisms. Theor Appl

Genet 120: 93–115.

MANGIN B., SIBERCHICOT A., NICOLAS S., DOLIGEZ A., THIS P., CIERCO-AYROLLES C., 2012 Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. Heredity (Edinb) **108**: 285–291.

MATSUOKA Y., VIGOUROUX Y., GOODMAN M. M., SANCHEZ J., BUCKLER E., DOEBLEY J., 2002 A single domestication for maize shown by multilocus microsatellite genotyping. Proc. Natl. Acad. Sci. **99**: 6080–6084.

MIKEL M.A., 2006 Availability and analysis of proprietary dent corn inbred lines with expired US plant variety protection. Crop Sci. **46**: 2555.

NEI M., 1973 Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. **70**: 3321–3323.

NEI M., 1978 Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics **89**: 583.

NELSON P.T., N.D. COLES, J.B. HOLLAND, D.M. BUBECK, S. SMITH, *et al*., 2008 Molecular Characterization of Maize Inbreds with Expired U.S. Plant Variety Protection. Crop Sci. **48**: 1673.

OZAKI K., OHNISHI Y., IIDA A., SEKINE A., YAMADA R., TSUNODA T., SATO H., SATO H., HORI M., NAKAMURA Y., TANAKA T., 2002 Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. Nat. Genet. **32**: 650–654.

PEIFFER J. A., M. C. ROMAY, M. A. GORE, S. A. FLINT-GARCIA, Z. ZHANG *et al.* 2014 The genetic architecture of maize height. Genetics (in press).

PETTEM F. 1956 Dwarfs. Maize Genetics Cooperation Newsletter **30**: 9-10.

PRITCHARD J. K., STEPHENS M., DONNELLY P., 2000 Inference of population structure using multilocus genotype data. Genetics **155**: 945.

R development Core Team 2013 R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

RATH J., H. HEUWINKEL, A. HERRMANN 2013 Specific biogas yield of maize can be predicted by the

interaction of four biochemical constituents. BioEnergy Res. **6**(3):939-952.

REBOURG, C., M. CHASTANET, B. GOUESNARD, C. WELCKER, P. DUBREUIL *et al*., 2003 Maize introduction into Europe: the history reviewed in the light of molecular data. Theor. Appl. Genet. **106**: 895–903.

RIEDELSHEIMER C., A. CZEDIK-EYSENBERG, C. GRIEDER, J. LISEC, F. TECHNOW, *et al*., 2012 Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat. Genet. **44**: 217–220.

RINCENT R., LALOE D., NICOLAS S., ALTMANN T., BRUNEL D., *et al*. 2012 Maximizing the Reliability of Genomic Selection by Optimizing the Calibration Set of Reference Individuals: Comparison of Methods in Two Diverse Groups of Maize Inbreds (Zea mays L.). Genetics **192**: 715–728.

RINCENT R., L. MOREAU, H. MONOD, E. KUHN, A.E. MELCHINGER *et al.* 2014 Recovering power in association mapping panels with variable levels of linkage disequilibrium. Genetics. In press

ROMAY M. C., M. J. MILLARD, J. C. GLAUBITZ, J. A. PEIFFER, K. L. SWARTS *et al.*, 2013 Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biology **14**:R55

SALVI S., CASTELLETTI S., TUBEROSA R., 2009 An updated consensus map for flowering time QTLs in maize. Maydica **54**: 501.

SALVI S., SPONZA G., MORGANTE M., TOMES D., NIU X., FENGLER K. A., MEELEY R., ANANIEV E. V., SVITASHEV S., BRUGGEMANN E., 2007 Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. Proc. Natl. Acad. Sci. **104**: 11376–11381.

SAS Institute Inc. 2008 SAS/STATÒ 9.2 User's Guide. Cary, NC, USA.

TENAILLON M. I. and A.CHARCOSSET, 2011 A European perspective on maize history. C. R. Biol. **334**: 221–228.

THORNSBERRY J. M., GOODMAN M. M., DOEBLEY J., KRESOVICH S., NIELSEN D., BUCKLER E. S., 2001 Dwarf8 polymorphisms associate with variation in flowering time. Nat. Genet. **28**: 286–289.

TRUNTZLER M., RANC N., SAWKINS M., NICOLAS S., MANICACCI D., LESPINASSE D., RIBIÈRE V., GALAUP P., SERVANT F., MULLER C., *et al.* 2012 Diversity and linkage disequilibrium features in a composite public/private dent maize panel: consequences for association genetics as evaluated from a case study using flowering time. Theor. Appl. Genet.**125**(4):731-47.

VANRADEN P., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci. **91**: 4414–4423.

VEYRIERAS J.-B., GOFFINET B., CHARCOSSET A., 2007 MetaQTL: a package of new computational methods for the meta-analysis of QTL mapping experiments. BMC Bioinformatics **8**: 49.

WALD A., 1943 Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large. Trans. Amer. Math. Soc. **54**: 426-482.

YANG J, BENYAMIN B, MCEVOY BP, GORDON S, HENDERS AK, *et al.* 2010 Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42: 565–569.

YU J., PRESSOIR G., BRIGGS W. H., VROH BI I., YAMASAKI M., DOEBLEY J. F., MCMULLEN M. D., GAUT B. S., NIELSEN D. M., HOLLAND J. B., KRESOVICH S., BUCKLER E. S., 2006 A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. **38**: 203–208.

Chapter 3

Maximizing the Reliability of Genomic Selection by Optimizing the Calibration Set of Reference Individuals: Comparison of Methods in Two Diverse Groups of Maize Inbreds (*Zea mays* L.)

R. Rincent,*,†,‡,§ D. Laloë,** S. Nicolas,* T. Altmann,†† D. Brunel,‡‡ P. Revilla,§§ V. M. Rodríguez,§§ J. Moreno-Gonzalez,*** A. Melchinger,††† E. Bauer,‡‡‡ C-C. Schoen,‡‡‡ N. Meyer,‡ C. Giauffret,§§§ C. Bauland,* P. Jamin,* J. Laborde,**** H. Monod,†††† P. Flament,§ A. Charcosset,*,1 and L. Moreau*

*Unité Mixte de Recherche (UMR) de Génétique Végétale, Institut National de la Recherche Agronomique (INRA), Université Paris-Sud, Centre National de la Recherche Scientifique (CNRS), 91190 Gif-sur-Yvette, France, †BIOGEMMA, Genetics and Genomics in Cereals, 63720 Chappes, France, ‡KWS Saat AG, Grimsehlstr 31, 37555 Einbeck, Germany, §Limagrain, site d'ULICE, av G. Gershwin, BP173, 63204 Riom Cedex, France, **UMR 1313 de Génétique Animale et Biologie Intégrative, INRA, Domaine de Vilvert, 78352 Jouy-en-Josas, France, ††Max-Planck Institute for Molecular Plant Physiology, 14476 Potsdam-Golm, Germany, and Leibniz-Institute of Plant Genetics and Crop Plant Research (IPK), 06466 Gatersleben, Germany, ‡‡Unité de Recherche (UR), 1279 Etude du Polymorphisme des Génomes Végétaux, INRA, Commisariat à l'Energie Atomique (CEA) Institut de Génomique, Centre National de Génotypage, 91057 Evry, France, §§Misión Biológica de Galicia, Spanish National Research Council (CSIC), 36080 Pontevedra, Spain, ***Centro de Investigaciones Agrarias de Mabegondo, 15080 La Coruna, Spain, †††Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70599, Stuttgart, Germany, ‡‡‡Department of Plant Breeding, Technische Universität München, 85354 Freising, Germany, §§§INRA/Université des Sciences et Technologies de Lille, UMR1281, Stress Abiotiques et Différenciation des Végétaux Cultivés, 80203 Péronne Cedex, France, ****INRA, Stn Expt Mais, 40590 St Martin De Hinx, France, and ††††INRA, Unité de Mathématique et Informatique Appliquées, UR 341, 78352 Jouy-en-Josas, France

ABSTRACT Genomic selection refers to the use of genotypic information for predicting breeding values of selection candidates. A prediction formula is calibrated with the genotypes and phenotypes of reference individuals constituting the calibration set. The size and the composition of this set are essential parameters affecting the prediction reliabilities. The objective of this study was to maximize reliabilities by optimizing the calibration set. Different criteria based on the diversity or on the prediction error variance (PEV) derived from the realized additive relationship matrix–best linear unbiased predictions model (RA–BLUP) were used to select the reference individuals. For the latter, we considered the mean of the PEV of the contrasts between each selection candidate and the mean of the population (PEVmean) and the mean of the expected reliabilities of the same contrasts (CDmean). These criteria were tested with phenotypic data collected on two diversity panels of maize (*Zea mays* L.) genotyped with a 50k SNPs array. In the two panels, samples chosen based on CDmean gave higher reliabilities than random samples for various calibration set sizes. CDmean also appeared superior to PEVmean, which can be explained by the fact that it takes into account the reduction of variance due to the relatedness between individuals. Selected samples were close to optimality for a wide range of trait heritabilities, which suggests that the strategy presented here can efficiently sample subsets in panels of inbred lines. A script to optimize reference samples based on CDmean is available on request.

doi: 10.1534/genetics.112.141473
Manuscript received May 1, 2012; accepted for publication July 19, 2012
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.112.141473/-/DC1.
[1]Corresponding author: UMR de Génétique Végétale, INRA, Univ Paris-Sud, CNRS, AgroParisTech, Ferme du Moulon, F-91190, Gif-sur-Yvette, France. E-mail: alain.charcosset@moulon.inra.fr

AMONG the different methods that use molecular markers for selection, genomic selection (GS) has received considerable attention in the last decade. The objective of this approach is to predict the breeding values of candidates based on their molecular marker genotypes. A prediction formula is developed using the genotypes and phenotypes of reference individuals forming a calibration set (Meuwissen

et al. 2001). The GS formula potentially includes all the marker effects, without preselection based on a significance threshold. If the marker density is sufficient, this permits the model to capture an important part of the genetic variance (Yang *et al.* 2010). Compared to traditional marker-assisted selection (MAS), the efficiency of which is limited by the power of marker-trait association tests, GS is expected to be more efficient, especially for highly polygenic traits (Bernardo and Yu 2007). GS was first used in animal breeding, particularly dairy cattle, and its use clearly improved the selection efficiency (Hayes *et al.* 2009a). It is now also widely studied by plant breeders, and interesting results were obtained (Jannink *et al.* 2010; Crossa *et al.* 2010; Albrecht *et al.* 2011).

Powerful statistical tools and relevant data sets (genotypes and phenotypes to train the prediction model) are key factors for the predictive efficiency. There are two ways to use the genotypic data in genomic selection. The first way is to estimate the marker effects in the calibration set and then to predict the breeding values of the selection candidates by multiplying their genotypes by the marker effects. This approach is used, for example, in the mixed model called random regression–best linear unbiased predictions (RR–BLUP; Whittaker *et al.* 2000; Meuwissen *et al.* 2001). The second approach is to use the marker genotypes to estimate a relationship matrix between phenotyped individuals of the reference population and nonphenotyped individuals, candidates to selection. This relationship matrix can then be used to estimate a variance/covariance matrix between the genetic values in a mixed model called RA–BLUP (RA for realized additive relationship matrix; Zhong *et al.* 2009), or G–BLUP. It has been proven that RR and RA–BLUP are statistically equivalent under conditions presented by Habier *et al.* (2007), Goddard (2009), and Hayes *et al.* (2009b).

The implementation of genomic selection is facilitated by recent advances in genotyping. We now have access to genotyping arrays, which provide genotypes of very good quality at low cost. The costs of sequencing are also decreasing and it is, or will soon become, possible to genotype the genetic material by sequencing (Huang *et al.* 2009; Metzker 2009; Elshire *et al.* 2011). In plant breeding, large collections of individuals are usually available to the breeder, corresponding to germplasm released by public institutes, private germplasm released at the end of their protection by patent (PVP), and individuals that have been used as parents of the current breeding program. All this material can be easily genotyped and potentially used to create the calibration set. Conversely, although there have been very important advances in the automatization of phenotyping, it is still very expensive to obtain relevant phenotypes with a high heritability for a large set of individuals. In addition, multi-environment trials are needed to test individuals under different conditions and estimate the genotype x environment interactions (GEI). As a result, it is now clearly admitted that the collection of phenotypic data relevant in terms of traits and environmental conditions with respect to the breeding objectives is the most limiting factor for running genomic selection and that it is also a key factor that needs to be optimized, with the constraint of a limited budget. Beyond plant breeding, this issue extends to a large extent to animal selection for traits that are either destructive or costly to measure, such as traits related to disease resistance or fertility (Boichard and Brochard 2012).

The question is then how to choose the reference individuals (calibration set) to phenotype, to maximize the reliability of the prediction of nonphenotyped individuals that are candidates to selection. Indeed, it has been shown that the accuracy of genomic predictions (that is the correlation between predicted and true breeding values) is highly influenced by the population used to calibrate the model (Albrecht *et al.* 2011; Pszczola *et al.* 2012). In a situation in which a large collection of individuals is available, one objective is to define which ones must be included in the calibration set to discriminate as accurately as possible which individuals from the selection population are the best ones (Figure 1). A first way to perform sampling could be to choose the individuals that capture most of the diversity present in the population. Another criterion could be to select the calibration set that minimizes the prediction error variance (PEV) of the genetic values. This criterion is valid at the individual level but does not take into account the genetic variance of the contrasts between individuals and may result in the sampling of close relatives. One classical way of evaluating the efficiency of a given selection method is to compute its accuracy, defined as the correlation between predicted and true values, which is an important factor of the expected genetic gain. This criterion is directly available in simulation studies in which true genetic values are known or can be indirectly measured by using cross-validation approaches in experimental data.

A few studies have used the expected accuracy, estimated as $\sqrt{1 - \mathrm{PEV}/\sigma_g^2}$ (where σ_g^2 is the additive genetic variance, and PEV represents the part of σ_g^2 that is not accounted for by the predictions) to compare experimental designs and statistical models for dairy cattle (VanRaden 2008; Hayes *et al.* 2009c; Pszczola *et al.* 2012). In these articles, individuals were assumed to be unrelated. As a consequence this criterion has the same disadvantage as PEV: it doesn't consider the decrease of genetic variance when close relatives are sampled.

To account for this possible decrease in genetic variance, it is possible to directly maximize the expected reliabilities of the contrasts between each selection candidate and the population mean. It can be implemented with the generalized coefficient of determination (Laloë 1993), which expresses the precision of any contrast between individuals. This criterion is the squared correlation between the true and the predicted contrast of genetic values. It is a function of the PEV and of the genetic variance. The generalized coefficient of determination (CD) is used by animal geneticists to optimize experimental designs. In particular it can be used to track disconnectedness, *i.e.*, individuals that cannot

be compared because they (or their relatives) were not phenotyped at least once in the same environment. The generalized CD was used, for example, to compare the efficiency of testing designs in beef cattle (Laloë and Phocas 2003) and sheep (Kuehn *et al.* 2007).

In plant breeding, the generalized CD was used by Maenhout *et al.* (2010) to get the most accurate BLUPs from phenotypic data available from a breeding company. The phenotypic data of breeding companies are very unbalanced, some phenotypes being disconnected from the others. Maenhout *et al.* (2010) assumed that the genotyping budget was limited, and they wanted to use the phenotypes already available for predicting the value of untested hybrids. Their challenge was, then, how to choose the individuals to genotype in order to optimize the use of available phenotypes. With this exception, to our knowledge, this criterion was paid little attention in plant breeding so far and it could be used for different applications such as the optimization of the sampling of the calibration set in genomic selection.

Since phenotyping is now the limiting factor in genome-wide analysis, we consider the case in which all the individuals are genotyped but only a proportion is going to be phenotyped (calibration set). In this article, we propose a method based on the generalized CD to optimize the sampling of the calibration set for predicting as accurately as possible the nonphenotyped individuals (Figure 1). To validate our optimization algorithm, we used phenotypic data for flowering time, plant biomass, and dry matter content, collected on two maize inbred panels for which genotypic information is available and compared several strategies for selecting the calibration set.

Materials and Methods

Genetic material

Our optimization procedure was evaluated on two maize diversity panels developed for the European program "Corn-Fed." These are composed respectively of 300 Flint lines and 300 Dent lines. This material includes 242 lines from the panel presented by Camus-Kulandaivelu *et al.* (2006) and lines derived from recent breeding schemes: 58 Dent lines from PVP (Mikel 2006; Nelson *et al.* 2008), 128 from the University of Hohenheim (Riedelsheimer *et al.* 2012), 81 from the Misión Biológica de Galicia and the Estación Experimental de Aula Dei, Spain (CSIC), 35 from the Centro Investigacións Agrarias de Mabegondo, Spain (CIAM), 23 from the Eidgenössische Technische Hochschule Zürich (ETHZ), and 33 from the Institut National de la Recherche Agronomique (INRA). This collection was created with the objective of covering European and American diversity of interest for temperate climatic conditions, as available from public institutes. Choice was guided by pedigree to avoid as far as possible overrepresentation of some parental materials.

Field data

The Flint and Dent lines were respectively crossed to a Dent and a Flint tester. The two panels were evaluated separately for flowering time and biomass production in two adjacent trials at five locations in 2010: Mons (France), Pontevedra and Mabegondo (Spain), and Roggenstein and Einbeck (Germany). The hybrids within each panel were divided into two groups according to their expected precocity. These two groups were evaluated as two blocks. A small number of randomly chosen entries was replicated within blocks (18 entries) and across blocks (18 entries) to estimate experimental error and an eventual block effect. Male flowering time (Tass_GDD6), plant dry matter yield (DM_Yield), and dry matter content (DMC) were registered for each plot. DMC and DM_Yield were observed at only four of the five locations for the Flint panel. Male flowering time was registered when 50% of the plants were shedding pollen and then converted into growing degree days (GDD) in base 6°, using the mean daily air temperature measured at each location. These traits were used here as examples, to test the optimized sampling algorithm. Plants with obviously extreme phenotypes were excluded from the study (between 2.2 and 2.8% of the data were removed for each trait).

Least-squares means were calculated with the GLM procedure (SAS Institute, 2008) by adjusting for block and trial effects (the phenotypes are compiled in File S1 and File S2). Trait heritability at the level of the experimental design was estimated with a mixed model (Trial as fixed effect, genotypes and genotypes × trial as random effects) after removing the block effects. Heritability was calculated as

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_{g\times E}^2/\text{nTrial} + \sigma_E^2/\text{nRep}},$$

where σ_g^2 is the additive genetic variance, σ_E^2 is the environmental variance, $\sigma_{g\times E}^2$ is the interaction variance, nTrial is the number of trials, and nRep is the mean number of replicates over the whole experimental design.

Genotyping, diversity, and relationship matrix

The two diversity panels were genotyped with the 50k SNPs array described by Ganal *et al.* (2011). This Illumina array includes 49,585 SNPs. Individuals, which had marker missing rate and average heterozygosity >0.1 and 0.05, respectively, were eliminated. Markers, which had missing rate and average heterozygosity >0.2 and 0.15, respectively, were eliminated. In total, 261 Flint lines and 261 Dent lines passed the genotyping and phenotyping filter criteria. To avoid the bias noted by Ganal *et al.* (2011) in the diversity analysis, we used only the markers that were developed by comparing the sequences of nested association mapping founder lines (PANZEA SNPs; Gore *et al.* 2009) to estimate Nei's index of diversity (Nei 1978) and relationship coefficients (30,027 and 29,094 markers passed the filter criteria for the Dent and the Flint lines, respectively, see File S1 and

File S2). Nei's index of diversity of each Panzea SNP was calculated and averaged over the genome to estimate diversity in the two panels.

One easy way to estimate the relationship between individuals with molecular markers is to calculate for each pair of individuals the proportion of shared alleles, also called identity-by-state (IBS). With biallelic markers it can be calculated as

$$\text{A\_IBS} = \frac{\mathbf{GG}' + \mathbf{G_2G_2}'}{K},$$

where $\mathbf{G}$ is the matrix of genotypes (with dimension number of individuals × number of markers) coded as 0, 0.5, and 1 for the homozygote, the heterozygote, and the other homozygote, respectively, K is the total number of markers, and $\mathbf{G_2} = \mathbf{1} - \mathbf{G}$, where $\mathbf{1}$ is a matrix of ones.

In this formula, a same weight is given to all markers. Another formula was proposed by Leutenegger *et al.* (2003), Amin *et al.* (2007), and Astle and Balding (2009) in which a particular weight, depending on the allele frequency, is given to each marker,

$$\text{A\_freq}_{i,j} = \frac{1}{K}\sum_{k=1}^{K}\frac{(G_{i,k} - p_k)(G_{j,k} - p_k)}{p_k(1-p_k)},$$

where i and j indicate individuals, $G_{i,k}$ is the genotype of individual i at marker k, and p_k is the frequency of the allele coded 1 of marker k in the panel. This estimator attributes a higher weight to similarity for rare alleles and to markers with low diversity. The allele frequencies p_k are estimated in a reference population (here each panel). We consider here the diversity panel as the base population; as a result the mean of the values of genomic relationship matrix **A_freq** is equal to zero. This formula can give negative estimates of relationship coefficient. Negative coefficients have no sense in terms of probability, but can be interpreted as negative correlations. These two genomic relationship matrices are positive semidefinite (Astle and Balding 2009) and invertible when the number of markers is sufficient and identical individuals are removed. Genomic relationship matrices, as described above, were estimated independently in both panels.

Statistical model

The genomic predictions were based on the RA–BLUP model, which allows a more direct derivation of PEV and CD for the breeding values (see below), using the following mixed model

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e},$$

where $\mathbf{y}$ is a vector of phenotypes, $\boldsymbol{\beta}$ is a vector of fixed effects (in our case only the intercept), $\mathbf{u}$ is a vector of random genetic values, and $\mathbf{e}$ is the vector of residuals. $\mathbf{X}$ and $\mathbf{Z}$ are design matrices.

The variance of the random effects $\mathbf{u}$ is $\text{var}(\boldsymbol{u}) = \boldsymbol{A}\sigma_g^2$, where $\mathbf{A}$ is the genomic relationship matrix and σ_g^2 is the additive genetic variance in the panel. The variance of the residuals $\mathbf{e}$ is $\text{var}(\boldsymbol{e}) = \boldsymbol{I}\sigma_e^2$, where $\mathbf{I}$ is the identity matrix. The prediction of $\mathbf{u}$ is obtained by solving Henderson's (1984) equations

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \lambda\mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix},$$

where $\lambda = \sigma_e^2/\sigma_g^2$ is the ratio between the residual and the additive variances in a simplified situation; in our case

$$\lambda = \frac{\sigma_E^2/\text{nRep} + \sigma_{g\times E}^2/\text{nTrial}}{\sigma_g^2}.$$

$\mathbf{A}$ is the genomic relationship matrix. Note that in this model we consider that a trait is determined by a large number of genes, each having small and independent effects. Genetic effects are assumed to follow a Gaussian distribution according to the central limit theorem (Fisher 1918).

Optimization criteria and CD

The final objective is to identify the individuals from the population that are best suited to build the calibration panel. One strategy for reaching this objective is to maximize the precision of the prediction of the difference between the value of each nonphenotyped individual and the mean of the total population of candidate individuals, which includes the phenotyped and the nonphenotyped individuals. This difference can be viewed as a specific contrast between genetic values of individuals.

A classical approach for this is to compute the expected PEV of each individual, which can be obtained from

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \lambda\mathbf{A}^{-1} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{C_{11}} & \mathbf{C_{12}} \\ \mathbf{C_{21}} & \mathbf{C_{22}} \end{bmatrix},$$

where $\text{PEV}(\hat{\mathbf{u}}) = \text{Var}(\hat{\mathbf{u}} - \mathbf{u}) = \text{diag}(\mathbf{C_{22}}) \times \sigma_e^2$.

More generally, the PEV of any contrast $\mathbf{c}$ of the predicted performances can be calculated as

$$\text{diag}\left[\frac{\mathbf{c}'\left(\mathbf{Z'MZ} + \lambda\mathbf{A}^{-1}\right)^{-1}\mathbf{c}}{\mathbf{c'c}}\right] \times \sigma_e^2,$$

where $\mathbf{c}$ is a contrast, *i.e.*, $\mathbf{1}'\mathbf{c} = 0$. $\mathbf{M}$ is an orthogonal projector on the subspace spanned by the columns of $\mathbf{X}$: $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X'X})^{-}\mathbf{X}'$ and $(\mathbf{X'X})^{-}$ is a generalized inverse of $\mathbf{X'X}$ (Laloë 1993).

A complementary approach to optimizing the choice of individuals to be phenotyped is to estimate the expected reliability of the prediction of contrasts. Laloë (1993) expressed the precision of any contrast with the generalized CD, defined as the squared correlation between the true and the predicted contrast of genetic values. This CD is equivalent to the expected reliability of the contrast

$$\text{CD}(\mathbf{c}) = \text{diag}\left[\frac{\mathbf{c}'\left(\mathbf{A} - \lambda\left(\mathbf{Z'MZ} + \lambda\mathbf{A}^{-1}\right)^{-1}\right)\mathbf{c}}{\mathbf{c'Ac}}\right].$$

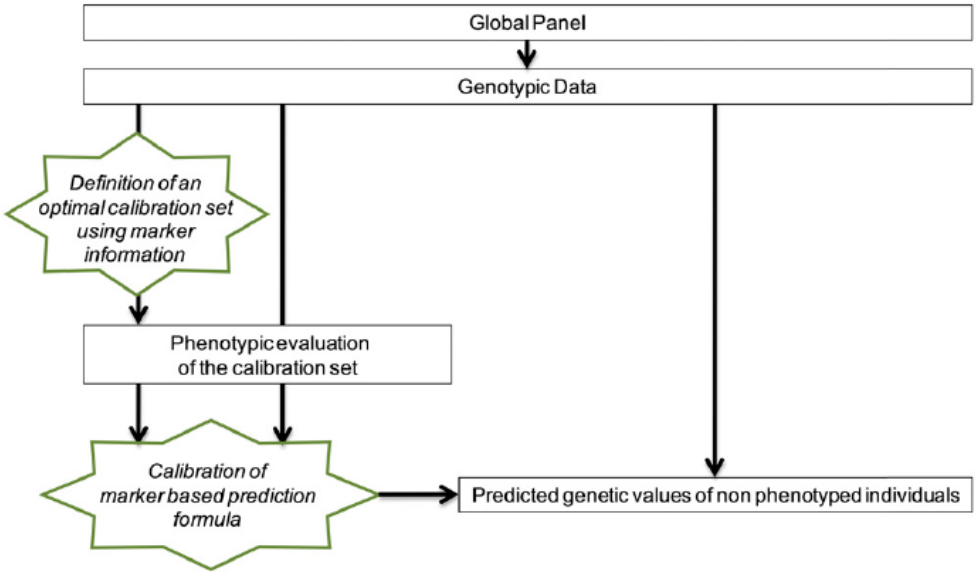

Figure 1 Optimization of calibration set to implement genomic selection in a diversity panel. This procedure was tested on two independent maize diversity panels.

The CD takes values between 0 and 1, a CD close to 0 meaning that the prediction of the contrast is not reliable, whereas CD close to 1 means that the prediction is highly reliable. The CD is a balance between PEV and the genetic variance (of the contrast), which takes into account relationship (Laloë *et al.* 1996).

Note that compared to the approach of Hayes *et al.* (2009c) who considered $\sqrt{1 - \mathrm{PEV}/\sigma_g^2}$ an estimation of accuracy, the term $\mathbf{c}'\mathbf{A}\mathbf{c}$ in the CD takes into account covariances between the candidate individuals. The use of generalized CD instead of PEV as optimization criterion is expected to prevent the selection of very closely related individuals.

The set of individuals to phenotype within each panel (Dent or Flint) was optimized by minimizing the mean of the PEVs of the contrast between each nonphenotyped individual and the mean of the panel: PEVmean = mean[diag (**PEV**(**C**))], where **C** is a matrix of contrasts: each column is a contrast between an unphenotyped individual and the mean of the population. Dimensions of **C** are total number of individuals × number of nonphenotyped individuals.

We also optimized the sampling by maximizing the mean of the CDs of the contrast between each nonphenotyped individual and the mean of the panel: CDmean = mean[diag(**CD**(**C**))]. In this case, the individuals that we decide not to phenotype are those that are the most reliably predicted with those that are phenotyped. In other words, we optimize the choice of individuals to phenotype, so that their phenotypes are as useful as possible to predict the unphenotyped individuals (Figure 1). We expect this strategy to sample key individuals that cover the panel variability as well as possible.

These approaches based on PEVmean or CDmean were used with the two relationship matrices described above: the IBS matrix **A_IBS** and the genomic relationship matrix **A_freq**.

These criteria, PEVmean and CDmean, were compared to other criteria expected to improve the calibration set sampling: we also considered as selection criteria the mean and the maximum of the genomic relationship matrix **A_freq** between the individuals in the calibration set (respectively denoted by Amean and Amax). These two criteria Amean and Amax were minimized to maximize the variability in the calibration set.

Optimization algorithm

Several exchange algorithms and simulated annealing (Kirkpatrick *et al.* 1983; Černý 1985) classically used to optimize experimental designs (Atkinson *et al.* 2007) were implemented in R 2.14.0 to optimize the different criteria. A simple exchange algorithm, further referred to as Algo1, was retained. At each step the random exchange of one individual between the calibration set and the set of nonphenotyped individuals is accepted if the criterion were improved and was rejected otherwise. More complex algorithms did not give significantly better results and needed more iterations to converge. They were therefore not retained for further investigations.

For each panel, we used Algo1 50 times to select a certain number of individuals (10, 30, 50, 70, 100, 150, or 200) for phenotyping, each time with a different random initial sample. Preliminary tests showed that 50 repetitions were sufficient to obtain stable results. We then used the true phenotypes of these individuals (calibration set) to predict the remaining individuals (validation set). We compared results obtained for optimized calibration sets with those obtained for randomly determined calibration sets (50 random sets for each calibration set size). This procedure was applied to each trait in each panel.

Observed prediction reliability and robustness of the optimization to variation of heritability

To compare the ability of the phenotyped individuals to predict the unphenotyped individuals (the validation set of individuals), we calculated the observed reliability of the predictions. The genomic selection reliability is defined by the square correlation between the genomic estimated breeding

values (GEBV) and the true breeding values (TBV): $\text{corr}^2(\text{GEBV}, \text{TBV})$, which is the square of the genomic selection accuracy (Dekkers 2007). We do not have access to the TBV of the candidate plants. Considering that $\text{corr}(\text{GEBV}, \text{Y}) = \text{corr}(\text{GEBV}, \text{TBV}) \times \text{corr}(\text{Y}, \text{TBV})$, where Y stands for the observed phenotypic performance, we estimated the genomic selection reliability as $\text{corr}^2(\text{GEBV}, \text{Y})/h^2$, since $h^2 = \text{corr}^2(\text{Y}, \text{TBV})$. For each panel and each calibration set size we compared the observed prediction reliabilities using the optimized or the random set.

In the CD calculation, the only parameter that is related to the trait is the variance ratio λ. This parameter is related to the heritability of the trait: $\lambda = (1 - h^2)/h^2$. We need to set a specific value for λ to use the sampling algorithm. But in practice, the calibration set will probably be phenotyped for traits of different heritabilities. It is thus important to know, for a set optimized with a specific value of λ, for which range of heritabilities it is optimum. To answer this question, we compared the CDmean of selection candidates obtained after sampling the calibration set with different values of lambda. If the CDmean obtained with different lambda values are correlated, one can assume that close subsets of individuals would be selected by the sampling approach.

For this, random sets of individuals were successively selected, and each time the CDmean was calculated (with the genomic relationship matrix) using three different values for λ: 4, 1, and 0.25 corresponding to heritabilities of 0.2, 0.5, and 0.8. The correlations between the three series of CDmean were then calculated.

Link between the PEV and the observed prediction error

For the Flint and the Dent panels independently, 50 sets of 150 individuals were sampled randomly or with the optimization algorithm (CDmean). These calibration sets were used to predict the genetic values of the unphenotyped individuals from the same panel. We calculated the PEVs of the contrasts between each predicted individual and the mean of the population (using a λ corresponding to the estimated heritability) and compared it to the observed prediction error (defined as the difference between the observation and the prediction). This comparison is interesting to check if our statistical model gives good estimates of the PEV and then indirectly if the estimated variance/covariance matrix fits the true variance/covariance matrix.

Genetic properties of optimized calibration sets

To visualize the genetic properties of the calibration sets optimized with CDmean, two kinds of tools were used: a principal coordinates analysis (PCoA) on the distance matrices (Gower 1966), and a network representation of the genomic relationship matrix.

A PCoA was performed on the distance matrix of each panel (we considered the distance between two individuals by one minus their relationship coefficient A_freq_{ij}). The individuals were then plotted using their coordinates on the two axes of the PCoA explaining most of the total variance. This representation gives an idea of the variability present in each panel. Using these graphs, we visualized the individuals selected by the sampling algorithm based on CDmean. It gives a rough idea of the variability of the panel captured by the calibration set.

To further understand how the individuals selected to be part of the calibration set relate to the other individuals of the population we used a visualization of the genomic relationship matrix. We represented the individuals in a network, in which two individuals are linked when their relationship coefficient (A_freq_{ij}) is >0.2, unlinked otherwise (Rozenfeld *et al.* 2008; Thomas *et al.* 2012). For this, the genomic relationship matrix was transformed in a matrix of Boolean indicating if the coefficients were >0.2 or not. The networks of the two panels were drawn with a Fruchterman and Reingold's force-directed placement algorithm (Fruchterman and Reingold 1991) with the package "network" in R.

Results

Trait variation

Tass_GDD6, DMC, and DM_Yield have an important variability in the two panels (Table 1). The average of these traits are only slightly different between the two panels because the Dent lines (usually late lines) were crossed to a Flint tester (early lines) and the Flint lines to a Dent tester. The genotype × environment interaction and the residual variances were low compared to the genetic variances for Tass_GDD6. The residual and interaction variances are relatively more important for DMC but remain below genetic variance. The residual variance was greater than the genetic variance for DM_Yield and the interaction variance was equal to the genetic variance in the Dent panel. The heritability of these traits is between 0.65 (DM_Yield in the Dent panel) and 0.95 (Tass_GDD6 in both panels).

Description of the diversity and of the genomic relationship matrix

The index of diversity (Nei 1978) in the Dent and the Flint panels was 0.34 and 0.32, respectively, leading to a mean A_IBS of 0.66 and 0.68, respectively. Histograms of the genomic relationship coefficients A_freq_{ij} in the Flint and the Dent panels show that most of the coefficients are <0.1, but some pairs of individuals are closely related in particular in the Flint panel (Figure 2). For these individuals the identity-by-state can be up to 0.99. The coefficient A_freq_{ij} of these pairs of individuals can almost reach 2 if the two individuals share many rare alleles. Three Dent and five Flint pairs were almost identical despite all the care that was used to create these diversity panels.

Observed prediction reliability and robustness of the optimization to variation of heritability

The reliabilities were lower in the Flint than in the Dent panel for the three traits and particularly for DM_Yield. For

Table 1 Statistics on Flowering time (Tass_GDD6, growing degree days), dry matter yield (DM_Yield, $t \times ha^{-1}$), and dry matter content (DMC, %) in the two panels of hybrids

	Dent			Flint		
	Tass_GDD6	DM_Yield	DMC	Tass_GDD6	DM_Yield	DMC
Mean	864.5	17.0	33.4	872.4	15.9	32.4
Genotypic variance	1354.5 ***	1.9 ***	13.0 ***	1692.1 ***	2.1 ***	8.6 ***
Trial × genotype variance	77.5 ***	1.9 ***	4.1 ***	95.8 ***	0.7 *	6.1 ***
Residual variance	292.2 ***	3.6 ***	6.5 ***	355.2 ***	3.9 ***	8.1 ***
Heritability	**0.95**	**0.65**	**0.87**	**0.95**	**0.67**	**0.72**

The variances were estimated in a mixed model with Genotype, Trial × genotype and Residual as random effects, $*P < 0.05$, $***P < 0.001$. The observations were previously corrected by block effects. The heritability corresponds to the broad-sense entry-mean heritability.

DM_Yield in the Flint panel the reliabilities are <0.3 even with a calibration set of size 200 (Figure 3). As expected the observed reliability increased with the size of the calibration set. For the random samples, an increase of the calibration set size generates an increase of the reliability following the law of diminishing returns (Figure 3). For the set optimized with PEVmean and CDmean, this trend is less clear. Within calibration set sizes, there were clear differences between the reliabilities obtained with the different approaches. All the approaches except the minimization of Amax gave better reliabilities than the reliabilities obtained after random sampling. The approach based on PEVmean was better than random sampling most of the time, but it was equivalent or worse than random sampling in few situations (particularly for DMC in the Flint panel). The reliabilities obtained by minimizing Amax in the calibration set were always lower or equivalent to those obtained by random sampling, whereas the minimization of Amean always gave higher reliabilities than random sampling (Figure 3).

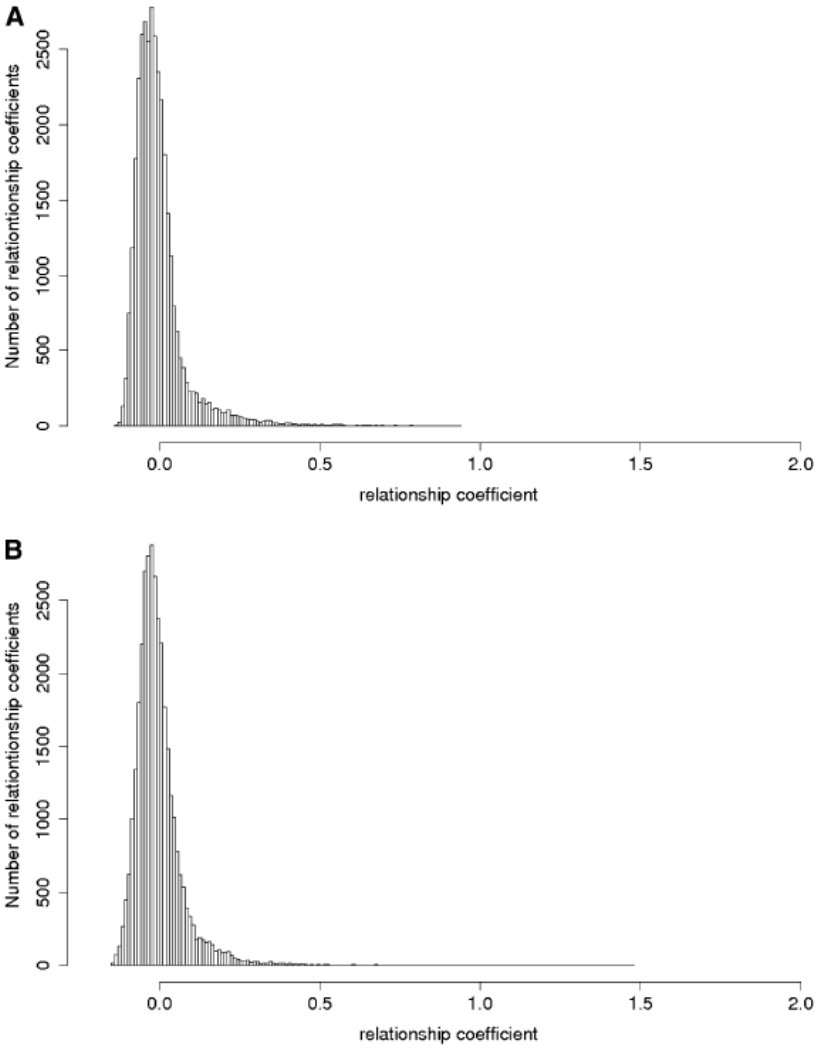

Figure 2 Histograms of the relationship coefficients between pairs of individuals. (A) Dent and (B) Flint. The relationship coefficients were extracted from **A_freq**. The two panels are considered as the reference populations; as a consequence the mean of the relationship coefficients is equal to zero in each panel.

The approach based on CDmean always gave higher reliabilities than random sampling. The use of **A_IBS** as variance/covariance matrix gave lower reliabilities. Considering the results obtained in the two panels with the different calibration set sizes, CDmean with **A_freq** was the best method.

The correlations between the CDmeans computed for the three levels of heritability were >0.90 most of the time (Table 2) and always >0.70. The CDmeans calculated with the intermediate value of h^2 ($h^2 = 0.5$) had minimum correlations of 0.86 and 0.91 with the CDmeans calculated with the two extreme heritabilities (0.2 and 0.8), for the Flint and Dent panels, respectively.

Link between the PEV and the observed prediction error

Another way of checking the reliability of our statistical models was to compare the expected PEVs and the observed prediction errors (Table 3 and Figure 4). Figure 4 illustrates the results obtained after 1 of the 50 repetitions of the algorithm on Tass_GDD6. This showed that the larger observed prediction errors mostly corresponded to high PEV, particularly for Flints.

The PEVs obtained with the approach based on CDmean were lower than the PEVs obtained with a random calibration set. This expectation was validated by the observed prediction errors, which were lower with CDmean than with random sampling.

Genetic properties of optimized calibration sets

The two first PCoA axes represented, respectively, 16.4 and 15.8% of the total variability in the Dent and the Flint panels (Figure 5). When the calibration set was small, the algorithm tended to select individuals on the extremities of the graph. When the calibration set was larger, the algorithm selected representative individuals. For example, in A2 many individuals were selected from the lower left cluster, where most individuals were placed. These patterns were stable across runs.

Figure 6 presents pairs of individual with a genomic relationship coefficient >0.2 (A_freq_{ij}) as linked by an edge. This visual representation gives a global idea of the relationships in the panels: individuals related to others are clustered into groups, while more originals lines are isolated on the graph. When few lines were phenotyped, the algorithm selected individuals representing the biggest clusters. But when the calibration set size was bigger, it was composed of few individuals in the clusters and many isolated individuals. At a given calibration set size, the algorithm selected all the "isolated" lines and few lines in the kinship clusters. When increasing even further the calibration set size, the few individuals that were not in the calibration set were located at the center of the kinship clusters.

Discussion

The objective of this study was to maximize the reliability of genomic predictions by optimizing the composition of the calibration set of individuals based on genotypic data only (Figure 1). To do so, we used different criteria that were expected to be related to the reliability of the genomic prediction. These criteria can be used before collecting phenotypic data to optimize the calibration set. The algorithms based on these criteria were tested on two independent panels that included inbred lines of different origins and on three traits with heritabilities ranging from 0.65 to 0.95. There were clear differences of observed reliabilities between the two panels and between the three traits (Figure 3). The limited number of degrees of freedom available for estimating error variance may affect the estimation of heritabilities, which may affect the scale of observed reliabilities for a given panel–trait combination (through the division by h^2). The low reliabilities obtained for the Flint panel for DM_Yield may be explained by a combination of (i) low precision of data used for prediction (similar, however, to that of Dent panel for the same trait), (ii) looser pedigree structure than in the Dent panel, and (iii) larger nonadditive effects possibly related to more important plant lodging, which deserve further investigations.

Whatever the differences in reliability range among panel–trait combinations, all the optimization criteria except Amax (the maximum of the relationship coefficients between the reference individuals) increased the observed reliability compared to random sampling.

The only exception to this was PEVmean for intermediate calibration set sizes for DMC in Flint panel. In particular, the approaches based on CDmean and Amean always gave higher reliabilities than random sampling whatever calibration set sizes. For Amean this is in accordance with Pszczola *et al.* (2012), who showed that the relatedness between the reference individuals and between the candidates and the reference individuals has a strong effect on the accuracy. For calibration sets of reduced size, Amean and CDmean yielded similar reliabilities because they both sampled the less-related individuals. For larger calibration sets, the approach based on CDmean gave better results, which can be explained by the consideration of the whole network of kinship, whereas Amean considers only the mean. CDmean explicitly takes into account the information brought by the experiment.

The optimization based on PEV was one of the most efficient approaches. However, the approach uniquely based on PEV (PEVmean) has two important drawbacks, which can explain why it can sometimes be worse than random sampling (Figure 3): (i) it doesn't take into account the decrease of genetic variance due to kinship, (ii) and it is highly dependent on the trait heritability. The first point can be neglected if all the individuals are independent. In this case the approaches based on PEVmean and on CDmean are equivalent. But most of the time the individuals considered by breeders are to some extent related, even in diversity panels like those considered in the present study. Not considering these relationship coefficients can lead to biased estimation of accuracy. This can partly explain why the

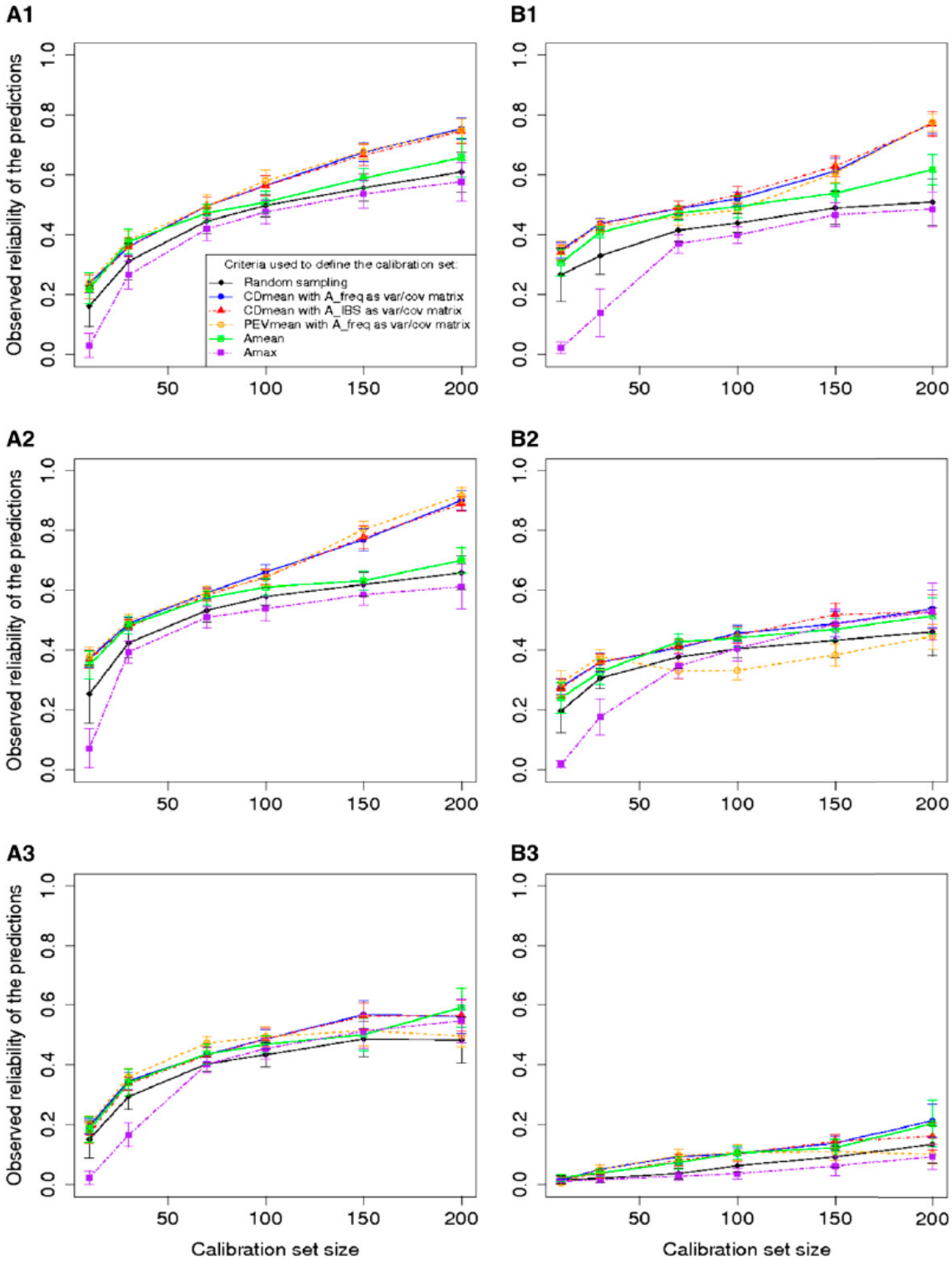

Figure 3 Reliability of the predictions of Tass_GDD6 (A1 and B1), DMC (A2 and B2) and DM_Yield (A3 and B3) using different sampling algorithms on the Dent panel (A1, A2, and A3), and the Flint panel (B1, B2 and B3). The calibration sets were randomly sampled or defined by maximizing CDmean with a relationship matrix based on the IBS or weighted by the allelic frequencies; minimizing PEVmean with a relationship matrix weighted by the allelic frequencies; minimizing the mean (Amean) or the maximum (Amax) of the relationship coefficient between the reference individuals. The individuals that are not in the calibration set are in the validation set. As a consequence, for each calibration set size the reliability is calculated with a different number of individuals. For each point, the vertical line indicates an interval of $2\sigma_R$ (σ_R being the standard deviation of observed reliabilities over the 50 runs). Optimization of PEVmean and CDmean was made with h^2 corresponding to the heritability measured for each trait in each panel.

Table 2 Correlation between the CDmeans calculated with different values of λ

Calibration set	Dent			Flint		
Size	λ=4 ; λ=1	λ=4 ; λ=0.25	λ=1 ; λ=0.25	λ=4 ; λ=1	λ=4 ; λ=0.25	λ=1 ; λ=0.25
10	0.99	0.98	1.00	0.99	0.97	0.99
50	0.93	*0.82*	0.97	0.91	*0.81*	0.98
70	*0.86*	*0.71*	0.97	0.95	*0.89*	0.99
100	0.93	*0.86*	0.98	0.97	0.94	0.99
200	0.99	0.96	0.99	0.97	0.93	0.99

For each calibration set size, the CDmeans of 200 random samples were calculated with three different values of λ. Each value of the table indicates the correlation between CDmeans calculated with two values of λ. The values in italics are the correlations <0.9. The three values of λ (4, 1, 0.25) are, respectively, equivalent to heritabilities of 0.2, 0.5, and 0.8.

formulas used in animal genetics, which consider the individuals as unrelated, overestimate accuracy compared to what is found by using cross-validation (VanRaden 2008; Hayes *et al.* 2009c; Pszczola *et al.* 2012). In the CD calculation, the covariance between the candidate individuals is taken into account by $\mathbf{c}'\mathbf{A}\mathbf{c}\sigma_g^2$, and as a result the reliability is better estimated.

The second point, sensitivity to heritability, is very important because the calibration set is often phenotyped for many traits of interest with different heritability levels. The calibration set has thus to be optimal for a wide range of heritability levels. Both PEV and CD depend on λ, which is directly related to the trait heritability. To test the effect of λ on the different methods, we used the algorithm on Tass_GDD6 with a λ of 1 corresponding to a heritability of 0.5. The reliabilities obtained with CDmean with the two λ values are very close, whereas PEVmean can be less accurate than random sampling if the λ value used for the optimization is different from the true λ (Supporting Information, Figure S1). The robustness of CDmean to variation of heritability is confirmed in Table 2, which shows that if an intermediate value of λ is chosen, the calibration set is close to optimality for a wide range of heritabilities. In fact this second point is related to the first one: the reduction of variance due to relationship is not taken into account in the PEV calculation, which makes it highly dependent on the trait heritability. For example, if the set is optimized by minimizing the PEV with a very low heritability, the calibration set is composed only of highly related individuals (results not shown), whereas if the heritability is high, the calibration set would explore the whole variability of the panel. In the CD calculation the term $\mathbf{c}'\mathbf{A}\mathbf{c}$ prevents selection of individuals too closely related.

The absence of a clear plateau for CDmean method according to calibration size in Figure 3 leads us to check whether improvement in reliability observed with CDmean-based optimization may be partly explained by the selection of validation sets (the complement to calibration set in our main approach) presenting a broad variation. To address this issue, we performed a different cross-validation procedure on Tass_GDD6. We considered here validation sets determined *a priori*. In a first step 30 individuals were randomly sampled to define the validation set. In a second step calibration sets were sampled from the remaining individuals at random or using different approaches to optimize the prediction reliability for the validation set. Although a diminishing return according to calibration population size increase was observed, the ranking in methods (Figure S2) was consistent with what was found before (Figure 3). This shows that an increase in reliability for CDmean cannot be attributed mostly to the extraction of an "easy to predict" validation set. We also performed the optimization on the adjusted means of DMC and DM_Yield of each single trial and found consistent results: the different approaches were ranked in the same order except for one trial for which the reliabilities were very low whatever the calibration set size and the method (results not shown).

Previous elements show that CDmean is preferable to PEVmean and is a criterion of choice to predict reliability and to optimize the calibration set. Under our conditions, using the optimized sampling algorithm based on CDmean and using **A_freq** as variance/covariance matrix, an optimized set of approximately 100 lines can reach the same reliability as random samples of approximately 200 lines. Cost of heavy phenotypic evaluations could therefore be substantially reduced by using an optimized calibration set.

This approach can also be used to estimate the precision of a particular prediction after collecting phenotypic data (Figure 4). This information is important because it would help the breeders to select the best individuals considering

Table 3 Means of the expected and observed error variances in the Dent and Flint panels for Tass_GDD6

	Dent		Flint	
	Mean PEVmean	Observed prediction error variance	Mean PEVmean	Observed prediction error variance
Random set	865.6	654.7	1204.1	973.8
Optimized set	610.8	367.9	857.9	699.8

The calibration set was composed of 150 individuals randomly sampled, or sampled with the algorithm based on CDmean. The procedure was repeated 50 times.

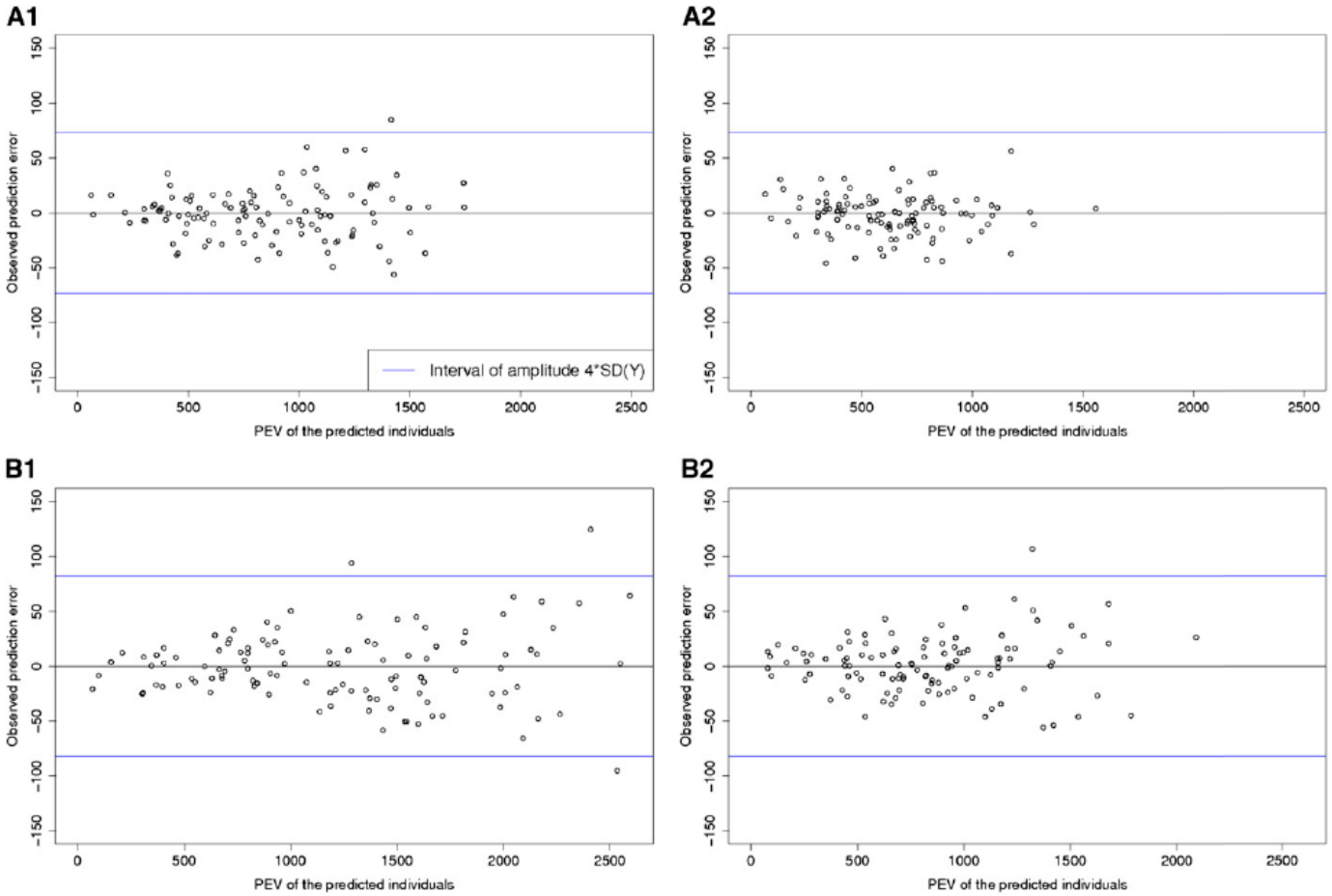

Figure 4 PEV and observed prediction errors for Tass_GDD6 (calibration set size, 150 individuals). (A1 and A2) Dent panel (261 hybrids), calibration set randomly sampled (A1) or optimized with CDmean (A2). (B1 and B2) Flint panel (261 hybrids), calibration set randomly sampled (B1) or optimized with CDmean (B2). The blue lines indicate an interval of 4SD(*Y*) [SD(*Y*) being the standard deviation of the adjusted means]. The PEVs were calculated with a λ value corresponding to the estimated heritability of each panel.

not only the best predicted values but also associated reliabilities. This information would also be useful to identify situations in which a complementary sampling of the calibration data set is needed to increase the reliability of the predictions of original individuals that were poorly predicted with the initial calibration set.

When the calibration set is small, it appears that the algorithm based on CDmean samples individuals that are "extreme" on the PCoA representation (Figure 5). As a consequence, the variability explained by the main axes is well captured by the calibration set. When the calibration set is larger, the selected individuals are spread across the whole graph, and they are always separated by a minimum distance. When two individuals are highly related, the algorithm never selects both of them as clearly illustrated by network visualizations (Figure 6). The number of clusters depends on the threshold used to determine if two individuals appear related or not. We used a threshold on A_freq_{ij} of 0.2 because the clusters of related lines were then clearly visible. When the calibration set is small, the individuals selected are in the biggest clusters. This choice permits reliable prediction of more individuals than if isolated lines were selected. If the calibration set becomes larger, both isolated and linked individuals are selected. It can be explained by the fact that when the clusters are represented by a sufficient number of phenotyped individuals, it brings more information to phenotype an isolated individual than an additional one in the clusters. At a certain calibration set size, the only lines that are not in the calibration set are in the center of the clusters. These lines are among the most typical of each group; they are also the most easily predicted when many genetically close lines are phenotyped.

In addition to these general trends, we showed that the selection of the reference individuals by the approaches based on CDmean or PEVmean depends on the method used to estimate the variance/covariance matrix. This relationship matrix should reflect the variance/covariance between individuals at the QTL positions. It is thus possible that the best formula with which to estimate A is not the same for different traits, according to the weight that is given to the markers. The use of **A_freq** instead of **A_IBS** slightly increased the observed reliability of the predictions. It shows that **A_freq** gave better estimates of the relationship coefficient between individuals than **A_IBS**, at least with our data.

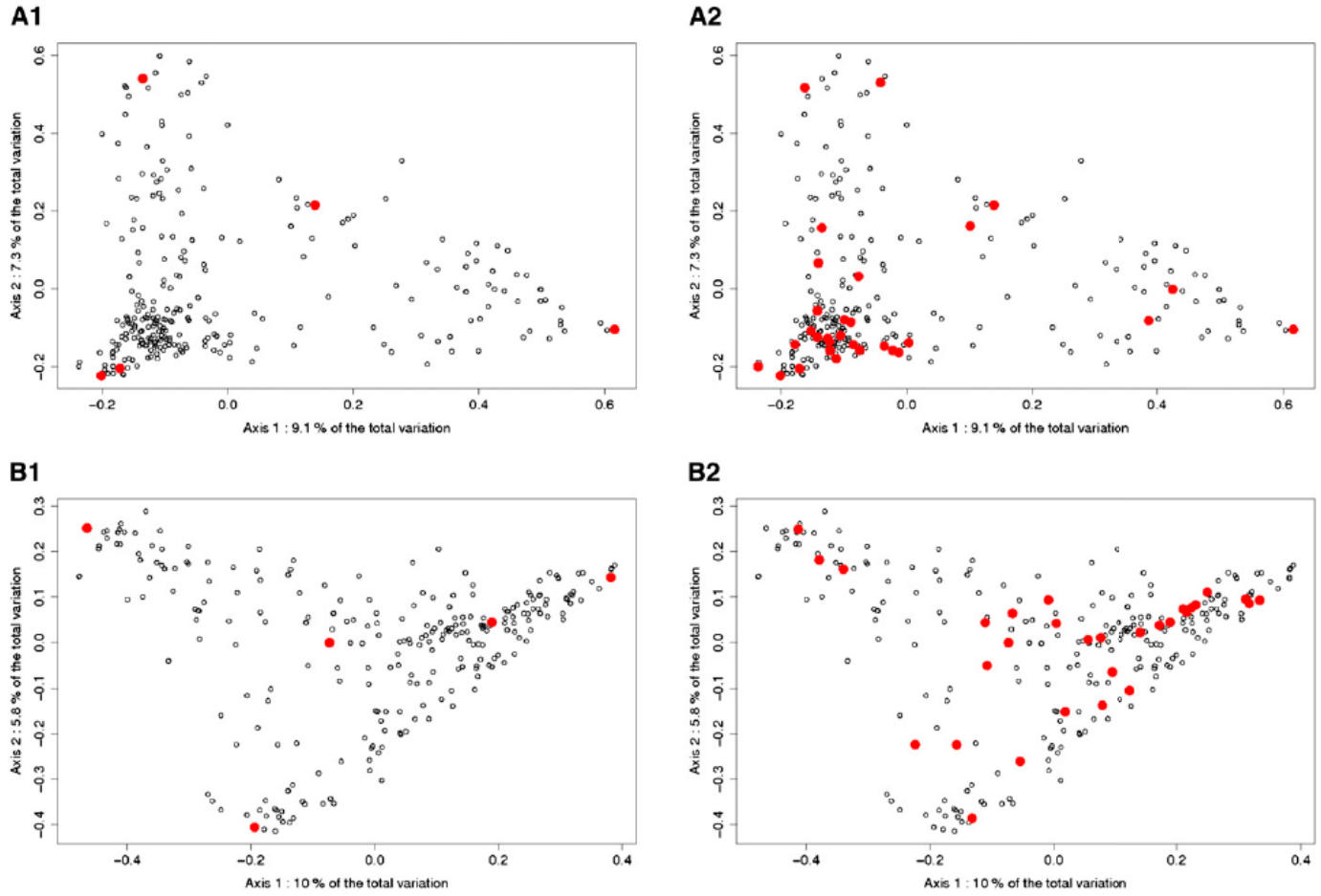

Figure 5 Principal coordinates analysis on the Dent and the Flint panel. Axis1 and Axis2 are the two first components of a PCoA on the distance matrix of the corresponding panel. The individuals selected by the algorithm based on CDmean are represented by red dots, other by circles. A1 and A2: PCoA on the Dent panel, calibration set composed of 5 individuals (A1) and 30 individuals (A2). B1 and B2: PCoA on the Flint panel, calibration set composed of 5 individuals (B1) and 30 individuals (B2).

In the case of highly polygenic traits, we consider that the QTL are spread on the whole genome, and so we use markers covering the whole genome to estimate the variance/covariance matrix. We need a number of markers high enough to have at least one marker in high linkage disequilibrium (LD) with each QTL. Goddard *et al.* (2011) showed that an incomplete coverage of the genome by markers can be a cause of overestimation of the accuracy. CDmean and PEVmean could be subject to this bias because we used a variance/covariance matrix estimated with markers to calculate these criteria. Goddard *et al.* (2011) proposed calculating a variance/covariance matrix based on the genomic relationship matrix and on the pedigree to predict accuracy without bias. In our case the pedigree was not available and so we could not use their correction. However, our marker density compared to LD was such that a risk of having an important bias was limited.

The approaches we proposed were tested on two independent diversity panels and three traits and globally consistent results were obtained. It would be interesting to test these approaches on other types of populations, in particular in the presence of strong population structure. We have considered here two heterotic groups separately. It may be interesting to test the approach to optimizing samples including lines of different heterotic groups, with the objective of obtaining accurate predictions across and within heterotic groups. It would then be required to have an important coverage of the genome to capture ancestral LD, otherwise the reliability would be overestimated as discussed before. Breeders are also interested in applying genomic selection in multifamilial populations (Albrecht *et al.* 2011; Zhao *et al.* 2012). Albrecht *et al.* (2011) showed that in such situations the prediction reliabilities are highly dependent on the composition of the calibration set. In particular, if few families are not represented in the calibration set, the observed reliabilities are lower than if few individuals are sampled in each family. Optimizing the calibration set therefore deserves specific attention in this case. CDmean could be used to optimize the sampling if the proper contrasts are considered: between each individual and its family mean, between each individual and the mean of the population, and between each family. These questions deserve consideration in future studies. Our study was based on diversity panels, and we could not evaluate how the

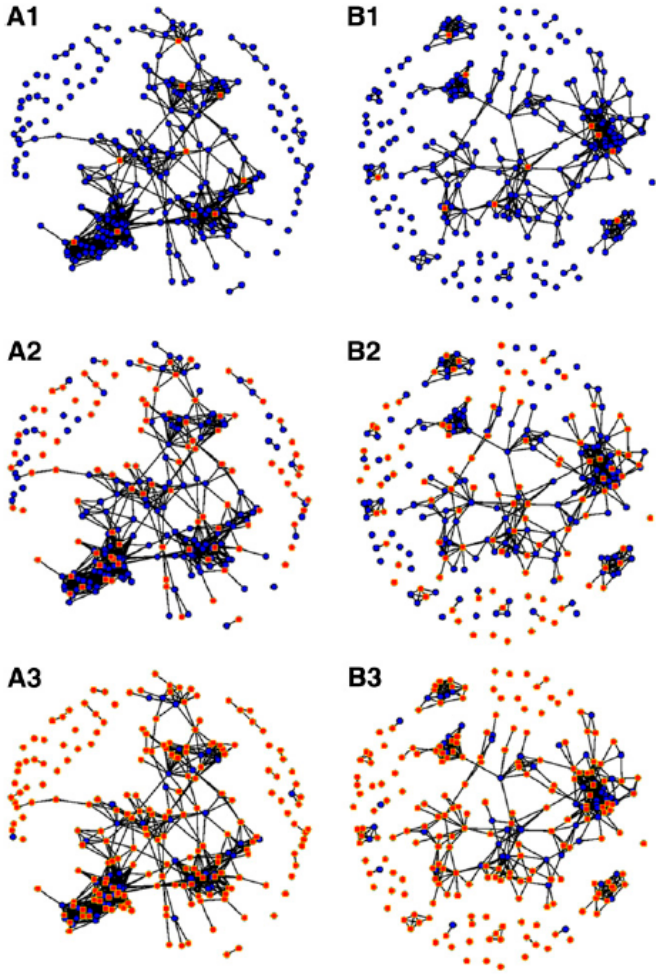

Figure 6 Network representation of the genomic relationship coefficients. (A1, A2, and A3) Dent panel, 3 calibration set sizes: 10 (A1), 100 (A2), and 200 (A3). (B1, B2 and B3) Flint panel, 3 calibration set sizes: 10 (B1), 100 (B2), and 200 (B3). These networks are drawn with a Fruchterman and Reingold's force-directed placement algorithm. Each node represents an individual; the pairs of individuals with a relationship coefficient >0.2 are linked by an edge. The individuals selected by the CDmean algorithm are represented by red squares and others by blue points.

reliability would evolve across the next generations derived from these materials. This aspect also has to be studied, because the gain of time due to selection on predicted values instead of phenotypic observations is the main interest of genomic selection. It would therefore be important to evaluate how often the prediction formula must be recalibrated.

Finally, although displaying contrasted heritabilities and possibly different contribution of nonadditive effects (see above), the three traits considered here are known to be highly polygenic (see Chardon *et al.* 2004 and Buckler *et al.* 2009 for Tass_GDD6), which justified the choice of the RA–BLUP model. For traits depending on major genes, this model might be inappropriate or nonoptimal and it may be preferable to use Bayesian or neural network models (Jannink *et al.* 2010). Our optimization criterion is based on the BLUP theory and so would be inappropriate if major genes are involved. It is, however, possible that CDmean would also be to some extent useful in increasing the reliability of Bayesian methods. It would be interesting to derive a similar criterion from the Bayesian theory to predict reliability before collecting phenotypes.

Acknowledgments

We are very grateful to those who made possible the gathering of inbred lines to our panels, in particular the following: Candice Gardner from United States Department of Agriculture North Central Regional Plant Introduction Station of Ames, Geert Kleijer from Agroscope Changins-Wädenswil of Nyon, Switzerland, Wolfgang Schipprack from Universität Hohenheim of Eckartsweier, Germany, Amando Ordás from Misión Biológica de Galicia of Pontevedra, Spain, Ángel Álvarez from Estacion Experimental de Aula Dei of Zaragoza, Spain, José Ignacio Ruiz de Galarreta from Centro Neiker de Arkaute of Vitoria, Spain, Laura Campo from Centro de Investigación Agraria Mabegondo of La Coruna, Spain, and Jacques Laborde and colleagues from Institut National de la Rercherche Agronomique of Saint Martin de Hinx, France. The authors thank the reviewers and the editor for their comments, which improved the manuscript. This research was jointly supported as "Cornfed project" by the French National Agency for Research (ANR), the German Federal Ministry of Education and Research (BMBF), and the Spanish Ministry of Science and Innovation (MICINN). R. Rincent is jointly funded by Limagrain, Biogemma, Kleinwanzlebener Saatzucht AG (KWS), and the Association Nationale de la Recherche et de la Technologie (ANRT).

Literature Cited

Albrecht, T., V. Wimmer, H.-J. Auinger, M. Erbe, C. Knaak *et al.*, 2011 Genome-based prediction of testcross values in maize. Theor. Appl. Genet. 123: 339–350.

Amin, N., C. M. van Duijn, and Y. S. Aulchenko, 2007 A genomic background based method for association analysis in related individuals. PLoS ONE 2: e1274.

Astle, W., and D. J. Balding, 2009 Population structure and cryptic relatedness in genetic association studies. Stat. Sci. 24: 451–471.

Atkinson, A. C., A. N. Donev, and R. D. Tobias, 2007 *Optimum Experimental Designs, With SAS*. Clarendon Press, Oxford.

Bernardo, R., and J. Yu, 2007 Prospects for genomewide selection for quantitative traits in maize. Crop Sci. 47: 1082.

Boichard, D., and M. Brochard, 2012 New phenotypes for new breeding goals in dairy cattle. Animal 6(544): 550.

Buckler, E. S., J. B. Holland, P. J. Bradbury, C. B. Acharya, P. J. Brown *et al.*, 2009 The genetic architecture of maize flowering time. Science 325: 714–718.

Camus-Kulandaivelu, L., and J.-B. Veyrieras, D. madur, V. Combes, M. Fourmann *et al.*, 2006 Maize adaptation to temperate climate: relationship between population structure and polymorphism in the Dwarf8 gene. Genetics 172: 2449–2463.

Černý, V., 1985 Thermodynamical approach to the traveling salesman problem: an efficient simulation algorithm. J. Optim. Theory Appl. 45: 41–51.

Chardon, F., B. Virlon, L. Moreau, M. Falque, J. Joets *et al.*, 2004 Genetic architecture of flowering time in maize as inferred from quantitative trait loci meta-analysis and synteny conservation with the rice genome RID G-3710–2010. Genetics 168: 2169–2185.

Crossa, J., G. de los Campos, P. Perez, D. Gianola, J. Burgueno *et al.*, 2010 Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186: 713–724.

Dekkers, J. C. M., 2007 Prediction of response to marker-assisted and genomic selection using selection index theory. J. Anim. Breed. Genet. 124: 331–341.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto *et al.*, 2011 A robust, simple genotyping-by-sequencing (gbs) approach for high diversity species. PLoS ONE 6: e19379.

Fisher, R. A., 1918 The correlation between relatives on the supposition of Mendelian inheritance. T. Roy. Soc. Edin. 52: 399–433.

Fruchterman, T. M. J., and E. M. Reingold, 1991 Graph drawing by force-directed placement. Softw. Pract. Exper. 21: 1129–1164.

Ganal, M. W., G. Durstewitz, A. Polley, A. Bérard, E. S. Buckler *et al.*, 2011 A large maize (Zea mays L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. PLoS ONE 6: e28334.

Goddard, M., 2009 Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136: 245–257.

Goddard, M., B. Hayes, and T. Meuwissen, 2011 Using the genomic relationship matrix to predict the accuracy of genomic selection. J. Anim. Breed. Genet. 128: 409–421.

Gore, M. A., J.-M. Chia, R. J. Elshire, Q. Sun, E. S. Ersoz *et al.*, 2009 A first-generation haplotype map of maize. Science 326: 1115–1117.

Gower, J. C., 1966 Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53: 325–338.

Habier, D., R. L. Fernando, and J. C. M. Dekkers, 2007 The impact of genetic relationship information on genome-assisted breeding values. Genetics 177: 2389–2397.

Hayes, B., P. Bowman, A. Chamberlain, and M. Goddard, 2009a Invited review: genomic selection in dairy cattle: progress and challenges. J. Dairy Sci. 92: 433–443.

Hayes, B. J., P. M. Visscher, and M. E. Goddard, 2009b Increased accuracy of artificial selection by using the realized relationship matrix. Genet. Res. 91: 47.

Hayes, B. J., P. J. Bowman, A. C. Chamberlain, K. Verbyla, and M. E. Goddard, 2009c Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet. Sel. Evol. 41: 51.

Henderson, C. R., 1984 *Applications of Linear Models in Animal Breeding*. University of Guelph Press, Guelph, Ontario, Canada.

Huang, X., Q. Feng, Q. Qian, Q. Zhao, L. Wang *et al.*, 2009 High-throughput genotyping by whole-genome resequencing. Genome Res. 19: 1068–1076.

Jannink, J. L., A. J. Lorenz, and H. Iwata, 2010 Genomic selection in plant breeding: from theory to practice. Brief. Funct. Genomics 9: 166–177.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi, 1983 Optimization by simulated annealing. Science 220: 671.

Kuehn, L. A., D. R. Notter, G. J. Nieuwhof, and R. M. Lewis, 2007 Changes in connectedness over time in alternative sheep sire referencing schemes. J. Anim. Sci. 86: 536–544.

Laloë, D., 1993 Precision and information in linear models of genetic evaluation. Genet. Sel. Evol. 25: 557–576.

Laloë, D., and F. Phocas, 2003 A proposal of criteria of robustness analysis in genetic evaluation. Livest. Prod. Sci. 80: 241–256.

Laloë, D., F. Phocas, and F. Ménissier, 1996 Considerations on measures of precision and connectedness in mixed linear models of genetic evaluation. Genet. Sel. Evol. 28: 1–20.

Leutenegger, A. L., B. Prum, E. Génin, C. Verny, A. Lemainque *et al.*, 2003 Estimation of the inbreeding coefficient through use of genomic data. Am. J. Hum. Genet. 73: 516–523.

Maenhout, S., B. De Baets, and G. Haesaert, 2010 Graph-based data selection for the construction of genomic prediction models. Genetics 185: 1463–1475.

Metzker, M. L., 2009 Sequencing technologies: the next generation. Nat. Rev. Genet. 11: 31–46.

Meuwissen, T., B. Hayes, and M. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819.

Mikel, M. A., 2006 Availability and analysis of proprietary dent corn inbred lines with expired US plant variety protection. Crop Sci. 46: 2555.

Nei, M., 1978 Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics 89: 583.

Nelson, P. T., N. D. Coles, J. B. Holland, D. M. Bubeck, S. Smith *et al.*, 2008 Molecular characterization of maize inbreds with expired U.S. plant variety protection. Crop Sci. 48: 1673.

Pszczola, M., T. Strabel, H. Mulder, and M. Calus, 2012 Reliability of direct genomic values for animals with different relationships within and to the reference population. J. Dairy Sci. 95: 389–400.

R development Core Team, 2006 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Riedelsheimer, C., A. Czedik-Eysenberg, C. Grieder, J. Lisec, F. Technow *et al.*, 2012 Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat. Genet. 44: 217–220.

Rozenfeld, A. F., S. Arnaud-Haond, E. Hernández-García, V. M. Eguíluz, E. A. Serrão *et al.*, 2008 Network analysis identifies weak and strong links in a metapopulation system. Proc. Natl. Acad. Sci. USA 105: 18824.

SAS Institute, 2008 SAS/STATÒ 9.2 User's Guide. SAS, Cary, NC.

Thomas, M., E. Demeulenaere, J. Dawson, A.R. Khan, N. Galic *et al.*, 2012 On-farm dynamic management of genetic diversity: the impact of seed diffusions and seed saving practices on a population variety of bread wheat. Evol. Appl. (in press).

VanRaden, P., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci. 91: 4414–4423.

Whittaker, J. C., R. Thompson, and M. C. Denham, 2000 Marker-assisted selection using ridge regression. Genet. Res. 75: 249–252.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders *et al.*, 2010 Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42: 565–569.

Zhao, Y., M. Gowda, W. Liu, T. Würschum, H. P. Maurer *et al.*, 2012 Accuracy of genomic selection in European maize elite breeding populations. Theor. Appl. Genet. 124: 769–776.

Zhong, S., J. C. M. Dekkers, R. L. Fernando, and J.-L. Jannink, 2009 Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. Genetics 182: 355–364.

Communicating editor: J. B Holland

General discussion

The technological progresses achieved in the last decades allowed geneticists to go much deeper in the analysis of complex traits. In particular, the increasing availability of molecular markers at low cost makes it easier to decipher the relationship between genotypes and phenotypes. Molecular markers contributed to the cognitive understanding of the traits genetic architecture (with QTL detection, linkage based and GWAS), and to the predictions of genetic values of possibly unphenotyped individuals (genomic predictions). GWAS and GS are based on close statistical models (MEUWISSEN *et al.* 2001; YU *et al.* 2006), but have different objectives (detection versus prediction). A huge number of research programs and papers in plant, animal and human genetics are devoted to these approaches. These two kinds of tools are of considerable interests in plant genetics and in particular in plant breeding. They bring great extensions to the breeders' toolbox, and seem to be promising for crop breeding (BERNARDO and YU 2007; JANNINK *et al.* 2010). However, for these approaches to be efficient, it is necessary to be careful at different methodological critical steps (efficiency being defined as detection power in GWAS and prediction accuracy in GS). The main objective of this thesis was to optimize the use of genotypic information in GWAS and GS to maximize their efficiency. A key step in these approaches is the estimation of kinship between individuals with molecular markers. The kinship matrix is involved in the most commonly used statistical models for GWAS and GS. It permits the control of false positive rate in GWAS, and to infer genetic information from relatives in GS. We studied the parameters affecting power in GWAS and proposed new marker based kinship estimators to increase power and control false positives efficiently (chapter 1). These methods were compared through simulations based on true genotypes (chapter 1) and used to detect QTLs related to flowering time or biomass in two maize diversity panels (chapter 2). In GS, some papers had highlighted the important effect of relationship between the calibration set and the predicted individuals on the accuracies (HABIER *et al.* 2007; HEFFNER *et al.* 2009; ZHONG *et al.* 2009). Considering that phenotyping is likely to remain more limiting than genotyping, we proposed an algorithm based on the genotypic data to optimize the composition of calibration sets. The parameters used in this algorithm were derived from the G-BLUP model and we compared its efficiency to that of more common approaches, based on true datasets (chapter 3). These studies were mostly based on the genotypes and phenotypes collected on two maize diversity panels in the framework of the European project Cornfed described in chapter 2. In this last section we will discuss more globally these three chapters and propose perspectives.

Increasing power in association mapping

1. Kinship estimator

Previous analytical and empirical studies had revealed that GWAS based on panels of intermediate size (hundreds to thousands of individuals) could only capture QTLs of intermediate to big size (LONG and LANGLEY 1999; ZHAO *et al.* 2007). This was confirmed in our simulations based on the Cornfed and Camus-Kulandaivelu (2006) genotypes (chapter 1). It is therefore necessary to optimize power in these designs to detect as many QTLs as possible. Analytical derivation of power revealed that allele frequencies and kinship between individuals could affect power in addition to population size (chapter 1). In practice true kinship is unknown and has thus to be estimated. We could show that the way of estimating it could affect power. Classical kinship estimators such as those proposed by ASTLE and BALDING (2009), VANRADEN (2008), or the simple Identity By State resulted in low power in regions of high LD. This is due to the fact that in these kinship estimators, markers are assumed to be independent so LD is not taken into account. As a consequence, regions with strong LD have a higher contribution in the kinship estimation and are overcorrected. We proposed two alternative ways of estimating kinship and compared their efficiency to detect QTLs through simulations. It revealed that these approaches could control false positive rate efficiently and were more powerful than classical approaches. In particular, the approach consisting in removing the markers physically linked to the tested position from the kinship estimation permitted the detection of more QTLs. This was shown by simulations and for real phenotypes from the Cornfed data, with an increase of about 40% of significant SNPs (chapter 2).

2. Marker density

For GWAS to be efficient, we need a sufficient genotyping density to have at least one marker in high LD with each QTL. From our estimations of LD (0.1 to 0.2cM to reach an r^2K of 0.1 depending on the chromosome and the panel, see Table 2), the 50k SNP-array used in this thesis (GANAL *et al.* 2011) is already a good basis (as shown in BOUCHET *et al.* 2013 for the C-K panel), but additional markers would be highly beneficial, as mentioned in chapter 2. This will be soon achieved by combining SNP-arrays, Genotyping-By-Sequencing and sequencing approaches. Other factors contributing to phenotypic variations such as Copy Number Variants (CNV), or epigenetics variants are also expected to be characterized soon

(CNV-arrays, methylome...) to track more genetic variations. This increase of available molecular polymorphisms leads to an increase in the number of tests, so that significance thresholds need to be adapted accordingly to limit the number of false positives in the detection. We need to consider for this the number of independent tests and not the total number of markers. LD between markers has again to be taken into account for this, as for example two markers in complete LD correspond to only one test. Different ways of estimating this number of independent tests were proposed (CHEVERUD 2001; LI and JI 2005). The approach of Li and Ji estimated around 3600 independent tests in both panels, which is more than 10 times lower than the number of tested SNPs. This equivalent number of independent tests should be revised on a regular basis when considering additional marker information but it is expected that, considering the LD of our panels, it should stabilize at some step before the total number (millions) of polymorphisms is reached.

Our results also illustrate a strong effect of relatedness on LD between distant polymorphisms. This illustrates that relatedness needs to be taken into account in association genetics models to prevent false positives. Even if this was not approached in the thesis, it is interesting to further analyze the local structure and organization of LD in genomic regions. Softwares as Fastphase (SCHEET and STEPHENS 2006) or Clusthaplo (LEROUX *et al.* 2014) were developed to infer local haplotypes. The analysis of the Cornfed Dent panel with Fastphase revealed long haplotypes in regions near centromeres (GIRAUD 2012). This information about ancestry is interesting to consider to detect associations between ancestral haplotypes and phenotypes. It allowed the detection of additional QTLs in multiparental connected populations (BARDOL *et al.* 2013) and in association panels (DUPUIS *et al.* 2011; ZHANG *et al.* 2012, GIRAUD 2012) and would deserve further consideration on the data presented in our study.

3. Population size and diversity

One efficient way of increasing power is to increase the size of the diversity panels. In theory this is possible, but the description of the Cornfed panels in chapters 1 and 2 showed that it may be difficult to sample large numbers of independent individuals among the genetic materials presently available. A few lines (eg. B73, Mo17, Ph207) were intensively used as parents of breeding programs in maize, which generated groups of related individuals (DUBREUIL *et al.* 1996; ROMAY *et al.* 2013), clearly identifiable in the Cornfed panels (chapter 2). Our results in chapter 1 suggest that adding related individuals only leads to marginal

improvement in power. This highlights how important it is to go back to old landrace populations to increase genetic diversity in our material. However, at this step one has to be careful not to sample too distant individuals, which would result in introducing structure in the panel. Another way to increase power may therefore be to develop new lines from parents belonging to different sub-groups, which converges in a way towards multiparental designs (eg., NAM, MAGIC). Optimizing such designs calls for further investigations.

Note also that panels of important size make it difficult to evaluate all the genotypes in a same experimental design (because of a difference in precocity for instance) and call for specific experimental planning. We have to consider in particular that the genotypes need to be evaluated in various environments to estimate the genotype*environment interactions (and more precisely QTL*environment interactions). The instability of most QTLs detected in chapter 2 clearly highlights that this has to be taken into account, as already observed in various studies (MOREAU *et al.* 2004; BOER *et al.* 2007). Considering the high cost of phenotyping in field network and/or platforms, we believe it is of high interest to develop sampling algorithms for optimizing the composition of association mapping panels to maximize their detection power at a given population size. This could be possibly formalized using the analytical study of power developed in chapter 1.

Using molecular information to maximize GS efficiency: optimizing the sampling of the calibration set

Although improvements can be expected from higher marker densities, the reduced number and size of the QTLs identified in this study (chapter 2) illustrates the limits of GWAS for highly polygenic traits. Genomic selection, which estimates all the marker effects simultaneously, allows the breeder to work on a much higher proportion of the genetic variance. It was shown in simulations and on true phenotypes that high prediction accuracies could be reached (JANNINK *et al.* 2010; CROSSA *et al.* 2010; ALBRECHT *et al.* 2011), potentially leading to great genetic progress. This was confirmed on the Cornfed datasets with reliabilities close to 0.8 for flowering time (chapter 3).

The optimal use of GS in the selection schemes depends on the species and on the breeder's strategy, but we believe that most of the cultivated species could benefit from this approach at some step(s). Breeding of long cycle species such as trees could be greatly improved by GS,

which could considerably reduce their breeding cycle, and thus increase genetic progress even with prediction accuracies lower than typical heritabilities. For other species with short breeding cycle, GS can reduce the cycle to a lesser extent but it could also be used to reduce the amount of phenotyped individuals and thus reduce the costs. It is particularly interesting for traits difficult and/or expensive to measure. Another use of GS is to eliminate individuals with poor expected performances at an early step in the breeding program, to focus phenotyping evaluation only on the most promising individuals. In all cases, genetic progress is highly influenced by the prediction accuracies. As shown in a few studies, the calibration of the prediction formula is one of the critical step in GS (HABIER *et al.* 2007; HEFFNER *et al.* 2009; ZHONG *et al.* 2009). We confirmed on the Cornfed datasets that accuracies of the selection candidate predictions are highly influenced by the composition of the Calibration Set (CS). Inadequate CS can potentially lead to accuracies close to zero or even negative which would be disastrous for breeders. On the opposite, optimizing the composition of the CS would allow the breeder to intensify the phenotypic effort on key individuals. We developed an algorithm based on the G-BLUP framework to optimize the composition of the CS in order to maximize prediction accuracies. This algorithm requires the genotypes of all the individuals but no phenotypes. It is based on the expected reliability of the predictions (or generalized CD, LALOË 1993). This algorithm was very efficient in both Cornfed panels for various traits such as flowering time or dry matter yield, for which it gave higher accuracies than random sets in all the considered scenarios. We showed that a same genetic progress could be potentially reached with half of the phenotyping cost when this algorithm is used instead of random sampling. One other potential use of this algorithm would be the optimal sampling of reference individuals to be re-sequenced or densely genotyped for imputing other individuals. As the algorithm samples the most informative individuals with regard to the predicted set, this seems reasonable, but still has to be tested.

Even though GS in panels can have practical applications like prescreening of materials for selecting parents of breeding programs, next steps of breeding programs involve in general families of full or half sibs. Our sampling approach therefore has to be validated on other genetic material with various diversity levels. We applied this approach to more structured populations (multiparental connected populations as commonly used by breeders) in collaboration with J. Crossa (results not shown here). First results on this type of dataset seem encouraging, but need further investigations. One of the major issues is to both define the composition of the optimal CS but also to define its optimal size. Although CD seems

promising, it has to be noted that experimental studies revealed that the use of the phenotypes of distant individuals could decrease prediction accuracy (RIEDELSHEIMER *et al.* 2013). This cannot be explained by the CD, which always increases when additional phenotypes are used. This is because the CD doesn't take into account the fact that distant individuals can bring more noise than information. One possibility to take this into account is to weight the information used to predict GV by considering both the CD and the correlation between the LD phases in the different populations (GIBBS *et al.* 2009, LEGARRA *et al.* in press). Related to this idea, and, similarly to what was done on the alternative kinship estimators in GWAS to optimize power, it may be important to take LD into account in GS models. In the classically used GS models, one assumes that the markers are independent. Introducing a covariance matrix between the markers seems encouraging (CHIQUET *et al.* 2013). Note that the different objectives of GWAS and GS may lead to different ways of taking LD into account to optimize their efficiency. In GWAS, we want to limit confounding between fixed (tested marker) and random (polygenic effect) effects. In GS we want to regularize (i.e. constraint the variation of) the effects attributed to SNPs in an efficient way.

Finally, it should be noticed that considerations above apply well in the context of highly polygenic traits (infinitesimal model). When some QTL have noticeably stronger effects, kinship could be improved for both GWAS and GS by being estimated at the causal genes. This supposes knowing their positions and having markers in complete LD with these genes. This is not possible, but prior knowledge on the genetic architecture could potentially be used to improve kinship estimate. In a Bayeasian framework, this can be achieved by taking into account prior knowledge. An alternative in a mixed model framework is to consider known QTLs as putative fixed effects in the model (BARDOL *et al.* submitted).

Diversity analysis and association mapping in the Dent and Flint Cornfed panels

The different history that the CF-Dent and CF-Flint panels have undergone was highlighted in the diversity analysis (chapter 2). It resulted in different structure, LD extent and phenotypic variability. We could show with simulations (chapter 1) and true datasets (chapter 2) that these characteristics lead to different levels of power in association mapping, CF-Dent being more powerful than CF-Flint. Associations were found for all traits in both panels (Tables S4 and S5). Although promising QTL were detected for biomass yield, most of the strongest associations (around 70%) were found for flowering traits and plant height (Tables 7 and 8).

This suggests different genetic architectures with bigger QTL for flowering time. We believe that this could be explained by the different types of selection that were applied to these traits. Optimal flowering time and plant height depends on the local conditions and breeding strategies, which results in stabilizing selection. The QTL-allele conferring a higher genetic value is not always the same, depending on the breeding strategy and genetic value determined by other QTL. This can maintain polymorphism, even for QTL with strong effects. On the opposite, we suppose that most of the breeding strategies have led to higher biomass productivity (as main breeding objective, or as correlative response of breeding for grain yield). This directional selection may have resulted in the fixation of the favorable alleles, in particular for the strongest QTLs, which would explain why we found less strong associations for biomass traits than for flowering and height traits. Some QTLs for biomass were also associated with flowering traits. In the context of multitrait selection for biomass increase at constant flowering time, they were submitted to a "less directional" selection than QTLs purely related to biomass. If the effect of flowering time is strong enough relative to that on biomass yield, this is expected to prevent fixation at corresponding QTL. In addition to the reduced significance of the detected biomass QTL (chapter 2), one other major limit of using these QTLs in marker assisted selection is their strong instability in the different environments. We believe it asks for more integrated breeding approaches.

Towards an integrated approach in plant breeding

An important challenge in the future of plant breeding is to use GWAS and GS tools within more integrated approaches. If genotyping and sequencing costs continue to decrease, and more importantly if phenotyping relevant with respect to agronomical targets is automatized in some ways (phenotyping platforms, drones...), we can expect to go much deeper in the understanding of biological processes. This would permit for instance the study of interactions between genes and between genes and the environment (epistasis, dominance and qtl*environment interactions). These interactions are now highly simplified in our models or even not considered at all, although they substantially contribute to phenotypic variability, for example through the heterosis phenomenon (SHULL 1914). Considering these interaction effects is an important challenge in plant breeding, because the breeders want to estimate the total genetic value of the selection candidates. Animal breeders are more focused on selecting breeding animals (for reproduction) and as a result select individuals on their additive genetic

value (the breeding value). But in plant breeding, not considering these interactions limits the potential of breeding to some extent for many crops. Some approaches were proposed to study these interactions in the context of GWAS, in particular to detect epistatic interactions (VARGAS *et al.* 2006; BOER *et al.* 2007; LARIEPE *et al.* 2012; MACKAY 2014). In GS, dominance is sometimes introduced in the statistical model (MAENHOUT *et al.* 2009; TECHNOW *et al.* 2012; SU *et al.* 2012), and other studies aimed at predicting genotype by environment interactions (SCHULZ-STREECK *et al.* 2013; HESLOT *et al.* 2014). One other potential progress in integrated breeding would be to take advantage of the information brought by ecophysiological models. One promising way is to include genetic parameters in ecophysiological models and consider these as traits. This would help the breeders predict the specific response of a genotype to given environmental conditions, and thus to develop genotypes adapted to local environments. In the context of climate change, it would also help to develop varieties robust to environmental stresses. The decomposition of integrated traits as yield in more basic traits, would also have the advantage to base the predictions on biological factors and no more on a black-box. We could expect for instance that this would increase the validity of the predictions to more distant individuals (the next generations). Few authors combined QTL detection results to ecophysiological models (REYMOND *et al.* 2003; QUILOT *et al.* 2005; CHENU *et al.* 2009) and could efficiently predict relatively simple traits. This approach applied to more integrated traits, together with the characterization of groups of environments would be highly beneficial to plant breeding.

These perspectives don't question the interest of optimization procedures but rather call for the development of more elaborated algorithms. We have to keep in mind that the phenotyping and genotyping effort will always be limited to some extent, because resources are limited and in competition with other sectors. Other studies are thus required to enrich this field of investigation.

LITERATURE

ALBRECHT T., WIMMER V., AUINGER H.-J., ERBE M., KNAAK C., OUZUNOVA M., SIMIANER H., SCHÖN C.-C., 2011 Genome-based prediction of testcross values in maize.Theor. Appl. Genet. **123**: 339–350.

ASTLE W., BALDING D. J., 2009 Population Structure and Cryptic Relatedness in Genetic Association Studies. Stat. Sci. **24**: 451–471.

BARDOL N., VENTELON M., MANGIN B., JASSON S., LOYWICK V., COUTON F., DERUE C., BLANCHARD P., CHARCOSSET A., MOREAU L., 2013 Combined linkage and linkage disequilibrium QTL mapping in multiple families of maize (Zea mays L.) line crosses highlights complementarities between models based on parental haplotype and single locus polymorphism. Theor. Appl. Genet. **126**: 2717–2736.

BEADLE G. W.,1939 Teosinte and the origin of maize. J. Hered.**30**:245–247

BEAVIS W. D., 1998 QTL analyses: power, precision, and accuracy, pp. 145–162 in Molecular Dissection of Complex Traits, edited by A. H. Paterson. CRC Press, New York.

BELÓ A., ZHENG P., LUCK S., SHEN B., MEYER D. J., LI B., TINGEY S., RAFALSKI A., 2007 Whole genome scan detects an allelic variant of fad2 associated with increased oleic acid levels in maize. Mol. Genet. Genomics **279**: 1–10.

BERNARDO R., YU J., 2007 Prospects for Genomewide Selection for Quantitative Traits in Maize. Crop Sci. **47**: 1082.

BOER M. P., WRIGHT D., FENG L., PODLICH D. W., LUO L., COOPER M., EEUWIJK F. A. van, 2007a A Mixed-Model Quantitative Trait Loci (QTL) Analysis for Multiple-Environment Trial Data Using Environmental Covariables for QTL-by-Environment Interactions, With an Example in Maize. Genetics **177**: 1801–1813.

BOER M. P., WRIGHT D., FENG L., PODLICH D. W., LUO L., COOPER M., EEUWIJK F. A. van, 2007b A Mixed-Model Quantitative Trait Loci (QTL) Analysis for Multiple-Environment Trial Data Using Environmental Covariables for QTL-by-Environment Interactions, With an Example in Maize. Genetics **177**: 1801–1813.

BOUCHET S., SERVIN B., BERTIN P., MADUR D., COMBES V., DUMAS F., BRUNEL D., LABORDE J., CHARCOSSET A., NICOLAS S., 2013 Adaptation of Maize to Temperate Climates: Mid-Density Genome-Wide Association Genetics and Diversity Patterns Reveal Key Genomic Regions, with a Major Contribution of the Vgt2 (ZCN8) Locus. PLOS ONE **8**: e71377.

BUCKLER E. S., HOLLAND J. B., BRADBURY P. J., ACHARYA C. B., BROWN P. J., BROWNE C., ERSOZ E., FLINT-GARCIA S., GARCIA A., GLAUBITZ J. C., GOODMAN M. M., HARJES C., GUILL K., KROON D. E., LARSSON S., LEPAK N. K., LI H., MITCHELL S. E., PRESSOIR G., PEIFFER J. A., ROSAS M. O., ROCHEFORD T. R., ROMAY M. C., ROMERO S., SALVO S., VILLEDA H. S., SOFIA DA SILVA H., SUN Q., TIAN F., UPADYAYULA N., WARE D., YATES H., YU J., ZHANG Z., KRESOVICH S., MCMULLEN M. D., 2009 The Genetic Architecture of Maize Flowering Time. Science **325**: 714–718.

CAMUS-KULANDAIVELU L., 2006 Maize Adaptation to Temperate Climate: Relationship Between Population Structure and Polymorphism in the Dwarf8 Gene. Genetics **172**: 2449–2463.

CAVANAGH C., MORELL M., MACKAY I., POWELL W., 2008 From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. Curr.Opin. Plant Biol. **11**: 215–221.

CHENU K., CHAPMAN S. C., TARDIEU F., MCLEAN G., WELCKER C., HAMMER G. L., 2009 Simulating the Yield Impacts of Organ-Level Quantitative Trait Loci Associated With Drought Response in Maize: A “Gene-to-Phenotype” Modeling Approach. Genetics **183**: 1507–1523.

CHEVERUD J. M., 2001 A simple correction for multiple comparisons in interval mapping genome scans. Heredity **87**: 52–58.

CHIQUET J., T. MARY-HUARD and S. ROBIN 2013 Multi-trait genomic selection via multivariate regression with structured regularization.Proceedings of MLCB/NIPS'13 workshop.

BOVINE HAPMAP CONSORTIUM, 2009 Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. Science **324**: 528–532.

CROSSA J., CAMPOS G. d. l., PEREZ P., GIANOLA D., BURGUENO J., ARAUS J. L., MAKUMBI D.,

SINGH R. P., DREISIGACKER S., YAN J., ARIEF V., BANZIGER M., BRAUN H.-J., 2010 Prediction of Genetic Values of Quantitative Traits in Plant Breeding Using Pedigree and Molecular Markers. Genetics **186**: 713–724.

DARWIN C., 1859 On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life (1st ed.). London: J. Murray. ISBN 1-4353-9386-4.

DOEBLEY J.F. 2004 The genetics of maize evolution. Annu Rev Genet 38:37–59

DUBREUIL P., P. DUFOUR, E. KREJCI, M. CAUSSE, D. DE VIENNE, A. GALLAIS, A. CHARCOSSET 1996 Organization of RFLP diversity among inbred lines of Maize representing the most significant heterotic groups. Crop Sci. **36**: 790–799.

DUCROCQ S., GIAUFFRET C., MADUR D., COMBES V., DUMAS F., JOUANNE S., COUBRICHE D., JAMIN P., MOREAU L., CHARCOSSET A., 2009 Fine Mapping and Haplotype Structure Analysis of a Major Flowering Time Quantitative Trait Locus on Maize Chromosome 10. Genetics **183**: 1555–1563.

DUPUIS M.-C., ZHANG Z., DRUET T., DENOIX J.-M., CHARLIER C., LEKEUX P., GEORGES M., 2011 Results of a haplotype-based GWAS for recurrent laryngeal neuropathy in the horse.Mamm. Genome **22**: 613–620.

ELSHIRE R. J., GLAUBITZ J. C., SUN Q., POLAND J. A., KAWAMOTO K., BUCKLER E. S., MITCHELL S. E., 2011 A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species (L Orban, Ed.). PLoS ONE **6**: e19379.

EWENS W., SPIELMAN R., 1995 The Transmission Disequilibrium Test - History, Subdivision and Admixture. Am. J. Hum. Genet.**57**: 455–464.

FALCONER D. S., MACKAY T. F. C.1996 Introduction to Quantitative Genetics.4th edition. Longman, New York.

FISHER R.A. 1918 The Correlation between Relatives on the Supposition of Mendelian Inheritance Philosophical Transactions of the Royal Society of Edinburgh 52:399–433.

FLINT-GARCIA S. A., THORNSBERRY J. M., BUCKLER E. S., 2003 Structure of linkage disequilibrium in plants. Annu. Rev. Plant Biol. **54**: 357–374.

GALLAIS A. 1990 Théorie de la sélection en amélioration des plantes. Edition Masson.

GALTON F., 1869 Hereditary genius (reprinted 1962, Meridian books, NY).

GALTON F., 1879 The geometric mean in vital and social statistics. Proc. Royal Soc. Lond.**29**:365-367.

GANAL M. W., DURSTEWITZ G., POLLEY A., BÉRARD A., BUCKLER E. S., *et al.*, 2011 A Large Maize (Zea mays L.) SNP Genotyping Array: Development and Germplasm Genotyping, and Genetic Mapping to Compare with the B73 Reference Genome (L Lukens, Ed.). PLoS ONE **6**: e28334.

GIRAUD H., 2012, Modélisation haplotypique : comparaison d'approches, application au maïs et à l'analyse de caractères quantitatifs. Mémoire de fin d'études option APIMET, Montpellier SupAgro.

GIULIANI S., SANGUINETI M. C., TUBEROSA R., BELLOTTI M., SALVI S., LANDI P., 2005 Root-ABA1, a major constitutive QTL, affects maize root architecture and leaf ABA concentration at different water regimes. J. Exp. Bot. **56**: 3061–3070.

GODDARD M., 2009 Genomic selection: prediction of accuracy and maximisation of long term response. Genetica **136**: 245–257.

HABIER D., FERNANDO R. L., DEKKERS J. C. M., 2007 The impact of genetic relationship information on genome-assisted breeding values. Genetics **177**: 2389–2397.

HABIER D., TETENS J., SEEFRIED F.-R., LICHTNER P., THALLER G., 2010 The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet. Sel. Evol. **42**: 5.

HAYES B., BOWMAN P., CHAMBERLAIN A., GODDARD M., 2009a Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. **92**: 433–443.

HAYES B. J., VISSCHER P. M., GODDARD M. E., 2009b Increased accuracy of artificial selection by using the realized relationship matrix.Genet. Res. **91**: 47.

HEFFNER E. L., SORRELLS M. E., JANNINK J.-L., 2009 Genomic Selection for Crop Improvement. Crop Sci. **49**: 1.

HENDERSON C.R., 1963 Selection index and expected genetic advance. In W. D. Hanson and H. F. Robinson (Ed.).Statistical Genetics and Plant Breeding.National Academy of Sciences-National Research Council, Washington, DC, Pub. 982.

HESLOT N., AKDEMIR D., SORRELLS M. E., JANNINK J.-L., 2014 Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. TAG Theor. Appl. Genet. Theor.Angew.Genet.**127**: 463–480.

HESLOT N., JANNINK J.-L., SORRELLS M. E., 2013 Using Genomic Prediction to Characterize Environments and Optimize Prediction Accuracy in Applied Breeding Data. Crop Sci. **53**: 921.

HESLOT N., YANG H.-P., SORRELLS M. E., JANNINK J.-L., 2012 Genomic Selection in Plant Breeding: A Comparison of Models. Crop Sci. **52**: 146.

HOFHEINZ N., BORCHARDT D., WEISSLEDER K., FRISCH M., 2012 Genome-based prediction of test cross performance in two subsequent breeding cycles.Theor. Appl. Genet. **125**: 1639–1645.

HOSPITAL F. and A. CHARCOSSET, 1997 Marker-assisted introgression of quantitative trait loci. Genetics **147**: 1469–1485.

HOSPITAL, F., L. MOREAU, F. LACOUDRE, A. CHARCOSSET and A. GALLAIS, 1997 More on the efficiency of marker-assisted selection.Theor. Appl. Genet. **95**: 1181–1189.

HUANG N., PARCO A., MEW T., MAGPANTAY G., MCCOUCH S., GUIDERDONI E., XU J. C., SUBUDHI P., ANGELES E. R., KHUSH G. S., 1997 RFLP mapping of isozymes, RAPD and QTLs for grain shape, brown planthopper resistance in a doubled haploid rice population. Mol. Breed. **3**: 105–113.

HUXLEY J., 1942 Evolution: The Modern Synthesis. London: Allen and Unwin.

JANNINK J. L., 2010 Dynamics of long-term genomic selection. Genet. Sel. Evol. **42**: 35.

JANNINK J. L., LORENZ A. J., IWATA H., 2010 Genomic selection in plant breeding: from theory to practice. Brief. Funct. Genomics **9**: 166–177.

JONES P., CHASE K., MARTIN A., DAVERN P., OSTRANDER E. A., LARK K. G., 2008 Single-

Nucleotide-Polymorphism-Based Association Mapping of Dog Stereotypes. Genetics **179**: 1033–1044.

LALOË D., 1993 Precision and information in linear models of genetic evaluation.Genet. Sel. Evol. **25**: 557–576.

LARIÈPE A., MANGIN B., JASSON S., COMBES V., DUMAS F., JAMIN P., LARIAGON C., JOLIVOT D., MADUR D., FIÉVET J., GALLAIS A., DUBREUIL P., CHARCOSSET A., MOREAU L., 2012 The Genetic Basis of Heterosis: Multiparental Quantitative Trait Loci Mapping Reveals Contrasted Levels of Apparent Overdominance Among Traits of Agronomical Interest in Maize (Zea mays L.). Genetics **190**: 795–811.

LARSSON S. J., LIPKA A. E., BUCKLER E. S., 2013 Lessons from Dwarf8 on the Strengths and Weaknesses of Structured Association Mapping (JK Pritchard, Ed.). PLoS Genet. **9**: e1003246.

LAURIE C. C., 2004 The Genetic Architecture of Response to Long-Term Artificial Selection for Oil Concentration in the Maize Kernel. Genetics **168**: 2141–2155.

LEGARRA A., G. BALOCHE, F.BARILLET, J.M. ASTRUC, C. SOULAS, *et al.*(in press) Multiple breed genomic evaluation for Pyrenees dairy sheep Within and across-breed genomic predictions and genomic relationships for Western Pyrenees dairy sheep breeds Latxa, Manech and Basco-Bearnaise. Genomia (in press).

LI J., JI L., 2005 Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity **95**: 221–227.

LOISELLE B.A., SORK V.L., NASON J., GRAHAM C., 1995 Spatial genetic structure of a tropical understory shrub Psychotria officinalis (Rubiaceae). Am. J. Bot. **82**(11):1420–1425.

LONG A. D., LANGLEY C. H., 1999 The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Res. **9**: 720–731.

LY D., HAMBLIN M., RABBI I., MELAKU G., BAKARE M., GAUCH H. G., OKECHUKWU R., DIXON A. G. O., KULAKOW P., JANNINK J.-L., 2013 Relatedness and Genotype × Environment Interaction Affect Prediction Accuracies in Genomic Selection: A Study

in Cassava. Crop Sci. **53**: 1312.

MACKAY T. F. C., 2014 Epistasis and quantitative traits: using model organisms to study gene-gene interactions. Nat. Rev. Genet. **15**: 22–33.

MAENHOUT S., BAETS B. DE, HAESAERT G., 2009 Prediction of maize single-cross hybrid performance: support vector machine regression versus best linear prediction. Theor. Appl. Genet. **120**: 415–427.

MAHER, B., 2008 The case of the missing heritability. Nature **456**:18–21.

MANGIN B., SIBERCHICOT A., NICOLAS S., DOLIGEZ A., THIS P., CIERCO-AYROLLES C., 2012 Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. Heredity **108**: 285–291.

MATSUOKA Y., Y. VIGOUROUX, M.M. GOODMAN, J. SANCHEZ G., E. BUCKLER, J. DOEBLEY, 2002 A single domestication for maize shown by multilocus microsatellite genotyping Proc. Natl. Acad. Sci. USA, **99**:6080–6084

MENDEL J.G.,1866 Versuche über Pflanzenhybriden Verhandlungen des naturforschenden Vereines in Brünn, Bd. IV für das Jahr, 1865 Abhandlungen:3–47.

MEUWISSEN T., HAYES B., GODDARD M., 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics **157**: 1819.

MOREAU L., CHARCOSSET A., GALLAIS A., 2004 Use of trial clustering to study QTL × environment effects for grain yield and related traits in maize. Theor. Appl. Genet. **110**: 92–105.

OZAKI K., OHNISHI Y., IIDA A., SEKINE A., YAMADA R., TSUNODA T., SATO H., SATO H., HORI M., NAKAMURA Y., TANAKA T., 2002 Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. Nat. Genet. **32**: 650–654.

PRITCHARD J. K., STEPHENS M., DONNELLY P., 2000 Inference of population structure using multilocus genotype data. Genetics **155**: 945.

QUILOT B., KERVELLA J., GÉNARD M., LESCOURRET F., 2005 Analysing the genetic control

of peach fruit quality through an ecophysiological model combined with a QTL approach. J. Exp. Bot. **56**: 3083–3092.

RANDHAWA H. S., MUTTI J. S., KIDWELL K., MORRIS C. F., CHEN X., GILL K. S., 2009 Rapid and Targeted Introgression of Genes into Popular Wheat Cultivars Using Marker-Assisted Background Selection (BP Dilkes, Ed.). PLoS ONE **4**: e5752.

RESENDE M. F. R., MUNOZ P., RESENDE M. D. V., GARRICK D. J., FERNANDO R. L., DAVIS J. M., JOKELA E. J., MARTIN T. A., PETER G. F., KIRST M., 2012 Accuracy of Genomic Selection Methods in a Standard Data Set of Loblolly Pine (Pinus taeda L.). Genetics **190**: 1503–1510.

REYMOND M., MULLER B., LEONARDI A., CHARCOSSET A., TARDIEU F., 2003 Combining Quantitative Trait Loci Analysis and an Ecophysiological Model to Analyze the Genetic Variability of the Responses of Maize Leaf Growth to Temperature and Water Deficit. Plant Physiol. **131**: 664–675.

RIAR A. K., KAUR S., DHALIWAL H. S., SINGH K., CHHUNEJA P., 2012 Introgression of a leaf rust resistance gene from Aegilops caudata to bread wheat. J. Genet. **91**: 155–161.

RIEDELSHEIMER C., ENDELMAN J. B., STANGE M., SORRELLS M. E., JANNINK J.-L., MELCHINGER A. E., 2013 Genomic Predictability of Interconnected Biparental Maize Populations. Genetics **194**: 493–503.

RITLAND K., 1996 Estimators for pairwise relatedness and individual inbreeding coefficients. Genet Res. (Camb) **67**:175–185.

SALVI S., TUBEROSA R., 2005 To clone or not to clone plant QTLs: present and future challenges. Trends Plant Sci. **10**: 297–304.

SANZ-ALFEREZ S., RICHTER T. E., HULBERT S. H., BENNETZEN J. L., 1995 The Rp3 disease resistance gene of maize: Mapping and characterization of introgressed alleles. Theor. Appl. Genet. **91**: 25–32.

SAX K., 1923 The association of size differences with seed coat pattern and pigmentation in Phaseolus vulgaris. Genetics 8: 552-560.

SCHEET P., STEPHENS M., 2006 A Fast and Flexible Statistical Model for Large-Scale

Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. Am. J. Hum. Genet.**78**: 629–644.

SCHULZ-STREECK T., OGUTU J. O., GORDILLO A., KARAMAN Z., KNAAK C., PIEPHO H.-P., 2013 Genomic selection allowing for marker-by-environment interaction. Plant Breed.**132**: 532–538.

SERVIN B., 2004 Toward a Theory of Marker-Assisted Gene Pyramiding. Genetics **168**: 513–523.

SHULL, G. H., 1914 Duplicate genes for capsule-form in Bursa pastoris. Zeitschrift ind. Abst. u. Verebsgl. **12**: 97-149.

SU G., CHRISTENSEN O. F., OSTERSEN T., HENRYON M., LUND M. S., 2012 Estimating Additive and Non-Additive Genetic Variances and Predicting Genetic Merits Using Genome-Wide Dense Single Nucleotide Polymorphism Markers (AA Palmer, Ed.). PLoS ONE **7**: e45293.

TECHNOW F., RIEDELSHEIMER C., SCHRAG T. A., MELCHINGER A. E., 2012 Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. Theor. Appl. Genet. **125**: 1181–1194.

THABUIS A., PALLOIX A., SERVIN B., DAUBEZE A. M., SIGNORET P., LEFEBVRE V., 2004 Marker-assisted introgression of 4 Phytophthora capsici resistance QTL alleles into a bell pepper line: validation of additive and epistatic effects. Mol. Breed. **14**: 9–20.

THORNSBERRY J. M., GOODMAN M. M., DOEBLEY J., KRESOVICH S., NIELSEN D., BUCKLER E. S., 2001 Dwarf8 polymorphisms associate with variation in flowering time. Nat. Genet. **28**: 286–289.

TRUNTZLER M., BARRIÈRE Y., SAWKINS M. C., LESPINASSE D., BETRAN J., CHARCOSSET A., MOREAU L., 2010 Meta-analysis of QTL involved in silage quality of maize and comparison with the position of candidate genes. Theor. Appl. Genet. **121**: 1465–1482.

VANRADEN P., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci. **91**: 4414–4423.

VARGAS M., EEUWIJK F. A. van, CROSSA J., RIBAUT J.-M., 2006 Mapping QTLs and QTL ×

environment interaction for CIMMYT maize drought stress program using factorial regression and partial least squares methods. Theor. Appl. Genet. **112**: 1009–1023.

VISSCHER P. M., 2008 Sizing up human height variation. Nat. Genet. **40**: 489–490.

WEI F., STEIN J.C., LIANG C., ZHANG J., FULTON R.S., *et al.*, 2009a Detailed Analysis of a Contiguous 22-Mb Region of the Maize Genome. PLoS Genet 5: e1000728.

WEI F., ZHANG J., ZHOU S, HE R., SCHAEFFER M., *et al.*, 2009b The Physical and Genetic Framework of the Maize B73 Genome. PLoS Genet 5: e1000715.

WHITTAKER J. C., THOMPSON R., DENHAM M. C., 2000 Marker-assisted selection using ridge regression. Genet. Res. **75**: 249–252.

WINDHAUSEN V. S., ATLIN G. N., HICKEY J. M., CROSSA J., JANNINK J.-L., SORRELLS M. E., RAMAN B., CAIRNS J. E., TAREKEGNE A., SEMAGN K., BEYENE Y., GRUDLOYMA P., TECHNOW F., RIEDELSHEIMER C., MELCHINGER A. E., 2012 Effectiveness of Genomic Prediction of Maize Hybrid Performance in Different Breeding Populations and Environments. G358 Genes Genomes Genetics **2**: 1427–1436.

YANG J., BENYAMIN B., MCEVOY B. P., GORDON S., HENDERS A. K., NYHOLT D. R., MADDEN P. A., HEATH A. C., MARTIN N. G., MONTGOMERY G. W., GODDARD M. E., VISSCHER P. M., 2010 Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. **42**: 565–569.

YU J., HOLLAND J. B., MCMULLEN M. D., BUCKLER E. S., 2008 Genetic Design and Statistical Power of Nested Association Mapping in Maize. Genetics **178**: 539–551.

YU J., PRESSOIR G., BRIGGS W. H., VROH BI I., YAMASAKI M., DOEBLEY J. F., MCMULLEN M. D., GAUT B. S., NIELSEN D. M., HOLLAND J. B., KRESOVICH S., BUCKLER E. S., 2006 A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. **38**: 203–208.

ZHANG Z., GUILLAUME F., SARTELET A., CHARLIER C., GEORGES M., FARNIR F., DRUET T., 2012 Ancestral haplotype-based association mapping with generalized linear mixed models accounting for stratification. Bioinformatics **28**: 2467–2473.

ZHAO K., ARANZANA M. J., KIM S., LISTER C., SHINDO C., TANG C., TOOMAJIAN C., ZHENG

H., DEAN C., MARJORAM P., NORDBORG M., 2007 An Arabidopsis Example of Association Mapping in Structured Samples. PLoS Genet **3**: e4.

ZHAO Y., GOWDA M., LIU W., WÜRSCHUM T., MAURER H. P., LONGIN F. H., RANC N., REIF J. C., 2011 Accuracy of genomic selection in European maize elite breeding populations. Theor. Appl. Genet. **124**: 769–776.

ZHONG S., DEKKERS J. C. M., FERNANDO R. L., JANNINK J.-L., 2009 Factors Affecting Accuracy From Genomic Selection in Populations Derived From Multiple Inbred Lines: A Barley Case Study. Genetics **182**: 355–364.

ZHOU S., WEI F., NGUYEN J., BECHNER M., POTAMOUSIS K., *et al.*, 2009 A Single Molecule Scaffold for the Maize Genome. PLoS Genet 5: e1000711.

ZHU C., GORE M., BUCKLER E. S., YU J., 2008 Status and Prospects of Association Mapping in Plants. Plant Genome J. **1**: 5.

APPENDICES

Appendix I: supplemental chapter 1

Another simulation approach was used to compare the ability of the different models to detect QTLs. The genetic model was simulated as in the second step of simulations presented in the paper (the QTLs were sampled among all the PANZEA SNPs) but considering now that markers within a given genetic distance of a QTL were under H1 and the others under H0. We considered genetic distances of 1, 2, 3, 5 and 10 cM. For each genetic model (50 or 100 QTLs) and each panel, 200 runs were used to estimate the proportion of QTLs (PowerQTL), and the proportion of H1-markers (Power) declared significant at a realized FDR of 0.1. The realized false discovery rate (FDR) was defined as the proportion of markers under H0 among the markers declared significant. To estimate PowerQTL, we considered that a QTL was detected when at least one of the corresponding H1-markers had a significant Pvalue.

Table S1 Power of the QTL detections with $\mathcal{M}_{K\_Freq}$, $\mathcal{M}_{K\_Chr}$, and $\mathcal{M}_{K\_LD}$ at a realized FDR of 0.1. PowerQTL is the proportion of QTL discovered, Power is the proportion of H1-markers discovered.

			PowerQTL			Power		
	Nb QTLs	Window (cM)	M_{K_Freq}	M_{K_Chr}	M_{K_LD}	M_{K_Freq}	M_{K_Chr}	M_{K_LD}
C-K	50	1	0.08	0.11	0.10	0.0012	0.0028	0.0025
	50	2	0.11	0.14	0.13	0.0010	0.0024	0.0021
	50	3	0.12	0.16	0.15	0.0009	0.0021	0.0019
	50	5	0.15	0.21	0.19	0.0008	0.0019	0.0017
	50	10	0.24	0.32	0.29	0.0008	0.0019	0.0016
	100	1	0.03	0.05	0.04	0.0004	0.0011	0.0008
	100	2	0.05	0.07	0.06	0.0004	0.0010	0.0008
	100	3	0.06	0.10	0.08	0.0004	0.0010	0.0008
	100	5	0.09	0.15	0.13	0.0004	0.0011	0.0009
	100	10	0.21	0.32	0.27	0.0006	0.0017	0.0013
CF-Dent	50	1	0.09	0.12	0.11	0.0019	0.0052	0.0041
	50	2	0.11	0.17	0.15	0.0015	0.0052	0.0038
	50	3	0.13	0.21	0.19	0.0014	0.0054	0.0038
	50	5	0.17	0.28	0.26	0.0013	0.0053	0.0036
	50	10	0.26	0.46	0.40	0.0014	0.0065	0.0037
	100	1	0.04	0.07	0.06	0.0007	0.0030	0.0020
	100	2	0.05	0.12	0.09	0.0006	0.0032	0.0019
	100	3	0.07	0.17	0.12	0.0006	0.0036	0.0019
	100	5	0.11	0.26	0.19	0.0007	0.0045	0.0022
	100	10	0.24	0.54	0.42	0.0011	0.0081	0.0039
CF-Flint	50	1	0.09	0.10	0.09	0.0014	0.0026	0.0023
	50	2	0.11	0.14	0.12	0.0012	0.0023	0.0019
	50	3	0.13	0.17	0.15	0.0010	0.0022	0.0018
	50	5	0.16	0.22	0.19	0.0010	0.0022	0.0017
	50	10	0.25	0.35	0.30	0.0010	0.0024	0.0016
	100	1	0.03	0.05	0.04	0.0005	0.0013	0.0010
	100	2	0.05	0.08	0.06	0.0004	0.0013	0.0010
	100	3	0.06	0.10	0.08	0.0004	0.0013	0.0009
	100	5	0.09	0.16	0.13	0.0005	0.0015	0.0010
	100	10	0.18	0.34	0.27	0.0006	0.0023	0.0014

Appendix II: supplemental chapter 2

Table S1: Statistics on the hybrid and per se adjusted means in the CF-Dent and CF-Flint panels.

			Tass_GDD6	Silk_GDD6	ASI_GDD6	PLHT	DMC	DMCcorr	DMY	DMYcorr
CF-Dent	Hybrids	mean	888.1	906.1	18.1	255.8	34.0	0.0	16.0	0.0
		min	816.6	827.5	-3.9	225.2	26.0	-4.0	11.6	-3.8
		max	995.3	1008.2	42.7	286.8	40.9	4.2	19.7	3.0
		var	1165.0	1351.2	72.4	153.9	8.9	2.3	1.9	1.5
	Per se	mean	876.8	883.2	6.5					
		min	662.2	662.2	-76.9					
		max	1070.9	1115.7	156.8					
		var	7964.1	8776.8	634.5					
CF-Flint	Hybrids	mean	882.7	913.9	23.3	254.5	31.3	0.0	15.1	0.0
		min	803.4	841.4	1.6	217.3	24.3	-4.4	11.1	-4.2
		max	1034.7	1056.8	48.0	296.5	37.3	4.4	19.6	4.0
		var	1593.3	1535.4	68.0	237.9	5.9	2.5	2.3	1.9
	Per se	mean	1003.0	968.6	40.0	116.9				
		min	784.1	809.2	-80.3	53.6				
		max	1308.5	1305.0	189.9	351.4				
		var	11576.7	9096.2	1741.0	772.0				

Table S2: Admixture and performances of the Dent lines (crossed to the tester)

Accession	StiffStalk B73	Lancaster Mo17	UH_4068family	Iodent	StiffStalk B14	Minnesota13	Lancaster Oh43	F252family	Tass_GDD6	Silk_GDD6	ASI_GDD6	DMC	DMY	PLHT	DMCcorr	DMYcorr
B109_uh	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	915	931	16	31	19	275	-1,0	2,5
EC133A_ciam	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	886	894	4	34	17	258	-0,9	1,1
Lp5_usda	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	892	903	10	32	17	265	-1,8	1,3
B73_inra	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	909	920	14	31	17	271	-1,9	1,4
NS701_usda	0,81	0,00	0,00	0,00	0,19	0,00	0,00	0,00	919	922	2	33	17	276	-0,4	1,1
N192_inra	0,79	0,00	0,06	0,00	0,14	0,00	0,00	0,00	894	904	12	32	16	257	-1,3	0,0
NK764_usda	0,70	0,00	0,00	0,00	0,30	0,00	0,00	0,00	884	903	18	34	16	252	0,8	0,5
PHG86_usda	0,65	0,00	0,03	0,00	0,29	0,00	0,00	0,03	922	918	-2	32	18	271	-1,5	2,2
F7025_inra	0,64	0,00	0,00	0,00	0,00	0,00	0,00	0,36	900	926	24	33	17	266	0,1	1,0
EC136_ciam	0,63	0,00	0,01	0,01	0,00	0,27	0,04	0,05	873	889	16	34	15	252	-0,8	-1,2
DK78010_usda	0,61	0,00	0,00	0,00	0,39	0,00	0,00	0,00	898	900	7	34	17	255	-0,2	1,0
LH74_inra	0,59	0,00	0,00	0,00	0,41	0,00	0,00	0,00	939	931	-4	33	17	255	0,9	0,5
B84_inra	0,44	0,00	0,00	0,00	0,36	0,18	0,00	0,00	925	950	20	30	19	266	-1,1	2,2
EC326A_ciam	0,43	0,12	0,00	0,01	0,23	0,09	0,13	0,00	870	881	10	36	17	264	0,0	0,9
B104_inra	0,43	0,00	0,00	0,03	0,24	0,28	0,00	0,02	932	962	27	28	16	266	-2,4	-0,2
B110_uh	0,41	0,01	0,04	0,07	0,22	0,25	0,00	0,00	925	949	22	30	18	265	-1,1	1,7
F924_inra	0,31	0,10	0,02	0,02	0,24	0,25	0,00	0,05	908	914	6	33	17	267	-0,9	0,9
DKMBNA_usda	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	909	937	24	31	17	276	-0,9	0,8
F748_inra	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	917	924	7	33	18	270	0,4	2,0
LH65_usda	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	921	957	34	30	16	268	-1,0	-0,3
NC262B_uh	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	938	969	30	30	19	270	-1,0	1,9
Mo17_inra	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	922	962	37	30	16	272	-1,5	-0,4
NC290_uh	0,00	0,98	0,02	0,00	0,00	0,00	0,00	0,00	908	941	29	31	18	271	-1,9	1,4
NC258_usa	0,00	0,83	0,00	0,02	0,04	0,06	0,05	0,00	995	1005	9	28	20	269	-0,3	2,4
CR1Ht_usda	0,00	0,79	0,00	0,03	0,00	0,13	0,00	0,06	872	892	19	32	16	250	-1,7	0,0
LH59_usda	0,00	0,74	0,00	0,00	0,00	0,00	0,26	0,00	912	935	19	31	15	257	-1,7	-1,4
AS5707_usda	0,00	0,73	0,03	0,02	0,05	0,17	0,00	0,00	934	955	20	31	16	270	-1,0	-0,8
DKMDF-13D_USDA	0,00	0,69	0,00	0,00	0,00	0,00	0,31	0,00	956	966	12	32	18	278	1,0	1,0
LH60_usda	0,00	0,67	0,03	0,02	0,00	0,24	0,00	0,03	928	957	25	31	16	269	-0,6	-0,2
F816_inra	0,03	0,67	0,00	0,06	0,00	0,11	0,00	0,13	864	885	23	34	16	268	-1,5	0,3
PHK76_usda	0,00	0,64	0,02	0,02	0,09	0,00	0,23	0,00	910	919	11	33	16	269	-0,4	0,3
PHJ40_usda	0,04	0,47	0,00	0,03	0,21	0,14	0,06	0,04	878	893	13	35	16	248	0,4	-0,2
EP72_csic	0,00	0,43	0,05	0,04	0,11	0,18	0,08	0,10	922	945	19	30	18	265	-2,1	1,7
LAN496_inra	0,00	0,40	0,02	0,03	0,08	0,38	0,01	0,07	915	950	31	32	15	250	0,3	-1,5
B106_inra	0,01	0,39	0,06	0,06	0,02	0,34	0,03	0,09	933	963	26	31	17	271	-0,5	0,0
W602S_uh	0,07	0,36	0,10	0,05	0,02	0,19	0,16	0,05	891	936	43	32	17	254	-0,2	0,4
W604S_uh	0,06	0,34	0,03	0,01	0,06	0,16	0,29	0,04	932	946	15	31	17	259	-0,7	0,1
PHT77_usda	0,00	0,33	0,00	0,03	0,03	0,29	0,30	0,02	918	936	19	33	18	272	0,7	1,5
UH250_uh	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	877	883	11	37	15	248	2,0	-0,4
UH_P036_uh	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	848	852	5	38	16	258	1,1	0,4
UH_S033_uh	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	865	881	17	37	17	257	2,1	1,4
UH_S036_uh	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	860	877	15	38	16	249	2,0	0,6
UH_S040_uh	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	847	857	9	39	16	255	1,6	0,3
D06_uh	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	836	853	21	39	16	245	2,5	0,8
UH_P034_uh	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	863	878	16	37	18	267	1,4	1,8
UH_4068_uh	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	860	871	9	37	16	263	0,7	0,5
UH_S002_uh	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	866	881	16	36	15	251	0,8	-1,3
D09_uh	0,00	0,00	0,81	0,00	0,00	0,00	0,00	0,19	853	868	16	37	16	258	1,0	0,1
UH_P072_uh	0,00	0,01	0,76	0,21	0,01	0,00	0,00	0,00	873	880	9	38	17	244	2,5	1,1
UH_P017_uh	0,00	0,00	0,63	0,37	0,00	0,00	0,00	0,00	857	875	19	38	16	251	2,2	0,4
UH_P075_uh	0,00	0,00	0,62	0,38	0,00	0,00	0,00	0,00	864	880	15	37	17	251	1,1	1,3
UH_P046_uh	0,00	0,00	0,58	0,42	0,00	0,00	0,00	0,00	836	850	14	39	15	247	1,3	-0,5
UH_P087_uh	0,00	0,00	0,54	0,46	0,00	0,00	0,00	0,00	876	884	9	37	16	263	1,9	0,5
UH_P060_uh	0,00	0,00	0,50	0,50	0,00	0,00	0,00	0,00	877	892	17	37	17	255	2,6	1,3
UH_P104_uh	0,07	0,00	0,50	0,26	0,09	0,08	0,00	0,00	877	886	13	37	14	247	1,5	-1,5

UH_S015_uh	0,08	0,00	0,38	0,02	0,12	0,30	0,07	0,02	863	881	21	36	17	242	1,3	0,9
PH207_usda	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	886	913	24	34	16	253	-0,1	-0,1
PHH93_usda	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	891	907	17	35	16	262	1,0	0,0
UH_6173_uh	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	851	885	33	38	16	250	2,5	0,1
UH_P001_uh	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	863	887	22	39	16	259	3,8	0,4
UH_P024_uh	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	869	885	17	36	17	258	0,3	1,5
UH_P038_uh	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	848	861	16	36	16	256	-0,3	0,7
UH_P048_uh	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	877	905	34	36	12	236	1,5	-3,8
UH_P006_uh	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	853	864	11	37	16	255	0,4	0,3
UH304_uh	0,00	0,00	0,00	0,98	0,00	0,00	0,00	0,02	877	890	13	35	16	264	0,3	0,5
UH_6148_uh	0,00	0,00	0,00	0,95	0,05	0,00	0,00	0,00	859	877	18	37	16	253	0,9	0,1
UH_P033_uh	0,00	0,00	0,00	0,95	0,00	0,00	0,00	0,05	867	884	17	36	17	245	0,3	1,0
UH_6246_uh	0,00	0,00	0,00	0,92	0,00	0,00	0,00	0,08	864	880	18	36	15	245	0,6	-0,3
UH_6161_uh	0,00	0,00	0,02	0,89	0,08	0,00	0,00	0,01	864	878	16	36	16	252	0,7	-0,2
UH_P042_uh	0,00	0,06	0,00	0,82	0,05	0,04	0,00	0,03	878	890	12	36	16	247	0,8	0,1
UH_6145_uh	0,00	0,00	0,03	0,82	0,08	0,00	0,00	0,07	855	874	24	38	16	251	2,5	0,2
PHG50_usda	0,00	0,03	0,00	0,76	0,00	0,21	0,00	0,00	879	901	20	33	16	247	-1,6	0,3
FC1890_inra	0,00	0,09	0,00	0,69	0,00	0,22	0,00	0,00	899	914	14	34	17	257	0,8	1,2
UH_P066_uh	0,00	0,00	0,00	0,69	0,28	0,00	0,00	0,03	860	866	6	37	16	249	1,0	0,1
UH_P148_uh	0,00	0,00	0,31	0,69	0,00	0,00	0,00	0,00	879	888	11	38	16	262	3,4	0,1
PHG83_usda	0,00	0,14	0,00	0,69	0,00	0,17	0,00	0,00	922	939	18	32	18	261	-0,8	1,4
UH_P115_uh	0,00	0,00	0,33	0,67	0,00	0,00	0,00	0,00	874	889	13	37	17	248	2,5	1,3
UH_P074_uh	0,00	0,00	0,34	0,66	0,00	0,00	0,00	0,00	863	881	18	34	16	247	-0,7	0,5
11430_usda	0,00	0,03	0,00	0,66	0,00	0,01	0,30	0,00	898	906	9	35	15	255	1,6	-1,4
UH_P064_uh	0,03	0,02	0,10	0,66	0,07	0,09	0,03	0,00	869	884	13	37	17	270	2,0	1,4
UH_P136_uh	0,00	0,00	0,35	0,65	0,00	0,00	0,00	0,00	871	876	6	37	16	253	1,5	0,0
UH_P131_uh	0,00	0,00	0,28	0,65	0,00	0,00	0,00	0,07	860	870	11	37	17	244	1,1	1,0
Mo12_usa	0,00	0,04	0,03	0,65	0,00	0,19	0,02	0,08	851	884	30	37	13	249	1,5	-2,3
UH_P135_uh	0,00	0,00	0,26	0,63	0,00	0,00	0,00	0,11	853	867	17	38	16	259	2,3	0,9
UH_6179_uh	0,00	0,00	0,03	0,62	0,00	0,00	0,06	0,29	878	896	16	36	17	264	1,7	1,0
UH_6102_uh	0,00	0,00	0,39	0,61	0,00	0,00	0,00	0,00	876	895	20	34	16	256	-0,3	0,2
UH_P040_uh	0,00	0,00	0,39	0,61	0,00	0,00	0,00	0,00	855	858	8	39	16	252	2,4	0,0
UH_6103_uh	0,00	0,00	0,40	0,60	0,00	0,00	0,00	0,00	855	870	13	38	16	252	1,9	0,2
EC242C_ciam	0,00	0,08	0,00	0,59	0,20	0,07	0,06	0,00	861	889	32	36	15	236	1,2	-0,8
UH_P130_uh	0,00	0,01	0,36	0,58	0,04	0,00	0,00	0,00	851	852	1	40	17	241	3,5	1,1
UH_6132_uh	0,00	0,00	0,43	0,57	0,00	0,00	0,00	0,00	848	873	22	36	16	253	1,0	0,8
FV353_inra	0,00	0,00	0,00	0,54	0,00	0,18	0,07	0,21	867	879	13	35	16	258	-0,6	0,1
B103_inra	0,00	0,12	0,00	0,53	0,10	0,23	0,01	0,00	862	888	25	34	15	235	-0,9	-0,7
F912_inra	0,00	0,04	0,00	0,53	0,24	0,12	0,02	0,05	897	902	9	35	16	256	0,3	-0,1
UH_P089_uh	0,01	0,02	0,00	0,52	0,00	0,20	0,04	0,21	882	902	20	33	17	255	-1,1	0,6
UH_6110_uh	0,00	0,00	0,48	0,52	0,00	0,00	0,00	0,00	858	883	24	37	16	258	1,4	0,1
UH_P084_uh	0,49	0,00	0,00	0,51	0,00	0,00	0,00	0,00	873	883	12	34	17	258	-0,9	1,4
FC1819_inra	0,00	0,26	0,01	0,39	0,21	0,09	0,00	0,05	869	877	10	36	16	239	0,7	-0,1
PHV78_usda	0,00	0,31	0,00	0,35	0,01	0,31	0,02	0,00	952	960	9	30	18	271	-1,3	1,8
DKFAPW_usda	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	887	899	12	32	17	256	-2,6	1,0
EZ31_csic	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	911	937	23	34	16	271	1,9	0,2
EZ37_csic	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	908	922	18	34	17	271	0,6	0,8
A632_inra	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	888	917	24	35	17	268	1,4	1,3
B14a_inra	0,02	0,00	0,00	0,00	0,98	0,00	0,00	0,00	919	947	23	31	17	271	-0,6	0,7
PHG39_usda	0,00	0,12	0,00	0,00	0,88	0,00	0,00	0,00	943	940	-2	32	18	268	-0,3	1,2
LH145_usda	0,00	0,00	0,15	0,00	0,85	0,00	0,00	0,00	895	917	23	36	16	262	2,3	-0,2
PHT55_usda	0,00	0,13	0,00	0,03	0,84	0,00	0,00	0,00	951	955	6	31	18	270	0,0	1,8
EC151_ciam	0,00	0,00	0,06	0,00	0,72	0,00	0,22	0,00	898	918	19	32	15	247	-0,9	-0,7
EC334_ciam	0,00	0,05	0,04	0,01	0,66	0,17	0,02	0,05	914	936	20	33	16	253	0,0	0,0
B37_inra	0,00	0,00	0,00	0,00	0,63	0,21	0,08	0,07	939	968	26	29	18	267	-1,9	1,6
EC175_ciam	0,14	0,00	0,09	0,03	0,63	0,08	0,01	0,01	870	901	27	37	17	276	2,6	0,7
EC169_ciam	0,24	0,03	0,03	0,00	0,61	0,08	0,00	0,00	872	889	20	35	16	251	0,4	0,0
DKFBHJ_usda	0,09	0,00	0,00	0,07	0,60	0,19	0,06	0,00	891	901	14	34	15	260	0,2	-1,0
DE811_inra	0,04	0,26	0,00	0,00	0,59	0,07	0,03	0,01	932	951	16	31	16	271	-1,0	-0,2

F7019_inra	0,00	0,00	0,14	0,00	0,59	0,00	0,00	0,27	874	879	8	36	17	249	0,8	0,9
PHG80_usda	0,01	0,31	0,00	0,00	0,56	0,07	0,06	0,00	908	922	16	32	18	262	-2,0	1,5
PHG71_usda	0,00	0,00	0,00	0,45	0,55	0,00	0,00	0,00	880	894	17	35	17	253	0,8	0,8
LH146Ht_usda	0,35	0,00	0,12	0,00	0,53	0,00	0,00	0,00	869	887	20	35	17	256	0,6	1,0
F918_inra	0,00	0,00	0,00	0,02	0,52	0,41	0,00	0,04	926	954	26	30	17	265	-1,5	0,5
F894_inra	0,00	0,00	0,01	0,07	0,51	0,35	0,06	0,00	917	941	21	32	18	264	0,0	1,7
DK4676A_usda	0,05	0,01	0,02	0,05	0,49	0,22	0,09	0,06	886	881	-2	32	16	247	-2,7	0,1
F618_inra	0,00	0,00	0,00	0,07	0,48	0,39	0,07	0,00	905	918	12	33	18	259	0,0	1,9
NKH8431_usda	0,23	0,00	0,06	0,01	0,42	0,23	0,00	0,05	871	887	15	35	17	251	0,0	1,2
F584_inra	0,03	0,02	0,00	0,06	0,42	0,33	0,08	0,07	915	940	23	30	16	254	-1,9	-0,3
PHB09_usda	0,16	0,03	0,00	0,02	0,41	0,18	0,18	0,02	902	911	11	33	18	269	-0,6	1,6
EC130_ciam	0,00	0,07	0,01	0,00	0,40	0,34	0,15	0,03	869	878	11	34	15	230	-0,5	-1,1
EZ48_csic	0,19	0,00	0,00	0,05	0,37	0,30	0,09	0,00	945	966	22	30	18	275	-0,1	1,0
PHK29_usda	0,33	0,10	0,00	0,04	0,37	0,09	0,05	0,03	922	939	18	32	19	276	0,0	3,0
PHV63_usda	0,13	0,27	0,00	0,03	0,36	0,15	0,06	0,00	935	944	7	32	18	262	0,4	1,7
F1808_inra	0,25	0,03	0,02	0,23	0,36	0,11	0,00	0,00	901	907	6	33	18	255	-0,7	2,4
F7081_inra	0,00	0,00	0,06	0,02	0,35	0,24	0,04	0,30	904	917	12	33	16	265	-0,2	0,1
Wf9_inra	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	897	931	34	31	14	257	-1,7	-2,0
CG1_uh	0,03	0,00	0,00	0,00	0,00	0,97	0,00	0,00	859	895	33	32	15	260	-2,9	-1,4
A3_inra	0,05	0,00	0,01	0,01	0,01	0,92	0,00	0,00	889	922	32	33	16	265	-0,3	-0,7
B89_inra	0,05	0,00	0,00	0,06	0,00	0,83	0,06	0,00	878	914	33	34	16	253	0,1	-0,5
EM1163_inra	0,00	0,03	0,06	0,00	0,05	0,82	0,01	0,02	864	893	26	38	14	254	3,2	-2,0
A654_inra	0,00	0,00	0,06	0,03	0,00	0,81	0,00	0,09	864	881	17	35	16	242	-0,2	0,1
SDp254_inra	0,00	0,13	0,00	0,07	0,00	0,80	0,00	0,00	880	904	23	33	15	246	-1,1	-0,5
F604_inra	0,00	0,07	0,05	0,02	0,01	0,76	0,01	0,07	848	872	22	36	14	248	-0,5	-1,8
EP77_csic	0,00	0,02	0,05	0,01	0,06	0,75	0,02	0,10	879	888	13	36	16	270	1,5	-0,4
LH85_usda	0,00	0,00	0,06	0,05	0,00	0,75	0,03	0,12	857	872	19	36	15	237	0,7	-1,1
EZ5_csic	0,01	0,07	0,00	0,00	0,05	0,71	0,03	0,13	920	947	28	30	16	267	-1,9	-0,1
MS71_uh	0,02	0,04	0,04	0,05	0,08	0,71	0,02	0,03	898	921	27	32	15	250	-1,0	-0,9
A554_inra	0,00	0,05	0,09	0,00	0,00	0,71	0,00	0,14	860	873	14	37	13	246	1,4	-2,7
F496_inra	0,04	0,07	0,00	0,00	0,07	0,71	0,07	0,03	860	885	27	35	15	240	-0,2	-0,8
Pa374_inra	0,03	0,08	0,02	0,01	0,06	0,70	0,07	0,04	872	893	23	35	16	247	0,3	0,5
EA3076_csic	0,00	0,08	0,04	0,03	0,00	0,70	0,03	0,12	918	959	35	31	15	273	-0,9	-1,9
F608_inra	0,00	0,03	0,02	0,05	0,07	0,70	0,11	0,01	899	919	19	33	16	260	0,3	-0,1
NC358_uh	0,04	0,08	0,04	0,01	0,00	0,69	0,04	0,10	925	965	33	29	17	251	-1,4	0,2
Pa405_uh	0,06	0,03	0,04	0,00	0,05	0,68	0,05	0,09	886	926	38	35	16	255	1,7	-0,8
B113_uh	0,04	0,10	0,01	0,02	0,04	0,68	0,07	0,05	902	927	23	31	16	261	-1,5	-0,1
W23_uh	0,08	0,03	0,10	0,00	0,00	0,66	0,07	0,06	879	910	30	34	16	263	0,2	-0,2
CQ201_uh	0,04	0,07	0,01	0,02	0,06	0,64	0,12	0,05	848	873	28	33	15	233	-2,3	-0,9
N16_inra	0,05	0,05	0,11	0,04	0,04	0,63	0,00	0,08	961	993	30	28	18	272	-1,0	1,4
FV181_inra	0,10	0,08	0,08	0,01	0,01	0,63	0,01	0,08	830	839	11	38	15	233	0,2	-1,0
A374_inra	0,08	0,00	0,03	0,08	0,06	0,63	0,10	0,04	907	927	19	34	15	249	1,5	-1,0
Oh02_inra	0,05	0,05	0,03	0,04	0,05	0,62	0,06	0,10	947	965	16	33	17	274	2,4	0,1
Pa91_inra	0,02	0,01	0,00	0,00	0,01	0,61	0,35	0,00	939	977	36	28	16	273	-2,5	-0,4
PP147_inra	0,10	0,09	0,03	0,00	0,00	0,61	0,06	0,11	930	954	24	30	16	279	-1,0	-0,4
F7001_inra	0,02	0,01	0,01	0,05	0,12	0,61	0,10	0,08	901	921	19	32	18	273	-0,8	1,9
A375_inra	0,09	0,00	0,00	0,02	0,08	0,61	0,18	0,02	902	915	13	35	15	257	1,7	-0,9
W182B_inra	0,05	0,04	0,07	0,02	0,08	0,61	0,11	0,04	861	868	9	36	15	255	0,1	-1,1
YUBC1a_inra	0,01	0,03	0,00	0,01	0,14	0,61	0,10	0,10	871	904	24	31	14	245	-4,0	-2,3
EM1027_inra	0,02	0,13	0,03	0,03	0,05	0,60	0,06	0,08	922	940	17	31	17	267	-0,8	0,6
N6_inra	0,06	0,07	0,02	0,02	0,06	0,60	0,08	0,09	905	936	33	30	14	247	-1,7	-2,5
A340_inra	0,01	0,07	0,03	0,01	0,03	0,60	0,08	0,17	855	873	20	34	16	260	-1,8	0,0
FV335_inra	0,08	0,07	0,01	0,02	0,05	0,60	0,07	0,11	863	882	21	35	17	254	0,2	1,3
Ia153_inra	0,00	0,08	0,06	0,03	0,05	0,59	0,01	0,18	851	870	20	37	14	237	0,9	-1,0
PB7_inra	0,05	0,08	0,04	0,03	0,03	0,59	0,01	0,16	853	874	22	37	13	237	0,7	-2,6
Oh33_inra	0,01	0,06	0,04	0,09	0,10	0,59	0,02	0,08	919	930	10	34	15	266	1,5	-1,1
EP55_csic	0,04	0,02	0,03	0,01	0,14	0,59	0,10	0,06	879	908	26	36	15	257	2,0	-0,7
EP52_csic	0,06	0,04	0,05	0,00	0,04	0,59	0,16	0,06	910	927	17	31	18	260	-2,3	2,3
B108_uh	0,04	0,13	0,06	0,00	0,04	0,59	0,06	0,08	904	902	2	32	16	261	-1,4	0,3
A310_inra	0,02	0,06	0,05	0,05	0,00	0,59	0,07	0,17	895	916	26	31	14	258	-1,8	-2,0

Mo15W_inra	0,04	0,10	0,03	0,04	0,03	0,58	0,08	0,11	946	988	38	28	16	266	-1,9	-0,9
EC232_ciam	0,02	0,08	0,04	0,05	0,04	0,58	0,08	0,10	890	887	1	34	18	257	-1,0	2,2
B97_inra	0,09	0,04	0,09	0,04	0,01	0,58	0,07	0,07	910	935	25	30	18	272	-2,4	1,6
Mt42_inra	0,06	0,02	0,04	0,04	0,03	0,58	0,03	0,19	831	843	13	40	13	228	2,2	-2,0
WH_inra	0,09	0,03	0,10	0,02	0,00	0,57	0,09	0,09	833	850	14	38	13	238	0,2	-2,3
W182E_inra	0,02	0,04	0,08	0,05	0,06	0,57	0,09	0,10	856	874	17	33	15	245	-2,3	-0,3
PHG84_usda	0,02	0,22	0,00	0,14	0,00	0,57	0,04	0,00	941	952	14	32	18	269	0,0	1,5
PB98TR_inra	0,00	0,10	0,03	0,00	0,06	0,57	0,08	0,16	921	969	42	30	15	257	-0,7	-1,8
EP67_csic	0,01	0,08	0,12	0,02	0,03	0,57	0,10	0,08	872	893	21	35	16	260	0,3	0,4
W33_inra	0,11	0,00	0,04	0,00	0,00	0,57	0,03	0,26	856	861	5	37	13	246	1,3	-2,7
C105_inra	0,02	0,08	0,02	0,00	0,07	0,56	0,11	0,14	847	872	26	35	13	230	-0,9	-2,7
F7057_inra	0,08	0,08	0,02	0,00	0,02	0,55	0,08	0,17	890	908	19	34	15	243	0,1	-1,1
EP2_csic	0,08	0,07	0,02	0,01	0,02	0,55	0,06	0,18	831	861	30	33	13	238	-2,6	-2,8
A188_inra	0,03	0,12	0,02	0,04	0,04	0,55	0,04	0,16	851	866	15	36	16	262	0,0	0,9
EZ19_csic	0,11	0,00	0,05	0,05	0,20	0,55	0,05	0,00	974	1005	27	27	18	287	-1,6	0,3
F908_inra	0,07	0,00	0,04	0,02	0,06	0,55	0,17	0,08	859	883	22	35	15	259	0,1	-0,5
FV218_inra	0,00	0,00	0,00	0,01	0,10	0,55	0,00	0,33	868	882	16	35	15	242	0,3	-0,5
B98_uh	0,04	0,10	0,05	0,02	0,06	0,55	0,07	0,11	924	970	41	30	17	279	-0,6	0,8
W117_inra	0,08	0,11	0,02	0,00	0,00	0,55	0,04	0,20	849	869	21	36	15	238	-0,6	-0,3
EP56_csic	0,02	0,03	0,02	0,05	0,00	0,55	0,09	0,24	847	868	22	38	13	230	2,0	-2,6
LH82_inra	0,05	0,09	0,00	0,26	0,04	0,55	0,00	0,01	870	878	11	35	16	248	0,0	0,1
PHR36_usda	0,02	0,16	0,03	0,18	0,04	0,55	0,00	0,03	901	924	23	31	17	252	-1,7	0,6
N25_inra	0,12	0,09	0,06	0,04	0,03	0,54	0,07	0,05	934	973	35	29	15	270	-1,2	-1,1
PB116_inra	0,00	0,30	0,01	0,00	0,03	0,53	0,04	0,09	905	933	26	31	15	254	-1,3	-1,1
W9_inra	0,02	0,00	0,02	0,00	0,02	0,53	0,08	0,34	876	903	27	37	14	237	2,9	-1,8
CO151_uh	0,00	0,04	0,00	0,00	0,09	0,52	0,03	0,32	827	844	18	39	14	232	1,2	-1,4
A347_inra	0,10	0,01	0,05	0,11	0,07	0,52	0,06	0,08	913	919	5	35	14	261	1,1	-1,9
K55_inra	0,04	0,13	0,04	0,04	0,05	0,52	0,06	0,12	942	957	13	32	18	264	1,1	1,3
MS153_inra	0,03	0,01	0,07	0,05	0,00	0,52	0,16	0,15	883	923	33	32	16	256	-2,1	-0,5
F670_inra	0,06	0,04	0,03	0,06	0,01	0,51	0,09	0,19	866	887	21	35	14	246	0,3	-1,8
NC260_usa	0,03	0,32	0,01	0,05	0,08	0,51	0,00	0,01	943	961	18	31	18	285	-0,1	1,2
FV230_inra	0,07	0,06	0,09	0,03	0,00	0,51	0,01	0,23	844	863	18	35	16	250	-1,4	0,3
CO316_inra	0,00	0,08	0,13	0,10	0,05	0,51	0,00	0,12	841	847	9	40	15	238	2,1	-0,3
PHG35_usda	0,00	0,11	0,02	0,32	0,04	0,50	0,00	0,00	945	959	13	30	17	271	-1,2	0,1
T8_inra	0,03	0,22	0,04	0,04	0,03	0,50	0,05	0,10	979	1008	28	26	17	280	-3,0	-0,2
EZ47_csic	0,10	0,00	0,06	0,00	0,20	0,49	0,10	0,05	899	917	20	30	16	258	-2,7	0,3
EP29_csic	0,05	0,00	0,06	0,04	0,03	0,49	0,25	0,09	886	905	20	32	15	248	-0,6	-0,5
F904_inra	0,00	0,02	0,00	0,22	0,00	0,49	0,03	0,25	865	895	30	33	14	234	-1,7	-2,4
W95115_inra	0,05	0,18	0,05	0,00	0,08	0,48	0,11	0,05	881	900	17	32	15	244	-1,8	-1,5
CL29_inra	0,03	0,01	0,02	0,00	0,21	0,47	0,20	0,06	817	828	12	38	12	232	-0,2	-2,5
LH93_usda	0,02	0,10	0,06	0,04	0,00	0,46	0,20	0,12	933	964	29	31	17	261	0,1	0,6
B111_uh	0,24	0,01	0,01	0,05	0,17	0,46	0,00	0,06	927	935	6	31	17	262	-0,3	0,4
N22_inra	0,05	0,15	0,03	0,10	0,06	0,45	0,04	0,12	963	984	16	29	16	265	-1,1	-1,3
F888_inra	0,05	0,03	0,10	0,05	0,18	0,45	0,08	0,06	906	926	19	32	16	265	-0,7	-0,1
EP51_csic	0,03	0,17	0,04	0,09	0,04	0,45	0,07	0,11	862	898	37	31	14	244	-3,1	-2,3
PHZ51_usda	0,03	0,39	0,03	0,06	0,01	0,45	0,01	0,03	933	958	23	31	18	280	-0,2	1,2
PHG47_usda	0,02	0,12	0,04	0,02	0,02	0,45	0,34	0,00	886	891	11	35	16	255	0,3	0,0
EP28_csic	0,05	0,04	0,03	0,00	0,12	0,45	0,19	0,12	837	844	8	38	13	234	0,1	-1,9
FP1_inra	0,00	0,06	0,04	0,04	0,26	0,45	0,06	0,08	889	912	21	33	16	264	-0,2	-0,4
F838_inra	0,06	0,09	0,03	0,03	0,10	0,44	0,15	0,10	929	947	19	33	18	256	0,9	1,9
Pa36_inra	0,03	0,01	0,04	0,03	0,00	0,44	0,08	0,38	861	885	26	36	12	244	1,3	-3,5
A148_inra	0,01	0,05	0,02	0,00	0,00	0,43	0,22	0,27	828	850	25	39	12	228	1,3	-3,8
EC140_ciam	0,41	0,05	0,00	0,01	0,01	0,43	0,05	0,04	879	889	15	32	16	258	-2,2	0,5
NK807_usda	0,05	0,03	0,00	0,00	0,30	0,42	0,06	0,13	892	898	7	33	17	253	-1,3	0,7
EZ11A_csic	0,07	0,00	0,09	0,03	0,11	0,41	0,21	0,07	926	944	14	28	17	272	-3,3	0,5
A158_inra	0,01	0,01	0,08	0,01	0,01	0,40	0,15	0,32	853	864	13	38	14	233	2,1	-1,9
UH_S020_uh	0,02	0,05	0,24	0,04	0,15	0,39	0,04	0,07	936	953	16	33	18	266	1,6	1,0
F752_inra	0,02	0,07	0,00	0,09	0,33	0,38	0,08	0,02	949	960	6	31	18	270	-0,6	1,4
DK2MA22_usda	0,00	0,15	0,10	0,01	0,02	0,38	0,27	0,07	900	940	33	32	15	258	-0,1	-1,1
FV356_inra	0,00	0,01	0,00	0,24	0,17	0,38	0,07	0,13	856	858	6	37	16	248	0,8	0,3
DKMBPM_usda	0,02	0,10	0,00	0,00	0,07	0,38	0,37	0,06	963	968	10	31	17	276	1,3	0,9

UH_S067_uh	0,04	0,03	0,28	0,01	0,13	0,37	0,03	0,12	865	874	13	36	15	242	0,5	-0,3
EZ46_csic	0,10	0,00	0,10	0,04	0,14	0,37	0,20	0,06	896	906	12	31	16	265	-2,8	0,4
W401_inra	0,01	0,00	0,11	0,02	0,00	0,36	0,15	0,36	828	837	11	39	14	231	0,9	-1,2
PHW65_usda	0,00	0,28	0,02	0,19	0,00	0,34	0,17	0,00	931	951	18	32	18	265	0,0	1,1
CO158_inra	0,01	0,08	0,31	0,02	0,00	0,34	0,10	0,14	854	862	11	37	15	255	0,5	-0,8
DKHBA1_usda	0,00	0,29	0,01	0,05	0,29	0,32	0,00	0,04	990	996	9	29	19	272	0,5	1,9
EP27_csic	0,00	0,00	0,06	0,04	0,20	0,32	0,29	0,09	839	858	20	36	13	239	0,4	-1,9
B100_uh	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	905	919	15	31	16	251	-2,1	-0,3
B102_uh	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	901	927	23	29	17	234	-1,6	0,5
DKMBST_usda	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	925	943	17	30	17	262	-1,7	0,4
H99_inra	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	892	917	26	32	15	243	-1,7	-1,4
ML606_usda	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	891	916	26	31	17	255	-2,1	0,8
A619_inra	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	861	882	21	32	16	245	-3,0	0,3
LH38_usda	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	888	905	17	33	16	249	-1,7	0,4
Oh43_inra	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	901	919	17	32	16	259	-1,7	-0,4
DK78371A_usda	0,00	0,20	0,00	0,00	0,00	0,00	0,80	0,00	918	927	9	31	17	269	-2,3	0,8
Va26_inra	0,04	0,00	0,00	0,00	0,02	0,14	0,80	0,00	908	928	22	32	16	265	-0,3	-0,7
H95_inra	0,00	0,03	0,00	0,01	0,02	0,24	0,69	0,01	955	977	19	31	17	260	0,0	0,0
Oh40B_inra	0,01	0,11	0,00	0,00	0,01	0,19	0,67	0,00	898	923	25	33	16	268	0,1	-0,4
LH123Ht_usda	0,02	0,41	0,00	0,01	0,09	0,00	0,47	0,00	938	952	14	29	17	284	-1,7	0,6
F7059_inra	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	861	879	20	37	16	256	1,9	0,6
FV292_inra	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	845	857	17	39	15	251	2,5	-0,6
FV252_inra	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	848	865	22	38	15	251	1,3	-0,9
F7038_inra	0,03	0,00	0,00	0,00	0,00	0,00	0,00	0,97	870	898	25	37	16	261	1,9	0,2
CO125_inra	0,00	0,00	0,04	0,00	0,00	0,00	0,00	0,96	852	867	13	38	14	256	1,8	-1,4
FV288_inra	0,00	0,07	0,00	0,00	0,00	0,00	0,00	0,92	864	874	13	37	16	243	0,9	-0,1
UH_2500_uh	0,01	0,00	0,13	0,00	0,00	0,02	0,02	0,82	863	889	23	37	14	246	1,9	-1,4
UH_1603B_uh	0,00	0,00	0,11	0,00	0,00	0,04	0,07	0,79	862	870	9	37	16	246	0,9	-0,3
F7028_inra	0,15	0,12	0,00	0,00	0,00	0,00	0,00	0,73	893	912	24	38	16	252	4,2	-0,1
FV317_inra	0,00	0,17	0,02	0,00	0,12	0,00	0,00	0,69	874	889	16	37	15	245	2,0	-0,8
UH_8513_uh	0,00	0,00	0,12	0,06	0,00	0,11	0,07	0,65	856	861	7	37	16	258	0,7	0,4
FV284_inra	0,00	0,00	0,05	0,00	0,03	0,26	0,03	0,62	841	864	23	36	14	225	0,3	-1,3
F7009_inra	0,00	0,00	0,03	0,00	0,05	0,27	0,05	0,61	849	862	19	39	14	225	2,0	-1,4
FV113_inra	0,00	0,01	0,03	0,00	0,02	0,35	0,01	0,59	863	875	14	37	16	242	1,1	0,4
FV277_inra	0,00	0,00	0,01	0,00	0,00	0,41	0,01	0,58	841	845	6	35	15	247	-1,8	-0,4
F922_inra	0,00	0,02	0,00	0,10	0,01	0,25	0,06	0,57	889	895	10	38	18	272	3,1	2,4
FV354_inra	0,03	0,16	0,00	0,00	0,02	0,23	0,06	0,50	880	886	6	37	16	234	1,3	-0,8
FV330_inra	0,00	0,04	0,01	0,22	0,00	0,26	0,00	0,47	844	854	13	36	15	239	-1,4	-0,3
UH_2551_uh	0,00	0,00	0,15	0,06	0,00	0,26	0,14	0,39	839	856	15	37	16	248	0,4	0,7
UH_1595_uh	0,01	0,09	0,12	0,07	0,06	0,26	0,00	0,38	856	879	23	35	15	240	-1,0	-0,7
UH_1675_uh	0,00	0,04	0,20	0,05	0,29	0,05	0,00	0,37	837	852	17	41	15	243	3,5	-0,4

Table S3: Admixture and performances of the Flint lines (crossed to the tester).

Accession	Hohen. Fl.	UH_F047fam.	Lacaune	CIAMAranga	Ital.OP	Spanish	Pyren.	NF	Tass_GDD6	Silk_GDD6	ASI_GDD6	DMC	DMY	PLHT	DMCcorr	DMYcorr
UH_5271_uh	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	846	875	21	34	14	247	1,3	-0,4
UH_F016_uh	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	864	885	15	33	17	253	0,9	1,9
UH_L054_uh	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	873	895	13	34	16	253	1,9	1,5
UH_5231_uh	0,93	0,06	0,00	0,00	0,00	0,00	0,01	0,00	867	886	14	31	15	251	-1,0	0,4
UH_5250_uh	0,82	0,00	0,18	0,00	0,00	0,00	0,00	0,00	841	887	38	30	15	241	-2,0	-0,2
UH_L038_uh	0,82	0,00	0,00	0,00	0,00	0,14	0,00	0,05	870	902	22	31	15	256	0,0	-0,1
FV362_inra	0,78	0,00	0,22	0,00	0,00	0,00	0,00	0,00	850	870	13	35	15	244	2,5	0,2
UH_1107_uh	0,78	0,00	0,00	0,00	0,00	0,00	0,15	0,08	846	884	32	31	15	247	-1,4	0,5
UH_5248_uh	0,76	0,00	0,09	0,00	0,00	0,00	0,01	0,14	825	864	35	34	13	237	0,6	-1,1
UH_5267_uh	0,75	0,25	0,00	0,00	0,00	0,00	0,00	0,00	881	899	16	33	16	262	0,7	1,6
F02803_inra	0,75	0,01	0,24	0,00	0,00	0,00	0,00	0,00	876	891	11	33	15	258	1,5	0,7
UH_2109_uh	0,74	0,26	0,00	0,00	0,00	0,00	0,00	0,00	836	868	29	32	15	242	-0,9	0,0
UH_L031_uh	0,74	0,00	0,00	0,00	0,00	0,26	0,00	0,00	885	916	20	32	13	237	0,7	-1,8
UH_1224_uh	0,73	0,00	0,00	0,00	0,00	0,00	0,23	0,04	864	893	20	32	15	244	0,4	-0,3
FV361_inra	0,73	0,14	0,00	0,00	0,02	0,11	0,00	0,00	857	883	20	35	15	250	3,0	0,6
UH_F050_uh	0,71	0,29	0,00	0,00	0,00	0,00	0,00	0,00	859	886	22	34	15	256	1,2	0,7
UH_8007_uh	0,66	0,00	0,11	0,00	0,00	0,00	0,19	0,04	869	890	18	31	13	230	-1,0	-1,7
F363_inra	0,62	0,12	0,23	0,00	0,00	0,00	0,03	0,00	855	871	15	35	15	253	2,0	0,3
UH_F035_uh	0,62	0,23	0,10	0,00	0,00	0,00	0,06	0,00	842	866	17	34	14	244	1,4	-0,2
UH_5172_uh	0,59	0,02	0,07	0,00	0,00	0,09	0,19	0,05	876	898	16	32	17	257	0,5	1,7
UH_5264_uh	0,58	0,27	0,15	0,00	0,00	0,00	0,01	0,00	858	878	13	34	13	237	1,0	-1,3
F364_inra	0,55	0,00	0,41	0,00	0,00	0,00	0,00	0,04	855	896	30	34	16	254	2,1	0,9
UH_L016_uh	0,48	0,39	0,00	0,01	0,00	0,12	0,00	0,00	878	913	26	31	15	284	-0,8	0,1
UH_L021_uh	0,45	0,27	0,00	0,00	0,00	0,28	0,00	0,00	864	896	24	31	15	273	-0,3	0,4
F03802_inra	0,45	0,00	0,32	0,00	0,00	0,22	0,00	0,02	846	875	23	37	17	263	4,4	2,1
UH_L042_uh	0,44	0,33	0,00	0,01	0,00	0,19	0,00	0,04	882	906	16	32	16	253	0,5	0,7
UH_5206_uh	0,39	0,10	0,09	0,00	0,00	0,07	0,25	0,10	847	881	35	35	15	251	3,2	0,4
UH_5113_uh	0,38	0,00	0,00	0,00	0,00	0,00	0,25	0,36	856	883	20	31	16	247	-1,2	0,9
UH_7727_uh	0,35	0,00	0,00	0,00	0,00	0,33	0,00	0,32	848	879	24	36	15	243	3,3	0,2
UH_F023_uh	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	870	904	23	31	18	257	-0,8	2,7
UH_F070_uh	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	875	899	19	34	15	257	2,2	-0,2
UH_F084_uh	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	868	893	20	32	16	259	0,0	1,4
UH_F093_uh	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	869	894	25	32	17	271	-0,1	1,7
UH_F098_uh	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	841	864	14	35	16	251	2,2	1,7
UH_F105_uh	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	868	886	14	33	17	258	0,3	2,0
UH_L058_uh	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	893	920	14	30	14	278	-0,9	-0,8
UH_F082_uh	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	865	896	26	33	15	254	1,2	0,6
UH_F091_uh	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	875	907	24	32	13	248	0,0	-2,5
UH006_uh	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	867	894	23	33	17	255	1,1	2,3
UH_F043_uh	0,14	0,86	0,00	0,00	0,00	0,00	0,00	0,00	876	908	21	31	15	256	-0,6	0,1
UH_F106_uh	0,00	0,83	0,06	0,00	0,04	0,05	0,00	0,03	869	898	22	31	15	259	-0,9	0,0
UH_F038_uh	0,25	0,75	0,00	0,00	0,00	0,00	0,00	0,00	849	876	18	34	13	249	0,9	-1,3

UH_F048_uh	0,27	0,73	0,00	0,00	0,00	0,00	0,00	0,00	882	907	19	31	12	253	-0,5	-2,5
UH_5222_uh	0,31	0,69	0,00	0,00	0,00	0,00	0,00	0,00	860	892	27	32	16	256	-0,3	0,7
UH_F020_uh	0,20	0,67	0,07	0,00	0,00	0,00	0,00	0,05	873	885	18	35	11	223	2,1	-3,7
UH_F037_uh	0,26	0,56	0,00	0,00	0,00	0,00	0,01	0,16	857	890	25	33	15	257	1,1	0,8
UH_F027_uh	0,49	0,50	0,00	0,00	0,00	0,00	0,00	0,01	844	877	28	33	14	249	0,5	-0,3
UH_L048_uh	0,23	0,46	0,00	0,02	0,00	0,23	0,00	0,06	869	885	11	32	17	256	0,1	2,1
UH_2065_uh	0,30	0,37	0,05	0,01	0,02	0,00	0,08	0,17	848	871	17	36	14	254	3,0	-1,0
UH_F018_uh	0,21	0,31	0,04	0,00	0,00	0,00	0,17	0,26	865	894	24	32	17	251	0,3	2,1
F7012_inra	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	829	857	24	35	15	234	1,0	0,7
FV283_inra	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	866	904	28	31	16	245	0,2	1,2
FV286_inra	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	891	927	31	29	15	239	-1,8	-0,4
FV324_inra	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	856	898	35	32	16	238	-0,1	1,2
FV7_inra	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	831	873	29	34	15	230	1,0	0,3
FV226_inra	0,00	0,00	0,99	0,00	0,00	0,00	0,00	0,01	832	858	20	35	15	243	1,0	0,6
UH_1118_uh	0,05	0,00	0,95	0,00	0,00	0,00	0,00	0,00	879	907	23	32	15	242	0,2	0,3
FV1_inra	0,00	0,00	0,93	0,00	0,00	0,00	0,07	0,00	803	841	29	37	13	217	2,9	-1,1
EZ59_csic	0,00	0,01	0,81	0,00	0,00	0,00	0,18	0,00	871	915	32	30	16	246	-0,2	1,4
F564_inra	0,00	0,00	0,80	0,00	0,02	0,19	0,00	0,00	914	945	21	29	17	259	-1,4	1,9
UH_1199_uh	0,03	0,01	0,78	0,00	0,00	0,18	0,00	0,00	859	876	15	33	17	257	0,5	2,0
PB268_inra	0,00	0,09	0,73	0,00	0,00	0,00	0,18	0,00	854	874	15	33	17	247	0,3	1,8
FV373_inra	0,06	0,00	0,68	0,00	0,00	0,21	0,00	0,05	857	877	13	35	15	241	2,2	0,7
FV85_inra	0,00	0,04	0,65	0,00	0,00	0,00	0,31	0,00	858	901	36	31	14	224	-1,1	-0,6
FV355b_inra	0,03	0,00	0,64	0,01	0,02	0,23	0,02	0,05	847	868	8	33	15	241	0,4	0,4
UH_3056_uh	0,09	0,00	0,64	0,00	0,00	0,15	0,09	0,04	873	891	13	32	17	265	0,0	1,6
F657wx_inra	0,00	0,10	0,61	0,00	0,00	0,13	0,12	0,04	900	937	27	29	16	265	-0,8	0,6
FV160_inra	0,00	0,01	0,59	0,00	0,01	0,32	0,07	0,00	838	871	31	33	12	226	0,0	-2,1
UH_5257_uh	0,36	0,07	0,57	0,00	0,00	0,00	0,00	0,00	829	859	25	35	14	239	1,7	-0,2
FV344_inra	0,00	0,02	0,52	0,01	0,03	0,26	0,10	0,05	872	882	2	34	16	256	2,2	0,8
F902_inra	0,00	0,16	0,50	0,00	0,00	0,00	0,33	0,01	854	880	19	31	15	259	-1,7	0,6
F350_inra	0,05	0,00	0,45	0,03	0,11	0,03	0,28	0,06	894	914	9	28	13	249	-3,0	-2,1
F337_inra	0,00	0,19	0,44	0,00	0,00	0,06	0,32	0,00	835	859	14	34	15	253	0,8	0,7
CO255_inra	0,15	0,00	0,43	0,00	0,00	0,16	0,25	0,00	854	883	26	33	14	253	1,0	-0,6
F03801_inra	0,33	0,07	0,42	0,00	0,00	0,19	0,00	0,00	843	870	21	36	15	260	3,1	0,0
F916_inra	0,00	0,00	0,40	0,00	0,07	0,38	0,07	0,09	919	954	24	30	18	275	0,6	2,9
F359_inra	0,16	0,00	0,40	0,01	0,03	0,34	0,06	0,00	856	883	20	36	17	269	3,3	1,9
EC35G_ciam	0,02	0,00	0,28	0,23	0,10	0,09	0,16	0,11	880	913	26	31	17	269	-0,5	1,7
EC209_ciam	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	862	908	34	33	16	245	2,4	0,4
EC218_ciam	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	936	979	32	29	16	276	-0,4	0,5
EC237_ciam	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	928	965	31	30	19	272	0,1	3,4
EC244_ciam	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	871	907	23	33	16	247	1,8	0,6
EC45_ciam	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	920	937	15	28	17	259	-2,1	1,4
EC46_ciam	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	944	956	6	29	16	268	0,0	0,2
EC49A_ciam	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	926	943	13	29	17	262	-1,7	1,9
EC50_ciam	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	915	952	27	30	20	287	0,4	4,0

EC214_ciam	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	920	953	17	31	18	261	1,1	2,9
EC212A_ciam	0,00	0,03	0,03	0,95	0,00	0,00	0,00	0,00	895	918	15	31	17	247	-0,5	1,5
EC246_ciam	0,00	0,00	0,00	0,76	0,00	0,00	0,24	0,00	844	876	27	35	13	235	2,5	-1,4
EC248_ciam	0,00	0,00	0,00	0,70	0,06	0,03	0,21	0,00	912	947	22	30	16	277	0,1	0,9
EC245B_ciam	0,00	0,00	0,02	0,64	0,07	0,26	0,00	0,00	831	865	27	31	14	237	-2,0	-0,8
EC51_ciam	0,03	0,02	0,24	0,52	0,00	0,07	0,12	0,00	885	918	25	28	16	235	-2,9	1,0
EC23A_ciam	0,00	0,00	0,00	0,52	0,00	0,00	0,48	0,00	862	890	24	33	16	250	1,3	1,4
CH10-4_ethz	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	908	943	23	31	16	257	0,9	0,3
EP86_csic	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	889	908	16	31	14	267	-0,7	-0,5
LO11_inra	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	949	982	22	29	12	243	0,1	-3,8
LO32_inra	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	976	982	5	30	16	266	0,5	0,0
LO33_inra	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	876	914	37	34	15	251	2,1	-0,1
LO3_inra	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	936	964	20	28	14	266	-1,8	-1,2
PB79-2_inra	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	897	949	41	27	16	264	-2,8	0,6
PB80_inra	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	909	948	26	29	15	259	-1,2	-0,4
PB97_inra	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	905	930	12	30	16	260	-0,5	0,4
PP87_inra	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	920	946	18	30	17	263	0,5	1,2
PB57_inra	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	901	929	12	29	16	264	-1,2	0,6
CH10-3_ethz	0,00	0,00	0,00	0,00	0,78	0,00	0,00	0,22	885	903	16	35	13	237	2,8	-1,8
F48_inra	0,00	0,00	0,05	0,00	0,71	0,09	0,05	0,09	884	920	20	31	14	252	0,0	-1,1
CH446A_ethz	0,00	0,00	0,00	0,00	0,57	0,21	0,04	0,18	860	890	29	33	14	245	0,4	-0,4
EM1301_inra	0,02	0,00	0,00	0,01	0,47	0,00	0,27	0,23	881	907	19	29	13	249	-3,1	-2,3
FC16_inra	0,00	0,00	0,01	0,04	0,39	0,19	0,26	0,12	895	926	24	31	14	246	0,5	-0,8
EA2841_csic	0,01	0,03	0,00	0,08	0,33	0,33	0,14	0,08	907	942	21	30	12	261	-1,2	-3,5
EZ38_csic	0,05	0,00	0,00	0,00	0,00	0,86	0,00	0,09	962	995	19	29	19	296	1,4	3,3
EZ22_csic	0,04	0,00	0,00	0,08	0,00	0,83	0,00	0,04	1019	1057	22	24	19	290	-1,9	1,6
EZ53_csic	0,00	0,03	0,03	0,00	0,00	0,82	0,00	0,12	894	931	24	30	14	247	-0,7	-1,2
EZ6_csic	0,00	0,00	0,00	0,01	0,18	0,81	0,00	0,00	969	990	11	29	18	274	0,7	1,5
EZ33_csic	0,00	0,00	0,00	0,00	0,19	0,81	0,00	0,00	965	994	14	29	18	270	0,6	2,1
CO114_inra	0,00	0,05	0,00	0,04	0,00	0,80	0,00	0,11	823	858	32	32	12	226	-1,8	-2,3
EZ32_csic	0,03	0,03	0,00	0,00	0,10	0,79	0,00	0,05	1035	1052	13	26	18	284	-0,6	1,1
P465P_inra	0,03	0,02	0,03	0,02	0,09	0,74	0,02	0,05	925	971	40	30	16	256	0,6	0,4
CO109_inra	0,00	0,03	0,00	0,02	0,01	0,73	0,01	0,21	898	943	34	28	11	253	-2,0	-4,2
F816_inra	0,15	0,00	0,00	0,00	0,00	0,72	0,03	0,09	890	914	16	31	16	268	0,1	0,5
EZ2_csic	0,00	0,00	0,00	0,00	0,29	0,71	0,00	0,00	987	1011	14	26	18	288	-1,4	2,0
EP69_csic	0,00	0,06	0,08	0,04	0,05	0,68	0,07	0,02	863	895	27	30	12	222	-1,5	-3,2
EZ30_csic	0,00	0,01	0,00	0,06	0,17	0,67	0,08	0,00	968	1009	26	25	17	281	-3,2	0,9
EP68_csic	0,00	0,09	0,04	0,00	0,12	0,67	0,03	0,05	928	949	15	29	13	257	-0,1	-2,5
UH_7057_uh	0,23	0,12	0,00	0,00	0,00	0,65	0,00	0,00	940	972	22	30	14	265	0,8	-1,8
PB53_inra	0,07	0,00	0,00	0,07	0,02	0,65	0,01	0,19	832	892	43	33	15	246	1,6	0,1
F920_inra	0,00	0,02	0,17	0,00	0,04	0,63	0,11	0,02	917	928	2	31	14	249	1,0	-1,7
EZ10_csic	0,00	0,01	0,01	0,00	0,24	0,62	0,12	0,01	942	972	19	27	17	256	-1,5	0,8
UH_7001_uh	0,18	0,02	0,04	0,00	0,00	0,61	0,08	0,07	905	922	10	33	15	250	2,5	-0,2
EZ51_csic	0,04	0,00	0,00	0,05	0,09	0,61	0,13	0,07	986	997	10	27	17	296	-0,1	1,1

EP73_csic	0,00	0,07	0,03	0,06	0,12	0,61	0,06	0,05	897	929	24	32	15	229	0,6	-0,4
UH_L007_uh	0,28	0,13	0,00	0,00	0,00	0,59	0,00	0,00	913	939	15	32	16	265	2,1	1,0
EA2024_csic	0,02	0,00	0,03	0,03	0,22	0,59	0,12	0,01	927	964	27	28	15	268	-1,5	-0,4
F64_inra	0,01	0,00	0,16	0,03	0,17	0,58	0,05	0,00	975	1007	28	28	17	287	0,5	1,1
FP1_inra	0,03	0,03	0,00	0,08	0,00	0,58	0,12	0,16	922	960	24	30	15	255	0,1	-0,9
PP85_inra	0,00	0,00	0,04	0,05	0,21	0,58	0,08	0,03	991	1006	14	28	18	293	-0,3	1,5
EP65_csic	0,00	0,02	0,00	0,04	0,00	0,57	0,24	0,13	976	1005	23	26	15	264	-2,0	-1,1
EZ49_csic	0,00	0,04	0,00	0,03	0,23	0,56	0,12	0,02	944	979	25	27	16	266	-1,4	-0,1
EP79_csic	0,07	0,00	0,11	0,08	0,00	0,56	0,16	0,01	830	860	16	33	15	230	0,0	0,1
EP64_csic	0,00	0,00	0,00	0,07	0,00	0,55	0,18	0,20	959	986	19	26	14	270	-3,1	-1,6
EP31_csic	0,06	0,03	0,04	0,09	0,01	0,54	0,07	0,15	865	896	24	32	15	259	0,5	-0,1
UH_L001_uh	0,09	0,13	0,03	0,02	0,06	0,53	0,00	0,14	877	920	31	33	13	242	2,0	-1,8
Ia2132_usa	0,00	0,03	0,00	0,00	0,00	0,52	0,00	0,46	890	923	18	31	12	246	0,1	-3,2
UH_L003_uh	0,34	0,00	0,04	0,00	0,00	0,51	0,00	0,11	934	965	23	28	15	266	-1,0	-0,9
EM1027_inra	0,00	0,03	0,00	0,05	0,18	0,50	0,16	0,09	952	988	19	29	17	282	-0,1	0,5
F591_inra	0,00	0,05	0,05	0,00	0,09	0,49	0,27	0,05	922	949	17	30	16	279	0,0	0,4
EA2000_csic	0,00	0,00	0,00	0,04	0,36	0,48	0,12	0,00	862	912	41	30	14	250	-1,9	-0,8
F473_inra	0,00	0,02	0,00	0,00	0,08	0,47	0,14	0,28	883	907	16	33	14	249	1,4	-1,6
EP53_csic	0,00	0,03	0,00	0,11	0,05	0,47	0,07	0,27	863	894	25	28	13	236	-4,4	-2,1
F9003_inra	0,01	0,13	0,12	0,00	0,00	0,46	0,26	0,02	908	940	27	27	14	264	-3,1	-1,6
UH_L010_uh	0,25	0,00	0,00	0,02	0,00	0,44	0,00	0,29	916	941	15	29	17	266	-0,8	1,3
UH_L012_uh	0,30	0,17	0,00	0,03	0,00	0,44	0,00	0,06	919	951	22	30	17	281	0,2	1,3
C105_inra	0,01	0,02	0,02	0,03	0,00	0,43	0,05	0,43	884	918	27	32	14	238	1,0	-1,1
EC22_ciam	0,04	0,03	0,03	0,15	0,04	0,43	0,22	0,06	916	951	21	31	17	273	0,6	1,5
EP32_csic	0,03	0,00	0,00	0,22	0,09	0,42	0,18	0,06	881	915	25	32	15	251	0,8	-0,2
EP39_csic	0,05	0,00	0,00	0,09	0,25	0,42	0,13	0,06	877	929	39	32	13	240	0,9	-1,9
F61_inra	0,03	0,05	0,00	0,03	0,17	0,41	0,07	0,25	939	967	16	28	16	266	-1,4	0,0
EA1070_csic	0,00	0,00	0,00	0,00	0,35	0,39	0,26	0,00	913	956	33	29	15	256	-0,8	-0,4
EP66_csic	0,00	0,00	0,00	0,10	0,00	0,39	0,31	0,19	940	959	11	27	16	265	-2,7	0,6
F347_inra	0,01	0,00	0,33	0,05	0,14	0,39	0,02	0,06	906	944	30	31	17	260	0,7	1,2
PB261_inra	0,00	0,00	0,00	0,10	0,00	0,38	0,32	0,21	929	953	16	29	16	270	-0,7	-0,1
EP71_csic	0,00	0,00	0,00	0,11	0,04	0,36	0,27	0,22	904	926	12	29	15	255	-2,1	-0,9
PB6R_inra	0,00	0,00	0,01	0,15	0,08	0,35	0,22	0,19	895	925	21	28	15	232	-2,5	-0,1
EP45_csic	0,01	0,02	0,03	0,01	0,20	0,32	0,12	0,28	939	964	21	30	16	278	-0,1	0,2
FV281_inra	0,14	0,00	0,28	0,00	0,00	0,31	0,27	0,00	829	865	26	33	14	241	0,1	-0,2
EP80_csic	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	928	955	20	30	18	285	0,7	2,8
FV70_inra	0,00	0,00	0,00	0,00	0,00	0,00	0,99	0,01	865	901	28	32	15	252	0,0	-0,3
FV71_inra	0,00	0,00	0,03	0,00	0,00	0,00	0,97	0,00	837	877	30	33	14	236	0,5	-0,6
F47_inra	0,01	0,00	0,00	0,00	0,03	0,00	0,89	0,07	863	884	15	29	15	251	-3,4	0,4
FV69a_inra	0,02	0,00	0,03	0,01	0,13	0,00	0,79	0,03	832	849	21	32	15	246	-1,9	0,6
F759_inra	0,00	0,03	0,13	0,00	0,00	0,00	0,78	0,06	924	961	33	30	16	274	0,9	0,5
FC1772_inra	0,00	0,00	0,00	0,00	0,10	0,09	0,73	0,08	893	937	40	29	16	260	-1,6	0,7
CH36_ethz	0,00	0,00	0,00	0,10	0,12	0,01	0,72	0,05	856	876	15	32	13	243	-1,1	-2,1
F45_inra	0,00	0,00	0,09	0,05	0,12	0,00	0,72	0,02	884	907	15	30	15	255	-2,0	-0,3
FV72_inra	0,00	0,03	0,06	0,09	0,11	0,00	0,72	0,00	846	888	31	32	14	233	-0,4	-0,8

F41_inra	0,00	0,02	0,15	0,00	0,09	0,00	0,71	0,04	874	915	32	30	15	243	-1,0	0,0
FV83_inra	0,29	0,00	0,00	0,00	0,00	0,00	0,71	0,00	875	902	24	30	14	246	-1,6	-1,2
EP1_inra	0,24	0,05	0,00	0,00	0,00	0,00	0,71	0,00	850	890	40	29	13	239	-3,2	-1,4
FC201_inra	0,00	0,00	0,00	0,00	0,07	0,00	0,69	0,24	887	931	39	30	15	259	-0,8	-0,1
FV65_inra	0,00	0,00	0,05	0,08	0,15	0,00	0,69	0,04	822	846	20	36	13	240	2,3	-1,3
F39_inra	0,00	0,02	0,06	0,00	0,13	0,02	0,66	0,10	914	945	23	30	15	268	0,1	-0,4
FV75_inra	0,00	0,04	0,11	0,00	0,03	0,00	0,62	0,19	864	890	28	34	15	261	1,4	0,5
FC673_inra	0,00	0,00	0,14	0,00	0,06	0,00	0,62	0,19	917	935	11	30	16	279	-0,4	0,7
FC209_inra	0,00	0,00	0,00	0,00	0,04	0,00	0,61	0,35	901	938	27	32	15	247	1,2	-0,4
F476_inra	0,05	0,01	0,00	0,01	0,00	0,29	0,60	0,04	925	950	17	30	16	262	-0,1	0,9
W121_inra	0,23	0,03	0,00	0,04	0,00	0,11	0,60	0,00	842	880	28	30	12	234	-1,9	-2,4
FV74_inra	0,00	0,00	0,10	0,02	0,05	0,00	0,59	0,24	862	889	27	30	14	252	-2,7	-1,0
FV76_inra	0,03	0,00	0,05	0,05	0,21	0,03	0,58	0,06	839	862	18	32	13	223	-1,3	-1,9
FC46_inra	0,00	0,00	0,00	0,00	0,04	0,00	0,56	0,40	909	955	34	30	15	258	0,1	-0,7
FV18_inra	0,00	0,04	0,16	0,00	0,00	0,00	0,56	0,24	897	929	26	30	16	262	-0,5	0,5
F810_inra	0,00	0,02	0,00	0,00	0,03	0,35	0,54	0,05	931	949	15	31	16	259	0,8	0,3
EZ3_csic	0,00	0,04	0,17	0,00	0,26	0,01	0,52	0,00	931	959	18	29	15	249	-0,6	-0,6
EZ4_csic	0,00	0,00	0,21	0,03	0,25	0,00	0,51	0,00	904	937	20	30	15	241	-0,2	-0,6
FC23_inra	0,00	0,00	0,00	0,00	0,02	0,00	0,50	0,47	875	908	28	34	17	254	2,5	1,6
EM1349_inra	0,00	0,02	0,06	0,07	0,17	0,17	0,49	0,04	933	966	25	28	16	281	-1,4	0,5
FV2_inra	0,00	0,34	0,18	0,00	0,00	0,00	0,48	0,00	828	859	23	33	15	242	-0,3	0,9
FC21_inra	0,00	0,00	0,00	0,00	0,08	0,00	0,47	0,45	868	913	39	33	13	240	1,2	-2,3
FV268_inra	0,00	0,01	0,45	0,00	0,00	0,07	0,47	0,00	828	873	38	35	14	228	2,1	-0,7
FC42_inra	0,00	0,01	0,01	0,02	0,08	0,00	0,46	0,42	900	928	18	32	15	252	1,2	0,0
EP40_csic	0,00	0,01	0,01	0,24	0,12	0,17	0,44	0,01	906	944	25	27	15	272	-3,2	-0,6
F7048_inra	0,00	0,00	0,00	0,00	0,03	0,43	0,43	0,11	949	985	28	29	19	287	0,3	2,7
PLS42_inra	0,00	0,00	0,00	0,00	0,15	0,00	0,43	0,42	842	878	32	32	14	232	-0,3	-0,2
EP42_csic	0,00	0,05	0,00	0,28	0,06	0,11	0,41	0,08	914	949	29	31	16	265	0,9	0,2
PV139_inra	0,00	0,00	0,00	0,13	0,11	0,13	0,41	0,21	874	917	30	31	14	254	-0,4	-1,1
FV77_inra	0,00	0,00	0,09	0,01	0,26	0,02	0,37	0,25	856	888	23	32	14	233	-0,4	-0,6
FV79_inra	0,00	0,00	0,05	0,03	0,24	0,00	0,37	0,30	843	871	20	33	14	226	-0,1	-0,7
UH_1494_uh	0,09	0,36	0,18	0,00	0,00	0,00	0,37	0,00	873	906	30	30	14	242	-2,0	-0,6
FV345_inra	0,11	0,06	0,27	0,01	0,00	0,14	0,36	0,05	837	884	39	34	16	244	1,5	1,4
PB86_inra	0,00	0,05	0,03	0,16	0,03	0,26	0,36	0,10	883	922	30	29	14	254	-1,9	-1,6
PV135_inra	0,00	0,00	0,00	0,17	0,12	0,15	0,35	0,20	834	868	24	37	16	252	3,7	1,3
PV125_inra	0,01	0,00	0,01	0,18	0,10	0,13	0,35	0,21	853	878	17	30	14	227	-2,4	-0,9
PB40_inra	0,00	0,00	0,00	0,08	0,06	0,27	0,35	0,23	912	947	25	28	15	270	-1,4	-0,4
EP37_csic	0,11	0,00	0,21	0,20	0,00	0,15	0,34	0,00	849	894	38	31	14	229	-0,7	-0,4
PB18_inra	0,00	0,00	0,03	0,12	0,08	0,15	0,33	0,29	960	991	18	27	18	293	-1,7	1,8
F7032_inra	0,11	0,00	0,23	0,00	0,03	0,27	0,30	0,07	840	887	35	31	15	241	-1,3	0,1
UH_1102_uh	0,22	0,24	0,20	0,00	0,00	0,05	0,29	0,00	852	878	17	34	15	242	0,8	0,8
CH19-1_ethz	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	820	859	34	34	14	251	1,0	-0,4
CH19-3_ethz	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	850	883	28	33	16	260	0,2	1,6
CH27-12_ethz	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	829	874	38	32	14	258	-1,3	-0,3
CH27-17_ethz	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	829	872	37	31	14	251	-1,6	-0,6

CH28-2_ethz	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	837	883	34	28	14	266	-4,3	-0,6
CH34_ethz	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	834	880	42	33	14	242	0,0	-1,1
CH4-2_ethz	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	911	939	19	28	14	266	-2,1	-1,6
CH5-2_ethz	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	849	876	12	33	14	242	0,0	-1,0
CH8-6_ethz	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	884	909	18	28	15	258	-3,5	-0,6
CH8-7_ethz	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	891	919	21	30	14	252	-1,0	-0,8
FC24_inra	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	881	911	24	28	14	264	-3,3	-1,1
CH7-1_ethz	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	865	904	28	33	17	278	0,8	1,7
CH17-3_ethz	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	897	918	7	31	16	257	0,4	1,0
CH16-1_ethz	0,00	0,00	0,00	0,00	0,04	0,00	0,00	0,96	880	915	28	32	15	274	1,1	-0,1
UH_1101_uh	0,00	0,00	0,04	0,00	0,00	0,00	0,00	0,96	840	881	35	33	16	256	0,7	1,8
ND33_inra	0,00	0,00	0,01	0,01	0,00	0,04	0,03	0,90	850	877	22	34	15	245	1,4	-0,2
PLS27_inra	0,00	0,00	0,00	0,00	0,03	0,00	0,08	0,90	819	857	31	34	13	220	0,5	-0,9
RT9_inra	0,00	0,02	0,00	0,00	0,08	0,00	0,00	0,89	825	858	29	33	14	245	-0,2	-0,2
NYS302_inra	0,00	0,00	0,00	0,00	0,00	0,12	0,00	0,88	905	922	10	31	14	255	-0,3	-1,1
D105_inra	0,12	0,00	0,00	0,00	0,00	0,00	0,00	0,88	860	909	33	33	17	260	1,3	2,1
UH_FF0721H_uh	0,01	0,00	0,00	0,00	0,01	0,00	0,11	0,87	835	862	25	35	15	241	1,4	0,2
IL101_usa	0,00	0,00	0,00	0,00	0,00	0,16	0,00	0,84	857	907	30	29	13	235	-2,0	-2,4
P39_usa	0,00	0,00	0,01	0,01	0,00	0,15	0,00	0,83	868	907	30	30	12	235	-2,1	-2,8
DK105_uh	0,20	0,00	0,00	0,00	0,00	0,00	0,00	0,80	868	898	22	34	16	266	1,8	0,8
FV10_inra	0,00	0,01	0,03	0,00	0,00	0,00	0,23	0,73	843	880	28	32	14	253	-1,0	-0,9
EP4_csic	0,00	0,00	0,00	0,00	0,00	0,27	0,00	0,73	934	973	34	30	16	268	0,4	0,0
IL14H_usa	0,00	0,00	0,00	0,01	0,00	0,26	0,01	0,72	931	955	14	29	17	278	-1,0	1,2
Ak3_inra	0,00	0,02	0,00	0,00	0,13	0,01	0,12	0,71	896	932	28	30	15	259	-0,5	-0,6
EP47_csic	0,00	0,00	0,00	0,00	0,00	0,30	0,00	0,70	925	964	26	28	17	288	-1,1	1,3
W617_inra	0,00	0,02	0,00	0,00	0,00	0,29	0,03	0,66	849	883	28	35	13	257	2,5	-2,0
NY303_inra	0,00	0,01	0,02	0,00	0,00	0,31	0,00	0,66	859	901	32	32	16	266	0,6	1,1
RT10_inra	0,03	0,00	0,03	0,03	0,00	0,28	0,00	0,62	854	898	28	36	15	257	4,4	0,0
UH_1203_uh	0,26	0,00	0,00	0,00	0,00	0,14	0,00	0,60	860	881	15	32	15	246	-0,5	-0,3
UH_1278_uh	0,35	0,00	0,00	0,00	0,00	0,00	0,05	0,60	873	903	18	34	17	265	1,7	1,6
CoeSt6_inra	0,00	0,02	0,02	0,00	0,00	0,31	0,09	0,56	886	919	23	33	15	248	1,5	-0,2
EP16_csic	0,00	0,02	0,00	0,00	0,00	0,35	0,07	0,56	894	934	24	30	15	270	-0,5	-0,8
FC1571_inra	0,00	0,04	0,12	0,00	0,00	0,00	0,29	0,54	862	891	18	33	15	252	1,4	0,8
EP46_csic	0,00	0,01	0,03	0,00	0,00	0,45	0,00	0,51	880	944	48	29	15	258	-0,5	-0,2
ND1_inra	0,00	0,04	0,00	0,03	0,00	0,33	0,10	0,50	840	880	30	32	11	230	0,1	-3,4
UH_1291_uh	0,48	0,00	0,00	0,00	0,00	0,00	0,03	0,50	836	877	35	34	14	235	1,5	-0,4
FC25_inra	0,00	0,00	0,04	0,04	0,08	0,11	0,24	0,49	919	962	37	28	15	273	-1,6	-0,4
YUBR6_inra	0,04	0,00	0,05	0,00	0,00	0,41	0,08	0,43	892	937	31	32	13	246	1,4	-2,5
FC30_inra	0,00	0,00	0,09	0,00	0,11	0,00	0,39	0,41	859	876	12	35	14	252	2,5	-0,3

Table S4: Significant associations in the CF-Dent panel.

Trait	Name	Chr	Pos	MAF	-logP_K_Freq	-logP_K_Chr	Max-logP	effet
ASI_GDD6	PZE-101089397	1	80963531	0,07	2,79	2,92	4,88	3,19
ASI_GDD6	SYN25693	1	107883631	0,19	0,17	0,14	5,10	-0,29
ASI_GDD6	PZE-101143122	1	184242633	0,14	3,16	3,21	6,83	2,59
ASI_GDD6	PZE-102076429	2	57661247	0,33	2,82	2,84	6,59	1,87
ASI_GDD6	SYN5941	2	178262299	0,22	0,51	0,53	4,95	0,75
ASI_GDD6	PZE-102186878	2	230928395	0,08	4,19	4,12	7,54	4,31
ASI_GDD6	SYN31045	3	219899838	0,29	1,91	1,98	6,00	1,50
ASI_GDD6	SYN24156	4	1254252	0,13	0,06	0,09	4,89	0,12
ASI_GDD6	PZE-104082520	4	156714293	0,20	1,99	1,99	5,27	1,83
ASI_GDD6	PZE-104099884	4	176896645	0,13	3,76	3,77	4,91	3,11
ASI_GDD6	PZE-105041198	5	27247368	0,15	2,66	2,78	5,29	2,19
ASI_GDD6	SYN17491	5	142480380	0,19	1,43	1,49	5,21	1,60
ASI_GDD6	PZE-105108793	5	165742446	0,17	0,62	0,61	5,11	0,87
ASI_GDD6	SYN19125	5	178322096	0,41	2,47	2,33	5,45	-1,59
ASI_GDD6	PZE-106058040	6	106964908	0,10	3,74	3,76	5,64	-2,97
ASI_GDD6	SYN35090	6	107754798	0,08	1,10	1,12	5,13	1,74
ASI_GDD6	SYN34382	6	163178351	0,49	2,26	2,30	4,98	1,64
ASI_GDD6	PZE-107063100	7	120203200	0,29	2,36	2,31	5,42	1,81
ASI_GDD6	SYN14551	7	144205589	0,12	2,36	2,46	5,20	2,47
ASI_GDD6	PZE-108076600	8	131968792	0,16	0,79	0,83	5,20	1,12
ASI_GDD6	PZE-110044134	10	84047481	0,16	1,90	1,92	4,95	1,86
ASI_GDD6	PZE-110066651	10	122809735	0,10	0,17	0,14	5,09	0,44
DMC	PZE-102028065	2	13287473	0,39	1,75	1,78	5,05	0,31
DMC	PZE-102115993	2	153437719	0,24	3,41	4,19	7,22	-0,62
DMC	PZE-103019324	3	11529830	0,13	3,14	3,23	4,91	-0,79
DMC	PZE-103091384	3	150832948	0,48	5,26	5,41	5,54	0,74
DMC	SYN16049	3	160514830	0,23	1,71	1,86	5,15	-0,42
DMC	PUT-163a-148946654-487	3	211146412	0,17	1,21	1,24	5,14	0,33
DMC	PZE-103176862	3	222726491	0,47	1,54	1,55	4,91	-0,27
DMC	PZE-104093308	4	169801533	0,40	2,89	3,19	7,58	-0,51
DMC	PZE-104117587	4	193892648	0,16	1,54	1,76	5,07	0,40
DMC	PZE-104117602	4	193998611	0,16	1,54	1,76	5,07	0,40
DMC	SYN36949	4	233828118	0,47	4,70	4,99	4,99	-0,62
DMC	PZE-104146792	4	234820975	0,17	0,61	0,64	5,08	0,19
DMC	PZE-105056998	5	55222259	0,31	1,55	1,29	5,03	0,34
DMC	PZE-105134814	5	190732112	0,19	2,22	2,36	5,20	-0,56
DMC	PZE-106001730	6	2722897	0,07	1,07	1,05	4,89	-0,65
DMC	SYN37180	7	115969947	0,13	1,50	1,40	5,19	-0,48
DMC	PZE-107085004	7	140712419	0,41	3,02	3,16	5,42	-0,46
DMC	SYN2781	8	132204010	0,36	2,95	3,49	5,00	-0,44
DMC	PZE-109003341	9	3897669	0,45	1,09	1,12	5,12	0,24
DMC	PZE-110022209	10	31035820	0,29	2,42	2,59	5,01	-0,51
DMC	PZE-110022293	10	31219126	0,11	4,95	5,35	5,35	-1,05
DMCcorr	PZE-102028065	2	13287473	0,39	2,55	2,66	5,41	0,24

DMCcorr	PZE-102105700	2	132572326	0,31	0,51	0,61	4,87	0,10
DMCcorr	PZE-102146070	2	193189169	0,34	1,27	1,31	4,92	0,17
DMCcorr	PZE-103018185	3	10518979	0,07	4,73	4,87	4,87	-0,77
DMCcorr	PZE-103091384	3	150832948	0,48	1,75	1,78	5,31	0,24
DMCcorr	PZE-106013540	6	34603600	0,48	0,71	0,80	4,95	-0,12
DMCcorr	PZA00250.1	6	132889427	0,26	1,99	2,18	5,15	-0,27
DMCcorr	SYN35963	6	151581241	0,32	0,83	1,06	5,14	0,13
DMCcorr	PZE-107085004	7	140712419	0,41	4,35	4,51	5,91	-0,34
DMY	SYN11901	1	8515602	0,05	1,63	1,72	5,95	0,35
DMY	SYN28649	1	20216839	0,17	1,07	1,06	5,75	-0,19
DMY	SYN24828	1	20419410	0,17	1,07	1,06	5,75	-0,19
DMY	PZE-101032846	1	20529530	0,17	1,07	1,06	5,75	-0,19
DMY	SYN27887	1	20642806	0,17	1,07	1,06	5,75	-0,19
DMY	PZE-101197402	1	245790961	0,16	0,62	0,72	5,02	0,14
DMY	PZE-102140170	2	188360287	0,24	2,33	2,53	5,74	-0,28
DMY	SYN25546	3	158889565	0,36	2,29	2,30	5,16	0,26
DMY	PZE-104087710	4	162559844	0,45	0,91	0,95	5,57	-0,13
DMY	PZE-105006142	5	3090863	0,06	0,56	0,67	5,30	0,16
DMY	PZE-105076197	5	84345125	0,27	3,28	3,52	6,67	-0,31
DMY	PZE-105078037	5	88044014	0,17	0,87	0,95	7,41	-0,16
DMY	PZE-105134814	5	190732112	0,19	6,54	6,70	6,70	0,56
DMY	PZE-106025056	6	59883557	0,26	1,74	2,18	6,25	-0,22
DMY	PZE-106025268	6	60194317	0,26	1,97	2,42	6,65	-0,24
DMY	PZE-106025544	6	61014030	0,26	1,90	2,34	6,68	-0,23
DMY	PZE-106025785	6	61607500	0,23	2,27	2,72	4,93	-0,27
DMY	PZE-106025894	6	61754377	0,23	2,10	2,55	5,84	-0,25
DMY	PZE-106025970	6	61857563	0,23	2,26	2,73	6,27	-0,26
DMY	PZE-106026150	6	62179786	0,23	2,79	3,26	5,99	-0,30
DMY	SYN31342	8	25128932	0,41	0,70	0,91	5,20	0,11
DMY	PZE-108043407	8	72173318	0,44	2,07	2,35	4,96	-0,21
DMY	PZE-108056027	8	100939003	0,29	3,14	3,28	5,24	0,31
DMY	PUT-163a-76908411-3976	9	106399234	0,40	2,42	3,10	6,18	0,26
DMY	SYN8176	10	85188524	0,38	1,32	1,48	4,96	0,20
DMY	PZE-110048719	10	91358505	0,21	1,06	1,22	5,17	-0,18
DMY	SYN17973	10	95382609	0,39	2,54	2,83	5,52	-0,26
DMY	PZE-110054571	10	103737016	0,45	1,05	1,34	5,30	0,16
DMY	PZE-110054698	10	104287640	0,45	1,26	1,56	5,00	0,18
DMY	PZE-110054906	10	104920707	0,45	1,05	1,34	5,30	0,16
DMY	PZE-110054954	10	105232535	0,46	1,31	1,61	5,27	0,18
DMY	PZE-110055076	10	105787691	0,48	1,41	1,70	6,41	0,19
DMY	PZE-110062376	10	117571189	0,49	2,50	2,84	5,99	-0,25
DMYcorr	SYN28649	1	20216839	0,17	1,01	1,00	5,75	-0,17
DMYcorr	SYN24828	1	20419410	0,17	1,01	1,00	5,75	-0,17
DMYcorr	PZE-101032846	1	20529530	0,17	1,01	1,00	5,75	-0,17
DMYcorr	SYN27887	1	20642806	0,17	1,01	1,00	5,75	-0,17
DMYcorr	SYN20148	1	68555034	0,29	2,68	2,65	5,15	0,25

DMYcorr	PZE-101137930	1	178689818	0,07	3,12	3,24	5,86	-0,45
DMYcorr	PZE-101178657	1	223626792	0,36	0,80	0,86	5,00	0,12
DMYcorr	PZE-102097611	2	113206951	0,06	3,02	3,12	5,19	-0,52
DMYcorr	PZE-102140170	2	188360287	0,24	1,76	1,91	5,13	-0,22
DMYcorr	PZE-104087710	4	162559844	0,45	0,75	0,76	5,28	-0,10
DMYcorr	PZE-105006142	5	3090863	0,06	0,40	0,47	5,90	0,11
DMYcorr	SYN6476	5	59995577	0,22	1,44	1,51	5,00	0,19
DMYcorr	PZE-105076197	5	84345125	0,27	3,74	3,98	6,57	-0,30
DMYcorr	PZE-105078037	5	88044014	0,17	1,00	1,07	7,31	-0,16
DMYcorr	PZE-105134814	5	190732112	0,19	6,00	6,07	6,07	0,49
DMYcorr	SYN18865	6	62193454	0,23	2,82	3,25	5,48	-0,27
DMYcorr	PZE-108056042	8	100949959	0,30	3,02	3,14	5,08	0,28
DMYcorr	PUT-163a-76908411-3976	9	106399234	0,40	2,40	3,08	6,17	0,23
DMYcorr	PZE-110019212	10	23250814	0,06	2,86	2,87	5,81	-0,48
DMYcorr	SYN8176	10	85188524	0,38	1,98	2,25	4,87	0,23
DMYcorr	SYN17973	10	95382609	0,39	3,19	3,58	5,24	-0,27
DMYcorr	PZE-110054571	10	103737016	0,45	1,41	1,76	5,22	0,18
DMYcorr	PZE-110054698	10	104287640	0,45	1,62	1,98	4,93	0,19
DMYcorr	PZE-110054906	10	104920707	0,45	1,41	1,76	5,22	0,18
DMYcorr	PZE-110054954	10	105232535	0,46	1,67	2,03	5,20	0,20
DMYcorr	PZE-110055076	10	105787691	0,48	1,74	2,08	6,37	0,20
DMYcorr	PZE-110062376	10	117571189	0,49	2,39	2,67	6,04	-0,22
PLHT	PZE-101187908	1	233000449	0,09	0,55	0,62	4,99	-1,41
PLHT	SYN5941	2	178262299	0,22	4,88	5,10	5,41	4,11
PLHT	PZE-102137600	2	186447969	0,14	4,59	5,19	6,90	4,28
PLHT	PZE-102140170	2	188360287	0,24	3,61	3,97	5,87	-3,31
PLHT	PZE-105037558	5	22552184	0,48	3,28	3,42	6,16	2,66
PLHT	SYN10486	6	154322421	0,13	1,45	1,41	5,29	2,11
PLHT	PZE-107080558	7	135572251	0,18	3,34	3,43	5,25	3,62
PLHT	PZE-109070379	9	114964362	0,20	4,02	4,38	4,95	3,65
Silk_GDD6	PZE-102003082	2	2020244	0,43	1,85	2,06	5,53	-4,48
Silk_GDD6	SYN5941	2	178262299	0,22	7,42	7,57	7,57	12,97
Silk_GDD6	PZE-102137574	2	186441433	0,13	2,38	2,81	5,17	7,78
Silk_GDD6	PZE-102167180	2	211498614	0,18	2,67	2,72	5,79	-7,88
Silk_GDD6	SYN24935	2	234407011	0,37	3,71	3,87	4,94	6,46
Silk_GDD6	PZE-103091384	3	150832948	0,48	4,08	4,15	6,06	-8,50
Silk_GDD6	PZE-103091660	3	151341647	0,33	1,55	1,48	5,56	4,66
Silk_GDD6	PZE-103093822	3	154669325	0,07	4,06	4,22	5,22	14,89
Silk_GDD6	SYN25546	3	158889565	0,36	4,45	4,49	5,42	9,00
Silk_GDD6	PZE-104005694	4	1512170	0,07	3,67	3,57	4,96	14,09
Silk_GDD6	SYN4676	4	158146519	0,14	3,05	3,11	4,90	10,05
Silk_GDD6	PZE-104093308	4	169801533	0,40	4,73	4,98	5,21	8,77
Silk_GDD6	PZE-104104590	4	180785427	0,20	4,28	4,41	5,83	9,42
Silk_GDD6	PZE-104144719	4	233541810	0,34	2,42	2,56	6,14	5,57
Silk_GDD6	SYN36949	4	233828118	0,47	4,65	4,96	6,06	8,13
Silk_GDD6	PZE-105028662	5	14841088	0,06	2,55	2,82	5,05	11,57

Silk_GDD6	PZE-107065530	7	122130497	0,05	5,22	5,31	5,31	17,80
Silk_GDD6	PZE-108064653	8	115446396	0,14	5,42	5,90	7,34	12,59
Silk_GDD6	SYN3437	8	120459721	0,30	2,59	3,29	5,56	-6,29
Silk_GDD6	PZE-108068741	8	120768244	0,36	3,27	3,95	6,58	-6,50
Silk_GDD6	SYN10628	8	122465125	0,20	3,13	3,69	5,08	7,93
Silk_GDD6	PZE-108070194	8	123056834	0,47	3,39	4,05	5,08	-6,55
Silk_GDD6	PZE-108070380	8	123506141	0,27	7,83	8,86	8,86	11,97
Silk_GDD6	SYN23066	9	118046086	0,45	1,57	1,62	5,15	4,46
Tass_GDD6	PZE-101071092	1	53621515	0,37	2,28	2,47	5,40	5,08
Tass_GDD6	SYN39244	1	217212143	0,37	2,25	2,40	5,38	4,52
Tass_GDD6	PZE-102067477	2	44992791	0,31	1,97	2,42	4,96	4,82
Tass_GDD6	PZE-102075412	2	55829631	0,07	3,30	3,79	5,48	12,07
Tass_GDD6	SYN5941	2	178262299	0,22	7,07	7,22	7,54	11,68
Tass_GDD6	PZE-102167180	2	211498614	0,18	1,96	1,98	5,37	-6,06
Tass_GDD6	PZE-103012466	3	6649723	0,30	3,81	4,06	6,09	6,51
Tass_GDD6	PZE-103091384	3	150832948	0,48	5,11	5,23	6,12	-8,86
Tass_GDD6	PZE-103091660	3	151341647	0,33	2,94	2,82	6,05	6,31
Tass_GDD6	SYN25546	3	158889565	0,36	4,49	4,57	5,34	8,36
Tass_GDD6	SYN16049	3	160514830	0,23	1,81	1,80	4,89	5,34
Tass_GDD6	PZE-104093308	4	169801533	0,40	4,25	4,50	5,02	7,64
Tass_GDD6	PZE-104121926	4	198929997	0,30	2,94	3,36	5,24	6,44
Tass_GDD6	SYN11060	4	232934677	0,24	4,40	4,60	6,78	-7,09
Tass_GDD6	PZE-104144719	4	233541810	0,34	1,99	2,16	5,70	4,56
Tass_GDD6	SYN36949	4	233828118	0,47	5,78	6,29	7,69	8,36
Tass_GDD6	SYN1693	5	59291752	0,15	2,57	3,15	4,90	8,05
Tass_GDD6	SYN16484	5	188467017	0,41	1,16	1,29	5,51	2,83
Tass_GDD6	PZE-106041442	6	90548488	0,31	0,86	0,86	5,25	2,71
Tass_GDD6	SYN17287	7	119042427	0,20	3,80	3,88	5,92	7,53
Tass_GDD6	PZE-107065530	7	122130497	0,05	5,71	5,79	5,79	17,30
Tass_GDD6	PZE-107074700	7	130495196	0,07	3,46	3,46	4,93	10,55
Tass_GDD6	PZE-108000403	8	503482	0,29	1,06	0,94	5,23	3,14
Tass_GDD6	PZE-108064653	8	115446396	0,14	5,37	5,93	6,21	11,58
Tass_GDD6	SYN27932	8	118188472	0,39	5,36	5,76	5,76	7,80
Tass_GDD6	PZE-108070194	8	123056834	0,47	3,71	4,30	5,26	-6,34
Tass_GDD6	PZE-108070380	8	123506141	0,27	8,81	9,98	9,98	11,65
Tass_GDD6	PZE-108072699	8	126077120	0,37	4,56	5,74	5,74	6,88
Tass_GDD6	PZE-108072730	8	126287026	0,36	3,88	5,00	5,00	6,41

Table S5: Significant associations in the CF-Flint panel.

Trait	Name	Chr	Pos	MAF	-logP_K_Freq	-logP_K_Chr	Max-logP	effet
ASI_GDD6	SYN36074	1	44802906	0,14	3,35	3,42	5,21	2,67
ASI_GDD6	PZE-102110668	2	143259003	0,12	1,06	1,12	6,42	1,37
ASI_GDD6	SYN30953	2	202643907	0,11	1,49	1,46	5,61	-1,81
ASI_GDD6	PZE-104067263	4	133252441	0,41	0,70	0,74	5,17	0,71
ASI_GDD6	SYN2340	4	156955443	0,26	2,32	2,42	5,12	-1,77
ASI_GDD6	PZE-106063139	6	114481013	0,18	0,70	0,70	5,81	0,99
ASI_GDD6	PZE-107027539	7	32478358	0,08	5,40	5,68	5,68	-4,56
ASI_GDD6	PZE-107050502	7	99894530	0,24	4,79	5,09	5,09	-2,77
ASI_GDD6	SYN17065	7	140572906	0,34	2,18	2,48	4,96	-1,65
ASI_GDD6	PZE-108068669	8	120610321	0,33	0,64	0,75	5,20	-0,82
ASI_GDD6	SYN32327	9	27015269	0,16	1,50	1,48	5,69	1,56
ASI_GDD6	PZE-109051312	9	88794757	0,10	1,04	1,05	8,62	1,52
ASI_GDD6	PZE-109080576	9	128645534	0,06	1,72	1,74	5,17	2,75
ASI_GDD6	PZE-110036842	10	70312175	0,09	0,55	0,53	5,66	1,03
ASI_GDD6	PZE-110054216	10	102987075	0,09	0,90	0,93	5,70	-1,65
DMC	PZE-101070781	1	53414468	0,24	3,18	3,37	5,42	-0,62
DMC	PZE-101085247	1	74605720	0,47	1,90	1,92	4,91	-0,43
DMC	PZE-101103268	1	102950723	0,32	3,08	3,21	5,93	-0,58
DMC	PZA03580.2	1	175296774	0,35	3,78	3,85	5,46	-0,59
DMC	PZE-101160270	1	202420962	0,48	2,98	3,26	5,56	-0,50
DMC	PZE-104019337	4	19944254	0,06	2,54	2,62	5,69	0,84
DMC	PZE-104078745	4	152972399	0,17	5,30	5,33	6,28	-0,97
DMC	PZE-107045416	7	92726573	0,08	1,11	1,07	5,29	0,45
DMC	PZE-109044922	9	76536460	0,15	2,77	2,76	5,08	0,64
DMC	SYN1108	10	113164779	0,49	1,67	1,91	4,94	0,33
DMC	PZE-110085234	10	136958512	0,19	0,42	0,49	4,88	-0,16
DMCcorr	SYN3797	4	63825919	0,14	1,40	1,42	6,06	0,32
DMCcorr	ZM012702-0484	4	184782238	0,09	1,43	1,33	5,20	0,40
DMCcorr	PZE-107045416	7	92726573	0,08	1,19	1,22	6,20	0,35
DMCcorr	SYN18508	7	156213202	0,44	0,56	0,53	4,98	0,12
DMCcorr	PZE-108047916	8	80390227	0,41	0,79	0,98	6,02	-0,15
DMCcorr	SYN23829	9	37286714	0,43	2,52	2,75	4,87	-0,33
DMY	SYN9368	1	3407925	0,36	0,86	0,86	4,87	-0,15
DMY	SYN10537	1	17966974	0,23	4,62	4,66	5,20	-0,52
DMY	PZE-101122758	1	153344342	0,25	4,91	5,39	5,39	0,53
DMY	SYN13856	1	154077833	0,17	6,00	6,42	6,42	0,65
DMY	PZE-101182771	1	227438388	0,13	4,13	4,16	6,12	0,61
DMY	PZE-101205141	1	253793852	0,27	1,84	1,88	5,28	0,28
DMY	PZE-102118123	2	158375830	0,27	2,85	3,18	5,56	-0,36
DMY	PZE-106112449	6	159681296	0,13	0,91	0,92	5,42	-0,23
DMY	PZE-107106303	7	158237314	0,12	2,12	2,12	5,14	-0,40
DMY	SYN35860	8	22173140	0,23	1,84	1,80	6,06	-0,28
DMY	PZE-109091780	9	138892323	0,06	3,86	3,87	6,79	-0,77

DMY	SYN13602	10	85642012	0,41	3,24	3,34	5,45	0,35
DMY	PZE-110065697	10	121515142	0,31	2,73	2,88	5,67	0,33
DMYcorr	SYN9368	1	3407925	0,36	0,82	0,82	4,93	-0,13
DMYcorr	SYN10537	1	17966974	0,23	5,60	5,76	6,27	-0,52
DMYcorr	SYN101	1	18143655	0,27	3,83	3,84	4,87	-0,40
DMYcorr	SYN13856	1	154077833	0,17	5,20	5,45	5,45	0,55
DMYcorr	PZE-101182771	1	227438388	0,13	3,60	3,63	5,64	0,51
DMYcorr	PZE-101205141	1	253793852	0,27	1,37	1,40	5,24	0,21
DMYcorr	PUT-163a-71764007-3479	2	189633700	0,19	1,13	1,15	4,89	-0,21
DMYcorr	SYN10842	5	173161573	0,18	0,81	0,81	5,10	0,16
DMYcorr	PZE-105146456	5	199643263	0,41	4,66	4,70	5,26	0,38
DMYcorr	PZE-106112449	6	159681296	0,13	0,80	0,80	5,38	-0,19
DMYcorr	PZE-107091664	7	146548837	0,13	0,83	0,83	5,18	0,21
DMYcorr	PZE-107106303	7	158237314	0,12	1,97	1,97	5,17	-0,35
DMYcorr	SYN35860	8	22173140	0,23	2,18	2,17	6,25	-0,28
DMYcorr	PZE-109091780	9	138892323	0,06	4,40	4,41	6,71	-0,75
DMYcorr	PZE-110065697	10	121515142	0,31	2,67	2,77	5,73	0,29
PLHT	PZA03613.2	1	2941215	0,41	2,65	2,71	5,39	2,82
PLHT	PZE-101070781	1	53414468	0,24	4,63	5,05	5,43	5,30
PLHT	PZE-101102330	1	100904886	0,46	2,44	3,18	5,38	-3,26
PLHT	PZE-101122758	1	153344342	0,25	4,62	5,64	5,64	5,11
PLHT	SYN13856	1	154077833	0,17	5,31	6,10	6,51	6,03
PLHT	PZE-101129465	1	165231088	0,09	2,26	2,64	4,91	4,36
PLHT	PZE-101130308	1	166772606	0,08	3,60	4,02	5,57	6,53
PLHT	PZE-101199628	1	248736328	0,21	1,97	1,94	5,91	-3,02
PLHT	PUT-163a-4226354-2040	1	277740868	0,19	3,50	3,67	4,90	-4,20
PLHT	PUT-163a-74241827-3665	1	278514200	0,26	3,62	3,81	4,95	-4,01
PLHT	PZE-102068428	2	46205534	0,45	3,34	3,58	5,13	3,46
PLHT	SYN29939	3	216270627	0,43	2,55	2,42	5,33	3,12
PLHT	SYN34674	7	9541223	0,33	1,28	1,22	6,45	-2,07
PLHT	PZE-107028382	7	33915862	0,11	1,85	2,01	4,93	-4,24
PLHT	PZE-107099933	7	154625997	0,37	4,45	4,35	5,26	-4,11
PLHT	PZE-108049320	8	84808001	0,06	4,83	5,03	5,03	8,79
PLHT	SYN17872	8	101237704	0,14	5,96	6,12	6,12	6,43
PLHT	PZE-109073790	9	119310870	0,13	5,88	5,78	5,88	-7,25
Silk_GDD6	PZE-101070781	1	53414468	0,24	6,15	6,72	9,09	12,41
Silk_GDD6	PZE-101163301	1	206514625	0,13	1,21	1,33	5,13	5,11
Silk_GDD6	SYN16123	1	299406832	0,50	1,54	1,43	5,00	-4,38
Silk_GDD6	SYN300	1	300441295	0,36	5,10	5,44	5,92	-8,52
Silk_GDD6	PZE-102068514	2	46437351	0,08	3,33	3,74	5,02	13,30
Silk_GDD6	SYN14630	2	232987692	0,28	3,60	3,86	5,01	-6,95
Silk_GDD6	SYN10208	3	2745080	0,07	1,64	1,70	5,40	7,10
Silk_GDD6	PZE-103073710	3	122074590	0,07	2,40	2,44	5,41	14,19
Silk_GDD6	PZE-104050441	4	78545218	0,19	2,70	2,86	5,52	8,75
Silk_GDD6	PZE-107013193	7	9531124	0,10	3,32	3,42	5,08	13,13
Silk_GDD6	PZE-107126988	7	169443263	0,44	2,95	2,91	4,93	-7,21

Silk_GDD6	PZE-108105367	8	159981703	0,11	1,39	1,38	5,07	6,75
Silk_GDD6	SYN1030	8	162222740	0,13	4,65	4,75	6,67	13,59
Silk_GDD6	PZE-109081270	9	129469502	0,18	2,85	2,84	4,96	9,01
Silk_GDD6	PZE-110050012	10	94200010	0,34	3,27	3,68	4,91	7,85
Silk_GDD6	PZE-110051214	10	96393513	0,47	2,31	2,64	5,77	-5,99
Tass_GDD6	PZE-101070781	1	53414468	0,24	5,37	5,75	5,93	12,14
Tass_GDD6	PZE-101180507	1	225044054	0,15	1,64	1,78	5,25	6,77
Tass_GDD6	PUT-163a-74241827-3665	1	278514200	0,26	3,77	3,78	5,36	-8,56
Tass_GDD6	SYN300	1	300441295	0,36	4,45	4,68	5,74	-8,38
Tass_GDD6	PZE-103042081	3	39396493	0,17	3,47	3,52	4,97	10,49
Tass_GDD6	PZE-103052172	3	58297700	0,41	4,58	4,54	5,43	9,05
Tass_GDD6	PZE-103116753	3	176260144	0,21	3,37	3,36	5,26	9,69
Tass_GDD6	PZE-104010147	4	7174847	0,25	4,83	4,87	5,29	-10,03
Tass_GDD6	PZE-104050441	4	78545218	0,19	3,05	3,24	5,39	9,76
Tass_GDD6	PZE-107013193	7	9531124	0,10	3,28	3,38	5,53	13,50
Tass_GDD6	PZE-107071389	7	127608827	0,08	3,45	3,69	5,15	15,15
Tass_GDD6	PZE-107099933	7	154625997	0,37	3,44	3,50	5,11	-7,61
Tass_GDD6	SYN1030	8	162222740	0,13	4,51	4,55	8,02	13,95
Tass_GDD6	PZE-109038492	9	57188600	0,14	2,98	3,18	5,20	11,12

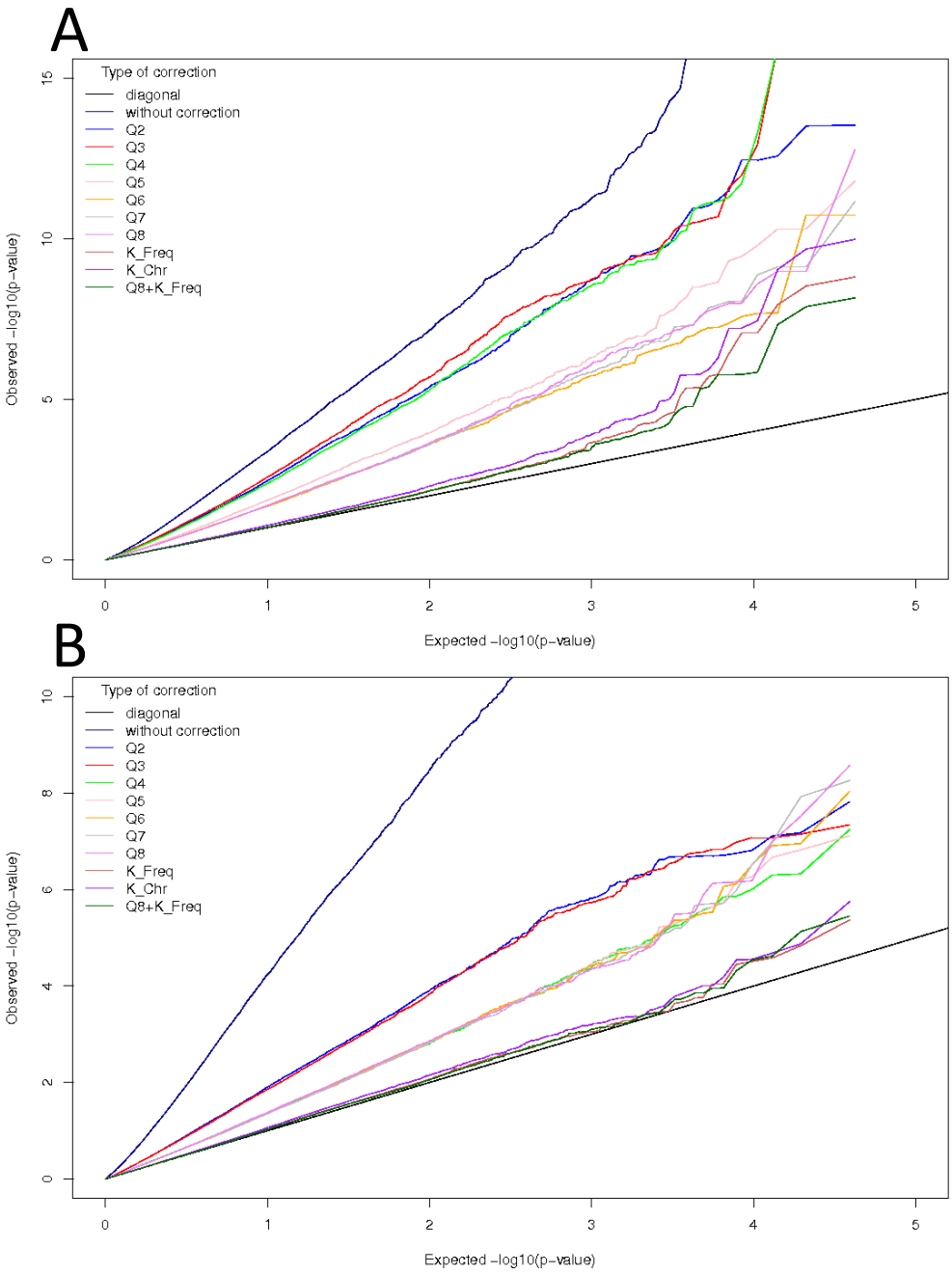

Figure S1: QQ-plot of Tass_GDD6 in the CF-Dent (A) and CF-Flint (B) panels, using different (Q+K) models and two different ways of estimating the kinship matrix (K_Freq and K_Chr).

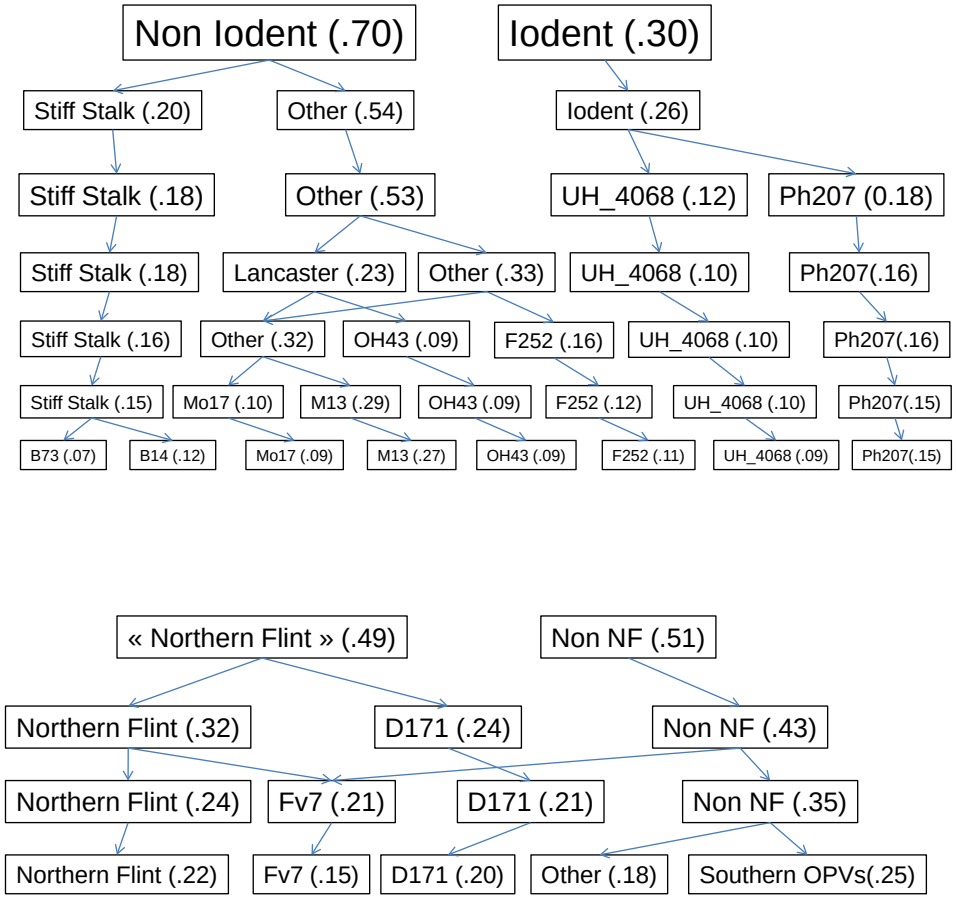

Figure S2: Admixture in the CF-Dent and CF-Flint panels from N_Q=2 to N_Q=8. Each group was called according to the pedigree of the lines. Frequency of each group are indicated in bracket. Arrows were drawn between groups sharing a high proportion of lines.

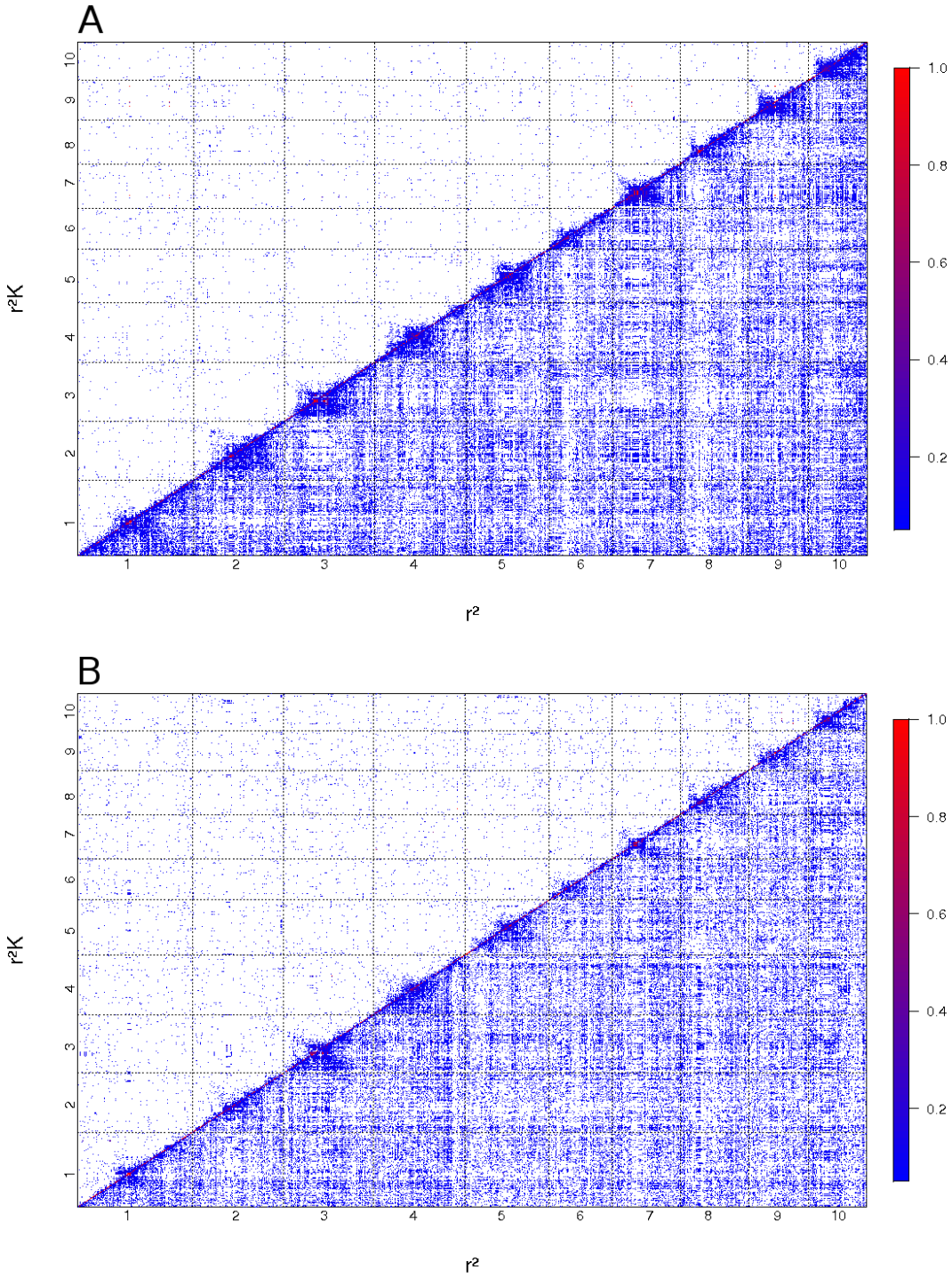

Figure S3: CF-Dent (A) and CF-Flint (B) panels estimated with 4000 markers sampled according to their physical position. Raw squared correlations (r^2) are represented below the diagonal, and r^2 corrected by relatedness (r^2K) estimated as K_Freq are presented above the diagonal. Cells corresponding to LD below 0.05 are in white. Markers were ordered according to their physical position.

Appendix III: supplemental chapter 3

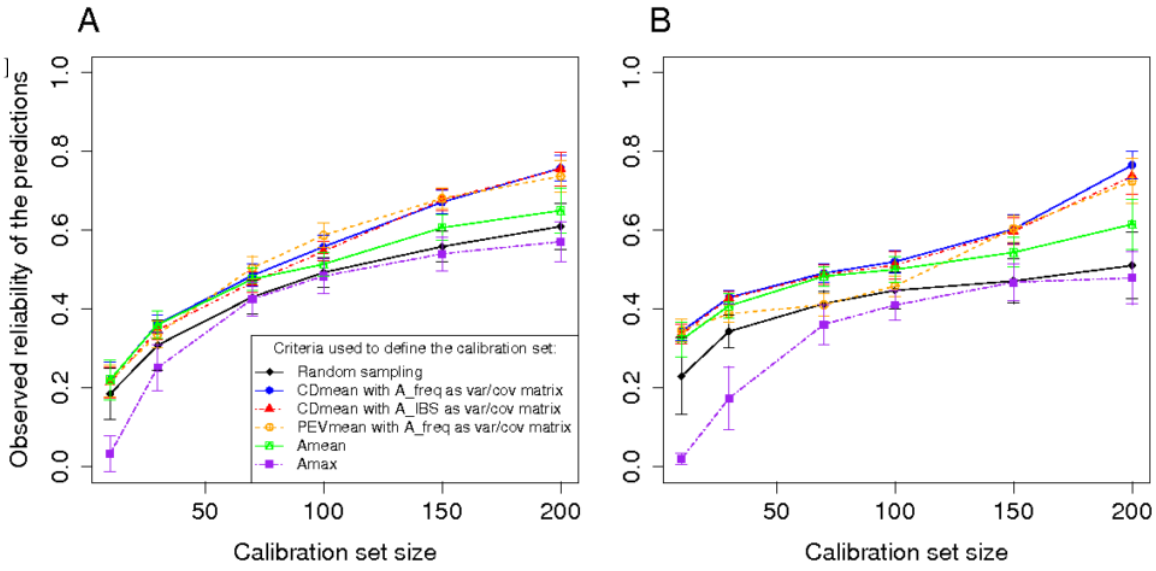

Figure S1 Reliability of the predictions of Tass_GDD6 using different sampling algorithms on the Dent panel (A) and the Flint panel (B) using a λ value corresponding to an heritability of 0.5. The calibration sets were randomly sampled, or defined by: maximizing CDmean with a relationship matrix based on the IBS or weighted by the allelic frequencies; minimizing PEVmean with a relationship matrix weighted by the allelic frequencies; minimizing the mean (Amean) or the maximum (Amax) of the relationship coefficient between the reference individuals. The individuals that are not in the calibration set are in the validation set. As a consequence for each calibration set size the reliability is calculated with a different number of individuals. For each point, the vertical line indicates an interval of $2\sigma_R$ (σ_R being the standard deviation of observed reliabilities over the 50 runs).

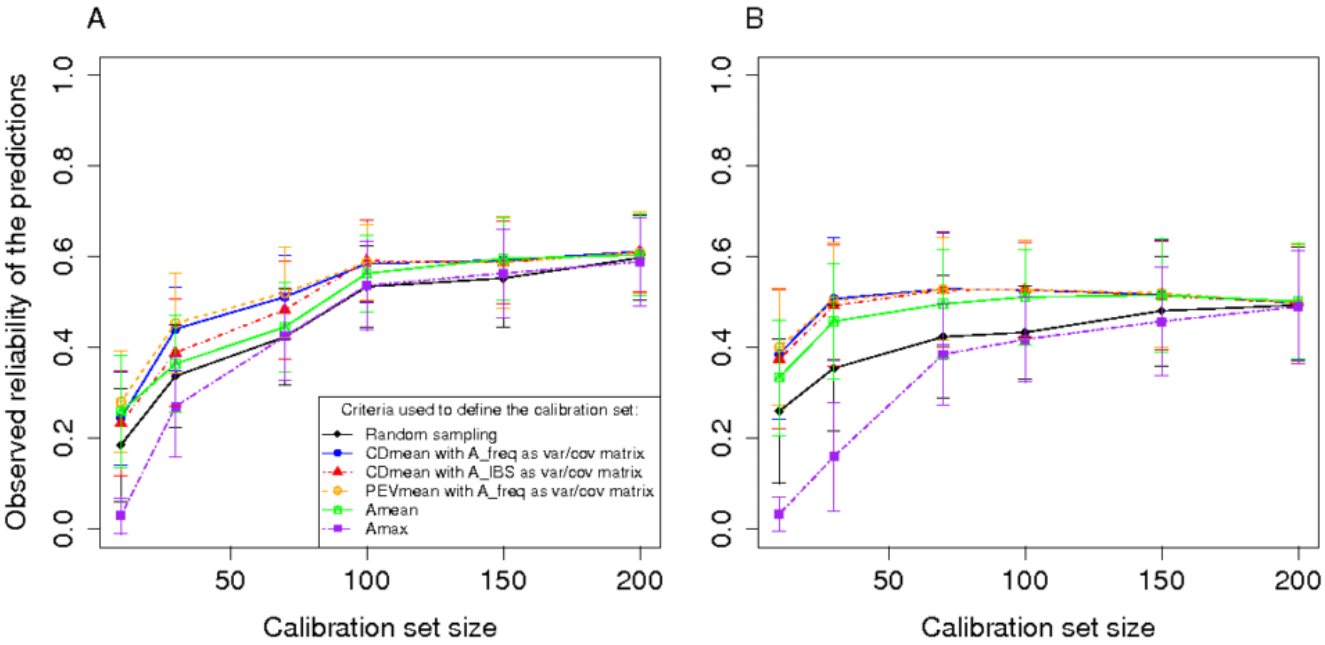

Figure S2 Cross-validation on the predictions of flowering time using different sampling algorithms in the Dent panel (A) and the Flint panel (B). In a first step 30 individuals are randomly sampled to constitute the validation set. In a second step calibration sets are sampled from the remaining individuals using different approaches to optimize the prediction reliability of the validation set. These calibration sets were randomly sampled, or defined by: maximizing CDmean with a relationship matrix based on the IBS or weighted by the allelic frequencies; minimizing PEVmean with a relationship matrix weighted by the allelic frequencies; minimizing the mean (Amean) or the maximum (Amax) of the relationship coefficient between the reference individuals. For each point, the vertical line indicates an interval of $2\sigma_R$ (σ_R being the standard deviation of observed reliabilities over the 50 runs). Optimization of PEVmean and CDmean was made with h^2=0.95.