Etude du déterminisme génétique de caractères quantitatifs chez les végétaux: Méta-analyse de QTL et études d'association.

Jean-Baptiste Veyrieras

Station de Génétique Végétale
UMR INRA CNRS INA P-G UPS-XI
Ferme du Moulon
Gif-sur-Yvette

Unité de Biométrie et
d'Intelligence Artificielle
INRA
Castanet-Tolosan

THÈSE

Présentée par

Jean-Baptiste Veyriéras

En vue de l'obtention du

Doctorat de l'Institut National Agronomique Paris-Grignon

ÉTUDE DU DÉTERMINISME GÉNÉTIQUE DE CARACTÈRES QUANTITATIFS CHEZ LES VÉGÉTAUX : MÉTA-ANALYSE DE QTL ET ÉTUDES D'ASSOCIATION

Soutenue le 27 Février 2006 devant le jury composé de :

André Gallais	Professeur à l'INA P-G	Président
Bernard Prum	Professeur à l'Université d'Evry	Rapporteur
Jean-Luc Jannink	Assistant-Professeur à l'Iowa State University	Rapporteur
Lounès Chikhi	Chargé de recherche au CNRS	Examinateur
Oliver Martin	Professeur à l'Université d'Orsay	Examinateur
Stéphane Robin	Professeur à l'INA P-G	Examinateur
Bruno Goffinet	Directeur de recherche à l'INRA	Directeur de thèse
Alain Charcosset	Directeur de recherche à l'INRA	Directeur de thèse

Table des matières

Résumé

L'étude de l'hérédité de caractères quantitatifs est au coeur de la génétique depuis l'avènement de cette science au siècle dernier. La génétique quantitative dispose désormais d'un cadre expérimental et théorique bien établi pour étudier les facteurs génétiques sous-jacents à l'hérédité de caractères complexes, que ce soit dans un but cognitif ou pour épauler les programmes d'amélioration variétale chez les espèces d'intérêt agronomique. Généralement, l'étude du déterminisme génétique de caractères quantitatifs s'articule autour de trois étapes majeures : i) déterminer les principaux locus impliqués dans la variation observée (dénommés QTL pour « Quantitative Trait Loci »), ii) identifier les gènes en cause, iii) discriminer les principales formes alléliques de ces gènes et évaluer leurs effets.

L'avènement des marqueurs moléculaires et de la génomique au cours des vingt dernières années a facilité la mise en oeuvre de dispositifs expérimentaux de cartographie de QTL en génétique végétale. Il est désormais possible d'avoir accès, via les bases de données publiques, à une masse importante de résultats de détection de QTL. Parallèlement, pour certaines espèces végétales, l'annotation de leurs génomes fournit une information de plus en plus riche et précise sur la structure et la fonction des gènes.

L'objectif de cette thèse était double. D'une part nous avons cherché à développer une nouvelle approche de type méta-analyse pour optimiser le croisement entre les données de QTL et celles issues de la génomique, dans le but de faciliter la recherche de gènes candidats. Une fois ces derniers identifiés, l'association entre leur diversité allélique et la variation du caractère peut être évaluée à l'aide de populations aux bases génétiques larges. Ces dernières approches étant relativement récentes chez les végétaux, la thèse a visé, dans un deuxième temps, à élaborer des méthodes adéquates au type de matériel généralement utilisé dans ce contexte : de la modélisation de la structuration génétique en passant par celle du déséquilibre de liaison intragénique, à leur intégration conjointe dans les tests d'association.

Mots clés : marqueurs moléculaires, QTL, méta-analyse, déséquilibre de liaison, structure génétique, études d'association, bioinformatique.

Abstract

The study of quantitative trait heredity has been a major goal of genetics since the beginning of the 20^{th} century. Quantitative genetics provides nowadays a well-established theoretical framework to explore the genetic factors underlying quantitative traits and its use in breeding programs has become commonplace for species of agronomical interest. Generally, the study of quantitative trait genetic determinism consists in 3 main steps : i) detect the major loci involved in trait variation, called Quantitative Trait Loci (hereafter QTL), ii) identify the underlying genes, iii) characterize the allelic diversity at these genes and evaluate their effects.

The advent of molecular markers and genomics since the 80's has tremendously enhanced the use of QTL mapping experiments in plants. Nowadays, a large number of QTL results has been made available via public databases. At the same time, the ongoing annotation of plant genomes provides a valuable information on the structure and function of genes.

First, this thesis has aimed at developping a meta-analysis framework in which both QTL and genomic data can be crossed in order to improve "candidate genes" selection. Then, association between candidate gene allelic diversity and trait variation can be evaluated using diverse germplasm collection. As association mapping techniques are quite recent in plants, new methodological developments have been done in order to deal with plant material typically involved in these studies : this ranges from genetic structuration and intragenic linkage disequilibrium modeling to the integration of these models in association mapping strategies.

Key words : molecular markers, QTL, meta-analysis, linkage disequilibrium, population structure, association study, bioinformatics.

Première partie

Introdution Générale

Chapitre 1

Prolégomènes

> "On ne peut être assez admiratif devant la simplicité des moyens par lesquels la nature s'est dotée de la capacité de varier à l'infini ses productions et d'éviter la monotonie. Un petit nombre d'entre eux, l'union et la ségrégation des caractères, combinés de diverses façons, peuvent conduire à un nombre infini de variétés."
>
> *Augustin Sageret (1763-1851)*

En 1866, lorsque le moine augustinien, Gregor Mendel (1822-1884), établit à l'aide de populations contrôlées de pois - méticuleusement développées dans les jardins du monastère de Brno en Moravie - les lois de l'hérédité de caractères discrets, la notion de déterminisme génétique demeure percluse par la concurrence de plusieurs théories de l'hérédité. A cette époque, la notion d'hérédité était souvent confondue avec des idées plus globales, et notamment celles relatives à l'évolution. Cela tient principalement à ce que l' "on ne parvenait pas alors à expliquer l'être vivant, et notamment sa formation, par le seul jeu actuel des lois physiques [...]." Pichot (1999). Cette singularité de l'être vivant, qui le différencie fondamentalement du monde des objets inanimés, et qui résiste tant aux assauts réductionnistes de la physique, a pour corollaire la variation. Autrement dit, définir l'hérédité ne pouvait se faire sans préciser au préalable l'origine et la nature de la variabilité, et des espèces, et des individus au sein de celles-ci. Certes, d'un chien naît un chien, et d'un chat un chat. Chaque être vivant est conditionné par les caractéristiques communes de son espèce. Mais si chaque enfant ressemble à ses parents, il s'en différencie également. Dès lors, comment rendre compte à la fois des ressemblances et des différences ?

Bien que sa nature et ses mécanismes sous-jacents n'aient été compris que tardivement, il est vraisemblable que, très tôt déjà, l'homme eut conscience de la relation particulière entre hérédité - en ce qu'elle constitue un phénomène apparent - et variation. Lorsque celui-ci se mit à cultiver des plantes ou à élever des animaux et que progressivement, désireux d'améliorer les récoltes

ou la production de viande ou de lait, il vint à sélectionner les meilleurs "souches" (végétales ou animales), il ne pouvait faire autrement que de postuler une hérédité. La Bible même témoigne de ce postulat lorsqu'elle met en scène la ruse de Jacob aux dépens de Laban (Genèse, chapitre XXX). Jacob propose ainsi à son beau-père de séparer de son troupeau, puis de lui confier, les brebis dont le lainage est de différentes couleurs ; et il conclut avec lui que tout ce qui naîtra d'un "noir mêlé de blanc" lui reviendra en récompense. Et celui-ci de s'arranger à ce qu'à la saison où les brebis de Laban sont en chaleur, elles ne se reproduisent qu'avec ses boucs tachetés. Le stratagème conduisit ainsi à la naissance d'un nombre toujours plus important de brebis au lainage tacheté, et "Il devint de cette sorte extrêmement riche, et il eut de grands troupeaux, des serviteurs et des servantes, des chameaux et des ânes". On trouve également d'autres exemples explicites de transferts de "qualités" entre procréateurs et enfants dans la mythologie et les rhapsodies grecques.

Bien que l'origine du mot "hérédité" revienne au latin (*hereditas* désignait les biens laissés par un romain à sa mort ainsi que leur processus de transmission), ce sont les Grecs qui furent les premiers à penser l'hérédité en tant qu'objet d'étude scientifique et qui proposèrent des théories faisant moins appel au sens commun et à la science populaire des mythes. La première, qui connut un vif succès au cours des siècles suivants, fut la théorie de la pangenèse[1] (étymologiquement : "engendrement par le tout") enseignée par le célèbre médecin grec Hippocrate (environ 460-370 avant J.-C.), qui s'inspira de la pensée du philosophe présocratique Anaxagore (environ 500-428 avant J.-C.). L'idée forte de la pangenèse repose sur le mélange entre les matériaux séminaux du père et de la mère, matériaux émis par toutes les parties de l'organisme. L'hérédité pour la première fois prenait "corps". Plus particulièrement cette conception évoquait déjà l'aspect particulaire ou atomiste de l'hérédité. Idée qui se diffusa aussi chez Epicure qui, dans son Jardin, évoquait à ses disciples ces "très petites particules" vectrices des caractères individuels.

Aristote s'opposa à cette conception de l'hérédité qui selon lui "réduisait le tout aux parties". Que ce soit dans *De generatione* ou *De partibus*, le philosophe macédonien défendit ce que plus tard l'on a appelé une vision holistique [2] de l'hérédité. Pour Aristote, la semence mâle, qui donnait forme à la substance inorganisée de la femelle (*catamenia*), apportait le principe générateur de la forme (*eidos*). Selon lui, cette *eidos* était immatériel - à rapprocher de l'âme. Bien que ce principe générateur ne soit pas sans rappeler les conceptions modernes de la fécondation, la théorie d'Aristote ne fut reprise que tardivement, vers la fin du XIXème.

Ni la période romaine, ni le Moyen-Âge, ne vit se développer de nouvelles

[1] dénommée aussi théorie de la panspermie.

[2] Holisme : Doctrine ou point de vue qui consiste à considérer les phénomènes comme des totalités

théories et il faut attendre la science officielle de la Renaissance pour voir ressurgir timidement un questionnement des principes fondamentaux de l'hérédité et de la fécondation. Au début du XVIIème siècle, bien que méconnue, l'embryologie cartésienne développée en marge du *Traité de l'Homme* et du *Discours de la méthode*, a conduit à une théorie de l'hérédité assez singulière. L'animal-machine de Descartes rencontra un vif succès auprès des "mécanistes" du XVIIème et du XVIIIème. Ces derniers, persuadés que l'"agitation" des particules séminales évoquée par le philosophe n'expliquait pas de manière satisfaisante la formation d'une chose aussi complexe et aboutie que l'être vivant, optèrent pour une théorie de la "préformation". Selon eux, l'individu était déjà préformé dans les "germes" (spermatozoïdes et ovules). Il ne lui restait plus alors qu'à grandir. Cette théorie fut complétée par celle de l'emboîtement des germes (dénommés homoncules chez l'homme) qui décrit l'hérédité comme un système vertigineux de "poupées russes". Le préformationnisme, qui aujourd'hui nous apparaît bien fantaisiste, fut soutenu par des scientifiques éminents comme Nicolas Malebranche (1638-1715) et Gottfried Wilhelm von Leibniz (1646-1716) et conserva des partisans jusqu'au début du XIXème. Paradoxalement, cette théorie s'apparente presque à une négation de l'hérédité : Dieu, qui depuis la nuit des temps aurait déjà tout "emboîté", est ici substitué à l'individu.

Exception faite du préformationnisme, aucun autre système élaboré ne fut alors proposé pour expliquer la nature de l'hérédité. Cette apparent désintérêt s'explique surtout par la conception dominante du monde visible empruntée à la philosophie "essentialiste" de Platon. Pour le philosophe, la variabilité apparente n'est que le reflet d'un nombre limité de formes invariables, appelées *eide*. Au Moyen-Âge, les disciplines de Saint-Thomas d'Aquin - les thomistes - substituèrent au mot grec le substantif d'*essence*. L'essentialisme prédomina fortement dans la théologie et la philosophie jusqu'au XIXème. Dans ce contexte philosophique, l'hérédité est une évidence et non un problème scientifique : tous les membres d'une même espèce partagent la même essence, les cas atypiques et rares (c'est à dire les "variants") ne traduisant au fond qu'une manifestation imparfaite de celle-ci.

Un des plus célèbres essentialistes fut le naturaliste suédois Carl von Linné (1707-1778). Homme de science très pieux, partisan du fixisme [3], il fut l'un des pères fondateurs de la taxonomie. Pour les essentialistes, l'enjeu résidait davantage dans la classification des espèces visibles que dans la compréhension des mécanismes ontologiques (surtout chez les fixistes pour qui l'origine est une préoccupation théologique et non pas scientifique). Bien que Linné qualifiait la variabilité au sein d'une espèce, d'accidentelle et sans conséquence sur l'essence même de celle-ci, il formalisa tout de même le concept de variété : "Il y a autant de variétés qu'il y a de plantes différentes

[3]Fixisme : Théorie selon laquelle les espèces vivantes sont immuables parce que dotées, dès l'origine [par Dieu], de tous les mécanismes nécessaires à leur mode de vie

produites par les graines d'une espèce. Une variété est une plante changée par une cause accidentelle : le climat, le sol, la température, les vents, etc." (*Philosophia Botanica*, cité par MAYR (1982)).

A la même époque, le mathématicien et astronome Pierre Louis Moreau de Maupertuis (1698-1759), lui aussi essentialiste, et farouche opposant à la théorie de la préformation, insista sur la contribution égale du père et de la mère à l'hérédité, et tâcha de formaliser le concept de la pangenèse. Remarquons qu'il adopta une démarche expérimentale proche de la génétique dans son acception méthodologique la plus moderne, en étudiant la transmission de la polydactylie [4] au sein de différentes généalogies (sur quatre générations).

Un siècle plus tard, l'avènement de la cytologie et l'histologie, en partie grâce aux travaux du zoologiste Théodor Schwann (1810-1882) et du botaniste M.J. Schleiden (1804-1882) - discipline qui doit beaucoup à l'invention du microscope optique vers 1668 par le néerlandais Antoni van Leeuwenhoek (1622-1723) - permit à Charles Darwin (1809-1882) d'affiner la thèse de Maupertuis en l'immergeant dans le cadre de la théorie cellulaire. Il introduisit la notion de gemmules, décrivit comment ces gemmules étaient "transportées" au sein de l'organisme jusqu'à leur accumulation dans les cellules germinales. Lors de la fécondation, l'enfant héritait des gemmules de son père et de sa mère et leurs expressions déterminaient ainsi ses caractéristiques[5]. Darwin formula sa théorie sous le titre prudent d'"hypothèse provisoire à la pangenèse". En fait, il s'agissait d'avantage d'une synthèse de plusieurs théories en vogue à son époque que d'un programme ambitieux d'explication de l'hérédité.

En particulier, Darwin souscrivait en partie à la thèse de l'hérédité par mélange, soutenue fortement par Karl Wilhelm von Nägeli (1817-1891) [6]. Cette théorie, qui ne rejettait pas le caractère particulaire, atomiste des déterminants de l'hérédité, permettait d'expliquer l'apparence intermédiaire des caractères de certains descendants, dits hybrides (par exemple, les éleveurs d'animaux savaient qu'en croisant des espèces géographiquement différentes, elles se "mélangeaient"). Dans la théorie pangénétique de Darwin, cela se traduisit par la possibilité, lors de la fécondation, de la fusion de gemmules maternelles et paternelles, conduisant ainsi à la création d'un caractère intermédiaire. Mais si Nägeli prônait une hérédité exclusivement par mélange, Darwin lui demeurait plus incertain : "il serait plus juste de dire que les éléments des deux espèces parentales se présentent dans chaque hybride sous deux états, c'est à dire, soit fusionnés et mélangés, soit complètement

[4]Polydactylie : Malformation héréditaire caractérisée par la présence d'un ou de plusieurs doigts ou orteils en surnombre

[5]Cette formulation lui permettait également d'expliquer l'atavisme, phénomène cher à ses yeux, par le truchement de gemmules "dormantes" qui pouvaient se "réveiller" après plusieurs générations.

[6]L'histoire retient que de la correspondance assidue que Nägeli entretint avec Mendel, le botaniste suisse ne daigna pas accorder d'importance aux résultats révolutionnaires du moine morave.

séparés." (*Variation of Animals and Plants*, cité par MAYR (1982)).

Il n'est pas étonnant que l'auteur de *The Origin of Species*, qui marqua et marque encore la pensée contemporaine, se fut ainsi intéressé au problème de l'hérédité. S'il était nécessaire de fournir une explication plus empirique que philosophique à l'origine de la diversité des espèces, il fallait également préciser de quelle manière cette variabilité avait pu se perpétuer dans le temps. L'hérédité assurait alors ce lien essentiel entre variabilité et continuité. Toutefois, si le concept particulaire adopté par Darwin expliquait comment les caractères d'une espèce se transmettait d'une génération à l'autre, il ne rendait pas compte de l'apparition de nouvelles "variétés", notion pourtant capitale dans sa théorie évolutive. Pour Darwin les causes possibles de la variation se divisaient en deux facteurs principaux : l'effet de l'environnement et celui de l'usage et du non-usage (sous entendu d'un organe). Comme beaucoup de ces contemporains, il souscrivait ainsi à la notion de ce que l'on a plus tard qualifié d' "hérédité flexible", selon laquelle le "bagage" héréditaire pouvait être éventuellement modifié sous l'effet du milieu ou de l'environnement.

Le cousin de Darwin, Francis Galton (1822-1911), père de la biométrie, est peut-être le premier qui, en son temps, remit le plus ouvertement en cause cette conception de l'hérédité. Persuadé du caractère "inflexible" de l'hérédité il développa des théories innovantes. Par exemple, sans avoir connaissance des travaux de Mendel, il proposa une explication des caractères des hybrides en introduisant le concept de ségrégation entre des unités particulaires de l'hérédité (les *strip*). Insistant sur l'unicité des individus, Galton développa une pensée populationnelle qui lui permit de poser les bases d'une authentique statistique des populations (en inventant par la même occasion un concept majeur en statistiques : la régression). Mais ni Galton, ni Darwin n'eut, à cette époque, connaissance des avancées spectaculaires de la cytologie en Allemagne. Avancées qui allaient s'avérer décisives dans l'établissement d'une nouvelle science : la génétique.

Les travaux du biologiste et médecin allemand, Friedrich Leopold August Weismann (1834-1914) constituent sûrement le point marquant du progrès qui s'opéra alors, autour de la conception de l'hérédité. Distinguant avec lucidité les constituants du "génotype" de ceux du "phénotype" (deux mots non encore forgés, mais qui se rapprochent des notions de *soma* et de *germen* proposées par Weismann), il définit en 1899 les déterminants génétiques comme "des unités actives du vivant, intervenant de manière spécifique dans le développement, c'est à dire d'une façon telle que soit produit le caractère dont ils sont les déterminants." (*Das Keimplasma*, cité par MAYR (1982)). Il dénomma biophore (étymologiquement "porteuse de vie") cette unité fondamentale dont l'expression détermine un caractère bien précis et parla de *déterminant* pour désigner une composition spécifique de biophores.

Parallèlement aux travaux de Weismann, et à son concept de plasma germinatif (*Das Keimplasma*), le botaniste hollandais Hugo de Vries (1848-

1935), introduisait dans *La Pangenèse intracellulaire* la théorie des pangènes, unités fondamentales et supports matériels de l'hérédité qui, selon lui, permettaient d'expliquer la nature composite des caractères d'une espèce : "Ces facteurs [les pangènes] sont les unités que la science de l'hérédité a à étudier. Exactement comme la physique et la chimie remontent aux molécules et aux atomes, les sciences biologiques ont à pénétrer jusqu'à ces unités pour expliquer, par leurs combinaisons, les phénomènes du monde vivant." (*La Pangenèse intracellulaire*, cité par PICHOT (1999)). La théorie de De Vries est sans doute plus proche des conceptions actuelles qu'aucune de celles qui l'avaient précédée. Puis au printemps de l'année 1900, De Vries, le botaniste Carl Erich Correns (1864-1933) et l'agronome autrichien Erich von Tschermak (1871-1962), publièrent de manière rapprochée des articles dans lesquels tous trois affirmaient avoir découvert les lois de l'hérédité, tout en précisant que ces lois avaient déjà été constatées 35 ans auparavant par un dénommé Gregor Mendel. Ce printemps 1900, et la redécouverte des lois de l'hérédité de Mendel, marqua ainsi la naissance de la science génétique - le mot en revanche ne fut inventé que plus tard, en 1906, par le généticien anglais, William Bateson (1861-1926).

Mais les lois de la ségrégation mendélienne n'expliquaient pas l'apparition de "nouvelles variations". Pourtant, la clef de voûte de la théorie de l'évolution de Darwin, désormais admise et reconnue, reposait sur l'abondance de variations à partir desquelles la sélection avait prise. Réfutant l'hérédité "flexible" dans leurs théories, ni De Vries, ni Weissman n'avançaient pourtant des arguments satisfaisants pour expliquer les variations continues. Il restait donc à expliquer les allèles, et plus particulièrement, la variabilité allélique[7].

La solution apparue par le biais d'études, de prime abord plutôt opposées, sur les variations rares. De Vries observa lors d'essais expérimentaux en champs, l'apparition d'individus aberrants. Il les isola et montra que leurs caractères singuliers étaient héritables lors de croisements successifs. Toutefois, les déductions qu'en fit De Vries peuvent, rétrospectivement, nous apparaître surprenantes. En effet, il ne présenta pas sa *Théorie de la mutation* (1901 pour le tome I et 1903 pour le tome II) [8] comme une théorie de la variation héréditaire mais comme une nouvelle théorie de l'évolution, concurrente à celle de Darwin. Ainsi De Vries distinguait les variations continues et quantitatives (variations qui suivent la loi de Adolphe Quételet (1796-1874)[9]), qu'il dénomma "fluctuations", des variations brusques et qualitatives, qu'il qualifia de "mutations". Selon lui, les variations continues n'interviennent

[7]Pour éviter tout anachronisme, il faut rappeler que le mot *allèle* fut inventé plus tard, en 1906, par Bateson. Plus précisément, il introduisit le terme d'*allélomorphe* qui se simplifia, par l'usage, en *allèle*.

[8]De Vries n'inventa pas le mot mutation, ce terme était déjà utilisé dès le XVIIème pour désigner des changements d'aspects des fossiles

[9]Qui n'est autre que la loi de Gauss-Laplace, aujourd'hui communément appelée loi normale.

pas dans l'évolution, s'opposant ainsi à Darwin chez qui seules les petites variations continues sont à prendre en compte. Pour de Vries au contraire, l'évolution ne peut se faire que par saut qualitatif, de manière saltatoire. De plus, cela le conduisit à amoindrir le rôle de la sélection : chez Darwin, pour avoir prise, la sélection opère sur une importante variabilité continue, alors que pour le botaniste hollandais, la sélection n'intervient qu'après la mutation, et seulement si elle est nécessaire. Autrement dit, à la continuelle "lutte pour la vie", De Vries substitue le hasard et la discontinuité des mutations.

Pourtant, aussi étonnant que cela puisse nous paraître aujourd'hui, De Vries n'établit pas de lien direct entre la mutation et les lois de Mendel qu'il venait de redécouvrir. Il faut ici rappeler que De Vries, faute de délimiter clairement le phénotype du génotype, nommait en fait mutation la modification de la "forme" de l'être. Selon lui, la mutation apparaissait de manière unique, isolée, et se traduisait par la création *de novo* d'un pangène, et non par l'altération d'un pangène préexistant.

La *Théorie de la mutation* constitua un tournant dans la manière d'étudier l'hérédité, changement radicale qui se mesure à l'aune de la polémique que cette théorie allait alors déclencher. Mais les progrès concomitants de la biochimie ne facilitèrent pas *a priori* sa défense. En effet, les conclusions récentes de la biochimie firent tomber en désuétude la notion de "particules élémentaires vivantes", et les pangènes se trouvèrent alors privé de "réalité", faute de support physique. Jusqu'à présent ils étaient porteurs de l'hérédité dans la mesure où ils étaient les composants élémentaires de la matière vivante. Dès lors, le pangène ne tenait plus sa réalité que du caractère qui lui correspond. Sur le plan scientifique, cela conduisit à un renversement étonnant de la définition de l'hérédité : sa nature physique se dissolva - pour un temps - au profit de sa discrétisation en autant de caractères observables.

Cette nouvelle voie de l'étude de l'hérédité ouverte par De Vries entre 1880 et 1910, va être suivie notamment par le danois Wilhelm Johannsen (1857-1927). En 1909, soucieux de désigner plus distinctement les composantes de l'hérédité, il forgea le mot gène en sacrifiant le préfixe "pan" du pangène de De Vries. Il insista alors sur la nature "calculatoire" du gène, proscrivant ainsi toutes velléités d'interprétation spéculative. En outre, il introduisit deux notions, désormais au coeur même de la génétique moderne, et qui sont fondamentales à la compréhension de l'hérédité et du développement des organismes : le phénotype (à partir du grec "paraître") et, de manière similaire, le génotype (mot formé de "gène" et de "type"). Le génotype devint ainsi le patrimoine génétique d'un individu et son phénotype le corps dans lequel ce patrimoine a été traduit au cours du développement. Ces définitions dérivaient des travaux que Johannsen effectua sur les lignées pures (notion aussi qu'il inventa) de haricots (*Phaseolus vulgaris*) : après de nombreuses générations d'auto-fécondations successives qui auraient dû conduire à des haricots génétiquement identiques, il continuait cependant à

observer de la variabilité : il en déduisit que le phénotype ne représentait pas de manière fidèle le génotype. Aussi, féru des méthodes quantitatives et de l'analyse statistique, il définit le phénotype comme la valeur statistique moyenne de l'échantillon. Implicitement, il mit en évidence l'importance des interactions entre le génotype et l'environnement. Autrement dit, Johannsen fut le premier qui départagea ce qui relevait de l'apparence mesurable et ce qui relevait de l'hérédité sous-jacente.

Bien qu'à la lumière des connaissances d'aujourd'hui, le travail et la pensée de Johannsen nous apparaissent comme les prémisses prometteurs de la génétique moderne, il dut affronter les critiques vigoureuses des biométriciens et de leur éminent chef de fil, Galton. La controverse entre les disciples de Galton (de tendance holiste) et les "mendélo-mutationnistes" auxquels appartenait Johannsen, s'alimentait principalement de l'apparente incapacité de ces derniers à expliquer la variation de caractères continus. On trouvait bien chez De Vries une tentative d'explication de l'hérédité de caractère quantitatif par un nombre d'exemplaires fluctuant du pangène correspondant, mais cette conception était désormais totalement incompatible avec la notion de gène telle que l'avait introduite Johannsen. Ainsi, la polémique enfla autour de l'interprétation de la distribution d'un caractère selon la loi de Quételet. Pour Johannsen, le fait que la descendance des individus choisis à l'extrémité de la courbe tendait statistiquement à revenir vers la moyenne indiquait clairement une origine non génétique de la variation. Galton, lui, était persuadé du contraire et chercha à expliquer cette continuité par une théorie que, plus tard, Pearson dénomma la théorie de "l'hérédité ancestrale".

A l'instar de son cousin Darwin, Galton était fasciné par la résurgence des caractères ancestraux. Aussi, sa théorie de l'hérédité ancestrale se voulait à la fois une explication de l'atavisme et du phénomène de "régression" des extrêmes vers la moyenne de la population. Selon sa théorie, incompatible avec le mendélisme, un individu n'héritait qu'en proportion $1/2$ des caractères de ses parents directs, l'autre moitié se décomposant en contributions successives des ancêtres en proportions $(1/2)^n$ (les quatre grands-parents contribuent à $1/4$ de l'hérédité, les huit arrières grands-parents à $1/8$, et ainsi de suite). D'autres que Galton, à la même époque, tenteront d'expliquer l'hérédité des caractères à variation continue . L'américain William Ernest Castle (1867-1962) développa une théorie de la contamination à partir d'observations empiriques faites sur des cobayes albinos (théorie définitivement réfutée en 1919). Forte des avancées de la cytologie, une théorie cytoplasmique de l'hérédité vit aussi le jour : elle suggérait que la variation continue d'un caractère émanait du cytoplasme. Plus particulièrement, elle était due à une substance diffuse, particulière à chaque espèce, présente dans le cytoplasme, et indépendante des gènes mendéliens discontinus. Elle fut également rapidement réfutée par de nombreux arguments en sa défaveur. Notamment, ces opposants mirent en avant que les contributions cytoplasmiques des parents pouvaient être très différentes, et cependant leurs contributions génétiques rester strictement

identiques.

Tandis que l'une après l'autre, ces théories non mendeliennes de l'explication de l'hérédité de caractères quantitatifs étaient réfutées, se dessina l'idée qu'à un phénotype pouvait correspondre plusieur gènes. Ironie à nouveau de l'histoire, cette hypothèse multifactorielle de l'hérédité avait déjà été évoquée par Mendel lui même. Autrement dit, on démontra alors que la combinaison discontinue de différents gènes pouvait contrôler la variation continue d'un phénotype. En 1910, l'agronome suédois Nilsson-Ehle fut le premier à démontrer expérimentalement la transmission héréditaire de caractère quantitatif chez le blé (à partir d'une étude sur la couleur des grains), et fut suivi de peu par E.M. East chez le maïs, et de Charles Benedict Davenport (1866-1944) chez l'homme. Désormais, la génétique mendélienne était en mesure d'expliquer et d'étudier la variation continue des biométriciens. L'étude de l'hérédité multifactorielle (également dénommée polygénisme) s'est alors constituée comme une discipline a part entière à partir de 1918, grâce notamment aux travaux de Sir Ronald Aylmer Fisher (1890-1962). La génétique quantitative "formelle" était née.

Il ne manquait plus alors qu'une théorie unifiée dans laquelle les lois de Mendel et les découvertes récentes en histologie de la chromatine (le terme de "chromatine" avait été donné par Walter Flemming (1843-1905) en 1879 au matériel contenu dans le noyau de la cellule) puissent se rejoindre. C'est Thomas Hunt Morgan (1866-1945) qui paracheva l'édifice en publiant *La Théorie du gène* en 1926. Morgan et son équipe, évitant toute spéculation sur la nature physico-chimique du gène, imposèrent, par une démarche expérimentale convaincante et ingénieuse sur la drosophile (*Drosophila melanogaster*), une vision linéaire des chromosomes sur lesquels chaque gène occupe une place identifiable, dénommée *locus*. Plus précisément, voici comment Morgan lui-même introduit sa théorie : "La théorie établit que les caractères de l'individu se rapportent à des paires d'éléments (gènes) dans le matériel germinal, éléments qui sont réunis en un nombre défini de groupes de liaison. Elle établit que les membres de chaque paire de gènes se séparent lors de la maturation des cellules germinales en accord avec la première loi de Mendel, et qu'en conséquence chaque cellule germinale contient seulement un jeu. Elle établit que les membres appartenant à différents groupes de liaison s'assortissent indépendamment les uns des autres an accord avec la seconde loi de Mendel. Elle établit qu'un échange ordonné - le crossing-over - se produit parfois entre les éléments de groupes de liaison correspondants ; et elle établit que la fréquence des crossing-over met en évidence l'ordre linéaire des éléments dans chaque groupe de liaison, et la position relative des éléments les uns par rapport aux autres." (*La Théorie du gène*, cité par Pichot (1999)).

La génétique progressa alors très rapidement jusque dans les années 50. A cette date, ses acquis principaux pouvaient se résumer en sept grands points :

1. le matériel génétique (le génotype) se décompose en unités appelées gènes,
2. les caractères (le phénotype) résultent de l'expression d'un ou plusieurs gènes situés à certains locus sur les chromosomes,
3. une vision linéaire des chromosomes,
4. le principe diploïde : un individu hérite d'une paire de chromosomes, l'un de la mère, l'autre du père,
5. la notion de génome : "Les gamètes, contenant un lot de chromosomes, portent un jeu complet de gènes, un génome. L'œuf et toutes les cellules somatiques, contenant deux chromosomes de même sorte, portent deux génomes complets, l'un d'origine paternelle, l'autre d'origine maternelle" (Jean Rostand (1894-1977), *Introduction à la génétique*, 1936).
6. une mutation est un changement brusque d'un gène,
7. plusieurs gènes peuvent contribuer à l'expression d'un seul caractère (polygénisme), et un seul gène peut affecter plusieurs caractères (pléiotropie).

Mais si le formalisme apporté par Morgan à la théorie de l'hérédité dotait la génétique d'un cadre expérimental séduisant, en excluant toute spéculation physico-chimique sur la nature du gène, il risquait de cantonner cette science à sa seule dimension empirique, voire "semiologique". Or, dès 1869, les travaux de Friedrich Miescher (1844-1895) avaient permis d'isoler et d'identifier, au sein des cellules, le matériel du noyau, et plus particulièrement les acides desoxyribonucléique (ADN). Quelques années plus tard, Flemming avancera l'hypothèse que l'ADN est le constituant principal des chromatines, et au début du siècle, E.B. Wilson abondera dans son sens lorsque dans son ouvrage *The Cell* (1900), il écrira : "La chromatine n'est probablement rien d'autre que la nucléine [...]. Les données relatives à la maturation, à la fécondation et à la division cellulaire soutiennent l'idée que la substance nucléaire, et surtout la chromatine, est un facteur déterminant de l'hérédité." (cité par Mayr (1982)). Enfin en 1924, Feulgen et Rossenbeck confirmeront que l'ADN est l'unique constituant des chromosomes.

Ainsi, entre le début du siècle et les années 50, s'accomplit la convergence, qui sera décisive par la suite, entre la génétique formalisée par Morgan et la biochimie. Cette correspondance conceptuelle se traduisit notamment par le désormais célèbre aphorisme du microbiologiste Archibald Garrod (1857 - 1936), généralisé par George Wells Beadle (1903 - 1989) : à "un gène, une enzyme". Puis en 1928, les recherches de Frederik Griffith (1877-1941) sur un vaccin contre la pneumonie le conduisirent à observer que des souris à qui on inoculait à la fois une forme non virulente de pneumocoques et des pneumocoques morts, tués au préalable par la chaleur, ces souris en mouraient. Il supposa alors que la résurgence de la forme virulente des pneumocoques se faisait par l'entremise d'une substance physico-chimique dénommée "virule". L'expérience fut reproduite quelques années plus tard *in*

vitro, une première fois par Martin H. Dawson et Richard H.P. Sia en 1931, et surtout par Oswald Avery (1877-1955) en 1933. Ce dernier, aidé de ces collaborateurs, démontra que ces virules n'étaient autres que des fragments d'ADN. Rétrospectivement, ces premières manipulations génétiques *in vitro* constituent la première preuve incontestable que l'ADN est le support de l'hérédité. Mais, Avery n'osera pas tirer de ses résultats une conclusion aussi nette et restera prudent quant à leur généralisation.

Cette prudence s'explique principalement par l'"enzymomanie" de l'époque. Les contemporains d'Avery ne voyaient dans l'ADN qu'un polymère "monotone" qu'ils supposaient sans forte variabilité entre espèces. De plus, le trop faible poids moléculaire des acides nucléiques, comparé à celui des acides aminés des protéines, le rendait impropre à leurs yeux pour expliquer la complexité de la structure des êtres vivants. Des mauvaises langues iront même jusqu'à contester les résultats d'Avry en arguant d'une possible contamination protéique de l'ADN purifié qu'il utilisait dans ses expériences. Aussi pendant plusieurs années, on éluda la question en parlant de "nucléo-protéine". Il faudra alors attendre d'autres expériences de transformations bactériennes et notamment les travaux de A. Boivin (qui montra que si la quantité de protéines était variable selon les cellules, en revanche celle de l'ADN restait constante, et qu'elle était divisée en deux dans les cellules germinales) pour que l'ADN s'imposa comme l'unique vecteur physique de l'hérédité.

L'étude du gène n'appartint désormais plus aux seuls biologistes et se positionna à la frontière entre physique, chimie et biologie. Le gène apparaissait ainsi de plus en plus comme la "matrice" à partir de laquelle la structure complexe des êtres vivants s'établissait. L'influence des physiciens fut alors décisive pour en élucider les mécanismes. Tout d'abord, bien que peu lu par les biologistes dans un premier temps, le livre de Erwin Rudolf Josef Alexander Schrödinger (1887-1961), *What is life ?* (1942), fut sûrement la première tentative en date pour remettre la science génétique "sur ses pieds" : face à la notion "molle" du support de l'hérédité implicitement introduite par Morgan et inhérente à son concept empirique de cartographie (qui va du phénotype au génotype), le physicien se soucia moins des aspects expérimentaux que d'établir une théorie cohérente pour rendre compte du phénotype à partir du génotype : "la fibre chromosomique contient, chiffré dans une sorte de code miniature, tout le devenir d'un organisme, de son développement, de son fonctionnement." (cité par Pichot (1999)). Si Morgan proposait de découvrir statistiquement comment se transmettait, de génération en génération, un caractère particulier, Schrödinger eut l'ambition d'établir les lois de la transmission de l'organisation biologique des êtres vivants. A la question "Comment expliquer la persistance au cours de la vie individuelle et au fil des générations de l' ordre apparent des structures vivantes ?", Schrödinger répondit par la physique : en supposant que le support de l'hérédité était bien une substance, elle devait alors avoir des caractéristiques physiques propres aux structures stables et pérennes. Le paradigme physique

d'un ordre défini et strict, se traduisit alors par un cristal d'atomes, seule structure ordonnée capable de rendre compte à la fois de l'aspect moléculaire de l'hérédité et de sa résistance au cours du temps. Nonobstant la pertinence de la correspondance biunivoque entre la structure microscopique et macroscopique des êtres vivants, Schrödinger demeura vague sur les mécanismes qui permettaient à la première de prendre "corps".

Le physicien autrichien fut toutefois le premier à énoncer le problème sous l'angle informationnelle d'un programme codé dans une structure moléculaire stable et héritable. Bien que le terme d'information n'était pas encore forgé à la parution de son ouvrage - la *Théorie mathématique de le communication* de Claude Elwood Shannon (1916 - 2001) date de 1948 - on attribue néanmoins à ce dernier la notion d'information génétique : l'ordre microscopique du modèle se "traduit" en instructions par lesquelles les gènes commandent l'élaboration et le maintien de la structure de tout être vivant. Le sacre de la théorie ne se fera pas trop attendre puisque 11 années plus tard, la découverte de la structure en double hélice de l'ADN par James Dewey Watson (1928 -) et Francis Harry Compton Crick (1916 - 2004) confirma l'incroyable robustesse et stabilité du support de l'hérédité - ce succès devait notamment beaucoup aux très belles images de diffraction de l'ADN, par cristallographie aux rayons X, issues des travaux de Rosalind Frankline (1920 - 1958). Ces derniers n'hésiteront pas à faire de leur découverte universelle le "dogme centrale de la biologie moléculaire".

Tout alla alors très vite, et la biologie moléculaire s'imposa comme la "reine triomphante" à qui les autres disciplines de la génétique devaient désormais allégeance. La théorie de Schrödinger devint le modèle phare vers lequel les découvertes qui s'accumulaient progressivement devaient converger. Le concept d'information envahit le vocabulaire du biologiste et le "décryptage" de l'ADN et des ses produits s'imposa comme l'unique mystère à résoudre. Dans les années 1960, les principales fonction du génome étaient élucidées : mécanismes de duplication de l'ADN, existence et rôle des ARN messagers et de transfert, le code génétique, mécanisme de synthèse des protéines (Crick et les "codons" (1961)), principes généraux de la régulation de cette synthèse. En immergeant la notion de gène dans la structure physico-chimique de l'ADN (à l'aphorisme "un gène, une enzyme" se substitua la conjonction "un gène, un segment d'ADN"), la biologie moléculaire se fit forte de réaliser ainsi la synthèse de qui n'avait été qu'"un biophore (Weismann), un pangène (De Vries), une unité de calcul (Johannsen), un locus (Morgan), une protéine (presque tout le monde après De Vries), et enfin un ordre physique (Shrödinger). Le gène des années 60 était la synthèse de toutes ces théories successives." (Pichot (1999)).

Les mécanismes de l'hérédité dés lors se déplacèrent : le gène, en correspondance linéaire avec l'ADN et les protéines, perdit sa causalité directe avec le phénotype que l'on supposa désormais déterminé par le complexe protéine-enzyme. La biologie moléculaire, en découvrant la nature informationelle du gène,

modifia du même coup sa définition : de structurale elle devint fonctionnelle. L'hérédité "inflexible" incarnée jusqu'alors par la correspondance biunivoque entre gène et phénotype perdit de sa rigidité, et se retrouva obscurcie par le brouillard opaque des modes d'expressions et des régulations extra-géniques. De plus, la découverte chez les eucaryotes du caractère morcelé des gènes (découpés en introns et exons) et des phénomènes d'épissage alternatifs lors de la transcription, acheva les dernières croyances d'une relation purement bijective entre génotype et phénotype. Au déterminisme génétique, la biologie moléculaire répondit alors par la complexité des phénomènes de régulations entre les gènes et leurs produits : le gène perdit ainsi définitivement sa nature autonome dans la cellule, et François Gros de trancher : "Ce qui caractérise le mieux le gène aujourd'hui, ce n'est pas sa matérialité physique et chimique au niveau de l'ADN (le gène apparaît en effet de moins en moins comme un segment particulier et continu de l'ADN), ce sont bien davantage les produits qui résultent de son activité : ARN cytoplasmique et protéine." (F.Gros, Les secrets du gène, cité par Pichot (1999)). Pour tâcher de combler la perte du lien direct entre patrimoine génétique et phénotype, de nombreuses théories se développèrent. Par exemple, Henri Atlan proposera en 1972 le concept d' "héritabilité épigénétique" : "Ce qui est transmis n'est pas seulement une structure moléculaire statique, mais un état d'activité fonctionelle, c'est-à-dire une certaine expression de la signification fonctionnelle de l'ensemble des structures cellulaires.". Le génome n'apparaissait plus comme un simple programme linéaire mais plutôt comme une banque de données dans laquelle la machinerie cellulaire puiserait pour se maintenir et évoluer. S'imposa alors une vision dynamique de l'hérédité, dans laquelle l'aspect statique de l'hérédité chromosomique semblait s'effacer. Le passage du locus-échantillon au génome-information était désormais consommé. Chassés par Morgan, les vieux démons spéculatifs de la biologie resurgirent à nouveau.

Bien qu'ayant acquis l'apparence d'une science "dure" en disséquant à une échelle microscopique le génome, la biologie moléculaire se heurta elle aussi au caractère fuyant du mystère des êtres vivants, qui ne semblaient décidément pas vouloir se réduire à la seule chimie. Ernst Mayr (1904-2005) fut aussi plus humble et plus modeste lorsqu'il dressa le bilan des avancées de la génétique entre 1965 et 1980 : "

1. La découverte la plus spectaculaire et, jusque dans les années 1940, la plus inattendue, est que le matériel génétique, que l'on sait à présent consister en ADN, ne participe pas lui-même à l'édification du corps d'une nouvel individu, mais sert simplement de plan de construction, de catalogue d'instructions, désigné du nom de "programme génétique".
2. Le code, grâce auquel le programme est traduit chez les organismes individuels, est le même dans tout le monde vivant, des micro-organismes inférieurs aux plantes et animaux supérieurs.
3. Le programme génétique (génome), chez tous les organismes diploïdes

se reproduisant sexuellement, consiste en un jeu d'instructions reçu du père et un jeu de reçu de la mère. Les deux programmes sont en général homologues et agissent de concert.

4. Le programme génétique consiste en molécules d'ADN associées chez les eucaryotes à certaines protéines (histones), dont la fonction précise est encore mal connue, mais qui, apparemment, relève d'une participation à la régulation de l'activité de différents loci dans différentes cellules.
5. La voie menant de l'ADN du génome aux protéines du cytoplasme (transcription et traduction) est à sens unique. Les protéines du corps ne peuvent ainsi induire aucun changement dans l'ADN. L'hérédité des caractères acquis est donc une impossibilité chimique.
6. Le matériel génétique (ADN) est absolument constant ("inflexible") de génération en génération, sauf dans le cas de très rares "mutations" (c'est-à-dire erreurs de duplications).
7. Tout individu d'une espèce se reproduisant sexuellement est génétiquement unique, parce que plusieurs allèles différents peuvent être représentés à des dizaines de milliers de loci distincts dans une population ou une espèce donnée.
8. Cet énorme stock de variation génétique offre continuellement et quasi indéfiniment des matériaux à la sélection naturelle." Mayr (1982) :

Ce dernier point trahit aussi la sympathie de Mayr pour la théorie darwinienne, en excluant au passage les découvertes récentes de Motoo Kimura (1924-1994) - qui publia la somme de ses travaux en 1983 dans un recueil intitulé la *Théorie neutraliste de l'évolution.*

En définitive, si les composantes de l'hérédité sont désormais identifiées, l'étude du déterminisme génétique demeure encore principalement une démarche cognitive visant à circonscrire le plus finement possible la complexité vertigineuse des processus d'expression des gènes. La naissance de l'ingénierie génétique dans les années 70, puis celle de la génomique et ses dérivés en "-iques" dans les années 80 et 90, offrent peut-être désormais la dimension et la puissance opératoire qui permettront d'élucider complètement cette énigme vieille de plusieurs siècles : l'hérédité.

Chapitre 2

Introduction

Formalisée par R.A. Fisher, prolongée par les travaux ultérieurs de K. Mather (1911-1990) et des sélectionneurs comme I.M. Lerner (1910-1977), la génétique quantitative fit de rapides progrès des années 1900 aux années 50. Puis, en l'espace d'à peine 30 ans, l'avènement de la biologie moléculaire et de la génomique a fait faire un bond spectaculaire à cette discipline. Si les concepts forts de la génétique quantitative n'ont pas été remis en cause par cette révolution "opératoire", ses méthodes expérimentales et statistiques ont bénéficié d'un véritable enrichissement. Qui mieux que les marqueurs moléculaires permettent une étude fine de l'hérédité de caractères complexes et de ses composantes ? L'arrivée de ces derniers dans le champ expérimental de la génétique quantitative, en facilitant l'élaboration de modèles prédictifs génotype-phénotype et l'optimisation des processus de création variétale, a contribué à affermir le potentiel de cette discipline dans la conduite des programmes d'amélioration des plantes et des animaux domestiques. Mais au delà de ces objectifs appliqués, de par les allers et retours constants entre expérimentation, analyse et modélisation, ces nouvelles techniques ont aidé à affiner les hypothèses sur la nature des causes sous-jacentes à l'hérédité multifactorielle.

Dans ce cadre scientifique désormais bien établi, étudier le déterminisme génétique de caractères quantitatifs c'est chercher à identifier les facteurs génétiques impliqués dans la variation du caractère. Ces facteurs ont généralement une architecture allélique et épistatique complexe que seules les techniques actuelles rendent analysable. Plus particulièrement, les avancées récentes de la post-génomique permettent d'envisager à présent des études à l'échelle unitaire et physique du gène. Mais, revenons tout d'abord aux fondamentaux. Par définition, un caractère quantitatif résulte de la ségrégation de nombreux gènes dont la plupart ont un effet individuel faible. Quelques gènes seulement présentent des effets "mesurables" : on parle alors de gène majeur. La question première "combien de gènes contrôlent l'expression d'un caractère quantitatif" se meut ainsi en une autre question, formulée sous un angle

plus pragmatique, mais aussi de par le caractère expérimental inhérent à son éventuelle réponse, plus statistique : "combien de gènes - ou de locus - expliquent une proportion mesurable et significative de la variabilité phénotypique ?".

Depuis les travaux de Morgan, la réponse à cette question s'est essentiellement structurée autour d'une vision linéaire et opérationnelle du génome (ce dernier renvoie de prime abord à l'ensemble des locus qui coségrègent dans la population d'étude, plutôt qu'à une vision physique et fonctionnelle telle qu'elle s'est imposée ces dernière années par le biais de la génomique). Dès lors l'étude de l'hérédité de caractères quantitatifs - mais également discrets - s'articule autour de deux étapes : i) établir une carte génétique à l'aide de locus marqueurs, ii) identifier, sur cette carte, les locus significativement corrélés à la variation du caractère. Lorsque ce dernier est quantitatif, ces locus se dénomment QTL (acronyme anglais pour "Quantitative Trait Locus").

La détection de QTL à l'aide de marqueurs moléculaires a été développée initialement chez les plantes dans les années 80 (la première en date étant celle publiée par Paterson *et al.* (1988)). Chez les végétaux, la possibilité de créer des lignées génétiquement homogènes et stables (c'est à dire homozygotes à tous les locus) a facilité la mise en place de populations dites "contrôlées", d'effectifs importants et pour lesquelles chaque individu à une généalogie connue. Le principe général repose sur l'hypothèse que si deux lignées présentent des expressions contrastées pour le caractère étudié, c'est qu'elles présentent des facteurs génétiques également contrastés, et plus précisément des configurations alléliques distinctes aux locus causaux. Des populations hybrides, dénommées F_1, peuvent alors être créées en croisant deux lignées parentales (issues de la diversité génétique disponible au sein de l'espèce). Les individus d'une population F_1 ayant tous le même génotype aux locus (ils sont tous hétérozygotes), des générations supplémentaires sont nécessaires pour pouvoir mesurer la "liaison" entre marqueurs et QTL (voir Encadré 1).

Sur le plan théorique et statistique, le principe de la détection de QTL a suscité un vif intérêt, et de nombreux développements méthodologiques ont vu le jour dans les années 90 - toujours en grande partie chez les végétaux. Il est désormais possible de tester des modèles génétiques élaborés, avec plusieurs QTL sur un même chromosome, et incluant éventuellement des effets épistatiques entre eux Kao *et al.* (1999); Kao and Zeng (2002); Zeng *et al.* (2005). Ces avancées méthodologiques, combinées à la facilité relative de mise en œuvre des dispositifs expérimentaux, ont favorisé l'essor des approches de cartographie de QTL chez les plantes. Pour différentes espèces végétales, les bases de données publiques offrent désormais la possibilité de consulter et de comparer une masse importante de résultats issus de diverses expériences de détection de QTL [1]. Par son principe séduisant et grâce à

[1] Par exemple, chez le maïs, la base de données publique *maizegdb*, accessible via `http://www.maizegdb.org`, regroupe 56 expériences de détection de QTL différentes relatives à 573 caractères distincts. Soit un total de 1504 QTL à ce jour.

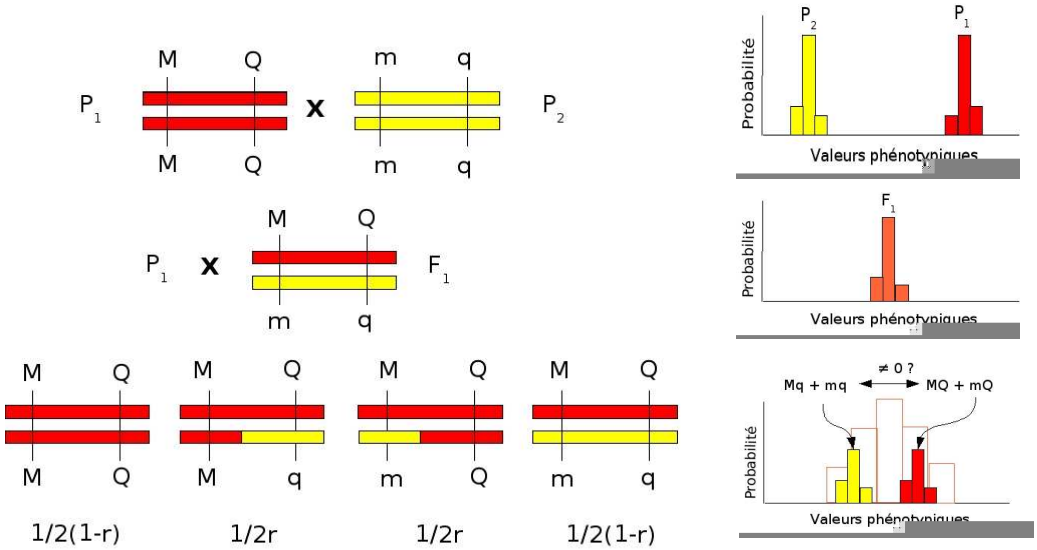


FIG. 2.1 – Représentation schématique d'un plan de croisement de type "backcross" entre deux lignées pures P_1 et P_2. Les individus de la génération F_1 sont rétrocroisés avec le parent P_1 afin de faire ségréger le marqueur et le QTL en 4 classes génotypiques. La probabilité de chaque classe dépend explicitement du "degré" de liaison entre le marqueur et le QTL. Elle est fonction du taux de recombinaison noté r.

Encadré 1 : Principe de la détection de QTL

La méthode la plus simple pour détecter une liaison entre un marqueur moléculaire et un QTL consiste à croiser entre elles des lignées pures (homozygotes à tous les locus) et à analyser les ségrégations entre le marqueur et le QTL dans les générations ultérieures.
L'un des croisements les plus couramment utilisé en génétique végétale consiste à rétrocroiser les individus d'une population F_1 avec l'un des deux parents. Un exemple schématique de "backcross" est illustré dans la Figure 2.1. Supposons qu'un individu F_1 ait un génotype hétérozygote MQ/mq, où M est l'allèle au marqueur lié à l'allèle Q au QTL chez l'un des deux parents. On note alors Y_M, Y_m, Y_Q et Y_q la moyenne phénotypique des individus issus du backcross et qui ont hérité respectivement de l'allèle M, m, Q et q. On note r le taux de recombinaison entre le marqueur et le QTL. On a alors la relation suivante :

$$Y_M = (1-r)Y_Q + rY_q$$
$$Y_m = (1-r)Y_q + rY_Q$$

L'approche usuelle pour cartographier un QTL est de calculer soit un ratio de log-vraisemblance, soit une F-statistique à chaque marqueur le long de la carte génétique. De manière similaire, on peut utiliser une t-statistique (élevée au carré) donnée par

$$T_M = \frac{(Y_M - Y_m)^2}{\mathrm{var}(Y_M - Y_m)}$$

avec $\mathrm{var}(Y_M - Y_m) \approx 4\sigma_e^2/n$, où n est la taille du backcross et σ_e la variance environnementale, incluant potentiellement les effets d'autres QTL. L'idée étant que plus le marqueur sera proche physiquement du QTL (c'est à dire r petit), plus le numérateur de T_M capturera les paramètres génétiques au QTL. En effet, dans le cas d'un backcross, si on note a l'effet additif au QTL, on a la relation suivante :

$$\mathsf{E}(Y_M - Y_m) = a(1-2r)$$

un coût expérimental abordable, cette approche s'est alors progressivement imposée comme le préalable indispensable à l'étude du déterminisme de caractères quantitatifs. Cependant, de par la structure même du dispositif expérimental, les méthodes de détection de QTL ont malheureusement un niveau de résolution trop faible pour pouvoir identifier sans ambiguïté le ou les gènes sous-jacents aux QTL : en moyenne, chez les végétaux, la résolution d'une détection de QTL "classique" est d'environ 10 à 30 cM (Kearsey and Pooni (1996); Chardon *et al.* (2004)). Pour avoir un ordre de grandeur en "nombre de gènes putatifs", 10 cM correspondent à peu près chez le maïs à $18600/30000 \times 10 \approx 160$ gènes (voir le Tableau 2.1).

Cette limitation s'explique à la fois par le faible nombre de générations et d'individus techniquement étudiables. Une première possibilité pour affiner la localisation de QTL est de procéder à des croisements plus efficaces pour accumuler les recombinaisons entre marqueurs et QTL au cours des générations. Ces dispositifs, comme les "Advanced Intercross Lines" développés par Darvasi and Soller (1995); Wang *et al.* (2003), permettent d'"allonger" la carte génétique et de positionner ainsi avec une meilleure précision les QTL. Dans le même esprit, le développement de "Near Isogenic Lines", suivi de la recherche d'individus recombinant dans la région introgréssée, a permis chez certaines espèces, la localisation à l'échelle même du gène de quelques QTL majeurs : on parle alors de clonage positionnel (à ce jour, 3 chez la tomate et le riz et 2 chez arabidopsis et le maïs Paran and Zamir (2003)). Mais, le développement et l'évaluation de telles populations contrôlées demeurent encore assez onéreux. Mais surtout, ces avancées expérimentales se restreignent à une configuration allélique particulière au QTL et seule la localisation de QTL majeurs peut être ainsi affinée.

En définitive, même les dispositifs expérimentaux les plus "précis" ne renseignent que sur un sous-ensemble de QTL et de configurations alléliques à ces QTL. Or, l'architecture allélique sous-jacente au déterminisme de caractères quantitatifs est probablement plus complexe à l'échelle de l'espèce entière. La diversité réduite des populations contrôlées est donc un handicap sérieux à une étude "globale" des composantes du déterminisme. Bien que des schémas de croisement multi-parentaux permettent à présent d'étudier la coségration de plusieurs allèles distincts aux QTL (Rebai *et al.* (1997); Blanc *et al.* (2003)), il est impensable de développer autant d'expérimentations que de nombre de configurations alléliques possibles aux QTL - d'autant que ce nombre est en grande partie inconnu. Notons tout de même que des apports méthodologiques récents permettent d'envisager la cartographie de QTL dans des schémas multi-parentaux complexes et diversifiés (Jannink *et al.* (2001); Jansen *et al.* (2003); Crepieux *et al.* (2004)). Mais, du fait d'un nombre restreint de générations considérées, ces dispositifs souffrent de la même limitation en terme de résolution que ceux évoqués précédemment. Comment, dès lors, parvenir à augmenter simultanément la diversité étudiée et le degré de résolution ?

Historiquement, les premiers travaux sur la relation entre diversité génétique et variation phénotypique, ont été conduits en génétique humaine. On a alors parlé d'"études d'association". Le premier avantage de ces approches, comparées à la cartographie de QTL classique, est de considérer une collection d'individus représentant au mieux la diversité génétique de l'espèce. Ainsi, la majorité, voire l'intégralité des allèles, aux QTL coségrègent potentiellement dans la population étudiée. Deuxièmement, cette dernière résultant d'un échantillonnage réalisé à l'échelle de l'espèce, on peut vraisemblablement espérer obtenir un "bon" niveau de résolution grâce aux nombreux événements de recombinaisons accumulés au cours de son histoire évolutive. Autrement dit, à l'inverse des populations contrôlées, les individus devraient présenter - idéalement - un faible degré d'apparentement.

Mais si, dans le cadre des populations contrôlées, la précision attendue sur la localisation d'un QTL est une fonction explicite des paramètres du croisement considéré, de sa structure, ainsi que des effets génétiques au QTL (voir Darvasi and Soller (1997); Visscher and Goddard (2004)), la complexité et l'opacité des mécanismes évolutifs à l'échelle d'une espèce rendent la détermination du pouvoir résolutif des études d'association plus délicate. Pourtant, ce prérequis est crucial en cela qu'il conditionne non seulement la faisabilité de l'étude mais aussi la stratégie qu'il conviendra d'adopter pour la réaliser. Si la génétique des populations s'est considérablement enrichie sur le plan théorique et méthodologique ces 20 dernières années (notamment avec l'apparition de la théorie de la coalescence), il demeure encore difficile d'inférer, sur la base de données moléculaires seules, la structure des scénarios évolutifs ainsi que leurs paramètres (taille effective, taux de recombinaison, taux de mutation, etc). Aussi, des méthodes plus pragmatiques se sont développées afin d'évaluer non pas l'histoire cryptique qui relie les individus présents - c'est à dire la cause - , mais ses conséquences matérialisées par la structure des corrélations entre locus, et cela à une double échelle : celle de l'espèce et celle du génome.

Le degré de précision d'une étude d'association est ainsi lié en grande partie au degré de liaison statistique entre locus. Le cas idéal étant que la population ait connu suffisamment d'événements de recombinaison pour que des locus physiquement très proches sur le génome ne présentent plus qu'un faible degré de liaison. Sur le plan statistique, au sein d'une population d'étude, ce degré de liaison se mesure par l'appariement non aléatoire des allèles entre paire de locus distincts, dénommé déséquilibre de liaison (DL) (voir Encadré 2). Notons que sous certaines hypothèses, quelques mesures du DL présentent l'avantage de s'exprimer analytiquement en fonction des paramètres évolutifs (voir Encadré 3), offrant ainsi la possibilité de traduire explicitement le lien entre une histoire évolutive hypothétique et le pouvoir résolutif des études d'association.

Connaître la structure du DL au sein de l'espèce étudiée devient alors la condition *sine qua non* à toute étude d'association. En génétique humaine,

Encadré 2 : Mesure du déséquilibre de liaison (DL)

Le DL est défini comme l'association non aléatoire entre allèles à différent locus. Si la notion de DL date du début du siècle (voir par exemple Jennings (1917)), la première mesure couramment utilisée fut introduite par Lewontin (1964) il y a environ 40 ans. Considérons deux locus bi-alléliques A/a et B/b. On note p_x la fréquence de l'allèle $x = A, a, B, b$ dans la population et $\tilde{p}_x$ la fréquence estimée à partir d'un échantillon de N individus. De manière similaire, p_{xy} désigne la fréquence de la co-occurence des allèles x et y aux deux locus dans la population et $\tilde{p}_{xy}$ celle estimée à partir de l'échantillon. Lewontin (1964) proposa alors de mesurer le DL par la statistique :

$$\tilde{D} = \tilde{p}_{AB} - \tilde{p}_A \tilde{p}_B$$

qui n'est autre que la covariance empirique entre les variables indiquant la présence ou l'absence de l'allèle A et l'allèle B aux deux locus. On parle de déséquilibre de liaison entre les locus si $\tilde{D}$ diffère significativement de zéro. Cet écart à indépendance peut être testé par un simple test du χ^2 :

$$\chi^2 = \frac{N\tilde{D}^2}{\tilde{p}_A(1-\tilde{p}_A)\tilde{p}_B(1-\tilde{p}_B)}$$

Toutefois cette première mesure, parce qu'elle dépend des fréquences alléliques, n'est pas très pratique pour effectuer des comparaisons entre plusieurs paires de locus. En remarquant que la valeur maximale du DL est donnée par $D_{max} = min(p_A p_b, p_a p_B)$, et sa valeur minimale $D_{min} = max(-p_A p_B, -p_a p_b)$, Lewontin (1964) suggéra de normaliser $\tilde{D}$ afin d'obtenir une autre mesure, indépendante des fréquences alléliques :

$$D^{'} = \tilde{D}/\tilde{D}_{max}$$

Découlant plus naturellement des statistiques, Hill and Robertson (1968) proposa pour sa part d'utiliser :

$$r^2 = \frac{\tilde{D}}{\sqrt{\tilde{p}_A(1-\tilde{p}_A)\tilde{p}_B(1-\tilde{p}_B)}}$$

qui n'est autre que le coefficient de corrélation empirique (noté parfois Δ dans la littérature) et qui s'obtient également à partir du tableau de contingence des allèles aux deux locus (comme implicitement évoqué par le test du χ^2 proposé précédemment). Notons que d'autres mesures ont été introduites ultérieurement (Nei and Li (1980), Yule (1900)). Enfin, certaines d'entre elles ont été étendues au cas multi-allélique.

Espèce	Taille physique (Gb)	Taille génétique (cM)	ADN/cM (Mb)	Nombre de chromosomes	Nombre de gènes
Humain	3.00	3000	1	23	30000
Drosophile	0.17	340	0.5	4	15000
Arabidobsis	0.15	630	0.14	5	25000
Riz	0.43	1570	0.28	12	50000
Maïs	2.50	1860	1.40	10	$\sim$ 30000
Blé	16.00	3500	4.60	21	-

TAB. 2.1 – Taille des génomes de 6 espèces et relation entre longueur physique et génétique.

Encadré 3 : DL et mécanismes évolutifs

Parmi les mécanismes évolutifs qui sont susceptibles de générer du DL, les plus couramment reportés dans la littérature (voir par exemple JANNINK and WALSH (2002); FLINT-GARCIA *et al.* (2003); GAUT and LONG (2003)) sont :
- La dérive génétique : DL introduit par l'effet d'échantillonnage à chaque génération.
- Le système de reproduction : chez les espèces autogames (comme *arabidopsis*) la réduction du taux de recombinaison effectif est supérieur à 10 pour une autogamie de 95%, d'environ 50 pour 99% (NORDBORG (2000)). Moins de recombinaisons induit un maintien du DL au fil des générations.
- La migration.
- La sélection (alliée également à des interactions épistatiques : les haplotypes combinant les allèles favorables aux différent locus causaux sont préférentiellement sélectionnés).
- la mutation.

Pour des hypothèses évolutives particulières, les moments théoriques de certaines mesures du DL peuvent s'exprimer simplement en fonction des paramètres évolutifs. Sous un modèle de Wrigth-Fisher, sans mutation ni sélection, l'espérance, au cours des générations, du DL (mesuré par D) entre deux locus dans une population isolée s'écrit :

$$\mathsf{E}[D(t)] = D(0)(1-r)^t \left(1 - \prod_{i=1}^{t} \frac{1}{2N(i)}\right)$$

où t est la génération courante, $D(0)$ le déséquilibre de liaison à la génération initiale, r le taux de recombinaison entre les deux locus, et $N(i)$ la taille de la population à la génération $i = 1, \ldots, t$. Dans le même cadre théorique, en supposant atteint l'équilibre entre la recombinaison et la dérive, et pour une grande taille de population N, HILL and WEIR (1994) montra que :

$$\mathsf{E}[r^2] = \frac{1}{1+4Nr}$$

Enfin, toujours pour une population de grande taille mais en introduisant un modèle de mutation infinitésimal aux locus, HILL (1975) proposa l'approximation suivante :

$$\mathsf{E}[r^2] = \frac{10+\rho+4\theta}{22+13\rho+\rho^2+32\theta+6\theta\rho+8\theta^2}$$

avec $\rho = 4Nr$, et $\theta = 4N\mu$ où μ est le taux de mutation par génération et par locus.

la littérature est désormais riche d'articles discutant des patrons de DL observés dans différentes régions du génome. En particulier, la synthèse de Pritchard and Przeworski (2001) a permis de préciser les structures de DL attendues selon les hypothèses les plus compatibles avec les schémas évolutifs pressentis chez l'homme. En génétique végétale, les études sur le DL sont plus tardives, dû en parti à un intérêt récent pour les études d'association. Chez le maïs, Tenaillon *et al.* (2001) a observé une décroissance rapide du DL en fonction de la distance physique (DL non significatif au delà d'environ 200b) laissant augurer d'un avenir prometteur pour les études d'association. Avec un jeu de données plus conséquent, Remington *et al.* (2001) a confirmé cette décroissance rapide au sein de gènes suspectés d'être impliqués dans l'adaptation du maïs au climat tempéré, bien que selon Remington *et al.* (2001) elle ne devienne significative qu'à une distance d'environ 1500b. Le Tableau 2.1 résume, pour six espèces différentes, la décroissance moyenne du DL reportée dans la littérature. On remarque notamment, comme attendu, l'influence du système de reproduction sur cette décroissance. Mais revenons au cas du maïs : d'après le Tableau 2.1, 1500b correspondent environ à 0.001 cM, et comparé aux 10 à 30 cM des intervalles de confiance obtenus pour les QTL détectés dans des populations contrôlées, l'alternative offerte par les études d'association paraît presque miraculeuse !

Ainsi, si nous suivons le classement de la Figure 2.2, les études d'association réalisées sur des collections aux bases génétiques larges occupent (idéalement) une place de choix dans la cartographie fine de QTL. Comment expliquer alors une émergence aussi tardive de ces approches en génétique végétale ? On pourrait être tenté d'imputer aux seules avancées techniques récentes (essentiellement les méthodes de séquençage haut-débit), sans lesquelles ces études réalisées au niveau intragénique n'auraient pu être possibles, la responsabilité de ce retard. Mais, ces techniques naviguent si facilement entre espèces que le décalage entre génétique humaine et génétique végétale ne peut s'expliquer par ce seul facteur. La raison, bien-sûr, est plus profonde.

D'une part, en génétique humaine, les résultats issus des études d'association sont parfois soumis à controverse : ce que détecte une étude, une autre indépendante ne le révèle pas. Ce problème de reproductibilité des résultats fut en premier lieu imputé aux processus d'échantillonnage ainsi qu'aux choix des seuils d'erreur de première espèce (Lander and Kruglyak (1995)). Mais, le même problème se pose en cartographie classique de QTL (Keightley and Knott (1999)), et les difficultés rencontrées pour reproduire des expériences reposent davantage sur des hétérogénéités entre dispositifs expérimentaux que sur la seule dimension statistique Sillanpaa and Auranen (2004). Enfin des synthèses par méta-analyse ont récemment confirmé la cohérence de plusieurs résultats publiés chez l'homme Lohmueller *et al.* (2003), témoignant ainsi en faveur de cette démarche.

Au-delà de la polémique suscitée par la question - et l'enjeux - de la

Espèce	Système de reproduction	Étendue du DL
Humain	Allogame	
Africain		5kb
Européen		80kb
Drosophile	Allogame	$<$ 1kb
Arabidopsis	Autogame	250 kb
Riz	Autogame	100kb
Maïs	Allogame	
Populations		1kb
Lignées		1.5kb
Lignées élites		>100kb
Blé (dur)	Autogame	10-20 cM

TAB. 2.2 – Décroissance du DL en fonction de la distance physique pour certaines espèces (adapté de FLINT-GARCIA *et al.* (2003))

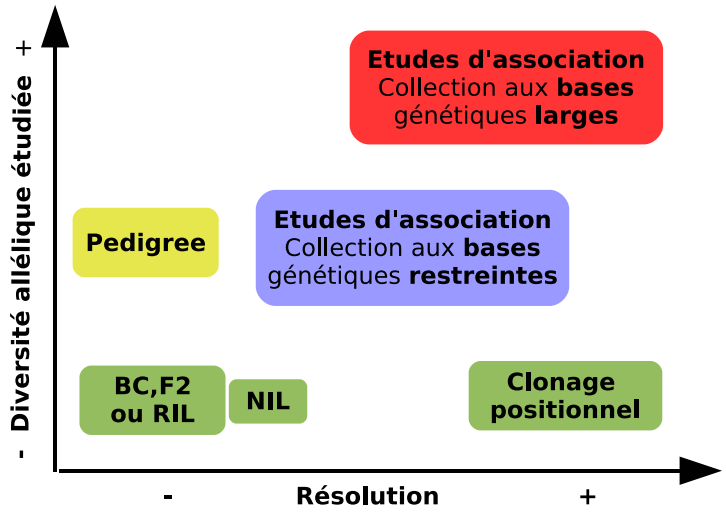


FIG. 2.2 – Diagramme idéal de la répartition des différentes méthodes de cartographie de QTL en fonction de la diversité allélique considérée et du pouvoir résolutif (adapté de FLINT-GARCIA *et al.* (2003)).

reproductibilité des résultats, le retard de la génétique végétale dans ce domaine s'explique, à notre avis, par la structuration génétique souvent "cryptique" des collections. La génétique humaine n'est pourtant pas exempte des effets confondant induits par la structuration génétique des populations d'études : une mise en garde sérieuse avait déjà été énoncée en 1988 par Knowler *et al.* (1988) suite à une étude sur le diabète de type 2. Les auteurs montrèrent que l'apparente association entre un variant génétique particulier et la susceptibilité au diabète de type 2 était due en réalité, aux fréquences contrastées de ce variant entre les individus d'origine nord américaine et ceux d'origine européenne (pour chaque origine prise séparément aucune association significative n'avait été alors trouvée). Or, du fait de la domestication et des pressions de sélection exercées par l'homme sur les espèces cultivées, ces dernières (comme le blé, le maïs ou le riz) ont une histoire évolutive impliquant des schémas de croisement complexes et des flux de gènes restreints, conduisant à la création d'une structure génétique compliquée. Ces effets de structuration ne se limitent pas qu'aux plantes domestiquées par l'homme, et l'étude de Sharbel *et al.* (2000) sur arabidobsis a montré également la nature complexe des origines génétiques interférant dans les populations actuelles.

Ces phénomènes d'"admixture" (mot forgé a partir de l'anglais "mixture" pour désigner des structures de mélange complexes) ne sont pas le seul fait des plantes. Tout événement de migrations entre des populations historiquement isolées est susceptible d'avoir créé de l'admixture, et il semble désormais admis que la population humaine ait connu de tels processus au cours de son histoire (par exemple, diverses origines européennes et africaines ont contribué, conjointement aux origines vernaculaires, à la diversité actuelle de la population sud-américaine Chikhi *et al.* (2001)). Avec l'essor des marqueurs s'est rapidement développée l'idée que l'information moléculaire, recueillie chez des populations humaines contemporaines, devait contenir la trace de ces événements démographiques passés. Il a fallu toutefois attendre jusqu'à ces 10 dernières années, avant que ne se développent, en génétique humaine, des méthodes exploitant au mieux cette information. En particulier, en proposant un compromis raisonnable entre des hypothèses évolutives simples et une modélisation individu "centrée" des conséquences génétiques des phénomènes d'admixture, la méthode de Pritchard *et al.* (2000a) a connu depuis un vif succès.

Succès renforcé par un article complémentaire publié la même année (Pritchard *et al.* (2000b)). Les auteurs y montrent comment les résultats obtenus au préalable à l'aide du logiciel structure, issu de leur premier article, peuvent être intégrés dans les tests d'association effectués sur des échantillons structurés. L'impact sur la génétique végétale fut quasi-immédiat. A une année d'intervalle, Thornsberry *et al.* (2001) publia la première étude d'association chez le maïs : suivant les conseils de Pritchard *et al.* (2000b), les auteurs prirent soin d'étudier la structuration génétique de

leur collection de lignées pour intégrer les "proportions d'admixture" de ces dernières dans les tests. La région génique étudiée par Thornsberry *et al.* (2001) se nomme *dwarf8*, et la génétique végétale vient de faire un grand pas vers la mise en œuvre d'études d'association à grande échelle. Notons que bien que leurs conclusions soient plus prudentes, Andersen *et al.* (2005) puis Camus-Kulandaivelu *et al.* (2006) ont depuis confirmé cette association à l'aide de jeux de données plus conséquents et plus diversifiés.

Le problème de la structuration génétique désormais en parti résolu, la question se pose à nouveau de savoir à quelle échelle peut-on espérer raisonnablement travailler chez les plantes. Reprise sous un autre angle, la question du DL, si elle détermine de prime abord le niveau de résolution à l'échelle physique du génome, conditionne également la stratégie a adopter pour les études d'association : autrement dit, entre étude "génome complet" ou étude locale, que choisir ? La première approche a cela de séduisante qu' à l'instar de la cartographie classique de QTL on analyse simultanément tous les marqueurs distribués le long du génome. Mais elle soulève une question épineuse : quelle densité de marqueurs est requise pour obtenir une couverture optimale de l'ensemble des chromosomes ? Fondés sur les études empiriques des patrons du DL, quelques chiffres ont été avancés : environ 70000 marqueurs seraient requis chez l'homme, 2000 pour arabidopsis, 750000 pour les populations traditionnelles contre 50000 pour les lignées de maïs (estimation proposée par Flint-Garcia *et al.* (2003)). Ces derniers chiffrent donnent le vertige tant sur le plan technique [2] que sur la taille de l'échantillon nécessaire pour garantir les niveaux de seuil des tests d'association.

En outre, l'aspect novateur des études d'association chez les végétaux invite à la prudence et l'on voit difficilement comment un programme "génome complet" pourrait être lancé sans des études préliminaires à une échelle plus modeste. Une première alternative a été proposée en génétique animale par Meuwissen and Goddard (2000) : à partir de résultats préliminaires de détection de QTL sur de nombreuses familles , les auteurs suggèrent de densifier la région en marqueurs (avec un pas entre 0.25 et 1 cM) afin de localiser finement le QTL sur la base d'un pronostic sur la généalogie reliant les familles entre elles (et donc sur la base du DL entre marqueurs). Mais, d'une part l'on dispose rarement de tels dispositifs expérimentaux chez les végétaux, et d'autre part, il est vraisemblable que même pour des schémas multi-parentaux existants, la diversité allélique au QTL soit sous-représentée.

En génétique humaine, bien avant que le débat "genome-wide" soit ouvert Carlson *et al.* (2004); Hirschhorn and Daly (2005); Marchini *et al.*

[2] Remarquons néanmoins que le coût du génotypage est de moins en moins élevé. Chez l'humain, les dernières évaluations du projet HapMap Consortium. (2003) tablent sur un coût de 0.01U.S$. Un individu pourrait ainsi être caractérisé à 50000 marqueurs pour 500U.S$. Selon Hirschhorn and Daly (2005), une fois que le coût par marqueur sera tombé à 0.001US$, une stratégie génome complet sera envisageable chez l'homme.

(2005); de Bakker *et al.* (2005), l'approche ciblée "gène candidat" a été et demeure largement exploitée. L'idée repose sur une sélection *a priori* des régions d'étude, incluant un voire plusieurs gènes ; *a priori* fondé sur la présomption, préalablement acquise en confrontant si possible plusieurs sources d'information, que cette région est impliquée dans le déterminisme du caractère d'intérêt. Or, chez les végétaux, la masse des résultats acquis en cartographie de QTL classique fournit une base solide pour orienter la définition des zones d'études prioritaires le long du génome. De plus, la possibilité de croiser l'information entre différentes espèces grâce aux outils de la génomique comparative, permet, pour certains caractères modèles au déterminisme transversal, de mieux circonscrire l'ensemble des gènes candidats (voir par exemple Chardon *et al.* (2004)). C'est l'une des raisons pour lesquelles cette approche a suscité un vif intérêt en génétique végétale.

Cependant, le lien entre gènes candidats et données de cartographie de QTL demeure encore laborieux, dû principalement aux larges intervalles de confiance (nous rappelons que chez le maïs un intervalle - petit - de 10cM correspond environ à une centaine de gènes). Face aux perspectives prometteuses des études d'association, il serait utile de définir des stratégies moins empiriques, permettant notamment de tirer avantage du nombre important de résultats de détection de QTL obtenus, au sein d'une espèce, pour des caractères ontologiquement proches. Mieux cerner l'ensemble des gènes candidats, en excluant en amont la plus grande part de l'information non pertinente, c'est faciliter et accélérer par la suite les études d'association.

En conclusion, l'analyse de gènes candidats via les études d'association ouvre désormais une nouvelle voie en génétique végétale pour étudier plus finement le déterminisme de caractères quantitatifs. Au début des années 2000, en s'appuyant sur les ressources génétiques maintenues à l'échelle internationale, des premières collections diversifiées ont vu le jour afin de répondre aux enjeux de la génétique d'association chez les espèces végétales modèles (par exemple, la constitution du panel de maïs à l'UMR du Moulon, qui inclut à la fois des origines américaines et européennes, date de 2002). C'est dans ce contexte émergeant, à l'interface entre génomique et génétique quantitative, que notre travail s'est positionné. Deux problématiques se sont alors immédiatement imposées : i) comment identifier un premier jeu pertinent de gènes candidats ? ii) comment conduire par la suite les études d'association dans ce type de collection ?

Le premier objectif de cette thèse a donc visé à élaborer une nouvelle méthode de méta-analyse de QTL, afin de circonscrire le mieux possible, chez une espèce donnée, les régions chromosomiques impliquées dans la variation d'un caractère d'intérêt - pour lequel suffisamment d'analyses indépendantes ont été reportées dans la littérature. La première partie de la thèse présente l'article correspondant.

La conduite d'études d'association "gène candidat centrées" dans des collections structurées chez les végétaux a constitué le deuxième objectif

de la thèse. Comme évoqué précédemment, l'analyse de la structuration génétique est un préalable indispensable pour limiter par la suite ses effets confondants dans les tests d'association. Bien que le logiciel STRUCTURE, récemment enrichi par l'apport de FALUSH *et al.* (2003), propose une modélisation bayésienne convaincante, nous avons souhaité développer une alternative plus simple, moins onéreuse en temps de calcul, et également plus explicite sur la stratégie de choix de modèle. Cette méthode fait l'objet d'un projet d'article qui constitue la deuxième partie de la thèse.

Le concept de DL étant au coeur des études d'association, inspirés par des observations empiriques faites sur des données de séquençage de gènes de maïs (mais aussi reportées chez l'humain DALY *et al.* (2001)), nous introduisons dans une troisième partie une nouvelle méthode pour modéliser le DL sous l'hypothèse que la population d'étude ait connu, dans un passé "pas trop lointain", un événement de fondation. Cette méthode est décrite dans un projet d'article. Nous le complétons par une brève revue des méthodes pour effectuer des études d'association ainsi que d'une réflexion sur la possibilité d'intégrer cette modélisation du DL dans une nouvelle démarche de cartographie fine.

Dans une dernière partie nous discutons des perspectives possibles à l'ensemble de ce travail. Enfin, notre souci constant de rendre disponible les outils que nous avons été amené à développer au cours de la thèse, nous a conduit à développer deux bibliothèques d'analyse pour chacun des objectifs. Nous proposons donc en annexe deux projets de note. La première décrit le module de méta-analyse de QTL. La deuxième présente la bibliothèque sur laquelle repose les outils dédiés aux études d'association.

Deuxième partie

Méta-analyse de QTL

Chapitre 3

Meta-Analysis of QTL Mapping Experiments

Authors : Jean-Baptiste Veyrieras [a,1], Bruno Goffinet[2] and Alain Charcosset[1]

[1] UMR, INRA UPS-XI INAPG CNRS Génétique Végétale, Ferme du Moulon, 91190 Gif-sur-Yvette, France
[2] MIA, INRA, Chemin de Borde Rouge BP52627, 31326 Castanet Tolosan Cedex, France

Keywords : genetic-map, QTL, meta-analysis.

Submitted to *Genetics*

[a]to whom correspondence should be addressed

Abstract

In this article we describe a new method to perform integration of QTL mapping experiments in order to establish a consensus model for both the marker and the QTL positions on the genome. First we present an original statistical approach to merge simultaneously distinct genetic maps into a single consensus map which is optimal in terms of weighted least squares and which can be used to investigate recombination rate heterogeneity between studies. Secondly, assuming that QTL can be projected on the consensus map, we propose a new clustering approach based on a Gaussian mixture model to decide how many QTL underly the distribution of the observed QTL. We demonstrate by means of simulations that usual model choice criteria from mixture model literature perform relatively well in this context. Simulations also show that this meta-analysis procedure leads to a reduction

of the length of the confidence interval of QTL location provided that the number of observed QTL is not too small with regard to the number of "true" QTL locations. Finally we illustrate our new approach on previously published QTL detection results relative to flowering time in maize.

Introduction

The advent of linkage mapping experiments using molecular markers in the 1980s has tremendously increased the potential of quantitative genetics by identifying regions of the genome the polymorphism of which affects the variation of quantitative traits, called Quantitative Trait Loci (QTL). Since the last two decades a large set of methods and algorithms have been developed in order to facilitate and to improve the localization of QTL for various kinds of population pedigree. Although a large number of powerful methods have been developed to tackle the problem of QTL detection, the limited number of recombination events available in routinely used pedigree designs for QTL mapping Kearsey and Pooni (1996), mainly due to both a few mating generations and a restricted number of sampled individuals (generally a few hundreds), lead essentially to an approximate mapping of the QTL. From results of QTL experiments gathered over a wide range of plant species, Kearsey and Farquhar (1998) have shown that confidence intervals around most likely QTL positions are, on average, approximately 10 cM, which usually includes several hundred of genes. More recent advent in the area of molecular biology have allowed researchers to carry out positional cloning of QTL (e.g. in maize Salvi *et al.* (2002) have investigated the region of *vgt1*) but this approach still remains extremely expensive both in terms of time and resources.

Also several authors (see for instance Kearsey and Farquhar (1998), Xu (2003)) have pointed out that QTL detection is statistically biased both relatively to the true number of QTL, which is underestimated since only the few QTL with large effects are detected, and relatively to the QTL effects which are biased towards larger values as only significant effects are reported (a phenomenon has commonly referred to as the Beavis effect Beavis (1994)). Even though they must be considered with the awareness of these limitations, QTL mapping experiments have become commonplace and have greatly contributed to improve the knowledge about the genetic determinism of complex traits.

Therefore since the first publication of a QTL localization using molecular data Paterson *et al.* (1988) more and more species and traits have been studied and a large part of these results has been made available via public databases. One of the main purposes of these databases was to help researchers to compare results from different QTL studies, to study the congruency of QTL locations. In other words, it aims to address the following question : "do

QTL identified for a given trait in a population correspond to those detected in other populations ?". In theory one would expect that the variation of a quantitative trait within a species is explained by a finite number of genes. Thus QTL congruency investigation might be a relevant approach to improve knowledge on trait genetics and several publications have pointed out its usefulness (Chardon *et al.* (2004); Khatkar *et al.* (2004); Lin *et al.* (1995); Keightley and Knott (1999); Mihaljevic *et al.* (2004), Paterson *et al.* (1995)). Nevertheless the combination of results from linkage studies can be tedious since, even if several studies focus on the same trait within the same species, family structures, sample sizes, marker maps, or QTL detection methods may differ between studies.

These impediments were partially removed by recent developments. First, integration of genetic maps and QTL locations by iterative projections on a reference map is now widely used to position both markers and QTL on a single consensus map (see for instance Arcade *et al.* (2004)). However this process yields a consensus marker map which both statistical properties and biological "reality" can't be clearly assessed, even if a robust ordered marker map was used as reference. Yap *et al.* (2003) proposed an original approach using graph theory to integrate various types of maps (genetic, physical or sequence-based) but it mainly dealt with dissection* of marker order inconsistencies between maps. From up to now it seems that there is not any efficient methodological framework to build reliable consensus marker map on which markers and candidate genes from different mapping experiments can be both ordered and positioned (except by merging raw mapping data from multiple populations as proposed by Stam (1993) and Schiex (1997)).

Second, in order to study QTL congruency Goffinet and Gerber (2000) proposed an original approach based on a meta-analysis strategy. Meta-analysis, which is mainly used in medical, social, and behavioural sciences, aims to take benefit from pooling results across independent studies in order to combine them in a single result or estimate. The relevance of meta-analysis investigations in genetics and evolution has been discussed and pointed out by several authors in the last decade (see for instance Allison and Heo (1998), Lohmueller *et al.* (2003); Van Zandt and Mopper (1998); Vollestad *et al.* (1999)). More recently Etzel and Guerra (2003) developed another meta-analysis based approach to overcome the between-study heterogeneity and to refine both QTL location and the magnitude of the genetic effects. Yet both the method of Goffinet and Gerber (2000) and Etzel and Guerra (2003) are limited to a small number of underlying QTL positions (from one to four for the former and only one for the later) which is a hard limitation to genome-scale study of QTL congruency. Even if the average number of QTL per experiment is around four in plants (Chardon *et al.* (2004); Kearsey and Farquhar (1998)), one would expect that more than four genes can be implied in the trait

variation on a single chromosome.

To remove these impediments we have developed a new 2-stage meta-analysis procedure which makes it possible to integrate multiple independent QTL mapping experiments. Our aim was to elaborate a global framework to evaluate the homogeneity of both genetic marker and QTL mapping results from literature and public data bases. The first part of our meta-analysis procedure consists in building a consensus genetic marker map that takes into account the statistical properties of genetic distance estimates using a Weighted Least Squares (WLS) strategy, in order to test the consistency of both the order and the marker interval distances in different mapping experiments. The validity of this approach was evaluated by means of simulations for different kinds of usual pedigree commonly used in genetic mapping experiments in plant. Secondly, once consensus marker map has been built the QTL locations can be projected on it. The QTL meta-analysis can then be carried out using a new clustering algorithm based on a Gaussian mixture model leading to the identification of a limited number of underlying QTL which best explain the observed distribution of QTL positions in the mapping experiments. As it has been emphasised by Goffinet and Gerber (2000), the crucial point at this step is to find an unbiased criterion to select the correct number of QTL. To do so we have studied by means of simulations the properties of usual model choice criteria in the context of Gaussian mixture. Finally, as an illustration, we applied our new approach to QTL detection results gathered for flowering time in maize.

Meta-analysis of genetic maps

Genetic map information

Consider a set of n genetic mapping experiments concerning the same linkage group. These different experiments may involve different kinds of population pedigree. We consider that for each experiment $i = 1, \ldots, n$ only the estimated distances between ordered marker along the linkage group are available. We denote c_i, N_i, M_i the population cross design, the population size and the number or markers on the i^{th} genetic map, respectively. Let's suppose that two markers m_j and m_k have been positioned on the i^{th} map, $\hat{r}_{i,jk}$ stands for the estimated recombination rate between markers m_j and m_k and $\hat{d}_{i,jk} = f(\hat{r}_{i,jk})$ the corresponding estimated distance, where f is the mapping function which is assumed to be the same in the n mapping experiments. Applying the classical asymptotic Gaussian distribution of the maximum-likelihood estimation of the parameter we assume that the $\hat{r}_{i,jk}$ are normally distributed around the true recombination rate $r_{i,jk}$ between markers m_j and m_k with a variance $\text{var}(\hat{r}_{i,jk}) = \gamma^2_{i,jk}$. This variance $\gamma^2_{i,jk}$ depends on the pedigree c_i, the value of $r_{i,jk}$, the sample size N_i and the amount of information supplied by the marker pair m_j and m_k in the

sampled population (see Appendix A).

Since mapping function are generally bijective functions, functional invariance property of the maximum-likelihood estimate can be applied. So it comes that $\hat{d}_{i,jk}$ is also normally distributed around the true distance denoted $d_{i,jk} = f(r_{i,jk})$. To obtain the variance of $\hat{d}_{i,jk}$ we use the first term of the Taylor expansion of the inverse of the mapping function leading to the approximation :

$$\text{var}(\hat{d}_{i,jk}) \approx \text{var}(\hat{r}_{i,jk}) \times \left[\frac{\partial f(\hat{r}_{i,jk})}{\partial r}\right]^2$$

Now suppose the n experiments are consistent with the following hypotheses :

- Hypothesis 1 : they come from independent population samples. This implies that $\text{cov}(\hat{r}_{i,jk}, \hat{r}_{i',jk}) = 0$ and $\text{cov}(\hat{d}_{i,jk}, \hat{d}_{i',jk}) = 0$ for any pair of markers m_j and m_k which have been mapped in population i and i', $i \neq i'$ and $(i, i') \in [1..n]^2$.
- Hypothesis 2 : there is no interference, i.e in each mapping experiment the recombination events occur independently in each marker interval.This is equivalent to say that for a given mapping experiment i both the ordered marker interval recombination rate and distance estimates are independent, i.e $\text{cov}(\hat{r}_{i,j(j+1)}, \hat{r}_{i,(j+1)(j+2)}) = 0$ and $\text{cov}(\hat{d}_{i,j(j+1)}, \hat{d}_{i,(j+1)(j+2)}) = 0$ for $i = 1, \ldots, n$ and $j = 1, \ldots, M_i - 2$.
- Hypothesis 3 : the "true" marker order and recombination rate are supposed to be the same in the different populations, i.e $r_{i,jk} = r_{i',jk}$ if markers m_j and m_k have been mapped in population i and i', $i \neq i'$ and $(i, i') \in [1..n]^2$.
- Hypothesis 4 : we assume that all the genetic maps are connected. Mathematically, this means that if we consider maps as vertices and common markers as edges, then the corresponding graph is supposed to be connected.

The meta-analysis model

Let's define $\hat{\mathbf{D}} = (\hat{d}_{i,jk})$ and $\mathbf{\Sigma} = \text{diag}(\sigma^2_{i,jk})$ the vector of ordered marker interval distance estimates and the diagonal variance covariance matrix of $\hat{\mathbf{D}}$. We assume that a total of M distinct markers have been mapped in the n populations. The aim of the meta-analysis is to cross all the available information on marker order and positions in order to build a consensus linkage group on which the M markers are positioned. To do so we introduce $\mathbf{X} = (x_1, \ldots, x_M)$ the vector of the "true" positions of these M markers on the consensus linkage group, where the x_i's can be either positive or negative depending on an arbitrary zero-reference on the chromosome (hereafter we suppose $x_1 = 0$). If the n mapping experiments are consistent with the previous hypotheses and assuming that the distances on the linkage group

are additive we propose to estimate X by solving the following linear system :

$$\hat{d}_{i,jk} = x_k - x_j + \epsilon_{i,jk}$$

where $(j,k) \in [1,\ldots,M]^2$, $i \in [1,\ldots,n]$, $\hat{d}_{i,jk}$ is the distance estimate of the interval between marker m_j and m_k consecutive on the i^{th} linkage group, $x_k - x_j$ is the true distance between these markers, $\epsilon_{i,jk} \sim \mathcal{N}(0, \sigma_{i,jk})$ is the expected standard deviation of the distance estimate $\hat{d}_{i,jk}$. If hypothesis 4 holds we are ensured that this system has at least one solution. Applying a classical weighted least squares (WLS) strategy, the optimal solution is the one which minimizes the target function,

$$\chi = \sum_{i=1}^{n} \sum_{jk} \frac{[\hat{d}_{i,jk} - (x_k - x_j)]^2}{\sigma^2_{i,jk}}$$

Let's introduce $\mathbf{A}$ a design matrix such that $\chi = {}^t(\hat{\mathbf{D}} - \mathbf{AX})\mathbf{\Sigma}^{-1}(\hat{\mathbf{D}} - \mathbf{AX})$ Then the value of $\mathbf{X}$ which minimizes χ is given by :

$$\hat{\mathbf{X}} = ({}^t\mathbf{A}\mathbf{\Sigma}^{-1}\mathbf{A})^{-1}\ {}^t\mathbf{A}\mathbf{\Sigma}^{-1}\hat{\mathbf{D}}$$

which is also a maximum-likelihood estimation of $\mathbf{X}$ with variance-covariance matrix given by $({}^t\mathbf{A}\mathbf{\Sigma}^{-1}\mathbf{A})^{-1}$. Finally $\hat{\mathbf{X}}$ gives both the marker positions and the marker order along the consensus linkage group. The goodness-of-fit of the model can be evaluated by the means of a chi-square test as $\chi \sim \chi^2_{q-M+1}$ where q is the length of the vector $\hat{D}$, i.e the number of marker intervals over the n linkage groups.

Interpretation of the WLS Model

Consider the following idealized scenario : suppose that the n gathered genetic maps share the same markers, i.e $M_i = M$ for $i = 1,\ldots,n$. In this simple case the computation of $\hat{\mathbf{X}}$ is straightforward :

$$\begin{cases} \hat{x}_1 & = 0 \\ \hat{x}_{j+1} - \hat{x}_j & = \dfrac{\sum\limits_{i=1}^{n} \sigma^{-2}_{i,j(j+1)} \hat{d}_{i,j(j+1)}}{\sum\limits_{i=1}^{n} \sigma^{-2}_{i,j(j+1)}} \quad j = 1,\ldots,M-1 \end{cases}$$

and χ is the sum of $M-1$ terms each distributed as a chi-square with $n-1$ degree of freedoms. This is equivalent to go along the marker intervals and for each one to test if the distances are homogeneous between populations using a classical test of equal means. This WLS approach not only provides a simple framework to create a consensus marker map but also makes it possible to test for the homogeneity of distance between several mapping experiments. This can be viewed as an alternative to the M-test devised by Morton (1956) when raw data are not available.

Simulation study

Scenario 1: As Σ is based on a simple Taylor expansion at the first order, the meta-analysis result could suffer from a lack of precision of the variance estimates. Moreover the variance estimates are generally computed using the maximum-likelihood estimate of the recombination rate, which is also an approximation. A simple scenario was explored in order to investigate the impact of these approximations on the consensus linkage group construction. We considered a single chromosome on which 21 markers were spread by randomly drawing 20 marker interval distances in a Gaussian distribution with mean equals to 10cM and standard deviation of 2cM. This Gaussian distribution of marker interval distances allows to create a variety of marker configurations with intervals of reasonable length. For a given pedigree n population data sets were simulated for each marker configuration (all the markers were assumed to be fully informative). For all the mapping experiments we fixed the number of individuals to 200. For each data set the recombination fractions were estimated using a usual maximum likelihood procedure. The order of markers was assumed to be known. Finally the consensus linkage group was build by our WLS strategy. 50 marker configurations were drawn for a given pedigree and 100 replicates were done for each marker configuration.

For each individual mapping experiment we define the Interval Mean Square Error (IMSE) as $\text{IMSE}(i) = \frac{1}{J}\sum_{j=1}^{J} E(\hat{r}_{i,j} - r_j)^2$ where J is the number of intervals (here $J = 20$), $\hat{r}_{i,j}$ the estimation of the recombination rate in the j^{th} marker interval in the i^{th} mapping experiment and r_j is the true recombination rate. This indicator measures the quality of an estimated marker map relatively to the "true" marker map. It comes immediately that the expected IMSE(i) is $\frac{1}{J}\sum_{j=1}^{J}\gamma_{i,j}^2$ where $\gamma_{i,j}^2$ is the variance of the recombination rate estimate in the j^{th} marker interval in the i^{th} mapping experiment. Using the distance estimates obtained by the WLS approach, it is also possible to compute this indicator for the consensus linkage group, denoted IMSE(c). Therefore in order to evaluate the quality of the consensus linkage group we computed the quantity $\overline{\text{IMSE}} = \frac{1}{n}\sum_{i=1}^{n} \text{IMSE}(i)/\text{IMSE}(c)$. If the n mapping experiments share the same family structure and if the approximation made on the variance estimates are not too crude, $\overline{\text{IMSE}}$ should be equal to n.

Scenario 2: Secondly, mapping experiments generally do not have all their markers in common. In this case, the proportion, p, of common markers between the mapping experiments can be a limiting factor to carry out the meta-analysis. In order to study the impact of p on the quality of the consensus linkage group, we investigated a 200cM long chromosome covered by 2001 markers equally spaced by 0.1 cM. Each mapping experiment consists

in a scattering view of the original chromosome with a limited number of markers randomly picked but subject to a constraint on marker interval distances in order to avoid both too tiny or too large intervals. This constraint was set so that all the mapping experiments have marker intervals with distances laying between 5 and 30 cM. Finally for a given number n of mapping experiments p was defined as the ratio between the average number of common markers over all the pairs of individual maps and the total number of markers M. The number of markers per mapping experiments was set to 20. In our simulation study we focused on 4 common marker proportion configurations : $p = 0.15, 0.25, 0.50$ and 0.75 which correspond to respectively $3, 5, 10$ and 15 common markers between pairs of mapping experiments. For a given type of population, a given number of populations n and a given value of p, 25 marker configurations were generated. Each replicate consisted in 100 simulated data sets and for each data set the linkage group of the n mapping experiments were constructed as for the first simulation procedure.

For this second simulation scenario there are two ways to evaluate the quality of the consensus linkage group. As previously proposed, we first computed the quantity $\overline{\text{IMSE}}$. Note that in this case IMSE(i) and IMSE(c) are not computed on the same marker intervals and we cannot presume the value of the ratio of these two quantities. Nevertheless it is a practical way to evaluate the mean square error of the consensus linkage relatively to the ones of the individual mapping experiments. We can also look at the ability of the consensus model to predict the marker interval recombination fractions in the individual mapping experiments. This can be done by substituting in each IMSE(i) the recombination rate estimates computed from the experiment data set by those deduced from the consensus linkage group. This leads to the Interval Mean Square Error of Prediction (IMSEP). Thus the ability of the consensus model to predict the recombination rate estimates in each mapping experiment can be evaluated by the indicator $\overline{\text{IMSEP}} = \frac{1}{n} \sum_{i=1}^{n} \text{IMSE}(i)/\text{IMSEP}(i)$.

Results: The results of simulations for the first scenario are depicted in Figure 3.1 for three different population types. As expected, the observed ratio increases with n, despite slightly lower than the value expected if the true variances for distance estimation would be used. Whatever the kind of population, this indicates that our approximation is not too crude and that further theoretical developments would only have a minor effect.

In Table 3.1 we reported the results of the second simulation scenario. This shows that the proportion of common markers can have a strong impact on the estimations of the consensus marker interval distances. For example when 2 mapping experiments have less than 75% of common markers both $\overline{\text{IMSE}}$ and $\overline{\text{IMSEP}}$ indicate a substantial loss in quality on the recombination rate estimates of the consensus linkage group. This is partially removed when n increases. However the results indicates that when n increases, although

the WLS approach leads generally to a consensus model with a good quality of prediction (measured by $\overline{\text{IMSEP}}$) whatever the proportion p of common markers, the intrinsic quality of the result (measured by $\overline{\text{IMSE}}$) strongly depends on this proportion (for $n = 10$ and $p = 0.15$ and 0.25 then the consensus linkage group have in average a lower IMSE than the individual mapping experiments). This can be explained by the fact that, for a given proportion of common markers p, the number of markers to position on the consensus linkage groups increases with n. In other word the gain due to the amount of information brought by the combination of common markers over the experiments is balanced by the markers which are only observed in a single mapping experiment (e.g. in our simulation the average number of distinct markers for $n = 2, 5, 10$ and $p = 0.75$ was $M = 30, 32, 36$).

Meta-analysis of QTL

QTL experiment summary

Now suppose that for a given trait a QTL detection has been carried out in the n mapping experiments. The minimal information supplied by a QTL detection consists of a set of estimated positions of the QTL ,denoted $\{\hat{x}_{ij}\}_{j=1,\ldots,q_i}$, and the corresponding proportion of variance explained by each QTL, the r-square values $\{\lambda_{ij}\}_{j=1,\ldots,q_i}$. Here q_i is the number of QTL detected for mapping experiment i on the linkage group (generally $q_i = 1$ or possibly 2). The confidence intervals (CI) of the $\hat{x}_{ij}$'s, denoted γ_{ij}, can also be reported. The construction of the CI may have been performed by different approaches :

- support interval : it is the most popular approach. Confidence interval is set as the map interval corresponding to a l loglikelihood odd ratio (LOD) decline either side of the LOD peak and one speaks of a LOD minus l (LOD$-l$) support region (Conneally *et al.* (1985), Lander and D. (1989)). By the mean of simulations Mangin *et al.* (1994) showed that this approximation is not correct for QTL having small effect leading to a biased CI. More recently Dupuis and Siegmund (1999) have shown that $l = 1$, and $l = 1.5$, support interval corresponds in fact to 90%, and 95%, confidence regions, respectively, in the case of a dense map of 1 cM spaced markers.
- likelihood method : Dupuis and Siegmund (1999)
- bootstrapping : Darvasi *et al.* (1993), Kao *et al.* (1999)

When the CI is not available it is possible to obtain an approximation of the CI by applying the empirical formula proposed by Darvasi and Soller (1997). By means of intensive simulations they showed that for either a backcross or a F_2 population the expected CI at 95% level can be expressed as $\text{CI}(95) \approx 530/(N\lambda)$ where N is the population size and λ the proportion of variance explained by the QTL. More recently Visscher and Goddard

(2004) have derived simple analytical equations which are in good agreement with the formula of Darvasi and Soller (1997).

Whatever the method used to estimate the uncertainty on the QTL locations we assume that the $\hat{x}_{ij}$'s are normally distributed around the true position x_{ij} of the QTL : $\hat{x}_{ij} \sim \mathcal{N}(x_{ij}, \sigma_{ij}^2)$ where σ_{ij}^2 is the variance of the estimated position which can be deduced from the confidence interval γ_{ij}. For a CI of β% (β depends on the method used to compute the CI), the standard deviation σ_{ij} can be estimated as $\sigma_{ij} = \gamma_{ij}/(2u_\beta)$ where u_β is the double-sided β-percentile of a centered normalized gaussian. This Gaussian approximation based on the classical asymptotic theory has been suggested by Goffinet and Gerber (2000), even though this is not perfectly correct for QTL with small effects (Mangin *et al.* (1994)).

Furthermore the n QTL mapping experiments are assumed to be consistent with the following assumptions :

- Hypothesis 1 : they are independent. This can be considered as correct when the individuals measured in the different populations have been generated independently. Independence between experiments i and $i^{'}$ means independence between $\hat{x}_{ij}$ and $\hat{x}_{i^{'}j}$.
- Hypothesis 2 : for a given trait there is a finite number of underlying QTL which cosegregate in the mapping experiments : this means that the populations share the same trait determinism with potentially different allelic configurations at the QTL. In other word there is a finite number of true QTL positions on the linkage groups, i.e $\{x_{ij}\}$ can potentially contain redundancy.

In addition to the two previous hypotheses we also assume that the detected QTL locations are independent within experiments. This is not really true when the QTL detection does not properly take into account linked QTL. But with the advent of composite interval mapping strategy (Jansen (1993), Zeng (1994)) multiple-QTL model can now be fitted by adding properly chosen cofactors which limit the impact of linkage between QTL on the position estimates. Therefore we assume that $\hat{x}_{ij}$ and $\hat{x}_{ij^{'}}$ are independent for all $j \neq j^{'}$.

Pre-processing

First the WLS strategy proposed in the previous section is applied to the n mapping experiments in order to build a consensus linkage group. Then the QTL locations are projected on the consensus linkage group using a simple homothetic rule between the original QTL flanking marker interval and the corresponding one on the consensus chromosome. For a given QTL location the new confidence interval (if available) on the consensus linkage group is computed by taking into account the average dilatation (expansion or contraction) between the original and the consensus chromosome. This is done by computing the sum over the common marker intervals of the ratio

of the interval lengths weighted by the probability that the QTL position might be in this interval. There are two possible strategies to approximate this probability. The first one relies on a rough approximation using a Gaussian distribution around the most likely position $\hat{x}_{ij}$ of the QTL, namely $\Pr(\text{QTL } j \text{ in } m) = \frac{\int_{y_m}^{y_{m+1}} \phi[(y-\hat{x}_{ij})/\sigma_{ij}]dy}{\int_0^L \phi[(y-\hat{x}_{ij})/\sigma_{ij}]dy}$ where $\phi[x]$ is the density function of a centered normalized Gaussian distribution, m is the index of the marker interval, y_m and y_{m+1} are the absolute positions of the flanking markers on the original map of total length L. If the LOD score profile is available, a more accurate strategy can be applied by substituting ϕ for the density function which best fits the profile.

The meta-analysis model

The purpose of the QTL meta-analysis is to evaluate the degree of congruency of QTL detected in the n mapping experiments and related to the same trait. By assuming that there is a finite number of true QTL locations Goffinet and Gerber (2000) proposed a clustering based approach to both classify the observed QTL and estimate the positions of the underlying QTL. Their method proceeds by testing all the possible QTL combinations and then choosing the one which maximizes a penalized log-likelihood. Although original, this method suffers from a categorical repartition of the QTL in the clusters, which is a limit case of Gaussian mixture models. We propose to adopt a similar clustering strategy but with a more standard Gaussian mixture model which allows QTL to be probabilistically distributed into clusters.

A Gaussian mixture model

In order to lighten the notation we note q the total number of observed QTL locations and we ignore the mapping experiment subscripts so that $\hat{X} = (\hat{x}_1, \ldots, \hat{x}_q)$ and $\Sigma = (\sigma_1, \ldots, \sigma_q)$. Then let's suppose there are $K \geq 1$ true QTL located at $X^{[K]} = (x_1, \ldots, x_K)$ which segregate in at least one of the n QTL mapping experiments. Since the QTL position estimates $\hat{X}$ are normally distributed around their true positions, the problem of finding the K underlying true positions can be viewed as a particular Gaussian mixture problem where the variances of each observation are known. Thus the log-likelihood of the observations can be written as follows :

$$L(\hat{X}, \Sigma; \Theta^{[K]}) \propto \sum_{i=1}^{q} \log \left[\sum_{j=1}^{K} \pi_j^{[K]} \phi \left(\frac{\hat{x}_i - x_j^{[K]}}{\sigma_i} \right) \right] \tag{3.1}$$

where $\Theta^{[K]} = (X^{[K]}, \Pi^{[K]})$ denotes the parameters of the model, $\Pi^{[K]} = (\pi_1^{[K]}, \ldots, \pi_K^{[K]})$ are the mixing proportions, summing to one, and $\phi(x)$ is

the density function of a centered normalized Gaussian distribution. We assume without loss of generality that $x_1^{[K]} < x_2^{[K]} < \ldots < x_K^{[K]}$ and that $\pi_j^{[K]} \neq 0$, $j = 1, \ldots, K$. In other word the distribution of the observed QTL locations is shaped by a mixture density where the components $x_j^{[K]}$ are the positions of the true QTL on the linkage group and the mixing proportions π_j represent the proportion of QTL related to the j^{th} true QTL which have been detected in the n mapping experiments.

Maximizing 3.1 can be achieved via a standard EM algorithm (Dempster *et al.* (1977)) by using the following parameter updates (M-step) :

$$x_j^{[K]} = \frac{\sum_{i=1}^{q} t_{ij}^{[K]} \hat{x}_i \sigma_i^{-2}}{\sum_{i=1}^{q} t_{ij}^{[K]} \sigma_i^{-2}} \quad \text{and} \quad \pi_j^{[K]} = \frac{1}{q} \sum_{i=1}^{q} t_{ij}^{[K]}$$

where $t_{ij}^{[K]}$ is the conditional probability that $\hat{x}_i$ belongs to the j^{th} meta-QTL. This conditional probability $t_{ij}^{[K]}$ is obtained by applying a simple Bayes' rule evaluated at the current parameter estimates (E-step) :

$$t_{ij}^{[K]} = \frac{\pi_j^{[K]} \phi\left(\frac{\hat{x}_i - x_j^{[K]}}{\sigma_i}\right)}{\sum_{j'=1}^{K} \pi_{j'}^{[K]} \phi\left(\frac{\hat{x}_i - x_{j'}^{[K]}}{\sigma_i}\right)}$$

The EM-algorithm is run until reaching convergence : this yields the maximum-likelihood estimate denoted $\tilde{\Theta}^{[K]} = (\tilde{X}^{[K]}, \tilde{\Pi}^{[K]})$. Finally, once $\tilde{\Theta}^{[K]}$ has been obtained the variance-covariance matrix of the parameter estimates, conditionally to the current model, can be computed by applying the Supplemental EM (SEM) strategy proposed by Meng and Rubin (1991).

Model choice

The problem is that we do not know K, i.e the number of true QTL positions. Since the mixture model of K components is nested into the model with $K+1$ components, likelihood ratio test (LRT) should appear suitable. However, as discussed by many authors (see for instance Aitkin and Rubin (1985), Titterington *et al.* (1985)) the LRT statistic does not follow the usual χ^2 distribution due to testing a null hypothesis on the boundary of the parameter space (i.e the regularity conditions on the loglikelihood do not hold). Another strategy is to use the Kullback-Leibler information in order to derive the so called information criterion which is widely used to select statistical model. In particular Kullback-Leibler information can be viewed as a measure of goodness-of-fit of a statistical model. Here for a given

value of K, minimizing the Kullblack-Leibler information is equivalent to maximizing the negentropy KL,

$$\mathrm{KL} = -\int_{\hat{X}} g(\hat{X}, \Sigma) L(\hat{X}, \Sigma; \Theta^{[K]}) d\hat{X}$$

where $g(\hat{X}, \Sigma)$ is the the true underlying density function. Thus, from the point of view of the negentropy maximization principle, the goodness of the model can be evaluated by the expected log-likelihood. Note that the negentropy maximization principle naturally leads to the maximization of the log-likelihood. However the maximized log-likelihood is a naive estimate of the expected loglikelihood : since the same data set $\hat{X}$ is used for both the estimation of the parameter and the estimation of the expected log-likelihood, $L(\hat{X}, \Sigma; \Theta^{[K]})$ is a biased estimator of the expected loglikelihood. Its bias is defined by,

$$B = E_{\hat{X}}\left[L(\hat{X}, \Sigma; \tilde{\Theta}^{[K]}) - E_Y\left[L(Y, \Sigma; \tilde{\Theta}^{[K]})\right]\right]$$

and the use of $L(\hat{X}, \Sigma; \tilde{\Theta}^{[K]}) - B$ is justified as an estimate of KL. There are different strategies to estimate this bias and several information based criteria have been reported in the mixture model literature in order to tackle the issue raised by choosing the number of components (see Appendix B). Here, we propose to evaluate some of these information based criteria to determine the optimal number of QTL.

Simulation study

Since the goal of QTL meta-analysis is to obtain a better predictive inference of the true QTL locations we have compared the two alternative strategies :

- Strategy 1 : choose the model with as many true QTL as the number of observed QTL. It is the naive model, $\tilde{x}_i(1) = \hat{x}_i$
- Strategy 2 : choose the best model K according to the model choice criterion, $\tilde{x}_i(2) = \sum_{j=1}^{K} t_{ij}^{[K]} \tilde{x}_j^{[K]}$.

For each strategy $s = 1, 2$ the measure of performance used was the mean squared error of prediction

$$\mathrm{MSEP}(s) = \frac{1}{q}\sum_{i=1}^{q} E[(x_i - \tilde{x}_i(s))^2]$$

Absolute values of these MSEP are not of interest here because our goal is comparison of strategies ; hence, we consider the ratios MSEP(2)/MSEP(1) for 5 different criterion : AIC, AIC_c, AIC3, BIC and EIC. This last criterion was obtained by means of simulations (data not shown) leading to a simple relation between EIC and AIC, EIC $\approx$ AIC $-K + 1$. Note that here EIC is not the empirical information criterion defined by Ishiguro *et al.* (1997).

Generating data based on the Gaussian mixture model

We assume that the complexity which shapes the distribution of the observed QTL along the chromosome can be represented by our mixture model. In order to explore mixture configurations which are realistic we have assumed that the QTL effects have a L-shaped distribution (i.e most of the detected QTL in mapping experiments have a small effect and only a few show a strong effect, or in other words, most of the detected QTL have large confidence intervals). Consequently this implies that Σ^{-1} has also a L-shaped distribution (i.e the smaller the effect of the QTL the larger the confidence interval of the estimated QTL position). Then for a given value of the number of true QTL, K, we randomly generated configurations as follows :

1. Draw Σ from a inverse gamma distribution with shape parameter $\beta = 4$ and a scale parameter $\alpha = 2$ (this simply mimics a L-shaped distribution).
2. Generate the mixing proportions by choosing them over the discrete uniform $]0, 1[$ distribution subject to constraint $\sum_{k=1}^{K} \pi_k = 1$ and $0.1 < \pi_k < 0.9$ for $k = 1, \ldots, K$.
3. Draw from the mixing proportions the origins of the observed QTL $Z = (z_1, \ldots, z_q)$ where $z_{ij} = 1$ if the i^{th} observed QTL belongs to the j^{th} true QTL, 0 otherwise.
4. Generate the true QTL positions, $X = (x_1, \ldots, x_k)$, subject to constraint $x_k + \tau_{min} < x_{k+1} < x_k + \tau_{max}$ where τ_{min} and τ_{max} are defined so that the distance between x_k and x_{k+1} lies between δ_{min} and δ_{max}. The distance δ is defined as the mahalanobis distance between x_k and x_{k+1} : $\delta = \sqrt{\frac{(x_k - x_{k+1})^2}{a_k^2 + a_{k+1}^2}}$ where $a_k = (\sum_{i=1}^{q} z_{ik}\sigma_i)/(\sum_{i=1}^{q} z_{ik})$ is the average standard deviation for the k^{th} true QTL. This measures the separation between consecutive true QTL relatively to the precision of the experiments : $\delta \leq 2$ corresponds to tightly or moderately separated QTL, while $\delta \geq 3$ corresponds to widely separated QTL.

We stress that this process is not an attempt to describe reality, nevertheless it makes it possible to cover a large range of possible repartitions of the QTL. Finally, for each of the 4 distance constraints considered ($\delta_{min} = 1, 2, 3, 4$ and $\delta_{max} = \delta_{min} + 1$), 50 configurations were generated. For a given value of K, 200 MSEP(s) values were computed by repeated 100 times the following scenario :

1. Draw a sample $\hat{X}$ of size q.
2. Run the EM-algorithm to obtain $\tilde{\Theta}^{[K]}$ for $K = 1, \ldots, q$.
3. Choose the best model according to each criterion.

Results

In Figure 3.2 we summarized the result of simulations for several values of K and q by averaging over the distance constraint configurations ($\delta_{\min} = 1, 2, 3$ and 4). At first sight the 5 criteria seem to have the same behavior whatever the configuration, except for AIC3 which crucially underperforms for small values of q. For reasonable sample size relatively to the true number of components the meta-analysis appears to be more efficient than strategy 1. Since the AIC criterion have relatively good performance in these simulations we assume that there is no need for a specific theory to deal with this kind of mixture models and that this criterion can be used to carry out model selection in this context. So in Figure 3.3 we focus on the AIC criterion for the different distance configurations $\delta_{\min} = 1, 2, 3$ and 4. This clearly shows that for configurations with reasonable separation between the true positions of the QTL, the meta-analysis performs relatively well in most of cases. It is worth noting that the better the probability to choose the true model, the better the quality of the QTL position estimates. In order to evaluate the ability of the meta-analysis to improve the precision on the "true" QTL locations we computed the quantities $|x_i - \hat{x}_i(s)|$ and calculated the quantiles at 95 and 90% of its empirical distribution over all the QTL for the two strategies. The smaller this confidence interval, the better the estimated position $\hat{x}_i(s)$. We reported in Table 3.2 the average ratios of these quantities between the two strategies. Hence, if there are actually one, two, three or four different QTL locations with a reasonable separation ($\delta_{\min} \geq 2$), we can see that the meta-analysis not only gives better estimates of the QTL locations but also makes it possible to reduce the length of the 95% CI (in most of the situtations this length is halved). According to Darvasi and Soller (1997) to halve a CI in a QTL experiment, one needs to use at least two times the initial number of individuals.

Finally, since simulations presented above have been done assuming independence between experiments and known variances, we carried out additional simulations to evaluate the impact of nonindependent observations and to consider the effects of imperfect knowledge of the variance : the results indicated that the meta-analysis is quite robust in these cases (data not shown).

Application to flowering time in maize

QTL mapping experiments

Recently Chardon *et al.* (2004) made a bibliographical review of QTL studies relative to 4 traits related to flowering time in maize : days to pollen shed (DPS), silking date (SD), plant height (HT) and leaf number (LN). From the 22 QTL studies they reported, we excluded 6 experiments for

which QTL detection was based on ANOVA with a low density of markers and 2 others for which it was not possible to get exact information on both genetic linkage map and QTL locations. In addition to these 15 mapping experiments we considered 3 other recent experiments. Details of these 18 QTL studies are given in Table 3.3. We focus here on chromosome 8.

Consensus map for chromosome 8

Among the 153 distinct markers which have been positioned over the 18 mapping experiments on the chromosome 8 only 53 markers are observed in at least two different mapping experiments. We restricted the meta-analysis to these 53 markers. Only one order inconsistency was detected between Poupard *et al.* (2001) and Mechin *et al.* (2001) concerning markers *umc89a* and *umc12a*. As in Poupard *et al.* (2001) *umc12a* is very close to *umc89a* (less than 2 cM) we have decided to ignore this marker in this mapping experiment. Over the 18 mapping experiments the mean interval distance is about 18.9 cM with an average of 8.7 markers per mapping experiment and it exists at least one common marker path which connected all the mapping experiments together (insuring that the WLS can be applied). The consensus linkage group of chromosome 8 is depicted in Figure 3.4. The goodness-of-fit of the consensus chromosome is relatively bad : χ=365.31 with $\chi \sim \chi^2_{87}$. It may be due to some recombination rate heterogeneities between mapping experiments, may be located in the filled marker intervals of Figure 3.4. Note that variability of recombination rate in maize was first reported by Stadler (1925) and more recently Williams *et al.* (1995) demonstrated that exotic inbred lines exhibit higher recombination rate that U.S. inbreds origin along chromosome 1 (see also Ji *et al.* (1999)). On the other hand, since no information about the marker configurations in each individual mapping experiment was available, we have computed the variances of the distance estimates by assuming no missing data and no ambiguities (dominance) in the original data sets. This is surely too optimistic and some data sets may have included missing data and/or dominant markers. Therefore the precision on the distance estimate can have been overestimated for some marker intervals.

QTL meta-analysis

From the 18 QTL studies we projected 34 QTL on the consensus chromosome 8. Among these 34 QTL, 16 (47%) are related to SD, 10 (29%) to DPS and 8 (24%) to HT. The distribution of the r-square values clearly show a L-shape : 75% of the QTL have r-square values lower than 12%. For 17 QTL a CI was reported (build from a 1-LOD support) from which we computed the standard deviations assuming that a 1-LOD support corresponds in fact to a 90% CI. For the other QTL we derived the standard deviations from

the formula proposed by Darvasi and Soller (1997). Then models from $K = 1$ to $K = 10$ QTL were considered and their parameters estimated by applying the EM-algorithm as previously defined.

In Table 3.4 we give the Δ_K and the w_K values (see Appendix B) for the criteria AIC, AIC_c, AIC3 and BIC for the different values of K explored. This clearly shows that the model with 5 QTL is the best one. Apart from models with 6 and 7 QTL the Δ_K values suggest that using another model to fit the data leads to a substantial loss in information. For the model with 5 QTL the parameter estimates are listed in Table 3.5 and depicted in Figure 3.5. First, 3 QTL (1,4 and 5) have been detected in only 22% of the mapping experiments. At least two observed QTL are assigned to each of these 3 QTL without ambiguity.

Secondly, two closely linked QTL (2 and 3) contribute to 75% of the reported QTL. This is strongly consistent with the knowledge of this region where a major QTL, *vgt1*, is tightly linked to another QTL, *vgt2* (Vladutu *et al.* (1999), Salvi *et al.* (2002)). It is worth noting that the confidence interval of the QTL corresponding to *vgt1* (around 3.8 cM) encompasses a marker interval of approximately 2 cM in which this QTL has been finely mapped by Salvi *et al.* (2002) using NIL lines (result not included in our analysis). This congruency lends further credence to the meta-analysis approach.

Discussion and Conclusion

Nowadays more and more studies concerning QTL detection are available via public databases and the number of articles dealing with the comparison and/or integration of these results increases Khavkin and Coe (1997, 1998); Chardon *et al.* (2004, 2005). We believe that our meta-analysis procedure can greatly contribute to facilitate the elaboration of such syntheses by providing a simple statistical framework to establish consensus models for both linkage maps and QTL locations.

First, the WLS strategy we proposed is a step forward to integrate several genetic marker maps. Contrary to iterative projection procedures, this approach provides a well-established statistical machinery (WLS) to assess the goodness-of-fit of the consensus model. It can also be used to test the homogeneity of the distance estimates among different mapping experiments. This can be usefull to investigate the possible variation of recombination rate among genotypes (as reported by Williams *et al.* (1995)). As pointed out in the application, this method can suffer from the lack of knowledge about the effective precision on the marker interval distances in each individual mapping experiment due to possible missing data and/or the type of scoring of individual markers (codominance vs dominance). This could be improved by asking researchers to supply the variance estimates

of the marker interval distances when they submit their results to a public database. These variance estimates could be used to improve the weight factors in the WLS model. Also, as sometimes robust framework maps are available in the literature or via public databases (e.g the IBM2 map in maize available at `http://www.maizegdb.org`), this method can be easily modified in order to fix a genetic map as a reference (i.e for which the distances between ordered markers are assumed to be the "actual" distances). In this case only the positions of the markers which are not reported on the reference have to be estimated.

Secondly, for the QTL meta-analysis itself, the Gaussian mixture model used to fit the distribution of the observed QTL locations on the chromosome provides a well-studied statistical inference technique. In this model-based clustering, each "true" QTL is mathematically represented by the Gaussian distribution of its detected positions, which leads to a probabilistic classification of the observed QTL. Simulation results reveal that usual model choice criteria give relatively good performances in this context. In particular, it brings out that the well-established AIC criterion can be used in most of the cases. Contrary to Goffinet and Gerber (2000) who proposed a specific model choice criterion, the results show that AIC gives relatively good performances in the QTL meta-analysis context. This difference with regard to the conclusions of Goffinet and Gerber (2000) may be explained by their discrete formulation of the problem (recall that, instead of using a usual Gaussian mixture likelihood to evaluate the probability of the data, they assumed that the observations could be categorically assigned to the mixture components). Parameter estimates obtained by this approach were not really the maximum-likelihood estimates of the underlying mixture model. This may have added a bias in the evaluation of the AIC criterion, which could explain the bad performances they obtained for this criterion in their simulations.

Thus, our mixture-modelling approach makes it possible to go beyond the limits encountered by Goffinet and Gerber (2000) : the Akaike like criterion they proposed was limited to models from 1 to 4 QTL. As a consequence, Chardon *et al.* (2005) who used the method of Goffinet and Gerber (2000), was obliged to break chromosomes on distinct segments to carry out the meta-analysis. This subjective division of the chromosome can now be avoided thanks to our method. Simulations have shown that the ratio between the number of observed QTL and the number of "true" QTL is one of the main limiting factor. The number of "true" QTL which can be assessed by the meta-analysis must be reasonable compared to the number of observed QTL (at least between 5 or 10 observed QTL per actual location). Note that this also depends on the distance between true QTL. But since there are more and more QTL locations reported for a given trait and since the real number of distinct QTL locations which can be detected with usual experimental designs is limited (only QTL with relatively large effects can be found), we assume that in many cases the ratio between observed and "true"

QTL locations will steadily increase and should generally be reasonable. It is worth noting that, provided that the number of observed QTL is not too small, the meta-analysis is able to separate "true" QTL locations even if they are closely linked (as illustrated with *vgt1* and *vgt2* in the application, and the consistency of the *vgt1* estimated position with fine mapping results of Salvi *et al.* (2002) not included in the meta-analysis).

The ultimate step toward a more accurate identification of QTL relies on finding the underlying genes. Up to now, the majority of QTL isolated in plants have been cloned via positional cloning (see for instance Salvi and Tuberosa (2005)). However positional cloning of QTL is quite expensive both in terms of time and resources due to the necessity to screen recombinant individuals within large population (typically several hundreds) and to characterize these individuals with a very dense set of molecular markers. As an alternative and thanks to the advent of structural and functional genomic, QTL can also be resolved through association mapping of candidate genes. Candidate genes identification is based on a assumption that the polymorphism of the gene is associated with the variation of the trait of interest. Both function and mapping information have to be crossed to establish this assumption. The function of the gene may have been determined in the species of interest, based for instance on mutant analysis. More often, function is hypothesized based on sequence homology with genes the function of which has been established in model species, including possible positional cloning of QTL. Gene mapping information may have been obtained in the species of interest, but may have been also inferred from synteny based projections, as illustrated by Chardon *et al.* (2004) for rice to maize. Relevancy of the colocalization between QTL and candidate genes crucially depends on the confidence interval of the QTL positions. For this purpose the reduction of the confidence interval of the QTL is an important goal Kearsey and Farquhar (1998). The ability of our method to reduce the QTL confidence interval by taking advantage of pooling QTL results could contribute in an increased resolution in selecting candidate genes. It is worth noting that candidate genes are generally mapped on a framework map used as reference for the species of interest (e.g in maize Falque *et al.* (2005)), while the QTL detections are carried out specific populations (generally obtained by crossing parents contrasted for the trait(s) of interest). Therefore, the selection of candidate genes which colocalize with QTL depends also on the process used to merge these different maps. Up to now, no statistical method had been proposed to combine candidate genes and QTL mapped in independent experiments. Our WLS strategy should increase the precision of the integration of candidate gene mapping information.

Finally once candidate genes have been selected and their different haplotypes defined, association studies can be carried out. The identification of a statistically significant association between haplotype variation at a candidate gene and the target trait gives further credence on the role of this gene in the trait

variation. Since the last 5 years more and more association studies have been reported in plants GUPTA *et al.* (2005). It would be interesting to integrate these new results into a global meta-analysis framework. Further developments are needed to combine onto a synthetic model the different scale of mapping : from linkage mapping (QTL) to fine mapping (association studies).

Appendix A

Let's assume that g classes of genotypes are expected in the frequencies $\{p_j\}_{j=1,\ldots,g}$ which are function of r, the recombination fraction. MATHER (1936) discussed in details how to estimate r from the derivative of the log-likelihood,

$$\frac{\partial L}{\partial r} = \sum_{j=1}^{g} a_j \frac{\partial p_j}{\partial r}$$

where a_j is the observed number of genotypes for the j^{th} class, $j = 1, \ldots, g$. He also defined the mean amount of information, i_r, supplied by a single individual as,

$$i_r = \sum_{j=1}^{g} \left[\frac{1}{p_j} \left(\frac{\partial p_j}{\partial r} \right)^2 \right]$$

from which the standard error of r, σ, can be derived : $\sigma = (N i_r)^{-1}$ where N is the number of individuals in the population.

Let's consider the results of crossing two parents $AABB$ and $aabb$. For backcross design $g = 4$ classes can be discriminated, namely $AABB$, $AABb$, $AaBB$, $AaBb$, and the individual information is given by

$$i_r = \frac{1}{r(1-r)}.$$

For selfed populations, usual marker techniques (e.g RFLP markers showing codominant segregation) make it possible to distinguish nine genotypic classes over the ten possible genotypic configurations : the $AaBb$ class generally includes the two double heterozygous genotypes AB/ab and Ab/aB unless the phase can be resolved. The expected frequencies $p_1, p_2, \ldots, p_9$ can be expressed in terms of the HALDANE and WADDINGTON (1930) zygotic proportions :

$$\begin{cases} C_t & = AABB + aabb = p_1 + p_2 \\ D_t & = AAbb + aaBB = p_3 + p_4 \\ 2E_t & = AABb + AaBB + Aabb + aaBb = p_5 + \ldots + p_8 \\ \frac{1}{2}(F_t + G_t) & = ABab = p_9 \end{cases}$$

where t is the number of selfing generations. The classes C_t, D_t, E_t, F_t and G_t are subject to the constraint $2C_t + 2D_t + 4E_t + F_t + G_t = 2$ so that

$C_1 = D_1 = E_1 = G_1 = 0$ and $F_1 = 2$ (HALDANE and WADDINGTON (1930)). It follows that the mean average amount of information in a selfed population following t generations of self-fertilization is given by,

$$i_r = \frac{1}{C_t}\left(\frac{\partial C_t}{\partial r}\right)^2 + \frac{1}{D_t}\left(\frac{\partial D_t}{\partial r}\right)^2 + \frac{2}{E_t}\left(\frac{\partial E_t}{\partial r}\right)^2 + \frac{1}{2(F_t + G_t)}\left(\frac{\partial (F_t + G_t)}{\partial r}\right)^2$$

where the derivatives of the expected frequencies of each class with respect to r can be obtained by derivation of the recurrence equations.

Self-fertilized recombinant inbred line population is a limit case of selfed population when $t \to \infty$. Haldane and Waddington HALDANE and WADDINGTON (1930) have demonstrated that the fraction of crossover events observed, R, is related to the recombination frequency r per meiosis by the formula,

$$r = D(R) = \frac{R}{2(1-R)}$$

In term of R the mean amount of information is similar to the backcross case and leads to,

$$i_R = \frac{1}{R(1-R)}$$

It can be showed that i_r is directly related to i_R by the equation,

$$\begin{aligned} i_r &= \left(\frac{\partial D(R)}{\partial R}\right)^{-2} i_R \\ &= \frac{2}{r(1+2r)^2} \end{aligned}$$

More recently LIU *et al.* (1996) and WINKLER *et al.* (2003) have extended the equations of HALDANE and WADDINGTON (1930) to the case of intermated populations. It comes from these results that i_r of a single lineage in an population following t generations of random mating is given by,

$$i_r = \frac{(1-r)^{2t-2}[2(1-r) + t(1-2r)]^2[1 + 3(1-2r)^2(1-r)^{2t}]}{[1 - (1-2r)^4(1-r)^{4t}]}$$

and that the relation between the fraction of crossover events observed in a self-fertilized intermated recombinant inbred population to the recombination frequency per meiosis is,

$$R = \frac{1}{2}\left(1 - \frac{1-2r}{1+2r}(1-r)^t\right).$$

Then $D(R)$ is the function which values are the solutions of the equation

$$2[1 - D(R)]^{t+1} - [1 - D(R)]^t + [2 - 4R][1 - D(R)] + 3[2R - 1] = 0$$

Although in this case there are no analytical formula for both $r = D(R)$ and $i_r = \left(\frac{\partial D(R)}{\partial R}\right)^{-2} i_R$, this quantities can be evaluated using standard numerical methods. Finally if a pair of markers is fully informative or if there are enough individuals with no missing information, σ can be consistently estimated by applying the above strategy according to the family structure. Otherwise the number of classes to consider is the number of observed classes $\tilde{g} < g$ and σ can be computed using the same procedure by substituting g by $\tilde{g}$ (or more advanced strategies can be used such as applying the SEM of Meng and Rubin (1991) approach if a EM algorithm was used to infer the recombination rate from data).

Appendix B

Assuming that model is true, regularity conditions of the loglikelihood and asymptotic normality of the MLE, Akaike (1973) proved that B can be asymptotically approximated by the number of free parameters in the model. This leads to the well-established expression,

$$\mathrm{AIC} = -2L(\hat{X}, \Sigma; \tilde{\Theta}^{[K]}) + 2\nu$$

where $\nu = 2K - 1$. When ν is large relative to the sample size q there is a small-sample version of AIC called AIC_c,

$$\mathrm{AIC}_c = -2L(\hat{X}, \Sigma; \tilde{\Theta}^{[K]}) + 2\nu + \frac{2\nu(\nu+1)}{q-\nu-1}$$

which should be used unless $q/\nu >$ about 40 for the model with the largest value of ν (see Sugiura (1978)). These easily computable information criteria are also an extension of Fisher's loglikelihood theory Akaike (1992). It is worth mentioning that without assuming that the model is true Takeuchi (1976) derived an asymptotically unbiased estimator of expected log-likelihood. His method, which requires quite large sample sizes to reliably estimate the bias adjustment term (details in Burnham and Anderson (2002)), leads in many cases to a correction approximately equal to ν giving further credence to use AIC and AIC_c in practice.

Since AIC and AIC_c rely on the usual asymptotic theory of the MLE and that the regularity conditions do not hold when comparing two contrasted mixture models, they are not a correct way of comparing models in the case of mixture (see Titterington *et al.* (1985), Aitkin and Rubin (1985)). However due to its simplicity and its compelling concept AIC have been widely used in mixture model applications and seems to yield relatively good performances in simulation studies (see for instance Biernacki and Govaert (1998)).

By means of a Monte-Carlo approach Wolfe (1971) obtained an approximation of the null distribution of the loglikelihood ratio test (LRT) when testing two

contrasted hypotheses on the number of components in a Gaussian mixture. Bozdogan (1987) proposed to use this approximation leading to a modified AIC criterion, namely AIC3 defined by

$$\text{AIC3} = -2L(\hat{X}, \Sigma; \tilde{\Theta}^{[K]}) + 3\nu$$

More recently Bozdogan (1990) proposed an informational complexity criterion, called ICOMP, for choosing parsimonious models. It requires to compute the Fisher information matrix of the model which can make its computation tedious (Cutler and Windham (1993) suggested to approximate the Fisher information matrix with its empirical mean to derive ICOMP). It is worth mentioning that Windham and Cutler (1992) have also introduced another criterion, called MIR, which is based on the smallest eigenvalue of the ratio of Fisher information matrices (MIR can be computed from the EM convergence rate).

Another widely used criterion in both frequentist and bayesian mixture context was originally proposed by Schwarz (1978), the bayesian information criterion defined as

$$\text{BIC} = -2L(\hat{X}, \Sigma; \Theta) + \nu \log(n)$$

which is based on an approximation of the Bayes factor. Note that BIC can also be derived as a non-Bayesian result and that like AIC the BIC approximation is only valid when standard regularity conditions regarding the loglikelihood are verified (Burnham and Anderson (2004) give more detail on the deep foundations of both BIC and AIC).

Finally, whatever the information criterion used to select the best model the individual criterion values are not generally interpretable. It is imperative to rescale its values. For example, the Akaike information criterion, AIC, can be rescaled as follows :

$$\Delta_K = AIC_K - AIC_{K^*}$$

where K^* is the value of K which gives the minimal value of AIC for the $K_{\max}$ different AIC_K values. Δ_K is easy to interpret as the information loss experienced if we are using a model with K components rather than the best model with K^* components for inference. Hence the Δ_K's allow a quick strength-of-evidence comparison and ranking candidate models. In particular one can compute the useful "weights of evidence" w_K given by,

$$w_K = \frac{\exp(-\Delta_K/2)}{\sum_{j=1}^{K_{\max}} \exp(-\Delta_j/2)}$$

which can be interpreted as the probability that model K is in fact the best model for the data.

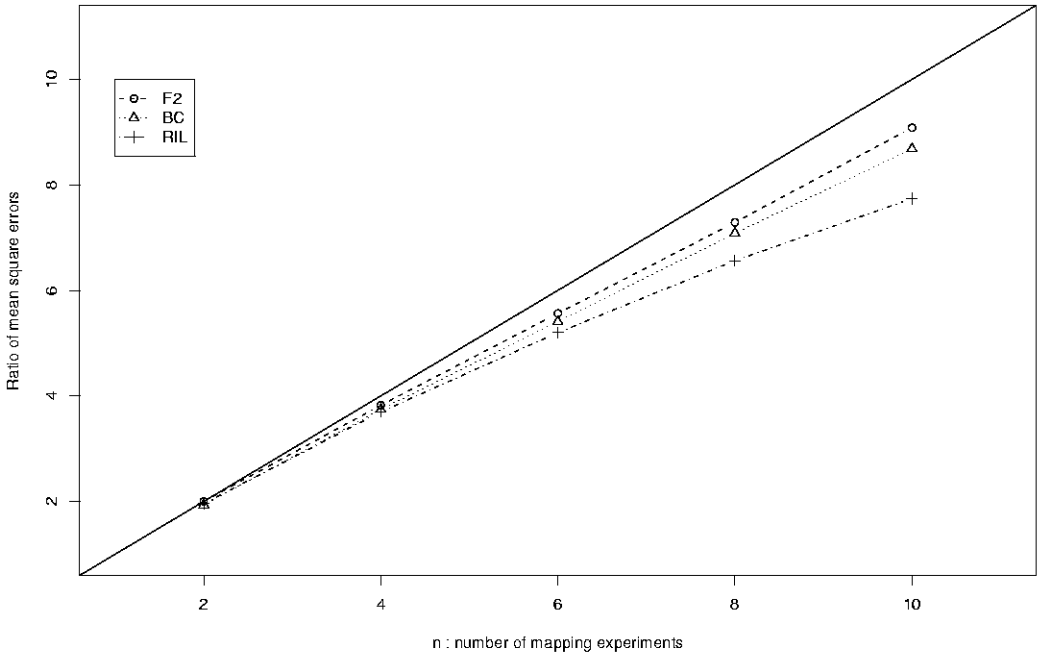


FIG. 3.1 – Average values of $\overline{IMSE}$ over 50 marker configurations (scenario 1) for 3 kinds of pedigree : backcross (BC), F2 and recombinant inbred lines via selfing (RIL). The solid line represents the expected values of $\overline{IMSE}$. See the simulation section for the definition of $\overline{IMSE}$.

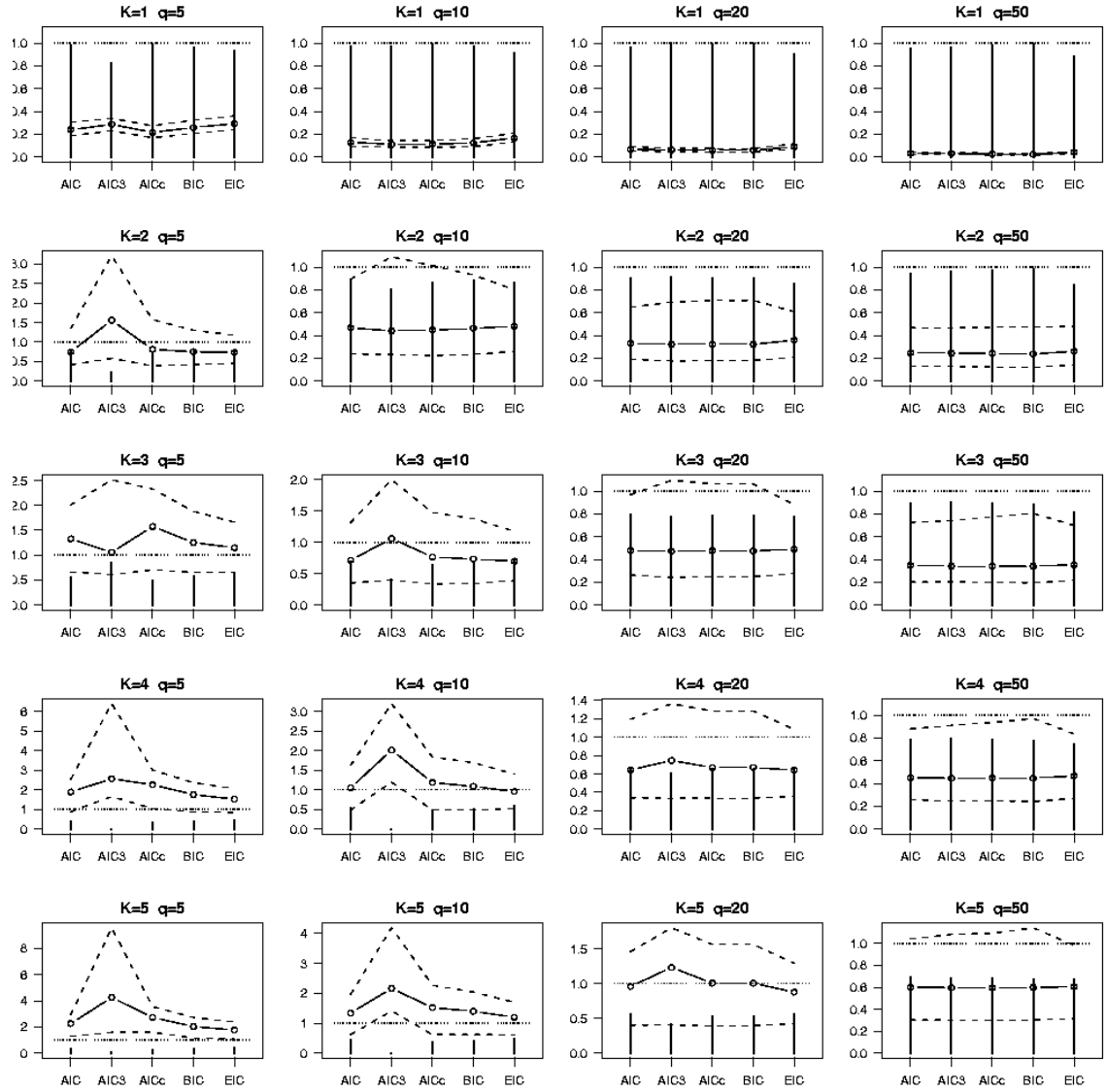


FIG. 3.2 – Simulation results for different values of the true number of QTL, K, and the number of observed QTL, q. The vertical bars indicate the probability that the best model selected by the criterion is the true model. The open circles, respectively the dotted lines, represent the mean, respectively the 0.1% and 0.9% quantiles, of the ratio between MSEP(2)/MSEP(1) for each criterion.

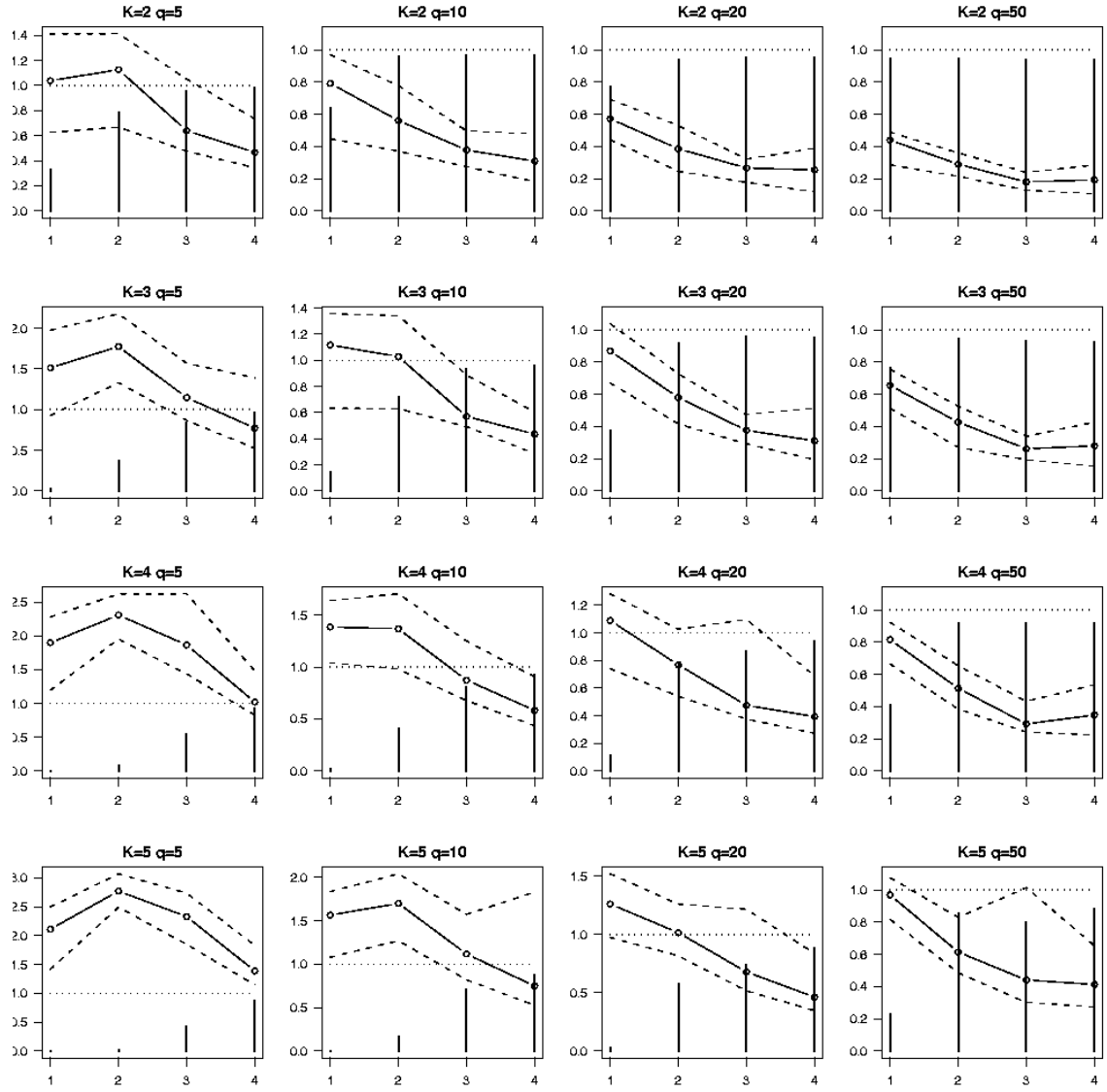


FIG. 3.3 – Behavior of the AIC criterion for the 4 distance constraints, $\delta_{\min} = 1, 2, 3$ and 4. The vertical bars indicate the probability than the AIC criterion has selected the true model. The open circles, respectively the dotted lines, represent the mean, respectively the 0.1% and 0.9% quantiles, of the ratios between MSEP(2)/MSEP(1).

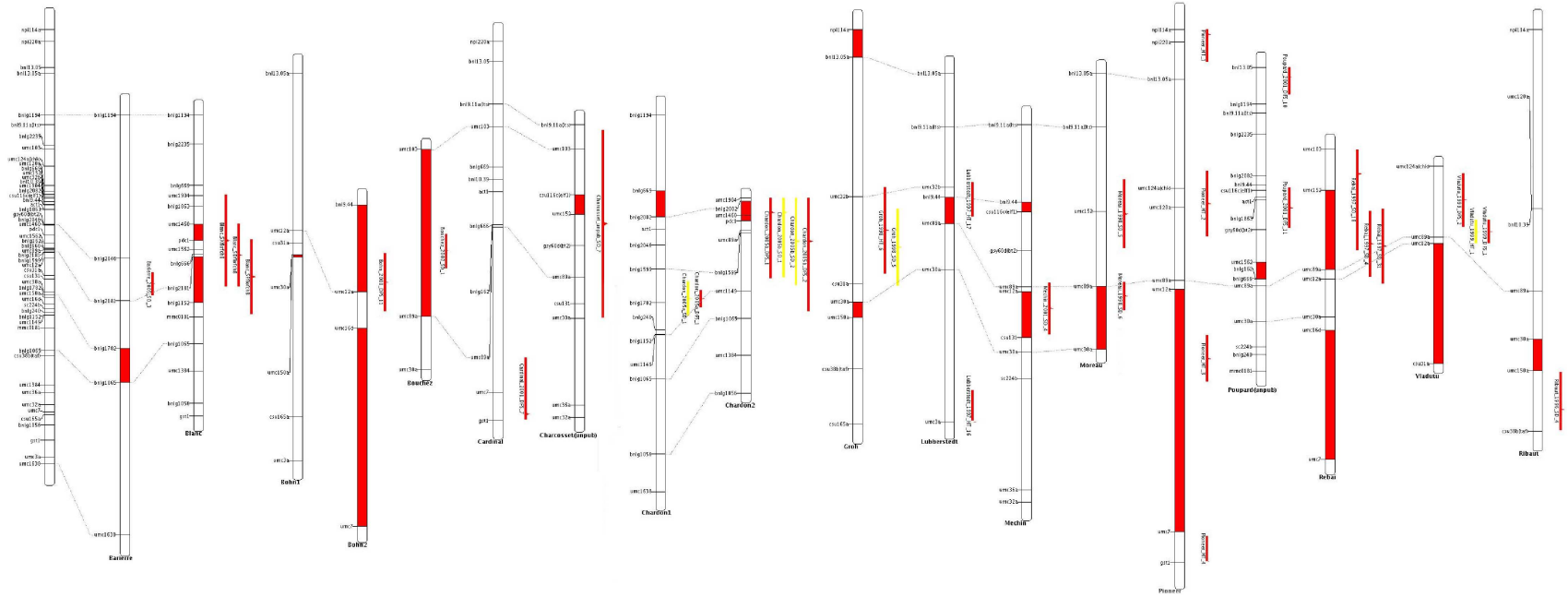


FIG. 3.4 – Overview of chromosome 8 for the 18 mapping experiments involved in the meta-analysis of flowering time in maize. The first chromosome at the left represents the consensus chromosome obtained by applying the WLS approach as described in the first section of the article. The filled marker intervals indicates that the standardized residual between the interval distance estimates of the original chromosome and the consensus one exceeded the double-sided 95% percentile of a normalized centered gaussian.

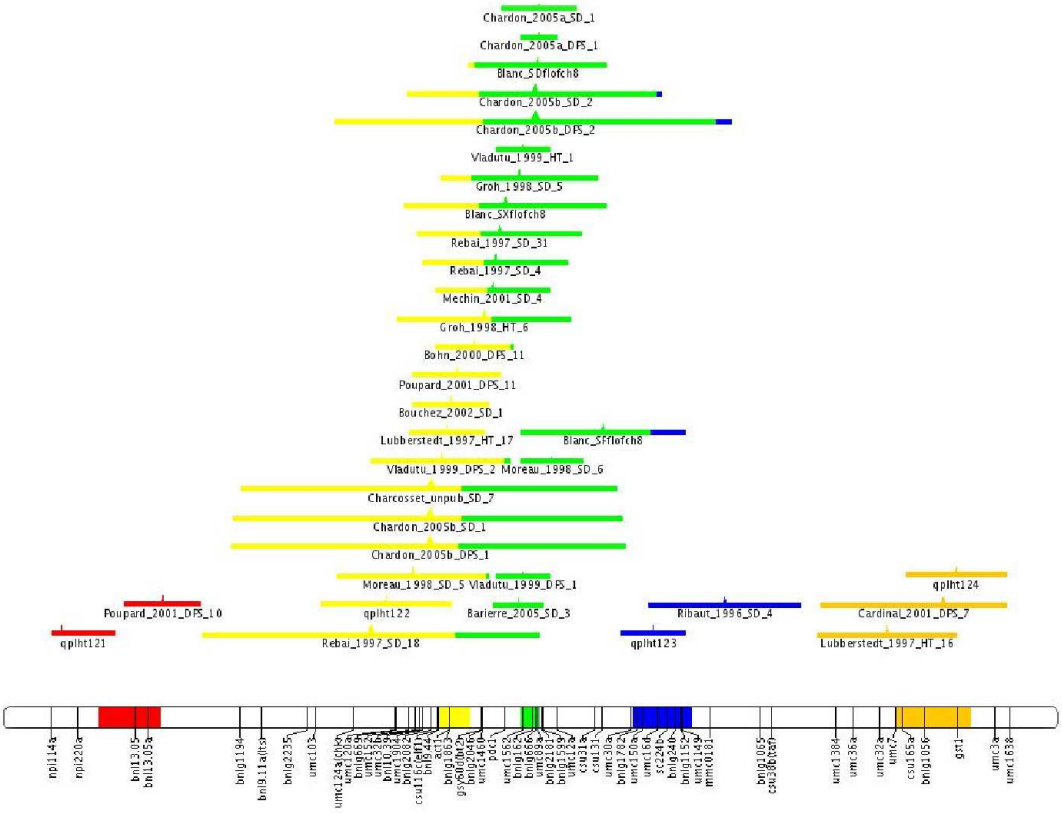


FIG. 3.5 – Result of the meta-analysis of the 34 QTL projected on the consensus chromosome 8. The CI of the meta-QTL positions are indicated on the chromosome by the filled area. The observed QTL positions are depicted by their most probable position (triangle) and the CI of the QTL are quantitatively colored according to the membership probabilities of the QTL.

	n	2				5				10			
	p	0.15	0.25	0.50	0.75	0.15	0.25	0.50	0.75	0.15	0.25	0.50	0.75
BC	$\overline{\text{IMSE}}$	0.28	0.44	0.82	**1.24**	0.33	0.52	**1.01**	**1.62**	0.43	0.64	**1.02**	**1.65**
	$\overline{\text{IMSEP}}$	0.74	0.78	0.95	**1.06**	0.97	**1.17**	**1.62**	**2.08**	**1.32**	**1.67**	**2.45**	**3.30**
F2	$\overline{\text{IMSE}}$	0.24	0.54	0.87	**1.39**	0.33	0.58	**1.22**	**1.97**	0.46	0.76	**1.40**	**3.01**
	$\overline{\text{IMSEP}}$	0.71	0.76	0.90	**1.02**	0.94	**1.16**	**1.60**	**1.97**	**1.30**	**1.67**	**2.35**	**2.80**

TAB. 3.1 – Average values of $\overline{\text{IMSE}}$ and $\overline{\text{IMSEP}}$ over 25 marker configurations (scenario 2) for a given proportion p of common markers between n mapping experiments, for backcross (BC) and F2 designs. The values are indicated in bold when the meta-analysis leads to a consensus linkage groups with a better quality in terms of recombination rate estimates than the individual mapping experiments. See the simulation section for the definition of both $\overline{\text{IMSE}}$ and $\overline{\text{IMSEP}}$.

δ		1		2		3		4	
K	q	q90	q95	q90	q95	q90	q95	q90	q95
2	20	**0.75 (100%)**	**0.84 (100%)**	**0.52 (100%)**	**0.66 (98%)**	**0.39 (100%)**	**0.46 (100%)**	**0.29 (100%)**	**0.30 (100%)**
	50	**0.62 (100%)**	**0.74 (100%)**	**0.38 (100%)**	**0.55 (100%)**	**0.26 (100%)**	**0.33 (100%)**	**0.18 (100%)**	**0.19 (100%)**
3	20	0.94 (72%)	1.02 (38%)	**0.68 (100%)**	**0.84 (92%)**	**0.50 (100%)**	**0.59 (100%)**	**0.36 (100%)**	**0.38 (100%)**
	50	0.81 (100%)	0.91 (94%)	**0.53 (100%)**	**0.71 (100%)**	**0.34 (100%)**	**0.45 (100%)**	**0.23 (100%)**	**0.26 (100%)**
4	20	1.06 (18%)	1.10 (8%)	0.85 (86%)	1.01 (50%)	**0.63 (94%)**	**0.76 (90%)**	**0.42 (100%)**	**0.46 (100%)**
	50	0.92 (98%)	0.99 (56%)	**0.63 (100%)**	**0.82 (96%)**	**0.40 (100%)**	**0.51 (98%)**	**0.27 (100%)**	**0.33 (100%)**
5	20	1.13 (6%)	1.15 (2%)	1.03 (36%)	1.18 (6%)	0.74 (92%)	0.94 (70%)	**0.48 (100%)**	**0.57 (96%)**
	50	1.00 (40%)	1.05 (8%)	0.72 (98%)	0.92 (84%)	0.54 (96%)	0.73 (84%)	**0.33 (100%)**	**0.41 (94%)**

TAB. 3.2 – Mean ratio of the length of the confidence interval at 90% and 95% between strategy 2 and 1. The values between brackets indicate the number of times the meta-analysis approach led to a lower value of the quantile. The values are indicated in bold when in at least 90% of times the meta-analysis improved the precision on the QTL location, in italic otherwise.

QTL experiments	Parents	Type of population	Population size	Traits	Reference
Barière	F838 × F286	RIL	242	SD	Barriere *et al.* (2005)
Bohn1	CML131 × CML67	F_2	215	HT	Bohn *et al.* (1996)
Bohn2	B73 × Mo17	F_2	226	DPS	Bohn *et al.* (2000)
Bouchez	F2 × MBS847	BC_3S_1	217	SD	Bouchez *et al.* (2002)
Cardinal	B73 × B52	RIL	200	DPS,HT	Cardinal *et al.* (2001)
Charcosset	F2 × F252	F_5	129	SD	Charcosset *et al.* (2000)
Chardon1	F7p × F2	$F_{2:3}$	150	DPS,SD	Chardon *et al.* (2005)
Chardon2	F7p × Gaspé	$F_{2:3}$	150	DPS,SD	Chardon *et al.* (2005)
Groh	CML131 × CML67	RIL	166	HT	Groh *et al.* (1998)
Lubberstedt	KW1265 × D146	$F_{2:3}$	380	HT	Lubberstedt *et al.* (1997)
Mechin	F2 × MBS847	F_5	100	SD,HT	Mechin *et al.* (2001)
Moreau	F2 × F252	F_3	300	SD	Moreau *et al.* (2004)
Blanc	DE,F283,F810,F9005	F_2	150 (per population)	SD,DPS	Blanc *et al.* (2003)
Pioneer	Unknown	$F_{4:5}$	976	HT	`http://www.maizegdb.org`
Poupard	F2 × MBS847	RIL	86	SD	Poupard *et al.* (2001)
Rebai	Unknown	$F_{2:3}$	1200	SD	Rebai *et al.* (1997)
Ribaut	Tropical	F_2	272	DPS,SD	Ribaut *et al.* (1996)
Vladutu	E20 × N28	F_2	88	DPS,HT,LN	Vladutu *et al.* (1999)

Tab. 3.3 – 18 QTL mapping experiments related to flowering time in maize.

K	Δ_K (w_K)			
	AIC	AIC_c	AIC3	BIC
1	1096.32(0)	1088.94(0)	1088.32(0)	1084.11(0)
2	497.22(0)	490.52(0)	491.22(0)	488.06(0)
3	139.15(0)	133.79(0)	135.15(0)	133.05(0)
4	34.73(0)	31.53(0)	32.73(0)	31.67(0)
5	**0.00(0.86)**	**0.00(0.99)**	**0.00(0.95)**	**0.00(0.97)**
6	4.00(0.12)	8.50(0.01)	6.00(0.05)	7.05(0.03)
7	8.00(0.02)	18.70(0)	12.00(0)	14.11(0)
8	12.00(0)	31.17(0)	18.00(0)	21.16(0)
9	16.00 (0)	46.75(0)	24.00(0)	28.21(0)
10	19.54(0)	66.33(0)	29.54(0)	34.81(0)
34	51.42(0)	43.92(0)	76.42(0)	89.58(0)

TAB. 3.4 – Model choice on chromosome 8 for flowering time in maize. All the criteria select the model with 5 QTL (in bold on the table). Except for model with 6 QTL (and 7 for AIC), the Δ_K values suggest that using another model to fit the data rather leads to a substantial loss in information. The w_K values between brackets represent the weights of evidence which can be interpreted as the probability that model K is in fact the best model for the data.

QTL	Position $(\tilde{X})$	Weight $(\tilde{\Pi})$	Mahalanobis distance to next QTL	95% CI
1	14.6	0.06	6.21	11.7
2	75.4	0.35	1.26	6.2
3	89.5	0.44	2.64	3.8
4	114.5	0.07	5.20	11.1
5	165.2	0.09	-	13.9

TAB. 3.5 – Parameter estimates of the model with 5 QTL on chromosome 8 for the application to flowering time in maize. The positions and the lengths of the CI are given in cM. The 95% CI is derived from the conditional variance estimate of the position estimates obtained by applying the SEM strategy MENG and RUBIN (1991). Following VLADUTU *et al.* (1999) QTL 2 and 3 corresponds to *vgt2* and *vgt1*, respectively.

Troisième partie

Etude de la structure génétique

Chapitre 4

Mining Population Structure using Principal Component Analysis Framework

Authors : Jean-Baptiste Veyrieras [a,1], Letizia Camus-Kulandaivelu[1], and Alain Charcosset[1]

[1] UMR, INRA UPS-XI INAPG CNRS Génétique Végétale, Ferme du Moulon, 91190 Gif-sur-Yvette, France

Keywords : multinomial, binary data, mixture, admixture, PCA, DPCA.

To be submitted to *Bioinformatics*

[a]to whom correspondence should be addressed

Abstract

In this article we describe a new process to investigate population structure using multilocus genotype data. Based on recent theoretical development of discrete principal component analysis, we present a EM-algorithm procedure in order to "extract" from marker data population-components on which individuals are probabilistically assigned. Then we propose a simple strategy to select the minimal number of population-components which "captures" most of the linkage disequilibrium due to population structure. Using simulated data, we show that our approach yields relatively good performances both to determine the "actual" number of underlying populations and to accurately estimate population-components. Results from the analysis of a data set of 153 maize inbred lines genotyped at 55 SRR marker loci illustrate the

ability of our approach to track down structuration and to reveal hidden evolutionary events.

Introduction

Studying population structure has become commonplace in genetic data analysis. Interest ranges from population genetics to more applied issues : (i) learning about the evolutionary relationships of modern populations Cavalli-Sforza *et al.* (1994), (ii) studying the inheritance of complex genetic diseases by gene mapping in structured populations Stephens *et al.* (1994); McKeigue (1998) (iii) controlling of confounding effect due to population stratification in association mapping Ewens and Spielman (1995); Pritchard *et al.* (2000b); Hoggart *et al.* (2003).

Usually two kinds of population structure are distinguished. First the population under study may result from sampling individuals in different populations, each with its own characteristic set of allele frequencies. This is the mixture case. Second, the individuals may be sampled from a single population which has experienced "admixture" events. The term of admixture covers a large variety of scenarios of past introduction of individuals from one or several genetically distinct population(s) into another. This phenomenon, quite frequent in human populations, seems to "occur widely and in many species" Chikhi *et al.* (2001). In this case, stratification appears when individual admixture proportions, i.e the proportion of genome inherited from each population which have contributed to the present one, vary between individuals.

The advent of molecular markers since the last two decades has tremendously contributed to facilitate population structure analysis, by allowing researchers to characterize large collections of individuals using neutral marker loci. Markers are said to be neutral when they are not associated with traits subject to selection or adaptation. Using neutral marker data, one would like to be able to not only detect the presence of population structure but also identify underlying populations to which individuals could be assigned. Up to now, the statistical toolbox for population structure analysis has been enriched by several approaches : first, introduced by Menozzi *et al.* (1978), the use of principal component analysis (PCA) offers a well-established statistical framework to both evaluate and visualize the nature of the correlations among individuals. Biplots of the first axes of PCA have been largely reported in the literature to illustrate genetic differentiation between populations. However, when PCA is based on covariance or correlation matrix between marker loci, direct interpretation becomes difficult since negative values appear in the component matrix so they cannot be interpreted as "typical population" in any usual sense. Another limit of PCA in this context is that clusters of individuals can only be identified by eye. Secondly, matrices of

pairwise distances between individuals can be derived from the marker data set. These matrices may then be explored using some convenient graphical representation such as tree (e.g by applying neighbor-joining clustering) or PCA-like analysis via principal coordinate analysis (PCoA), or multidimensional scaling (MDS). Conversely, classical nonhierarchical distance-based clustering methods have been rarely used in this context. Due to their simplicity and their appealing graphical representation distance-based methods have been widely used in population genetics (see for instance Mohammadi and Prasanna (2003)). However, hierarchical clustering methods are not adapted when one aims to study population admixture since they rely on a categorical modelling of the repartition of the individuals into clusters (for example, individuals which derive from recent admixture event may "move" from one cluster to another depending on the markers used).

Therefore several authors Pritchard *et al.* (2000a); Falush *et al.* (2003); Hoggart *et al.* (2003) have recently proposed new model-based clustering strategies in order to carry out finer statistical inference about population structure and take admixture into account. The aim of these approaches is to establish a soft clustering of the individuals by both identifying the underlying populations and providing a probabilistic classification of the individuals. As pointed out by Buntine and Jakulin (2004, 2005), ignoring notations, they are equivalent to discrete principal component analysis (DPCA) - a good introduction to DPCA can be found in Buntine (2002). DPCA attempts to decorrelate discrete multivariate data set by finding independent compositional components (therefore DPCA can also be interpreted as a particular version of independent component analysis (ICA) Buntine and Jakulin (2004) applied to discrete multivariate data set).

In this article, we present a new procedure based on both PCA and DPCA in order to study population structure in a sample. First, we recall how to proceed with PCA to explore correlation pattern due to structuration and how it can be used to test whether there is structuration in the sample. Then by restricting DPCA to binary case we present a EM-algorithm Dempster *et al.* (1977) strategy to maximize the likelihood of the marker data with either a mixture or an admixture model. Contrary to Pritchard *et al.* (2000a); Hoggart *et al.* (2003) which used Bayesian approaches, we give analytical formula to iteratively estimate the parameters of both mixture and admixture DPCA models. We devote particular attention to the problem of choosing the number of components, i.e the number of populations, by providing a simple parsimonious criterion in order to select the minimal number of populations which "capture" most of the correlation due to structuration. We then evaluate the performance of our method using simulated data sets. Finally we apply it on a collection of 153 maize inbred lines which should reflect past admixture events in maize populations mainly due to the process of domestication and further historical events.

Notation and Background

Suppose we genotype N diploid individuals at L loci. We assume that haplotypes can be resolved for the N individuals. We discuss in appendix A how to proceed when genotype phases are unknown. We note J_l the number of distinct alleles observed at locus $l = 1, \ldots, L$. Let's $\mathbf{x}_i$ denote the i^{th} haplotype, by the vector $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iL})$ where $\mathbf{x}_{il} = (x_{il1}, \ldots, x_{il(J_l-1)})$ and $x_{ilj} = 1$ if and only if the haplotype i has allele j at locus l, 0 otherwise. Let the matrix $\mathbf{X}$ denote the observed haplotypes such as the i^{th} row of the matrix being the transpose of $\mathbf{x}_i$. We note $M = \sum_{l=1}^{L}(J_l - 1)$ the number of columns of $\mathbf{X}$. Note that $\mathbf{X}$ is generally called the allelic state matrix and this matrix is sparse and discrete. Let's introduce the vector $\hat{\mathbf{p}}$ of the mean estimates of the columns of $\mathbf{X}$, which is also the maximum-likelihood estimator of the allele frequencies. Finally, define $\mathbf{R}$ the matrix obtained by adjusting $\mathbf{X}$ as follows,

$$\mathbf{R} = \frac{1}{\sqrt{2N}}(\mathbf{X} - \mathbb{1}.^{\mathrm{T}}\hat{\mathbf{p}})$$

where $\mathbb{1}$ is the $\mathbf{R}^{2N}$ vector which all elements are equal to 1. It comes immediately that

$$^{\mathrm{T}}\mathbf{R}_{lj}\mathbf{R}_{l'j'} = \hat{p}_{ljl'j'} - \hat{p}_{lj}\hat{p}_{l'j'}$$

where $\mathbf{R}_{jl}$ is the column of matrix $\mathbf{R}$ corresponding to locus l and allele $j \in [1, \ldots, J_l - 1]$ and $\hat{p}_{ljl'j'}$ is the frequency estimate of the joint occurrence of alleles j and $j^{'}$ at loci l and $l^{'}$. Thus $^{\mathrm{T}}\mathbf{R}_{jl}\mathbf{R}_{j'l'}$ is the linkage disequilibrium (LD) estimate between allele j of locus l and allele $j^{'}$ of locus $l^{'}$.

Suppose we have sampled individuals from a single isolated population in both linkage equilibrium (LE) and Hardy-Weinberg equilibrium (HWE). Then it comes that $\mathsf{E}[^{\mathrm{T}}\mathbf{R}_{lj}\mathbf{R}_{l'j'}] = 0$, i.e the expected value of the LD between loci over all samples of size N is zero. And it follows that,

$$\mathsf{E}[^{\mathrm{T}}\mathbf{R}\mathbf{R}] = \mathbf{\Sigma}$$

where $\Sigma = (\Sigma_1, \ldots, \Sigma_L)$ is a bloc diagonal matrix where $\Sigma_l(j, j) = p_{lj}(1-p_{lj})$ and $\Sigma_l(j, j^{'}) = -p_{lj}p_{lj'}$ for $j \in [1, \ldots, J_l - 1]$ and $j^{'} \neq j$ where p_{lj} is the frequency of allele j at locus l in the population. In order to illustrate how structuration leads to create "artefactual" LD between physically independent loci, let's consider the two following scenarios.

First let's assume that the sampled population is in fact stratified in two distinct populations. In each underlying population we still assume that the marker loci segregate independently from each other (LE) and that the haplotypes are also independent (HWE). In this case, Nei and Li (1973) have shown that,

$$\mathsf{E}[^{\mathrm{T}}\mathbf{R}_{lj}\mathbf{R}_{l'j'}] = q(1-q)\delta_{lk}\delta_{l'j'}$$

where q is the contribution of the first population to the whole population and δ_{lj} is the difference of allele frequencies between the two populations, i.e $\delta_{lj} = p_{1lj} - p_{2lj}$ where p_{klj} is the frequency of allele j at locus l in population $k = 1, 2$. It follows that

$$\mathsf{E}[{}^{\mathrm{T}}\mathbf{RR}] = \mathbf{\Sigma} + \mathbf{\Delta}(q, \mathbf{P})$$

where $\mathbf{P} = (\mathbf{p}_1, \mathbf{p}_2)$ and $\mathbf{\Delta}(q, \mathbf{P})$ is obtained by computing all the pairwise products of locus allele frequency differences $\delta_{lj}\delta_{l'j'}$ multiplied by $q(1-q)$. Note that here $p_{lj} = qp_{1jl} + (1-q)p_{2jl}$.

Secondly, let's now consider a hybrid isolated population (as depicted in Figure 4.1). Suppose we sample individuals from this population at a given generation. Then, assuming that the hybrid isolated population evolves as an ideal Wright-Fisher population (with no mutation), it can be demonstrated that :

$$\mathsf{E}[{}^{\mathrm{T}}\mathbf{R}_{lj}\mathbf{R}_{l'j'}] = q(1-q)\delta_{lk}\delta_{l'j'}(1-c)^t[1 - \prod_{g=1}^{t}\frac{1}{2N(g)}]$$

where t is the number of generations from the mixture event, c the recombination fraction between loci l and $l^{'}$ and $N(g)$ the population size at generation $g = 0, \ldots, t$. If $t = 0$ we get the formula obtained for a mixture of two populations. Since loci are assumed to be independent ($c = 1/2$) and provided that the size of the population is quite large over time, we get $\mathsf{E}[{}^{\mathrm{T}}\mathbf{R}_{lj}\mathbf{R}_{l'j'}] \approx (1/2)^t q(1-q)\delta_{lk}\delta_{l'j'}$. Then it comes that

$$\mathsf{E}[{}^{\mathrm{T}}\mathbf{RR}] \approx \mathbf{\Sigma} + \mathbf{\Delta}(t, q, \mathbf{P}) \tag{4.1}$$

This formula can be easily extended to more than 2 "ancestral" populations. Thus it comes out that the expected covariance matrix is composed by two terms : the first one, $\mathbf{\Sigma}$, can be interpreted as a sampling variance component. The second one, $\mathbf{\Delta}(t, q, \mathbf{P})$, reflects the magnitude of covariance between loci due to the past event of mixture. The hybrid isolation model is obviously an idealization but it provides a simple way to understand how the information about population structuration can be tracked down from the marker data set by means of PCA and DPCA.

First Step : PCA

Singular Value Decomposition

There is a direct relation between PCA and singular value decomposition (SVD) in the case where principal components are computed from the covariance matrix. Then applying PCA on ${}^{\mathrm{T}}\mathbf{RR}$ is equivalent to find the SVD of $\mathbf{R}$:

$$\mathbf{R} = \mathbf{US}^{\mathrm{T}}\mathbf{V}$$

where $\mathbf{U}$ is an $2N \times M$ matrix (assuming without loss of generality that $2N > M$), $\mathbf{S}$ is a $M \times M$ diagonal matrix and ${}^{\mathrm{T}}\mathbf{V}$ is a $M \times M$ matrix. The matrix $\mathbf{US}$ contains the principal component scores, which are coordinates of the individuals in the space of the principal components, ${}^{\mathrm{T}}\mathbf{V}$ yields the principal components and $\mathbf{S}^2$ the eigenvalues of ${}^{\mathrm{T}}\mathbf{RR}$, from which the variance explained by each component can be derived. If the sampled population is not structured, then the biplot of the first columns of $\mathbf{US}$ should not show structuration of the cloud of points (a point here stands for a haplotype). Otherwise the cloud shape should reflect the degree of structuration contained in the marker data set.

Indicator of structuration

From the SVD of $\mathbf{R}$ we can assess its spectral norm, namely λ, by taking the highest singular value,

$$\lambda = ||\mathbf{R}||_2 = \max_{m=1,\ldots,M}(s_m)$$

If we assume that the sampled population is not structured, the expected value of λ over all samples of size N is given by,

$$\lambda_0 = \mathsf{E}[\lambda] = \mathsf{E}[||\mathbf{R}||_2]$$

The choice of λ_0 comes from the fact that the hypothesis of no structuration can be viewed as the null hypothesis. In order to evaluate λ_0, we propose to carry out a parametric boostrap under the null hypothesis (i.e a single population in LE and HWE). This provides a simple way to compare the observed value λ to its empirical distribution under the null hypothesis.

Second step : DPCA

Although PCA offers a well-established Gaussian framework to track down correlation in multivariate data, when it is applied to sparse and discrete matrices, interpretation of components cannot be made in term of "typical" populations. The idea of DPCA is to avoid Gaussian modelling of the data by providing a more adapted probabilistic context. In particular DPCA attempts to model correlation pattern between discrete factors by assuming an underlying mixture model. Here we propose a simplified version of DPCA restricted to binary data.

Mixture model

In the mixture case, the haplotypes are assumed to be drawn from a mixture of independent multinomial, each with its own characteristic set of

frequencies :

$$\mathbf{x}_i \sim \sum_{k=1}^{K} q_k \mathcal{M}(\mathbf{P}_k)$$

where $\mathbf{P} = (\mathbf{P}_1, \ldots, \mathbf{P}_K)$ is the matrix of allele frequencies in the K populations and $\mathbf{q} = (q_1, \ldots, q_K)$ are the mixing proportions, i.e the contribution of each population to the sampled population. Then the loglikelihood of the marker data set can be written as follows,

$$L(\mathbf{X}; \mathbf{q}, \mathbf{P}) \propto \sum_{i=1}^{2N} \log \left(\sum_{i=1}^{K} q_k \mathsf{Pr}(\mathbf{x}_i | \mathbf{P}_k) \right) \tag{4.2}$$

where $\mathsf{Pr}(\mathbf{x}_i|\mathbf{P}_k) = \prod_{l=1}^{L} p_{kl\zeta_{il}}$ is the probability of the haplotype in population k and ζ_{il} indicates the index of the allele at locus l for haplotype i. Maximizing 4.2 can be achieved via a standard EM-algorithm by applying the following E and M steps :

- **E-step** : let's denote $\mathbf{Z} = (z_1, \ldots, z_{2N})$ the vector of unknown (or hidden) populations of origin of the haplotypes. The expression of the distribution of $\mathbf{Z}$ is obtained by applying a simple Bayes' rule,

  $$\mathsf{Pr}(z_i = k | \mathbf{x}_i, \mathbf{q}, \mathbf{P}) = \frac{q_k \mathsf{Pr}(\mathbf{x}_i|\mathbf{P}_k)}{\sum_{k'=1}^{K} q'_k \mathsf{Pr}(\mathbf{x}_i|\mathbf{P}_{k'})}$$

  We denote $\mathbf{\Pi}$ the matrix of the posterior probabilities of origin of the haplotypes, such that $\mathbf{\Pi}(i,k) = \mathsf{Pr}(z_i = k|\mathbf{x}_i, q, \mathbf{P})$.
- **M-step** : $\mathbf{q}$ and $\mathbf{P}$ can be updated together as follows,

  $$\mathbf{P} = {}^{\mathrm{T}}\mathbf{X}.\mathbf{\Pi}.({}^{\mathrm{T}}\mathbb{1}.\mathbf{\Pi}.\mathbf{I})^{-1}$$
  $$\mathbf{q} = {}^{\mathrm{T}}\mathbb{1}.\mathbf{\Pi}(2N)^{-1}$$

  where $\mathbf{I}$ is the identity matrix in $\mathbf{R}^K$.

Admixture model

As for the mixture case, the haplotypes are still assumed to be drawn from a mixture of independent multinomial

$$\mathbf{x}_i \sim \sum_{k=1}^{K} q_{ik} \mathcal{M}(\mathbf{P}_k)$$

but each haplotype $\mathbf{x}_i$ has now its own mixing proportions defined by the vector $\mathbf{q}_i = (q_{i1}, \ldots, q_{ik})$. This means that each haplotype has potentially experienced an original history through generations. Then the loglikelihood of the admixture model is given by

$$L(\mathbf{X}; \mathbf{Q}, \mathbf{P}) \propto \sum_{i=1}^{2N} \sum_{l=1}^{L} \log \left(\sum_{i=1}^{K} q_{ik} \mathsf{Pr}(\mathbf{x}_{il}|\mathbf{P}_k) \right) \tag{4.3}$$

where $\mathsf{Pr}(\mathbf{x}_{il}|\mathbf{P}_k) = p_{kl\zeta_{il}}$. Here we modify $\mathbf{Z}$ such that vector $\mathbf{z}_i = (z_{i1}, \ldots, z_{iL})$, i.e the transpose of the i^{th} row of $\mathbf{Z}$, is the hidden ancestral path through haplotype i, and we denote $\mathbf{Q}$ the matrix the i^{th} row of which is given by the transpose of the vector $\mathbf{q}_i$. We propose to estimate the parameters $\mathbf{P}$ and $\mathbf{Q}$ of the admixture model by maximizing 4.3 via the following EM-algorithm :

– **E-step** : applying a simple Baye's rule yields :

$$\mathsf{Pr}(z_{il} = k|\mathbf{x}_i, \mathbf{q}_i, \mathbf{P}) = \frac{q_{ik}\mathsf{Pr}(\mathbf{x}_{il}|\mathbf{P}_k)}{\sum_{k'=1}^{K} q_{ik'}\mathsf{Pr}(\mathbf{x}_{il}|\mathbf{P}_k)}$$

Let's denote $\mathbf{\Pi}_l$ the $2N \times K$ matrix which contains the posterior probabilities of origin of locus l.

– **M-step** : $\mathbf{Q}$ and $\mathbf{P}$ can be updated together as follows,

$$\mathbf{p}_l = {}^{\mathrm{T}}\mathbf{X}_l.\mathbf{\Pi}_l.({}^{\mathrm{T}}\mathbb{1}.\mathbf{\Pi}_l.\mathbf{I})^{-1} \; l = 1, \ldots, L$$

$$\mathbf{Q} = \left(\sum_{l=1}^{L} \mathbf{\Pi}_l\right)(L)^{-1}$$

where $\mathbf{X}_l$ is the sub matrix of $\mathbf{X}$ formed by the columns related to locus l and $\mathbf{p}_l$ the sub matrix of $\mathbf{P}$ obtained by getting the rows related to locus l.

Interpretation

To use a homogeneous notation between the two models we denote $\hat{\mathbf{W}}$ the $2N \times K$ matrix obtained at the last iteration of the EM-algorithm such that in the mixture case this matrix $\hat{\mathbf{W}}$ is equal to $\hat{\mathbf{\Pi}}$ and in the admixture case to $\hat{\mathbf{Q}}$. In the mixture model this matrix gives the *a posteriori* membership probabilities of the haplotypes into the K populations, and in the admixture model $\hat{\mathbf{W}}$ contains the proportion of locus inherited from each of the K populations. Then whatever the model fitted to the marker data, this provides an approximation of the sparse and discrete data matrix $\mathbf{X}$ by a product of smaller matrices, $\hat{\mathbf{W}}$ and $\hat{\mathbf{P}}$:

$$\mathbf{X} \approx \hat{\mathbf{W}}.{}^{\mathrm{T}}\hat{\mathbf{P}}$$

$\hat{\mathbf{W}}$ can be interpreted as the *score* matrix while the matrix $\hat{\mathbf{P}}$ which contains the allele frequency estimates of the populations can be interpreted as the *principal component* matrix.

Choosing the number of components

Consider now the error of the approximation by taking the scaled difference :

$$\mathbf{R}(K) = \frac{1}{\sqrt{2N}}(\mathbf{X} - \hat{\mathbf{W}}.{}^{\mathrm{T}}\hat{\mathbf{P}})$$

It comes immediately that for $K = 1$, $\mathbf{R}(1) = \mathbf{R}$. In this case both mixture and admixture model lead naturally to $\hat{\mathbf{P}} = \hat{\mathbf{p}}$ and $\hat{\mathbf{W}} = \mathbb{1}$. The element of $\sqrt{2N}\mathbf{R}(K)$ corresponding to haplotype i and allele j at locus l is given by $x_{ilj} - \hat{\omega}_{ilj}$ where $\hat{\omega}_{ilj} = \sum_{k=1}^{K} \hat{w}_{ik}\hat{p}_{klj}$ is the predicted frequency of allele j at locus l for haplotype i. The matrix $\mathbf{R}(K)$ can thus be interpreted as the error matrix of prediction of the model. In other words, this reflects the ability of the model to correct for structuration of the sample by centering each matrix element relatively to its expected frequency.

Let's assume that $\mathbf{X}$ comes from a (ad)mixture of independent multinomial with K components. Here we note $\mathbf{Z}$ the $2N \times L$ matrix of the "actual" hidden origins (among K) of the loci for each haplotype. Note that this matrix is a random matrix and that the join density of a complete data set $(\mathbf{X}, \mathbf{Z})$ is given by :

$$\mathsf{Pr}(\mathbf{X}, \mathbf{Z}|\mathbf{P}, \mathbf{Q}) = \mathsf{Pr}(\mathbf{X}|\mathbf{Z}, \mathbf{P}).\mathsf{Pr}(\mathbf{Z}|\mathbf{Q})$$

Under this assumption, the elements of the observed data matrix $\mathbf{X}$ have the following properties :

$$\mathsf{E}_{\mathbf{X}|\mathbf{Z}}[x_{ilj}] = p_{z_{il}lj}$$

$$\mathsf{E}[x_{ilj}] = \mathsf{E}_{\mathbf{Z}}[\mathsf{E}_{\mathbf{X}|\mathbf{Z}}(x_{ilj})] = \sum_{i=1}^{K} q_{ik}.p_{klj}$$

where $\mathsf{E}_{\mathbf{X}|\mathbf{Z}}[.]$ stands for expectation over all possible samples $\mathbf{X}$ conditionally to $\mathbf{Z}$, and $\mathsf{E}_{\mathbf{Z}}[.]$ for expectation over all possible locus origins. The last relation insures that $\mathsf{E}[\mathbf{R}(K)] = 0$.

Suppose a model with K^* components is fitted to the data. Let's introduce the following quantities,

$$\epsilon_i(lj) = \mathsf{E}_{\mathbf{X}|\mathbf{Z}}[\hat{w}_{ilj}] - p_{z_{il}lj}$$

$$\epsilon_i(lj, l^{'}j^{'}) = \mathsf{E}_{\mathbf{X}|\mathbf{Z}}[\hat{w}_{ilj}.\hat{w}_{il^{'}j^{'}}] - p_{z_{il}lj}.p_{z_{il^{'}}l^{'}j^{'}}$$

where $\epsilon_i(lj)$ measure the "individual bias" of the model predicted frequency of occurrence of allele j at locus l and $\epsilon_i(lj, l^{'}j^{'})$ the "individual bias" of the model predicted frequency of co-occurrence of alleles j and $j^{'}$ at locus l and $l^{'}$. After some calculations, it can be shown that :

$$\mathsf{E}_{\mathbf{X}|\mathbf{Z}}\left[{}^{\mathrm{T}}\mathbf{R}_{jl}(K^*)\mathbf{R}_{j^{'}l^{'}}(K^*)\right] = \epsilon(lj, l^{'}j^{'}) - \epsilon(lj) - \epsilon(l^{'}j^{'})$$

where

$$\epsilon(lj, l^{'}j^{'}) = \frac{1}{2N}\sum_{i=1}^{2N} \epsilon_i(lj, l^{'}j^{'})$$

$$\epsilon(lj) = \frac{1}{2N}\sum_{i=1}^{2N} \epsilon_i(jl).p_{z_{il^{'}}l^{'}j^{'}}$$

By taking now expectation over $\mathbf{Z}$ we get $\mathsf{E}_{\mathbf{Z}}[\epsilon(lj)] = 0$ and it follows that :

$$\begin{aligned}\mathsf{E}\left[{}^{\mathrm{T}}\mathbf{R}_{jl}(K^*)\mathbf{R}_{j'l'}(K^*)\right] &= \mathsf{E}_{\mathbf{Z}}\left[\mathsf{E}_{\mathbf{X}|\mathbf{Z}}\left[{}^{\mathrm{T}}\mathbf{R}_{jl}(K^*)\mathbf{R}_{j'l'}(K^*)\right]\right] \\ &= \mathsf{E}_{\mathbf{Z}}[\epsilon(lj, l^{'}j^{'})] - \mathsf{E}_{\mathbf{Z}}[\epsilon(lj)] - \mathsf{E}_{\mathbf{Z}}[\epsilon(l^{'}j^{'})] \\ &= \mathsf{E}_{\mathbf{Z}}[\epsilon(lj, l^{'}j^{'})]\end{aligned}$$

For example, suppose $K^* = 1$ and that the "true" model is a simple mixture model with $K = 2$. It comes immediately that

$$\begin{aligned}\mathsf{E}_{\mathbf{Z}}[\epsilon(lj, l^{'}j^{'})] &= q(p_{lj}p_{l'j'} - p_{1lj}p_{1l'j'}) + (1-q)(p_{lj}p_{l'j'} - p_{2lj}p_{2l'j'}) \\ &= q(1-q)\delta_{lj}\delta_{l'j'}\end{aligned}$$

which is, as shown previously and recalling that $\mathbf{R}(1) = \mathbf{R}$, the expected linkage disequilibrium Nei and Li (1973).

Finally we have the following relation :

$$\mathsf{E}\left[{}^{\mathrm{T}}\mathbf{R}(K^*)\mathbf{R}(K^*)\right] = \mathbf{\Sigma} + \mathbf{E}(K^*)$$

where $\mathbf{E}(K^*)$ is the covariance component accounting for the error of prediction of the model obtained from the $\mathsf{E}_{\mathbf{Z}}[\epsilon(lj, l^{'}j^{'})]$ values. As for $\mathbf{R}(1) = \mathbf{R}$, we can assess the spectral norm of $\mathbf{R}(K^*)$, denoted $\lambda(K^*)$, by taking the highest singular value of ${}^{\mathrm{T}}\mathbf{R}(K^*)\mathbf{R}(K^*)$

$$\lambda(K^*) = ||\mathbf{R}(K^*)||_2 = \max_{m=1,\ldots,M}\{s_m(K^*)\}$$

where the $s_m(K^*)$'s are the singular values obtained by applying SVD on $\mathbf{R}(K^*)$. Here, $\lambda(K^*)$ can be viewed as a measure of goodness-of-fit of the model since it reflects the magnitude of the covariance term $\mathbf{E}(K)$. In other words, $\lambda(K)$ measures the residual LD in the marker data set when correcting for structuration assuming K^* subpopulations. If $K^* = K$ it comes that $\mathbf{E}(K^*) = 0$, $\mathsf{E}[\lambda(K^*)] = \lambda_0$, and for $K^* \leq K$ we get the following inequality :

$$\lambda_0 \leq \mathsf{E}[\lambda(K^*)] \leq \mathsf{E}[\lambda(1)]$$

Therefore, let's define the following ratio,

$$\Gamma(K^*) = \frac{\lambda(1) - \lambda(K^*)}{\lambda(1) - \lambda_0}$$

where λ_0 can be substituted by its estimate $\tilde{\lambda}_0$. Thus $\Gamma(K^*)$ is related to the proportion of LD explained by the model with K^* populations. Coupling $\lambda(K^*)$ and $\Gamma(K^*)$ provides a simple and readable criterion to select the value K^* which best explains the covariance component due to population stratification. The decision of when to stop "extracting" populations basically

depends on when there is only very little covariance left. This offers a straightforward parallel with PCA in which the number of components is usually selected by keeping only the first axes which account for a given proportion (typically 90 or 95%) of the variance. Scree test strategy Cattell (1966) can also be used : this consists in plotting the $\lambda(K)$'s values and then choosing the value of K where the smooth decrease of the $\lambda(K)$'s values appears to level off to the right of the plot.

Implementation

Simulations

To evaluate the performance of our approach, a hybrid isolated population (Figure 4.1) was simulated : two distinct populations of equal size N=5000 were fused to produce a single random-mating population in which 100 bi-allelic marker loci were assumed to segregate independently. The allele frequencies of the first "ancestral" population were randomly drawn from an uniform distribution between 0 and 1. For the other population, the allele frequencies were generated so that the allelic difference between the two populations was in average 0.5. In Figure 4.2 we have represented the evolution of the proportions of "ancestry" over generations for this scenario. Our analysis consisted in two phases : first we looked at the behavior of the structure indicator obtained by PCA for different generations and then we studied the performance of DPCA to both choose the right number of populations and to classify individuals.

First, in Figure 4.3 we depict the results obtained by applying PCA on several samples randomly drawn at different generations with $N = 100$ individuals and $L = 50$ marker loci. This clearly shows that when the mixture event is not too far in time, PCA makes it possible to clearly visualize the structuration via standard biplot of the first axes of PCA. However for $t > 2$ the cloud of points is difficult to analyze and no obvious clusters of individuals can be defined by eye. As expected, the summary statistics λ (Figure 4.3 B) reflects the magnitude of the covariance component due to structuration. More precisely, we can see that the observed values of λ are largely outside the 95% probability support of λ_0 for $t < 5$, and converge asymptotically to λ_0 when t increases. This illustrates the usefulness of this indicator to address the question :"Is the sample structured ?" or "Does there remain significant LD to track down population structure via DPCA ?".

Secondly, in Figure 4.4 we plot the behavior of $\lambda(K)$ when fitting either admixture or mixture model to data sets obtained by randomly sampling $N = 100$ individuals from the hybrid-isolated population at different generations and for four different numbers of marker loci $L = 10, 25, 50$ and 100. At $t = 0$ we can see that both admixture and mixture models lead to a good modelling of the covariance component for $K = 2$ populations in all cases. At $t = 2$

the admixture model always lead to choose $K = 2$ populations. Conversely, the mixture model tends to overestimate the number of populations when L increases. We know that at $t = 0$ we have a simple mixture of genotypes meanwhile at $t = 2$ the distribution of the proportion of ancestry is multimodal (see Figure 4.2) due to admixture. The model choice strategy based on the scree plot of the values of $\lambda(K)$ not only provides a simple way to select the optimal value of K for a given model (mixture or admixture) but also makes it possible to compare the results obtained by fitting the two models on the data set (this may be usefull since sometimes no information is available about the underlying evolutionary scenario). Hence, a parsimonious rule leads to select $K = 2$ populations and a mixture model at $t = 0$, and an admixture model at $t = 2$.

The case $t = 5$ appears to be a much harder problem than for the previous generations. In this case for $L = 10$ the difference between $\lambda(1)$ and λ_0 is not really significant and then our strategy fails to detect structuration. This indicates a lack of power of the method for such a limited number of loci. For higher number of loci the admixture model with $K = 2$ appears generally to be the best one, although for some samples our criterion may lead to select $K = 3$ populations.

Having shown that our model choice strategy performs relatively well, we now examine the performance of DPCA to cluster individuals in their appropriate populations. In the case of simple mixture of haplotypes ($t = 0$ and $t = 1$) the clustering performs very well even with a small number of loci ($L = 10, 25$), and perfectly for $L = 50$ and $L = 100$ (Figure 4.5). It is worth noting that when L increases the EM-algorithm convergence is achieved in a very few number of iterations (less than 10). In the admixture case, the scatter plots of the inferred ancestral proportions against the true ancestral proportions of individuals (Figure 4.6) show that these proportions are hard to estimate with a small number of loci. When the number of loci increases the EM-algorithm leads to very close estimates of the individual admixture proportions even for samples drawn at $t = 5$, in which there remains only a small fraction of LD and no longer "pure" individuals (as illustrated in Figure 4.2 and Figure 4.3). Comparison between actual allele frequencies in the two ancestral populations and the inferred ones also indicates that the EM-algorithm yield relatively good performances even when almost of the individuals are admixed ($t \geq 2$). Finally we have simulated supplemental hybrid isolated populations using different allelic contrast between the two ancestral populations (in particular lower than 0.5). Results of analyzes of samples drawn in these populations have confirmed the ability of the EM-algorithm to correctly infer the population structure (data not shown).

Application

We now apply our method on a maize data set which consists in 153 inbreds directly obtained from traditional landraces. This inbred line collection represents the ancestral inbred gene pool used for modern selection in temperate regions and was used by Camus-Kulandaivelu *et al.* (2006) to explore the diversity and the genetic structure of maize. Each inbred was characterized by 55 SSR loci leading to 331 distinct alleles. Hence here the data matrix $\mathbf{X}$ is a 153×276 rectangular matrix. By applying PCA on the adjusted matrix $\mathbf{R}$ we obtained $\lambda = 1.88$. This value of λ was compared with its empirical distribution under the null hypothesis (i.e a single population) obtained from 1000 random samples. This clearly suggests the presence of structuration in the data set (see Figure 4.7 for $K = 1$), which is confirmed by biplot of the two first axes of the PCA (see Figure 4.8 and Figure 4.9).

Since PCA reveals the presence of a covariance component which may be explained by hidden population structuration, we applied the EM-algorithm for both mixture and admixture models on the marker data set from $K = 2$ to $K = 5$. For each value of K we selected among 10 independent runs of the EM-algorithm the one with the best loglikelihood and computed the corresponding $\lambda(K)$ value. Each time, the EM-algorithm was run with a convergence tolerance of 10^{-8} and a maximal number of iterations of 2000. For a given value of K, the 10 repetitions generally yielded very close parameter estimates and loglikelihood values. The scree plot of the $\lambda(K)$'s values is depicted in Figure 4.7. We can see that under the admixture model, $K = 3$ populations are enough to explain the LD due to structuration (meanwhile for $K = 3$ the mixture model explains only 65.7%). In Figure 4.8 we represent the results of the admixture model with $K = 3$ populations together with PCA biplots and Neighbor-Joining (NJ) tree based on the distance matrix derived from $\mathbf{X}$. This admixture model with $K = 3$ is consistent with the hypotheses about the processes involved in maize domestication and later adaptation to temperate climate. Three main historical maize origins were expected for temperate maize in our panel : NorthernFlint (NF), Corn-Belt Dent (CBD) and European Flint (EF), which also included a limited number of tropical materials (TR) and popcorn (PC). It is known that part of EF material was derived from NF and Caribbean germplasms Rebourg *et al.* (2003) meanwhile part of DE material was derived from NF and non Caribbean Tropical germplasms Anderson and Brown (1952); Liu *et al.* (2003); Ho *et al.* (2005). Figure 4.8, along with this *a priori* classification, shows that the first PCA axis is strongly determined by the differentiation of the NF group, which is identified as one of the three populations by $K = 3$ model. This strong differentiation is consistent with results of Doebley *et al.* (1986). Considering the proportions of admixture for EF and CBD inbreds, shows that the two other groups revealed by the EM algorithm can be interpreted as the non NF contributors of EF

(Figure 4.8B), and the non NF contributors of CBD (Figure 4.8C). Although the model with $K = 3$ captures the main part of LD, it can be interesting to explore models with a higher number of populations, namely $K = 4$ and $K = 5$. In Figure 4.9 we depict the classification obtained for the successive values of K together with the *a priori* classification of the inbreds, adding Tropical (TR) and Popcorn (PC) origin. This illustrates that going on "extracting" populations leads to subdivide a previous population in order to reduce the intra population diversity. For $K = 4$, a new population including PC and TR materials is "extracted" from the population interpreted at $K = 3$ as the non NF progenitors of CBD. This new population is in turn subdivided for $K = 5$ according to the two origins. These two last groups seems to have had a very limited contribution to the genetic diversity of CBD and EF.

We have also compared our approach to STRUCTUREPRITCHARD *et al.* (2000a) (recall that STRUCTUREcan be viewed as a Bayesian implementation of DPCA). For each model (from $K = 2$ to $K = 5$) we explored a series of 10 independent runs of the Gibb's sampler with 10^5 iterations following a burn-in of 30000 iterations. The result for the best outputs are displayed in Table 4.1. For $K = 2$ and $K = 3$ the two approaches yielded very close results, as illustrated by similar $\lambda(2)$ and $\lambda(3)$ values. As for the EM-algorithm, the $\lambda(3)$ obtained from STRUCTUREoutput clearly indicates that $K = 3$ populations are enough to explain the LD due to structuration. On the other hand, the model choice criterion proposed by STRUCTURE, which is based on a *ad hoc* Bayesian deviance criterion (SPIEGELHALTER *et al.* (1998)), suggests to select model with $K = 4$ populations. Although several supplemental runs of the Gibb sampler have been done with different burn-in and sampling iteration values, it appeared that for $K \geq 4$, STRUCTUREgot stuck into a unique mode as if it was enable to "extract" a new population from the data set. This was removed when assuming correlation between population allele frequencies as done by CAMUS-KULANDAIVELU *et al.* (2006).

Discussion

In this article we have described a procedure to study population structure using multilocus genotype data. We pointed out that this approach, as well as the previous Bayesian implementation of PRITCHARD *et al.* (2000a), can be interpreted as a particular DPCA problem. We have also illustrated how it is usefull to couple PCA and DPCA results into a single data analysis framework to both select the optimal number of populations and to visualize results. As for PCA for which scores and components can be computed using an EM-algorithm strategy TIPPING and BISHOP (1998), we give analytical formula to iteratively estimate both "population-components" and individual "admixture-scores". Comparison with the Bayesian approach of PRITCHARD

et al. (2000a) reveals that the EM-algorithm yields quite similar results and that both methods provide an efficient way to infer population structure using multilocus genotype data (data not shown). It is worth noting that the convergence of the EM-algorithm occurs in a relative small number of iterations and it is generally much faster than Gibb's sampling strategies (this is based on comparisons between our implementation of the EM-algorithm and STRUCTURE). Rather to contrast our "Frequentist" point of view with previous "Bayesian" modelling, we think that both statistical frameworks have their own assets - balanced by their own intrinsic limits - and must be considered with care regarding the data analysis context in which they are applied. On the one hand, for example, the EM-algorithm is not able to deal with *a priori* information on the origin of the individuals although this information can be sometimes available as discussed by PRITCHARD *et al.* (2000a). On the other hand, it appears that the EM-algorithm seems to be more able to discriminate close ancestral populations as it brings out in the application on the maize inbreds for $K \geq 4$.

More recently, FALUSH *et al.* (2003) and HOGGART *et al.* (2003) have proposed a more sophisticated Bayesian approach to tackle admixture mapping by allowing for linkage among marker loci, provided that the map order of marker loci is known. When marker loci are linked our admixture model might attempt to explain the residual covariance component due to linkage by admixture, which may lead to spurious allele frequency and admixture proportion estimates. There is no strong limit to adapt the EM-algorithm for such situation but we do not deal with it here since it is a more complex problem and different than the one addressed by DPCA.

In this article we have paid particular attention to the problem of choosing the optimal number of populations, K, which is necessary to model the covariance component due to structuration. First, we propose a novel approach to test for the presence of population structure ($K = 1$ v.s $K > 1$) using a simple parametric bootstrap of the spectral norm of the covariance matrix. Second, our procedure provides a easy to understand rule to select the minimal number of populations to "extract" from the data set and can be applied whatever the algorithm machinery used to estimate (ad)mixture parameters (even when genotype phases are unknown as described in Appendix A). Simulations have shown that this strategy yields relatively good performances provided that the event of mixture or admixture is not too far in time and that the number of marker loci used to infer model parameters is sufficient. Thus we always recommend to compute the structure indicator λ before fitting either mixture or admixture model to the data set and to compare it to its empirical distribution obtained by parametric bootstrap. We stress that care should be taken when dealing with model choice and that our strategy may be understood for what it does, i.e trying to minimize the residual LD in the data set when correcting for structuration for a given value of K. Therefore K may not have always a "biological" meaning and

may depend on the sampling of both individuals and marker loci. However, when one aims to control for spurious LD in association studies, this strategy provides a parsimonious rule to choose the minimal number of covariates to include into the association study model in order to prevent from false positive results due to structuration (Pritchard *et al.* (2000b); Satten *et al.* (2001); Hoggart *et al.* (2003, 2004)). It is worth noting that in this case, it can also be viewed as a data reduction problem (similar to data reduction with PCA) where we attempt to capture the maximal information contained in the data set by a reduced number of independent population-components on which each individual can be probabilistically "scored". Thus this makes it possible to model the information on structuration brought by numerous marker loci with only a small number of variables (the individual "scores" or admixture), and so to preserve reasonable power for testing for association.

In summary, our method provides a homogeneous framework based on standard and advanced PCA methodology which can be extended to other kinds of sparse and discrete data matrix reduction problems, being aware that the underlying probabilistic model is relatively simple and obviously an idealization. Nevertheless we think that it is flexible enough to be applied in a wide range of fine clustering problems.

Appendix A

There are two possible alternatives for diploid data when the genotype phases are unknown. First, the haplotypes can be resolved using standard haplotype reconstruction strategy Niu (2004) and the matrix $\mathbf{X}$ can then be build by using the best haplotypic configuration obtained for each genotype. Otherwise, the data matrix $\mathbf{X}$ must be recoded as follows :

$$x_{ilj} = \begin{cases} 2 & \text{if genotype is } jj \\ 1 & \text{if genotype is } jj^{'} \text{ and } j^{'} \neq j \\ 0 & \text{if genotype is } j^{'}j^{''} \text{ and } j^{'}, j^{''} \neq j \end{cases}$$

where each line now stands for a genotype. It can easily be shown that, ${}^{\mathrm{T}}\mathbf{R}_{jl}\mathbf{R}_{j'l'} = 2\Delta_{ljlj'}$ where $\Delta_{ljl'j'}$ is the composite LD reported by Weir (1996). The lack of knowledge of the phases leads to an additional covariance component. Thus direct interpretation of PCA biplots may be tedious in this case. However the summary statistic λ can still be compared to its empirical distribution under the assumption that there is no structuration. Note that the EM-algorithm can be run on a arbitrary haplotypic representation of the genotypes and do not require that the phases are known to estimate the mixture or the admixture parameters (recall that the likelihood of any given genotype is simply proportional to the product over the phases and the loci of the probabilities of the observed alleles). Finally the $\lambda(K)$ values

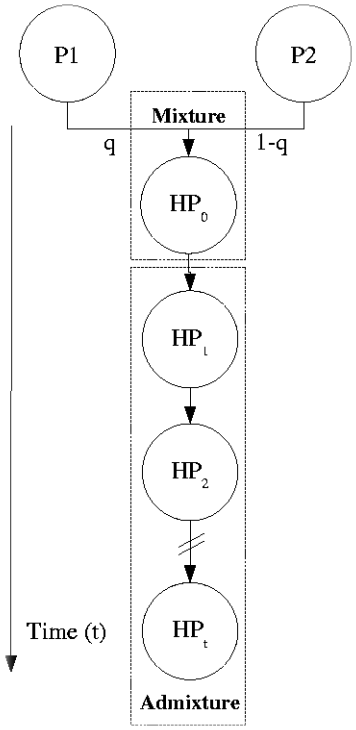


FIG. 4.1 – Hybrid Isolation model : a proportion q, respectively $1-q$, of individuals from population P1, respectively P2, emigrate to form a new isolated population. Thus HP_0 is a mixture of "pure" individuals from the two original populations P1 and P2. Going forward in time, $t \geq 1$, the recombination events "mixes" individual genotypes creating "mosaics" of loci with different origins (admixture). This simple scenario is one possible cause of admixture (figure adapted from LONG (1991)).

can also be obtained from the covariance matrix derived from $\mathbf{R}(K)$ which is now computed by adjusting the recoded data matrix $\mathbf{X}$ relatively to the predicted values of the number of alleles (0,1 or 2) at each locus for each genotype.

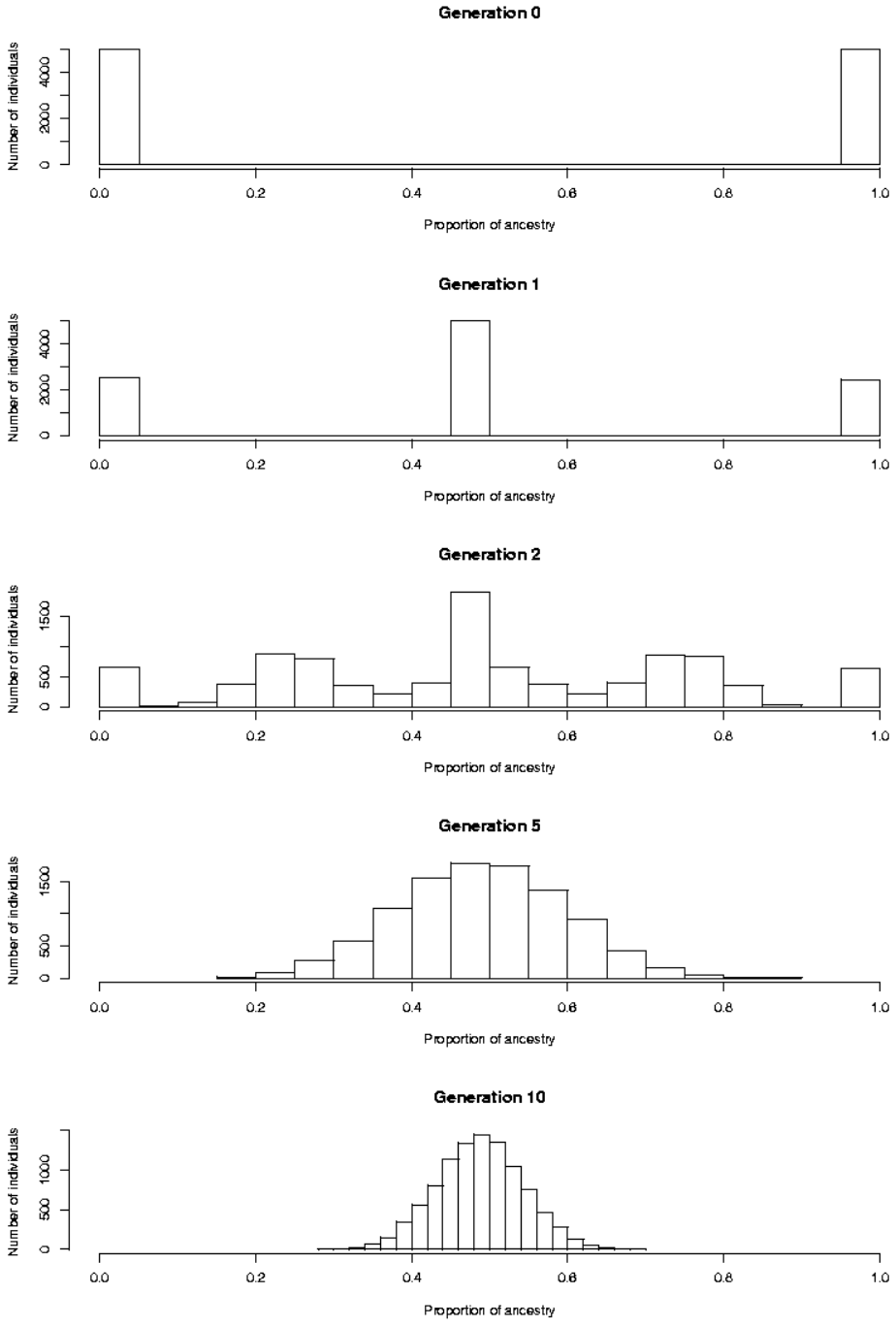


FIG. 4.2 – Evolution of the proportion of ancestry over five generations $t = 0, 1, 2, 5, 10$ in a hybrid-isolated population obtained by merging two populations of equal size N=5000.

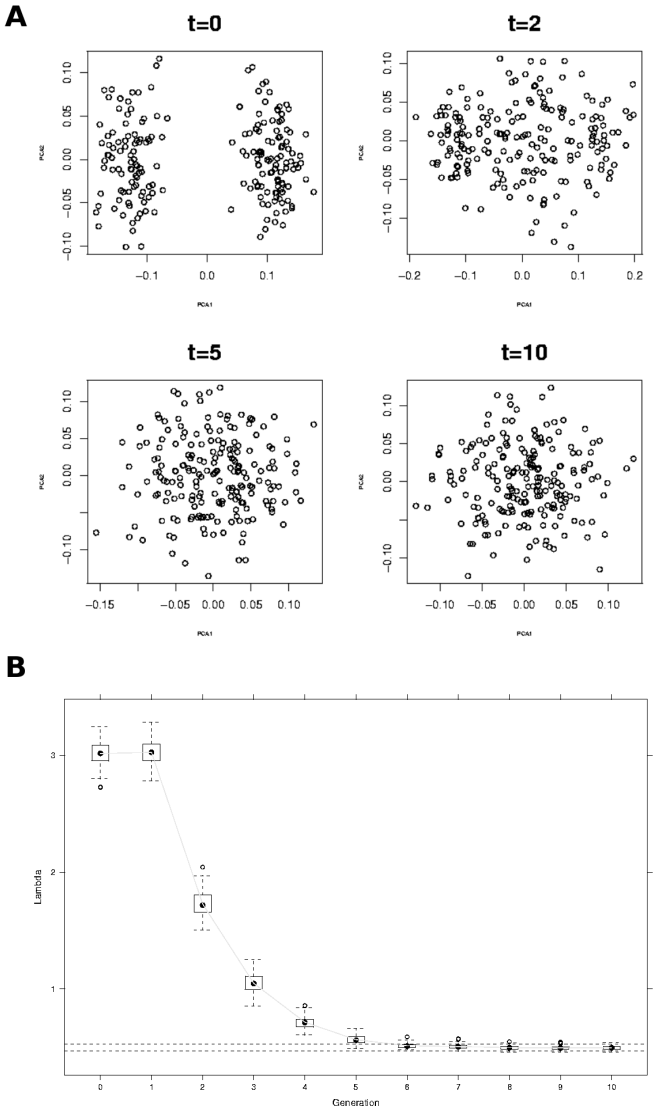


FIG. 4.3 – Illustration of the PCA pre-analysis of marker data set at different generations after the creation of the hybrid-isolated population. A : biplots of the two first axes of PCA for 4 samples of $N = 100$ individuals at generations $t = 0, 2, 5, 10$. B : whisker plot of the structure indicator λ at each generation. The dashed lines indicates the 95% probability support of λ_0 obtained by parametric bootstrap assuming a single population. For B, moments of the structure indicator were obtained by randomly drawing 100 samples from the hybrid-isolated population each of $N = 100$ individuals assuming $L = 50$ bi-allelic marker loci.

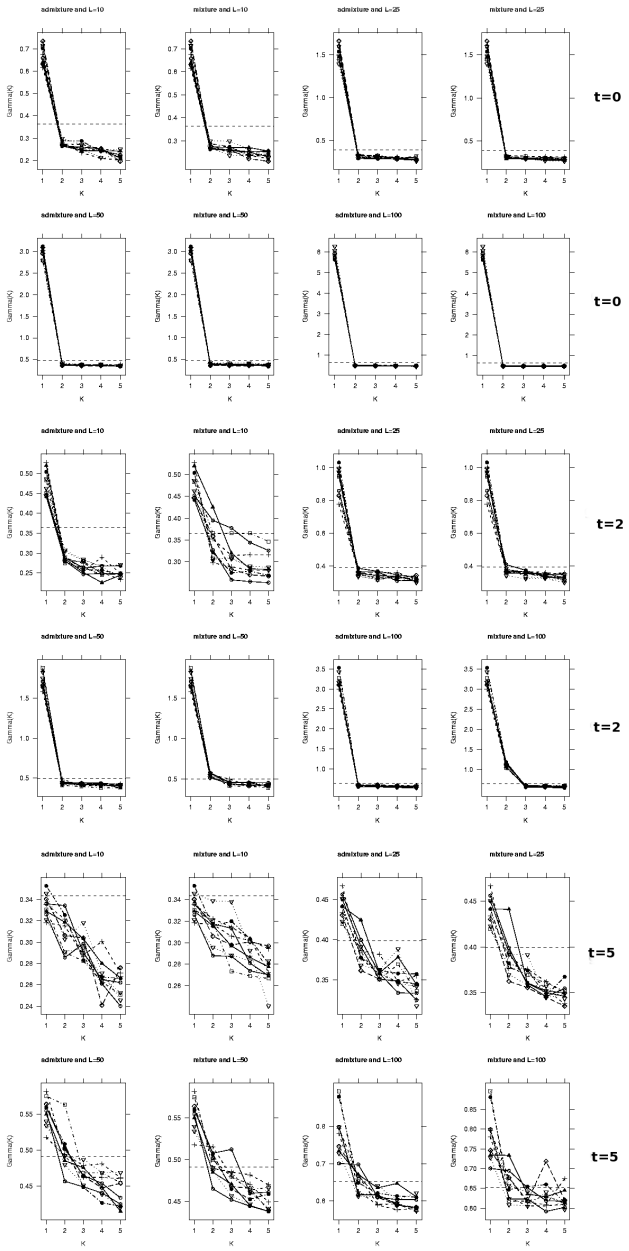


FIG. 4.4 – Scree plot of the $\lambda(K)$'s values obtained by fitting either mixture (right) or admixture (left) models at different generations, $t = 0, 2, 5$, and for different number of marker loci, $L = 10, 25, 50, 100$. The model and the number of loci used are indicated at the top of each plot, in which results from 10 random samples are displayed. The horizontal dashed line indicates the value of λ_0, which is function of L. It appears that $K = 2$ populations are enough to explain the LD due to structuration under the mixture model at $t = 0$ and the admixture model at $t = 2$, whatever the number of marker loci used. At $t = 5$ there are no longer "pure" individuals in the samples, which makes the problem much more harder than for generations $t = 0$ and $t = 2$, and the number of marker loci considered has a strong impact on the ability of the method to select the right number of populations.

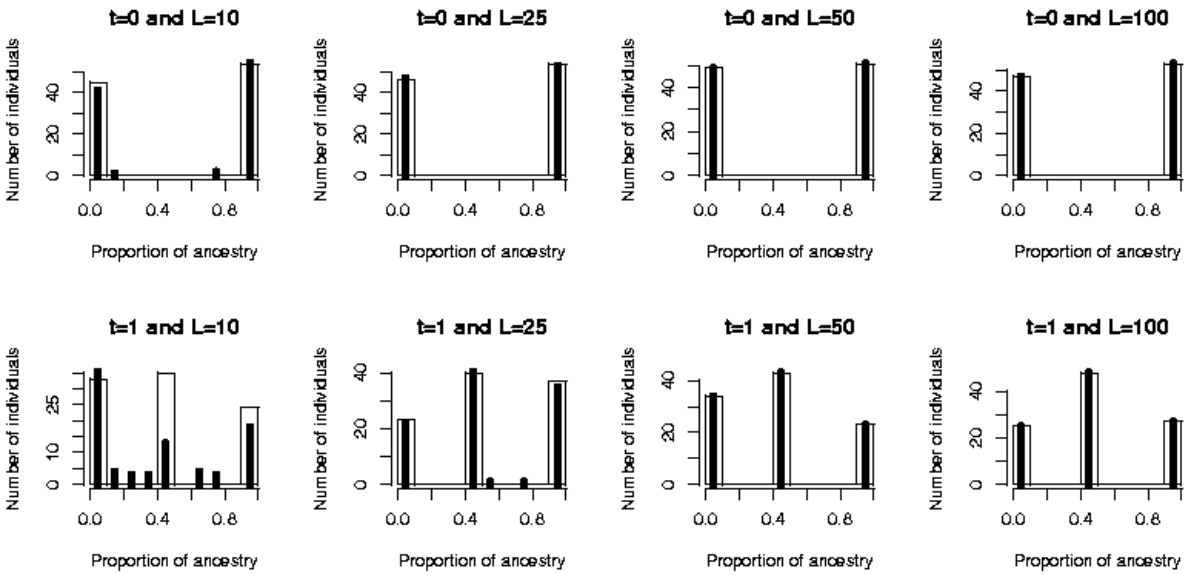


FIG. 4.5 – Summary of DPCA results on data sets of size $N = 100$ randomly drawn from the hybrid isolated population at $t = 0$ and $t = 1$ for different number of bi-allelic marker loci L. The bars stands for the histogram of the "actual" proportions of ancestry while the solid black lines represent the histogram of the inferred proportions of ancestry. For $L \geq 50$ DPCA perfectly cluster haplotypes in their appropriate population of origin.

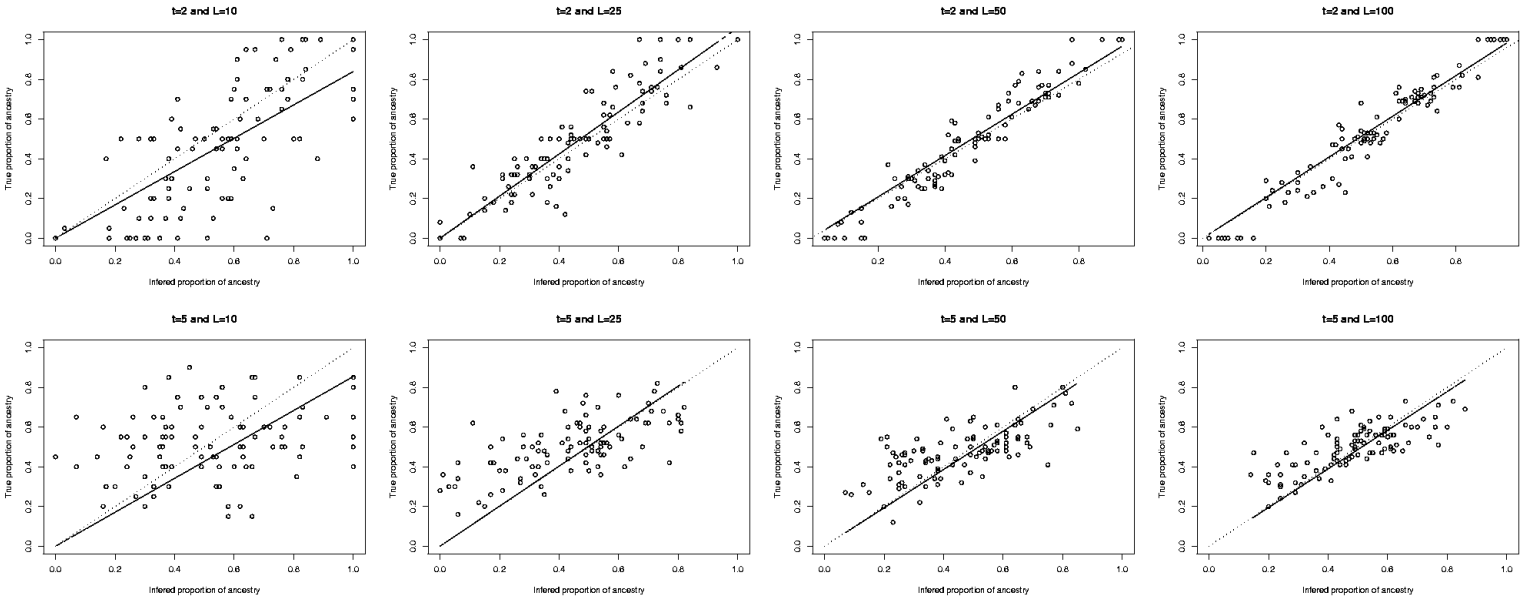


FIG. 4.6 – Summary of DPCA results on data sets randomly drawn from the hybrid isolated population at $t = 2$ and $t = 5$, with size $N = 100$, and for different number of bi-allelic marker loci L. Each plot consists of a scatter view of the estimated value of the ancestry proportions (i.e its mixing proportion $\mathbf{q}_i$) against the true ancestry proportion (i.e the proportion of alleles from each ancestral population). The dashed line stands for the bisectrix ($y = x$) and the solid line for the regression line assuming no intercept. For $L = 100$ and $t = 2$ we can clearly identify 5 clusters which represent the 5 modes of the expected ancestry proportion distribution and which can also be interpreted as the 5 possible allelic contributions of the 4 grandparents from the two ancestral populations. (PRITCHARD *et al.* (2000a)).

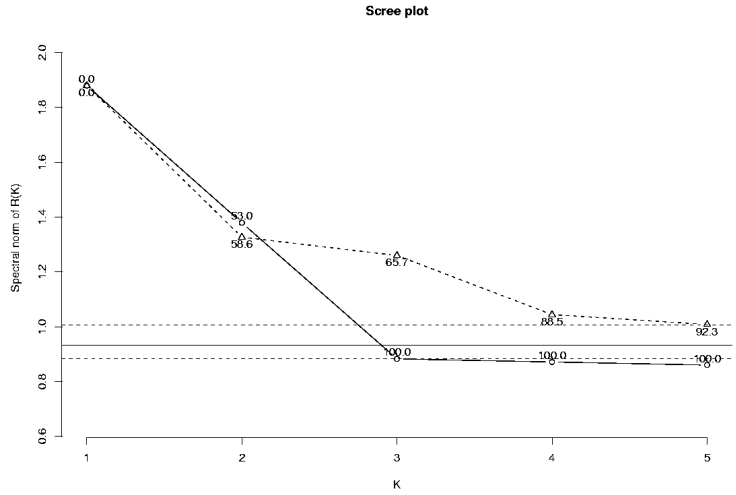


FIG. 4.7 – Scree plot of the $\lambda(K)$'s values obtained by applying DPCA on the 153 maize inbreds with 55 SSR loci. For each point we indicate the corresponding value of $\Gamma(K)$ which represents the percentage of LD explained by the model. The open circles connected by a solid line correspond to $\lambda(K)$'s values derived from the admixture model while the triangles connected by a dashed line represent the ones obtained under the mixture model. The horizontal dashed lines represent the 95% probability support of λ_0 and the horizontal solid line its mean value estimated by parametric bootstrap with 1000 replicates. The admixture model with $K = 3$ populations explain 100% of the covariance component due to structuration.

K	STRUCTUREno corr [a]			STRUCTUREcorr [b]	EM
	$\log \Pr(\mathbf{X}\|K)$ [c]	$\Pr(K\|\mathbf{X})$ [d]	$\lambda(K)$	$\lambda(K)_{\text{corr}}$	$\lambda(K)_{\text{EM}}$
2	-9572.7	~ 0	1.4107	1.4185	1.3793
3	-9466.3	~ 0	0.8975	0.9182	0.8827
4	-9383.6	~ 0.997	0.9026	0.9068	0.8715
5	-9395.7	~ 0.003	0.9034	0.8709	0.8602

TAB. 4.1 – Results of STRUCTUREon the 153 maize inbreds.

[a] Model assuming no correlation between populations

[b] Model assuming correlation between populations

[c] *ad hoc* Bayesian deviance

[d] Estimated posterior probabilities of K assuming an uniform prior for K between 2 and 5.

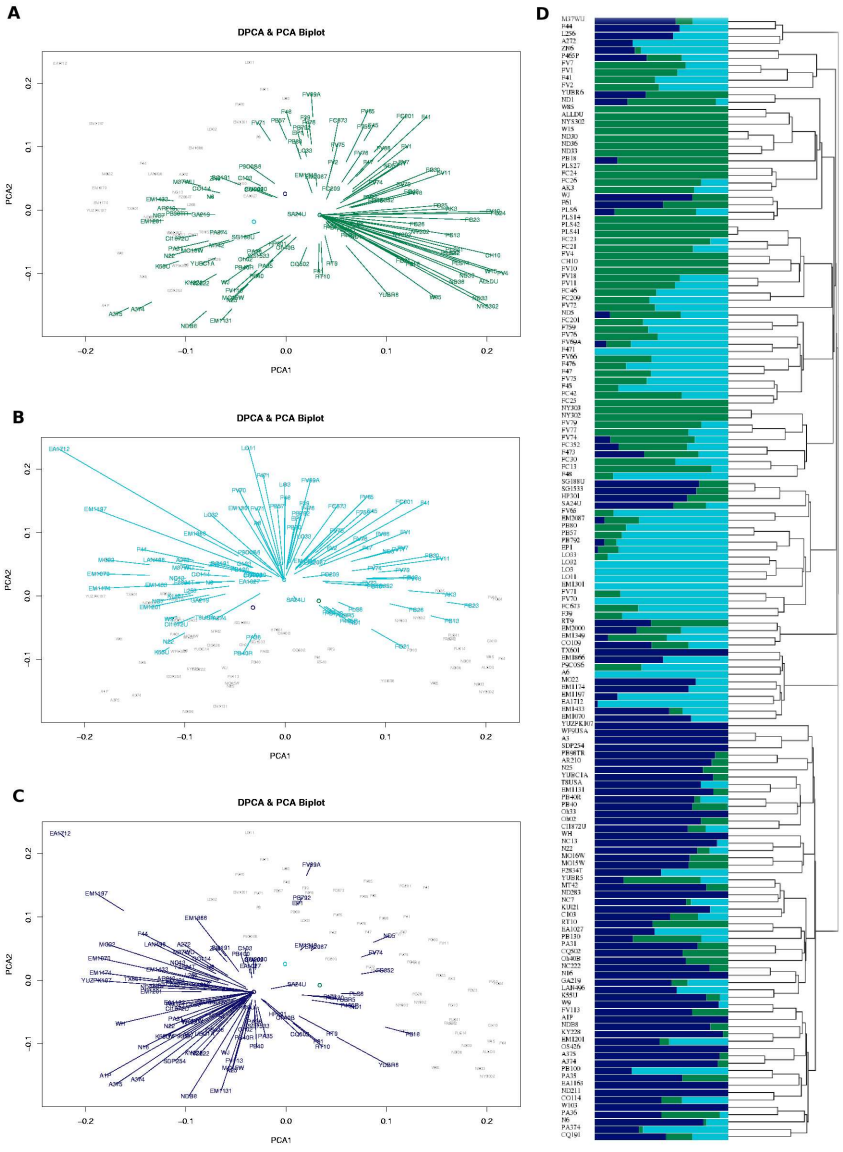


FIG. 4.8 – Join results of PCA, DPCA (for $K = 3$) and NJ hierarchical clustering on the 153 maize inbreds with 55 SSR loci. On the left side (A,B and C), the segments on the PCA biplots represent the weighted distances between each inbred to the centroid of the population cluster. The centroid is obtained by summing the scores of the inbreds weighted by their admixture proportions. On the right side (D), the admixture proportions inferred by the EM-algorithm are plotted together with the tree obtained by NJ on the distance matrix. Although the NJ clustering succeeded to separate Corn-Belt Dent material to other materials, going forward through the tree leads to successive subdivisions which are more difficult to interpret. This comparison with DPCA emphasizes the intrinsic limit of hierarchical clustering when there is admixture. The green cluster of DPCA can be interpreted as the contribution of Northern-Flint, the bright blue as the contribution of the non NF origin of European-Flint and the dark blue as the contribution of the non NF origin of Corn-Belt Dent.

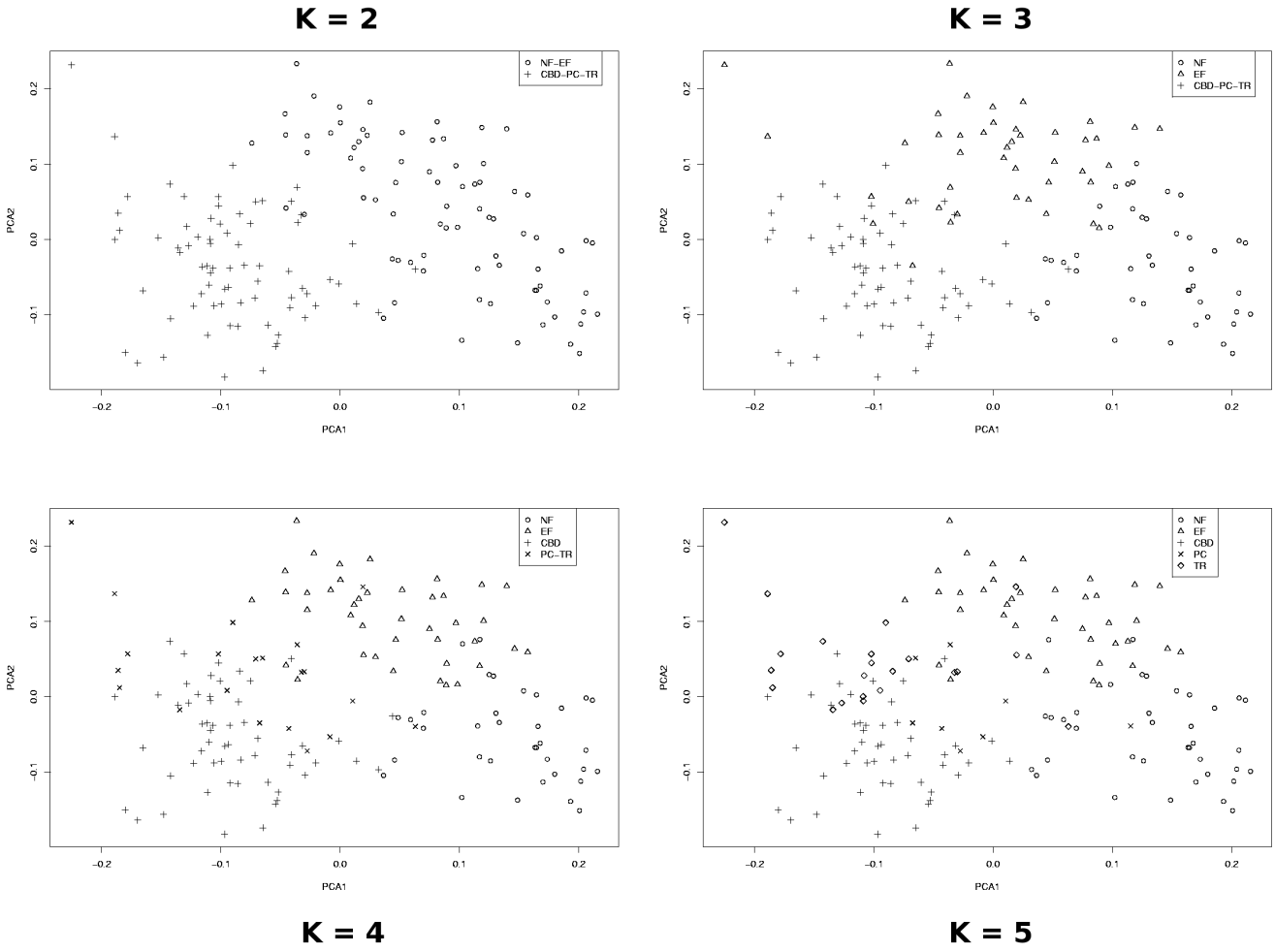


FIG. 4.9 – Illustration of the way that DPCA "extracts" population structure from the 153 maize inbred data set from $K = 2$ to $K = 5$. We plot the two first axes of the PCA together with the classification obtained from DPCA results by assigning individuals to population-group according to their maximum admixture proportions. Along with a *a priori* classification of some inbreds of our panel, we interpreted each population-group as 5 main distinct origins : Northern-Flint (NF), European-Flint (EF), Corn-Belt Dent (CBD), Popcorn (PC) and Tropical (TR).

Quatrième partie

Déséquilibre de liaison et haplotypes ancestraux

Chapitre 5

Modeling Background Linkage Disequilibrium by Ancestral Haplotype Structure

Authors : Jean-Baptiste Veyrieras [a,1], and Alain Charcosset[1]

[1] UMR, INRA UPS-XI INAPG CNRS Génétique Végétale, Ferme du Moulon, 91190 Gif-sur-Yvette, France

Keywords : linkage-disequilibrium, recombination, haplotype, HMM.

To be submitted to *Genetics*

[a]to whom correspondence should be addressed

Abstract

Recent developments of linkage disequilibrium (LD) based studies have created a pressing need for statistical methods that could take advantage of LD patterns between markers to extract useful information with regard to the hidden evolutionary process which have shaped the data. Generally, marker data can be summarized into a set of distinct haplotypes. One of the most challenging part of the issue raised by haplotypic data analysis is the identification of past recombination events. Through generations, recombination acts as a "fragmentation" process, so that each current haplotype can be viewed as a mosaic of fragments inherited from ancestral progenitors and these fragments may have been blurred by rare mutation events. In

this article we introduce a new statistical framework based on a hidden Markov model (HMM) to detect recombinant haplotypes with regard to a limited number of ancestral haplotypes. Making little assumption on these ancestral "templates", we present a 2-step algorithm which makes it possible to infer the set of ancestral haplotypes which best captures the mutation and recombination pattern among the observed haplotypes. Moreover, we show how this modelling can be interpreted as a coloring problem where one aims to minimize the number of colors in order to visualize the diversity by assigning a same color to the haplotype fragments which exhibit identical or similar conserved pattern of mutations. Finally, based on this coloring formulation we proposed a lossless compression strategy to select the optimal set of ancestral haplotypes which minimizes the number of required historical events to explain the data.

By means of simulations we show that our method yields good performances both to infer the actual ancestral haplotypes and to assign correctly ancestral fragments to the observed haplotypes. Results from the analysis of four intragenic region of candidate genes in maize illustrate the usefulness of our approach to discriminate different evolutionary histories.

Introduction

Linkage Disequilibrium (LD), which represents, at a population level, the non-random assortment of alleles at different loci along the genome, offers a valuable information in which past demographic events may have been "trapped" Nordborg and Tavare (2002). Thus, the understanding of observed LD patterns could help researchers to build up hypotheses on the underlying interplay between genetic factors and evolutionary process. In particular, LD based studies may inform us on population history (see for instance Hill (1981); Vitalis and Couvet (2001); Beaumont (2004); Wang (2005)), and can provide a useful framework to finely explore, at a whole species scale, the genetic determinism of complex traits Jorde (1995); Terwilliger and Weiss (1998); Kruglyak (1999); Jorde (2000). Historically, statistical methods aiming at exploring LD structure are based on pairwise marker loci studies and suffer of the following limitations : i) they do not consider all marker simultaneously, ii) they often yield "noisy" results which are difficult to directly interpret in terms of evolutionary process, iii) they hide recombination events.

Therefore, rather than focusing on pairwise marker analyses, better modeling of genetic variation can be achieved by considering marker haplotypes. Empirical studies of haplotype diversity along the genome have suggested that LD patterns seem to be broken into a finite number of blocks of strong haplotype structure. This "blocky" view of the genome was first reported in human (e.g. Daly *et al.* (2001); Patil *et al.* (2001)). A block is characterized

by a low haplotype diversity and it is separated from another block by shorter regions of shattered haplotype structure with higher diversity. This phenomenological interpretation of LD patterns has triggered a controversy in human : does block structure come from particular historical events or from empirical artefact ? To investigate the impact of demographic parameters on the block structure, several authors have recently carried out simulation studies. It comes out that haplotype block structures seem to be shaped by peculiar past demographic events (such as population growth) Stumpf and Goldstein (2003) and/or by the presence (or not) of recombination hotspots Wall and Pritchard (2003). Although block structure brings reliable information on the underlying evolutionary parameters, its modeling still remains subtle and its inference from marker data have two main drawbacks. First, it strongly depends on the model used so that different algorithms or diversity measures lead to different block boundaries. Second, it can fail to capture additional correlation across blocks or/and can hide sub-structure in blocks Gabriel *et al.* (2002).

On the other hand, coalescent models offer a well-established and rich theoretical framework that helps building up a better understanding of the haplotype diversity patterns expected under various demographic models. However, in the context of data analysis, coalescent-based approaches have been hampered by their computational burden (see for instance the full-likelihood methods of Griffiths and Marjoram (1996); Kuhner *et al.* (2000); Fearnhead and Donnelly (2001)). Recently, Li and Stephens (2003) proposed an intermediate but tractable approximate conditional likelihood based on coalescence theory to infer recombination rate from marker data. Using computer simulations, Li and Stephens (2003) demonstrated the competitiveness of their approach with the full-likelihood methods. They also showed the flexibility of their modeling to detect recombination hotspots.

The key point of the modeling of Li and Stephens (2003) is the interpretation of recombination as a fragmentation process so that the observed haplotypes can be viewed as a "mosaic" of fragments inherited from progenitors. Based on this representation, alternative methods have been recently developed. They rely on a haplotype by haplotype parsing scheme such that a given haplotype can be viewed as a concatenation of a limited number of conserved mutation motives. In particular, Schwartz *et al.* (2002) provided an original dynamic algorithm, called the alignment method, to parse a set of observed haplotypes using a given set of "ancestral" haplotypes which represent the ancestral source of each polymorphic sites. Simultaneously, Ukkonen (2002) proposed a dynamic algorithm to find the minimal number of founder haplotypes from a set of recombinant haplotypes. The assumption that observed haplotype samples are generally shaped by a few ancestral haplotypes has been also used by El-Mabrouk and Labuda (2004) to reconstruct minimal pathways of recombinations. More recently, this "ancestral fragmentation" model was implemented in a hidden Markov model (HMM) framework by Koivisto

et al. (2004), extending the previous work of Ukkonen (2002).

HMM offers a powerful statistical technique to tackle haplotype parsing issues. In particular, as suggested by several authors (see for instance McPeek and Strahs (1999); Li and Stephens (2003)), recombination events can be interpreted as a Markovian process, and the hidden part of the mechanism as the ancestral part of the observed diversity. In this article, we present a new HMM based approach to capture background LD by means of a limited number of ancestral (or founder) haplotypes. Our model relies on the assumption that the observed population was founded some generations ago by a limited group of ancestors. Thus, the current haplotypes result from iterated recombinations between ancestral haplotypes and may have been blurred by rare mutation events. This model is free from global block structure and flexible enough to parse long range haplotypes. After introducing the HMM, we describe an heuristic procedure to identify a limited number of ancestral haplotype templates which best capture the haplotype diversity. Then we present a lossless data compression method aiming at finding the most parsimonious ancestral haplotype composition. The ability of our approach to extract the actual ancestral fragments from data sets is evaluated by means of simulations. Finally results on four data sets, previously used by Remington *et al.* (2001) to investigate LD structure in maize genome, demonstrate the usefulness of our approach to discriminate among contrasted evolutionary scenarios.

Methods

Notation and background

The input X consists of a $N \times M$ haplotype-SNP matrix. The rows of X correspond to distinct haplotypes, where each haplotype X_i has multiplicity n_i in the original sample (i.e. $\sum_{i=1}^{N} n_i$ haplotypes have been sampled). Following Ukkonen (2002), let A the $K \times M$ ancestral haplotype matrix where each row A_k, $k = 1, \ldots, K$, is a founder haplotype, so that X has a parse in terms of A. This means that each X_i has a decomposition into $1 \leq m_i \leq M$ ancestral fragments a_{im}, such that $X_i = a_{i1}a_{i2}\ldots a_{im_i}$ and a_{ij} occurs in at least one A_k in the same location as in X_i. We implicitly assume that two successive fragments a_{ij} and a_{ij+1} are from different ancestral haplotypes of A. Here we extend the formulation of Ukkonen (2002) by allowing for imperfect fragmentation of observed haplotypes, so that $X_i = a^*_{i1}a^*_{i2}\ldots a^*_{im_i}$ and a^*_{ij} occurs in at least one A_k but the match can be imperfect at some "rare" SNP sites. This "imperfect" modelling of ancestral haplotype fragments accounts for potential mutation events.

Hidden Markov model

This ancestral fragmentation model can be interpreted as a copying process : each observed haplotype derived from the ancestral haplotypes by copying a given fragment, namely a^*_{im}, of an ancestral haplotype with imperfections in the copying process. To model this mosaic-like process (as illustrated in Figure 5.1), we introduce the following HMM. Let the hidden random variable Z_{ij} denote which ancestral haplotype X_i copies at site j (so $Z_{ij} \in \{1, \ldots, K\}$). To mimic the effect of recombination between ancestral haplotypes the Z_{ij} can be modelled using a first order Markov chain on $\{1, \ldots, K\}$ such that $\mathsf{Pr}(Z_{i1} = k) = \mathsf{Pr}(A_k)$ and :

$$\begin{aligned}&\mathsf{Pr}(Z_{ij+1} = k^{'} | Z_{ij} = k) \\ &\quad = \begin{cases} \exp(-\rho_j d_j) + (1 - \exp(-\rho_j d_j))\mathsf{Pr}(A_{kj+1}) & \text{if } k^{'} = k \\ (1 - \exp(-\rho_j d_j))\mathsf{Pr}(A_{k^{'}j+1}) & \text{otherwise} \end{cases}\end{aligned}$$

where d_j is the physical distance between SNP j and $j+1$ (assumed known), ρ_j is a parameter which reflects the intensity of recombination between sites j and $j+1$ and $\mathsf{Pr}(A_{kj+1})$ is the probability to have recombined with the ancestral haplotype A_k between j and $j+1$. Note that the probability of having not recombined, $\exp(-\rho_j d_j)$, is derived from a simple Poisson modelling of the occurrence of recombination events along the sequence. The interpretation of ρ should be taken with caution as this mosaic process is rather phenomenological than related formally to the actual genealogical tree connecting the observed haplotypes. Simply, this Poisson modelling translates the following property : if the sites j and $j+1$ are very close genetically, they are likely to copy the same ancestral haplotype, i.e $Z_{ij} = Z_{ij+1}$. In other word, the closer the SNP sites, the lower the probability to copy a different ancestral haplotype. Here we restrict the model to the case where $\rho_j = \rho$ for all j.

To mimic the fact that the copying process may be imperfect, we assumed that given the copying process $Z_{i1}, \ldots, Z_{iM}$, the alleles $X_{i1}, \ldots, X_{iM}$ are independent, with

$$\begin{aligned}&\mathsf{Pr}(X_{ij} | Z_{ij} = k) \\ &\quad = \begin{cases} 1 - \epsilon & \text{for } X_{ij} = A_{kj} \\ \epsilon & \text{for } X_{ij} \neq A_{kj} \end{cases}\end{aligned}$$

where ϵ is the error, or mutation rate, parameter.

This HMM can be understood as follows : a given haplotype X_i choose its first SNP value from the given distribution of ancestral haplotypes. Then each subsequent SNP site j may follow a recombination event with probability $1 - \exp(-\rho d_j)$. If there is no recombination event then the haplotype goes on copying the same ancestral haplotype. Otherwise the new ancestral state is sampled among the ancestral haplotypes according to a

site specific distribution $Pr(A_{kj})$, $k = 1, \ldots, K$. Note that this implies some assumptions on the underlying events. First, recombination and mutation events are assumed to be independent and constant along the region. This may be reasonable in the absence of additional information. Second, the recombination points are assumed to be independent between haplotypes.

Computation

For a given value of K, to parse the haplotypes using the above HMM framework, we should *a priori* know :

- the ancestral haplotype matrix A.
- the frequencies from which ancestral fragments are sampled following a recombination event, namely $\mathsf{Pr}(Z_{ij} = k)$, for all i, j and k.
- the recombination rate parameter, ρ.
- the mutation parameter ϵ.

First, let's assume that for a given value of K the ancestral haplotype matrix A is known. Finding the optimal parse of the haplotypes can be viewed as a maximum likelihood procedure in which target parameters are the sampling frequencies of the ancestral fragments after a recombination, namely $\mathsf{Pr}(A_{kj})$, the recombination rate parameter, ρ, and the mutation rate parameter ϵ. Let's denote $F_k = \{\mathsf{Pr}(A_{kj})\}$ the $N \times M$ matrix of the k^{th} ancestral fragment frequencies at each putative recombination site, and $F = (F_1, \ldots, F_k)$. Then the likelihood of the data can be written as follows,

$$L(X; F, \rho, \epsilon | A) = \prod_{i=1}^{N} \mathsf{Pr}(X_i; F, \rho, \epsilon | A)^{n_i}$$

where $\mathsf{Pr}(X_i; F, \rho, \epsilon | A) = \sum_{k=1}^{K} \alpha_M^{(i)}(k)$ where the $\alpha_j^{(i)}(k)$'s terms, $j = 1, \ldots, M$ and $k = 1, \ldots, K$ are the forward variables Rabiner (1989) and can be recursively computed using the following induction relation :

$$\begin{cases} \alpha_{i1}^{(i)}(k) & = \mathsf{Pr}(A_k)\gamma_1^{(i)}(k) \\ \alpha_{j+1}^{(i)}(k) & = \gamma_{j+1}^{(i)}(k)\left((1-\theta_j)\alpha_j^{(i)}(k) + \theta_j \zeta_j^{(i)}\right) \end{cases}$$

where

$$\begin{aligned} \gamma_j^{(i)}(k) &= \mathsf{Pr}(X_{ij} | Z_{ij} = k) \\ \theta_j &= 1 - \exp(-\rho d_j) \\ \zeta_j^{(i)} &= \sum_{k'=1}^{K} F_{kj}^{(i)} \alpha_j^{(i)}(k') \end{aligned}$$

and $F_{kj}^{(i)} = \mathsf{Pr}(Z_{ij} = k)$, i.e the term at row i and column j of F_k. In order to make the above maximization problem tractable, we must reduce

the parameter space. In particular, we assume that at any recombination point, the probability to have recombined with a given ancestral haplotype A_k, can be roughly approximated by the ancestral haplotype frequency, i.e $F_{kj}^{(i)} \approx \mathsf{Pr}(A_k)$ for all j and i. This means that for each haplotype, at each recombination site, the probability to choose an ancestral fragment is given by the average contribution of the corresponding ancestral haplotype. Therefore we suggest to reduce the F matrix to the vector of the frequencies of the ancestral haplotypes. This leads to the following transition matrix :

$$\mathsf{Pr}(Z_{ij+1} = k^{'} | Z_{ij} = k) = \begin{cases} (1-\theta_j) + \theta_j F_k & \text{if } k^{'} = k \\ \theta_j F_{k'} & \text{otherwise} \end{cases}$$

where F_k is the frequency of the k^{th} ancestral haplotype. Thus, given the ancestral haplotype matrix A, the parameter space is reduced to $K+1$ free parameters : the recombination rate ρ, the mutation parameter ϵ and the $K-1$ independent ancestral haplotype frequencies, F.

Up to now, we have assumed that the ancestral haplotypes have been previously defined. Generally, the diversity pattern observed along a region is shaped by a few frequent haplotype variants which dominate over a "flat" distribution of rare halotypes due to recent recombinations or new mutations. The frequent common haplotypes are likely to represent the founder ancestral haplotypes. Therefore, for a given value of K, one could build the ancestral haplotype matrix A by choosing K haplotypes among the most frequent ones. This implies that we should fix a threshold beyond which haplotypes are said to be rare, and above which they are "eligible" as ancestral haplotype. However, real data sets do not always present such a clear haplotype frequency distribution and even when it is possible to classify the haplotypes into these two categories, eligible ancestral versus rare haplotypes, to find the the best ancestral requires to test all the possible combinations of ancestral haplotypes.

Here, for a given value of K, we suggest to find a "reasonable" guess of the ancestral haplotypes matrix A by applying a proportional membership fuzzy clustering approach J.C. (1981). Let's denote Q the $N \times K$ matrix of the probabilistic contribution of the K ancestral haplotypes to each observed halotype. And let's consider A^* the $K \times M$ matrix such that the element (k, j) of A^* is given by $\mathsf{Pr}(A_{kj} = 1)$, i.e the probability that ancestral haplotype k has allele 1 at SNP j. We call A^* the probabilistic template of A. Then, the fuzzy algorithm is composed of the following steps :

1. Initialize the matrix $Q = Q^{(0)}$.
2. At the m^{th} step, obtain the probabilistic ancestral haplotype template A_k^* with :

$$A_{kj}^{*(m)} = \frac{\sum_{i=1}^{N} n_i X_{ij} Q_{ki}^{(m-1)}}{\sum_{i=1}^{N} n_i Q_{ki}^{(m-1)}}$$

3. Update Q as follows :

$$Q_{ki}^{(m+1)} = \frac{\sum_{j=1}^{M} Q_{ki}^{(m)} A_{kj}^{*(m)}}{\sum_{j=1}^{M} \sum_{k'=1}^{K} Q_{k'i}^{(m)} A_{k'j}^{*(m)}}$$

4. Stop if $||Q^{(m+1)} - Q^{(m)}|| \leq \varepsilon$, where ε is a small positive constant.

This algorithm aims to minimize the objective function :

$$J_1(X; A^*, Q) = \sum_{i=1}^{N} n_i \sum_{j=1}^{M} \log(\sum_{k=1}^{K} Q_{ik} \mathsf{Pr}(A_{kj} = X_{ij}))$$

where $\mathsf{Pr}(A_{kj} = X_{ij}) = A_{kj}^*$ if $X_{ij} = 1$, otherwise $1 - A_{kj}^*$. This function can be interpreted as a binomial log-likelihood. In other words, each SNP is assumed to be independently sampled from a mixture of binomials (the above algorithm is similar to an Expectation-Maximization (EM) procedure Dempster *et al.* (1977)). Note that this algorithm depends on the starting points $Q^{(0)}$ and the stationary point of the process may fail to give the global solution. However, each generated solution always converges to local minima or saddle points of J_1. At the last step of the algorithm, we define the ancestral haplotype matrix A by taking the best path through each probabilistic ancestral template (i.e by taking the most probable state at each SNP) :

$$A_{kj} = \begin{cases} 1 & \text{if } A_{kj}^* > 0.5 \\ 0 & \text{otherwise} \end{cases}$$

We can also derive from the membership matrix Q a starting point for the ancestral haplotype frequencies as follows :

$$F_k = \frac{1}{N} \sum_{i=1}^{N} Q_{ki}$$

In practice, we found that it yields good starting points for both the ancestral matrix A and the ancestral haplotype frequencies F. At the end of the fuzzy clustering, the problem is reduced to the maximization of a second objective function

$$J_2(X; \rho, \epsilon | A, F) = \sum_{i=1}^{N} n_i \log \left(\sum_{k=1}^{K} \alpha_M^{(i)}(k) \right)$$

which is derived from the previous likelihood and which can be maximized by applying standard numerical maximization strategies (see for instance Press *et al.* (1992)). Finally the steps of the algorithm are :

1. Initialize the matrix $Q = Q^{(0)}$.
2. Maximize $J_1(X; A^*, Q)$ w.r.t A^* and Q.

3. Compute A and F from A^* and Q.
4. Maximize $J_2(X;\rho,\epsilon|A,F)$ w.r.t ρ and ϵ.

We called this 2-step algorithm BSAH, for Blind Separation of Ancestral Haplotypes by analogy with signal and image processing (each ancestral haplotype can be though as an original signal and each observed haplotype as a mixture of this ancestral signals potentially corrupted by noise, i.e mutations). "Blind" stands for the little *a priori* knowledge (indeed nothing) on the ancestral haplotypes and their respective frequencies. In practice we found that step 4 of BSAH can be improved by iteratively i)maximizing $J_2(X;\rho,\epsilon|A,F)$ and ii)updating F using the forward-backward algorithm Rabiner (1989). It is important to note that there are no unique solution to the problem since a change in A can be balanced by a change in F, ρ or ϵ. Therefore BSAH must be thought as an heuristic algorithm.

Model selection by lossless compression

This ancestral haplotype parsing problem can be seen as a coloring problem Schwartz *et al.* (2002); Ukkonen (2002). Finding the ancestral fragments involved in the observed haplotypes is similar to finding consistant coloring of these haplotypes, where each color stands for a particular ancestral haplotype fragment. Once A, F,ρ and ϵ have been estimated via the BSAH algorithm, finding the optimal coloring of a given haplotype is equivalent to find the path through the HMM with the highest emission probability. This is the Viterbi path which can be found by applying the Viterbi algorithm Viterbi (1967); Forney (1973); Rabiner (1989). The parse returned by the algorithm is the most probable decomposition of the haplotype in ancestral fragments and can be visualized by associating a unique color to each ancestral haplotype.

For a given value of K, the above coloring problem can also be interpreted as a binary matrix decomposition procedure :

$$X = \mathbf{F}\,[A, V] + E \tag{5.1}$$

where V is the Viterbi path matrix, E the error matrix and $\mathbf{F}$ the "mapping" application which takes as input the ancestral haplotype matrix A and the Viterbi path matrix V to output a $N \times M$ matrix where each element (i, j) is given by $A_{iv_{ij}}$ and $v_{ij} \in \{1, \ldots, K\}$ is the element (i, j) of V. Then the elements of the error matrix E, which is easily computed by taking the difference between X and $\mathbf{F}\,[A, V]$, have values in $\{-1, 0, 1\}$ and E reflects the imperfection of the copying process from the ancestral haplotypes due to mutation.

Let's consider the $N \times (M-1)$ matrix R which elements are given by :

$$r_{ij} = \begin{cases} v_{ij} & \text{if } v_{ij} \neq v_{i(j-1)} \text{ for } j = 2, \ldots, M; \\ 0 & \text{otherwise.} \end{cases}$$

and the the first column of V, namely $v = (v_{11}, v_{21}, \ldots, v_{N1})$. So, the non zero values in the matrix R indicate the recombination points which are required to optimally color each haplotype, and the vector v indicates the ancestral fragment origin of the first SNP site. For example, suppose that haplotype X_i has an optimal parse with only one ancestral fragment A_k, i.e $v_{ij} = k$ for all j, then $v_i = v_{i1} = k$ and $r_{ij} = 0$ for $j = 1, \ldots, M-1$. In other word, for a given haplotype, the number of non zero values in the corresponding row of R plus one indicates the number of ancestral fragments or colors used to parse the haplotype. Thus, we can define another mapping application, $\mathbf{G}$, which used v and R instead of V to map the halotypes, leading to :

$$X = \mathbf{G}\left[A, v, R\right] + E \tag{5.2}$$

Then the set (A, v, R, E) fully describes the initial haplotype matrix X, i.e that given this set and the mapping application $\mathbf{G}$, we can entirely rebuild the SNP data matrix X. This can be viewed as an attempt to decompose the sparse and binary data matrix X using a smaller matrix A, one vector v and two sparse matrix R and E. The more R and E are sparse (i.e contains zero values), the more efficient is the decomposition. In other word, the more R is sparse , the smaller the number of recombination points. Similarly, the more E is sparse, the smaller the number of mutation or errors. Thus the set (A, v, R, E) directly reflects the desired properties of the coloring such as small number of colors or small number of recombination points or mutation points.

In order to to compare the decompositions obtained for different values of K, we then suggest to use a lossless Harwell-Boeing (HB) Duff (1986) compression strategy. A HB compression of $N \times M$ sparse matrix X consists in storing only the non zero values using a column, respectively a row, compression scheme, and thus requires $M + 2|X|_{nz}$, respectively $N + 2|X|_{nz}$ memory units, instead of NM, where $|X|_{nz} = |\{(i,j)|X_{ij} \neq 0\}|$. Without loss of generality we assume that here $M > N$ and so compressing the sparse data matrix X has a memory cost denoted $|X|_{hb} = N + 2|X|_{nz}$. Finally, the HB compression can be improved by discarding the rows or the columns which elements are all zeros leading to the cost function $|X|_{hb} = N_{nz} + 2|X|_{nz}$, where N_{nz} is the number of rows with non zero values. So, our model selection strategy consists in the following rule :

- $\mathbf{K = 1}$: the compression is obtained by defining a single ancestral haplotype A_1 obtained by taking the allele at each SNP with the highest occurrence in the matrix X (note that it is not necessary the allele with the highest frequency). This insures that the resulting HB compression of the error matrix E is minimal. So, we get :

  $$\mathbf{ZIP} = M + |E|_{hb}$$

  where **ZIP** is the memory size requirement for the compression. The first term M in the right hand side of the relation means that we have

to store the allelic state of the ancestral haplotype A_1 at the M SNP sites.

– $\mathbf{K > 1}$: in this case we have to store the decomposition (A, s, R, E). For the ancestral matrix A we have to keep in memory each allelic state for each ancestral template, leading to a cost of MK. Follows the vector v for which we have to store the N elements. Finally, come the sparse matrices of recombination R and mutation E. Thus we have the relation :

$$\mathbf{ZIP} = MK + N + |R|_{hb} + |E|_{hb}$$

This lossless compression procedure (which is trivially illustrated in Figure 5.2) offers a simple parsimonious criterion to select among several ancestral models, i.e value of K, the one which yields the coloring with the minimal number of "historical" events. Here, there are three kinds of historical events, each with its own memory cost :i) the creation of a new ancestral haplotype has cost M, ii) a single recombination between two ancestral haplotypes has cost 1, iii) a single mutation of an ancestral haplotype site has cost 1.

Implementation

Simulation

To explore the ability of BSAH to infer and detect underlying ancestral haplotypic structure, we focused on the following simulation scenario :

1. Let A_1 an ancestral haplotype such that $A_{1j} = 0$ for $j = 1, \ldots, M$.
2. Generate A_2 with a given proportion of mutations, η, from A_1. Here, a mutation consists in changing a zero into a one in A_1.
3. Create a population of size N by merging N_1 diploid individuals which genotypes are homozygote $A1$ and N_2 diploid individuals which genotypes are homozygote $A2$.
4. Make the population evolved forward in time (t) with a constant size N, a given mutation parameter μ per generation and per site, and a given recombination rate between adjacent sites per generation and per meiosis, namely c.

In our simulation we used $M = 100$ SNP sites, N_1=N_2=N/2=50, $\mu = 0.005$ and 0.01, genetic distance between adjacent SNP sites of 0.1 and 0.25 cM (note that this is approximately the recombination rate). This allowed us to simulate a "micro population" forward in time which simply mimics the effect of foundation followed by mutation and recombination events. Thus each simulated data set consists in $N = 100$ haplotypes for which we assumed that the $M = 100$ SNP lie in a 10 kbp region (i.e that the recombination rate is $\sim c/100$ per bp). We stress that this scenario is not an

attempt to model a real evolutionary process, but rather it offers a readable and flexible simulation framework to evaluate the performance of BSAH (which is illustrated assuming $K = 2$ in Figure 5.3). Here, for each simulation parameter configuration, we explored 50 simulated data sets obtained at generations $t = 5$ and $t = 10$. The results discussed below are based on a single run of BSAH by value of K and data sets. Points i),ii) and ii) focus on quantitative evaluation of our method while iv) and v) explore the consistency of the coloring.

i) First, in Figure 5.4A, the histograms of the values of K selected by the **ZIP** criterion are displayed for $c = 0.1$ and $\mu = 0.005$. This shows that for all the simulated data sets under this configuration, the criterion always detected the haplotypic structuration. For some rare data sets, the model with $K = 3$ was selected but this was mainly due to an incorrect convergence of BSAH for $K = 2$. In fact, another run of the BSAH algorithm for this data sets made it generally possible to correct the value of the **ZIP** criterion and then to choose the right number of ancestral haplotypes. For $c = 0.25$ and $t = 10$ (see Figure 5.4B), the performance of the **ZIP** criterion slightly degrades, but still remains acceptable. Note that with $c = 0.25$ and $t = 10$, the proportion of recombinant haplotypes in the simulated data sets was in average about 60%, so that, as indicated by **ZIP** criterion, it can be more parsimonious to assume more than $K = 2$ ancestral haplotypes to parse the data. Finally, simulations with a higher mutation rate, $\mu = 0.01$, have confirmed the good performance of the **ZIP** criterion (data not shown).

ii) Second, we studied the quality of the ancestral templates as returned by the BSAH algorithm for $K = 2$. In Table 5.1 we displayed the average mean squared error over SNP between the ancestral haplotypes inferred by the algorithm and the actual ones for each simulation configuration. Obviously, even for both a high proportion of recombinant haplotypes in the data set and a small proportion of mutations between the ancestral haplotypes, the algorithm was able to correctly separate these later. Again,for some simulated data sets, the error was generally due to an imperfect convergence of the algorithm and a supplemental run made it possible to resolve it.

iii) For each simulated data sets we then looked at the estimated values of the recombination rate, namely $\hat{\rho}$, and the mutation rate, $\hat{\epsilon}$, assuming $K = 2$. Their mean values over the 50 simulated data sets for each simulation parameter configuration are displayed in Table 5.2. In our simulation scenario (assuming no interference) the recombination points occur as a Poisson process of rate 1 per Morgan. At the current generation t the recombination points form a Poisson process of rate $t \times 1$ per Morgan (this comes from standard results on the addition of Poisson variables). However, since at generation $t = 0$ the population consists in only homozygote individuals A_k/A_k, the following recombination events, from $t = 0$ to $t = 1$, cannot be observed (a recombination in a genotype homozygote A_k/A_k is "transparent"). The expected rate of observable recombination is given by $c \times (t - 1)$ rather

than by $c \times t$. The mean values of the estimated recombination rate, $\hat{\rho}$, are consistent with these expected values, even if for some configurations they exhibit a small systematic bias toward lower values. It is worth noticing that decreasing η does not seem to affect the quality of the estimation. Similarly, the same remarks can be done for the estimated mutation rate $\hat{\epsilon}$ (see Table 5.3). Nevertheless, for $t = 10$ and $\mu = 0.01$, $\hat{\epsilon}$ shows a higher bias than for lower mutation rate and number of generations. Note that increasing t and μ leads to increase the probability of recurrent mutations, so that the expected proportion of observable mutations may be lower than the expected rate computed as the product of the mutation rate μ and the number of generations t.

iv) We now looked at the consistency of the coloring obtained assuming $K = 2$. A coloring is said to be consistent if it is able to detect the recombinant haplotypes and to correctly assign colors to ancestral fragments. We first focused on the distribution of the ratio between the number of recombination points detected by the coloring and the actual number of recombination points in each simulated data sets. This is depicted in Figure 5.5 for simulated data sets assuming $\mu = 0.01$. We can see that, when the distance between ancestral haplotypes decreases, the mean of the distribution tends to shift toward negative values (a negative value indicates that the coloring has underestimated the actual number of recombination points). This was expected since decreasing η leads to "hide" some recombinations for which the ancestral fragments involved are very close (even similar). The same phenomenon occurs when both c and t increase. In fact, for small ancestral fragments the cost of recombination is balanced by the cost of mutation. Thus, if small ancestral fragments are very close (i.e just a few mutations distinguished them), then the cost of a double recombination can be higher than the cost of mutation (this phenomenon also occurs at the "borders" of the sequence). Note that this explains the apparent negative correlation between η and the mean values of $\hat{\epsilon}$ in Table 5.3.

v) Another indicator of the ability of BSAH to infer recombination points is to look at the probabilities of recombination at each site. These probabilities can simply be derived between each SNP site as the mean over haplotypes of the probability to have recombined. This later is derived as the fraction between the probability of emitting the haplotype along paths having a recombination in that point and the total probability of emitting the haplotype along any path. This is illustrated in Figure 5.6 in which we have plotted these probabilities together with the recombination points found by the coloring for two simulated data sets with $\mu = 0.005$, $c = 0.1$ (Figure 5.6A) and $c = 0.25$ (Figure 5.6B). We can see that BSAH yields consistent results with regard to the actual pattern of recombinations in the data sets. This also illustrates how recurrent mutation events affect the result of BSAH : some recombination points are shifted to a few SNP sites near, but generally are very close from the actual points. In Figure 5.6A

there is also a "border" effect : due to the slight difference between the ancestral haplotypes at both the first and the last SNP sites, recombination points at the extremities of the region are not detected by the coloring.

Finally, in order to study the impact on BSAH of heterogeneous recombination rate along the sequence, we also investigated simulated data sets assuming a single recombination hotspot in the middle of the sequence. In Figure 5.7 we plotted the profile of the probabilities of recombination obtained after a run of BSAH assuming $K = 2$ ancestral haplotypes on a simulated data set with a hotspot intensity of 10 and a background recombination rate $c = 0.1$ (i.e that the genetic distance between sites in the hotspot region was equal to $10 \times c = 1$ cM). The profile clearly indicates the presence of the hotspot. Other simulation results have confirmed that the algorithm was flexible enough to cope with heterogeneous pattern of recombination rate along the sequence (data not shown).

Application

To illustrate how BSAH works on biological data we investigated 4 intragenic aligned maize DNA sequence data sets downloaded from the panzea data base (http ://www.panzea.org/). These maize inbred DNA sequences correspond to flanking and coding regions of 4 candidate genes in maize : indeterminate1 (*id1*), dwarf8 (*d8*), dwarf3 (*d3*) and teosinte branched 1 (*tb1*). These genes are candidate for variation in plant height and/or flowering time in maize and have been previously studied by Remington *et al.* (2001) in order to explore the structure of LD in the maize genome. For each data set, we considered only the bi-allelic SNP sites and treated contiguous indel sites showing identical patterns of variation as single polymorphism. We also removed SNP sites for which there were more than 50% of missing data sites. The total number of maize inbreds for *id1*, *d8*, *d3* and *tb1* were 37,53,71 and 73, and they clusterized into 34,30,49 and 69 distinct haplotypes, respectively.

Thus, for each gene we ran 10 times the BSAH algorithm for $K = 2, 3, 4, 5$ and 6. For a given value of K, we selected the output of the BSAH algorithm with the smallest value of the **ZIP** criterion (note that generally the ten outputs were very close, even identical). These values are displayed in Table 5.4. If 4 ancestral haplotypes seem to best capture the haplotypic diversity for *id1* and *d3*, 3 are enough for *d8* and only 1 for *tb1*. It has been shown that *tb1* exhibits evidence of deviation from neutral equilibrium evolution due to its key role in the maize domestication (Wang *et al.* (1999, 2001); Tenaillon *et al.* (2001); Przeworski (2003); Clark *et al.* (2005)) : the *tb1* locus is responsible for the short lateral branches that distinguish maize from teosinte, its wild progenitor. This is evidenced both by the low diversity observed along the sequence and by the **ZIP** criterion (see Table 5.4). In fact, the observed haplotypes for *tb1*

can be explained by some rare mutations rather than by extracting distinct haplotypic patterns. This result suggests a unique haplotypic origin of *tb1* in maize, which might have been maintained through generations due to the high human artificial selection during the domestication process of maize. On the other hand results suggest that several ancestral haplotypes were transmitted from teosinte to maize for d8, id1 and d3 and were maintained throughout the post domestication selection process.

Consistently, these three other genes show a higher level of diversity than *tb1* (see Table 5.4). For *d8*, the coloring of the gene suggests a low proportion of recombinant haplotypes (88% of the haplotypes directly derived from the three ancestral templates) with regard to *id1* and *d3*. This is illustrated in Figure 5.8B which gives the distribution of the ancestral fragment lengths as found by the coloring of the three genes. The persistence of LD with distance seems also to be more important for *d8* than for the two other genes (see Figure 5.8A). Similarly, the estimated values of ρ for each gene highlights this difference : $\hat{\rho} = 5.75$ for d3, $\hat{\rho} = 1.61$ for id1 and $\hat{\rho} = 0.25$ for d8 (these values are given per kbp). This apparent singularity of *d8* was also pointed out by Remington *et al.* (2001) which hypothesized that, due to its role in flowering time variation Thornsberry *et al.* (2001); Camus-Kulandaivelu *et al.* (2006), *d8* may have been under strong divergent selection for adaptation to contrasted environments.

Finally, the optimal coloring of *d3*, *id1* and *d8* are depicted in Figure 5.9. This figure is also based on the haplotype block structure found by applying the algorithm of Anderson and Novembre (2003). This shows that the inferred ancestral fragments do not necessarily line up at block boundaries, and suggests that the data might be more parsimoniously described by not assuming an identical discrete block partition among haplotypes. Besides, for both *d3* and *id1*, one ancestral haplotype is fragmented over haplotypes and, contrary to the three other ones, it is not represented by at least one continuous haplotype along the entire region. For these two genes, this fragmented ancestral haplotype captures original mutation patterns which occur in the three other ancestral haplotypic backgrounds. This illustrates the flexibility of BSAH to separate contrasted haplotypic substructures and its ability to extract haplotype "motives" according to their distribution and their correlation in the data. We also compared the distribution of the observed haplotypes in the ancestral groups derived from the coloring and the classification of the inbreds into the 3 categories defined by Remington *et al.* (2001) : tropical/semi-tropical (ST), Stiff Stalk (SS) and non-Stiff-Stalk (NSS) origins. Results are summarized in Table 5.5. For *d3* and *id1* the distribution of the ancestral fragments in each group is to a large extent similar to the frequency of the ancestral haplotypes in the whole data set. Conversely, the SS lines seem to diverge from the NSS and ST lines for *d8*. This result must be interpreted with caution as the SS lines are under represented in the *d8* data set (only 13% of the observed haplotypes belong

to the SS origin). Nevertheless, the divergent nature of SS lines regarding to the NSS and ST lines was pointed out by Remington *et al.* (2001) from results based on genetic diversity analysis using SSR loci and recent studies on *d8* have showed that *d8* diversity is highly correlated with the genetic structure of maize inbred lines (see for instance Camus-Kulandaivelu *et al.* (2006)). The apparent differentiation of the ancestral haplotypes in *d8* in the different groups could reflect the foundation and selection events related to the adaptation of maize to temperate climate.

Conclusion

We have presented a new method to detect haplotypic structure in set of haploid sequences. Our method does not rely on a prior assumption that haplotype diversity can be partitioned into blocks and makes it possible to infer recent historical events based on the hypothesis that the observed haplotypes were derived by iterative recombination and mutation events from a few number of ancestral haplotypes. We proposed an original algorithm which is able to separate the optimal set of ancestral haplotypes, called BSAH. This algorithm can be used together with a simple and readable parsimonious criterion to find the optimal solution which minimizes the number of "historical" events required to capture the observed haplotypic diversity. Simulations showed that BSAH associated with the model choice strategy, **ZIP**, yielded good performances provided that the patterns of mutations between the ancestral haplotypes are not too close. Besides, results of applying BSAH to 4 intragenic DNA sequence data sets in maize are in good agreement with the knowledge of these regions.

It is worth noting that BSAH is not restricted to SNP data analysis and can be easily extended to multiallelic loci by modifying A^*, the matrix of the probabilistic template of the ancestral haplotypes, in order to take into account additional alleles. Similarly, we can modify $\mathsf{Pr}(X_{ij}|Z_{ij}=k)$ so that :

$$\mathsf{Pr}(X_{ij}|Z_{ij}=k) = \begin{cases} 1-\epsilon \text{ if } X_{ij}=A_{kj} \\ \frac{\epsilon}{m_j-1} \text{ otherwise} \end{cases}$$

where m_j is the number of alleles at locus j and $X_{ij} \in \{1,\ldots,m_j\}$.

Finally, we stress that our approach is mainly phenomenological and that results must be taken with caution regarding to the underlying genealogical tree which connects the haplotypes. More experimental evaluation on both simulated and real data are needed to investigate the reliability of our method for inferring mutation and recombination parameters. Nevertheless we anticipate that BSAH can be useful in the ongoing analysis and understanding of the genetic diversity in several species. In particular, by reducing the apparent genetic diversity (and so the number of variables) around just a

TAB. 5.1 – Average mean error between the estimated ancestral haplotypes and the actual ones by applying BSAH on 50 simulated data sets per simulation parameter configuration (i.e a cell in the table) where η is the proportion of mutations between the two ancestral haplotypes, c the recombination rate, and μ the mutation rate. For each run, the error was computed as the mean squared error between the inferred ancestral haplotypes and the actual ones.

t	c (cM)	f_r [a]	μ ($\times 10^2$)	$\eta = 1.0$	$\eta = 0.75$	$\eta = 0.5$	$\eta = 0.25$
5	0.1	~ 0.17	0.5	0.0000	0.0000	0.0000	0.0000
			1.0	0.0000	0.0000	0.0000	0.0000
5	0.25	~ 0.36	0.5	0.0000	0.0000	0.0000	0.0000
			1.0	0.0000	0.0000	0.0000	0.0000
10	0.1	~ 0.34	0.5	0.0008	0.0015	0.0007	0.0018
			1.0	0.0047	0.0030	0.0033	0.0038
10	0.25	~ 0.63	0.5	0.0004	0.0008	0.0009	0.0069
			1.0	0.0042	0.0060	0.0040	0.0041

[a] Average frequency of recombinant haplotypes in the simulated data sets.

TAB. 5.2 – Mean values of the estimated recombination rate $\hat{\rho}$ obtained by applying BSAH for $K = 2$ on 50 simulated data sets per simulation parameter configuration (i.e a cell in the table) : μ is the mutation rate, c the genetic distance between SNP, t the number of generations, and η the proportions of mutated sites between ancestral haplotypes.

μ ($\times 10^2$)	c (cM)	t	$c \times (t-1)$ [a] ($\times 10^4$ per bp)	$\mathsf{E}(\hat{\rho})$ ($\times 10^4$ per bp)			
				$\eta = 1.0$	$\eta = 0.75$	$\eta = 0.5$	$\eta = 0.25$
0.5	0.1	5	~ 0.40	0.38	0.42	0.36	0.34
	0.1	10	~ 0.90	0.87	0.81	0.85	0.79
	0.25	5	~ 1.00	0.96	0.88	0.95	0.90
	0.25	10	~ 2.25	2.19	2.28	2.11	2.03
1.0	0.1	5	~ 0.40	0.36	0.37	0.38	0.36
	0.1	10	~ 0.90	0.84	0.94	0.80	0.86
	0.25	5	~ 1.00	0.99	0.99	0.95	0.95
	0.25	10	~ 2.25	2.30	2.21	2.21	2.29

[a] In our simulation scenario, the generation $t = 1$ consists in heterozygote individuals $A1/A2$ so that previous recombination events in homozygote individuals A_k/A_k, $k = 1, 2$, are not observable. This explains the use of $c \times (t-1)$ instead of $c \times t$ to define the expected recombination rate.

few number of components, i.e the ancestral haplotypes, BSAH might help to preserve power in association studies.

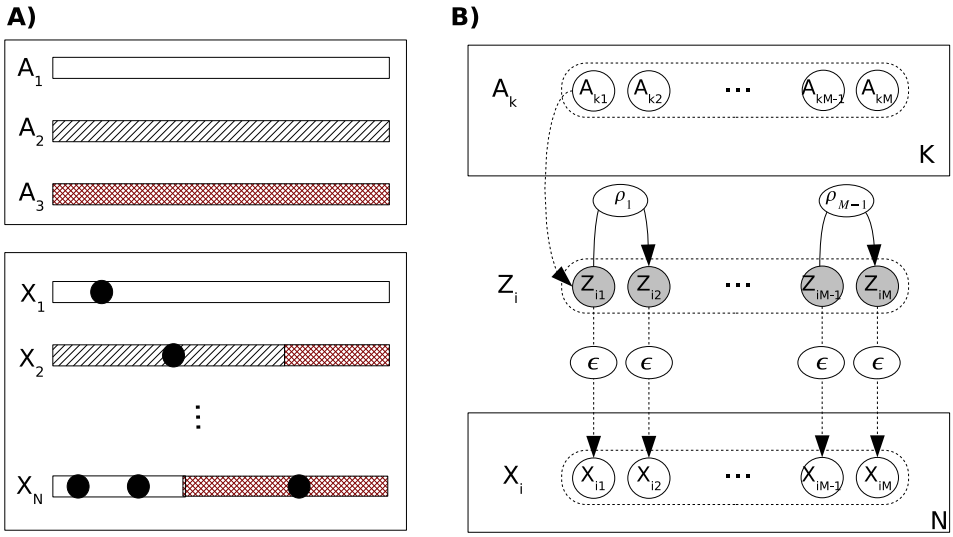


FIG. 5.1 – A) Illustration of how the sampled haplotypes X_1, X_2, ..., X_N, are build as imperfect mosaics of three ancestral haplotypes A_1, A_2 and A_3. The shading in each case shows which ancestral haplotype was copied at each position along the sequence. Black circles indicate that the copy was imperfect due to mutation events. This copying process can be thought as a Markov process along the sequence and is depicted in B) where open circles represent the state variables, filled circles the hidden variables which indicate to which ancestral haplotype each SNP belongs, and ovals the parameters. The parameter ϵ represents the mutation rate and the parameter ρ_l accounts for the recombination rate between adjacent sites.

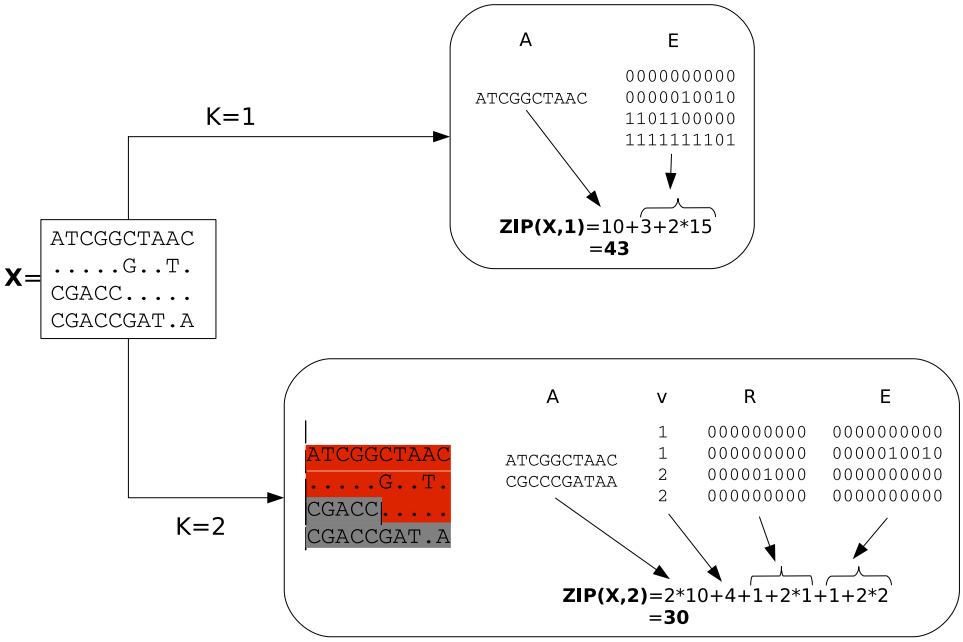


FIG. 5.2 – Interpretation of the haplotype coloring problem as a lossless compression mechanism. The **ZIP** criterion indicates the memory cost of the storage of the data matrix X according to its decomposition into the ancestral haplotype matrix, A, the vector of origin of the first SNP, v, the recombination matrix, R, and the error matrix, E. Here the **ZIP** criterion suggests to choose the model with $K = 2$ ancestral haplotypes.

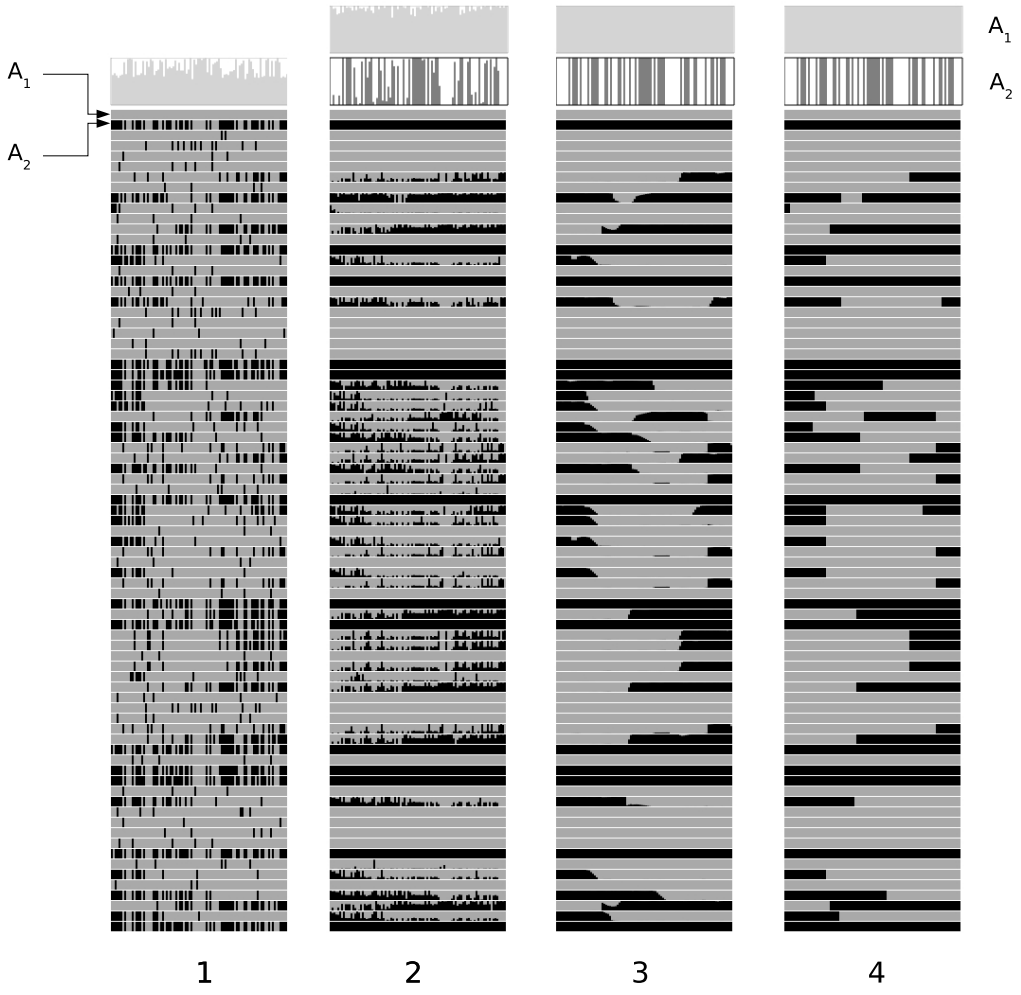


FIG. 5.3 – Illustration of the BSAH algorithm on a simulated data set with $t = 10$, $\eta = 0.6$, $\mu = 0.005$ and $c = 0.1$. 1) View of the distinct haplotypes in the data set : at the top, the box shows the SNP frequencies and the two first haplotypes are the two founder haplotypes (A_1 and A_2). 2) Output of step 1 of BSAH : the two boxes at the top represent the two probabilistic ancestral templates as computed by the fuzzy clustering. Then for each haplotype, the probabilities to belong to the two ancestral templates at each SNP site are given. 3) Output of step 2 of BSAH : for each haplotype the probability of origin of each SNP are displayed and, at the top, the two inferred ancestral halotypes are given. 4) Coloring of the data set obtained by applying the Viterbi algorithm.

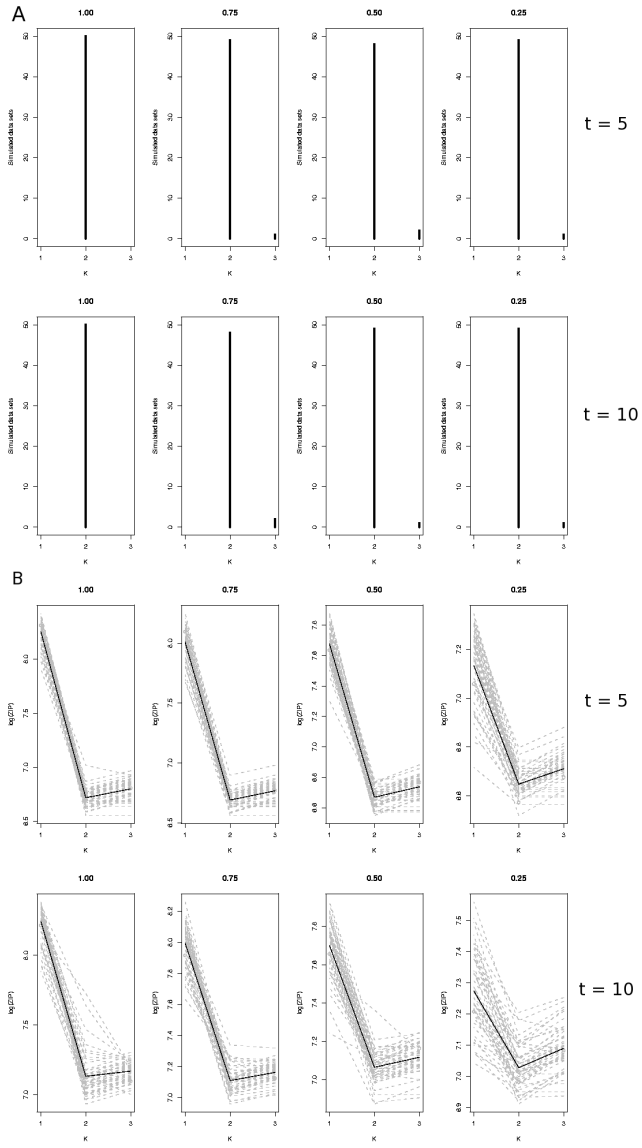


FIG. 5.4 – A) Histogram of the **ZIP** compression criterion results for simulated data sets with $c = 0.1$ and $\mu = 0.005$. Each plot consists in 50 independent simulations for different values of η, the proportion of mutations between the two ancestral haplotypes (the value is indicated at the top of each box). B) Plot of the **ZIP** compression criterion against the values of K for simulated data sets with $c = 0.25$ and $\mu = 0.005$ based on 50 replicates per value of η (gray dashed lines). The mean values of the criterion are connected by a black solid line. For A) and B), the t value at the right indicates the number of generations. For both configurations, the **ZIP** criterion was able to detect the haplotypic structure and in most of cases, it yielded to select the actual number of ancestral haplotypes, i.e $K = 2$.

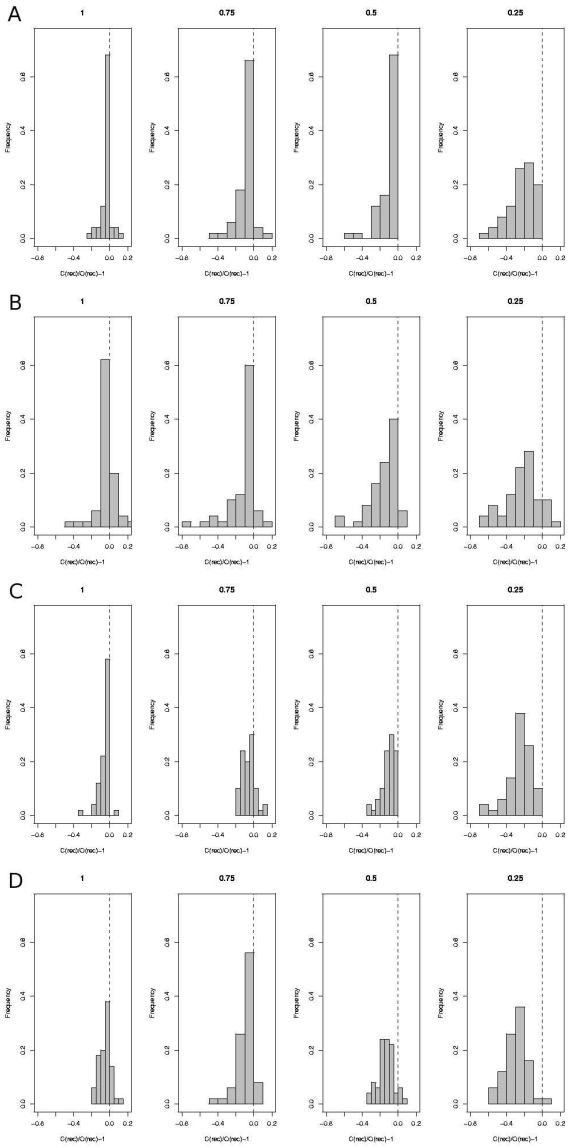


FIG. 5.5 – Histograms of the centered ratio between the number of recombination points detected by the coloring, namely C(rec), and the number of actual recombination points in the data set, O(rec). Then, a value of zero (visualized by a dashed line) indicates that the coloring has detected as many recombination points as the actual number, negative values reveals that the coloring has underestimated the number of recombination points, and vice versa. For each plot, the histogram was build from 50 coloring based on 50 independent simulated data sets. The value at the top of the box indicates the proportion of mutations between the two ancestral haplotypes, η. A)$c = 0.1$ and $t = 5$ B)$c = 0.1$ and $t = 10$, C)$c = 0.25$ and $t = 5$ and D) $c = 0.25$ and $t = 10$. For A),B),C) and D) simulations were done by assuming $\mu = 0.01$.

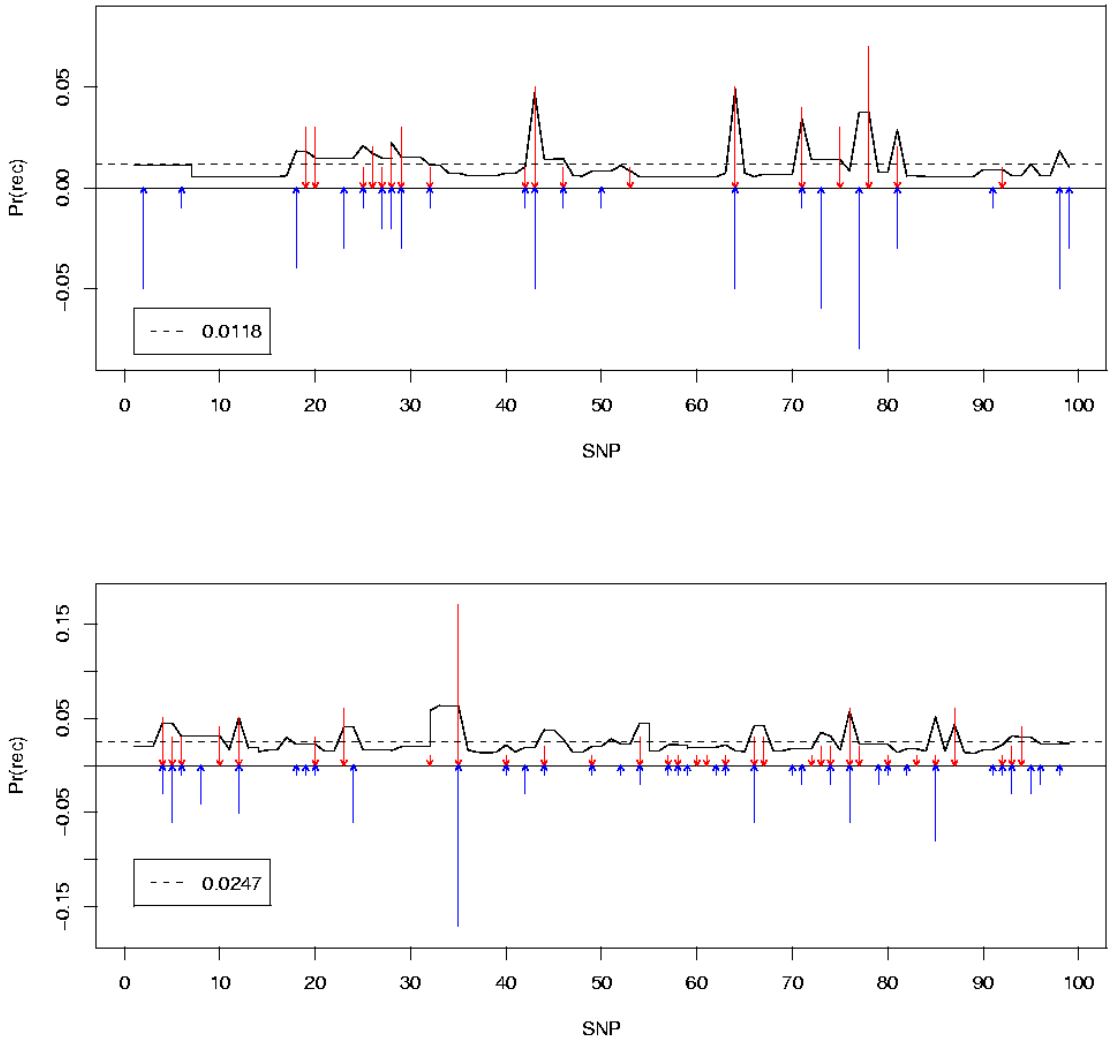


FIG. 5.6 – Illustration of the ability of BSAH to infer recombination points for two simulated data sets with $\eta = 0.5$, $t = 10$, $\mu = 0.005$ and genetic distances of A) $c = 0.1$ and B) $c = 0.25$ between adjacent SNP sites. The curve represents the probabilities of recombination at each site has outputted by BSAH, and the dashed line its mean value (which is reported in the legend). The arrows in the top part of the plot stand for the recombination points found by the coloring and the arrows at the bottom the actual recombination points. The length of the arrow indicates the proportion of corresponding haplotypes.

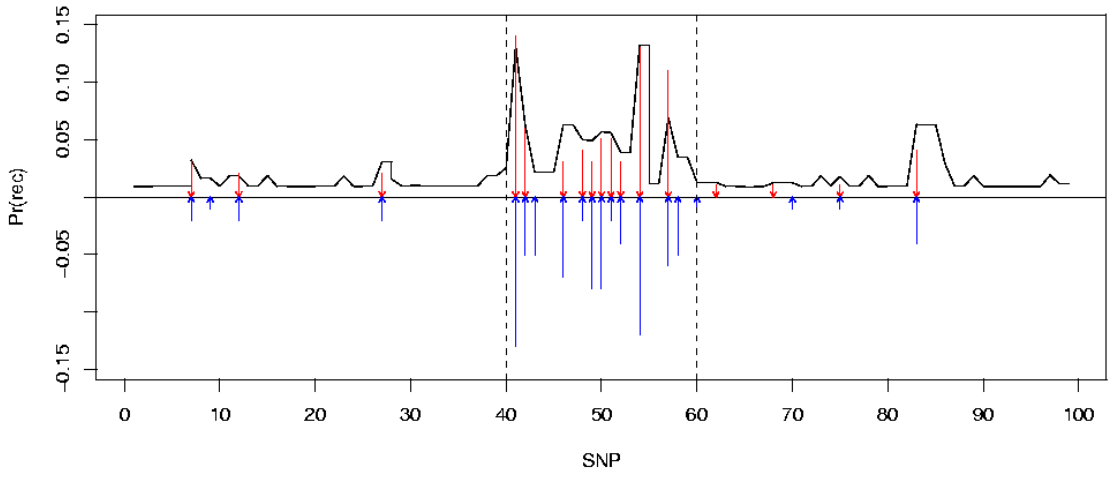


FIG. 5.7 – Illustration of the flexibility of BSAH when recombination rate is heterogeneous along the SNP sequence. The data set was simulated by assuming a single recombination hotspot at the middle of sequence and in which the recombination rate were 10 times higher than the background recombination rate, $c = 0.1$ (cM) (the hotspot is bounded between the two vertical dashed lines in the figure). The simulation parameters were $\eta = 0.5$, $t = 10$, and $\mu = 0.005$. See Figure 5.6 for the meaning of the curve and arrows in the figure.

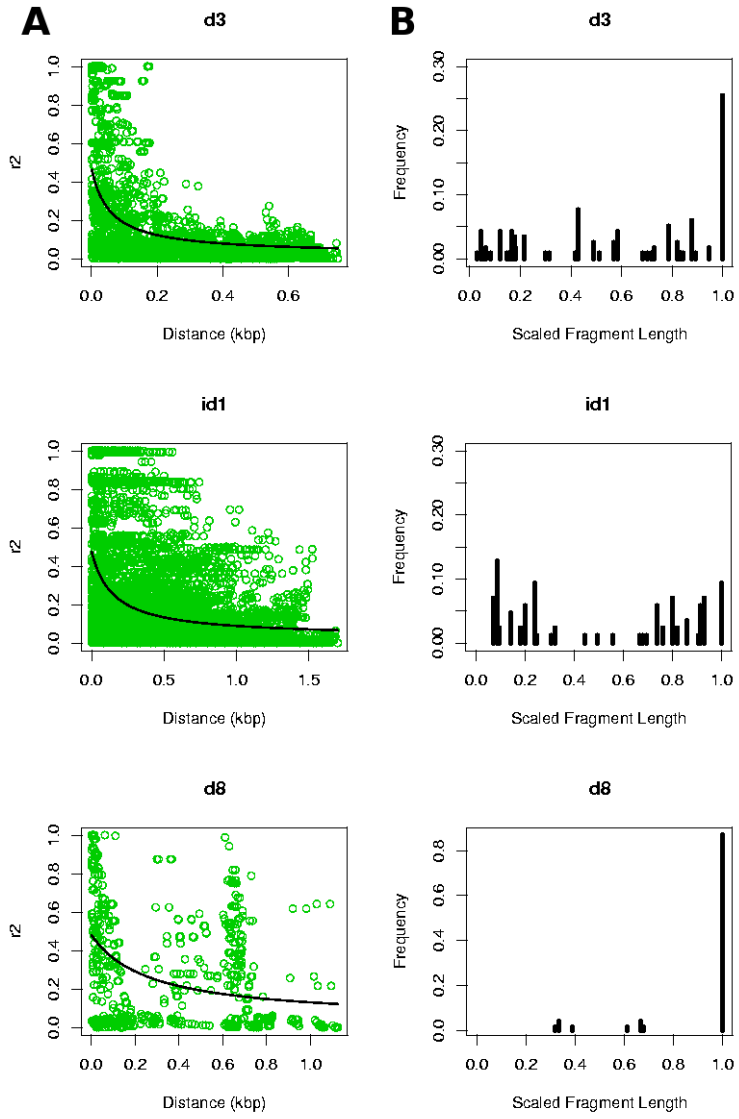


FIG. 5.8 – A) LD (r^2) decay as function of distance for *d3*, *id1* and *d8*. The solid line in each plot shows the non linear regression of r^2 on distance by using a mutation-recombination-drift model (see REMINGTON *et al.* (2001) for details on this regression). Regression coefficients were 35.75, 13.84 and 7.62 (per kbp) for *d3*, *id1* and *d8*, respectively. B)Distribution of the ancestral fragment lengths (rescaled comparing to the total length of the sequence) for the genes *d3*, *id1* and *d8* derived from the coloring of each gene with $K = 4$ ancestral haplotypes for *d3* and *id1*, and $K = 3$ for *d8*. The estimated values of the recombination rate between ancestral haplotypes were 5.75, 1.61 and 0.25 (per kbp) for *d3*, *id1* and *d8*, respectively. These values and the distribution of the fragment lengths clearly reflect the low level of recombination for *d8* regarding to the two other genes.

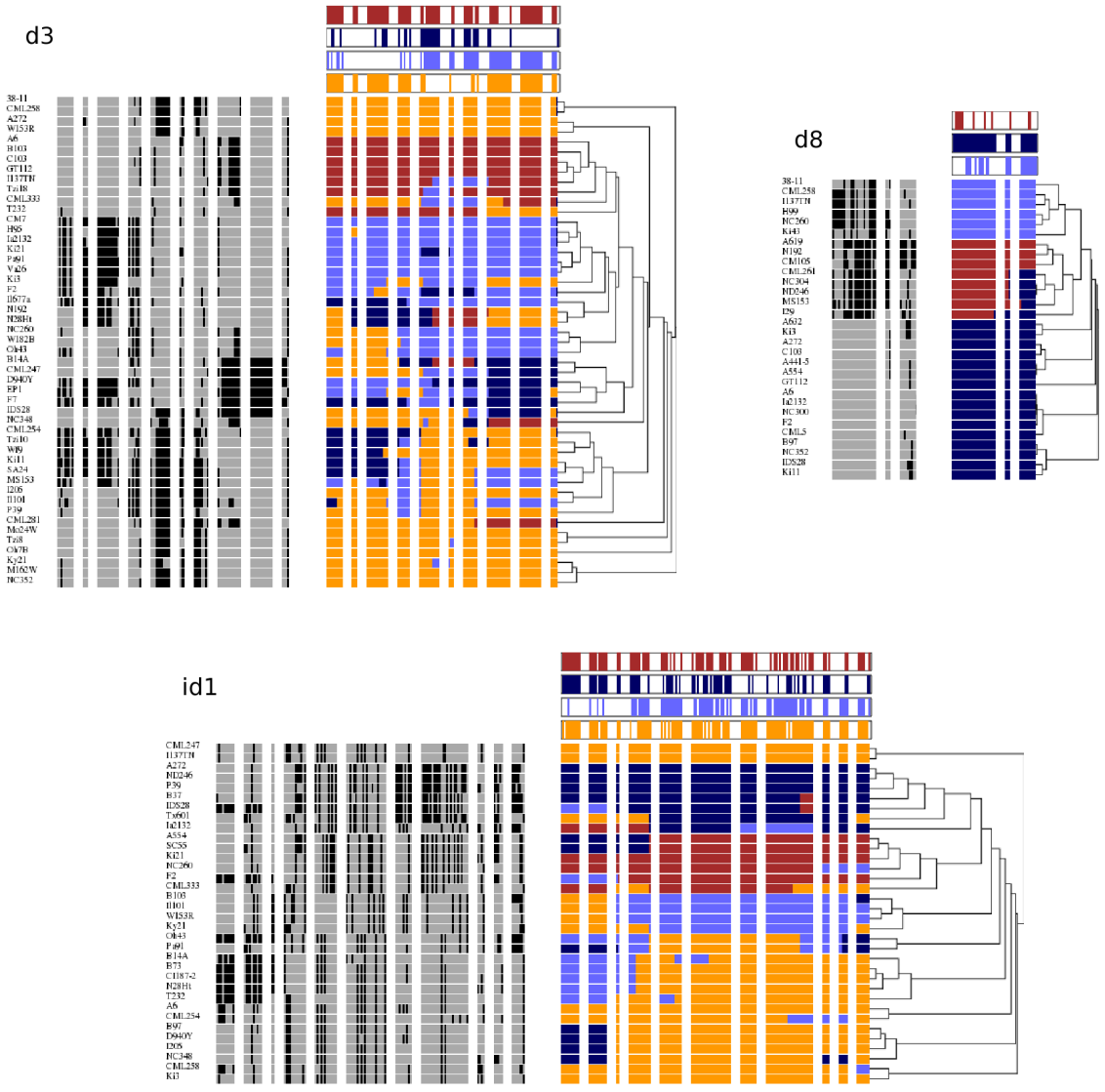


FIG. 5.9 – Coloring of the genes *d3*, *d8* and *id1*. The haplotypes are sorted according to their leaf position in a neighbor-joining tree based on the Euclidean distance matrix between haplotypes. For each gene the left part of the figure depicts the pattern of mutation in the raw data set and the right part the corresponding coloring derived from the best output of BSAH (i.e $K = 4$ for *d3* and *id1*, and $K = 3$ for *d8*). At the top of each coloring, the inferred ancestral haplotypes are displayed. The average mean distances between ancestral haplotypes for *d3*, *id1* and *d8* are 0.50, 0.48 and 0.61,respectively. The extra spaces between SNP sites are based on a block structure of the haplotype inferred by MDBlock ANDERSON and NOVEMBRE (2003). If some ancestral fragments line up with the block boundaries (for *d8* the block structure and the coloring give similar pattern), the "blocky structure" for *d3* and *id1* seems to ignore residual between and within block haplotype substructure.

TAB. 5.3 – Mean values of the estimated mutation rate $\hat{\epsilon}$ obtained by applying BSAH for $K = 2$ on 50 simulated data sets per simulation parameter configuration (i.e a cell in the table) : μ is the mutation rate, c the genetic distance between SNP, t the number of generations, and η the proportions of mutated sites between ancestral haplotypes.

c (cM)	μ ($\times 10^2$)	t	$\mu \times t$ ($\times 10^2$ per site)	$\mathsf{E}(\hat{\epsilon})$ ($\times 10^2$ per site)			
				$\eta = 1.0$	$\eta = 0.75$	$\eta = 0.5$	$\eta = 0.25$
0.1	0.5	5	2.5	2.30	2.48	2.62	2.94
	0.5	10	5.0	4.54	4.87	4.96	5.11
	1.0	5	5.0	4.60	4.80	4.77	4.85
	1.0	10	10.0	8.78	8.90	8.89	8.99
0.25	0.5	5	2.5	2.45	2.50	2.71	2.93
	0.5	10	5	4.60	4.82	4.95	5.17
	1.0	5	5.0	4.59	4.63	4.77	4.88
	1.0	10	10.0	8.86	8.97	8.90	8.85

TAB. 5.4 – Results of the lossless compression obtained by applying BSAH on the DNA sequence data sets for the four maize genes. (a) Nei diversity

Gene	N	M	Length (kbp)	$D^{(a)}$	K					
					1	2	3	4	5	6
id1	34	119	1.703	0.37	1669	1575	1392	**1245**	1383	1473
d3	49	82	0.75	0.38	1943	1743	1780	**1618**	1753	1861
d8	30	36	1.125	0.29	537	515	**504**	544	606	611
tb1	69	59	3.361	0.13	**726**	849	916	996	1064	1135

TAB. 5.5 – Distribution of the ancestral haplotype fragments into the 3 groups of origin of the maize inbred lines for the genes *d3*, *id1* and *d8*. For each ancestral haplotype we give its estimated frequency in the whole data set followed by the proportion of its corresponding fragments in each group of origin.

Gene		Ancestral Haplotypes			
	Origin	A1	A2	A3	A4
d3		0.18	0.14	0.28	0.40
	NSS	0.10	0.10	**0.40**	0.39
	SS	0.13	0.32	0.17	**0.37**
	ST	0.24	0.12	0.15	**0.49**
id1		0.17	0.21	0.18	0.44
	NSS	0.19	0.28	0.23	**0.29**
	SS	0.01	0.24	0.19	**0.56**
	ST	0.14	0.12	0.09	**0.65**
d8		0.20	0.59	0.21	-
	NSS	0.18	**0.59**	0.23	-
	SS	**0.69**	0.31	0.00	-
	ST	0.11	**0.66**	0.23	-

Chapitre 6

Études d'association : revue et perspectives

Le terme d'"étude d'association" regroupant plusieurs échelles et différents niveaux de traitement de l'information, nous avons souhaité, en premier lieu, clarifier cette problématique au travers d'une courte synthèse bibliographique. Comme nous le verrons dans cette première partie, la modélisation du déséquilibre de liaison (DL) s'impose progressivement comme un préalable indispensable à cette démarche. Nous proposons donc dans une seconde partie de ce chapitre une stratégie de test d'association fondée sur la modélisation des haplotypes proposée au chapitre précédent.

Synthèse bibliographique

A l'échelle d'une population d'étude aux bases génétiques larges, toute démarche de cartographie fine de gènes peut-être subsumée sous le concept d'"étude d'association". Ce dernier regroupe aussi bien des considérations étiologiques simples (caractères mendéliens) que complexes (caractères quantitatifs). Dans la littérature anglo-saxonne, le terme d'"étude d'association" revêt ainsi plusieurs acceptions : "association study", "association mapping", "gene mapping", "LD-mapping" ou encore "fine mapping". Ces notions sont souvent utilisées de manière interchangeable, confondant les différents objectifs et les différentes échelles d'étude qu'elles recouvrent. D'une manière générale, nous entendons ici par "étude d'association" toute approche dont le but est de détecter et/ou de localiser des variants génétiques causaux impliqués dans la variation d'un caractère d'intérêt à l'aide d'un échantillon d'individus **pour lesquels l'information généalogique n'est pas exploitable** (les relations d'apparentement étant trop "lâches" pour être valorisées). Les données utilisées pour mener cette étude se résument donc en un jeu de marqueurs caractérisant une région d'intérêt, des mesures phénotypiques et éventuellement des covariables (par exemple différents environnements

d'évaluation phénotypique, ou des proportions d'admixture). Cette région peut correspondre soit à un gène, soit à plusieurs gènes liés, soit à un chromosome entier, soit à une cartographie complète du génome. Quelle que soit l'échelle de l'étude, le but demeure le même : identifier la ou les zones significativement corrélées avec la variation du caractère au sein de la région.

En fonction de la nature des données, cette recherche peut avoir deux visées différentes :

- Existe-t-il une association entre la variabilité génétique de la région et la variation du caractère étudié ?
- Supposant *a priori* que la région contient un ou plusieurs variants génétiques causaux, quelle est, compte tenu de la densité de marqueurs disponible, la localisation la plus probable de ce ou ces variant(s) ?

Visées qui se distinguent à la fois sur le plan conceptuel et sur le plan statistique : Zollner and Pritchard (2005) proposent ainsi de distinguer la notion de "testing for association" de la notion de "fine mapping". Les développements récents dans cette discipline se répartissent habituellement selon ces deux objectifs qui se différencient essentiellement par la façon de déterminer le pouvoir résolutif de l'étude : dans le premier cas la résolution est déduite de mesures empiriques ou *ad hoc* sur la structure du déséquilibre de liaison (DL) dans l'espèce considérée, tandis que la deuxième approche vise à modéliser cette résolution conjointement à l'analyse de la variation du caractère. Et cette différence s'affirme dans la manière de traiter l'information fournie par les marqueurs le long de la région. Implicitement, cela sous-entend que l'enjeu ne réside peut-être pas dans la seule étude de la corrélation entre la variation phénotypique et la diversité génétique, mais aussi dans l'utilisation de cette dernière pour, sous certaines hypothèses, recouvrer la généalogie "manquante". Connaître celle-ci, n'est-ce pas retrouver le sens premier de toute étude de l'hérédité d'un caractère ?

L'approche "marqueur centrée"

Dans ce cas, on évalue marqueur par marqueur l'association entre le polymorphisme au marqueur et la variation phénotypique observée à l'aide de méthodes usuelles telles que l'ANOVA ou des modèles de régression (linéaire ou logistique, selon la nature des données phénotypiques, ou selon que le caractère ou le marqueur est choisi comme variable explicative). Dans le cadre de la régression linéaire, on peut facilement exprimer la relation entre les paramètres du modèle ajusté aux données, ceux du modèle génétique à un site causal putatif, et du DL entre ce site et le marqueur testé (voir Encadré 1). Cette approche, de par sa simplicité, à l'avantage de faciliter l'expression des statistiques de tests en fonction des paramètres d'intérêt - très souvent utile pour effectuer des calculs de puissances.

Cependant, lorsque le nombre de marqueurs augmente, on se trouve

Encadré 1 : Propriétés des modèles de régression marqueur par marqueur

Soient un échantillon de N individus, Y le vecteur de dimension N des phénotypes des individus, et M un marqueur caractérisé par m allèles dénotés $M_1, \ldots, M_m$, et de fréquences respectives, dans la population échantillonnée, $p_{M_1}, \ldots, p_{M_m}$. On considère alors les modèles linéaires suivant :

$$(H_0)\ Y = \mu + C\gamma + \epsilon \tag{6.1}$$

$$(H_1)\ Y = \mu + G\beta + C\gamma + \epsilon \tag{6.2}$$

$$(H_2)\ Y = \mu + A\alpha + C\gamma + \epsilon \tag{6.3}$$

où G est la matrice d'incidence représentant les génotypes observés au marqueur (potentiellement, G est une matrice de taille $N \times m(m+1)/2$), A la matrice d'incidence de taille $m \times N$ représentant les doses alléliques au marqueur, C une matrice de covariables et ϵ le terme d'erreur. Notons que si l'on suppose que l'effet génétique est additif, le modèle 6.2 est équivalent au modèle 6.3. Autrement dit, l'effet du génotype M_iM_j, dénoté β_{ij} se décompose simplement comme la somme des effets des allèles M_i et M_j : $\beta_{ij} = \alpha_i + \alpha_j$.
Considérons à présent un QTL bi-allélique Q/q de fréquences respectives p_Q et p_q dans la population. Soit a l'effet du génotype QQ au QTL, d celui du génotype Qq et $-a$ du génotype qq. Dès lors $\alpha_Q = a + (p_Q - p_q)d$ est l'effet moyen de substitution de l'allèle q par Q au QTL, $\delta_Q = 2d$ la déviation due à l'effet de dominance, et $\mu_Q = a(p_Q - p_q) + 2dp_Qp_q$ l'effet moyen du QTL dans la population. Enfin, on note $D_{M_iQ} = p_{M_iQ} - p_{M_i}p_Q$ le DL entre l'allèle Q au QTL et l'allèle M_i au marqueur M (p_{M_iQ} est la fréquence de l'haplotype M_iQ). Selon Fan *et al.* (2005), les coefficients des modèles de régression ci-dessus s'écrivent alors :

$$\begin{aligned}\beta_{ij} &= \mu_Q + \alpha_Q(D_{M_iQ}/p_{M_i} + D_{M_jQ}/p_{M_j}) \\ &\quad - \delta_Q D_{M_iQ} D_{M_jQ}/(p_{M_i}p_{M_j}) \\ \alpha_i &= \mu_Q/2 + \alpha_Q D_{M_iQ}/p_{M_i}\end{aligned}$$

On remarque que si $\delta_Q = 0$, c.a.d. s'il n'y a pas d'effet de dominance au QTL, on retrouve $\beta_{ij} = \alpha_i + \alpha_j$. Le degré de colinéarité entre le marqueur testé et le QTL étant bien représente ici par les composantes du déséquilibre de liaison entre les deux. D'autre part, dans le cas où $C = 0$ (modèle sans les covariables), les paramètres de non-centralité associés aux tests d'hypothèse H1 contre H0 et H2 contre H0 peuvent être approximés respectivement par (Fan *et al.* (2005)) :

$$\lambda_\beta \approx \frac{N}{\sigma_2}[\sigma^2_{ga}\Delta^2_{MQ} + \sigma^2_{gd}\Delta^4_{MQ}]$$

$$\lambda_\alpha \approx \frac{N\sigma^2_{ga}}{\sigma_2}\Delta^2_{MQ}$$

avec $\sigma^2 = \sigma^2_{ga} + \sigma^2_{gd} + \sigma^2_e$ est la variance totale, $\sigma^2_{ga} = 2p_Qp_q\alpha^2_Q$ la variance additive, $\sigma^2_{gd} = (p_Qp_q)^2\delta^2_Q$ la variance de dominance, et Δ^2_{MQ} la mesure du déséquilibre de liaison multiallélique entre le marqueur M et le QTL. Comme attendu, le test sera d'autant plus puissant que le marqueur est en déséquilibre de liaison (positif ou négatif) fort avec le QTL.

rapidement confronté au problème de tests multiples. Deux questions se posent alors : premièrement, tous les marqueurs sont-ils réellement informatifs ? Et pour ceux qui le sont, comment contrôler au mieux l'erreur de première espèce tout en préservant suffisamment de puissance pour les tests ? Le problème soulevé par la première question peut-être circonscrit par des processus amont de "filtre" sur les marqueurs :

- soit en sélectionnant les marqueurs qui capturent la plus grande diversité génétique dans la région considérée. Dans ce cas on utilise uniquement l'information fournie par le DL dans la région pour sélectionner les marqueurs. Cette sélection peut se faire à l'aide de différentes stratégies en fonction de la structure du DL observée : en supposant une structure en bloc Zhang *et al.* (2002, 2004a, 2005), en capturant les "motifs" de mutations les plus prédictifs (sous-entendu des génotypes aux autres marqueurs) Halperin *et al.* (2005); Schwartz (2004), ou par des approches fondées sur l'ACP Meng *et al.* (2003); Horne and Camp (2004). Cependant, cela présuppose qu'on préserve, simultanément, suffisamment de puissance pour les tests d'association. Or, cette hypothèse est encore sujette à controverse (voir en particulier Zhang *et al.* (2002, 2004b); de Bakker *et al.* (2005)).
- soit en combinant l'ensemble de l'information, marqueurs et phénotypes, dans un processus préliminaire de tests d'exclusion. Notons que des procédures d'exclusion avaient déjà été utilisées très tôt en cartographie de QTL classique (voir notamment Morton (1955) et plus récemment Boddeker *et al.* (2001)). Dans le cadre des études d'association, la procédure d'exclusion proposée par Hoh *et al.* (2000) - et reprise ultérieurement dans un autre article Hoh *et al.* (2001) ainsi que dans une revue Hoh and Ott (2003) - offre une stratégie assez souple pour s'adapter à différentes mesures d'association entre marqueurs et phénotypes (voir Encadré 2).

La deuxième question est de loin la plus délicate. La méthode de Bonferroni conduisant à des corrections trop conservatrices, des méthodes alternatives plus opérationnelles ont été suggérées récemment. Une première solution consiste à prendre en compte la non indépendance des tests du fait du DL au sein de la région étudiée. Autrement dit, les variables testées n'étant pas indépendantes, il serait préférable de définir une correction qui prenne en compte le degré de corrélation entre ces variables. De manière similaire à la sélection de marqueurs par ACP, Cheverud (2001) a proposé d'obtenir cette correction à l'aide d'une ACP de la matrice de covariance empirique entre marqueurs. Celle-ci permet de calculer le nombre effectif de marqueurs indépendants dans la région, donnant ainsi une approximation du nombre de tests indépendants. Toutefois, si le DL dans la région est faible, cette approche conduit à des valeurs voisines du nombre initial de marqueurs ; et le problème demeure entier lorsque ce nombre est élevé. La solution la plus satisfaisante à ce jour est sûrement la méthode dite du "false discovery rate"

Encadré 2 : Procédure préliminaire de sélection de marqueurs.

Soit un ensemble de M marqueurs parmi lequel on souhaite sélectionner le sous-ensemble le plus "informatif" pour effectuer des analyses ultérieures. Cette sélection peut être réalisée à l'aide de l'algorithme suivant (Hoh *et al.* (2000)) :

- **Étape 0** : calculer et ordonner les statistiques de test obtenues à chaque marqueur, notées $t_1 \leq \ldots \leq t_M$.
- **Étape 1** : échantillonner B_1 jeux de données à partir du jeu de données initial par "bootstrap". Pour chacun d'eux, calculer et ordonner les statistiques de test, notées $t_{1b_1} \leq \ldots \leq t_{Mb_1}$.
- **Étape 2** : permuter les données phénotypiques par rapport aux marqueurs dans le jeu de données initial ainsi que dans chacun des B_1 bootstraps précédents. Cela correspond à un deuxième niveau de bootstrap sous l'hypothèse nulle (pas d'association). Calculer et ordonner les statistiques de test obtenues dans chacune des B_2 permutations obtenues à partir du jeu de données initial, notées $t_{1b_2} \leq \ldots \leq t_{Mb_2}$ et à partir des B_1 échantillons, notées $t_{1b_1b_2}, \ldots, t_{Mb_1b_2}$.
- **Étape 3** : soit $j = 1$.
- **Étape 4** : retirer les $j-1$ marqueurs qui présentent la plus petite statistique de test à l'étape 0. Pour les $K-j+1$ marqueurs restants, calculer la somme de leurs statistiques de tests dans le jeu de données initial, $s_{[M-j+1]}$, puis dans les B_2 jeux permutés, $s_{[M-j+1]b_2}$. De manière similaire, retirer les $j-1$ marqueurs qui présentent la plus petite statistique de test à l'étape 1 et calculer pour les $K-j+1$ marqueurs restants les sommes correspondantes, $s_{[M-j+1]b_1}$ et $s_{[M-j+1]b_1b_2}$.
- **Étape 5** : Évaluer les quantités suivantes :

$$p_{[M-j+1]} = \frac{\sharp\{s_{[M-j+1]b_2} \geq s_{[M-j+1]}\}}{B_2}$$
$$p_{[M-j+1]b_1} = \frac{\sharp\{s_{[M-j+1]b_1b_2} \geq s_{[M-j+1]b_1}\}}{B_2}$$

- **Étape 6** : présélectionner les $M-j+1$ marqueurs dans le jeu de données initial si $p_{[M-j+1]} \leq \alpha$, un seuil prédéfini (par exemple $\alpha = 0.05$), sinon $j = j+1$ et retourner à l'étape 4. Appliquer la même règle de présélection aux B_1 jeux de données échantillonnés en utilisant $p_{[M-j+1]b_1}$.
- **Étape 7** : calculer la fréquence de présélection pour chaque marqueur à la fois dans le jeu de données initial et dans les B_1 jeu de données échantillonnés. Enfin, sélectionner les marqueurs dont la fréquence cumulée est supérieure à une seuil préétabli, par exemple 50%.

Enfin, au lieu de retirer successivement les marqueurs dont la statistique de test est la moins significative (mode "backward"), la procédure de sélection peut-être modifiée en incluant progressivement les marqueurs dont les statistiques de test sont les plus élevées (mode "forward").

(FDR) introduite en 1995 par Benjamini and Hochberg (1995) (et utilisée pour la première fois en cartographie génétique par Weller *et al.* (1998) dans le cadre de la détection multiple de QTL). L'efficacité de cette méthode a été illustrée par Sabatti *et al.* (2003) dans le contexte de cartographie fine de caractères discrets. Récemment, l'article de Benjamini and Yekutieli (2005) offre une évaluation rigoureuse, et non moins prometteuse, de l'utilisation du FDR dans les problématiques de cartographie de gènes. Pour un définition du FDR nous renvoyons à l'Encadré 3.

Encadré 3 : "False discovery rate"

Afin de résoudre le paradoxe entre un contrôle stricte de l'erreur de première espèce et la nécessité de préserver une puissance suffisante, un critère plus fonctionnel a été introduit en 1995 par Benjamini and Hochberg (1995). Supposons que l'on dispose de M marqueurs pour lesquels les tests d'association ont conduit à M p-values. Soit Q la proportion de faux positifs parmi les M_0 tests déclarés significatifs. Le "False discovery rate" (FDR) est défini comme étant la valeur attendue de Q. Le FDR peut alors être contrôlé à l'aide d'une procédure dénommée la procédure BH (pour Benjamini et Hochberg Benjamini and Hochberg (1995)), fondée uniquement sur la distribution des M p-values :

- **Étape 0** : Ordonner les p-values par ordre croissant, $p_{(1)} \leq \ldots \leq p_{(M)}$.
- **Étape 1** : Soit $i = M$.
- **Étape 2** : Si $p_{(i)} \leq q.i/M$, alors $k = i$. Sinon $i = i - 1$.

où q et le niveau de contrôle du FDR (par exemple $q = 0.05$). On espère donc au plus une proportion q de faux positifs dans les k premières p-values ainsi sélectionnées. D'autres procédures de contrôle ont été introduites par la suite et ont été discutées en détail par Benjamini and Yekutieli (2005). A l'aide de simulations, ces derniers ont montré que la procédure BH est particulièrement adaptée au problème de tests multiples dans le cadre de la cartographie de QTL.

Enfin, en ne considérant que l'information individuelle aux marqueurs, les approches marqueur par marqueur présupposent une relation causale simple entre les facteurs génétiques et la variation phénotypique. Or il est plus que vraisemblable que l'architecture génétique causale soit plus complexe, impliquant plusieurs locus avec éventuellement des effets épistatiques (Terwilliger and Weiss (1998); Hugot *et al.* (2001); Lohmueller *et al.* (2003); Hoh and Ott (2003)). Cette limitation peut-être en partie résolue en incluant simultanément dans le modèle plusieurs marqueurs (Fan *et al.* (2005)) et éventuellement les termes d'interaction correspondants. A l'instar de la détection de QTL classique, dans le cadre de la régression linéaire, des procédures de construction de modèle par sélection pas à pas de type "forward" et/ou "backward" peuvent ainsi être mises en œuvre pour identifier le modèle optimal. Bien qu'attractives, car simples à réaliser, ces procédures peuvent conduire à sélectionner des configurations de marqueurs non optimales. De plus le choix de leurs paramètres de contrôle n'est pas toujours aisé.

Mais surtout, en restreignant l'analyse à une démarche séquentielle, marqueur par marqueur, ces méthodes ignorent une information précieuse,

potentiellement contenue dans le jeu de données : l'histoire qui sous-tend conjointement la diversité génétique et la variation phénotypique.

L'approche "haplotype centrée"

Dans ce cas, l'information aux marqueurs est résumée en une liste d'haplotypes - c'est à dire une séquence singulière de polymorphismes contigus - qui représentent les allèles le long de la région étudiée. Notons que le nombre d'haplotypes observés est négativement corrélé au DL moyen au sein de la région (un DL fort se traduit par un petit nombre d'haplotypes). Dans un premier temps, en s'appuyant sur l'idée que la configuration multilocus capturée par les haplotypes représente mieux l'architecture allélique au sein de la région, pourquoi ne pas évaluer directement l'association entre la diversité haplotypique observée et la variation phénotypique ? Bien que des études théoriques aient suggéré que de telles approches n'étaient pas nécessairement plus puissantes que les modèles marqueur par marqueur Nielsen *et al.* (2004); Fan *et al.* (2005), des études réalisées à partir de données réelles ont mis en avant leur utilité pour détecter des associations que l'approche marqueur par marqueur seule n'aurait pu révéler Lu *et al.* (2003); Hagenblad *et al.* (2004); Buntjer *et al.* (2005), notamment lorsque l'on cherche à identifier un variant causal non observé Johnson *et al.* (2001); Gabriel *et al.* (2002) - sous-entendu non caractérisé par un marqueur.

Mais cette démarche n'est valable que pour des régions de petite taille (longueur relative au DL), pour lesquelles le nombre d'haplotypes observés est limité et ne "contrarie" pas trop la puissance des tests en abaissant dangereusement le nombre de degrés de liberté de la résiduelle. D'autre part, bien qu'elle utilise l'information conjointe aux marqueurs, elle n'en exploite pas vraiment tout le potentiel. Surtout, elle n'extrait pas explicitement la dimension supplémentaire apportée par la notion d'haplotype : la généalogie "cachée".

La possibilité d'intégrer dans les tests d'association l'histoire évolutive qui "relie" les haplotypes - mathématiquement cette histoire peut être visualisée par un graphe orienté, ou plus simplement, en ignorant la recombinaison, par un arbre - offre sur le plan conceptuel un atout majeur. L'idée maîtresse est ici guidée par l'hypothèse suivante : si on suppose qu'un allèle causal est autrefois apparu par mutation, celui-ci doit être alors "lié" à un contexte allélique particulier aux locus avoisinants. Autrement dit, la mutation causale survenant dans un haplotype singulier, c'est à dire dans un lignage particulier de la généalogie, les individus aujourd'hui porteurs de cette mutation sont probablement plus proches dans l'arbre que par le simple fait du hasard (hypothèse représentée schématiquement dans la Figure 6.1). L'étude d'association idéale se décomposerait ainsi en deux étapes : i) utiliser l'information multilocus afin de regrouper les haplotypes selon la généalogie la plus vraisemblable et ii) intégrer ce résultat afin de modéliser au mieux la relation "historique"

entre haplotypes et variation phénotypique.

L'idée d'intégrer la généalogie dans les tests d'association a été pour la première fois introduite en génétique humaine par Templeton *et al.* (1987). Dans une série de 5 articles s'étalant de 1987 à 1995, Templeton et son équipe (Templeton *et al.* (1987, 1988, 1992); Templeton and Sing (1993); Templeton (1995)) posèrent les bases de cette démarche en deux étapes. Tout d'abord, la généalogie des haplotypes est inférée à l'aide d'un cladogramme construit par des méthodes usuelles de phylogénie (par exemple, par maximum de parsimonie Aquadro *et al.* (1986); Templeton *et al.* (1992); Clement *et al.* (2000), ou par clustering hiérarchique à partir d'une matrice de distances entre haplotypes Saitou and Nei (1987)). Une fois le cladogramme construit, sa topologie est utilisée pour guider une série de tests d'hypothèse en groupant successivement les classes d'haplotypes voisines entre elles Templeton *et al.* (1987) (les tests étant évalués par ANOVA ou modèle de régression). Une méthode plus facilement automatisable et conduisant à des résultats similaires a été récemment développée par la même équipe Templeton *et al.* (2005); Posada *et al.* (2005) (voir Encadré 4). L'un des avantages de cette méthode est de réduire l'espace des tests d'hypothèse en le contraignant à la topologie de l'arbre : le nombre de tests effectués devient alors fonction du nombre de branches considérées dans l'arbre (au total, autant que d'haplotypes moins un) et non plus du nombre total de combinaisons possibles entre haplotypes (et du nombre de marqueurs). Cette stratégie permet ainsi d'explorer finement la décomposition de la variance phénotypique en fonction des classes haplotypiques observées, tout en tâchant de préserver le maximum de puissance pour les tests. Toutefois, bien qu'elle offre un cadre séduisant pour étudier des architectures génétique causales complexes, cette approche ne permet pas de localiser directement les mutations causales, à moins de réaliser des analyses supplémentaires au cas par cas, en s'appuyant sur la structure des effets significatifs détectés dans les sous-arbres. D'autre part, la construction de cladogramme n'est pas toujours aisée, notamment lorsque qu'il devient difficilement tenable de négliger l'effet de la recombinaison par rapport à la mutation dans la région étudiée.

Plus récemment, dans le contexte de cartographie fine de gènes de maladie, l'approche développée par Durrant *et al.* (2004) tâche de prévenir les effets dûs à la recombinaison à l'aide de fenêtres glissantes le long de la région. Dans chaque fenêtre, la généalogie des haplotypes est établie par une méthode classique de clustering hiérarchique et chaque partition dans l'arbre est testée successivement afin d'identifier celle qui s'ajuste le mieux aux données phénotypiques - dans l'esprit cette méthode doit beaucoup à Templeton *et al.* (1987). Cependant, le paramétrage de la fenêtre glissante peut être délicat (longueur fixe ou variable le long de la région ? ou comment prendre en compte l'hétérogénéité des patrons de DL) et pour les régions de grande taille, le problème de tests multiples est à nouveau posé. Sur le

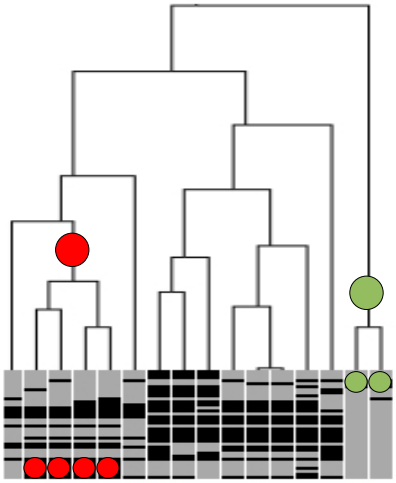


FIG. 6.1 – Illustration schématique et hypothétique d'une généalogie de 16 haplotypes dans une région causale pour un caractère cible. Chaque feuille de l'arbre représente un haplotype particulier échantillonné et les branches leurs relations ancestrales. Les deux cercles de couleurs indiquent deux événements de mutation indépendants. Ces mutations sont supposées contribuer à la variation du caractère, et sont héritées par les haplotypes situés à la terminaison des branches infra. Chacune des mutation est ainsi "enchâssée" dans un groupe d'haplotypes, ces derniers ayant "tendance" à se regrouper au sein d'un même "cluster".

Encadré 4 : "Tree Scanning"

Supposons que l'on ait construit un cladogramme reliant les haplotypes observés dans la région étudiée. Templeton *et al.* (2005) proposent alors une procédure itérative qui se décompose en quatre grandes étapes :

- **Étape 1** : pour chaque branche du cladogramme
  - Grouper les haplotypes en deux classes, nommées A et B, en "coupant" la branche courante.
  - Tester l'association entre le phénotype et les deux pseudo-classes alléliques A et B (par exemple, par une simple ANOVA).
  - Évaluer la p-value du test en permutant les données phénotypiques et les données haplotypiques (hypothèse nulle : pas d'association). Corriger éventuellement la p-value pour prendre en compte les tests multiples.
- **Étape 2** : Si au moins une branche a une p-value significative, aller à l'étape 3. Sinon s'arrêter.
- **Étape 3** : Pour chaque branche dont le test a été déclaré significatif, "partitionner" les haplotypes en deux classes, A et B, en "coupant" cette branche. Pour une des deux classes alléliques A ou B (ci-après A) :
  - "Partitionner" les haplotypes en trois classes, nommées $A^{'}$, $A^{''}$ et B, en "coupant" une branche dans le sous-arbre correspondant au pseudo-allèle A.
  - Retirer des analyses suivantes les individus ne portant que l'allèle B.
  - Tester l'association entre le phénotype et la partition courante.
  - Évaluer la p-value du test par permutation.
- **Étape 4** : Continuer si au moins une branche dans le sous-arbre a une p-value significative.

Des raffinements sont possibles, comme l'exclusion du processus d'analyse des haplotypes rares, ainsi que le choix *a priori* des coupures dans l'arbre.

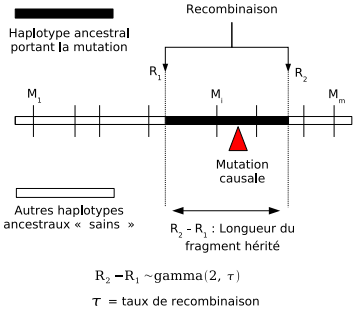
premier point, les auteurs ont suggéré d'inférer au préalable la structure en bloc de la région, puis d'ajuster la taille de la fenêtre en fonction du découpage obtenu. Quant au problème de tests multiples, le débat entre corrections de type Bonferroni, permutations ou FDR reste ouvert. Néanmoins, la souplesse de cette méthode et sa simplicité de mise en œuvre lui confère une réelle "attractivité".

Ces premières méthodes réalisent une avancée certaine dans la conduite des tests d'association. Mais en s'appuyant sur des techniques de clustering standard, elles laissent en suspens la question épineuse de la modélisation conjointe de la diversité génétique et de la variation phénotypique. A la question "comment modéliser la relation entre généalogie et phénotype", le travail de McPeek and Strahs (1999) a apporté, en 1999, une première réponse statistique qui aura un écho très favorable dans les années suivantes. Bien que limitée à l'étude de dispositifs cas-contrôle, cette méthode s'articule autour d'un concept novateur et fort : si un allèle causal, prédisposant à une maladie, est autrefois apparu par mutation dans un haplotype particulier, dit ancestral, les haplotypes des individus malades observés de nos jours doivent présenter un "déséquilibre ancestral" de part et d'autre du locus porteur de la mutation causale. Dès lors, localiser la position, le long de la région, autour de laquelle la diversité haplotypique sera plus restreinte chez les individus malades par rapport à celle observée chez les individus sains, conduira vraisemblablement à identifier le locus causal. L'aspect innovant de l'approche de McPeek and Strahs (1999) n'est bien sûr pas dans cette manière de présenter le problème (déjà évoqué supra en d'autres termes, nous tenions seulement à le repréciser ici dans le cadre expérimental particulier de McPeek and Strahs (1999)), mais dans la réponse statistique qu'elle lui donne (voir Encadré 5).

Prolongée par les travaux de Morris *et al.* (2000); Liu *et al.* (2001), cette méthode a inspiré par la suite d'autres approches plus élaborées, comme celles de Lam *et al.* (2000); Rannala and Reeve (2001); Morris *et al.* (2002) qui l'élargirent à des structures généalogiques plus complexes. Récemment, la méthode de Zollner and Pritchard (2005) réalise une synthèse remarquable de ces approches dans un modèle plus globale reposant sur la théorie de la coalescence. L'idée demeure la même (reconstruction locale de la généalogie la plus vraisemblable puis évaluation du modèle phénotypique) mais chez Zollner and Pritchard (2005) la modélisation conjointe de la généalogie et de la variation phénotypique offre une plus grande souplesse que celle proposée par ses prédécesseurs. Toutefois, côté machine, ces méthodes souffrent encore par leurs temps de calcul conséquents. D'autre part, si en génétique humaine les hypothèses sous-jacentes au modèle de coalescence sont acceptables, l'application de ces méthodes en génétique végétale est encore limitée par la complexité et l'opacité des scénarios évolutifs - bien que pour certaines espèces des travaux aient permis de commencer à éclaircir ces mécanismes Buckler and Thornsberry (2002); Rafalski and Morgante (2004).

Encadré 5 : Cartographie fine par généalogie en étoile

L'approche de McPeek and Strahs (1999) repose sur une vision originale du DL dans les dispositifs cas-contrôle : au lieu de considérer des statistiques calculées entre paires de marqueurs, les auteurs proposent d'évaluer autour d'un site donnée (observé ou non), la variabilité chez les patients malades de la longueur de l'haplotype ancestral dans lequel la mutation est supposée être apparue. L'hypothèse étant qu'à l'inverse des individus sains, les individus malades présentent de part et d'autre de la mutation causale, un "excès" de cet haplotype ancestral.
Cette hypothèse est illustrée de manière schématique dans la figure ci-dessous.



McPeek and Strahs (1999) supposent alors que les configurations alléliques observées entre haplotypes malades sont indépendantes une fois connue leur composition "ancestrale". On parle alors de généalogie en étoile : tous les haplotypes malades sont supposés avoir connu une histoire différente (en terme d'événements de recombinaison et de mutation) depuis l'apparition de la mutation dans l'haplotype ancestral (racine de l'étoile). Autrement dit, une fois identifiés les points de recombinaison "encadrant" la mutation (R_1 et R_2 dans la figure ci-dessus), la généalogie se décompose en autant de branches que d'haplotypes malades qui rayonnent autour de l'unique haplotype ancestral causal.
Le point fort de cette approche est de permettre une analyse pas à pas des intervalles de marqueurs afin d'identifier, chez les individus malades, la position la plus probable de la mutation causale. A chaque position (observée ou non), une vraisemblance est maximisée conduisant à l'estimation de l'haplotype ancestral causal et de sa longueur moyenne autour de la position testée. L'hypothèse de la généalogie en étoile permet d'écrire cette vraisemblance comme le simple produit des probabilités d'observations des haplotypes malades en fonction des paramètres du modèle, incluant éventuellement un paramètre mimant un taux de mutation. Le processus de recombinaison entre l'haplotype ancestral et les "autres" haplotypes (sous-entendu ceux observés chez les patients sains) étant supposé se comporter comme un processus Markovien le long de la région, la probabilité de chaque haplotype malade est calculée à l'aide d'une chaîne de Markov de premier ordre en utilisant l'information complète aux marqueurs (à droite et à gauche de la position considérée).
Il est intéressant de remarquer que dans les applications de cette méthode conduite par McPeek and Strahs (1999), quelque soit la position testée le long de la région, le même haplotype ancestral a été détecté. Enfin, Liu *et al.* (2001) a mis en œuvre une approche similaire dans le cadre Bayésien en intégrant dans le modèle la position de la mutation causale, ainsi qu'une possible hétérogénéité des haplotypes ancestraux chez les individus malades (plusieurs généalogies en étoile peuvent être ainsi considérées simultanément).

Enfin, ces méthodes ont été pour la plupart développées dans le cadre particulier des études cas-contrôle en génétique humaine, et par conséquent ne sont pas toujours aisément transposables à d'autres configurations expérimentales.

Aussi, des approches plus pragmatiques et plus transversales ont été développées ces dernières années. Tout d'abord une méthode de clustering multidimensionnel a été proposée par Molitor *et al.* (2003) afin d'explorer simultanément les données haplotypiques et phénotypiques. Autour d'une position donnée, on cherche ainsi à "clusteriser" les haplotypes à la fois sur la base de leur proximité génétique et phénotypique. Cette méthode a notamment été utilisée par Hagenblad *et al.* (2004) afin d'étudier l'association entre la précocité de floraison et la structure des haplotypes dans des régions candidates chez *arabidopsis*. D'autre part, les données haplotypiques présentant souvent des dimensions importantes, que ce soit en nombre de marqueurs et/ou nombre d'individus, des techniques inspirées par les méthodes de "data-mining" ont également été développées à la fois pour la cartographie fine de gènes de maladie et de QTL. Les plus remarquées à ce jour sont celles de Toivonen *et al.* (2000); Onkamo *et al.* (2002); Toivonen *et al.* (2004); Li and Jiang (2005).

Bilan

Les études d'association haplotype "centrées" ont suscité un vif intérêt ces dernières années et quelques revues font état de leurs avantages potentiels chez les végétaux, tant en terme de puissance que de lisibilité, relativement aux approches marqueur par marqueur (voir par exemple Flint-Garcia *et al.* (2003); Buntjer *et al.* (2005)). Cependant, dans les populations diploïdes les techniques de génotypage permettent rarement d'obtenir la phase des individus (à chaque marqueur, l'origine paternelle et maternelle des allèles est inconnue). Sous certaines hypothèses, il est possible de reconstruire les haplotypes à partir des données génotypiques seules (pour une revue des méthodes et leur comparaison, nous renvoyons à l'article de Niu (2004)). Mais la façon d'intégrer ces méthodes de reconstruction des haplotypes dans les études d'association demeure encore sujet à controverse. Si certains auteurs préconisent une approche en deux temps - i) reconstruction probabiliste des haplotypes et ii) utilisation de la reconstruction la plus probable pour le reste des analyses Templeton *et al.* (2005); Zollner and Pritchard (2005) (notons que la reconstruction probabiliste peut aussi être intégrée dans les tests Schaid *et al.* (2002)) - d'autres ont privilégié des approches où l'incertitude sur la phase est directement modélisée McPeek and Strahs (1999); Liu *et al.* (2001); Morris *et al.* (2003). Dans ce dernier cas les résultats paraissent mitigés : si Morris *et al.* (2004) suggèrent un léger gain comparé aux méthodes en deux temps, l'étude de Lu *et al.* (2003) conduit à la conclusion inverse. Enfin, pour les méthodes de cartographie fine complexes comme celle de Zollner and Pritchard (2005), la modélisation

de l'incertitude de phase pourrait se révéler être un vrai challenge.

D'autre part, aucune étude à ce jour n'intègre à la fois ces modélisations novatrices "haplotypes-phénotypes" et le problème lié à la structuration génétique. Après avoir au préalable établi les paramètres de la structure, Zollner and Pritchard (2005) propose bien une stratégie par permutation pour contrôler les associations fallacieuses, mais la procédure qu'il décrit n'est réalisable que pour des caractères discrets (par exemple. malade ou sain), et le coût machine des permutations dans ce cadre est de toutes manières trop conséquent pour imaginer une étude à grande échelle. Seuls les développements de Satten *et al.* (2001) et Hoggart *et al.* (2003) (repris et enrichis dans un second article Hoggart *et al.* (2004)) offrent des méthodes intégrant la modélisation de la structuration génétique dans les tests d'association. Mais le premier se limite à un modèle de mélange simple, qui est de toute évidence utopique pour la plupart des collections chez les végétaux, et si le cadre Bayésien du second est très proche du logiciel structure(Pritchard *et al.* (2000a); Falush *et al.* (2003)), il n'offre qu'une stratégie élégante pour intégrer l'incertitude sur les paramètres estimés de la structure dans des tests marqueur par marqueur. De plus, chaque étape d'analyse ayant sont propre lot de complexités, ces méthodes en une seule étape sont peut-être encore trop ambitieuses pour être appliquées sur des jeux de données pour lesquels la nature des mécanismes responsables des effets visibles de la structuration, ainsi que ceux impliqués dans le déterminisme des caractères étudiés, demeure encore incertaine - ce qui est le cas chez nombre d'espèces végétales.

C'est l'une des raisons pour lesquelles les premières études d'association dans des collections structurées chez les végétaux ont été réalisées à l'aide de modèles de régression linéaire (voir par exemple chez le maïs Thornsberry *et al.* (2001); Andersen *et al.* (2005); Camus-Kulandaivelu *et al.* (2006)). L'idée en est assez simple : les paramètres de la structure capturés par les proportions d'admixture des individus sont inclus dans le modèle de régression comme covariables. Ce modèle, avec les proportions d'admixture seules, forme ainsi l'hypothèse nulle par rapport à laquelle les tests d'association sont évalués (Pritchard *et al.* (2000b)). Idéalement, cette stratégie devrait permettre de corriger les corrélations "artéfactuelles" entre la variation phénotypique et le polymorphisme testé (voir Encadré 6). Sous cette hypothèse, nous introduisons dans la partie suivante une nouvelle approche pour conduire des tests d'association chez les végétaux qui tente d'allier les avantages d'une modélisation multilocus de la diversité génétique à la flexibilité des tests par régression linéaire.

Encadré 6 : Structure génétique et covariance "artéfactuelle"

Nous reprenons ici les notation de l'Encadré 1. A présent on suppose que la population d'étude est structurée en deux sous populations en proportion π et $1-\pi$. Les fréquences des allèles au QTL dans les deux sous-populations sont notées p_{Q1}, p_{Q2} et p_{q1}, p_{q2}. De manière similaire, celles des allèles au marqueur sont notées p_{M_i1}, p_{M_i2}, $i = 1, \ldots, m$. Notons que $p_Q = \pi p_{Q1} + (1-\pi)p_{Q1}$ et $p_{M_i} = \pi p_{M_i1} + (1-\pi)p_{M_i2}$. On suppose que le "vrai" modèle génétique s'écrit :

$$Y = Z + G + e$$

où

$$G = \begin{cases} a & \text{pour le génotype QQ} \\ d & \text{pour le génotype Qq} \\ -a & \text{pour le génotype qq} \end{cases}$$

et

$$Z = \begin{cases} \mu_1 & \text{pour les génotypes de la sous-population 1} \\ \mu_2 & \text{pour les génotypes de la sous-population 2} \end{cases}$$

La différence de moyenne entre les deux sous-populations pouvant résulter des différences de fréquences alléliques au QTL, mais aussi d'un fond génétique différent. Sous ce modèle on montre que :

$$\mathsf{Cov}(Y, X_{ij}) = H_{ij}(\Delta_{Pop} + \Delta_{QTL})$$

où X_{ij} est la variable indicatrice qui vaut un si le génotype au marqueur M est M_iM_j, zéro sinon, $H_{ij} = 1$ si $i = j$, sinon $H_{ij} = 2$, et :

$$\begin{aligned} \Delta_{Pop} &= \pi(1-\pi)(\mu_1 - \mu_2)(p_{M_i1}p_{M_j1} - p_{M_i2}p_{M_j2}) \\ \Delta_{QTL} &= (p_{M_i}p_{M_j}\mu_Q + 2\alpha_Q(D_{M_iQ}p_{M_i} + D_{M_jQ}p_{M_j}) - \delta_Q D_{M_iQ}D_{M_jQ}) \end{aligned}$$

où à présent le terme de déséquilibre de liaison s'écrit :

$$\begin{aligned} D_{M_iQ} = \pi D_{M_iQ1} &+ (1-\pi)D_{M_iQ2} \\ &+ \pi(1-\pi)(p_{M_i1} - p_{M_i2})(p_{Q1} - p_{Q2}) \end{aligned}$$

avec D_{M_iQz} le déséquilibre de liaison entre l'allèle Q et l'allèle M_i dans la sous-population $z = 1, 2$.

Ainsi la corrélation entre le génotype au marqueur et la variation du caractère est "augmentée" par la structuration, à la fois par un premier terme, Δ_{Pop} qui rend compte du contraste phénotypique entre les deux sous-populations et implicitement par le deuxième terme, Δ_{QTL}, via le déséquilibre de liaison "artéfactuel", D_{M_iQ}, entre les allèles au marqueur et le QTL.

Dans le cas idéal où les covariables du modèle de régression, C, indiqueraient sans erreurs et sans ambiguïtés l'origine des individus, l'artéfact en revanche "disparaîtrait" : nous serions en effet capable de décomposer la covariance $\mathsf{Cov}(Y, X_{ij})$ en deux composantes distinctes pour chacune des deux sous-populations.

Haplotypes ancestraux et études d'association

Nous esquissons dans cette partie une manière d'intégrer la modélisation du DL proposée au chapitre précédent dans la conduite de tests d'association. Nous présupposons donc une hypothèse évolutive forte : la structure en haplotypes ancestraux n'a un sens que si les haplotypes observés sont supposés dériver d'un petit nombre d'haplotypes fondateurs. L'événement de fondation est supposé également avoir eu lieu dans un passé "pas trop lointain" de manière à ce que son effet puisse encore être détectable aujourd'hui. Cette hypothèse réduit le champ d'application de notre approche à certaines espèces chez qui un récent et fort goulot d'étranglement a pu être documenté. Ce peut-être le cas, par exemple, des espèces végétales qui ont été domestiquées par l'homme, tel le maïs.

Cette hypothèse est schématisée dans la Figure 6.2. Dans cette figure la mutation causale est supposée être apparue avant le goulot d'étranglement dans un haplotype particulier, qui aurait réussi à "franchir" le goulot - soit du simple fait du hasard, soit par l'avantage sélectif apporté par la mutation. Vu sous un angle idéal, les haplotypes observés de nos jours et porteurs de la mutation devraient présenter autour du site causal un "excès" de l'haplotype ancestral dans lequel la mutation est apparue. Ainsi, en supposant que les événements de recombinaison et de mutation depuis le goulot d'étranglement n'aient pas "trop détériorés" la structure ancestrale, déterminer cette dernière devrait permettre d'identifier la mutation avec une précision fonction du nombre de recombinaisons accumulées au fil des générations de part et d'autre de cette mutation causale.

Notre approche s'articule donc autour de deux étapes : i) inférer la structure en haplotypes ancestraux dans la région étudiée, ii) utiliser la structure obtenue pour guider les tests. Chez les végétaux, les études d'association sont généralement réalisées au niveau d'une région génique incluant un ou éventuellement plusieurs gènes, mais souvent petite à l'échelle et du chromosome et du génome. L'hypothèse d'une structure ancestrale homogène au sein de la région n'est donc pas ici trop stricte (nous discuterons ultérieurement comment cette hypothèse peut être relâchée dans le cas de plus grandes régions d'étude). Par souci de flexibilité, nous avons préféré séparer le test, à proprement parler, de la modélisation de la diversité génétique au sein de la région. Cela notamment pour faciliter l'inclusion de variables supplémentaires dans les modèles de test (e.g. la structuration de la population).

Dans un premier temps, nous décrivons comment la structure en haplotypes ancestraux peut-être intégrée dans une démarche de test marqueur par marqueur dans le cadre d'un modèle linéaire standard. Puis, nous montrons comment effectuer des analyses "pas à pas", en calculant les probabilités qu'un marqueur "non observé", à une position donnée, appartienne aux différentes catégories ancestrales. Enfin nous étudions par simulation les propriétés de notre approche en la comparant à une simple stratégie de

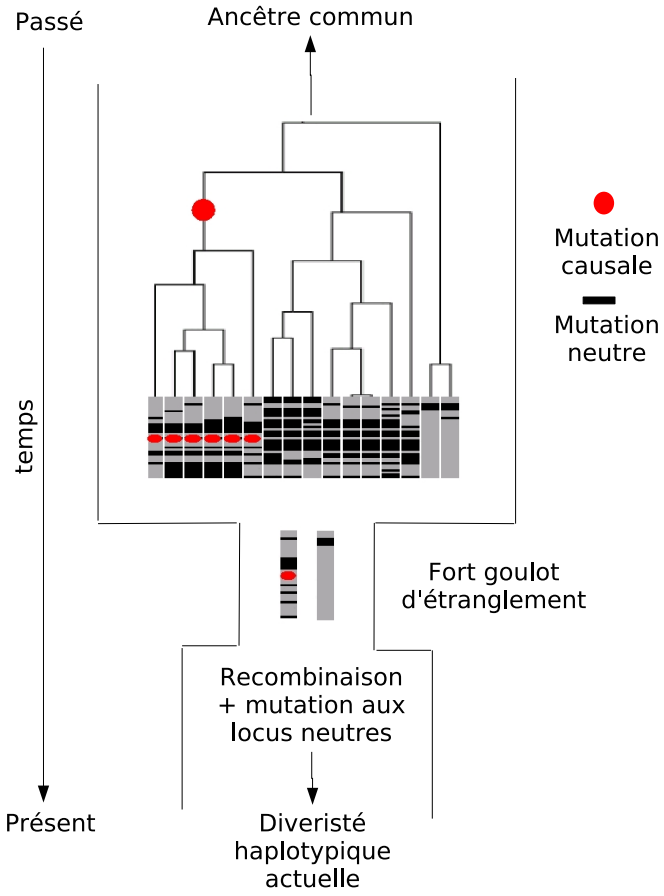


FIG. 6.2 – Illustration schématique de l'apparition d'une mutation causale *ante* goulot d'étranglement. Le goulot d'étranglement conduit à une nouvelle population fondée seulement à partir d'individus porteurs de l'haplotype causal et d'un autre haplotype. Si les événements de recombinaison et de mutation n'ont pas trop altéré les haplotypes ancestraux, les haplotypes actuels porteurs de la mutation présenteront, autour du site causal, un "excès" de l'haplotype ancestral correspondant.

test marqueur par marqueur.

Modèle

Tout d'abord, nous supposons que $K > 1$ haplotypes ancestraux ont été préalablement identifiés à l'aide de l'algorithme BSAH. Nous réutilisons ici la notation du chapitre précédent, à savoir :

- X est la matrice des haplotypes de taille $N \times M$, où N est le nombre d'individus haploïdes et M le nombre de marqueurs bi-alléliques.
- A est la matrice de taille $K \times M$ représentant les K haplotypes ancestraux détectés dans la région. L'état A_{kj} indique l'allèle porté par l'haplotype ancestral $k = 1, \ldots, K$ au marqueur $j = 1, \ldots, M$.
- Z_{ij} est la variable cachée qui indique l'indice de l'haplotype ancestral au marqueur $j = 1, \ldots, M$ pour l'haplotype $i = 1, \ldots, N$, et dont la distribution *a posteriori*, notée $\mathsf{Pr}(Z_{ij} = k)$, est obtenue à la dernière itération de l'algorithme BSAH.

Par souci de clarté dans la notation, nous considérons donc N individus haploïdes à M marqueurs bi-alléliques (la généralisation au cas diploïde - sous l'hypothèse que la phase est connue - et multiallélique s'effectuant sans difficultés majeures). Enfin le vecteur Y de taille N désigne le vecteur des phénotypes.

Marqueur observé

A chaque marqueur $j = 1, \ldots, M$, nous proposons d'utiliser le modèle de régression suivant :

$$Y = \mu + H\alpha + \beta X_j + C\gamma + e$$

où H est la matrice de taille $N \times K$ contenant les probabilités d'appartenance aux K haplotypes ancestraux, α le vecteur des effets du "fond ancestral", X_j la colonne j de la matrice X, β son effet, C une matrice de taille $N \times c$ contenant c covariables (par exemple, les proportions d'admixture des individus), et γ le vecteur des effets correspondant. L'élément H_{ik} de la matrice H s'obtient :

- soit par l'algorithme forward-backward : dans ce cas $H_{ik} = \mathsf{Pr}(Z_{ij} = k)$, où $\mathsf{Pr}(Z_{ij} = k)$ est la probabilité que le marqueur j de l'haplotype i provienne de l'haplotype ancestral k.
- soit par l'algorithme de Viterbi : dans ce cas $H_{ik} = 1$ si l'haplotype ancestral k est l'état le plus probable au marqueur j pour l'haplotype i, sinon $H_{ik} = 0$.

Dans les deux cas, nous rappelons que les éléments de la matrice H sont calculées conditionnellement à l'ensemble des marqueurs de part et d'autre du marqueur j courant.

A partir du modèle proposé, deux tests d'hypothèse peuvent être réalisés en considérant les cas suivants :

- $H_0 : Y = \mu + C\gamma + \epsilon$.
- $H_1 : Y = \mu + H\alpha + C\gamma + \epsilon$.
- $H_2 : Y = \mu + H\alpha + \beta X_j + C\gamma + \epsilon$.

Le test H_1 contre H_0 ($H_1 : H_0$) permet d'évaluer l'effet du "fond" ancestral au marqueur courant, $H_2 : H_1$ celui de son polymorphisme conditionnellement au fond ancestral. Finalement, Le modèle décrit supra se généralise aisément pour le cas diploïde en considérant les doses alléliques. Notons, que pour $K = 1$, le modèle se réduit de manière triviale à l'évaluation de l'effet allélique à chaque marqueur.

Marqueur non observé

Un sous-ensemble de marqueurs est parfois choisi pour caractériser la région afin de limiter les coûts de génotypage. Notre approche permet alors d'adopter une démarche "pas à pas" qui consiste à tester à intervalle de distance régulier l'effet du fond ancestral seul (sous-entendu $H_1 : H_0$). Cela implique que nous puissions établir, pour chaque haplotype, l'origine ancestrale à une position quelconque entre deux marqueurs. Autrement dit, quelle est la probabilité pour un marqueur "non observé" inclus entre deux marqueurs observés d'être originaire d'un haplotype ancestral donné. Considérons deux marqueurs consécutifs, j et $j+1$, séparés d'une distance d (supposée connue et exprimée en unité physique). Supposons que l'on veuille tester la position $j^{'}$ située à une distance $d1 < d$ de $M1$ et $d2 = d - d1$ de M_2. Dès lors, pour un haplotype donné i, la probabilité que le marqueur non observé $j^{'}$ provienne de l'haplotype ancestral k s'écrit :

$$\mathsf{Pr}(Z_{ij'} = k|X_i) = \frac{\mathsf{Pr}(X_i|Z_{ij'} = k^{'})}{\sum_{k'=1}^{K} \mathsf{Pr}(X_i|Z_{ij'} = k')}$$

où X_i désigne ici l'haplotype i, et :

$$\mathsf{Pr}(X_i|Z_{ij'} = k^{'}) = \sum_{k'=1}^{K} \sum_{k''=1}^{K} \mathsf{Pr}(Z_{ij'} = k|Z_{ij} = k^{'}, Z_{ij+1} = k^{''}) \times \alpha(k^{'})_j^{(i)} \times \beta(k^{''})_{j+1}^{(i)} \times \mathsf{Pr}(X_{ij+1}|Z_{ij+1} = k^{''})$$

avec :

- $\alpha(k)_j^{(i)}$ et $\beta(k)_j^{(i)}$ les variables "forward" et "backward" calculées pour l'haplotype i (voir le chapitre précédent),
- $\mathsf{Pr}(X_{ij+1}|Z_{ij+1} = k^{''})$ la probabilité d'observer l'allèle X_{ij+1} au marqueur $j+1$ sachant l'état ancestral $k^{''}$,
- et $\mathsf{Pr}(Z_{ij'} = k|Z_{ij} = k^{'}, Z_{ij+1} = k^{''})$ la probabilité que le marqueur non observé $j^{'}$ provienne de l'haplotype ancestral k sachant que les

marqueurs flanquants sont originaires des haplotypes ancestraux $k^{'}$ et $k^{''}$.

Cette dernière probabilité s'obtient de la manière suivante :

$$\mathsf{Pr}(Z_{ij'} = k | Z_{ij} = k^{'}, Z_{ij+1} = k^{''}) = \mathsf{Pr}(Z_{ij'} = k | Z_{ij} = k^{'}) \mathsf{Pr}(Z_{ij+1} = k^{''} | Z_{ij'} = k)$$

et nous rappelons que,

$$\mathsf{Pr}(Z_{ij'} = k | Z_{ij} = k^{'}) = e^{-\hat{\rho} d_1} I(k = k^{'}) + (1 - e^{-\hat{\rho} d_1}) \mathsf{Pr}(A_k)$$
$$\mathsf{Pr}(Z_{ij+1} = k^{''} | Z_{ij'} = k) = e^{-\hat{\rho} d_2} I(k = k^{''}) + (1 - e^{-\hat{\rho} d_2}) \mathsf{Pr}(A_{k^{''}})$$

où $I(\text{condition})$ vaut 1 si la condition est vraie, 0 sinon, $\hat{\rho}$ est le taux de recombinaison et $\mathsf{Pr}(A_k)$ la contribution globale de l'haplotype ancestral k. Ces deux quantités ayant été estimées au préalable par BSAH. Une fois $\mathsf{Pr}(Z_{ij'} = k)$ calculée pour chaque haplotype, on peut alors évaluer le modèle H_1 à la position courante. Ainsi, dans le cas où l'on ne dispose que de marqueurs encadrant une mutation causale supposée être apparue *ante* goulot d'étranglement, cette stratégie devrait permettre de la localiser.

Simulation

Afin d'évaluer notre méthode, nous avons considéré le scénario idéal suivant (équivalent à celui du chapitre précédent) :

1. soit un haplotype ancestral A_1 tel que $A_{1j} = 0$ pour $j = 1, \ldots, M$,
2. l'haplotype A_2 est obtenu en plaçant au hasard une proportion π de 1 dans A_1.
3. une population de taille N est créée à partir de N_1 individus homozygotes A_1 et N_2 individus homozygotes A_2.
4. la population évolue par panmixie avec un taux de mutation μ par génération et par site, un taux de recombinaison c par génération et par intervalle, pendant un nombre t de générations.

Nous avons alors considéré deux possibilités d'introduction de la mutation causale :

- mutation *ante* goulot d'étranglement : la mutation est introduite à l'étape 2 du scénario précédent en position centrale dans l'un des deux haplotypes.
- mutation *post* goulot d'étranglement : la mutation est introduite en position centrale dans les fragments d'un haplotype ancestral donné, à la dernière génération avec une probabilité q déterminée.

Enfin, la mutation causale a une valeur intrinsèque, notée a, de telle sorte que la valeur phénotypique des individus soit :

$$Y_i = aD_i + \epsilon_i$$

où D_i est la dose d'allèle causal de l'individu i, $\epsilon_i \sim \mathcal{N}(0, \sigma_e^2)$ avec σ_e^2 la variance environnementale. Enfin le site où se situe l'allèle causal n'est pas soumis à mutation pour les générations suivantes.

Par la suite, nous considérons des résultats obtenus pour les paramètres de simulation suivants : $M = 100$, N=100 individus et N_1=N_2, $\pi = 0.5$, $\mu = 0.005$, une distance de 0.025 cM entre les marqueurs (soit $c \approx 0.025$), et une variance environnementale unitaire, $\sigma_e^2 = 1$. Pour chaque jeu d'individus obtenu à l'issu de $t = 10$ générations, une seule analyse par BSAH a été effectuée. Les $M = 100$ marqueurs étaient supposés bi-alléliques et uniformément répartis sur une région de 10 kb. Un processus d'analyse est illustré par la Figure 6.3.

Mutation *ante* goulot d'étranglement

Dans un premier temps nous nous sommes intéressés au cas où l'on dispose d'une couverture complète de la région par les marqueurs (c'est à dire autant de marqueurs que de sites polymorphes observés, dont le site causal, comme illustré dans la Figure 6.3.

La Figure 6.4 montre les résultats obtenus pour 10 jeux de données simulés indépendamment, et pour différentes valeurs de l'allèle causal, $a = 0.5, 0.75, 1.0, 1.25$. Si l'approche marqueur par marqueur révèle bien la présence de la mutation causale au milieu de la région, le profil de la F-statistique est souvent "bruité", dû notamment au DL entre l'allèle causal et les autres allèles liés à la structure ancestrale (c'est à dire ceux discriminant A_1 de A_2). Même si des scénarios plus complexes restent à évaluer, dans le cas de marqueurs bi-alléliques, il est probable que ce "bruit" augmente avec le nombre d'haplotypes ancestraux non porteur de la mutation causale. En revanche, la Figure 6.4B , qui représente les profils de la F-statistique obtenus en testant $H_1 : H_0$ (c'est à dire l'effet ancestral seul au marqueur), indique sans ambiguïté la présence de la mutation causale en position centrale. Comme attendu, ce profil varie avec l'effet a de l'allèle causal : plus a est élevé et plus le profil se "resserre" autour de la position centrale. Notons que le profil plat de la F-statistique des tests $H_2 : H_1$ (Figure 6.4C) confirme la colinéarité entre la mutation causale et le fond ancestral.

Pour mimer une couverture partielle de la région par les marqueurs, nous avons également étudié des configurations avec $10, 25$ et 50 marqueurs répartis uniformément le long de la région (et n'incluant pas le site causal). Dans ce cas, une fois la structure ancestrale établie à l'aide de ce sous-ensemble de marqueurs, nous avons procédé à une analyse en considérant un pas mimant la densité de marqueur originale. Ici nous nous sommes limités au cas $a = 1$. Les résultats sont représentés dans la Figure 6.5. Le gain de l'approche pas à pas est net même quand seulement 10 marqueurs sont utilisés pour inférer la structure ancestrale.

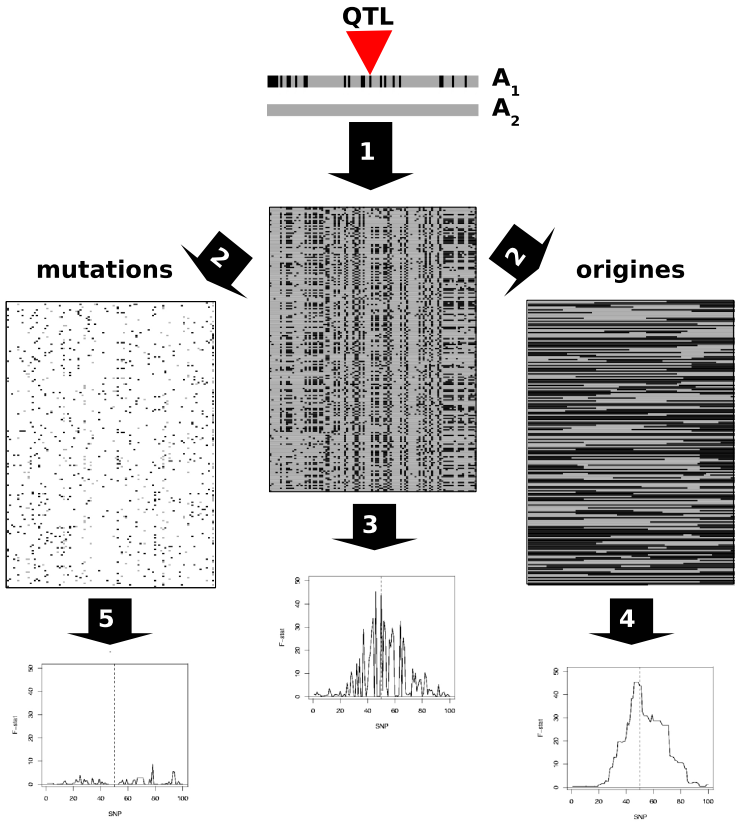


FIG. 6.3 – Illustration du scénario de simulation suivi des analyses comparatives de la région. Les deux haplotypes ancestraux sont représentés en haut de la figure. L'allèle causal (QTL) est supposé en complet déséquilibre de liaison avec l'haplotype A_1 (mutation *ante* goulot d'étranglement). Flèche 1 : après $t = 10$ générations de panmixie, les 200 haplotypes obtenus (N=100 individus diploïdes) sont analysés à l'aide de BSAH. Flèche 2 : à gauche, le profil de mutation détecté par l'algorithme. A droite, le coloriage des haplotypes en fonction des deux haplotypes ancestraux identifiés par BSAH. Flèche 3 : une régression simple marqueur par marqueur conduit à un profil de la F-statistique "bruité". Flèche 4 : la régression en fonction des probabilités d'appartenance des marqueurs aux haplotypes ancestraux permet de "lisser" le profil. Flèche 5 : l'effet des polymorphismes après avoir pris en compte l'effet du fond ancestral indique clairement que l'allèle causal est colinéaire à la structure ancestrale.

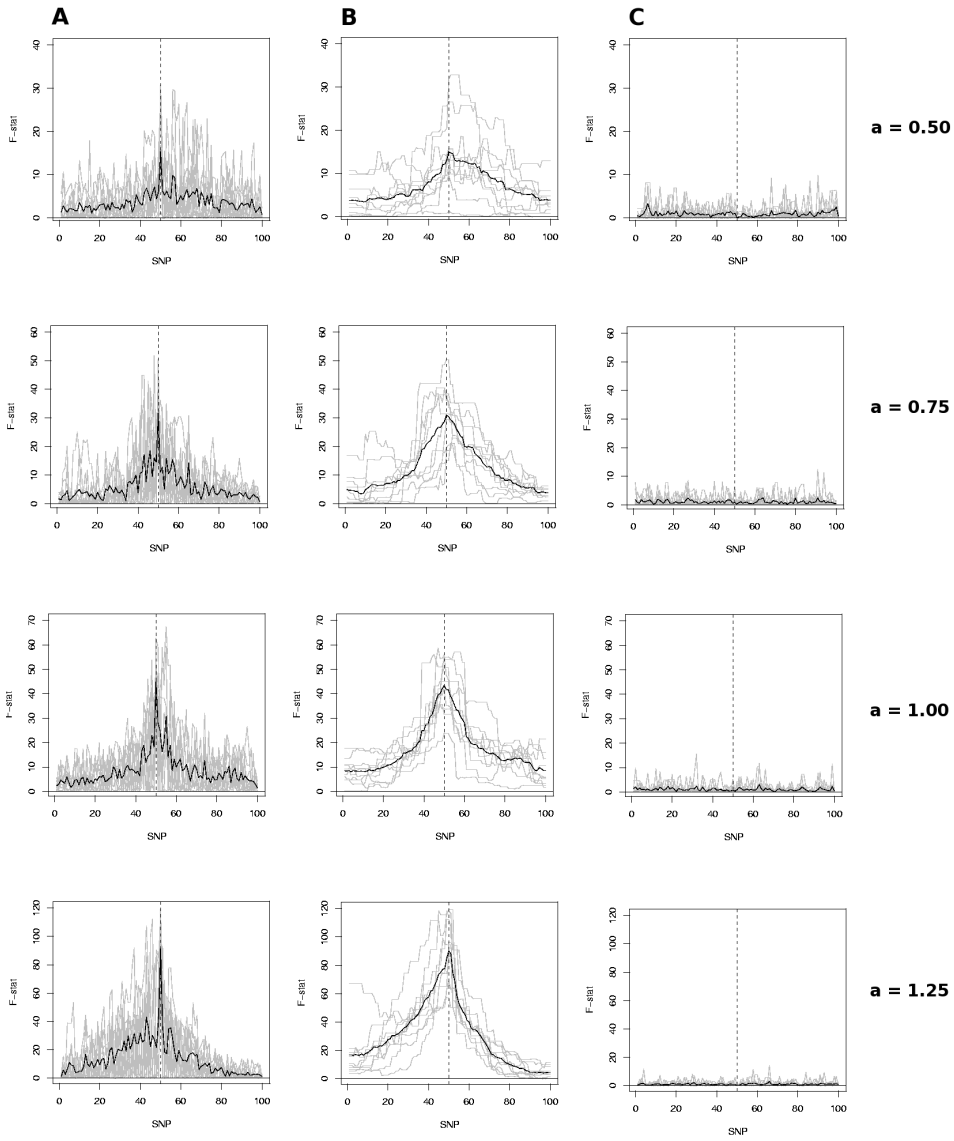


FIG. 6.4 – Profils de la F-statistique en considérant les $M = 100$ marqueurs dont le site causal. A) Régression simple marqueur par marqueur. B) Test H_1 : H_0 issu de la régression avec la décomposition ancestrale obtenue en supposant K=2 haplotypes ancestraux. C) Test H_2 : H_1 à partir de la même structure ancestrale. La valeur de a à droite de la figure indique l'effet de la mutation causale. Pour chacune de ces valeurs, 10 populations de $N = 100$ individus diploïdes ont été simulées en supposant une mutation causale *ante* goulot d'étranglement (courbes grises). La courbe en trait plein et noir indique le profil moyen obtenu à partir des 10 analyses indépendantes.

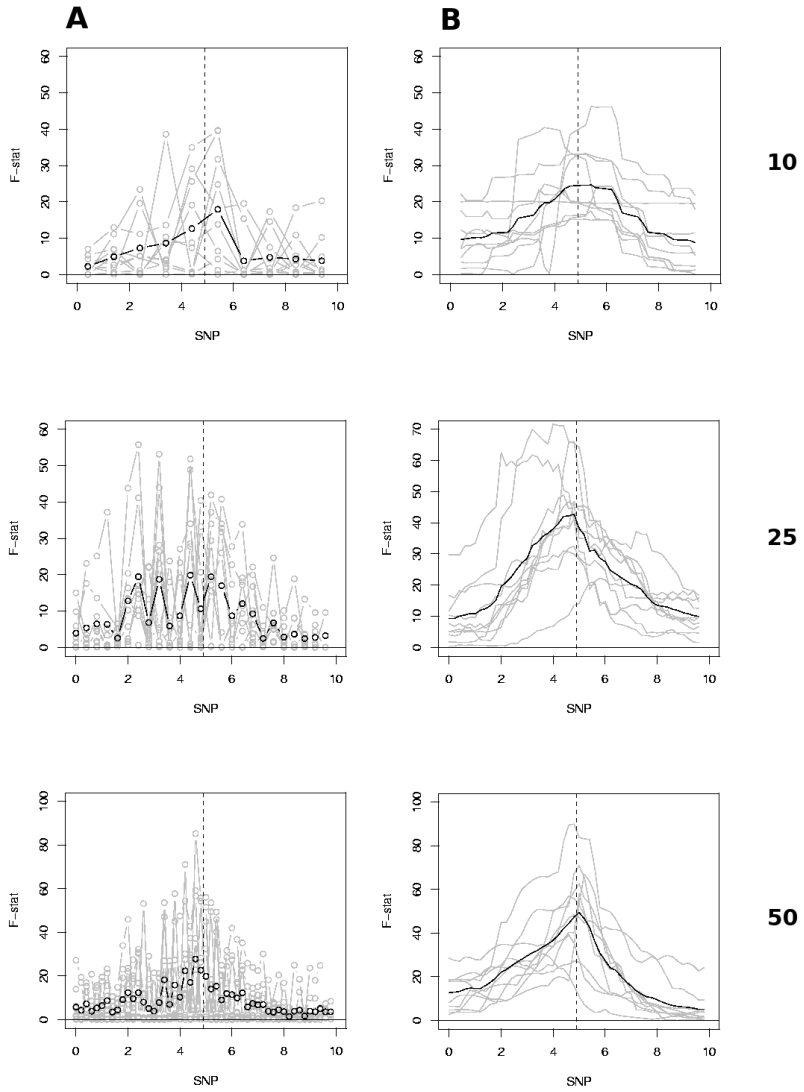


FIG. 6.5 – Profils de la F-statistique en considérant un sous-ensemble de marqueurs, uniformément répartis le long de la région. A) Approche simple marqueur par marqueur B) Approche "pas à pas" à l'aide de la structure ancestrale déterminée en supposant $K = 2$ haplotypes ancestraux. Le nombre de marqueurs dans la région est précisé à droite de la figure en face de chaque profil. Pour chaque analyse la pas a été choisi de manière à mimer la densité originale de marqueur. Pour chaque figure, 10 populations de $N = 100$ individus diploïdes ont été simulées en supposant une mutation causale *ante* goulot d'étranglement (courbes grises). La courbe en trait plein et noir indique le profil moyen obtenu à partir des 10 analyses indépendantes.

Mutation post goulot d'étranglement

En introduisant, à la dernière génération, la mutation causale de manière indépendante entre fragments ancestraux d'une même origine, notre scénario est équivalent à une apparition précoce de la mutation *post* goulot d'étranglement. Il a en outre l'avantage de faciliter le contrôle de la fréquence finale de la mutation. Sous l'hypothèse que celle-ci ne soit pas trop élevée, nous nous attendons à ce que le fond ancestral ne présente pas d'effet significatif le long de la région. En revanche, la mutation causale, apparue dans l'un des haplotypes ancestraux, devrait "ressortir" significativement. La Figure 6.6 représente les résultats obtenus pour 10 jeux de données simulés indépendamment et pour différentes probabilités d'apparition de la mutation causale dans un fond ancestral donné, $q = 0.15, 0.25, 0.35$ et 0.45. Enfin, l'effet de la mutation a été fixé à $a = 1$, et tous les marqueurs de la région (incluant le site causal) ont été utilisés dans les analyses. Premièrement, l'approche marqueur par marqueur simple détecte nettement la mutation causale quelle que soit la valeur de q. Deuxièmement, comme attendu, la structure ancestrale n'est pas significativement corrélée à la variation du caractère pour les faibles valeurs de q. Si pour $q = 0.35$ et $q = 0.45$, quelques profils suggèrent un léger effet du fond ancestral, en moyenne ce profil est quasi-plat. En revanche, le test $H_2 : H_1$ indique clairement la présence de la mutation causale en position centrale. L'analyse des sorties de BSAH pour chaque simulation a confirmé que l'allèle causal était bien spécifique de son fond ancestral (résultats non montrés).

Discussion et Conclusion

En décomposant la variabilité génétique selon deux origines potentielles, d'une part l'origine ancestrale des haplotypes, et d'autre part la possibilité d'un écart à cette origine par mutation, notre méthode permet de tester des hypothèses différentes quant à l'apparition de mutation causales dans la population étudiée. Nous insistons sur le fait que cette approche présuppose un scénario évolutif particulier, qui la restreint à des espèces ayant connu dans leur histoire un goulot d'étranglement récent. Dans le cas où cette hypothèse est peu vraisemblable à l'échelle globale de l'espèce, nous pensons qu'il pourrait être envisageable, localement, d'appliquer cette méthode aux régions dont les patrons de DL sont compatibles avec notre modélisation ancestrale. Nous invitons toutefois à la prudence : les scénarios étudiés par simulation étant très idéalisés, nous sommes conscients que des études supplémentaires, fondées sur des simulations plus réalistes, doivent être menées afin de valider définitivement notre démarche.

Nonobstant ces remarques, nous pensons que de par sa lisibilité et sa flexibilité, notre approche pourrait contribuer à faciliter la mise en œuvre d'études d'association chez les végétaux, et cela à plusieurs échelles. Lorsque

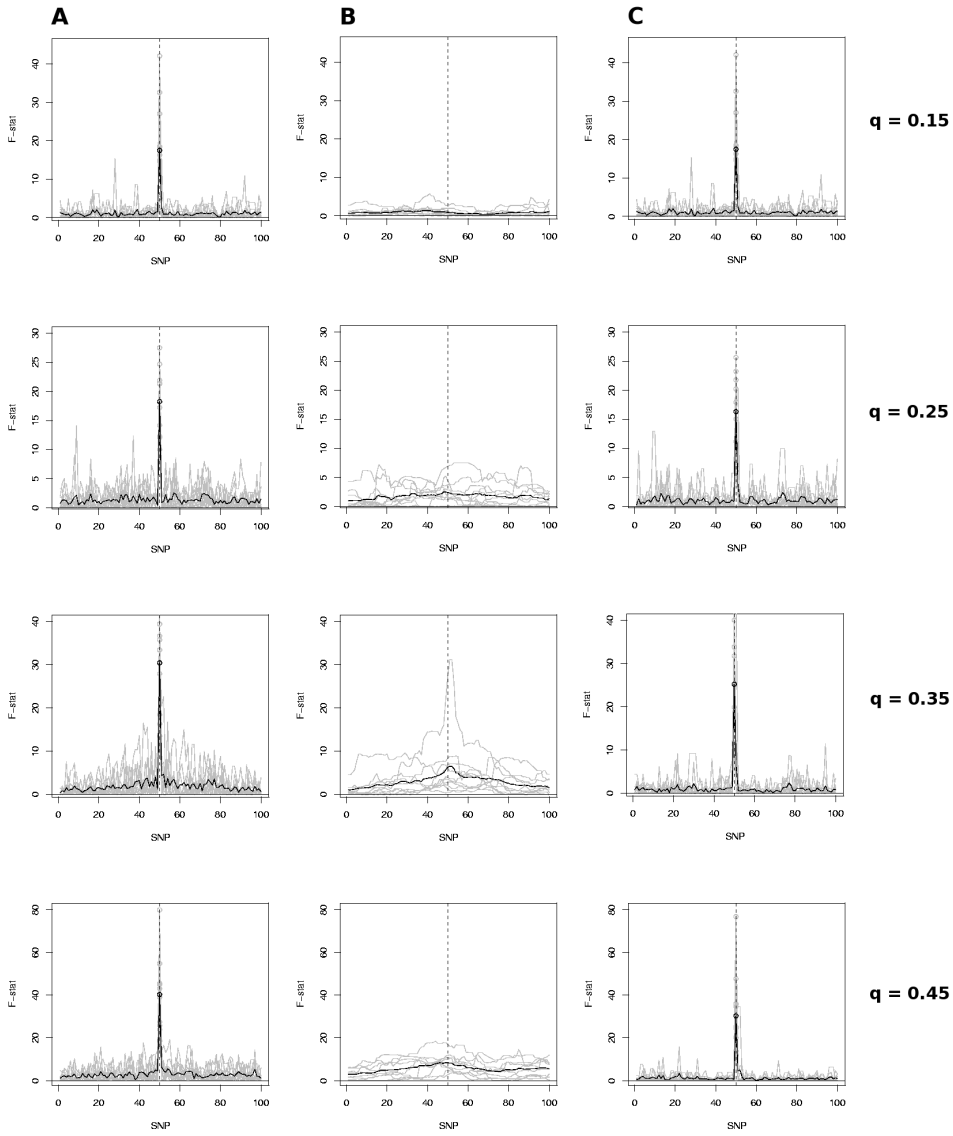


FIG. 6.6 – Profils de la F-statistique pour une mutation causale *post* goulot d'étranglement. A) Régression simple marqueur par marqueur. En supposant $K = 2$ haplotypes ancestraux, B) test $H_1 : H_0$, C) test $H_2 : H_1$. En face de chaque profil est indiqué la probabilité d'apparition de la mutation causale, q. Pour chacune de ces valeurs, 10 populations de $N = 100$ individus diploïdes ont été simulées (courbes grises). La courbe en trait plein et noir indique le profil moyen obtenu à partir des 10 analyses indépendantes.

les régions considérées se limitent à une zone intragénique, notre méthode permet non seulement de tester si cette région présente une association avec le caractère d'intérêt (dans l'idée du "testing for association" au sens de Zollner and Pritchard (2005)) mais aussi, sous l'hypothèse que la mutation causale est colinéaire au fond ancestral, d'adopter une démarche plus fine (via une approche de type "fine mapping", notamment lorsque les marqueurs ne recouvrent pas l'ensemble des sites polymorphes de la région). Dans ce dernier cas, à l'instar de la détection de QTL classique, il est possible de définir un intervalle de confiance autour la position causale la plus probable en utilisant les propriétés asymptotiques du ratio des log-vraisemblances ($H_1 : H_0$). Nous n'avons cependant pas étudié par simulation les propriétés d'un tel intervalle de confiance et son éventuelle utilisation doit donc être faite avec prudence. L'approche pas à pas pourrait être aussi utilisée comme un outil de diagnostic afin de conclure ou non à la présence du facteur causal dans la zone d'étude en fonction de la forme du profil de la statistique de test (un profil "remontant" sur les bords pourrait suggérer un facteur à l'extérieur de la région). Il est également possible d'étendre cette approche afin de pouvoir tester l'effet d'une mutation causale qui serait apparue *post* goulot dans un fond ancestral donné k à une position non observée $j^{'}$. Statistiquement, dans le cas haploïde, cela se traduit par l'introduction d'une variable cachée M_k^* au site non observé $j^{'}$ telle que cette variable vaut 1 si l'haplotype est porteur de la mutation, 0 sinon. L'idée étant que s'il y a effectivement une mutation causale d'effet suffisamment fort, on devrait pouvoir identifier les haplotypes "mutés" sur la base et de leur appartenance au type ancestral et de leur phénotype. Notons alors q_{M^*} sa fréquence dans la population ; la distribution de M_k^* est obtenue en appliquant la règle de Bayes suivante :

$$\mathsf{Pr}(M_{ki}^* = m|Y_i) = \frac{q_{M^*}\mathsf{Pr}(Y_i|M_{ki}^* = m)}{q_{M^*}\mathsf{Pr}(Y_i|M_{ki}^* = 1) + (1 - q_{M^*})\mathsf{Pr}(Y_i|M_{ki}^* = 0)}$$

où $m = 0, 1$, Y_i est la valeur phénotypique de l'individu i, et :

$$\mathsf{Pr}(Y_i|M_{ki}^* = m) = \frac{1}{\sqrt{2\pi}\sigma_e} \exp\left(\frac{Y_i - \mu_{im}}{2\sigma_e^2}\right)$$

avec,

$$\begin{cases} \mu_{i1} &= \mu + \alpha H_i + \beta_k M_k + \gamma C_i \\ \mu_{i0} &= \mu + \alpha H_i + \gamma C_i \end{cases}$$

avec $M_{ki} = \mathsf{Pr}(M_i^* = 1|Y_i)\mathsf{Pr}(Z_{ij'} = k)$, la probabilité que l'haplotype i soit porteur de la mutation. Dès lors au site $j^{'}$ la vraisemblance du modèle (qui est implicitement une vraisemblance de mélange gaussien) peut être maximisée à l'aide de l'algorithme EM Dempster *et al.* (1977) suivant :

– Étape 0 : initialise q_{M^*}.
– Étape $t+1$: soit $\theta^{(t)} = (q_{M^*}, \mu, \alpha, \beta_k, \gamma)^{(t)}$ les paramètres estimés à l'étape t.
  – Étape E : calculer $\mathsf{Pr}(M^*_{ki} = m | Y_i, \theta^{(t)})$, $m = 0, 1$.
  – Étape M : calculer $\theta^{(t+1)}$ à partir de :

$$\begin{cases} Y & = \mu + \alpha H + \beta_k M_k^{(t)} + C\gamma + e \\ q_{M^*}^{(t+1)} & = \frac{1}{N} \sum_{i=1}^{N} M_{ki}^{(t)} \end{cases}$$

  avec $M_{ki}^{(t)} = \mathsf{Pr}(M_i^* = 1 | Y_i, \theta^{(t)}) \mathsf{Pr}(Z_{ij'} = k)$.
  – Si $||\theta^{(t+1)} - \theta^{(t)}|| < \epsilon$ arrêter, sinon $t = t+1$ et retourner à l'étape E.

Dans le cas diploïde, la même procédure peut-être appliquée en considérant trois états pour la variable cachée : $2, 1$ et 0 correspondant aux génotypes possibles à la position $j^{'}$.

D'autre part, pour des régions plus importantes, et pour lesquelles une structure ancestrale homogène (c'est à dire une même valeur de K tout le long de la région) est guère probable, il conviendrait d'adopter une stratégie par "fenêtrage", où pour chaque fenêtre on admettrait une structure ancestrale homogène. Délimiter de telles zones ne paraît pas *a priori* aisé. Cependant, comme cela a été illustré au chapitre précédent, une valeur de K donnée n'implique pas obligatoirement que K halotypes ancestraux soient représentés tout le long de la région. Nous suggérons alors d'appliquer une première fois l'algorithme BSAH sur l'ensemble de la région et d'utiliser le résultat obtenu pour définir un découpage en sous régions entre lesquelles le nombre d'haplotypes ancestraux distincts diffère. Dans chaque sous région, la structure ancestrale peut alors être affinée par une analyse supplémentaire à l'aide de BSAH. Une autre possibilité consisterait à estimer au préalable le profil du taux de recombinaison le long de la région à l'aide de la méthode de Li and Stephens (2003), puis de définir les sous-régions en fonction de la présence ou non de "points chauds".

L'étude de grandes régions impliquant un nombre important de marqueurs (observés ou non) pose également le problème de tests multiples. Comme la structure ancestrale est établie indépendamment des tests d'association, des approches par permutation peuvent être mises en œuvre facilement. Une distribution empirique de la statistique de test (F-statistique ou ratio de vraisemblance) peut-être obtenue, sous l'hypothèse nulle, en permutant les données phénotypiques par rapport à la décomposition ancestrale (fond et/ou polymorphisme). Si des covariables sont incluses dans le modèle et font partie de l'hypothèse nulle, alors la distribution empirique s'obtient en permutant seulement la décomposition ancestrale par rapport au phénotype et aux covariables. Cependant, pour des dimensions élevées (grand nombre d'individus et de marqueurs) les tests par permutation peuvent se révéler

trop onéreux en temps machine. Comme alternative, et dans la mesure où une p-value peut-être dérivée de la statistique de test, nous conseillons d'appliquer la procédure de contrôle BH du FDR Benjamini and Hochberg (1995).

En résumé, notre méthode vise à offrir un cadre d'analyse complet et homogène afin de réaliser chez les végétaux des études d'association au sens large (du "testing for association" au plus élaboré "fine mapping"). Elle possède en outre l'avantage de pouvoir s'appliquer à des jeux de données conséquents. L'utilisation d'un modèle linéaire standard pour étudier l'association entre la décomposition ancestrale et la variation du caractère offre aussi suffisamment de souplesse pour envisager l'étude de modèles génétiques plus complexes, par exemple en incluant plusieurs sites à la fois (et éventuellement les termes d'interaction correspondants). Toutefois, nous sommes conscients que des études supplémentaires sont nécessaires afin de comparer les performances de cette approche à d'autres méthodes de la littérature (notamment la méthode de Durrant *et al.* (2004)). Dans ce but, nous avons intégré cette méthode aux autres outils d'analyse développés dans le cadre de la thèse, pour réaliser des études d'association (voir Annexe 2).

Cinquième partie

Discussion et Conclusion Générale

Chapitre 7

Discussion générale et perspectives

Chez les végétaux, la cartographie de QTL a constitué, à partir de la fin des années 1980, une étape clef dans l'étude du déterminisme de caractères quantitatifs. Comme nous l'avons vu au cours des chapitres précédents, la notion de cartographie de QTL se structure aujourd'hui, schématiquement, autour de deux stratégies complémentaires (voir la Figure 7.1) :

- **les études de liaison** : détection de régions chromosomiques corrélées à la variation du caractère d'intérêt à l'aide de populations contrôlées (rappelons que les méthodes de cartographie de QTL ont été développées initialement chez les plantes).
- **les études d'association** : recherche des variants génétiques causaux à l'échelle même du ou de plusieurs gènes dans des populations aux bases génétiques larges et pour lesquelles les relations d'apparentement entre individus sont souvent "cryptiques" ou peu informatives.

Cette dernière approche s'est développée plus récemment grâce à la systématisation de la recherche de gènes candidats (à partir notamment des données d'annotation issues de la génomique) et à la démocratisation du séquençage allélique. Face aux enjeux de cette démarche novatrice en génétique végétale, notre travail a tâché d'apporter quelques pistes méthodologiques, à l'interface entre génomique et génétique quantitative.

Dans un premier temps, la masse croissante des résultats issus des études de liaison nous a permis d'envisager une stratégie fondée sur le principe de la méta-analyse. La démarche décrite dans la première partie vise à la fois :

- à affiner la nature du déterminisme de caractères quantitatifs chez une espèce donnée en proposant un modèle synthétique à l'échelle du génome de la localisation conjointe des marqueurs et des QTL.
- à faciliter la sélection de "gènes candidats". Ces derniers étant supposés avoir été préalablement positionnés sur une carte génétique de référence (voir par exemple chez le maïs Falque *et al.* (2005)).

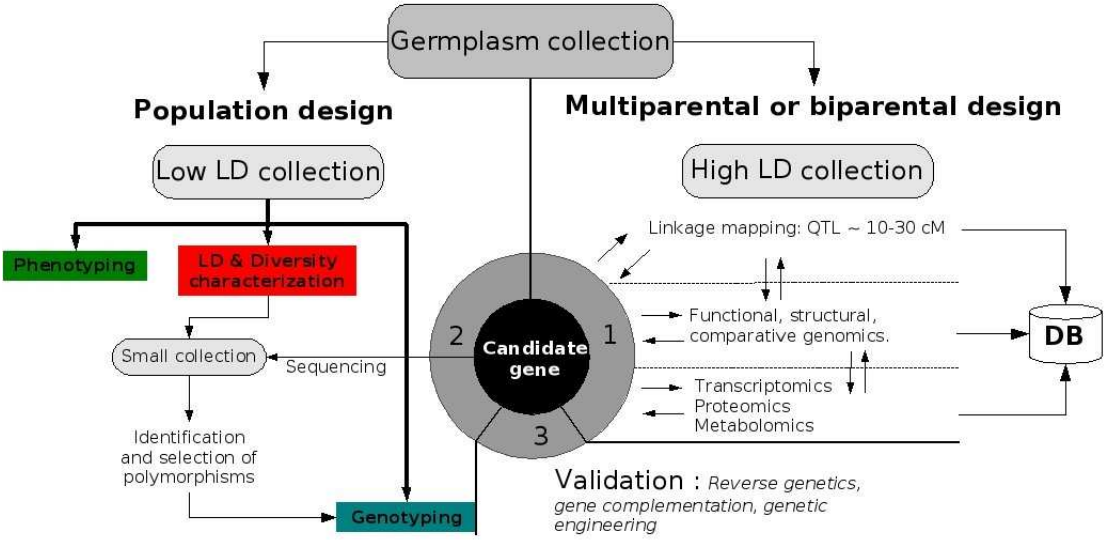


FIG. 7.1 – Représentation schématique d'un processus complet de cartographie fine de gènes impliqués dans le déterminisme de caractères d'intérêt. La première étape a conduit, ses dernières années, à enrichir considérablement les bases de données (DB) publiques. Les processus d'annotation fonctionnelle haut-débit ainsi que les outils de la génomique comparative ont notamment contribué à circonscrire l'espace des gènes candidats pour certains caractères. Dans un deuxième temps, l'association entre la variation du caractère et les polymorphismes des gènes candidats est évaluée à l'aide d'un dispositif de cartographie fine (faible DL). L'ultime étape consiste à valider biologiquement les associations détectées. Les développements méthodologiques de la thèse s'inscrivent dans les points 1 et 2 du schéma.

Ce dernier point est primordial dans le sens où un nombre croissant de données issues de la génomique sont disponibles, en particulier suite aux programmes de séquençage et d'annotation des génomes d'espèces modèles. Chez les végétaux, les études d'association étant jusqu'à présent principalement centrées "gènes candidats", l'étude des colocalisations entre QTL et gènes cartographiés est donc un enjeu majeur pour les années à venir. Pour ce faire, nous pensons que la méta-analyse pourrait constituer un appui intéressant aux outils de la génomique.

Une fois que des gènes candidats ont été identifiés, plusieurs méthodes peuvent être mises en œuvre afin d'effectuer leur validation fonctionnelle. Dans ce cadre, les études d'association offrent un premier niveau de validation dont les résultats peuvent servir à guider des approches plus biologiques telles que la transgénèse ou la mutagénèse. Pour la plupart des espèces végétales, les collections diversifiées créées pour effectuer les études d'association présentent des patrons de DL supposés assurer un "bon" degré de résolution à l'échelle physique du génome. Néanmoins, la structure génétique éventuelle de ces collections peut induire un DL "artéfactuel" susceptible de compromettre la validité de la démarche. Il est alors indispensable de réaliser des études préliminaires pour inférer, sur la base de marqueurs neutres, cette structure sous-jacente. Dans cet objectif, et bien que des méthodes plus complexes aient été développées ces dernières années, nous pensons que notre méthode d'analyse de la structuration pourrait constituer une alternative plus simple et plus lisible - notamment en explicitant la stratégie de choix de modèle en terme de DL résiduel.

Enfin, nous avons pu constaté au cours du chapitre précédent que la modélisation de la diversité génétique des régions candidates constituait un enjeu majeur afin d'explorer plus finement les associations entre variabilité génétique et variation phénotypique. Nous sommes conscients que la modélisation du DL que nous avons présentée dans la troisième partie de la thèse est très idéalisée. Néanmoins, nous pensons que sa souplesse et sa facilité de mise en œuvre - comparée aux approches fondées entièrement sur des modèles de coalescence - sont des atouts potentiels à son intégration aux tests d'association.

Dans ce chapitre, nous revenons dans une premier temps sur la méta-analyse et proposons des études supplémentaires à sa validation. Nous explicitons aussi comment effectuer la sélection de gènes candidats sur la base de ses résultats. Deuxièmement, nous discutons des études complémentaires à réaliser afin d'explorer plus finement le comportement de notre méthode d'étude de la structuration génétique. Troisièmement, nous esquissons la manière dont notre modélisation du DL peut-être intégrée à une autre problématique connexe à l'étude de la diversité génétique. Puis nous revenons brièvement sur la question des études d'association dans des populations structurées. Enfin, nous concluons dans un dernier chapitre par des considérations plus générales.

Méta-analyse de QTL

Les simulations présentées dans l'article de la première partie témoignent du bon comportement du modèle de mélange gaussien dans le cadre de la méta-analyse de QTL. En outre, l'étude des scénarios incluant la corrélation entre observations deux à deux ainsi qu'une connaissance imparfaite des variances a montré que notre approche était assez robuste (dans ce dernier cas, la probabilité de sélectionner le "vrai" modèle se dégrade légèrement, sans toutefois compromettre trop fortement la qualité de prédiction de la méta-analyse).

Néanmoins, ces simulations ne prennent pas en compte le possible écart à la normalité des positions de QTL détectés. Cet écart peut-être induit par les différentes sources d'erreur qui peuvent se cumuler lors de :

- la création de la carte consensus : suite à des hétérogénéités du taux de recombinaison entre populations et/ou à des incohérences d'ordonnancement local des marqueurs.
- la projection des QTL : par sa nature empirique, elle peut conduire à "bruiter" et la position la plus probable du QTL et son intervalle de confiance sur la carte consensus.

Aussi, il serait judicieux de compléter notre premier travail par des études supplémentaires plus proches du processus même de génération des données. Dans une premier temps, sous l'hypothèse que les taux de recombinaisons sont homogènes entre populations, nous proposons d'évaluer les performances de la méta-analyse selon le scénario détaillé dans l'Encadré 1. Des études préliminaires, au cas par cas (comme illustrées dans la Figure 7.2), semblent augurer d'une bonne tenue de notre approche dans ce contexte.

Toutefois, ce scénario de simulation ne prend pas en compte l'hétérogénéité du contenu des bases de données de QTL. Cette hétérogénéité se décompose généralement en :

- différents types de population de croisement.
- des cartes génétiques avec des densités de marqueurs différentes.
- des positions de QTL pour lesquelles on ne dispose pas toujours des intervalles de confiance.

Pour mimer au mieux ces sources d'hétérogénéité, nous proposons d'effectuer également des simulations fondées sur le même principe que celui évoqué supra, mais intégrant les distributions empiriques détaillées dans l'Encadré 2.

Conjointement au processus de validation par simulations que nous venons d'évoquer, la mise en œuvre de la méta-analyse chez différentes espèces et pour différents caractères pourrait constituer une validation empirique alternative. En particulier, chez les espèces qui disposent d'une cartographie physique complète (comme le riz) ou en voie d'être achevée (comme le maïs) et pour lesquelles, par conséquent, on dispose d'une densité élevée de gènes candidats, voire d'ores et déjà validés pour certaines. L'évaluation

Encadré 1 : Méta-analyse de QTL : scénario de simulations

Considérons un chromosome sur lequel sont ordonnés et positionnés M locus marqueurs. On se donne $K > 1$ QTL répartis le long du chromosome. Pour un type de croisement donné (backcross, F_2, etc), nous proposons le scénario suivant :

1. Pour une population p :
   - Soit z, l'indice d'un QTL tiré au hasard parmi $1, \ldots, K$ selon la loi multinomiale $\pi_1, \ldots, \pi_K$.
   - L'effet du QTL z est tiré au hasard dans une loi gamma. Les autres QTL ont un effet supposé nul. C'est-à-dire qu'un seul QTL coségrège avec les marqueurs par population.
   - Simuler N individus caractérisés au M marqueurs (supposés codominants).
   - Estimer la carte génétique en supposant l'ordre des marqueurs connu.
   - Estimer la position et l'intervalle de confiance du QTL par "Interval Mapping".
2. Répéter 1 pour $p = 1, \ldots, P$.
3. Construire la carte consensus par moindres carrés pondérés.
4. Effectuer la méta-analyse pour les P QTL projetés sur la carte consensus.

pourrait ainsi se fonder sur l'étude des colocalisations entre les positions des QTL estimées par la méta-analyse et la distribution des gènes candidats sur le génome. L'hypothèse étant que, si les positions de QTL obtenues par méta-analyse et que le choix des gènes candidats sont pertinents, alors la probabilité d'observer des colocalisations *a priori* fonctionnelles pour le caractère étudié devrait être supérieure à celle obtenue soit au hasard, soit en considérant indépendamment les QTL observés. Dans ce but, mais aussi en vu d'automatiser la sélection de gènes candidats à l'issu de la méta-analyse, nous proposons la procédure décrite dans l'Encadré 3. Bien que cette stratégie de sélection demeure assez empirique, nous pensons qu'en l'absence d'informations supplémentaires sur la relation entre gènes cartographiés et QTL, elle fournit une base raisonnable pour classer les données de gènes disponibles en intégrant à la fois leur *a priori* fonctionnel et leur degré de colocalisation avec les QTL. De plus, de par sa simplicité, elle a l'avantage de pouvoir s'appliquer à des banques de données de gènes conséquentes.

Étude de la structuration génétique

Une fois constituée la population d'étude en vue de conduire des études d'association, la première question qui doit s'imposer est celle relative à sa structure génétique. Pour tâcher d'y répondre, nous avons donc développé une approche qui a le mérite d'allier des techniques usuelles d'analyse multivariée à une modélisation souple des phénomènes de structuration.

Cependant, dans l'article que nous avons présenté, l'évaluation de notre méthode repose sur un scénario de mélange assez naïf et qui présuppose

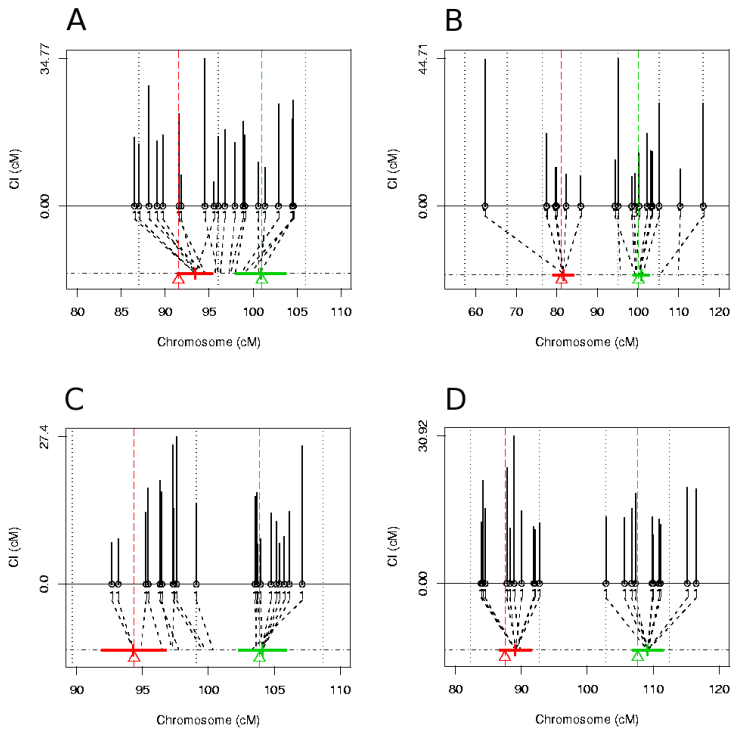


FIG. 7.2 – Illustration de résultats de la méta-analyse pour des données simulées selon le scénario décrit dans l'Encadré 1. A,B) backcross, C,D) F_2. Dans tous les cas, $P = 20$ populations ont été simulées pour lesquelles $M = 21$ marqueurs étaient répartis uniformément sur un chromosome de 200 cM de long. Les 2 vrais QTL étaient positionnés au centre du chromosome avec une distance de A,C) 10 cM et B,D) 20 cM. Les cercles indiquent les positions des QTL détectés dans chaque population par "Interval Mapping" et le trait plein au dessus la longueur de l'intervalle de confiance correspondant. Le chiffre en dessous des positions indiquent le nombre de QTL projeté à cette position (dans le cas où ce chiffre est supérieur à 1, seul le plus grand des intervalles de confiance est représenté). Le trait en pointillé indique la position prédite par le modèle pour chaque QTL observé. Enfin, les triangles vert et rouge indiquent les positions des "vrais" QTL sur le chromosome et le segment au-dessus, l'intervalle de confiance des QTL obtenus par la méta-analyse - la position est indiquée par un trait horizontal.

Encadré 2 : Méta-analyse de QTL : procédure empirique de simulation

Pour reproduire au mieux les configurations expérimentales de détection de QTL décrites dans les bases de données publiques, nous suggérons la démarche suivante. A partir d'une large collecte de données effectuée pour plusieurs caractères et plusieurs espèces, déterminer les distributions empiriques des paramètres suivants :
- type de population de croisement, notée $\mathcal{P}_C$,
- nombre d'individus par croisement, notée $\mathcal{P}_N$,
- nombre de marqueurs par croisement et par chromosome, notée $\mathcal{M}_n$,
- distance en cM entre marqueurs consécutifs, notée $\mathcal{M}_d$,
- nombre de marqueurs communs entre cartes, notée $\mathcal{M}_c$.
- nombre de QTL dont l'intervalle de confiance est connu, notée $\mathcal{Q}_c$.

Puis, pour une configuration donnée de "vrais" QTL, adopter la démarche suivante :

1. Pour une population $p = 1, \ldots, P$:
   - Tirer son type dans $\mathcal{P}_C$.
   - Tirer le nombre d'individus dans $\mathcal{P}_N$.
   - Tirer le nombre de marqueur M dans $\mathcal{M}_n$.
   - Tirer la distribution des marqueurs sur le chromosome dans $\mathcal{M}_d$.
   - Réaliser l' étape 1 de l'Encadré 1.
   - Reporter l'intervalle de confiance du QTL avec la probabilité $\mathcal{Q}_c$, sinon ne reporter que sa proportion de variance expliquée.
2. Répartir des marqueurs communs dans les P cartes en tirant leur proportion dans $\mathcal{M}_c$.
3. Effectuer les étapes 3 et 4 de l'Encadré 1.

une architecture simple de l'événement d'admixture. Or, les scénarios sous-jacents aux effets visibles de structuration génétique chez les végétaux sont sûrement plus complexes, impliquant potentiellement différents événements de mélanges dans le temps. De plus nous n'avons considéré que le cas de marqueurs bi-alléliques, or la plupart des analyses de structuration utilisent des marqueurs multiallélique qui conduisent souvent à des configurations hétérogènes en terme de nombre d'allèles et d'information aux marqueurs. Pour ces raisons, nous estimons qu'il serait souhaitable d'effectuer des simulations supplémentaires afin d'étudier le comportement de notre méthode dans des cas de figure plus proches à la fois des données expérimentales et de modèles d'admixture plus réalistes (comme le "continuous gene flow model" discuté par Pfaff *et al.* (2001)).

D'autre part, dans notre article, la comparaison avec structurese restreint à l'application. Or, il serait intéressant d'évaluer conjointement au cours d'un nouveau cycle de simulations le comportement des deux méthodes et d'identifier les limites et les avantages respectifs de chacune. Une première limite immédiate de notre approche, comparée au modèle bayésien de structure, est l'absence d'indicateurs de précision sur les quantités estimées à l'issue de l'algorithme EM. Le nombre de paramètres considérés étant généralement trop élevé pour envisager des stratégies usuelles d'estimation de matrice de variance-covariance, nous proposons de compenser cette apparente faiblesse

Encadré 3 : Méta-analyse de QTL : sélection de gènes candidats

On se restreint ici à un chromosome. Soit $q_1, \ldots, q_K$ les positions des K QTL estimés par la méta-analyse. On note $\sigma_1, \ldots, \sigma_K$ leurs écarts-types estimés conditionnellement à K. Soient N gènes candidats sur le chromosome aux positions $g_1, \ldots, g_N$. On note $\mathsf{Pr}(g_i)$ l'*a priori* fonctionnel de chaque gène (relativement au caractère d'intérêt). Cet *a priori* découle des informations d'annotation des gènes mais également de possibles comparaisons entre espèces (synthénie). La colocalisation d'un gène candidat g_i avec un QTL q_k s'évalue à l'aide d'une simple règle de Bayes :

$$\mathsf{Pr}(g_i = q_k) = \frac{\mathrm{Co}(g_i, q_k)}{\sum_{k'=1}^{K} \mathrm{Co}(g_i, q_{k'})}$$

avec $\mathrm{Co}(g_i, q_k) = \phi(\frac{q_k - g_i}{\sigma_k})$ où ϕ est la densité d'une loi normale centrée réduite. Conditionnellement à chaque QTL, le classement des gènes candidats s'obtient en appliquant à nouveau une règle de Bayes :

$$\mathsf{Pr}(g_i | q_k) = \frac{\mathsf{Pr}(g_i = q_k)\mathsf{Pr}(g_i)}{\sum_{i'=1}^{N} \mathsf{Pr}(g_{i'} = q_{k'})\mathsf{Pr}(g_{i'})}$$

Il est possible de rendre le classement non conditionnel aux QTL en intégrant les probabilités obtenues sur les K QTL :

$$\mathsf{Pr}(g_i | K) = \sum_{k=1}^{K} \pi_k \mathsf{Pr}(g_i | q_k)$$

où nous rappelons que π_k est la proportion de mélange correspondant au QTL k du modèle. Enfin, il est également possible de rendre ce classement non conditionnel à K en intégrant sur toutes les valeurs de K évaluées, c'est à dire pour $K = 1, \ldots, K_{\max}$ (éventuellement $K_{\max}$ peut-être égal au nombre de QTL observés initialement sur le chromosome). Pour cela on utilise le "weight of evidence", w_K, de chaque modèle :

$$\mathsf{Pr}(g_i | \mathrm{QTL}) = \sum_{K=1}^{K_{\max}} w_K \mathsf{Pr}(g_i | K)$$

où "QTL" désigne l'ensemble des QTL observés. Enfin, la position des gènes candidats sur le chromosome peut avoir été obtenue de deux manières :

– soit en considérant leur unique carte d'origine comme carte de référence (c'est-à-dire en supposant connue leur position). Dans ce cas la fonction de colocalisation $\mathrm{Co}(g_i, q_k)$ est celle décrite supra.
– soit en intégrant leurs cartes d'origine à l'ensemble des cartes de QTL.

Dans ce dernier cas la fonction de colocalisation $\mathrm{Co}(g_i, q_k)$ s'écrit :

$$\mathrm{Co}(g_i, q_k) = \frac{1}{2}\left(\phi(\frac{q_k - g_i}{\sigma_k}) + \phi(\frac{g_i - q_k}{\gamma_i})\right)$$

où γ_i est l'écart-type de la position du gène candidat g_i, estimé lors de la construction de la carte consensus.

par une stratégie *ad hoc* décrite dans l'Encadré 5. A la différence du "Gibbs sampling" de STRUCTURE, notre EM stochastique explore uniquement le voisinage du maximum de vraisemblance obtenu au préalable par l'algorithme EM. Une stratégie alternative pourrait consister à effectuer une analyse supplémentaire à l'aide de STRUCTUREen définissant les distributions *a priori* des paramètres à l'aide des valeurs estimées par l'algorithme EM.

Il serait également nécessaire d'étudier plus en détail l'effet, sur l'estimation des paramètres du modèle, d'autres mécanismes conduisant à la création de DL. Par exemple, dans le cas de mutations survenues *post* admixture, ou lorsque la liaison physique entre les marqueurs peut-être négligée (nous rappelons que notre modèle repose sur l'hypothèse que les marqueurs sont indépendants). Si l'on dispose d'une carte génétique des marqueurs, il est alors possible d'intégrer l'information de liaison dans le modèle pour prévenir des biais potentiels dus au déséquilibre de liaison supplémentaire induit par la liaison physique. Pour ce faire, chaque haplotype est modélisé par une chaîne de Markov cachée d'ordre un le long des marqueurs : les variables cachées correspondent aux populations d'origine à chaque marqueur, et la probabilité de transition d'un marqueur au suivant doit prendre en compte les proportions d'admixture de l'haplotype ainsi que la probabilité d'avoir recombiné entre les marqueurs. Cette probabilité est fonction de la distance génétique, supposée connue entre les marqueurs, et du nombre de générations survenues depuis l'événement d'admixture, capturé alors par un paramètre supplémentaire. Si ce modèle n'est pas trop difficile à réaliser (il a notamment été intégré dans STRUCTUREet dans le logiciel développé par HOGGART *et al.* (2003)), il pose toutefois deux problèmes : i) l'information de phase est rarement disponible pour les marqueurs couramment utilisés pour effectuer ce type d'analyse, ii) l'étude du choix de modèle devient moins immédiate, et l'on est désormais obligé d'avoir recours à des critères fondés sur une log-vraisemblance pénalisée qui nécessitent donc une évaluation supplémentaire. Notons que le premier point peut-être en partie résolu grâce aux outils désormais disponibles pour inférer la phase. A ce jour, aucune étude n'a cependant été publiée pour évaluer l'effet des erreurs de reconstruction de phase sur l'inférence des paramètres d'admixture.

Modélisation du DL et haplotypes ancestraux

Nous sommes conscients que notre modélisation du DL (décrite dans la partie précédente) est très empirique. De plus les simulations que nous avons réalisées dans l'article sont très minimalistes et les bonnes performances de notre méthode sont en grande partie dues à un scénario qui "colle" trop au modèle. Même si l'analyse de données de séquençage chez le maïs semblent indiquer que notre hypothèse évolutive n'est pas totalement fantaisiste, la réalité biologique de notre modèle demeure une question épineuse.

Encadré 4 : Étude de la structure : EM Stochastique

On suppose que les paramètres du modèle d'admixture ont été préalablement estimés par l'algorithme EM. Nous reprenons la même notation que dans l'article de la deuxième partie.

- **Étape 0** : A partir de $\hat{P}$ et $\hat{Q}$ obtenir la distribution *a posteriori* des variables cachées, $\mathsf{Pr}(Z|X,Q,P)$.
- **Étape 1** : pour chaque individu i et pour chaque locus l tirer au hasard son origine à partir de la distribution courante des variables cachées, $\mathsf{Pr}(z_{il}=k|x_i,q_i,P)$.
- **Étape 2** : Mettre à jour P et Q.
- **Étape 3** : Recalculer $\mathsf{Pr}(Z|X,Q,P)$.
- Répéter 1,2 et 3 B fois.

L'algorithme EM stochastique est similaire à une chaîne de Markov Monte-Carlo Celleux and Govaert (1992) et permet ainsi d'explorer le voisinage du maximum de vraisemblance. Nous proposons alors de lancer M chaînes indépendantes et de calculer les écarts-types d'un paramètre $\hat{\theta}$ (que ce soit une fréquence ou une proportion d'admixture) soit en utilisant directement l'estimateur empirique de la variance

$$\mathrm{var}(\hat{\theta}) = \frac{1}{MB}\sum_{m=1}^{M}\sum_{b=1}^{B}(\hat{\theta}-\theta_b^{(m)})^2$$

soit intégrant les pondérations fondées sur la vraisemblance obtenue à chaque étape du processus :

$$\mathrm{var}(\hat{\theta}) = \frac{1}{MB}\sum_{m=1}^{M}\sum_{b=1}^{B} w_b^{(m)}(\hat{\theta}-\theta_b^{(m)})^2$$

avec

$$w_b^{(m)} = \frac{\mathsf{Pr}(X|Q_b^{(m)},P_b^{(m)})}{\sum_{b'=1}^{B}\mathsf{Pr}(X|Q_{b'}^{(m)},P_{b'}^{(m)})}$$

soit en ne retenant que les points pour lesquels la log-vraisemblance des observations est supérieure à la log-vraisemblance maximale moins 2, c'est-à-dire, en reprenant la notion précédente, $w_b^{(m)} = 1$ si et seulement si $\log(\mathsf{Pr}(X|Q_b^{(m)},P_b^{(m)})) \geq \log(\mathsf{Pr}(X|\hat{Q},\hat{P}))-2$, sinon $w_b^{(m)}=0$.

D'une part, afin de mieux en évaluer la pertinence, il serait nécessaire de réaliser des études à partir de simulations plus réalistes. Par exemple, en considérant un scénario proche de celui proposé par Tenaillon *et al.* (2004) (afin d'étudier l'influence de la sélection et de la domestication sur la diversité génétique de certains gènes de maïs). Pour mimer la domestication, on considère ainsi une population ancestrale de taille N qui évolue en panmixie pendant un nombre de générations t_2, avant de connaître brutalement un goulot d'étranglement. Le goulot se caractérise par une réduction de la taille de la population Nb avec $b < 1$ pendant d générations. Puis à la génération $t_2 = t_1 + d$, la taille de la population augmente, soit instantanément, soit progressivement pour atteindre la taille actuelle Np. Le taux de recombinaison ainsi que le taux de mutations sont supposés constant au fil du temps.

Si sous l'angle de la génétique des populations l'algorithme BSAH s'apparente davantage à une solution pragmatique qu'à une modélisation satisfaisante des mécanismes évolutifs, nous pensons néanmoins que le taux de recombinaison et le taux de mutation estimés par BSAH peuvent être utiles pour effectuer des comparaisons entre différentes régions géniques (comme illustré dans l'application de notre article). Afin de faciliter cette démarche et de lui conférer une base plus statistique, nous proposons de calculer simultanément les intervalles de confiance des deux paramètres en considérant l'ensemble des couples (ρ, ϵ) situés dans la région autour du maximum de vraisemblance et délimitée par $\log(L(X; \rho, \epsilon|A)) - 2$, où $L(X; \rho, \epsilon|A)$ est la vraisemblance des observations conditionnelle aux haplotypes ancestraux (pour la notation voir l'article).

Enfin, cette modélisation du DL pourrait se révéler utile pour réaliser des processus de sélection de sous-ensemble de marqueurs dans des régions de grandes tailles dont la structure du DL est complexe. Nous avons vu au chapitre précédent, dans le cadre de l'approche marqueur par marqueur, que plusieurs méthodes avaient été proposées dans la littérature pour sélectionner au préalable les marqueurs les plus informatifs. Certaines de ces méthodes reposent sur une modélisation du DL en bloc d'haplotypes pour identifier ce que désormais on appelle communément des tagSNP : c'est-à-dire le sous ensemble minimal de marqueurs qui capture au mieux la structure en bloc (voir en particulier la synthèse de Zhang *et al.* (2005)). De manière similaire, BSAH pourrait être intégré dans une stratégie de sélection de marqueurs afin d'identifier le sous-ensemble qui représente le mieux possible la structure ancestrale détectée. Plus particulièrement, le cadre probabiliste de BSAH offre la possibilité de formuler ce problème sous un angle plus général en facilitant la construction de fonction prédictive (voir l'Encadré 6). BSAH pourrait ainsi être utilisé dans un premier temps sur un sous-échantillon d'individus afin de capturer les marqueurs les plus informatifs en vue de leur génotypage dans l'ensemble de la population d'étude. L'information fournie par l'ensemble des individus aux marqueurs sélectionnés pourrait être alors à nouveau analysé par BSAH afin de conduire des études d'association de

Encadré 5 : BSAH : sélection de marqueurs

Soient M marqueurs ordonnés le long d'une région donnée. On suppose que BSAH a détecté $K > 1$ haplotypes ancestraux dans cette zone. Nous reprenons alors la notation de l'article, à savoir : X est la matrice des haplotypes caractérisés aux M marqueurs, A la matrice des haplotypes ancestraux, $\hat{\rho}$ le taux de recombinaison et $\hat{\epsilon}$ le taux de mutation, tous deux estimés par BSAH. Pour alléger la notation on note par la suite $\hat{\theta} = (\hat{\rho}, \hat{\epsilon}, A)$. Supposons à présent que l'allèle au marqueur j de l'haplotype X_i soit manquant. On peut alors obtenir un pronostic sur l'allèle au marqueur j en appliquant la règle de Bayes suivante :

$$\mathsf{Pr}(X_{ij} = x | X_{il<j}, X_{il>j}, \hat{\theta}) = \frac{\mathsf{Pr}(X_{il<j}, X_{ij} = x, X_{il>j} | \hat{\theta})}{\sum_{x'} \mathsf{Pr}(X_{il<j}, X_{ij} = x', X_{il>j} | \hat{\theta})}$$

où $X_{il<j}$ désigne l'ensemble des allèles observés aux marqueurs à gauche de j et de manière similaire $X_{il>j}$ désigne l'ensemble des allèles observés aux marqueurs à droite de j, et $\mathsf{Pr}(X_{il<j}, X_{ij} = x, X_{il>j} | \hat{\theta})$ est la probabilité de l'haplotype $(X_{il<j}, X_{ij} = x, X_{il>j})$ sachant les paramètres du modèle.
Soit $S(M)$ un sous-ensemble de marqueurs de M. On introduit alors la fonction de prédiction f définie par :

$$f[X_{ij}|S(M)] = \begin{cases} X_{ij} & \text{si } j \in S(M) \\ \mathsf{Pr}(X_{ij} = x | S(M), \hat{\theta}) & \forall x \text{ , sinon} \end{cases}$$

où $\mathsf{Pr}(X_{ij} = x | S(M), \hat{\theta})$ est obtenue en appliquant la règle énoncée précédemment mais en ne considérant que les marqueurs à gauche et à droite du marqueur j appartenant au sous-ensemble $S(M)$. On définit alors l'erreur moyenne de prédiction individuelle $\eta_i = (1/M) \sum_{j=1}^{M} \eta_{ij}$, avec :

$$\eta_{ij} = \begin{cases} 0 & \text{si } j \in S(M) \\ \sum_x (\mathsf{Pr}(X_{ij} = x | S(M), \hat{\theta}) - I(X_{ij} = x))^2 & \text{sinon} \end{cases}$$

où $I(X_{ij} = x) = 1$ si $X_{ij} = x$, 0 sinon.
Le but est de trouver le sous-ensemble $S(M)$ de plus petite taille qui minimise l'erreur de prédiction moyenne $(1/N) \sum_{i=1}^{N} \eta_i$. Une solution exacte peut-être obtenue en appliquant l'algorithme dynamique proposé par Halperin *et al.* (2005) dont la complexité est polynomiale en $O(M^3 N)$. Notons que la prédiction des allèles aux marqueurs non sélectionnés peut-être améliorée en considérant un algorithme de type "forward-backward" qui permet d'obtenir la prédiction à un site donné en incluant celles obtenues aux autres marqueurs.
Enfin, remarquons que cette procédure de prédiction pourrait également être incluse à chaque itération de l'algorithme BSAH lui-même afin de gérer les données manquantes dans le jeu de donnée.

type "fine mapping" comme proposées au chapitre précédent. Notons que dans ce cas, nous pourrions étendre aux positions non observées la prédiction des allèles.

Structuration et études d'association

Nous n'avons malheureusement pas eu le temps, au cours de cette thèse, d'étudier plus en détail la prise en compte de la structuration génétique dans les études d'association. Jusqu'à présent nous avons admis au cours de nos analyses que, l'inclusion des estimations des proportions d'admixture comme covariables dans les modèles, permettait de prévenir au mieux les effets confondants de la structure. Nous sommes conscients que cette hypothèse nécessite d'être étayée par des simulations. Plus particulièrement, nous pensons qu'il serait utile de comparer cette approche à des méthodes intégrant l'incertitude des paramètres de la structure comme celle proposée par Hoggart *et al.* (2003) (voir Encadré 7). Et cela afin de déterminer notamment si ces dernières méthodes, généralement bien plus onéreuses en temps machine, apportent un gain substantiel. Cette réponse déterminera par le même coup la faisabilité de tests par permutation : il va de soi que si pour chaque marqueur à évaluer il est nécessaire de réaliser une intégration sur la distribution *a posteriori* des paramètres d'admixture, les tests par permutation seront bien trop conséquents à mettre en œuvre.

Néanmoins, il pourrait être raisonnable de débuter par une analyse simple en incluant directement l'estimation des paramètres d'admixture, puis pour les résultats significatifs obtenus, d'affiner les tests en appliquant la stratégie décrite dans l'Encadré 7. Cette idée s'appuie sur notre intuition que la méthode simple doit éventuellement conduire à un ensemble de tests significatifs un peu plus important que celui des méthodes plus élaborées (c'est-à-dire à un nombre de faux-positifs un peu plus élevé). Ainsi, l'approche simple pourrait servir de filtre amont avant d'envisager des approches plus fines.

Enfin, des problèmes peuvent également survenir lorsque le polymorphisme évalué est colinéaire à la structuration (tout du moins celle inférée). Ce peut-être le cas dans des populations où la structure reflète des répartitions géographiques corrélées à la variation du caractère d'intérêt (par exemple la précocité de floraison est souvent très liée aux zones géographiques Camus-Kulandaivelu *et al.* (2006)). Dans ce cas, la structuration expliquant à elle seule une grande part de la variation du caractère, les tests d'association seront d'autant moins puissant que le polymorphisme testé sera colinéaire à la structure ; il conviendrait alors de disposer de méthodes de test alternatives. Autrement dit, nous souhaiterions savoir si cette colinéarité résulte du seul fait du hasard ou si elle s'explique par une causalité sous-jacente ? Cette question est assez proche des problématiques de cartographie de gène dans des populations admixées utilisée en génétique humaine (voir la revue récente de Smith and O'Brien (2005)). Il serait alors intéressant d'étudier la possibilité d'intégrer ces méthodes dans ce contexte. D'autre part, des méthodes alternatives pourraient également s'appuyer sur l'étude de distributions empiriques de statistiques décrivant la répartition non-aléatoire de polymorphismes neutres en fonction de la structure. Ces distributions constituant un jeu d'hypothèses

Encadré 6 : Études d'association et structure : "Score test"

Si la modélisation de l'admixture par HOGGART *et al.* (2003) est purement bayésienne, ils proposent en revanche une stratégie hybride bayésienne/fréquentiste afin de prendre en compte l'incertitude sur les paramètres d'admixture dans les tests d'association. Cette approche repose sur la méthode du test du score ("score test") et permet d'intégrer numériquement le score sur la distribution *a posteriori* des variables cachées (c'est-à-dire l'origine des locus utilisées pour inférer la structure). Nous présentons ici une méthode similaire qui en résume l'esprit.
Soient X_m les données de marqueurs pour inférer la structure et X_g les données de marqueurs à tester. On notre Y le vecteur des phénotypes. Soient P et Q les paramètres de la structure et Z_m les variables cachées indiquant l'origine des locus marqueurs de X_m. La vraisemblance des observations s'écrit alors :

$$L_o(Y, X_g, X_m) = \sum_Z \mathsf{Pr}(Y, X_g, Q, P|Z)\mathsf{Pr}(Z|Q, P, X_m)$$

avec

$$\mathsf{Pr}(Y, X_g, Q, P|Z) = \prod_{i=1}^{N} \phi(\frac{Y_i - \mu - \alpha Q - \beta X_g}{\sigma_e})$$

où σ_e est la variance résiduelle et ϕ la densité d'une loi normale centrée réduite. On note alors la log-vraisemblance complète des observations :

$$L_c(Y, X_g, Q, P; \theta) = \log(\mathsf{Pr}(Y, X_g, Q, P|Z))$$

avec $\theta = (\mu, \alpha, \beta)$ les paramètres du modèle correspondant à l'hypothèse alternative ($\beta \neq 0$), et on note $\theta_0 = (\mu, \alpha)$ ceux du modèle sous l'hypothèse nulle ($\beta = 0$). L'idée du score test repose sur l'hypothèse qu'au voisinage du maximum de la log-vraisemblance, celle ci peut s'approximer par un développement de Taylor au deuxième ordre (approximation locale). Par conséquent, si θ_0 est bien le maximum de vraisemblance, le score, noté $U_{|\theta=\theta_0}$ (c'est à dire le gradient de la log-vraisemblance évalué en θ_0) sera "voisin" de zéro. Pour tester comment le score "voisine" zéro, on suppose qu'il est distribué normalement autour de zéro, et sous l'hypothèse nulle cette distribution normale à pour matrice de variance-covariance la matrice d'information notée $V_{|\theta=\theta_0}$ (c'est-à-dire moins la matrice hessienne, ou la matrice des dérivées secondes, de la log-vraisemblance évaluée en θ_0). La statistique de test s'écrit ainsi :

$$s = ({}^{\mathrm{T}}UV^{-1}U)_{|\theta=\theta_0}$$

et s est asymptotiquement distribuée selon un χ^2 avec $q = |\theta| - |\theta_0|$ degré de libertés. Pour obtenir le score test correspondant à la log-vraisemblance observée $L_o(Y, X_g, X_m)$, il nous faut donc intégrer sur toutes les configurations de Z possibles, pondérées par leurs probabilités *a posteriori*. Cela peut s'effectuer à l'aide d'un algorithme EM stochastique tel que décrit dans l'encadré 5. Dans ce cas, on note $U_1, \ldots, U_B$ et $V_1, \ldots, V_B$ les B matrices de score et d'information (complète) calculées à chaque itération de l'algorithme EM stochastique. Le score U et l'information observée V s'obtiennent alors de la manière suivante :

$$\begin{cases} U &= \frac{1}{B}\sum_{b=1}^{B} U_b \\ V &= \frac{1}{B}\sum_{b=1}^{B} V_b - \frac{1}{B-1}\sum_{b=1}^{B}(U - U_b)^2 \end{cases}$$

le deuxième terme dans V traduisant la perte d'information due aux données manquantes Z (c'est-à-dire due au fait que l'on ne connaît pas *a priori* l'origine des individus).
Remarquons que cette stratégie a été également utilisée par SCHAID *et al.* (2002) pour intégrer l'incertitude sur la phase dans les tests d'association.

nulles par rapport auxquelles les valeurs des polymorphismes testés pourraient être comparées.

Chapitre 8

Conclusion générale

Bien que les études d'association ne soient qu'à leur début chez les végétaux, les prochaines années s'accompagneront sûrement d'un nombre croissant de publications de résultats dans ce domaine, tout du moins chez les espèces modèles (Gupta *et al.* (2005)). Mais si elles ouvrent une voie prometteuse à l'étude du déterminisme génétique de caractères quantitatifs, à l'heure actuelle, l'empirisme des études sur le DL ainsi que l'opacité des structures génétiques chez les plantes, invitent à la prudence. Il est probable que, de manière concomitante à la publication de résultats et qu'à l'instar de la génétique humaine, la controverse sur leurs reproductibilités gagne progressivement la génétique végétale. Si nous continuons à suivre le fil méthodologique qui a guidé jusqu'à alors la génétique végétale sur les traces de la génétique humaine, la méta-analyse des données issues des études d'association pourrait s'imposer comme une réponse pertinente à la question de la reproductibilité - en génétique humaine, de plus en plus d'articles témoignent en faveur de cette démarche (Lohmueller *et al.* (2003); Munafo and Flint (2004)). La méta-analyse a ici un double objectif statistique et cognitif : d'une part évaluer le degré d'hétérogénéité des résultats, et d'autre part établir des pronostics sur la cause des "congruences" entre études d'association.

Chez les végétaux, la méta-analyse pourrait alors intégrer une dimension supplémentaire, en considérant, dans un seul et même processus d'analyse, les données issues de cartographie de QTL et celles provenant des études d'association. De prime abord, cette démarche pourrait paraître "tautologique", dans le sens où généralement, les gènes candidats sont sélectionnés à partir des données de QTL. Aussi, vouloir croiser à nouveau QTL et gènes candidats peut inspirer le sentiment désagréable de "boucler la boucle". Pour comprendre cette idée, il faut rappeler que, si l'étude des colocalisations entre QTL et gènes candidats procure une base sérieuse à la sélection de ces derniers, elle ignore la plupart du temps une dimension essentielle : la configuration allélique des gènes candidats dans les populations de croisement considérées.

Ce n'est que dans un deuxième temps, lors des études d'association à proprement parler, que le modèle allélique aux gènes candidats est établi à la fois en terme de variabilité et d'effets alléliques. Ainsi, pour l'ensemble des parents impliqués dans les croisements considérés, et en supposant que l'on dispose des haplotypes parentaux aux gènes candidats (HPGC), la méta-analyse pourrait alors explorer conjointement les données "QTL et HPGC" afin d'évaluer les congruences "gènes candidats - QTL" les plus vraisemblables. Nous pourrions ainsi établir, à partir des estimations des effets alléliques et de leurs configurations au sein des HPGC entre parents, un pronostic sur les profils de détection de QTL dans les croisements (ce pronostic intégrerait également les paramètres de croisement comme le type de famille et la taille). L'étude conjointe des profils estimés et des positions de QTL observées, pourrait ainsi aider à déterminer la nature des classes "gènes candidats - QTL" les plus probables.

Mais en supposant qu'une telle démarche soit envisageable dans les années à venir, son caractère séduisant ne doit pas masquer les difficultés qu'elle recouvre : les mécanismes sous-jacents au déterminisme de caractères complexes demeurant encore "cryptiques", il sera peut-être délicat, sur la seule base de modèles empiriques, de s'affranchir des effets confondants induits par l'hétérogénéité des données. Par exemple, certains QTL détectés peuvent ne pas avoir de correspondance causale directe avec les gènes candidats considérés.

Notamment, les détections de QTL étant conduites indépendamment dans des milieux différents, il pourrait être difficile de distinguer ce qui procède d'une interaction génotype x milieu, de ce qui résulte d'une absence ou d'une présence de signal de congruence entre le profil prédit et les QTL observés. Néanmoins, pour certains gènes majeurs, on peut s'attendre à ce que le bruit introduit par les interactions génotype x milieu ne masquent pas trop fortement l'effet du gène et que la plupart des profils prédits soient cohérents avec les QTL observés. D'autre part, la non prise en compte d'éventuelles interactions épistatiques peut également induire des différences entre profils prédits et QTL observés. Par exemple, on peut imaginer un cas où une configuration allélique particulière à deux gènes candidats distincts, dans une population, aurait conduit à la détection de deux positions de QTL distinctes. Mais que pris séparément, les profils prédits à chaque gène candidat ne témoignent d'aucun signal significatif.

Enfin, cette anticipation n'est qu'une manière possible d'interpréter les perspectives offertes par les études d'association et il est probable que les découvertes qui en découleront amèneront de nouvelles questions.

Sixième partie

Annexes

Chapitre 9

Annexe 1 : MetaQTL

MetaQTL : A Java package for meta-analysis of QTL mapping experiments

Authors : Jean-Baptiste Veyrieras [a,1], Julien Cornouiller[1], Bruno Goffinet[2], and Alain Charcosset[1]

[1] UMR, INRA UPS-XI INAPG CNRS Génétique Végétale, Ferme du Moulon, 91190 Gif-sur-Yvette, France
[2] MIA, INRA, Chemin de Borde Rouge BP52627, 31326 Castanet Tolosan Cedex, France

Keywords : genetic-map, QTL, meta-analysis, clustering, bioinformatics.

To be submitted to *Bioinformatics*

[a]to whom correspondence should be addressed

Abstract

Integration of results from multiple Quantitative Trait Loci (QTL) studies relative to a given trait or to several related traits is a key point to understand the genetic determinism of these traits. Up to now many efforts have been made to facilitate the storage, the compilation and the visualization of QTL detection results in public databases. Taking benefit from the amount of available results from QTL studies, we develop a new meta-analysis procedure that allows researchers to study QTL congruency into a well-established statistical framework. MetaQTL implements a series of programs which allow user both to carry out the meta-analysis and to display the results in various ways.

Introduction

Since the last decade, the advent of molecular markers have accelerated the pace of discovering the loci which are involved in the variation of quantitative traits. Quantitative Trait Loci (QTL) mapping usually begins with the collection of genotypic (based on molecular markers) and phenotypic data from a segregating population. First, from the genotypic data the markers are both ordered and positioned on a genetic map using standard linkage mapping approaches as implemented in valuable softwares (e.g Lander *et al.* (1987); Stam (1993); Schiex (1997)). Secondly, refinement of analytical methods have enabled QTL to be detected with more precision (see for instance Lander and D. (1989); Zeng (1994)). Nevertheless due to the limiting number of individuals and generations in usual experiments this approach generally leads to QTL locations with a confidence interval (CI) around 10 cM or more Kearsey and Farquhar (1998) which in plant genomes corresponds to thousand(s) of genes Sasaki *et al.* (2002).

Due to its relative simplicity and its compelling concept QTL mapping has been widely used and more and more QTL detection results are now available in public databases (e.g in maize at http ://www.maizegdb.org). One of the main purpose of these databases was to facilitate the comparison of different QTL detection results by providing both standard description of these results and ontologies (see for instance the trait ontology at http ://www.gramene.org/plant_onto
Relevance of comparative analysis of QTL studies have been illustrated by several authors Khavkin and Coe (1997, 1998); Lin *et al.* (1995). However this studies often relied on simple descriptive statistics.

This gap in QTL congruency study was partially bridged by Goffinet and Gerber (2000) who proposed a meta-analysis based approach to combine several QTL results. Their method makes it possible to evaluate how many "actual" QTL locations underlie the distribution of the observed QTL on the genome. This approach has been implemented in BioMercator by Arcade *et al.* (2004). This software first allows user to merge both markers and QTL onto a consensus map by means of an iterative projection procedure. Then the algorithm devised by Goffinet and Gerber (2000) can be applied to evaluate the likelihood of clustering the observed QTL in 1,2,3 or 4 groups. Afterward, the optimal number of clusters is selected by using a Akaike like criterion. Although original this approach suffers from the absence of indicator to assess the consensus map quality and from the limiting number of QTL clusters which can be explored.

Based on recent methodological developments, MetaQTL implements a series of Java programs in order to carry out more sophisticated QTL meta-analysis. All the programs in MetaQTL are command line programs. Each program does a small job and the user can combine the program as a group to do a complete analysis. Thanks to its flexible and modular implementation, MetaQTL could also be integrated in more elaborated softwares if needed.

QTL study database

First, before running meta-analysis one needs to store the different QTL studies into a database. To do this MetaQTL uses a simple multiple plain text files database. Each file corresponds to a table and the database is organized as follows :

- Experiment table : store descriptions on mapping experiments (name, population type and size, reference, ...).
- Genetic Map table : store for each mapping experiment the marker map.
- QTL table : store the detected QTL in each mapping experiment (position, confidence interval, r-square, ...).
- Trait Ontology table : use to describe how the traits are related together using a simple hierarchical relationship scheme.
- Marker Dictionary table : it is not uncommon in mapping experiments that a same marker is reported with different names. This table allows the user to specify a standard marker name to which several synonyms can be attached.

Once the database created, MetaQTL first checks if it is valid and then summarizes its contents in a set of XML files. All the programs of MetaQTL use these XML files as inputs. Utilities are provided to convert it in several plain text file formats if required.

QTL meta-analysis

The meta-analysis consists in three main steps.

Construction of a consensus marker map

MetaQTL implements an original approach based on a Weighted Least Square (WLS) strategy to integrate all the genetic marker maps into a single consensus map on which all the markers are ordered and positioned. A chi-square statistic is also returned which reflect the homogeneity of the marker interval distances among the experiments. Before performing the construction of the consensus map, MetaQTL allows user to compute some usefull statistics to assess the quality of the marker order between mapping experiments.

Projection of the QTL

QTL are generally projected by applying a simple homothetic rule. However when there are marker order or distance inconsistencies between input maps such a process can lead to dubious QTL projections. To remove this impediment MetaQTL proposed a dynamic algorithm which tracks the best QTL flanking

marker context for which the projection is optimal. The confidence interval of the QTL is resized according to a scaling ratio which takes into account the variation of the marker interval distances between the two maps conditionally to the QTL location on the original map.

QTL Clustering

MetaQTL implements two kinds of clustering algorithm. First, an EM-algorithm Dempster *et al.* (1977) based on a Gaussian mixture model can be applied to evaluate the likelihood of all the possible QTL clusterings. Contrary to Goffinet and Gerber (2000) this approach leads to a probabilistic QTL cluster memberships. Then usual model choice criteria in mixture context are proposed in order to determine the best QTL clustering. Second, standard hierarchical clustering algorithm are also provided, either by an average group linkage strategy or by a Ward's algorithm Ward (1963).

Visualization of the results

At each step the results of the analysis can be visualized : MetaQTL offers several programs to create figures from result files as illustrated in Figure 9.1.

This package (programs, sources and documentation) is distributed under the GNU Public License and any contribution to improve it is welcomed.

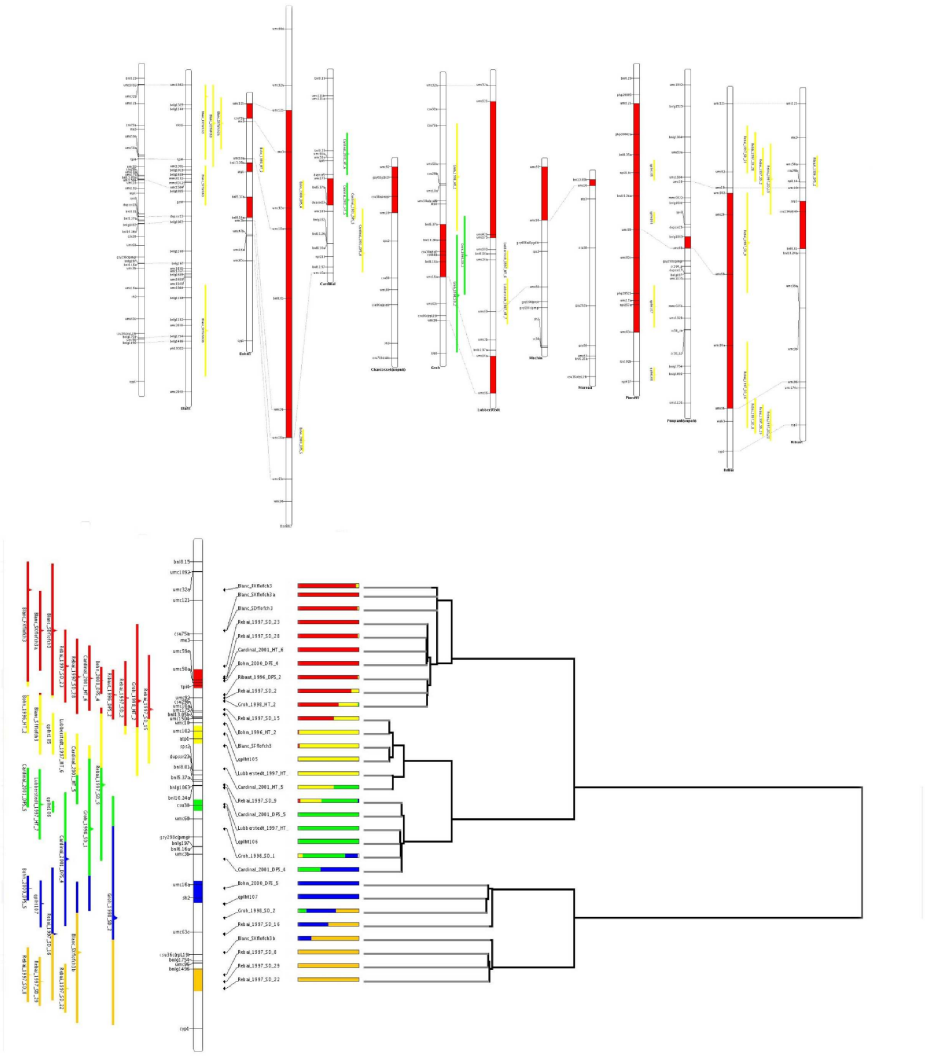


FIG. 9.1 – Overview of the meta-analysis on 30 QTL relative to flowering time in maize (chromosome 3) using MetaQTL. At the top left the consensus chromosome is displayed. The filled marker intervals on the input chromosomes stands for regions which showed significant deviation from the consensus model. Following, both Ward's algorithm and mixture based clusterings are depicted.

Chapitre 10

Annexe 2 : Libgda

Libgda : A C library and program package for statistical genetics

Authors : Jean-Baptiste Veyrieras [a,1]

[1] UMR, INRA UPS-XI INAPG CNRS Génétique Végétale, Ferme du Moulon, 91190 Gif-sur-Yvette, France

Keywords : statistical genetic, data analysis, bioinformatics, computational biology.

To be submitted to *BMC Bioinformatics*

[a]to whom correspondence should be addressed

Abstract

The C library libgda provides a high and low-level interface to handle and explore genetic data. It can be used to compute usual descriptive statistics (e.g allele or genotype frequencies, diversity indicators or pairwise linkage disequilibrium) or to build more sophisticated methods from the large variety of "bricks" provided by the library. The flexibility of the library is illustrated by a series of programs for which implemented procedures range from usual PCA, discrete PCA to advanced haplotypic data modelling. Libgda aims to help researchers to develop in-house methodologies in the field of statistical genetics.

Introduction

In genetic data analysis, the range of interest is vast and varies from population evolutionary studies to more applied issues such as association studies between phenotypic and genetic variation. Current open-source applications

generally focus on a peculiar issue rather than providing an unified framework in which different kinds of analyses can be performed. Nevertheless, more and more statistical genetic issues are now at the interface of several fields in genetics. For example, association studies may require to use theoretical developments from population genetics to better explain correlation patterns between traits and genotypes. Therefore, we anticipate that the use of open-source library which could provide a flexible interface to handle various kinds of genetic data, might help researchers to rapidly develop and distribute new prototype softwares for genetic data analysis.

In this article we present libgda, a library which aims at performing various manipulations of genetic data and carrying out different kinds of analysis procedures based on both low-level and high-level "bricks" provided by the library. This later is written in C, is compliant with the GNU standards and packaged as a dynamic library that can be installed on most Unix and Linux distribution. First, we briefly describe the core of the library and then we present a series of programs in order to carry out different kinds of statistical genetic analyses.

Implementation

Libgda is divided in several low-level modules which are dedicated to specific tasks. Apart from very low-level functions (such as memory management or IO utilities), each module is made up of a few C structures which help to facilitate the use of the functionalities of the library. In particular, libgda provides a very flexible representation of the fundamental "entities" involved in genetic data analysis : individual, locus, allele. Each of these entities is represented by a "node" like structure which can be pushed into a simple double chained list or included in a more subtle design such as tree or graph. Besides, each node can be potentially a container for a node of another type. For example, an individual can be represented by a single node which contains a list of locus nodes and each locus node can point to one or more allele nodes (this can be a way to describe the genotype of the individual).

Besides, the library offers some utilities to handle data matrix in several formats, in particular factor matrix where rows and columns can be classified according to relevant factors. Libgda also implements standard routines from Press *et al.* (1992) which can be used to develop new methodologies based on either usual linear algebra procedures, or on likelihood maximization strategy.

Since graphical representation is a major issue in data analysis, libgda provides an original graphic 2D interface which makes it possible to draw complex graphics using basic shapes such as line, rectangle, circle or ellipse. Currently only SVG output is available but development of other kinds of format are in progress (in particular postscript format).

Simulation are now widely used to explore the properties of evolutionary scenarios or to evaluate the performance of new data analysis methods. Therefore, libgda provides a forward-time population genetics simulation interface which can be easily used to create and investigate elaborated population evolutionary histories.

Finally, in order to facilitate the communication between the library and other external high-level procedures, the main structures of the library are binding to a XML representation by means of an internal implementation of marshalling and unmarshalling procedures. To do so, the library requires that libxml2 (http ://xmlsoft.org/) has been previously installed. We anticipate that this is likely to accelerate the pace of the integration of the library functionalities into more elaborated softwares.

Results

From the libgda low-level API, we have developed high-level structures in order to carry out specific genetic data analysis procedures. Since genetic data analysis may require as a preliminary to integrate several sources of information, we first describe how multiple data sets can be gathered into a single data analysis project. We then detail some of the data analysis procedures implemented from ligbda. For both the project management and the data analysis procedures correspond a series of programs which functions are summarized in Table 10.1.

Project

The concept of project is fundamental in data analysis. It allows the user to gather several data sets into a single data analysis framework. In the field of genetic data analysis, information can be classified into two main categories :

- genotypic data : results of genotyping individuals for a given set of molecular marker loci. These loci can be organized into a genetic or physical map. Besides, the type of marker and measurement may vary depending on the kind of population typed. Generally, two genotyping procedures can be distinguished :i) standard genotyping where an accession corresponds to a unique individual, ii) pool or frequency genotyping where an accession stands for a bulk of potentially distinct individuals.
- phenotypic data : measurement of traits in a population. This measurement may have been done using a particular “trait design” framework (e.g using several distinct environments).

Here, a project is specified using an XML document which is divided in different sections :

– `profile` : the profile allows the user to specify the name of the current project, the ploidy of the data to be analyzed and the default values of the attributes which describe the data sets (see below).
– `geno-data` : this section contains the description of the raw genotypic data. First, the data are organized in `locus-group`. A `locus-group` corresponds to a set of marker loci which have been typed together and are related to a same experiment or to a same region of the genome. This offers a convenient way to classify marker loci in the current project into a limited number of categories such as gene regions or marker types. If some marker loci have been positioned into a genetic or physic map, it can be specified using the `locus-map` section for which two kinds of format are possible : the positions on the marker map are given with regard to an arbitrary zero reference on each chromosome or the marker loci are ordered according to the map and adjacent distance between marker loci are given. The map can contain both genetic and physic positions by using the distance unit attribute. Then, each data set is embedded into a `sample` section which corresponds to the population of individuals that have been typed. This section, which must be defined by a unique name, can contain several `sample-data`. This later is linked to a given `locus-group`. This means that this population (or sample of individuals) has been characterized by several locus groups. Each data set for each locus group is described using the `data-description` tag for which the following properties can be specified :
  – `missing-data` : defines the string of characters used to encode missing data in the raw data set.
  – `gametic-phase` : 0 if gametic phase is not known, 1 otherwise. The default value is 1.
  – `txt-separator` : the character used to separate the genotypes at different loci (the locus separator). Possible values are `whitespace`, `tab` and `none` (e.g for long DNA sequence).

  Finally, comes the raw `data` set itself which can be directly included in the project file or, in order to clarify the structure of the project, an external link can be done to another file which contains the data set.
– `pheno-data` : this section can be used to specify phenotypic data which must be included into the current project. Here, the phenotypic data can consist in several `trait-measure`. A `trait-measure` corresponds to a measurement of a set of traits. Besides, by using the `trait-design` tag one can add to the trait variables some extra variables such as experimental factors or other covariates associated to this trait measurement. In this case both factors and trait variables must be specified into a single file which can be directly included in the project file or specified, like for genotypic data, using an external reference to a file. Note

that factor variables can be either categorical (e.g a given treatment), mixing (e.g probabilistic membership to populations) or continuous factors.

Once a project have been created (in Figure 10.1 we give an example of a XML project declaration), the program GDADB can be used to check its validity and to create a local data base, called the project data base, in which all the data are summarized. This local data base is encoded in full XML and all the other programs use it to get data content.

Data view

A central point of the architecture of the package is the concept of data view. A data view is a structure which makes it possible to handle and manage data points by merging in an unified representation both genotypic and phenotypic data. The program GDAVDB allows the user to make queries on the project data base in order to extract the corresponding genotypic data and to merge them into a single XML file. For example, one can merge data from several samples for a given locus groups or several locus group from the same sample. A large variety of filter procedures is available : remove a given set of loci and/or individuals, remove loci with a number of alleles higher than a given value, remove alleles at loci with a frequency beyond a given threshold, remove individuals and/or loci which show a proportion of missing data greater than a given value, compress adjacent loci which show identical pattern of variation into a single locus, clusterize individuals which exhibit identical genotypes at all loci. Finally, phenotypic data can be linked to the data view by using the program GDATDB .

Analysis result data base

Each data analysis procedure yields a "result object" which is stored into a XML result data base. Each result object contains a meta-information header which gives some details on the type of the result together with some of its basic parameters. The program GDARDB makes it possible to display the meta-information of any results stored in a given data base. These results may have been obtained from different data analysis programs. This offers a convenient way to group into a limited number of files all the analyses made on a data view. Finally, all the result objects can be translated into plain text file using the program GDAX2A .

PCA

The package provides a high-level interface to carry out principal component analysis (PCA) using : i)algorithm based on singular value decomposition (SVD), ii) EM-algorithm based on the Wiberg's method WIBERG (1976) in order to manage missing data. The program GDAPCA makes it possible to

carry out PCA on the empirical covariance or correlation matrix of marker loci or individuals. It is worth noting that the Wiberg's method can be used to infer missing data sites before performing other data analyses (all the programs of the package are able to deal with probabilistic description of marker data). Based on this PCA framework, the program GDAPCA-TG allows user to identify either marker loci or individuals which capture most of the genetic diversity in a given data view.

DPCA

Although PCA offers a well-established statistical framework to explore correlation structures in multivariate analysis, marker data are generally discrete or compositional. Therefore the program GDADPCA implements an original algorithm based on the recent development of discrete principal component analysis (DPCA). Here, the implementation is restricted to the binary case since either haplotypic or genotypic marker data can be encoded using a binary code. The idea of DPCA is to extract a given number of independent populations in which individuals are assigned probabilistically. For example, it can be used to infer population structure assuming either a mixture or an admixture model. In this case the program GDADPCA-R can be used to discriminate the minimal number of populations which are required to capture most of the correlation in the initial marker data. Lastly, the program GDADPCA-P allows the user to visualize DPCA results in various ways.

BSAH

One of the most challenging issue of genetic data analysis is the identification of the ancestral source of diversity observed in current data sets. The package implements a new algorithm, called BSAH, which aims to extract ancestral haplotypes from a set of observed haploid sequences. The program GDABSAH makes it possible to infer these ancestral haplotypes and to estimate the recombination rate and the mutation rate with regard to this ancestral haplotype model. Then GDAHZIP program can be plugged on the output of GDABSAH in order to evaluate the "entropy" of the ancestral model using an efficient lossless compression strategy. By checking the output of GDAHZIP one can then discriminate the ancestral model which best fits the observed haplotypes (recall that BSAH is an heuristic). Note that this procedure can also be used if one aims to minimize storage cost of long DNA sequence data sets in data base.

Association studies

The association study framework implemented here aims to be flexible. First, simple regression marker per maker can be performed by applying

GDACR on a single data view. Then, association can be tested using a usual F-test (see for instance FAN *et al.* (2005)) with regard to the following hypotheses :

$$\text{H1} : Y = \mu + aX_c + bX_g + \epsilon$$
$$\text{H0} : Y = \mu + aX_c + \epsilon$$

where Y is the vector of phenotypes for the current trait, X_c is the set of covariates which have been specified in the `trait-design` section (e.g sex or age), and X_g the set of variables corresponding to the polymorphism at the marker being tested (this can be either discrete indicators or probabilistic variables). The program allows user to control the type of encoding for X_g : i)genotypes, ii)allelic doses (similar to genotype encoding if genetic effects are additive).

Second, if one of the covariates is a mixture like factor such as population admixture, the association can be evaluated using a mixture of regression approach which likelihood is given by

$$L(Y, X_g; \mu, b) = \prod_{i=1}^{N} \left(\sum_{k=1}^{K} q_{ik} \mathsf{Pr}(Y^{(i)}, X_g^{(i)}; \mu_k, b) \right)$$

where q_{ik} is the membership probability of individual i for population k, K the total number of populations, $\mathsf{Pr}(Y^{(i)}, X_g^{(i)}; \mu_k, b) = \phi[(Y^{(i)} - \mu_k + bX_g^{(i)})/\sigma_e]$ is the probability of the observed phenotype of individual i in the k^{th} population, ϕ is the density function of a centered normalized Gaussian, and σ_e the residual standard deviation. Then the association test is based on a standard likelihood ratio :

$$\lambda = -2 \log \left[\frac{L_1(Y, X_g; \mu, b \neq 0)}{L_0(Y, X_g; \mu, b = 0)} \right]$$

where L_1, L_0 and their respective parameters are computed by applying an EM-algorithm DEMPSTER *et al.* (1977).

Haplotype based analyses can also be carried out inside GDACR . To do so, the program GDAHM must be first used to compute the haplotype model which will be used in the association test. Here, an haplotype model may consist in :

- a full range model : X_g is the set of observed haplotypes along the entire studied region.
- a sliding window model : X_g represents the haplotypes in a sliding window which size can be either defined by a given number of markers or by a given physical or genetic distance.
- a block model : X_g describes the haplotypes listed in each block. The block model may have been obtained either by applying standard haplotype block partition algorithms (see for instance ZHANG *et al.* (2005)) or from the ancestral partition returned by the BSAH algorithm.

Furthermore, for any haplotype model returned by GDAHM , the program GDATREE can be previously run to perform hierarchical clustering of the haplotypes by means of a neighbor-joining algorithm SAITOU and NEI (1987). It can be used by GDACR to carry out association test by moving up through the tree in order to find the partition of the haplotypes which best fits the data. This can be interpreted as a backward hierarchical procedure as follows :

- initial step : X_g includes all the observed haplotypes.
- i^{th} step : X_g is obtained by grouping the haplotypes at the i^{th} level in the tree (the first level is the level in which all the haplotypes are distinct, i.e the initial step).
- last step : select the i^{th} partition of haplotypes which yields the best p-value.

Note that this approach is similar to the cladistic analysis devised by DURRANT *et al.* (2004)).

Conclusion

The libgda library is actively under development and new updated versions will be released in the next months. In particular current work is focusing on the addition of usual statistical analyses in population genetics such as computation of diversity indices, linkage disequilibrium measures, test of Hardy-Weiberg equilibrium or Tajima's neutrality test. Another ongoing project consists in the development of a full XML interface to describe population simulation scenario.

Finally, external contribution are welcome and we hope that the library functionalities will match the interest of researchers involved in computational biology.

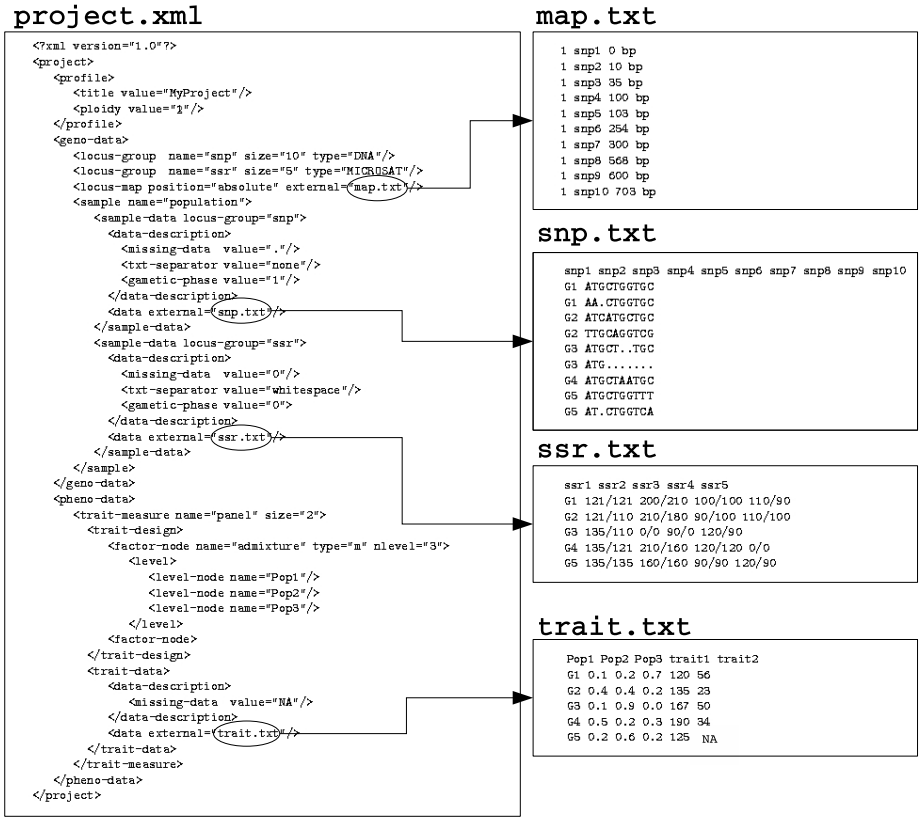


FIG. 10.1 – Illustration of a data analysis project.

Tab. 10.1 – Summary of the command line programs implemented with libgda.

Program	Definition
	Module
	Data management
GDADB	Convert project file into a XML data base
GDAVDB	Create a data view from the project data base
GDAVDB-I	Import genotypic data from a single data file
GDATDB	Link trait measures to a data view
GDAVDB-TR	Trim or Prune a data view
GDAA2X	Import plain text file into XML result
GDAX2A	Export XML result to a plain text file
	Result management
GDARDB	Display meta-information of a result data base
GDAVRDB	Convert results into a data view (if possible)
	Data analysis
GDAVDB-ST	Compute and display some basic statistics on the data view
GDAPCA	PCA on marker/individual data matrix
GDAPCA-TG	Select markers/individuals using a PCA-varimax procedure
GDADPCA	DPCA on marker data matrix
GDADPCA-R	Compute model choice criterion for DPCA results
GDADPCA-CR	Cross PCA and DPCA results
GDABSAH	Blind separation of ancestral haplotypes
GDAHZIP	Compress haplotypic data from BSAH results
GDATREE	Hierarchical clustering by neighbor joining
GDAHM	Compute haplotypic model from data view or results
GDACR	Linear and mixture regression model for association studies
	Data and Result Visualization
GDAVDB-P	Visualize data view
GDADPCA-P	Visualize DPCA results
GDABSAH-P	Visualize BSAH results

Bibliographie

Aitkin, M. and D. Rubin, 1985 Estimation and Hypothesis Testing in Finite Mixture Models. Journal of the Royal Statistical Society **47** : 67–75.

Akaike, H., 1973 Information theory and an extension of the maximum likelihood principle. 2nd Inter. Symp. on Information Theory : 267–281.

Akaike, H., 1992 *Breakthroughs in Statistics*, volume 1, chapter Information Theory and an Extension of the Maximum Likelihood Principle. Springer-Verlag, London, 610–624.

Allison, D. B. and M. Heo, 1998 Meta-analysis of linkage data under worst-case conditions : a demonstration using the human OB region. Genetics **148** : 859–865.

Andersen, J. R., T. Schrag, A. E. Melchinger, I. Zein and T. Lubberstedt, 2005 Validation of Dwarf8 polymorphisms associated with flowering time in elite European inbred lines of maize (Zea mays L.). Theor Appl Genet **111** : 206–217.

Anderson, E. and W. L. Brown, 1952 The history of the common maize varieties of the united states corn belt. Agricultural History **26** : 2–8.

Anderson, E. C. and J. Novembre, 2003 Finding haplotype block boundaries by using the minimum-description-length principle. Am J Hum Genet **73** : 336–354.

Aquadro, C. F., S. F. Desse, M. M. Bland, C. H. Langley and C. C. Laurie-Ahlberg, 1986 Molecular population genetics of the alcohol dehydrogenase gene region of Drosophila melanogaster. Genetics **114** : 1165–1190.

Arcade, A., A. Labourdette, M. Falque, B. Mangin, F. Chardon *et al.*, 2004 BioMercator : integrating genetic maps and QTL towards discovery of candidate genes. Bioinformatics **20** : 2324–2326.

Barriere, Y., G. Aurel, M. Briand, D. Denoue and A. Gueu, 2005 QTL mapping for cell wall constituents and cell wall digestibility in maize recombinant inbred line progeny F838 X F286 harvested at an early forage stage of maturity. Technical report, INRA.

Beaumont, M. A., 2004 Recent developments in genetic data analysis : what can they tell us about human demographic history ? Heredity **92** : 365–379.

Beavis, W., 1994 The power and deceit of QTL experiments : lessons from comparative QTL studies. In *Proceedings of the Forty-Ninth Annual Corn and Sorghum Industry Research Conference*. American Seed Trade Association, Washington, DC, 250–266.

Benjamini, Y. and Y. Hochberg, 1995 Controlling the False Discovery Rate : a pratical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B **57** : 289–300.

Benjamini, Y. and D. Yekutieli, 2005 Quantitative trait Loci analysis using the false discovery rate. Genetics **171** : 783–790.

Biernacki, C. and G. Govaert, 1998 Choosing Models in Model-based Clustering and Discriminant Analysis. Technical Report 3509, INRIA, France.

Blanc, G., L. Moreau, B. Mangin and A. Charcosset, 2003 QTL detection in connected populations of maize. In A. Caruna, editor, *12th Meeting of the EUCARPIA Section of Biometrics in Plant Breeding*. Spain.

Boddeker, I. R., H. H. Muller, R. Kress, F. Geller, A. Ziegler *et al.*, 2001 The use of sequential designs in genome scans for asthma susceptibility loci with affected sib pairs. Genet Epidemiol **21 Suppl 1** : 49–54.

Bohn, M., M. M. Khairallah, D. Gonzalez-de Leon, D. Hoisington, H. F. Utz *et al.*, 1996 QTL Mapping in Tropical Maize : I. Genomic Regions Affecting Leaf Feeding Resistance to Sugarcane Borer and Other Traits. Crop. Sci. **36** : 1352–1361.

Bohn, M., B. Schulz, R. Kreps, D. Klein and A. E. Melchinger, 2000 QTL Mapping of resistance agaisnt the European corn borer (Ostrinia nubilalis H.) in early maturing European dent germplasm. Theor. Appl. Genet. **1001** : 907–917.

Bouchez, A., F. Hospital, M. Causse, A. Gallais and A. Charcosset, 2002 Marker-assisted introgression of favorable

alleles at quantitative trait loci between maize elite lines. Genetics **162** : 1945–1959.

Bozdogan, H., 1987 Model Selection and Akaike Information Criteria (AIC) : The general theory and its analytic extensions. Psychometrika **52** : 345–370.

Bozdogan, H., 1990 On the Information-Based Measure of Covariance Complexity and its Application to the Evaluation of Multivariate Linear Models. Communication in Statistics, Theory and Methods **19** : 221–278.

Buckler, E. S. t. and J. M. Thornsberry, 2002 Plant molecular diversity and applications to genomics. Curr Opin Plant Biol **5** : 107–111.

Buntine, W., 2002 Variational extensions to EM and multinomial PCA. In *ECML 2002*.

Buntine, W. and A. Jakulin, 2004 Applying discrete PCA in data analysis. In Banff, editor, *UAI-2004*. Canada.

Buntine, W. and A. Jakulin, 2005 Discrete Principal Component Anlysis. Technical report, HIIT.

Buntjer, J. B., A. P. Sorensen and J. D. Peleman, 2005 Haplotype diversity : the link between statistical and biological association. Trends Plant Sci **10** : 466–471.

Burnham, K. P. and D. R. Anderson, 2002 *Model Selection and Multimodel Inference : A Pratical Information-Theoretical Approach*, volume 33. Springer-Verlag, New-York, 2 edition.

Burnham, K. P. and D. R. Anderson, 2004 Multimodel Inference, Understanding AIC and BIC in Model Selection. Sociological Methods & Research **33** : 261–304.

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.

Cardinal, A. J., M. Lee, N. Sharopova, W. L. Woodman-Clikeman and M. J. Long, 2001 Genetics mapping and analysis of quantitative trait loci for resistance to stalk tunneling by the European corn borer in maize. Crop. Sci. **41** : 835–845.

Carlson, C. S., M. A. Eberle, L. Kruglyak and D. A. Nickerson, 2004 Mapping complex disease loci in whole-genome association studies. Nature **429** : 446–452.

Cattell, R. B., 1966 The scree test for the number of factors. Multivariate Behavioral Research **1** : 245–276.

Cavalli-Sforza, L., P. Menozzi and A. Piazza, 1994 *The History and Geography of Human Genes*. Princeton University Press, Princeton, NJ.

Celleux, G. and G. Govaert, 1992 A classification EM algorithm and two stochastic versions. Comput. Stat. Data Anal. **14** : 315–332.

Charcosset, A., M. Causse, L. Moreau, D. Vienne and A. Gallais, 2000 Epistatic effect of genetic background on QTL expression in connected populations. Epistatis in Connected Population.

Chardon, F., D. Hourcade, V. Combes and A. Charcosset, 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.

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. Genetics **168** : 2169–2185.

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

Chikhi, L., M. W. Bruford and M. A. Beaumont, 2001 Estimation of admixture proportions : a likelihood-based approach using Markov chain Monte Carlo. Genetics **158** : 1347–1362.

Clark, R. M., S. Tavare and J. Doebley, 2005 Estimating a nucleotide substitution rate for maize from polymorphism at a major domestication locus. Mol Biol Evol **22** : 2304–2312.

Clement, M., D. Posada and K. A. Crandall, 2000 TCS : a computer program to estimate gene genealogies. Mol Ecol **9** : 1657–1659.

Conneally, P., J. Edwards, K. Kidd, J. Lalouel and N. Morton, 1985 Reports of the committee methods of linkage analysis and reporting. Cytogent. Cell. Genet. **40** : 356–359.

Consortium., T. I. H., 2003 The International HapMap Project. Nature **426** : 789–796.

Crepieux, S., C. Lebreton, B. Servin and G. Charmet, 2004 Quantitative trait loci (QTL) detection in multicross inbred designs : recovering QTL identical-by-descent status information from marker data. Genetics **168** : 1737–1749.

Cutler, A. and M. P. Windham, 1993 Information-Based Validity Functionals for Mixture Analysis. In B. H., editor, *Proceedings of the first US-Japan Conference on the Frontiers of Statistical Modeling*. Kluwer, Amsterdam, 149–170.

Daly, M. J., J. D. Rioux, S. F. Schaffner, T. J. Hudson and E. S. Lander, 2001 High-resolution haplotype structure in the human genome. Nat Genet **29** : 229–232.

Darvasi, A. and M. Soller, 1995 Advanced intercross lines, an experimental population for fine genetic mapping. Genetics **141** : 1199–1207.

Darvasi, A. and M. Soller, 1997 A simple method to calculate resolving power and confidence interval of QTL map location. Behav Genet **27** : 125–132.

Darvasi, A., A. Weinreb, V. Minke, J. Weller and M. Soller, 1993 Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated map. Genetics **134** : 943–951.

de Bakker, P. I. W., R. Yelensky, I. Pe'er, S. B. Gabriel, M. J. Daly *et al.*, 2005 Efficiency and power in genetic association studies. Nat Genet **37** : 1217–1223.

Dempster, A., N. Laird and D. Rubin, 1977 Maximum Likelihood from incomplete data via the EM algorithm (with discussion). J. Roy Stat. Soc. **B 39** : 1–38.

Doebley, J. F., M. M. Goodman and C. W. Stuber, 1986 Exceptional genetic divergence of Northern Flint Corn. Am. J. Bot. **73** : 64–69.

Duff, I. S., 1986 Users' Guide for the Harwell-Boeing Sparse Matrix Collection. Technical report, CERFACS, Toulouse, France.

Dupuis, J. and D. Siegmund, 1999 Statistical methods for mapping quantitative trait loci from a dense set of markers. Genetics **151** : 373–386.

Durrant, C., K. T. Zondervan, L. R. Cardon, S. Hunt, P. Deloukas *et al.*, 2004 Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. Am J Hum Genet **75** : 35–43.

El-Mabrouk, N. and D. Labuda, 2004 Haplotypes histories as pathways of recombinations. Bioinformatics **20** : 1836–1841. Evaluation Studies.

Etzel, C. and R. Guerra, 2003 Meta-analysis of genetic-linkage of quantitive trait loci. Am. J. Hum. Genet. **71** : 56–65.

Ewens, W. J. and R. S. Spielman, 1995 The transmission/disequilibrium test : history, subdivision, and admixture. Am J Hum Genet **57** : 455–464.

Falque, M., L. Decousset, D. Dervins, A.-M. Jacob, J. Joets *et al.*, 2005 Linkage mapping of 1454 new maize candidate gene Loci. Genetics **170** : 1957–1966.

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

Fan, R., J. Jung and L. Jin, 2005 High Resolution Association Mapping of Quantitative Trait Loci, A Population Based Approach. Genetics .

Fearnhead, P. and P. Donnelly, 2001 Estimating recombination rates from population genetic data. Genetics **159** : 1299–1318.

Flint-Garcia, S. A., J. M. Thornsberry and E. S. t. Buckler, 2003 Structure of linkage disequilibrium in plants. Annu Rev Plant Biol **54** : 357–374.

Forney, G. D., 1973 The Viterbi Algorithm. In *IEEE*, volume 61. 268-278.

Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy *et al.*, 2002 The structure of haplotype blocks in the human genome. Science **296** : 2225–2229.

Gaut, B. S. and A. D. Long, 2003 The lowdown on linkage disequilibrium. Plant Cell **15** : 1502–1506.

Goffinet, B. and S. Gerber, 2000 Quantitative trait loci : a meta-analysis. Genetics **155** : 463–473.

Griffiths, R. C. and P. Marjoram, 1996 Ancestral inference from samples of DNA sequences with recombination. J Comput Biol **3** : 479–502.

Groh, S., M. M. Khairallah, D. Gonzales-de Leon, M. Willcox, C. Jiang *et al.*, 1998 Comparison of QTLs mapped in RILs and their test-cross progenies of tropical maize for insect resistance and agronomic traits. Plant. Breed. **117** : 193–202.

Gupta, P. K., S. Rustgi and P. L. Kulwal, 2005 Linkage disequilibrium and association studies in higher plants : present status and future prospects. Plant Mol Biol **57** : 461–485.

Hagenblad, J., C. Tang, J. Molitor, J. Werner, K. Zhao *et al.*, 2004 Haplotype structure and phenotypic associations in the chromosomal

regions surrounding two Arabidopsis thaliana flowering time loci. Genetics **168** : 1627–1638.

Haldane, J. and C. Waddington, 1930 Inbreeding and Linkage. Genetics **16** : 357–374.

Halperin, E., G. Kimmel and R. Shamir, 2005 Tag SNP selection in genotype data for maximizing SNP prediction accuracy. Bioinformatics **21 Suppl 1** : i195–i203.

Hill, W., 1981 Estimation of effective population size from data on linkage disequilibrium. Genet. Res. **38** : 209–216.

Hill, W. G., 1975 Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. Theor Popul Biol **8** : 117–126.

Hill, W. G. and A. Robertson, 1968 Linkage disequilibrium in finite populations. Theor Appl. Genet. **38** : 226–231.

Hill, W. G. and B. S. Weir, 1994 Maximum-likelihood estimation of gene location by linkage disequilibrium. Am J Hum Genet **54** : 705–714.

Hirschhorn, J. N. and M. J. Daly, 2005 Genome-wide association studies for common diseases and complex traits. Nat Rev Genet **6** : 95–108.

Ho, J. C., S. Kresovich and K. R. Lamkey, 2005 Extent and Distribution of Genetic Variation in U.S. Maize : Historically Important Lines and Their Open-Pollinated Dent and Flint Progenitors. Crop Sci. **45** : 1891–1900.

Hoggart, C. J., E. J. Parra, M. D. Shriver, C. Bonilla, R. A. Kittles *et al.*, 2003 Control of confounding of genetic associations in stratified populations. Am J Hum Genet **72** : 1492–1504.

Hoggart, C. J., M. D. Shriver, R. A. Kittles, D. G. Clayton and P. M. McKeigue, 2004 Design and analysis of admixture mapping studies. Am J Hum Genet **74** : 965–978.

Hoh, J. and J. Ott, 2003 Mathematical multi-locus approaches to localizing complex human trait genes. Nat Rev Genet **4** : 701–709.

Hoh, J., A. Wille and J. Ott, 2001 Trimming, weighting, and grouping SNPs in human case-control association studies. Genome Res **11** : 2115–2119.

Hoh, J., A. Wille, R. Zee, S. Cheng, R. Reynolds *et al.*, 2000 Selecting SNPs in two-stage analysis of disease association data : a model-free approach. Ann Hum Genet **64** : 413–417.

Horne, B. D. and N. J. Camp, 2004 Principal component analysis for selection of optimal SNP-sets that capture intragenic genetic variation. Genet Epidemiol **26** : 11–21.

Hugot, J. P., M. Chamaillard, H. Zouali, S. Lesage, J. P. Cezard *et al.*, 2001 Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature **411** : 599–603.

Ishiguro, M., S. Yosiyuki and K. Genshiro, 1997 Bootrsapping log likelihood and EIC, an extension of AIC. Annals of the Institute of Statistical Mathematics **49** : 411–434.

Jannink, J., M. C. Bink and R. C. Jansen, 2001 Using complex plant pedigrees to map valuable genes. Trends Plant Sci **6** : 337–342.

Jannink, J.-L. and B. Walsh, 2002 Association mapping in plant populations. In C. International, editor, *Quantitative Genetics, Genomics and Plant Breeding*. 59–68.

Jansen, R. C., 1993 Interval mapping of multiple quantitative trait loci. Genetics **135** : 205–211.

Jansen, R. C., J.-L. Jannink and W. D. Beavis, 2003 Mapping Quantitative Trait Loci in Plant Breeding Populations : Use of Parental Haplotype Sharing. Crop. Sci. **43** : 829–834.

J.C., B., 1981 *Pattern Recognition with Fuzzy Objective Function Algoritms*. Plenum Press, New York.

Jennings, H., 1917 The numerical results of diverse systems of breeding, with respect to two pairs of characters, linked or independent, with special relation to the effects of linkage. Genetics **2** : 97–154.

Ji, Y., D. M. Stelly, M. De Donato, M. M. Goodman and C. G. Williams, 1999 A candidate recombination modifier gene for Zea mays L. Genetics **151** : 821–830.

Johnson, G. C., L. Esposito, B. J. Barratt, A. N. Smith, J. Heward *et al.*, 2001 Haplotype tagging for the identification of common disease genes. Nat Genet **29** : 233–237.

Jorde, L. B., 1995 Linkage disequilibrium as a gene-mapping tool. Am J Hum Genet **56** : 11–14. Comment.

Jorde, L. B., 2000 Linkage disequilibrium and the search for complex disease genes. Genome Res **10** : 1435–1444.

Kao, C., Z. Zeng and R. Teasdale, 1999 Multiple Interval Mapping for Quantitative Trait Loci. Genetics **152** : 1203–1216.

Kao, C.-H. and Z.-B. Zeng, 2002 Modeling epistasis of quantitative trait loci using Cockerham's model. Genetics **160** : 1243–1261.

Kearsey, M. and A. Farquhar, 1998 QTL analysis in plants ; where are we now ? Heredity **80** : 137–142.

Kearsey, M. and H. S. Pooni, 1996 *The genetical analysis of quantitative traits*. Chapman and Hall, London.

Keightley, P. and S. Knott, 1999 Testing the correspondance between map positions of quantitative trait loci. Genet. Res. Camb. **74** : 323–328.

Khatkar, M., P. Thomson, I. Tammen and H. Raadsma, 2004 Quantitative trait loci mapping in dairy cattle : review and meta-analysis. Genet. Sel. Evol. **36** : 163–190.

Khavkin, E. and E. H. Coe, 1997 Mapped genomic locations for developmental functions and QTLs reflect concerned groups in maize (Zea mays L.). Theor. Appl. Genet. **95** : 343–352.

Khavkin, E. and E. H. Coe, 1998 The major quantitative trait loci for plant stature, development and yield are general manifestations of developmental gene clusters. Maize Newslett. **72** : 60–66.

Knowler, W. C., R. C. Williams, D. J. Pettitt and A. G. Steinberg, 1988 Gm3 ;5,13,14 and type 2 diabetes mellitus : an association in American Indians with genetic admixture. Am J Hum Genet **43** : 520–526.

Koivisto, M., T. Kivioja, H. Mannila, P. Rastas and E. Ukkonen, 2004 Hidden markov modelling techniques for haplotype analysis. In *ALT 2004, LNAI 3244*. Springer-Verlag, Berlin Heidelberg, 37–52.

Kruglyak, L., 1999 Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet **22** : 139–144.

Kuhner, M. K., J. Yamato and J. Felsenstein, 2000 Maximum likelihood estimation of recombination rates from population data. Genetics **156** : 1393–1401.

Lam, J. C., K. Roeder and B. Devlin, 2000 Haplotype fine mapping by evolutionary trees. Am J Hum Genet **66** : 659–673.

Lander, E. and L. Kruglyak, 1995 Genetic dissection of complex traits : guidelines for interpreting and reporting linkage results. Nat Genet **11** : 241–247.

Lander, E. S. and B. D., 1989 Mapping Mendelien factors underlying quantitative traits using RFLP linkage maps. Genetics **121** : 185–199.

Lander, E. S., P. Green, J. Abrahamson, A. Barlow and D. M.J., 1987 MapMaker : an integrative computer package for constructing genetic linkage maps of experimental and natural populations. Genomics **1** : 174–181.

Lewontin, R., 1964 The interaction of selection and linkage. I. General considerations ; heterotic models. Genetics **49** : 49–67.

Li, J. and T. Jiang, 2005 Haplotype-based linkage disequilibrium mapping via direct data mining. Bioinformatics .

Li, N. and M. Stephens, 2003 Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics **165** : 2213–2233.

Lin, Y. R., K. F. Schertz and A. H. Paterson, 1995 Comparative analysis of QTLS affecting plant height and maturity across the poaceae, in reference to an interspecific sorghum population. Genetics **141** : 391–411.

Liu, J. S., C. Sabatti, J. Teng, B. J. Keats and N. Risch, 2001 Bayesian analysis of haplotypes for linkage disequilibrium mapping. Genome Res **11** : 1716–1724.

Liu, K., M. Goodman, S. Muse, J. S. Smith, E. Buckler *et al.*, 2003 Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. Genetics **165** : 2117–2128.

Liu, S. C., S. P. Kowalski, T. H. Lan, K. A. Feldmann and A. H. Paterson, 1996 Genome-wide high-resolution mapping by recurrent intermating using Arabidopsis thaliana as a model. Genetics **142** : 247–258.

Lohmueller, K. E., C. L. Pearce, M. Pike, E. S. Lander and J. N. Hirschhorn, 2003 Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat. Genet. **33** : 177–182.

Long, J. C., 1991 The genetic structure of admixed populations. Genetics **127** : 417–428.

Lu, X., T. Niu and J. S. Liu, 2003 Haplotype information and linkage disequilibrium mapping for single nucleotide polymorphisms. Genome Res **13** : 2112–2117.

Lubberstedt, T., A. Melchinger, C. Schon, H. F. Utz and D. Klein, 1997 QTL mapping in testcrosses of European flint lines of maize : I. Comparison of different testers for forage yield traits. Crop. Sci. **37** : 921–931.

Mangin, B., B. Goffinet and A. Rebaï, 1994 Constructing confidence intervals for QTL location. Genetics **138** : 1301–1308.

Marchini, J., P. Donnelly and L. R. Cardon, 2005 Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet **37** : 413–417.

Mather, K., 1936 Types of linkage data and their value. Ann. Eugenics **7** : 251–264.

Mayr, E., 1982 *Histoire de la biologie. Diversité, évolution et hérédité.*. The Bellknap Press of Harvard University Press. Editions Fayard, 1989, pour la traduction française.

McKeigue, P. M., 1998 Mapping genes that underlie ethnic differences in disease risk : methods for detecting linkage in admixed populations, by conditioning on parental admixture. Am J Hum Genet **63** : 241–251.

McPeek, M. S. and A. Strahs, 1999 Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. Am J Hum Genet **65** : 858–875.

Mechin, V., O. Argillier, Y. Hebert, E. Guingo, L. Moreau *et al.*, 2001 Genetic analysis and QTL mapping of cell wall degistibility and lignification in silage maize. Crop. Sci. **41** : 690–697.

Meng, X. and D. B. Rubin, 1991 Using EM to obtain asymptotic variance-covariance matrices : the SEM algorithm. J. Am. Stat. Assoc. **86** : 899–909.

Meng, Z., D. V. Zaykin, C.-F. Xu, M. Wagner and M. G. Ehm, 2003 Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. Am J Hum Genet **73** : 115–130.

Menozzi, P., A. Piazza and L. Cavalli-Sforza, 1978 Synthetic maps of human gene frequencies in Europeans. Science **201** : 786–792. Historical Article.

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

Mihaljevic, R., H. F. Utz and A. E. Melchinger, 2004 Congruency of quantitative trait loci detected for agronomic traits in testcrosses of five populations of european maize. Crop. Sci. **44** : 114–124.

Mohammadi, S. A. and B. M. Prasanna, 2003 Analysis of genetic diversity in crop plants - Salient statistical tools and considerations. Crop Sci. **43** : 1235–1248.

Molitor, J., P. Marjoram and D. Thomas, 2003 Application of Bayesian spatial statistical methods to analysis of haplotypes effects and gene mapping. Genet Epidemiol **25** : 95–105.

Moreau, L., A. Charcosset and A. Gallais, 2004 Use of trail clustering to study QTL x environment effects for grain yield and related traits in maize. Theor. Appl. Genet. **110** : 92–105.

Morris, A. P., J. C. Whittaker and D. J. Balding, 2000 Bayesian fine-scale mapping of disease loci, by hidden Markov models. Am J Hum Genet **67** : 155–169.

Morris, A. P., J. C. Whittaker and D. J. Balding, 2002 Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. Am J Hum Genet **70** : 686–707.

Morris, A. P., J. C. Whittaker and D. J. Balding, 2004 Little loss of information due to unknown phase for fine-scale linkage-disequilibrium mapping with single-nucleotide-polymorphism genotype data. Am J Hum Genet **74** : 945–953.

Morris, A. P., J. C. Whittaker, C.-F. Xu, L. K. Hosking and D. J. Balding, 2003 Multipoint linkage-disequilibrium mapping narrows location interval and identifies mutation heterogeneity. Proc Natl Acad Sci U S A **100** : 13442–13446.

Morton, N. E., 1955 Sequential tests for the detection of linkage. Am J Hum Genet **7** : 277–318.

Morton, N. E., 1956 The detection and estimation of linkage between the genes for elliptocytosis and the Rh blood type. Am J Hum Genet **8** : 80–96.

Munafo, M. R. and J. Flint, 2004 Meta-analysis of genetic association studies. Trends Genet **20** : 439–444.

Nei, M. and W. H. Li, 1973 Linkage disequilibrium in subdivided populations. Genetics **75** : 213–219.

Nei, M. and W. H. Li, 1980 Non-random association between electromorphs and inversion chromosomes in finite populations. Genet Res **35** : 65–83.

Nielsen, D. M., M. G. Ehm, D. V. Zaykin and B. S. Weir, 2004 Effect of two- and three-locus linkage disequilibrium on the power to detect marker/phenotype associations. Genetics **168** : 1029–1040.

Niu, T., 2004 Algorithms for inferring haplotypes. Genet Epidemiol **27** : 334–347.

Nordborg, M., 2000 Linkage disequilibrium, gene trees and selfing : an ancestral recombination graph with partial self-fertilization. Genetics **154** : 923–929.

Nordborg, M. and S. Tavare, 2002 Linkage disequilibrium : what history has to tell us. Trends Genet **18** : 83–90.

Onkamo, P., V. Ollikainen, P. Sevon, H. T. T. Toivonen, H. Mannila *et al.*, 2002 Association analysis for quantitative traits by data mining : QHPM. Ann Hum Genet **66** : 419–429.

Paran, I. and D. Zamir, 2003 Quantitative traits in plants : beyond the QTL. Trends Genet **19** : 303–306.

Paterson, A. H., E. S. Lander, J. D. Hewitt, S. Peterson, S. E. Lincoln *et al.*, 1988 Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. Nature **335** : 521–529.

Paterson, A. H., Y. R. Lin, K. F. Schertz, J. F. Doebley, S. R. M. Pinson *et al.*, 1995 Convergent domestication of cereal crops by independent mutations at corresponding genetic loci. Science **269** : 1714–1718.

Patil, N., A. J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi *et al.*, 2001 Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science **294** : 1719–1723.

Pfaff, C. L., E. J. Parra, C. Bonilla, K. Hiester, P. M. McKeigue *et al.*, 2001 Population structure in admixed populations : effect of admixture dynamics on the pattern of linkage disequilibrium. Am J Hum Genet **68** : 198–207.

Pichot, A., 1999 *Histoire de la notion de gène*. Flammarion, Paris.

Posada, D., T. J. Maxwell and A. R. Templeton, 2005 TreeScan : a bioinformatic application to search for genotype/phenotype associations using haplotype trees. Bioinformatics **21** : 2130–2132.

Poupard, B., L. Moreau and A. Charcosset, 2001 Analyse de l'épistatsie entre QTL pour 3 caractères agronomiques chez le maïs. Technical report, INRA.

Press, W. H., S. A. Teukolsky, W. T. Vetterling and B. P. Flannery., 1992 *Numerical Recipe in C : The Art of Scientific Computing*. Cambridge University Press, New York.

Pritchard, J. K. and M. Przeworski, 2001 Linkage disequilibrium in humans : models and data. Am J Hum Genet **69** : 1–14.

Pritchard, J. K., M. Stephens and P. Donnelly, 2000a Inference of population structure using multilocus genotype data. Genetics **155** : 945–959.

Pritchard, J. K., M. Stephens, N. A. Rosenberg and P. Donnelly, 2000b Association mapping in structured populations. Am J Hum Genet **67** : 170–181.

Przeworski, M., 2003 Estimating the time since the fixation of a beneficial allele. Genetics **164** : 1667–1676.

Rabiner, L. R., 1989 A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE **77** : 257–286.

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

Rannala, B. and J. P. Reeve, 2001 High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence. Am J Hum Genet **69** : 159–178.

Rebai, A., P. Blanchard, D. Perret and P. Vincourt, 1997 Mapping quantitative trait loci controlling silking date in a diallel cross among four lines in maize. Theor. Appl. Genet. **95** : 451–459.

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.

Remington, D. L., J. M. Thornsberry, Y. Matsuoka, L. M. Wilson, S. R. Whitt *et al.*, 2001 Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc Natl Acad Sci U S A **98** : 11479–11484.

Ribaut, J.-M., D. Hoisington, J. A. Deutsch, C. Jiang and D. Gonzalez-de Leon, 1996 Indentification of quantitative trait loci under drought conditions in tropical maize. I. Flowering parameters and the anthesis-silking interval. Theor. Appl. Genet. **92** : 905–914.

Sabatti, C., S. Service and N. Freimer, 2003 False discovery rate in linkage and association genome screens for complex disorders. Genetics **164** : 829–833.

Saitou, N. and M. Nei, 1987 The neighbor-joining method : a new method for reconstructing phylogenetic trees. Mol Biol Evol **4** : 406–425.

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

Salvi, S., R. Tuberosa, E. Chiapparino, M. Maccaferri, S. Veillet *et al.*, 2002 Toward positional cloning of Vgt1, a QTL controlling the transition from the vegetative to the reproductive phase in maize. Plant Mol Biol **48** : 601–13.

Sasaki, T., T. Matsumoto, K. Yamamoto, K. Sakata, T. Baba *et al.*, 2002 The genome sequence and structure of rice chromosome 1. Nature **420** : 312–316.

Satten, G. A., W. D. Flanders and Q. Yang, 2001 Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. Am J Hum Genet **68** : 466–477.

Schaid, D. J., C. M. Rowland, D. E. Tines, R. M. Jacobson and G. A. Poland, 2002 Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet **70** : 425–434.

Schiex, 1997 Carthagene : constructing and joining maximum likelihood maps. ISMB **5** : 258–267.

Schwartz, R., 2004 Algorithms for Association Study Design Using a Generalized Model of Haplotype Conservation. In *Computational Systems Bioinformatic Conference*. IEEE.

Schwartz, R., A. G. Clark and S. Istrail, 2002 Methods for Inferring Block-Wise Ancestral History from Haploid Sequences. The Haplotype Coloring Problem. In *WABI 2002*. Springer-Verlag, Berlin Heidelberg, 44–59.

Schwarz, 1978 Estimating the Dimension of a Model. Annals of Statistics **6** : 461–464.

Sharbel, T. F., B. Haubold and T. Mitchell-Olds, 2000 Genetic isolation by distance in Arabidopsis thaliana : biogeography and postglacial colonization of Europe. Mol Ecol **9** : 2109–2118.

Sillanpaa, M. J. and K. Auranen, 2004 Replication in genetic studies of complex traits. Ann Hum Genet **68** : 646–657.

Smith, M. W. and S. J. O'Brien, 2005 Mapping by admixture linkage disequilibrium : advances, limitations and guidelines. Nat Rev Genet **6** : 623–632.

Spiegelhalter, D. J., N. G. Best and B. P. Carlin, 1998 Bayesian deviance, the effective number of parameters, and the comparison of arbitrary complex models. Statistic computing **28** : 286–289.

Stadler, L. J., 1925 The Variability of Crossing Over in Maize. Genetics **11** : 1–37.

Stam, P., 1993 Construction of integrated genetic linkage maps by means of a new coputer package : JoinMap. Plant J. **3** : 739–744.

Stephens, J. C., D. Briscoe and S. J. O'Brien, 1994 Mapping by admixture linkage disequilibrium in human populations : limits and guidelines. Am J Hum Genet **55** : 809–824.

Stumpf, M. P. H. and D. B. Goldstein, 2003 Demography, recombination hotspot intensity, and the block structure of linkage disequilibrium. Curr Biol **13** : 1–8.

Sugiura, N., 1978 Further Analysis of the Data by Akaike's Information Criterion and the Finite Corrections. Communications in Statistics, Theory and Methods **A** : 13–26.

Takeuchi, K., 1976 Distribution of informational statistics and a criterion model fitting. Math. Sci. **153** : 12–18.

Templeton, A. R., 1995 A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping or DNA sequencing. V. Analysis of case/control sampling designs : Alzheimer's disease and the apoprotein E locus. Genetics **140** : 403–409.

Templeton, A. R., E. Boerwinkle and C. F. Sing, 1987 A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in Drosophila. Genetics **117** : 343–351.

Templeton, A. R., K. A. Crandall and C. F. Sing, 1992 A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. Genetics **132** : 619–633.

Templeton, A. R., T. Maxwell, D. Posada, J. H. Stengard, E. Boerwinkle *et al.*, 2005 Tree scanning : a method for using haplotype trees in phenotype/genotype association studies. Genetics **169** : 441–453.

Templeton, A. R. and C. F. Sing, 1993 A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analyses with cladogram uncertainty and recombination. Genetics **134** : 659–669.

Templeton, A. R., C. F. Sing, A. Kessling and S. Humphries, 1988 A cladistic analysis of phenotype associations with haplotypes inferred from restriction endonuclease mapping. II. The analysis of natural populations. Genetics **120** : 1145–1154.

Tenaillon, M. I., M. C. Sawkins, A. D. Long, R. L. Gaut, J. F. Doebley *et al.*, 2001 Patterns of DNA sequence polymorphism along chromosome 1 of maize (Zea mays ssp. mays L.). Proc Natl Acad Sci U S A **98** : 9161–9166.

Tenaillon, M. I., J. U'Ren, O. Tenaillon and B. S. Gaut, 2004 Selection versus demography : a multilocus investigation of the domestication process in maize. Mol Biol Evol **21** : 1214–1225.

Terwilliger, J. D. and K. M. Weiss, 1998 Linkage disequilibrium mapping of complex disease : fantasy or reality ? Curr Opin Biotechnol **9** : 578–594.

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.

Tipping, M. E. and C. M. Bishop, 1998 Principal Component Analysers. Technical report, Neural Computing Research Group.

Titterington, D., A. Smith and U. Markov, 1985 *Statistical Analysis of Finite Mixture Distributions*. John Wiley and Sons, New York.

Toivonen, H., P. Onkamo, P. Hintsanen, E. Terzi and P. Sevon, 2004 *Data mining for gene mapping*. IEEE Press.

Toivonen, H. T., P. Onkamo, K. Vasko, V. Ollikainen, P. Sevon *et al.*, 2000 Data mining applied to linkage disequilibrium mapping. Am J Hum Genet **67** : 133–145.

Ukkonen, E., 2002 Finding Founder Sequences from a Set of Recombinants. In *WABI 2002*. Springer-Verlag, Berlin Heidelberg, 277–286.

Van Zandt, P. and S. Mopper, 1998 A meta-analysis of adaptive deme formation in phytophagous insect populations. Am. Nat. **152** : 595–604.

Visscher, P. M. and M. E. Goddard, 2004 Prediction of the confidence interval of quantitative trait loci. Behavior Genetics **34** : 477–482.

Vitalis, R. and D. Couvet, 2001 Estimation of effective population size and migration rate from one- and two-locus identity measures. Genetics **157** : 911–925.

Viterbi, A. J., 1967 Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. IEEE Trans. Informat. Theory **IT-13** : 260–269.

Vladutu, C., J. McLaughlin and R. L. Phillips, 1999 Fine mapping and characterization of linked quantitative trait loci involved in the transition of the maize apical meristem from vegetative to generative structures. Genetics **153** : 993–1007.

Vollestad, L. A., K. Hindar and A. P. Moller, 1999 A meta-analysis of fluctuating asymmetry in relation to heterozygosity. Heredity **83** : 206–218.

Wall, J. D. and J. K. Pritchard, 2003 Assessing the performance of the haplotype block model of linkage disequilibrium. Am J Hum Genet **73** : 502–515. Evaluation Studies.

Wang, J., 2005 Estimation of effective population sizes from data on genetic markers. Philos Trans R Soc Lond B Biol Sci **360** : 1395–1409.

Wang, R. L., A. Stec, J. Hey, L. Lukens and J. Doebley, 1999 The limits of selection during maize domestication. Nature **398** : 236–239.

Wang, R. L., A. Stec, J. Hey, L. Lukens and J. Doebley, 2001 Correction : The limits of selection during maize domestication (col 398, pg 236, 1999). Nature **410** : 718–718.

Wang, X., I. Le Roy, E. Nicodeme, R. Li, R. Wagner *et al.*, 2003 Using advanced intercross lines for high-resolution mapping of HDL cholesterol quantitative trait loci. Genome Res **13** : 1654–1664.

Ward, J. H., 1963 Hierarchical Grouping to Optimize an Objective Function. J. Am. Stat. Assoc. **58** : 236–244.

Weir, B., 1996 *Genetic Data Analysis II*. Sinauer Associates, Sunderland, MA.

Weller, J. I., J. Z. Song, D. W. Heyen, H. A. Lewin and M. Ron, 1998 A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. Genetics **150** : 1699–1706.

Wiberg, T., 1976 Computation of principal components when data are missing. In *Proc. Second Symp. Computational Statistics*. Berlin, 229–236.

Williams, C. G., M. M. Goodman and C. W. Stuber, 1995 Comparative recombination distances among Zea mays L. inbreds, wide crosses and interspecific hybrids. Genetics **141** : 1573–1581.

Windham, M. and A. Cutler, 1992 Information Ratios for Validating Mixture Analyses. J. Am. Stat. Ass. **87** : 1188–1192.

Winkler, C. R., N. M. Jensen, M. Cooper, D. W. Podlich and O. S. Smith, 2003 On the determination of recombination rates in intermated recombinant inbred populations. Genetics **164** : 741–745.

Wolfe, J., 1971 A Monte Carlo study of sampling distribution of the likelihood ration for mixtures of multinormal distributions. Technical Bulletin STB **72-2**.

Xu, S., 2003 Theoretical basis of the Beavis effect. Genetics **165** : 2259–68.

Yap, I., D. Schneider, J. Kleinberg, D. Matthews, S. Cartinhour *et al.*, 2003 A graph-theoretic approach to comparing and integrating genetic, physical and sequence-based maps. Genetics **165** : 2235–2247.

Yule, G., 1900 On the association of attributes in statistics. Philos. Trans. R. Soc. London A. **194** : 257–319.

Zeng, Z. B., 1994 Precision mapping of quantitative trait loci. Genetics **136** : 1457–1468.

Zeng, Z.-B., T. Wang and W. Zou, 2005 Modeling quantitative trait Loci and interpretation of models. Genetics **169** : 1711–1725.

Zhang, K., P. Calabrese, M. Nordborg and F. Sun, 2002 Haplotype block structure and its applications to association studies : power and study designs. Am J Hum Genet **71** : 1386–1394.

Zhang, K., Z. Qin, T. Chen, J. S. Liu, M. S. Waterman *et al.*, 2005 HapBlock : haplotype block partitioning and tag SNP selection software using a set of dynamic programming algorithms. Bioinformatics **21** : 131–134.

Zhang, K., Z. S. Qin, J. S. Liu, T. Chen, M. S. Waterman *et al.*, 2004a Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies. Genome Res **14** : 908–916.

Zhang, W., A. Collins and N. E. Morton, 2004b Does haplotype diversity predict power for association mapping of disease susceptibility ? Hum Genet **115** : 157–164.

Zollner, S. and J. K. Pritchard, 2005 Coalescent-based association mapping and fine mapping of complex trait loci. Genetics **169** : 1071–1092.