Utilisation d'une population multi-parentale et hautement recombinante de blé tendre pour l'étude de l'architecture génétique de la précocité de floraison

Stéphanie Thépot

UNIVERSITÉ PARIS SUD

ÉCOLE DOCTORALE : SCIENCES DU VEGETAL

T H È S E

pour obtenir le titre de

Docteur

de l'Université de Paris Sud

Spécialité : Biologie

Présentée et soutenue publiquement par

Stéphanie Thépot

Utilisation d'une population multi-parentale et hautement recombinante de blé tendre pour l'étude de l'architecture génétique de la précocité de floraison

Thèse préparée à l'UMR 320 Génétique Végétale, Ferme du Moulon, F-91190 Gif-sur-Yvette

Soutenue le 13 mars 2014 devant le jury composé de :

Directeur (Invité) :	Jérôme Enjalbert
Co-encadrante :	Isabelle Goldringer
Rapporteurs :	Jacques David
	Fabrice Roux
Présidente :	Jacqui Shykoff
Examinateurs :	Pierre Dubreuil
	David Gouache

Utilisation d'une population multi-parentale et hautement recombinante de blé tendre pour l'étude de l'architecture génétique de la précocité de floraison

Résumé :
Aujourd'hui, alors que le nombre de marqueurs génétiques disponibles augmente rapidement, de nouvelles populations doivent être créées pour exploiter au mieux cette quantité d'informations dans le but de mieux comprendre l'architecture génétique de caractères complexes. Les populations de type MAGIC ont été créées pour rassembler les avantages des populations bi-parentales et des panels d'associations, la bonne puissance de détection et une localisation précise.

L'objectif de cette thèse était d'étudier l'intérêt de la population MAGIC INRA pour l'analyse de l'architecture génétique de la précocité de floraison. Cette population a été créée à partir de 60 parents brassés durant 12 générations de panmixie grâce à l'introduction d'un gène de stérilité mâle (*ms1b*). Cette étude a été réalisée sur 56 parents toujours disponibles en banque de graines et 380 lignées dérivées de la population après les 12 générations de recombinaison. Cette population a été génotypée avec la puce 9K iSelect, représentant environ 5 000 SNPs localisés sur tout le génome, additionnée de 14 marqueurs localisés dans des gènes candidats. Ce jeu de données moléculaires a été complété par des données fines de phénotypage de la précocité de floraison.

Suite aux 12 générations de panmixie, le DL de cette population a été très réduit, à longue comme à moyenne distance (<10cM). Ce faible DL nous a amené à développer un algorithme basé uniquement sur le DL qui ordonne les marqueurs de manière à avoir un DL décroissant monotone avec la distance. L'algorithme ordonne globalement de la même manière que la carte génétique les marqueurs à longue distance mais à courte distance le DL est moins lié à la distance génétique. La différence réside sur l'équilibre entre les effets de la recombinaison et de la dérive génétique sur le DL.

L'intérêt de la population MAGIC INRA pour détecter des QTLs a ensuite été étudié avec deux approches : une approche évolutionniste et une approche de génétique d'association. La première approche détecte les loci soumis à sélection par comparaison des fréquences alléliques de la population initiale (G0) et de la population évoluée (G12) grâce à une nouvelle méthode. La population initiale est composée des parents pondérés par une contribution estimée avec une nouvelle méthode bayésienne. 26 régions génomiques soumises à sélection ont été détectées. Une analyse de génétique d'association avec les marqueurs détectés sous sélection a montré que respectivement cinq et trois zones étaient associées à la précocité avec un semis d'automne et au caractère printemps/hiver. Une analyse phénotypique a effectivement mis en évidence la précocification de la date de floraison et une augmentation de la proportion de plantes de type printemps. Une analyse de génétique d'association a ensuite été réalisée sur les lignées SSD sur 12 caractères × environnements *i.e.* la date d'épiaison et le temps de remplissage du grain mesurés dans six environnements. Les tests d'association ont aussi été réalisés avec des

variables synthétisant l'information présente dans plusieurs traits phénotypiques soit avec une ACP, soit avec un modèle écophysiologique. Au total, toutes ces analyses ont détecté six QTLs dont trois correspondants à des gènes majeurs. Parmi ces six QTLs, deux sont spécifiques des caractères mesurés avec un semis d'automne et deux avec ceux mesurés avec un semis de printemps.

Mots clés : Multi-parentale, MAGIC, gestion dynamique, détection de QTLs, précocité de floraison, évolution expérimentale, *Triticum aestivum L.*

Studying flowering time genetics in wheat through the use of a multiparent advanced generation inter-cross population

Abstract :
Nowadays, with the dramatically increase of available molecular markers, there is a deep need for new populations allowing to exploit all of this information to better understand the genetic architecture of complex traits. MAGIC populations as they are built to bring together bi-parental populations and association panel advantages, provide such powerful detection and fine mapping capacities.

The aim of these PhD was to study the MAGIC INRA population usefulness for the study of genetic architecture of earliness. This population is derived from 12 cycles of random crosses between 60 founders, turning wheat from selfing to outcrossing thanks to the use of a nuclear male sterility gene (*ms1b*, Probus donor). This population is composed of 56 parents still available and 380 SSD lines. Parents and SSD lines were genotyped using the 9K iSelect SNPs array, providing around 5 000 SNPs on the whole genome, as well as 14 addition markers located in candidate genes. They were also finely phenotyped for earliness traits.

With the 12 panmictic generations, the population LD decreased strongly, especially at long and medium distance (<10cM). This allowed us to develop an algorithm mapping markers on the sole pairwise LD information, ordering markers in a way to have the LD decreasing along the distance. When considering long distances, overall the results were consistent with the order found on genetic maps while at short distance LD was poorly linked to genetic distance. These differences between long and short distances were linked to the balance between recombination and drift effects on LD.

The usefulness of the MAGIC INRA population for QTL detection was analyzed with two approaches : an evolutionary approach and an association genetics approach. The first one detects loci under selection by identifying high shift in allelic frequency with a new method. The initial population was composed of founders weighted by a contribution estimated with a new Bayesian method. 26 genomic areas under selection were detected. An association genetics analysis with the markers detected as under selection showed respectively five and three genomic regions associated with earliness and growth habit. Actually the G12 population was found phenotypically earlier than the G0 and with more spring individuals.

A broader association genetics analysis was performed on G12 population, studying 12 traits × environments *i.e.* heading date and grain filling time, both observed in six environmental conditions. Two additional integrated traits from either PCA or ecophysiological model were also analyzed. In all, these different analyses detected six QTLs, three of them corresponding to candidate genes. Among these six QTLs, two were specific to autumn sowing and two specific to spring sowing.

Keywords : multi-parental, MAGIC, dynamic management, QTL detection, flowering time, experimental evolution, *Triticum aestivum L.*

Remerciements

Je tiens tout d'abord à remercier Jérôme Enjalbert et Isabelle Goldringer qui m'ont fait confiance dès le début en me proposant cette thèse "sur-mesure" en fonction de mon projet professionnel. Ces trois années ont été très formatrices.

La bonne ambiance du Moulon m'a permis de travailler avec un grand nombre de personnes chacun spécialiste dans son domaine.

Merci à Sophie Pin, Nath Galic et Didier Tropée pour les expérimentations au champ. Je ne compterai pas les heures passées à semer, noter et récolter les 10 000 lignes de pépinières. Merci à Valérie, Carine et Xavier pour leur aide au labo, leurs conseils et leur gentillesse pour les extractions d'ADN.

Merci à Christine Dillmann toujours prêtes à m'aider pour les statistiques, R, et l'algorithme d'ordonnancement. ...

Merci à Mathieu Falques pour ses idées et conseils pour l'algorithme d'ordonnancement.

Merci à Laurence Moreau pour son expertise sur la génétique d'association.

Merci à Renaud Rincent pour m'avoir codé la fonction de calcul de puissance à la toute fin quand j'étais en plein rush.

Merci à Yannick De Oliviera pour le script de fusion de fichier Python et la tentative de créer un Thalia blé.

Merci à Pierre Rivière pour ces bons plans R et Latex (mais je ne suis pas sûre que Jérôme te remerciera sur ce dernier point ;-))

Je voudrais aussi remercier les temporaires comme Vincent en CDD et Caroline, Joséphine, Julien, Katia et Romain, les stagiaires que j'ai encadrés ; il ne faut pas oublier les MOMOs aussi. Ils m'ont tous été d'une grande aide pour les notations et les récoltes aux champs. Grâce à eux, chacun avec leur spécificité, j'ai beaucoup appris sur l'encadrement de personnes.

Je voudrais remercier Jacques David et Fabrice Roux d'avoir accepter d'être rapporteur de ma thèse, David Gouache, Pierre Dubreuil et Jacqui Shykoff d'avoir accepter de faire partie du jury.

Je voudrais remercier aussi toutes les personnes qui ont participé à mon comité de thèse pour leurs conseils : Vincent Allard, David Gouache, Stéphane Nicolas, Ian Mackay, Laurence Moreau, Etienne Paux.

Merci à Gwendal Restoux, Fred Hospital, Ian Mackay et David Gouache qui m'ont aidée à écrire mon premier papier scientifique et ce n'est pas une mince affaire ! !

Je voudrais remercier Vincent Allard et Matthieu Bogard pour leur implication dans l'estimation des paramètres écophysiologiques.

Merci évidemment à tous les amis que j'ai rencontrés au Moulon sans qui ces trois années n'auraient pas été pareils, rythmés par les goûters, les soirées jeux, les coinches, les bonnes bouffes, les séjours au ski, à Londres, à La Palmyre, à Toulouse, ... les heures à se défouler à la salle de sport. Ils ont dû se reconnaître mais je pensais evidemment à Adrien, Aude, Beatriz, Bub, Charlotte, Chris, Fabio, Héloïse, Margot-Alison, Nico, Paulina, Pierre, Pierrot, Sandra, Sara, Sosno, Véro, Yannick

Merci à toute ma famille, mes parents, mon frère et Thierry qui m'ont toujours soutenue dans les bons jours et comme dans les mauvais.

Table des matières

Liste des abréviations

ACP	Analyse en Composantes Principales
ANOVA	ANalysis Of VAriance
DAPC	Discriminant Analysis of Principal Components
DL	Déséquilibre de Liaison
FAO	Food and Agriculture Organization
GD	Gestion Dynamique
INRA	Institut National de Recherche Agronomique
MAGIC	Multi-parent Advanced Generation Inter-Cross
NAM	Nested Association Mapping
NIAB	National Institute of Agricultural Botany
QTL	Quantitative Trait Locus
RMSE	Root Mean Squared Error
SNP	Single Nucleotide Polymorphism
SSD	Single Seed Descent

Chapitre 1

Introduction

1.1 Contexte actuel de l'agriculture : la biodiversité, un réservoir d'espoir pour l'avenir

"Comment nourrir neuf milliards de personnes d'ici 2050 ?" est la question qui fait débat depuis déjà plus d'une vingtaine d'années [Evans, 1998; Waterlow et al., 1998; Parker, 2011; Shiferaw et al., 2013]. En effet, des prévisions estiment que la demande en céréales devrait doubler d'ici 2050 [Godfray et al., 2010]. Le principal problème est que l'accroissement de la population mondiale est accompagné par une hausse de la consommation de viande, de produits laitiers et de poisson [Godfray et al., 2010]. Une augmentation de la production animale a un fort impact sur les besoins de production végétale. En effet il faut 5kg de céréales pour produire 1kg de viande de porc et 13kg pour 1kg de viande de boeuf [Smil, 2002] ; ainsi en 2006, un tiers de la production globale de céréales était utilisé pour l'alimentation animale [Michéli et al., 2006]. En plus de cet accroissement de la demande alimentaire, l'urbanisation, la compétition avec les cultures non alimentaires (notamment les agro-carburants) et l'altération des terres agricoles (salinisation, érosion des sols) réduisent fortement les surfaces agricoles utilisées pour la production alimentaire [Beddington, 2010]. Une des solutions serait de limiter la hausse de la demande par la réduction des pertes post-récoltes qui concernent entre 30 et 40% de la production, grâce à l'amélioration des infrastructures (conditions de stockage, ...) et des transports mais aussi par le changement des habitudes alimentaires de chacun [Godfray et al., 2010]. Une ration de 3 000Kcal par jour et par personne étant nécessaire pour être en bonne santé, la production agricole actuelle est suffisante pour nourrir la population mondiale [FAO, 2013]. Mais si la production actuelle est suffisante, il reste cependant un problème de répartition de la production liée aux différences de rendement, même entre régions avec les mêmes conditions climatiques et surtout d'accès à la nourriture et d'accès à la terre pour les plus pauvres. Durant les 50 dernières années alors qu'en Asie, les rendements ont été multipliés par deux et par 1,6 en Amérique latine, en Afrique ils ont diminué depuis les années 70 et atteignent le même niveau que durant les années 60 [FAO, 2013]. Ce rendement est très dépendant notamment de la disponibilité en eau, de semences adaptées, d'intrants, ... [Godfray et al., 2010]. Le défit est au minimum, de maintenir le niveau actuel des rendements avec des environnements de plus en plus stressants, sachant qu'aujourd'hui les aléas climatiques ont tendance à faire stagner les rendements des principales céréales comme le blé, l'orge ou le riz depuis les années 90 [Ladha et al., 2003; Brisson et al., 2010; Finger, 2010].

Durant la révolution verte, le rendement était la seule variable à prendre en

compte pour établir sa stratégie de culture [Conway, 1998], l'agriculteur contrôlant très fortement l'environnement à l'aide d'intrants pour l'optimiser pour une culture donnée. Aujourd'hui, la rentabilité maximale n'est plus nécessairement associée à des rendements maximums, elle peut être atteinte en diminuant les dépenses dues aux intrants, ce qui contre-balance des rendements plus faibles [Chaumet et al., 2009]. De plus aujourd'hui la composante environnementale est devenue importante, notamment avec le plan Ecophyto [Butault et al., 2010] qui vise à réduire progressivement l'utilisation des produits phytosanitaires. Cette prise en compte de la composante environnementale fait suite à la prise de conscience de nombreuses conséquences néfastes de l'agriculture intensive. Elle a entraîné la pollution de l'eau, un épuisement des sols mais aussi une diminution massive de la biodiversité via la diminution voire la disparition d'espèces [Potts et al., 2010; Cardinale et al., 2012]. En plus de l'urbanisation qui détruit les éco-systèmes entiers, l'utilisation des pesticides, le labour, la diminution des rotations, la simplification des paysages (disparition de haies, champs plus grands, ...) et l'augmentation des rendements, ... ont contribué à diminuer la biodiversité [Reganold, 1988; Gabriel et al., 2013]. En effet une étude a montré que l'augmentation des rendements entraînait une diminution de l'abondance de certaines espèces d'insecte mobile comme les abeilles ou les papillons indépendamment de la pratique culturale [Gabriel et al., 2013]. La modernisation de l'agriculture s'est également traduit par une érosion génétique de la diversité cultivée due notamment au remplacement des variétés locales possédant une large base génétique par des variétés modernes à base génétique plus étroite [Roussel et al., 2004, 2005; Bonneuil and Thomas, 2009]. Pourtant certaines épidémies historiques comme celle due à l'helminthosporiose qui a ravagé les cultures de maïs aux États-Unis en 1970 ont montré que cette diversité avait mis en évidence le risque d'une dépendance à un petit nombre de variétés à haut rendement.

Les ressources génétiques végétales sont nécessaires pour l'amélioration des plantes [Frankel, 1995] que ce soit dans un but de productivité, de qualité ou d'extension de la zone de culture, mais elles sont surtout garantes du maintien du potentiel adaptatif [Gunderson, 2000]. Pour les plantes sauvages comme pour les plantes cultivées, l'environnement dans lequel elles se développent évolue constamment, notamment avec le réchauffement climatique. Pour ne pas disparaître, les populations doivent donc s'adapter. La diversité génétique d'une espèce constitue le réservoir de gènes sur lequel repose l'adaptation locale et le potentiel évolutif des plantes qui leur permettront de répondre aux éventuels changements environnementaux ou à l'évolution des besoins humains [Barrett and Schluter, 2008]. En effet, notamment pour l'agriculture biologique, il est important d'utiliser des plantes adaptées à leur milieu puisque les intrants sont limités voire nuls. L'utilisation des ressources génétiques végétales dans les programmes de sélection permettrait de créer les meilleures variétés pour une condition de culture (pratique culturale, environnement) et un débouché précis.

1.2 Conservation des ressources génétiques végétales cultivées

Suite à la conférence de Stockholm organisée par la FAO en 1972, deux approches de conservation de la biodiversité ont été définies : la conservation statique qui est décrite brièvement et la conservation dynamique qui sera présentée plus précisément car elle est à l'origine de mon sujet de thèse [Frankel, 1995; Maxted et al., 1997; Gepts, 2006].

1.2.1 Conservation statique

La conservation statique consiste à stocker à long terme, des semences ou des organes de multiplication ou plants en chambre froide, au congélateur ou en chambre de culture avec un degré d'hygrométrie approprié [Plucknett et al., 1987]. Ces banques sont constituées d'échantillons fixés d'accessions accompagnés de fiches de renseignements sur leur identité (nom, généalogie, ...), et leur condition de prélèvement (lieu, date, ...). Ce mode de gestion des ressources génétiques a été le premier utilisé, et est qualifié de statique car il vise à protéger une ressource de l'extinction en la conservant de façon à minimiser les effets de la sélection naturelle. Cette procédure de sauvegarde des ressources génétiques est très adaptée à la conservation des lignées fixées, mais présente cependant un certain nombres de limites en ce qui concernent les populations hétérogènes : i) la petite taille des échantillons ne leur permet pas d'être représentatifs d'une population hétérogène telle qu'une variété population ou une variété de pays (ou landrace) [Esquinas-Alcazar, 2005] ; ii) comme le pouvoir germinatif des graines diminue au cours du temps, il est nécessaire de les régénérer très régulièrement. La multiplication en faible effectif et par autofécondation, dans le cas des espèces autogames, entraîne à chaque fois une perte de diversité intra-échantillon par dérive génétique [Bretting and Duvick, 1997; Parzies et al., 2000; Soengas et al., 2009] ; iii) la régénération des semences est effectuée dans un environnement identique pour toutes les accessions. Le peu d'évolution possible se fera donc dans la même direction sans prendre en compte l'environnement d'origine [Esquinas-Alcazar, 2005]. Cette conservation est très utilisée puisqu'elle permet de garder un grand nombre d'échantillons de variétés actuelles ou anciennes facilement disponibles pour la recherche ou la production (après quelques générations de multiplication). En 2010, plus de 1 750 banques de graines sont répertoriées conservant au total plus de 7,4 millions d'échantillons parmi lesquelles 25 à 30% sont les accessions originales, les autres étant des copies de sauvegarde que les centres s'échangent entre eux. En plus de ces copies, un centre de sauvegarde contenant 412 000 accessions provenant du monde entier a été construit à Svalbard, proche du pôle Nord [FAO, 2010].

1.2.2 Conservation dynamique

Mise en avant par Brush [2000], la conservation dynamique, aussi appelée gestion dynamique (GD), vise à maintenir ou recréer des conditions où les forces évolutives maintiennent le potentiel adaptatif d'une espèce [Bretting and Duvick, 1997; Goldringer et al., 2001]. Le principe est de ressemer des populations d'une année sur l'autre pour leur permettre d'évoluer et de s'adapter de façon continue à un ou plusieurs environnements. Ce type de conservation permet à la fois de sélectionner différentes combinaisons de gènes dans chaque population x environnement, adaptées à un milieu donné, mais également de capter de nouvelles variations par mutation/recombinaison. Ces processus conduisent à une différenciation des populations locales tout en maintenant une variabilité génétique intéressante sur l'ensemble des populations d'un réseau. Pour limiter l'érosion génétique et l'accumulation d'un fardeau génétique trop important dans chaque population, résultat de la sélection naturelle et de la taille finie des populations, des flux de gènes entre populations sont nécessaires (migration). Un tel réseau de populations interconnectées correspond à une métapopulation, concept central de la biologie de la conservation qui permet de réfléchir sur le maintien de la diversité et de l'adaptation d'une espèce [Henry et al., 1991; Olivieri et al., 1990]. La structure en métapopulation permet aux différentes forces évolutives (décrites dans l'encadré 1.2.2) de maintenir le potentiel évolutif global. On peut identifier deux types de gestion dynamique : la gestion dynamique en conditions expérimentales (ex : station d'expérimentation) [Allard, 1988; Henry et al., 1991; Lavigne et al., 2001; Enjalbert et al., 2011] et la gestion dynamique à la ferme pour les espèces cultivées [Maxted et al., 2002; Enjalbert et al., 2011; Thomas et al., 2011].

La gestion dynamique à la ferme est réalisée par les agriculteurs eux-mêmes sur leurs propres parcelles [Maxted et al., 2002]. En plus des forces évolutives citées précédemment, les populations peuvent subir une sélection plus ou moins consciente de l'agriculteur en fonction de son système de culture, des débouchés de ses récoltes et de son environnement. En fonction des tailles d'échantillons manipulés, la sélection et la dérive peuvent diminuer la diversité génétique de ces populations, mais grâce à des échanges de semences la variabilité est maintenue à l'échelle globale sur le principe de la métapopulation [Louette, 2000]. Cette pratique est surtout développée dans des pays où l'agriculture traditionnelle est toujours pratiquée [Elias et al., 2001; Alvarez et al., 2005] ; mais elle se développe avec un regain d'intérêt pour la conservation de la biodiversité par l'intermédiaire de la formation d'association comme le réseau de semences paysannes (RSP) en France [Dawson et al., 2011]. Ces associations travaillent avec des chercheurs et/ou des semenciers sur la conservation de la biodiversité, l'étude de l'interaction génotype-environnement, et l'innovation variétale. Cette démarche s'intègre dans la recherche participative [Ceccarelli and Grando, 2007; Rivière, 2014].

En gestion dynamique expérimentale, une population hétérogène est semée dans un réseau de sites contrastés d'un point de vue biotique et abiotique [Allard, 1988; Henry et al., 1991; Lavigne et al., 2001; Enjalbert et al., 2011]. Ces populations sont

conduites sans sélection humaine consciente, par échantillonnage dans la récolte précédente. La reproduction se fait par croisements libres. Les populations sont ressemées chaque année sur une surface constante et suffisamment importante pour limiter les effets de la dérive génétique.

Encadré 1.2.2 : Description des forces évolutives :
La sélection naturelle correspond au fait que les individus les plus adaptés à leur environnement produisent plus de descendants, et donc transmettent plus efficacement les gènes qu'ils portent aux générations suivantes. Réciproquement, les gènes conférant un meilleur succès de reproduction aux individus qui les portent voient leur fréquence augmenter avec le temps. La sélection au sein d'une population peut être directionnelle, stabilisatrice ou diversifiante.
La dérive génétique correspond aux fluctuations aléatoires des fréquences alléliques du fait du tirage d'un nombre fini d'individus à chaque génération. Elle est d'autant plus importante que les populations sont petites.
La migration représente les mouvements d'individus ou de gamètes (pollen) entre populations. Elle permet l'introduction de nouveaux gènes, et/ou combinaisons de gènes dans une population.
La mutation est la première cause de création de variabilité génétique. Elle agit aléatoirement, avec une probabilité très faible mais très variable (entre $5,5.10^{-9}$ et $2,4.10^{-4}$ par base et par génération chez le blé [Thuillet et al., 2002; Akhunova et al., 2010]), et elle est plus souvent délétère que bénéfique. Elle provient de l'erreur de copie des bases de l'ADN lors de la réplication de l'ADN. Il existe plusieurs types de mutations : inversion de base (ou substitution), insertion, délétion, translocation, polyploïdie.

Les populations maintenues en gestion dynamique sont soumises de façon continue aux pressions de sélection liées à leur environnement. Ce type de gestion vise donc à maintenir un pool génétique en interaction avec l'environnement, afin de favoriser son adaptation. Cette pratique de gestion de la diversité soulève beaucoup de questions notamment sur la diversité initiale nécessaire pour que la population s'adapte aux différents milieux, la taille efficace de la population, et le système de reproduction de la population étudiée. L'importance de ces trois facteurs est développée dans les paragraphes suivants et illustrée à partir des trois programmes qui ont développé une approche expérimentale de gestion dynamique :

- Une population d'orge issue du croisement de 28 variétés, et maintenue sur plus de 60 générations d'évolution [Allard, 1988] ;
- 12 méta-populations d'*Arabidopsis thaliana* issues de la même population initiale résultant du croisement contrôlé de 16 accessions [Lavigne et al., 2001] ;
- trois méta-populations de blé tendre issues de deux populations autogames dérivées chacune du croisement pyramidal de 16 lignées et d'une population allogame de 60 parents brassée grâce à l'intégration d'un gène de stérilité mâle (*ms1b*) [David, 1992].

La diversité initiale d'une population

La diversité initiale d'une population représente son pouvoir d'adaptation aux nouvelles conditions de culture. En effet une population a deux moyens d'évoluer : soit grâce à la présence d'une diversité initiale, soit grâce à l'apparition de nouvelles mutations [Barrett and Schluter, 2008]. L'évolution en réponse à la sélection sur la base de la diversité initiale est le moyen le plus rapide pour une population de s'adapter [Barrett and Schluter, 2008]. Comme les études de populations de gestion dynamique sont étudiées pour le moment pour leur évolution à court ou moyen terme, il est d'autant plus important de prendre en compte la diversité initiale dans ces populations. Comment estimer cette diversité ? Quelle est la diversité idéale ? Ces questions sont fondamentales dans tous les programmes de gestion des populations. La diversité d'une population peut être quantifiée par le nombre de parents participant à l'élaboration de cette population, mais l'utilisation de parents apparentés ou génétiquement très distants n'apporte pas la même diversité. A l'échelle du génome, le nombre de marqueurs polymorphes peut être un indicateur intéressant et pour un même nombre de marqueurs polymorphes, le nombre moyen d'allèles et leurs fréquences indiquent la diversité génétique de la population. L'indice de diversité de Nei (He) [Nei, 1973] a été développé sur ce principe, il varie entre 0 et 1, et représente la probabilité à un marqueur donné de tirer deux allèles différents (Equation 1.1). Dans le cas des marqueurs bi-allélique, le He à un marqueur donné ne peut excéder 0,5.

$$He = 1 - \sum_{i=1}^{n} f_i^2 \tag{1.1}$$

La diversité d'une population peut être estimée par la moyenne des He à chaque locus. Plus le He est élevé et plus la population possède une grande diversité génétique neutre.

Les trois programmes de gestion dynamique expérimentale d'orges, blés et arabettes ont utilisé entre 15 et 60 parents [Allard, 1988; David, 1992; Lavigne et al., 2001], choisis pour apporter une grande diversité, ce qui se traduit par des valeurs de He comprises entre 0,35 et 0,47 [Lavigne et al., 2001; Rhoné et al., 2008]. Paillard et al. [2000a] ont montré que la diversité intra-population a tendance à diminuer au cours des générations à cause des pressions de sélection et de la dérive. Cette diminution est d'autant plus accentuée par les effets de goulots d'étranglement dans les populations de petite taille [Porcher et al., 2004]. A l'échelle de la méta-population, la diversité globale est maintenue [Paillard et al., 2000a] du fait que chaque population est soumise à des pressions de sélection spécifiques et sous réserve que le nombre de populations soit suffisant.

A partir de cette diversité initiale, les populations en gestion dynamique ont évolué au niveau phénotypique et moléculaire en réponse aux pressions de sélection locales : une augmentation de hauteur due à la compétition pour la lumière [Allard, 1988; Raquin et al., 2008a], une sélection divergente pour la précocité de floraison

[Allard, 1988; Lavigne et al., 2001; Goldringer et al., 2006; Rhoné et al., 2008, 2010], et une sélection des résistances aux maladies [Ibrahim and Barrett, 1991; Paillard et al., 2000a,b; Enjalbert et al., 2011].

La taille efficace

La taille efficace (Ne, [Wright, 1931]) est définie comme la taille d'une population idéale de Wright-Fisher qui aurait une variation temporelle de fréquence allélique ou une consanguinité équivalente à celle de la population d'étude [Wright, 1969]. La population idéale de Wright-Fisher est une population avec un sex-ratio de 1 et avec des générations non chevauchantes dans laquelle chaque individu peut se reproduire avec n'importe quel autre individu du genre opposé [Fisher, 1930; Wright, 1931]. Elle permet donc de rendre compte de l'évolution de la population sous l'effet de la dérive que ce soit en termes de perte de diversité, fixation d'allèle délétère et de consanguinité [Wright, 1969]. Plus la taille efficace est faible, plus la dérive est forte et plus la pression de sélection nécessaire pour fixer des allèles favorables doit être forte [Robertson, 1960]. La taille efficace peut être estimée de deux manières différentes soit à partir des données démographiques du nombre d'individus reproducteurs dans la population (Ne_d, Equation 1.2) soit à partir de données d'évolution de fréquences alléliques (Ne_g, Equation 1.3).

$$\frac{1}{Ne_d} = \frac{1}{4N_m} + \frac{1}{4N_f} \tag{1.2}$$

où N_m est le nombre de mâles reproducteurs et N_f est le nombre de femelles reproductrices. On fait l'hypothèse que les individus reproducteurs contribuent de la même manière.

$$Ne_g = \frac{t}{2(Fc + \frac{1}{2S_0} + \frac{1}{2S_t)})} \tag{1.3}$$

où t est le nombre de générations entre les deux échantillonages, S_0 la taille de l'échantillon initial et S_t la taille d'échantillonage de la même population après t générations. Fc est un estimateur de la variance de fréquence allélique entre la génération 0 et la génération t [Nei and Tajima, 1981].

Comme les populations naturelles ne satisfont pas les conditions de population idéale, la taille efficace (Ne) n'est pas égale à la taille démographique de la population ("census size" N), elle est toujours beaucoup plus faible. Le ratio Ne/N est dépendant de la méthode d'estimation du Ne ; la méthode démographique donne un ratio de 0,1 [Frankham, 1995] qu'avec la méthode génétique le ratio est plutôt de 0,03 [Goldringer et al., 2001]. Dans les populations de gestion de dynamique de blé, ces ratios calculés avec les tailles efficaces génétiques étaient de 0,05 [Goldringer et al., 2001] et 0,07 [Raquin et al., 2008b].

Le système de reproduction

Les plantes peuvent avoir un régime de reproduction autogame (autofécondation stricte), allogame (fécondation croisée) ou un régime de reproduction mixte. Il est possible de modifier le régime de reproduction des plantes, soit par manipulation physique soit par manipulation génétique. Chez les plantes allogames qui comme le maïs (*Zea mais*) ont généralement des fleurs mâles et femelles séparées, les autofécondations se font manuellement par ensachage des fleurs femelles pour éviter tout contact avec le pollen extérieur et des fleurs mâles pour récupérer le pollen et par pollinisation manuelle des fleurs femelles de la même plante. Pour les plantes autogames, qui comme le blé ont les organes mâles et femelles à proximité et parfois enfermés dans une fleur hermaphrodite (cléistogamie), les croisements sont possibles soit par une opération de castration manuelle soit par modification génétique : auto-incompatibilité, ou intégration de stérilité mâle (nucléaire et/ou cytoplasmique) [Rao et al., 1990].

Dans une population isolée sans pression de sélection, le régime de reproduction n'a pas d'effet sur l'évolution de la diversité de la population. Dans le cadre de la gestion dynamique, les populations sont soumises à des pressions de sélection. Comme dans les populations autogames, les descendants ressemblent beaucoup à leur parent, la présence d'une combinaison allélique favorable dans un environnement donné augmentera rapidement en fréquence. Dans ce cas, la réponse à la sélection est plus rapide dans une population autogame [Haldane, 1924], la perte de diversité sera donc aussi plus rapide par rapport aux populations allogames [Hamrick and Godt, 1996]. En cas de flux de gènes avec un allèle favorable ou une nouvelle mutation favorable, ils pourront être intégrés par recombinaison, les populations allogames intègrant plus rapidement ces nouveaux allèles. Dans une population autogame, la fréquence du migrant ou du mutant augmentera lentement si il a une "fitness" suffisamment élevée pour éviter élimination par la dérive [Bürger, 1999].

La gestion dynamique est complémentaire à la conservation statique puisqu'elle permet de garder un nombre plus limité de populations qui continuent à produire de la variabilité adaptée aux conditions actuelles.

1.3 Utilisation des ressources génétiques pour analyser les bases génétiques de caractères complexes

Les ressources génétiques sont peu utilisées dans la création variétale [Zamir, 2001] car i) elles ne sont pas assez caractérisées [Tanksley and McCouch, 1997] ; ii) même si elles présentent des caractères d'intérêt pas encore présents dans les variétés modernes, elles ont des fonds génétiques très éloignés des variétés modernes. L'intégration de ces caractères d'intérêt dans les variétés modernes est réalisée par croisements puis rétro-croisements des descendants avec la variété moderne pour éliminer le fond génétique de la variété donneuse. L'utilisation des ressources génétiques est pour le moment limitée à l'amélioration des variétés modernes pour des caractères qualitatifs (mono-génique). Les caractères quantitatifs sont suivant leur complexité contrôlés soit par un ou plusieurs gènes à effets forts et par de

nombreux gènes à effets plus faibles avec des effets additifs et/ou des interactions épistatiques entre eux [Manolio et al., 2009]. Dans ce cas, l'utilisation des marqueurs moléculaires est indispensable pour faire du pyramidage de QTLs et savoir quels sont les meilleurs individus à croiser. Une meilleure caractérisation des ressources génétiques permettrait leur plus grande utilisation dans l'élaboration des variétés commerciales. Pour le moment, elles servent essentiellement à la compréhension de l'architecture génétique des caractères complexes *via* la détection de QTLs [Li et al., 2011; Rousset et al., 2011]. Dans la suite du chapitre, la construction de carte génétique utilisée pour la localisation des QTLs et différentes méthodes de détections de QTLs sont présentées.

1.3.1 Construction de cartes génétiques

La localisation des QTLs se fait grâce à des cartes génétiques construites classiquement à l'aide de populations de lignées recombinantes. Dans ces populations, l'association non aléatoire entre deux allèles à différents locus (appelée déséquilibre de liaison (DL)) est une simple fonction décroissante continue du taux de recombinaison entre ces marqueurs. Les marqueurs qui co-ségrégent sont donc localisés dans la même zone du génome et leur fréquence de recombinaison permet d'estimer leur distance. A l'inverse des marqueurs qui ne co-ségrégent pas doivent être localisés sur des groupes de liaisons différents. Quand l'analyse est réalisée avec un grand nombre de marqueurs, les marqueurs peuvent être ordonnés les uns par rapport aux autres, leur distance est estimée et le nombre de groupes de liaison doit correspondre au nombre de chromosomes.

1.3.2 Détection de QTLs à l'aide d'une population de lignées recombinantes

Un QTL est un gène qui a un polymorphisme lié à la variation phénotypique d'un caractère d'intérêt. La puissance de détection d'un QTL dépend de son effet et de sa fréquence allélique (Figure 1.2). L'effet du QTL représente la variation phénotypique observée entre des individus qui portent des allèles différents à ce locus. Les QTLs à effet fort sont facilement détectables même avec peu d'individus et peu de marqueurs (considérant que le variant causal a une fréquence allélique proche de 0,5) [Mackay et al., 2009] . En augmentant le nombre de marqueurs et d'individus, la puissance de détection est plus grande et le nombre de QTLs détectés augmente (QTLs à effet plus faible en plus des QTLs à effet fort) [Flint and Mackay, 2009]. La fréquence allélique des marqueurs et variants causaux sont des facteurs très influents sur la puissance de détection. En effet si le variant causal a une fréquence très faible, il faudra un nombre d'individus suffisamment important pour observer un nombre significatif de fois l'allèle et tester son effet [Manolio et al., 2009]. Les marqueurs avec des fréquences alléliques faibles (inférieure à 0,05) ont des puissances de détection très faibles puisque juste par le hasard ces marqueurs peuvent être associés au phénotype [Carbone et al., 2006].

La précision de localisation des QTLs est dépendante du déséquilibre de liaison présent dans la population et de la densité de marquage. Plus le DL de la population est faible, plus le QTL sera localisé précisément (petit intervalle de confiance) [Carbone et al., 2006] pour peu que la densité de marquage soit adéquate. Dans une population avec un DL très faible, une densité de marquage trop faible entraîne un manque de puissance de détection. Le QTL ne sera détecté que si parmi les marqueurs génotypés, certains sont localisés à faible distance du variant causal.

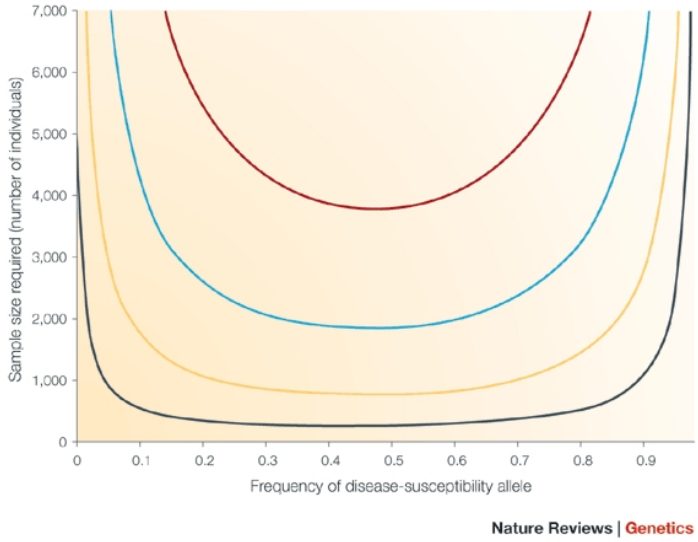


FIGURE 1.1 – Nombre d'individus nécessaires dans une étude d'association pour détecter des QTLs avec des ratios d'effets des allèles de 1,2 (rouge), 1,3 (bleu), 1,5 (jaune) et 2 (noire). Les chiffres correspondent à une puissance statistique de 80% avec un seuil de significativité de $p - value < 10^{-6}$ et un déséquilibre de liaison complet entre le QTL et le marqueur génotypé le plus proche.[Wang et al., 2005]

Classiquement la détection de QTLs est réalisée à l'aide de populations de lignées recombinantes initialement issues de croisements bi-parentaux [Haseman and Elston, 1972], dont les parents ont été choisis pour leur divergence pour le caractère étudié. Avec ce genre de population, le nombre de QTLs en ségrégation est assez limité (dépendant de la distance génétique entre les deux parents) et vu le faible nombre de générations de recombinaison, le DL est très fort et la localisation approximative. Par contre la puissance de détection des QTLs est très forte puisque les fréquences alléliques de chaque marqueur sont proches de 0,5. Pour réduire le DL dans ces populations et donc améliorer la précision de localisation sans perdre le pouvoir de détection, des populations hautement recombinantes de type "AIL" (Advanced Intercross Lines) ont été développées. Le principe est de croiser aléatoirement un certain nombre de générations les plantes F2 avant de fixer des lignées dites hautement recombinantes [Darvasi and Soller, 1995]. Puis dans le but d'avoir plus de QTLs en ségrégation, les populations multi-parentales et hautement recombinantes

sont apparues (populations MAGIC) [Churchill et al., 2004]. Leur grand nombre de parents augmente la diversité phénotypique et génétique, et donc le nombre de QTLs en ségrégation. De plus, leur nombre de générations de croisements réduit le DL et donc améliore la précision de la localisation. La limite de ce genre de population est qu'il est nécessaire de génotyper et phénotyper un grand nombre d'individus pour détecter tous les QTLs présents [Valdar et al., 2006; Cavanagh et al., 2013]. Avec ces populations de lignées recombinantes, la détection de QTLs se fait par analyse de liaison (linkage association) qui utilise les recombinaisons qui ont eu lieu lors de la création de la population.

1.3.3 Détection de QTLs à l'aide d'un panel d'association

En dehors de ces populations de lignées recombinantes, la détection de QTLs peut se faire à l'aide de panels de lignées par génétique d'association. Les lignées d'un panel sont soit des lignées issues de banque de graines choisies de manière à maximiser la diversité [Rousset et al., 2011], soit des lignées issues de populations naturelles [Slate, 2005]. Ces populations présentant une très large diversité ont un coût de création financier et temporel très limité par rapport aux populations de lignées recombinantes. Ces panels peuvent donc être utilisés pour un grand nombre de caractères d'intérêt. La génétique d'association se base sur l'utilisation des recombinaisons historiques d'un panel. Le DL inter-marqueur, beaucoup plus faible, permet une précision de localisation des QTLs bien meilleure qu'avec les populations bi-parentales. Par contre comme les lignées du panel sont assez distantes génétiquement, un certain nombre de marqueurs présente des allèles rares qui leur donnent une puissance de détection plus faible par rapport aux marqueurs avec des fréquences alléliques équilibrées. Ces panels peuvent présenter une structuration généralement liés à la provenance géographique des lignées [Rousset et al., 2011]. Cette structuration entraîne la détection de fausses associations. Cette structure peut être quantifiée soit i) à l'aide d'une matrice de structure (nommée Q) estimée par exemple par DAPC [Jombart et al., 2010] ou STRUCTURE [Pritchard et al., 2000] qui identifient un nombre de populations distinctes et assignent les individus aux populations avec un pourcentage d'appartenance soit ii) à l'aide d'une matrice d'apparentement (nommée K) calculée à partir de données généalogiques ou des données génétiques. Une correction de cette structure doit être alors intégrée dans des modèles statistiques [Yu et al., 2006]. Cette méthode utilisée avec un modèle d'analyse approprié par rapport à la population d'étude permet donc une détection plus précise mais moins puissante qu'une analyse avec une population de lignées recombinantes.

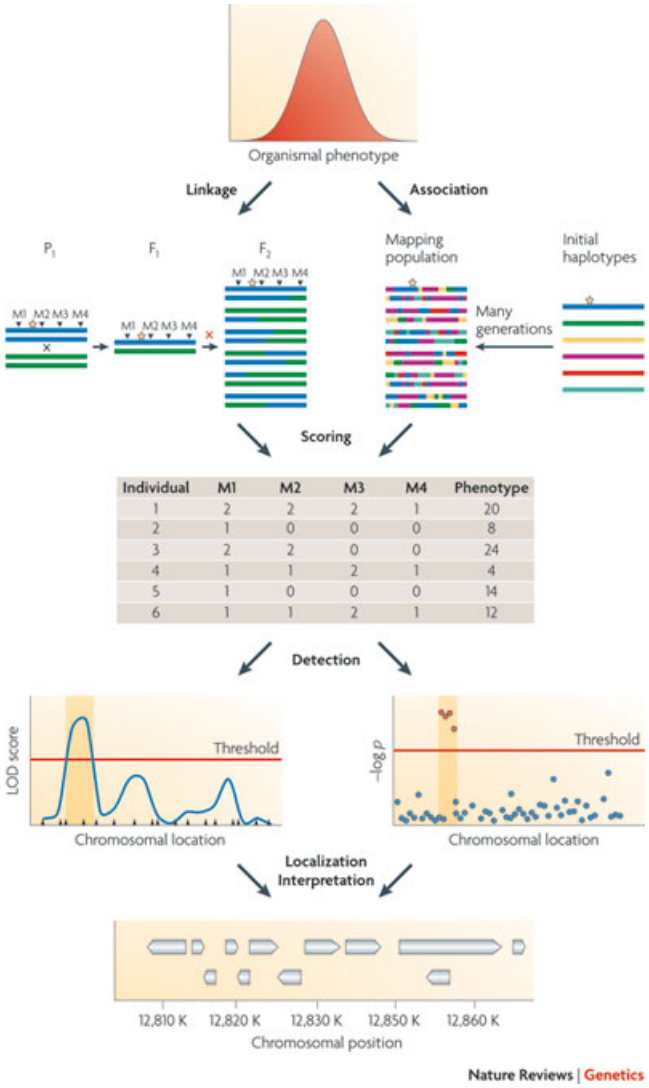


FIGURE 1.2 – Comparaison entre l'étude d'une population F2 et d'un panel d'association.
La figure montre une population F2 (à gauche) : Les parents (P1) sont deux lignées divergentes qui ont été croisées pour créer la génération F1. La génération F2 est issue de l'autofécondation des individus F1. Les marqueurs M1, M2, M3 et M4 sont des marqueurs polymorphes chez les parents. L'étoile jaune indique la position du QTL. La recombinaison a créé de nouveaux haplotypes chez les individus F2 qui permettent la localisation du QTL.
La génétique d'association (à droite) est aussi basée sur les recombinaisons, mais les recombinaisons ancestrales. Sur cette figure, six haplotypes de parents initiaux sont présentés. Les haplotypes de la population montrent comment les nombreuses générations de recombinaison aléatoires ont mélangé les haplotypes parentaux. Ce mélange permet de découpler tous les marqueurs qui ne sont pas liés au QTL. Le QTL sera détecté et localisé précisément seulement si un marqueur très lié a été génotypé. Quel que soit la stratégie, les populations de cartographie doivent fournir la variation génétique nécessaire pour expliquer le phénotype étudié. Les deux méthodes demandent un travail de phénotypage et de génotypage qui amèneront à déterminer s'il y a une différence significative du phénotype entre les individus qui portent des allèles différents à un marqueur. Si c'est le cas, le marqueur est lié au QTL.[Mackay et al., 2009]

1.3.4 Autres méthodes de détection de QTLs

Au début des années 2000, une nouvelle approche nommée LDLA (Linkage Disequilibrium Linkage Association) a été développée pour intégrer les avantages des deux méthodes décrites précédemment (génétique d'association et analyse de liaison) : augmentation de la précision de localisation et de la puissance de détection [Yu et al., 2006]. Cette méthode est utilisée avec des populations de lignées recombinantes de type bi-parentales connectées (NAM [Yu et al., 2008]) ou multiparentales et hautement recombinantes (MAGIC, [Kover et al., 2009]) pour lesquelles les parents sont connus. Elle est basée sur la reconstruction d'haplotypes parentaux chez les descendants et la prise en compte de l'apparentement entre les individus (matrice IBD (Identity By Descent)) [Meuwissen et al., 2002].

La détection de QTLs peut aussi être réalisée à l'aide de populations d'évolution expérimentales avec une sélection artificielle ou naturelle. Les populations avec sélection artificielle sont appelées populations de sélection récurrente. En exerçant une pression de sélection sur le même caractère sur plusieurs générations, les QTLs pourront être détectés à partir de l'évolution des fréquences alléliques entre la population initiale et la population sélectionnée [Foolad et al., 1997; De Koeyer et al., 2001; Thabuis et al., 2004]. Les populations avec sélection naturelle permettent quand à elles de détecter spécifiquement des QTLs associés à des caractères adaptatifs. En effet ces populations sont soumises aux pressions de sélection de l'environnement dans lequel elles sont cultivées, elles évolueront donc au niveau phénotypique et génotypique. En étudiant cette évolution soit par comparaison de deux échantillons d'une population cultivée dans une même condition à des générations différentes (évolution temporelle) soit d'une population cultivée dans deux conditions différentes (évolution spatiale), les zones soumises à sélection, c'est à dire des régions génomiques liées à l'adaptation, pourront être détectées [Allard, 1988; Goldringer et al., 2001; Porcher et al., 2004; Rhoné et al., 2007; Raquin et al., 2008a]. L'étude concomitante des changements de phénotypes et de fréquences alléliques et de l'association entre polymorphismes détectés et variation des caractères donne une validation de la détection du QTL.

Suivant la population d'étude et son histoire, il existe une diversité de méthodes pour détecter des QTLs. La population d'étude de cette thèse est une population multi-parentale et hautement recombinante de blé issue du croisement aléatoire de 60 parents, qui a évolué en panmixie durant 12 générations. Cette population peut s'apparenter à une population de type MAGIC mais le grand nombre de parents additionné aux multiples générations de croisements aléatoires nous a amené à l'étudier à la fois comme une population expérimentale qui a évolué sans sélection artificielle (Chap 4) et comme un panel de lignées (Chap 5.2). Le caractère d'étude choisi est un caractère majeur pour l'adaptation chez le blé, la précocité de floraison.

1.4 Etude d'un caractère adaptatif chez le blé tendre : la précocité de floraison

1.4.1 Le blé tendre : une céréale d'importance mondiale

Le blé tendre (*Triticum aestivum*) est la céréale la plus cultivée au monde. Elle est une des bases de l'alimentation humaine, animale et est maintenant transformée par l'industrie en bioéthanol. Avec plus de 220 millions d'hectares cultivés tous les ans [Shiferaw et al., 2013], elle a une zone de répartition très large sous des conditions climatiques très contrastées (Figure 1.3). Ce grand potentiel d'adaptation est peut-être dû à son allopolyploïdie [Dubcovsky and Dvorak, 2007]. Originaire du croissant fertile, l'ancêtre sauvage du blé tendre n'existe pas dans la nature. Il est le résultat d'une hybridation naturelle inter-specifique [Stebbins Jr, 1950] (Figure 1.4). Dans un premier temps, *Triticum urartu* (génome A) s'est hybridé avec *Triticum sect. Sitopsis* (génome B) il y a entre 150 000 et 500 000 ans donnant naissance au *Triticum turgidum spp. dicoccoïdes* (AABB), espèce sauvage, ancêtre du blé dur [Dvorak and Akhunov, 2005; Charmet, 2011]. Dans un deuxième temps *Triticum turgidum spp. dicoccum*, blé domestiqué, s'est croisé avec *Aegilops tauschii* (génome D) il y a entre 7 000 et 10 000 ans [Charmet, 2011]. Le blé tendre est donc une espèce allohexaploïde ce qui signifie qu'il possède trois génomes homéologues A, B et D contenant chacun sept paires de chromosomes (2n=6x=42). Le génome du blé est un génome de 17Gb correspondant à 20 fois celui du riz [Arumuganathan and Earle, 1991] ou cinq fois celui de l'homme [Gill et al., 2004]. A cause de son allopolyploidie et de la proportion de séquences répétées [Li et al., 2004], la séquence du génome du blé n'a été publiée qu'en 2012 [Brenchley et al., 2012] alors que 41 génomes de 36 espèces différentes avaient déjà été séquencés [Michael and Jackson, 2013], mais la pseudo-molécule, assemblage propre de tous les "scaffolds", n'est toujours pas disponible.

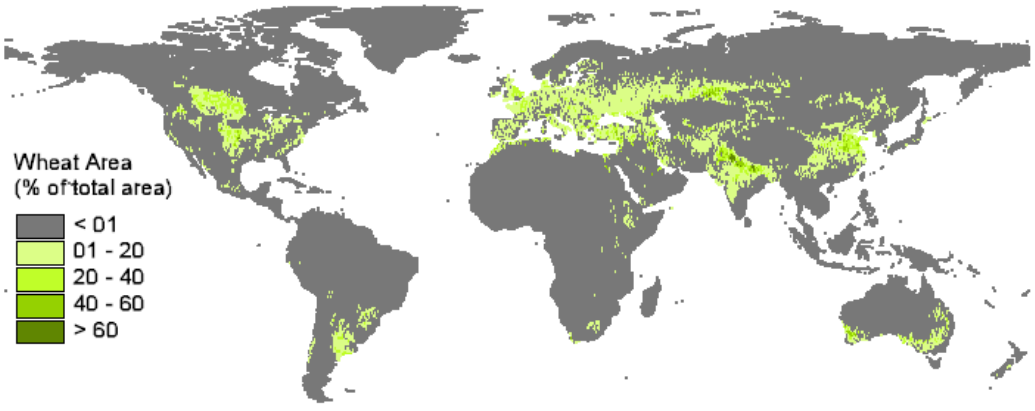


FIGURE 1.3 – Répartition mondiale des surfaces récoltées de blé. [Monfreda et al., 2008]

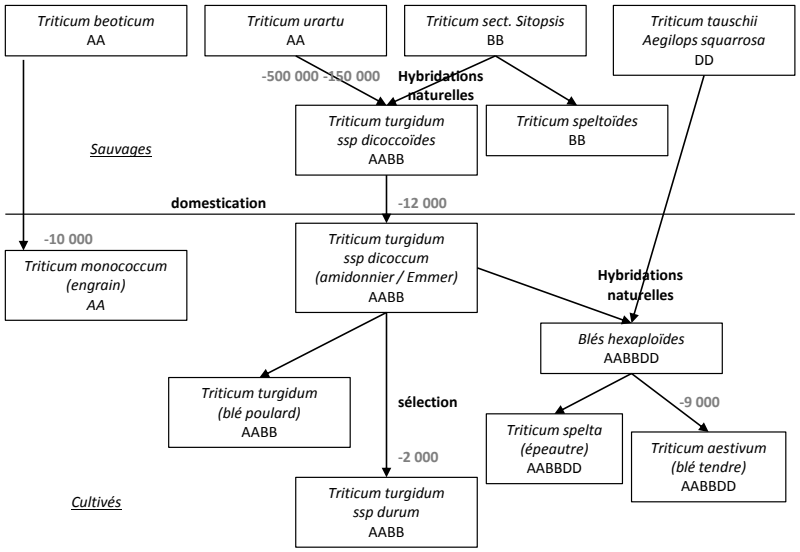


FIGURE 1.4 – Origine des blés domestiqués.

1.4.2 Importance de la précocité de floraison chez le blé tendre

La précocité de floraison chez le blé tendre, mesurée notamment par la date d'épiaison est un caractère adaptatif clé [Worland, 1996]. En effet elle synchronise le cycle reproducteur de la plante avec les conditions environnementales. Comme les plantes sont sensibles aux maladies, à la sécheresse et/ou aux fortes températures, et que ces sensibilités varient au cours de leur cycle de développement, cette coordination est importante. Elle est d'autant plus importante dans le cadre du changement climatique que la fréquence et l'intensité des sécheresses et des fortes chaleurs devraient augmenter [Lobell et al., 2011] et que les plantes devront éviter ces stress en ajustant leur phénologie [Gouache et al., 2012]. Cet évitement permet donc d'optimiser le rendement. La précocité de floraison joue un rôle important dans l'élaboration de ce caractère, en effet une période de sécheresse durant la floraison entraîne une diminution de rendement de 14% à plus de 60% [Ficher and Maurer, 1978]. Ce lien est également révélé dans des analyses génétiques, des effets pléiotropes ayant été mis en évidence par la co-localisation des gènes contrôlant la précocité et des QTLs liés au rendement [Kato et al., 2000].

1.4.3 Architecture génétique de la précocité de floraison chez le blé tendre

La précocité de floraison est un caractère complexe divisé en trois composantes : le besoin en vernalisation, la sensibilité à la photopériode et la précocité intrinsèque. Le besoin en vernalisation est la nécessité pour les plantes sensibles de subir une période de froid pour initier la floraison. Ce caractère quantitatif permet toutefois de catégoriser les blés en deux classes. Les plantes de type printemps n'ont pas besoin de période de vernalisation pour fleurir, alors que les plantes de type hiver en nécessite une d'une longueur génotype dépendant. La sensibilité à la photopériode

est le besoin ou non de subir des jours longs afin de pouvoir épier. Une sensibilité au jour long évite une floraison précoce avant la période hivernale même si les températures sont douces. La troisième composante, la précocité intrinsèque, est définie comme la précocité une fois que les besoins en vernalisation et la sensibilité à la photopériode sont satisfaits.

1.4.3.1 Le besoin en vernalisation

Le besoin en vernalisation est contrôlé par des gènes majeurs nommés *Vrn*. Les gènes homéologues *VrnA1, VrnB1 et VrnD1* sont localisés au même endroit sur le bras long des chromosomes 5A, 5B et 5D, respectivement [Law et al., 1976]. Il est possible qu'il s'agisse du même gène homéologue déjà présent chez les ancêtres diploïdes [Law and Worland, 1997]. Ces trois gènes ont une interaction épistatique avec dominance des allèles d'insensibilité [Snape et al., 2001] qui accélèrent le cycle de développement aussi bien de la phase végétative que de la phase reproductive [Snape et al., 2001]. Que ce soit après des conditions vernalisantes ou non, les plantes de type printemps fleurissent plus rapidement que les plantes de type hiver pour un même fond génétique [Snape et al., 2001]. En plus de ces trois gènes majeurs, les gènes *Vrn2* (*VrnA2, VrnB2, VrnD2*) et *Vrn3* (*VrnA3, VrnB3, VrnD3*) ont aussi été identifiés comme accélérant la floraison. Ils sont localisés respectivement sur les chromosomes 5A, 4B, 4D, 7A, 7B et 7D [Yan, 2003; Yan et al., 2004b, 2006; Bonnin et al., 2008]. Contrairement aux autres, le gène *VrnD4* localisé sur le chromosome 5D accélère la floraison des plantes de type hiver après une faible période de vernalisation (5-10 jours) [Yoshida et al., 2010; Kippes et al., 2013].

Au niveau moléculaire, les polymorphismes impliqués dans des variations phénotypiques sont : une longue délétion de 2,8kb localisée dans l'intron 1 des gènes *Vrn1* [Fu et al., 2005] et des délétions et mutations dans le promoteur de *VrnA1* [Yan et al., 2004a]. Tous ces polymorphismes modifient les sites de reconnaissance de *Vrn1* par *Vrn2*, qui est un répresseur. *Vrn2* est lui même réprimé par la vernalisation [Yan, 2003]. *Vrn2* peut aussi présenter une mutation sur son site de fixation qui inhibe son action sur *Vrn1* [Yan et al., 2004b]. *Vrn1* en jour long active *Vrn3* qui a une rétro-action positive sur *Vrn1* par l'intermédiaire de la répression de *Vrn2* (Figure 1.5).

1.4.3.2 La sensibilité à la photopériode

La sensibilité à la photopériode est régulée par une famille de gènes homéologues *Ppd-A1, Ppd-B1 et Ppd-D1*, respectivement localisés sur les chromosomes 2A, 2B et 2D à des positions similaires sur le bras court [Welsh et al., 1973]. Ils réduisent tous la phase de développement des primordium ce qui induit une montaison plus précoce [Snape et al., 2001]. Les polymorphismes responsables de l'insensibilité à la photopériode sont une délétion de 2kb dans le promoteur de *Ppd-D1* [Beales et al., 2007], deux délétions indépendantes de *Ppd-A1* [Wilhelm et al., 2009] et des mutations de *Ppd-B1*. La modification de *Ppd-D1* précocifie la plante de 6 à 14

jours [Snape et al., 2001] sachant que *Ppd-D1* a un effet plus fort que *Ppd-A1* qui lui même a un effet plus fort que *Ppd-B1* [Snape et al., 2001; Bentley et al., 2011]. La précocité de floraison est donc dépendante des polymorphismes aux locus *Ppd* et donc du nombre d'allèles d'insensibilité présents [Shaw et al., 2012].

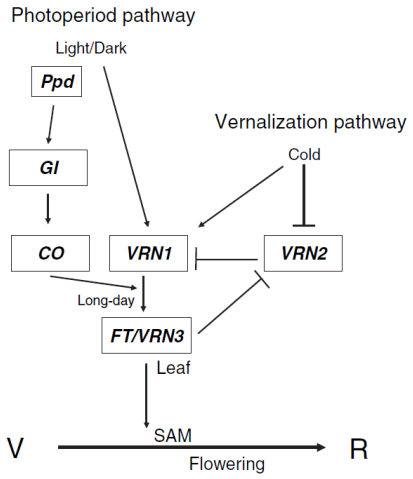


FIGURE 1.5 – Modèle d'interaction des gènes impliqués dans régulation de la précocité de floraison. [Shimada et al., 2009]

1.4.3.3 La précocité intrinsèque

La précocité intrinsèque est la composante la moins bien connue, elle est définie comme la variabilité restante quand les plantes ont leur besoin en vernalisation et en photopériode satisfaits. Il a été montré qu'elle était liée à la sensibilité à la température [Slafer and Rawson, 1995]. Une étude sur *Triticum monoccocum*, un blé diploïde (contenant un génome apparenté au génome A du blé hexaploïde), a mis en évidence les gènes $EpsA^m1$ et $Eps\text{-}3A^m$ localisés sur les chromosomes respectifs 1A et 3A [Bullrich et al., 2002; Gawroński and Schnurbusch, 2012]. $EpsA^m1$ entraîne un allongement des phases végétative et reproductrice.

1.4.3.4 Interactions entre les composantes

Il existe des interactions fortes entre les composantes de besoin en vernalisation et de sensibilité à la photopériode qui ne sont pas encore bien comprises dans leur globalité, mais quelques points ont été démontrés : i) la répression de *Vrn2* par la vernalisation peut-être remplacée par l'interruption des jours longs par six semaines de jours courts [Dubcovsky et al., 2006] ; ii) l'allèle insensible de *Ppd-D1* active l'expression de *Vrn3*, de manière directe ou indirecte [Shaw et al., 2012] et iii) l'allèle hiver de *Vrn1B* a une forte interaction avec *Ppd-D1* qui en jour long multiplie par 2,5 son effet [Zheng et al., 2013] (Figure 1.5).

1.4.4 Stratégie d'adaptation par la précocité de floraison chez le blé tendre

A l'origine le blé a été domestiqué dans le croissant fertile, dans le Sud Est de la Turquie et avait un phénotype de type hiver et sensible à la photopériode. Aujourd'hui, il représente 17% des surfaces mondiales cultivées [Gill et al., 2004] (Figure 1.3) et peut présenter une insensibilité à la vernalisation et/ou à la photopériode. Vu la multitude de polymorphismes associés à la précocité de floraison présents dans les gènes *Vrn* et l'association détectée entre ce caractère et la provenance géographique des accessions [Kato and Yokoyama, 1992; Matsuoka et al., 2008], nous nous sommes demandés si il y avait une stratégie d'adaptation particulière basée sur la présence d'allèle spécifique en fonction des régions de la planète et/ou des conditions de culture. Pour étudier cette hypothèse, j'ai réalisé en début de thèse une méta-analyse des données disponibles de répartition des plantes printemps/hiver et des allèles printemps de *Vrn1* (*VrnA1, VrnB1 et VrnD1*) dans le monde à partir de la base de données européenne du blé (ewdb, `http://genbank.vurv.cz/ewdb/`) développée par le programme coopératif européen pour les ressources génétiques végétales (ECPGR) et des publications suivantes : Worland et al. [1994]; Iwaki et al. [2000, 2001]; van Beem et al. [2005]; Moiseeva and Goncharov [2007]; White et al. [2008]; Herndl et al. [2008]; Zhang et al. [2008]; Sun et al. [2009]; Yang et al. [2009]; Guo et al. [2010]; Kolev et al. [2010]; Andeden et al. [2011]; Iqbal et al. [2011]. Avec 53 847 données récoltées initialement, 42 550 données localisées dans 83 pays ont été utilisées dans l'analyse (Tableau 1.1).

Tableau 1.1 – Tableau récapitulatif des données utilisées pour les analyses de distribution du caractère Printemps/Hiver, et des allèles des gènes *VrnA1*, *VrnB1*, *VrnD1* et *Ppd-D1*.

Trait	nb données	nb pays	références
Printemps/Hiver	38 038	81	ewdb ; Worland et al. [1994]; Iwaki et al. [2000, 2001]; White et al. [2008]; Sun et al. [2009]; Guo et al. [2010]
VrnA1	1 094	57	Iwaki et al. [2000, 2001]; van Beem et al. [2005]; Moiseeva and Goncharov [2007]; White et al. [2008]; Zhang et al. [2008]; Sun et al. [2009]; Kolev et al. [2010]; Andeden et al. [2011]; Iqbal et al. [2011]
VrnB1	1 046	55	Iwaki et al. [2000, 2001]; van Beem et al. [2005]; Moiseeva and Goncharov [2007]; White et al. [2008]; Zhang et al. [2008]; Andeden et al. [2011]; Iqbal et al. [2011]
VrnD1	1 094	58	Iwaki et al. [2000, 2001]; van Beem et al. [2005]; Moiseeva and Goncharov [2007]; White et al. [2008]; Zhang et al. [2008]; Kolev et al. [2010]; Andeden et al. [2011]; Iqbal et al. [2011]
Ppd-D1	1 278	11	White et al. [2008]; Herndl et al. [2008]; Yang et al. [2009]; Kolev et al. [2010]; Andeden et al. [2011]

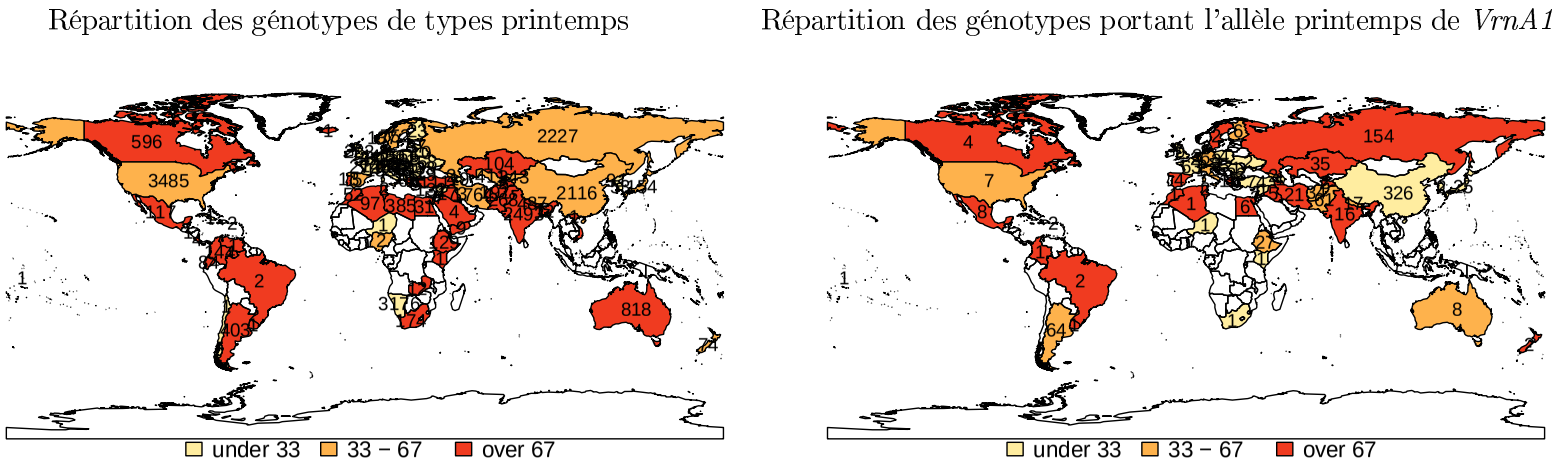
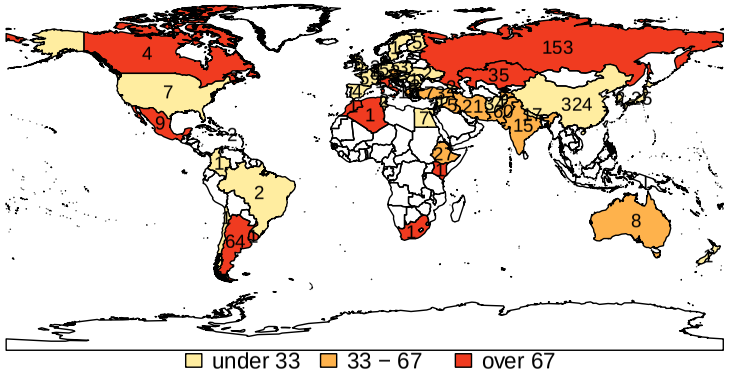


FIGURE 1.6 – Distribution mondiale des génotypes de type printemps (à gauche) et des génotypes portant l'allèle printemps de *VrnA1* (à droite). Les couleurs représentent la proportion par pays avec trois classes : $< 33\%$, entre 33% et 67% et $< 67\%$, avec une couleur plus foncée pour les pays avec des proportions plus élevées. Le chiffre correspond au nombre de données utilisées par pays.

Répartition des génotypes portant l'allèle printemps de *VrnB1*



Répartition des génotypes portant l'allèle printemps de *VrnD1*

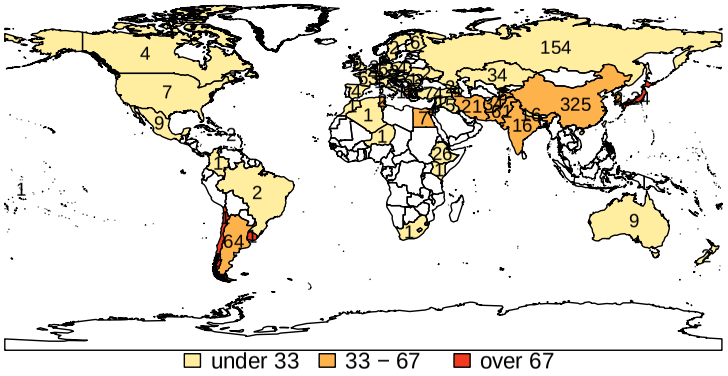


FIGURE 1.7 – Distribution mondiale des génotypes portant l'allèle printemps de *VrnB1* (à gauche) et des génotypes portant l'allèle printemps de *VrnD1* (à droite). Les couleurs représentent la proportion par pays avec trois classes : $< 33\%$, entre 33% et 67% et $< 67\%$, avec une couleur plus foncée pour les pays avec des proportions plus élevées. Le chiffre correspond au nombre de données utilisées par pays.

La carte 1.6 montre que les plantes de type printemps sont surtout présentes dans les régions tropicales (Afrique, Amérique Latine, Australie,...) ou alors dans des pays avec des températures très froides en hiver comme le Canada. Généralement les plantes de type printemps sont semées tard pour éviter le gel et récoltées tôt pour éviter la sécheresse estivale [Law and Worland, 1997]. Les plantes de type hiver sont plus cultivées dans des pays avec des hivers tempérés. Elles sont généralement semées tôt dans la saison [Law and Worland, 1997], avec un cycle long qui maximise le rendement [Worland, 1996]. Cette répartition est cohérente avec le fait que cette adaptation se fait en fonction des températures hivernales [Iwaki et al., 2000; Zhang et al., 2008] ; avec plus précisément les plantes de type printemps cultivées dans les zones où la température hivernale est inférieure à -7°C ou supérieure à 4°C, et les plantes de types hivers dans les zones avec des conditions intermédiaires [Iwaki et al., 2001]. A l'origine, le blé avait un phénotype de type hiver avantageux pour éviter de commencer la montaison trop tôt si l'hiver est doux et plus résistant à certaines maladies. Mais la forme printemps a été rapidement sélectionnée durant la période d'expansion des céréales post-domestication, car elle pouvait être semée et récoltée sur une courte période. Le cycle rapide des formes printemps permettait de faire deux récoltes par an. Comme les deux formes apportaient un avantage notable dans des conditions variées, elles ont persisté jusqu'à maintenant [Cockram et al., 2007].

Pour étudier plus précisément la stratégie d'adaptation des plantes au niveau moléculaire pour les besoins en vernalisation, nous avons regardé la distribution des allèles printemps de chacun des trois gènes homéologues de *Vrn1*. Une comparaison globale des trois cartes (Figure 1.6 & 1.7) montrent que les allèles printemps de *VrnA1* sont plus fréquents que ceux de *VrnB1*, et que l'allèle printemps de *VrnD1* est plutôt rare. L'allèle printemps de *VrnA1* est surtout présent en Asie excepté la Chine, en Amérique et en Europe. Celui de *VrnB1* est aussi très présent en Asie (excepté la Chine) mais beaucoup moins en Europe et en Amérique. L'allèle printemps de *VrnD1* est surtout présent dans les plantes cultivées au Japon, au Chili et en Uruguay et un peu dans le sud de l'Asie et en Argentine. La comparaison entre les cartes de *VrnA1* et *VrnB1* donne l'impression que l'allèle printemps de *VrnB1* est souvent présent de façon corrélée avec l'allèle printemps de *VrnA1*. Il n'y a que quelques pays d'Afrique pour lequel l'allèle de *VrnB1* est plus abondant que celui *VrnA1*.

La comparaison des quatre cartes (Figure 1.6 & 1.7) permet de voir la cohérence entre la présence des plantes de type printemps et la présence des allèles printemps des gènes de *Vrn1*, sachant que ces allèles sont dominants. Certains pays comme le Canada, l'Australie et l'Argentine ont une proportion très élevée de blé de printemps qui peut être liée à la présence d'allèles printemps de *VrnA1* ou de *VrnB1* voire de *VrnD1* pour l'Argentine. Dans d'autres pays comme les Etats-Unis, le Japon ou la Chine, pour lesquels la proportion de blé de printemps est plus faible, une stratégie allèle spécifique semble être en place. En effet respectivement l'allèle printemps de *VrnA1*, *VrnD1* et *VrnD1* est présent relativement plus souvent que les deux autres pour les Etats-Unis, le Japon et la Chine. En Europe, le peu de plantes de

type printemps semblent tous avoir l'allèle printemps de *VrnA1*. Tous ces résultats semblent cohérents mais pour la Russie, en comparaison avec la proportion moyenne (autour de 50%) de plantes de type printemps, les proportions d'allèle printemps de *VrnA1* et de *VrnB1* semblent très élevées. Les données utilisées pour créer les cartes ne proviennent pas toutes des mêmes publications et donc des mêmes plantes. Il peut y avoir des biais d'échantillonnage des génotypes utilisés pour certaines études. De plus, une étude uniquement des variétés de pays aurait été plus rigoureuse pour comprendre l'adaptation des plantes à leur milieu mais trop peu de données étaient disponibles.

Le même travail de recueil de données a été réalisé pour la sensibilité à la photopériode à travers le locus *Ppd-D1* mais trop peu de pays étaient couverts par les données récoltées (Tableau 1.1). L'allèle d'insensibilité à la photopériode est présent chez 42% des variétés de pays turques [Andeden et al., 2011] et chez 38,6% des variétés de pays chinoises [Yang et al., 2009]. Quelle que soit la région du monde, l'insensibilité à la photopériode a été introduite dans la majorité des variétés modernes (entre 60 et 91% suivant les pays) [Trethowan et al., 2007; Yang et al., 2009; Guo et al., 2010; Kolev et al., 2010; Andeden et al., 2011]. L'allèle d'insensibilité à la photopériode est souvent intégré dans les variétés modernes afin de les précocifier. Les variétés qui conservent l'allèle sensible sont destinées aux hautes latitudes d'Asie central et du Nord-Est de la Chine [Yang et al., 2009], dans les régions où les étés sont chauds et secs [Cockram et al., 2007]. Cet allèle d'insensibilité est chez l'orge une mutation apparue durant l'expansion des céréales post-domestication qui a été rapidement sélectionné pour leur avantage écologique dans les pays d'Europe du Nord [Cockram et al., 2007].

La connaissance des différents polymorphismes contrôlant la précocité de floraison et des combinaisons alléliques les plus adaptées à des environnements spécifiques permettront de fournir aux agriculteurs des variétés plus performantes pour leurs conditions de culture. Aujourd'hui ce contrôle se fait actuellement en faisant varier les allèles des gènes majeurs comme *Vrn1* ou *Ppd* mais pour avoir un contrôle plus fin il faudra aussi exploiter les allèles aux gènes mineurs et continuer à le travail de détection de QTLs car pour le moment les QTLs trouvés n'expliquent pas la totalité de l'héritabilité observée.

1.5 Objectifs et plan de la thèse

L'objectif de cette thèse est d'étudier l'intérêt d'une nouvelle population multi-parentale et hautement recombinante (de type MAGIC) issue du projet de gestion dynamique pour analyser l'architecture génétique d'un caractère complexe. Cette population est une population allogame de blé issue du croisement naturel aléatoire de 60 lignées parentales grâce à l'intégration d'un gène de stérilité mâle. Elle a été cultivée pendant 12 générations sur le site du Moulon (48,4°N, 21°E) (Chap 2). Le caractère d'étude choisi est la précocité de floraison. Ce caractère déjà très fortement étudié est contrôlé notamment par des gènes majeurs déjà identifiés. Mais ces gènes majeurs n'expliquent pas toute la variabilité phénotypique observée, un deuxième objectif était donc d'affiner les connaissances de l'architecture génétique de ce caractère.

Dans un premier temps, j'ai étudié la structure du déséquilibre de liaison présent dans la population. A partir de ce déséquilibre de liaison, j'ai développé un algorithme d'ordonnancement de marqueurs car avec des populations construites à partir de croisements non contrôlés, les algorithmes classiques d'ordonnancement de marqueurs ne peuvent être utilisés. Cet algorithme a été testé sur différentes populations pour définir ces conditions d'utilisation (Chapitre 3).

Puis, j'ai analysé l'évolution de cette population durant les 12 générations de panmixie en comparaison avec une population initiale dont la composition a été estimée grâce au développement d'une méthode Bayésienne. Cette comparaison a été réalisée aussi bien à l'échelle phénotypique pour les caractères précocité de floraison et le type printemps/hiver qu'à l'échelle génotypique. La comparaison des deux populations a aussi permis de détecter des régions génomiques soumises à sélection et en analysant leur association avec le phénotype de détecter des QTLs de la précocité. (Chapitre 4).

Enfin, j'ai décrit tous les caractères que nous avons phénotypés et créé des stratégies pour rassembler cette information en quelques variables, dans le but de faire de la détection de QTLs. Cette détection de QTLs a été réalisée classiquement par génétique d'association après avoir déterminé le modèle le plus approprié. Les résultats de cette analyse nous ont permis de discuter sur l'intérêt de cette population pour détecter des QTLs nouveaux et préciser la localisation de QTLs déjà connus (Chapitre 5.2).

Chapitre 2

Matériels et Méthodes

Ce chapitre présente le matériel végétal utilisé dans la thèse, ainsi que les méthodes utilisées pour l'acquisition des données phénotypiques et moléculaires. Les méthodes statistiques utilisées pour l'analyse des données sont détaillées dans les chapitres suivants.

2.1 Matériel végétal étudié

2.1.1 Création de la population d'étude

La population MAGIC INRA étudiée dans ce travail est constituée d'un ensemble de lignées issues de la population composite "allogamisée" (pool génétique PS) conduite en gestion dynamique sur le site du Moulon depuis 1984. La population initiale PS a été créée entre 1976 et 1980, suivant un plan de croisement partiellement détaillé dans les archives disponibles [Trottet, 1988] (Annexe A). Conçue pour avoir une base génétique très large, cette population est issue du croisement de 60 variétés ou lignées de pré-breeding des années 70, ou de variétés de pays, principalement d'origine européenne, mais également d'autres régions du monde (Tableau 2.1) [Le Boulc'h, 1994]. Ces lignées ont été choisies pour leur résistance à *Septoria nodorum*, ainsi qu'à d'autres maladies, ou pour leur courte paille et leur bonne valeur agronomique. Concernant la création de la population, les archives décrivent tout d'abord le croisement parallèle de 50 parents avec un mutant de la variété "Probus" [Fossati and Ingold, 1970] homozygote pour l'allèle de stérilité mâle nucléaire récessif *ms1b* [McIntosh, 1988]. Les plantes F1 obtenues ont été soit autofécondées, soit rétro-croisées avec les 50 parents auxquels ont été ajoutés neuf nouveaux parents (Probus exclu) pour diminuer la proportion du génome de Probus dans la population. Au total environ 5 000 plantes (F2, BC1 et nouveaux F1) ont été semées en mélange dans une parcelle isolée des autres parcelles de blé (distance supérieure à 100m) et entourée d'une bordure de seigle protégeant des flux de pollen extérieurs. A chaque génération, les individus mâles stériles, reconnaissables par leur phénotype au moment de la floraison (Figure 2.1), ont été marqués et récoltés à maturité. Les graines récoltées sur ces épis mâles stériles sont obligatoirement issues de croisements aléatoires par pollinisation anémophile. Comme les plantes mâles stériles (*ms1b/ms1b*) sont pollinisées par du pollen provenant d'une plante fertile hétérozygote (*Ms1b/ms1b*), les descendants ont une composition stable de 50% de mâles stériles (*ms1b/ms1b*) et de 50% de mâles fertiles (*Ms1b/ms1b*). Les grains issus de mâles stériles ont ainsi été récoltés puis été ressemés en isolement durant

12 générations successives entre 1984 et 1996, sans migration ni sélection humaine consciente. A chaque génération environ 3 000 épis mâles stériles ont été récoltés, permettant de ressemer environ 10 000 graines. En 2006, 1 026 descendants (S0) ont été échantillonnés et semés pour développer des lignées par la méthode de Single Seed Descend (SSD), la moitié a été fixée par l'équipe de Ian Mackay au NIAB (Cambridge, UK) et l'autre moitié à l'INRA du Moulon. Le nom des lignées est composé d'une lettre (A ou F) suivi d'un chiffre, la lettre correspondant au lieu de fixation (Angleterre ou France). Les familles S4 ont été semées en 2010-2011. Au total la population a subi 15 méioses efficaces. Pour notre étude, nous disposons de 1 026 lignées SSD et de 56 lignées parentales, au lieu des 60 lignées initiales, car quatre ne sont plus disponibles (Tableau 2.1).

FIGURE 2.1 – Photo d'une parcelle avec les épis mâle-stériles marqués par une laine rouge. Dans l'encadré, gros plan sur un épi fertile (à gauche) et sur un épi mâle stérile (à droite), reconnaissable grâce au bâillement de ses glumelles.

FIGURE 2.2 – Schéma représentant les différentes générations de fixation par la méthode Single Seed Descent. La moitié des fixations a été réalisée au NIAB en Angleterre et l'autre à l'INRA du Moulon en France.

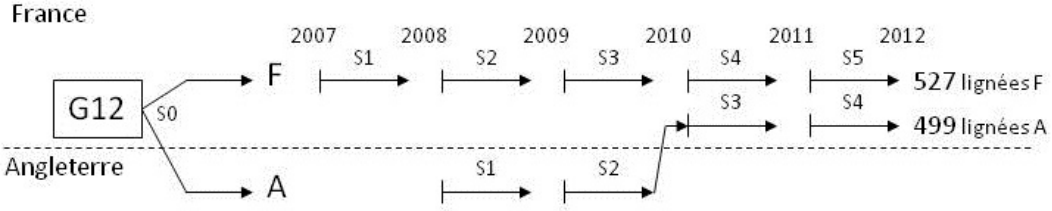


Tableau 2.1 – Liste des génotypes parentaux retrouvés dans les archives.

Génotype	Généalogie	Origine
DUCAT	Prieur / 3 / SW // 90 / Etoile de Choisy	France
LUTIN	Champlein / Cappelle // Versailles-B21	France
COMTAL	T.durum / Champlein // Dakota-51	France
COURTOT	Mexique-50 / Versailles-B21	France
TALENT	Champlein / 3 / Thatcher / Vilmorin27 // Fortunato	France
CA68	(Champlein x Aronde) 68	France
C 56-5	(2-7 x Cappelle) x 16-4-2) 56-5 (avec 16-4-2 lignée soeur de Courtôt)	France
REDON M4	Variété de pays de Bretagne	France
VPMM1-1-1-2 R4		France
(VPM x Moisson4) 3lignées		France
R 5-1	[(VPMM x Moisson) 9 x (US(60)43 x Prieur) 61] 5-1	France
V2D11	(VPMM 1.1.2.4 x D65.5) 11-3	France
mutant du 81-12	[(68-2 x Yga) 8-2 x (90-2 x Etoile de Choisy) 1-8]	France
CARALA	sélection dans Alabama bluestem	USA
HADDEN		USA
US 113		USA
US 117		USA
US 123		USA
US 125		USA
ATLAS 66	Frondosa / Redhart3-Noll28	USA
ANDES 56		Amérique du Sud
LAFRON		Amérique du Sud
GRANA	Etoile de Choisy x Wysokolitewka Sztywnosloma x Dankowskabiola	Pologne
MIRONOVSKAIA 808	sélection Artemovka	URSS
KAVKAZ	Lutenscens 314 H147 / Bezostaya 1	URSS
3596-58		Bulgarie
WEINSTEPHAN 1007-53	Heines Bart / seigle 4x // ? x blé tendre	Allemagne
DOMUS		Allemagne
ORLANDO	descent from rye with Neuzucht	Allemagne
LAPIS		Allemagne
MARIS HUNTSMAN	Ci 12633 / 5 x Cappelle // Hybride46 / Cappelle / 3 / 2 x Professeur Marchal	GB
SAPPO	Ci 12633 / 5 xRing // Els / 6 x Ring	GB
CHALK	Carsten VIII / H4255	GB
TJB 155 = KINSMAN	[(Ci 12633 x Cappelle) x (Hybride45 x Cappelle)] x (Professeur Marchal x Maris Ranger)	GB
TJB 240	Maris Envoy / TL365-A25	GB
TJB 251 *		GB
TJB 636		GB
TL 25-11 *		GB
TL 365 A34	TJB 16-18 / 3 / Cappelle // Vilmorin 29 / VG8058	GB
MARIS HOBBIT	(Professeur Marchal x (Marne x VG-9144)) x TJB 16	GB
CORIN	ST 102 x Tadorna	GB
CLEMENT	Hope / Timstein // 3x Heines VII / Riebesel 57-41 / 2 x Heines VII / 4 / Cléo	Hollande
NAUTICA	Mildress / Manella	Hollande
FERMO Mutant 9-14		Suisse
CARDENAL	Mengavi / 8156 // Jar / 3 / 8157	Australie
DARKAN	Eureka2 / Kenya C6041 // ?	Australie
OXLEY	Penjamo62 /4x Gabos56 // TPP / Nainari60 / 4 / 2 x Lerma Rojo // Norin10 / Brevor14 / 3 / 3 x Andes	Australie
CONDOR	Lerma Rojo // Norin10 / Brevor14 / 4 / Y54 // Norin10 / Brevor14 / 3 / 3 x Andes	Australie
HARUHIKARI		Japon
TOROPI	Petitblanco8 // Frontana 1971-37 / Quaderna	Brésil
LAGOA VERMELHA	Variété de pays	Brésil
IAS 63 *		Brésil
IAS 20	Iassul =Colonias // Kenya58 / Frontana	Brésil
V3D8 *	[(VPMM 1-1-2-4 x D65-5) 8-4	France
MINISTRE NAIN	Benoist40 / Professeur Delos // ? x lignée RHT3	Belgique
L707	Wakeland x Blueboy	USA
OASIS	Arthur / 5 / Arthur x 3 / 3 / Ribo // Riley x 2 / Riley67	USA
REDHART	sélection dans Southern Flint ou Red May	USA
MARQUILLO	Marquis / Uimillo	USA
PROBUS mutant	Trubillo / Platahof (lignée donneuse de la stérilité mâle)	France

Les génotypes suivis d'une * ne sont plus disponibles ou ont rencontré des problèmes de génotypage.

2.1.2 Choix d'un sous-échantillon de lignées

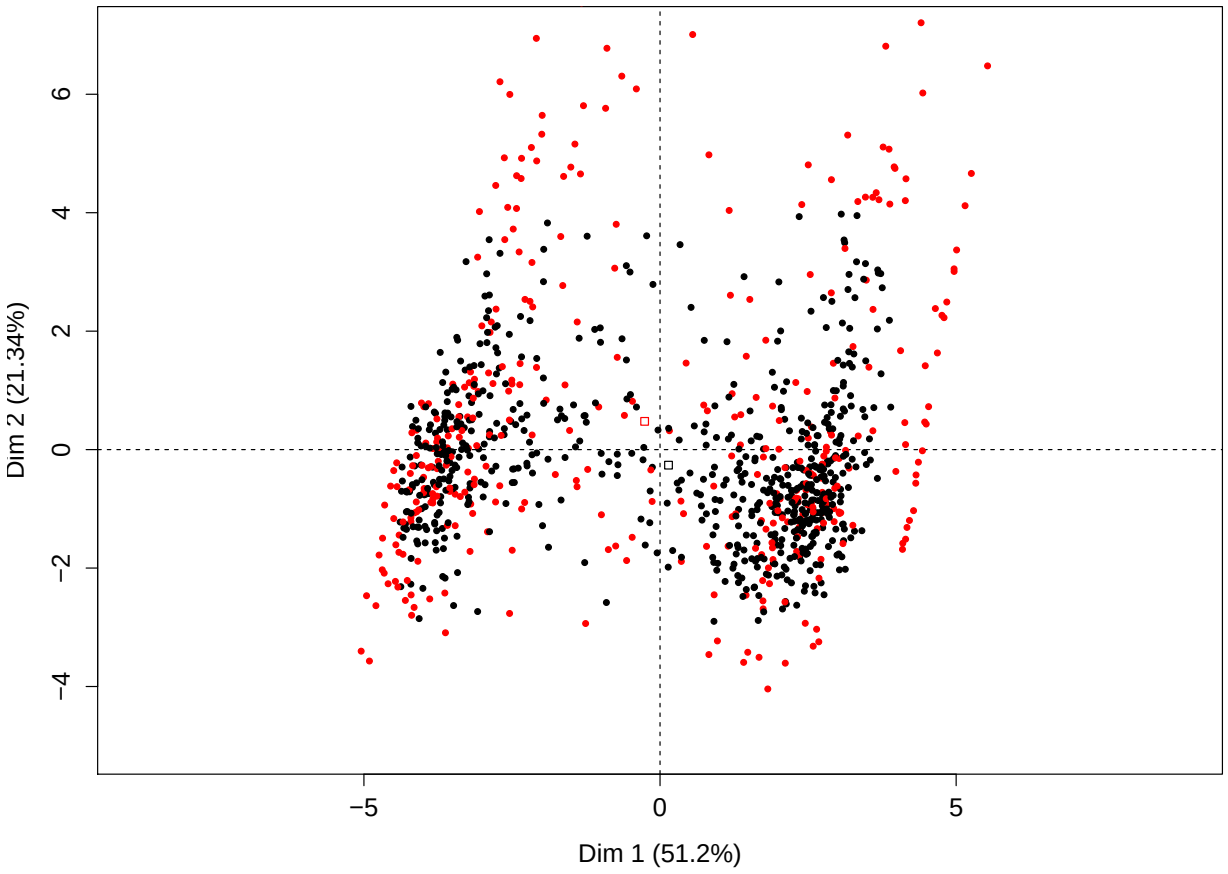
Dans le cadre de ce travail, nous avons développé un phénotypage ciblé et intensif sur la précocité de floraison, ainsi qu'un génotypage couvrant l'ensemble du génome. L'importance du travail engendré par le phénotypage et le coût des nouvelles technologies de génotypage haut débit nous ont amenés à travailler plus finement sur un sous-échantillon de lignées SSD. Ce sous échantillon a été constitué pour être représentatif de la diversité phénotypique de la population, sur la base des données de phénotypage de la première année (2010-2011). Une analyse en composantes principales (ACP) a été effectuée sur l'ensemble des notations des stades de développement. Les individus extrêmes ont été sélectionnés. Les autres individus ont été choisis par tirage aléatoire (1 sur 6, réalisé sur la base des coordonnées sur l'axe 1 de l'ACP). Ces individus couvrent donc l'intégralité de la gamme de variation phénotypique. Lors de cette sélection, nous avons écarté les génotypes peu fixés sur le plan phénotypique, présentant une forte proportion de mâle-stériles (plus de 20% de plantes mâles stériles), ou ceux pour lesquels nous ne disposions pas d'assez de semences. *In fine*, le sous échantillon MAGIC-INRA est composé de 380 lignées SSD (Figure 2.3).

2.2 Dispositif expérimental pour la caractérisation phénotypique du rythme de développement

2.2.1 Détail des dispositifs

La précocité de floraison et différents stades de développement ont été évalués en pépinière en plein champ sur le site du Moulon (48,4 °N, 21 °E) durant deux années consécutives (2010-2011 & 2011-2012). Les 1 026 lignées SSD, 56 parents ainsi que 24 témoins (variétés modernes actuelles) (Annexe B) ont été semés en pépinière, sur des lignes de 1,20m : un génotype par ligne et 20 grains par génotype, avec un inter-rang de 18cm. Trois dates de semis ont été réalisées afin d'étudier plus précisément l'effet de la vernalisation : un semis d'automne (novembre), et deux semis de printemps (mars et avril). Deux répétitions ont été réalisées pour les deux premières dates de semis et une seule pour celle d'avril. La première année, les 1 026 lignées ont été répétées en deux blocs complets randomisés, alors qu'en seconde année, seul le sous échantillon de 380 lignées SSD a été observé avec deux répétitions. En 2011-2012, le sous échantillon a aussi été étudié en condition de jour long (16h de jour). Cette expérimentation a été réalisée en extérieur, grâce à un éclairage artificiel faible intensité dès la levée (guirlande d'ampoules de 40W espacées de 1,20m à une hauteur de 1,50m [Ratet, 2011] ; Annexe C). L'éclairage faible intensité permet de rallonger la longueur du jour sans activer la photosynthèse afin de pouvoir comparer les plantes cultivées en plein champ avec une lumière naturelle et celles cultivées avec des jours longs artificiels [González et al., 2003]. En complément des pépinières, des essais en micro-parcelle ont été réalisés : en 2010-2011 à Cambridge en Angleterre par l'équipe de Ian Mackay

FIGURE 2.3 – Répartition des 1 026 lignées SSD selon les deux premiers axes de l'ACP avec en rouge les 380 individus du sous-échantillon.



(NIAB) avec l'ensemble des génotypes, sans répétitions, et en 2011-2012 au Moulon avec le sous échantillon dans un dispositif répété en deux blocs complets randomisés.

2.2.2 Méthodes de notation

Dans chaque condition de pépinière, une cinétique de développement à cinq points a été renseignée. Les dates des cinq stades suivant ont été notées : dernière feuille ligulée (Z39), épiaison (Z55), floraison (Z65), fin du stade laiteux (Z77) et maturité physiologique (Z87) (Figure 2.4). Ces notations ont été effectuées par observation de l'état moyen de chaque ligne de pépinière, avec généralement trois observations par semaine (lundi-mercredi-vendredi). Une mesure de l'étalement des stades à l'intérieur des lignes a été établie grâce à un suivi du début, milieu, et fin de stade. Les stades de maturité du grain (Z77 et Z87) ont été déterminés en fonction de l'aspect visuel de la plante : l'équipe de Ian Mackay a développé une échelle permettant de déterminer ces stades en fonction de l'état de sénescence de la feuille drapeau et de la couleur de l'épi (Annexe D). Pour les essais agronomiques

réalisés en Angleterre, seuls les stades de floraison (Z65) et de fin du remplissage du grain (Z87) ont été notés. Pour ceux réalisés en France, à l'exception du premier stade (dernière feuille ligulée : Z39), toute la cinétique a été phénotypée.

Toutes les données disponibles sont synthétisées dans le tableau 2.2. Les protocoles de notations ont été répertoriés dans l'annexe E. Ce travail de phénotypage a été possible grâce à l'implication de plusieurs notateurs. Afin de rassembler les notations des différentes personnes, nous avons développé un script de fusion de fichiers de notations (Annexe G). Un didacticiel d'apprentissage de notation des différents stades a aussi été réalisé afin de limiter l'effet notateur (Annexe F).

FIGURE 2.4 – Photo des différents stades de développement phénotypés.

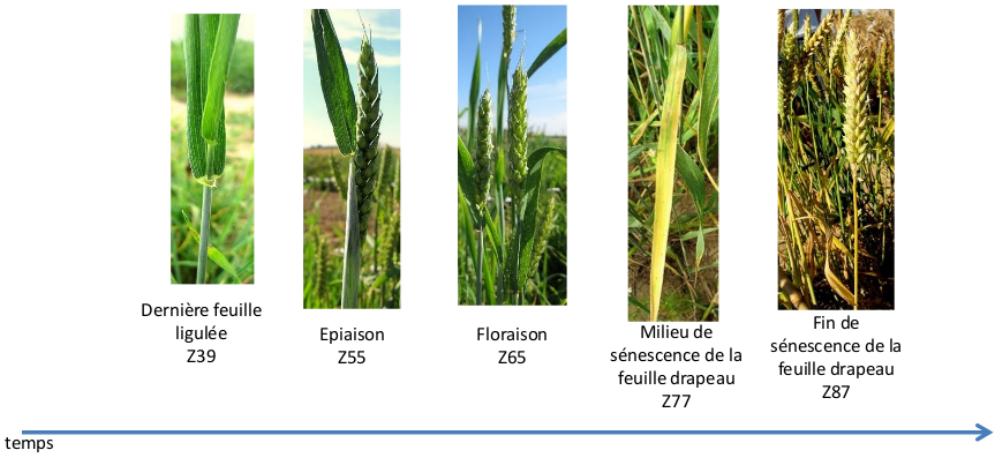


Tableau 2.2 – Description des données phénotypiques.

Année	Semis	Lieu	Dispositif	Pop	Nb répétition	Z39	Z55	Z65	Z77	Z87
2010-2011	novembre	Moulon	Pépinière	T	2	x	x	x	x	x
		Cambridge (UK)	Parcelles	T	1			x		x
	mars	Moulon	Pépinière	T	2	x	x	x	x	x
	avril	Moulon	Pépinière	T	1	x	x	x	x	x
2011-2012	novembre	Moulon	Pépinière	T	1	x	x	x	x	x
		Moulon	Pépinière	Ech	2	x	x	x	x	x
		Moulon + Jour Long	Pépinière	Ech	1	x	x	x	x	x
		Moulon	Parcelles	Ech	2		x	x	x	
	mars	Moulon	Pépinière	T	1	x	x	x	x	x
		Moulon	Pépinière	Ech	2	x	x	x	x	x
	avril	Moulon	Pépinière	T	1	x	x	x	x	x

T : représente les 1 026 lignées SSD + 56 lignées parentales
Ech : représente les 380 lignées SSD du sous-échantillon

2.3 Marquage génétique de la population

Les ADN des 1 026 lignées SSD et des 56 lignées parentales ont été extraits à partir de 500mg de feuille fraîche prélevée au champ en 2010-2011 dans la pépinière (plantes S4 autofécondées pour les SSD). Le protocole d'extraction est une adaptation de Dellaporta et al. [1983], qui inclut une précipitation des carbohydrates [Michaels et al., 1994].

2.3.1 Génotypage avec la puce 9K iSelect

Le sous-échantillon de 380 lignées SSD ainsi que les 56 parents ont été génotypés avec la puce SNP 9K iSelect [Cavanagh et al., 2013]. Ces 9 000 SNPs répartis sur tout le génome sont issus :

- de l'analyse transcriptomique de 26 accessions
- de la capture de séquences de 3 500 gènes polymorphes dans le croisement Synthétic x Opata
- d'un set de 655 SNPs découverts dans des variétés de pays.

Le génotypage a été réalisé par l'équipe de MJ Hayden au DPI Victoria à Bundoora (Australie) avec le BeadStation et l'iScan (technologie Illumina). A partir du signal de la puce 9K iSelect, l'assignation des allèles a été réalisée avec le logiciel GenomeStudio (`http://www.illumina.com/software/genomestudio_software.ilmn`). Les assignations ont ensuite été vérifiées visuellement sur l'ensemble des SNPs, et corrigées manuellement si nécessaire. Seuls les SNPs produisant une assignation non ambiguë et une ségrégation biallélique claire ont été retenus.

2.3.2 Génotypage avec la technologie KASPAR

Un set de 93 marqueurs a été génotypé sur l'ensemble des 1 082 individus (1 026 lignées SSD + 56 parents). Ce set est composé de 38 marqueurs neutres et de 54 marqueurs localisés dans des gènes candidats associés à la précocité de floraison.

2.3.2.1 Marqueurs localisés dans des gènes candidats

Parmi les marqueurs localisés dans les gènes candidats, 20 polymorphismes sont issus de la bibliographie ou de communications personnelles (Tableau 2.3).

Pour compléter, j'ai recherché les séquences orthologues aux gènes liés à la précocité de floraison chez les graminées tempérées, publiés dans Higgins et al. [2010], dans la base de données graingene (`http://wheat.pw.usda.gov/GG2/index.shtml`), ainsi que dans GeneBank (`http://www.ncbi.nlm.nih.gov.gate1.inist.fr/genbank/`) avec au total 74 séquences identifiées chez *T. aestivum* ou chez *H. vulgare*. Ces séquences ont été comparées par Blast [Altschul et al., 1990] aux séquences flanquantes des SNPs présents sur la puce 9K iSelect et avec les séquences de SNP publiées dans Chao et al. [2010]. Seuls les polymorphismes qui s'alignaient

avec une e-value inférieure à 10^{-3} et un pourcentage d'identité supérieure à 99% ont été retenus. Les SNPs non spécifiques d'une unique séquence candidate ont été éliminés. Au total 54 polymorphismes ont été génotypés dont 34 identifiés par Blast (Tableau 2.4).

Tableau 2.3 – Liste des marqueurs génotypés par la méthode KASPAR sur les 1 026 individus, déjà identifiés comme étant localisés dans des gènes candidats.

Nom du marqueur	Nom du gène candidat dans lequel il se localise	Polymorphisme	Référence bibliographique
DEAP_moulon_COAB_1	*CO-B*	SNP	Rhoné 2008
DEAP_moulon_FTA_2	*FTA*	SSR	Bonnin et al. 2008
DEAP_moulon_FTD_3	*FTD*	1bp indel	Bonnin et al. 2008
DEAP_moulon_LDDB_4	*LDDB*	SNP	Rhoné 2008
DEAP_moulon_Ppd-D1ex8_5	*Ppd-D1*	16bp deletion	Beales et al. 2007
DEAP_moulon_Ppd-D1prom_6	*Ppd-D1*	2kb indel	Beales et al. 2007
DEAP_moulon_Vrn-1Din1_7	*Vrn-1D*	4 kb indel	Fu et al. 2005
DEAP_moulon_Vrn1Apr-1/2_8∗	*Vrn1A*	indel	Yan et al. 2004a
DEAP_moulon_Vrn1Apr-2/3_9∗	*Vrn1A*	SNP	Yan et al. 2004a
DEAP_moulon_Vrn1Apr45_10	*Vrn1A*	4bp indel	Rhoné et al. 2008
DEAP_moulon_Vrn1Bin1_11	*Vrn1B*	4 kb indel	Fu et al. 2005
DEAP_moulon_PpdA1GS100_12	*Ppd-A1*	1027bp indel	Bentley et al. 2011
DEAP_moulon_PpdA1GS105_13	*Ppd-A1*	1117bp indel	Bentley et al. 2011
DEAP_moulon_Ppd-A1-CaDe_14	*Ppd-A1*	305bp indel	(S. Griffiths, pers. comm.)
DEAP_moulon_TaGW2_15	*TaGW2*	SNP	Su et al. 2011
DEAP_moulon_RHT-B1_16	*RHT-B1*	SNP	(S. Griffiths, pers. comm.)
DEAP_moulon_RHT-D1_17	*RHT-D1*	SNP	(S. Griffiths, pers. comm.)
DEAP_moulon_Ppd-B1-SNP_CT_18	*Ppd-B1*	SNP	(S. Griffiths, pers. comm.)
DEAP_moulon_Ppd-B1-SNP_GC_19	*Ppd-B1*	SNP	(S. Griffiths, pers. comm.)
DEAP_moulon_Vrn1A-ex8_20	*Vrn1A*	SNP	Sherman et al. 2004

Les marqueurs suivis d'une ∗ ont rencontré des problèmes techniques de génotypage.

2.3.2.2 Marqueurs neutres

Les marqueurs neutres (Tableau 2.5) ont été choisis parmi les polymorphismes de la puce 9K iSelect non liés à la précocité de floraison. Un échantillon aléatoire de 38 marqueurs non liés (DL<0.6) et présentant un indice de diversité (*He*) de 0,5 sur les lignées parentales a été sélectionné.

Tableau 2.4 – Liste des marqueurs génotypés avec la méthode KASPAR sur les 1 026 individus, identifiés par comparaison (Blast) comme homologues à une liste de gènes candidats trouvés dans les bases de données.

Nom du marqueur	Nom du gène candidat dans lequel il se localise	Nom du marqueur dans la puce 9K iSelect
DEAP_moulon_CA_7180_21	*LDDA*	wsnp_Ku_c5623_9966516
DEAP_moulon_CA_4699_22	*Vrn1B*	wsnp_Ex_c7546_12900094
DEAP_moulon_CA_7108_23	*CO1*	wsnp_Ku_c48167_54427241
DEAP_moulon_CA_6412_24	*LDDA*	wsnp_Ku_c1102_2211433
DEAP_moulon_CA_3677_25	*VIL2*	wsnp_Ex_c39304_46635517
DEAP_moulon_CA_5042_26	*VIL2*	wsnp_Ex_rep_c102044_87296690
DEAP_moulon_CA_923_27	*CO4*	wsnp_CAP12_c1461_744121
DEAP_moulon_CA_6905_28	*FTB*	wsnp_Ku_c3201_5970486
DEAP_moulon_CA_4805_29	*FT*	wsnp_Ex_c8424_14192191
DEAP_moulon_CA_5860_30	*CO1*	wsnp_JD_c15333_14824351
DEAP_moulon_CA_7896_31	*ZTL*	wsnp_Ra_c3766_6947263
DEAP_moulon_CA_4509_32	*Vrn1B*	wsnp_Ex_c645_1273901
DEAP_moulon_CA_44_33	*SMZ*	wsnp_BE403956B_Ta_2_3
DEAP_moulon_CA_4049_34	*CO3*	wsnp_Ex_c4921_8764088
DEAP_moulon_CA_7643_35	*SOC1*	wsnp_Ra_c16053_24607526
DEAP_moulon_CA_5440_36	*CO4*	wsnp_Ex_rep_c67690_66354931
DEAP_moulon_CA_5396_37	*TaGI3*	wsnp_Ex_rep_c67404_65986980
DEAP_moulon_CA_5656_38	*Vrn1B*	wsnp_Ex_rep_c69901_68864080
DEAP_moulon_CA_3583_39∗	*Vrn1B*	wsnp_Ex_c3670_6694480
DEAP_moulon_CA_6574_40	*TaHd1A*	wsnp_Ku_c15816_24541712
DEAP_moulon_CA_2045_41	*PHYA*	wsnp_Ex_c1563_2987002
DEAP_moulon_CA_5269_42	*PHYA*	wsnp_Ex_rep_c66600_64897324
DEAP_moulon_CA_2307_43	*ZTL*	wsnp_Ex_c18382_27210656
DEAP_moulon_CA_8172_44∗	*VIN3*	wsnp_Ra_rep_c70756_68675384
DEAP_moulon_CA_750_45	*SMZ*	wsnp_CAP11_c3346_1639010
DEAP_moulon_CA_7895_46	*ZTL*	wsnp_Ra_c3766_6947230
DEAP_moulon_CA_CH_47	*ZTL*	wsnp_Ra_c3766_6947953
DEAP_moulon_CA_CH_48∗	*SMZ*	wsnp_BE490226A_Ta_2_1
DEAP_moulon_CA_4974_49	*CDF1*	wsnp_Ex_c9872_16271161
DEAP_moulon_CA_4916_50	*CO1*	wsnp_Ex_c9440_15657149
DEAP_moulon_CA_2528_51∗	*SMZ*	wsnp_Ex_c20353_29417160
DEAP_moulon_CA_4872_52	*SMZ*	wsnp_Ex_c9063_15093396
DEAP_moulon_CA_2164_53	*Vrn1D*	wsnp_Ex_c16720_25268525
DEAP_moulon_CA_3692_54	*TOE2*	wsnp_Ex_c3977_7201781

Les marqueurs suivis d'une ∗ ont rencontré des problèmes techniques de génotypage.

Tableau 2.5 – Liste des marqueurs neutres, génotypés avec la méthode KASPAR sur les 1 026 individus avec leur nom de référence dans la puce 9K iSelect.

Nom du marqueur	Nom du marqueur dans la puce 9K iSelect
DEAP_moulon_NE_143_1	wsnp_BE443995B_Ta_2_2
DEAP_moulon_NE_182_2	wsnp_BE445506B_Ta_2_4
DEAP_moulon_NE_210_3	wsnp_BE489326B_Ta_2_1
DEAP_moulon_NE_513_4	wsnp_BF484606A_Ta_2_3
DEAP_moulon_NE_578_5	wsnp_BG606986A_Ta_2_4
DEAP_moulon_NE_605_7	wsnp_BM140362A_Ta_2_2
DEAP_moulon_NE_618_8	wsnp_BQ161779B_Ta_2_4
DEAP_moulon_NE_1460_15	wsnp_Ex_c11265_18216936
DEAP_moulon_NE_2328_17	wsnp_Ex_c18616_27481826
DEAP_moulon_NE_2353_18	wsnp_Ex_c18800_27681277
DEAP_moulon_NE_2366_19	wsnp_Ex_c18965_27868480
DEAP_moulon_NE_2821_20	wsnp_Ex_c23618_32855041
DEAP_moulon_NE_3630_24	wsnp_Ex_c38105_45710671
DEAP_moulon_NE_4113_26	wsnp_Ex_c5185_9189184
DEAP_moulon_NE_4816_27	wsnp_Ex_c8588_14419007
DEAP_moulon_NE_4929_28	wsnp_Ex_c9502_15748469
DEAP_moulon_NE_4961_29	wsnp_Ex_c9763_16125630
DEAP_moulon_NE_5083_30	wsnp_Ex_rep_c103087_88123733
DEAP_moulon_NE_6485_37	wsnp_Ku_c13204_21105694
DEAP_moulon_NE_6902_38	wsnp_Ku_c3151_5892200
DEAP_moulon_NE_6919_40	wsnp_Ku_c33335_42844594
DEAP_moulon_NE_7005_41	wsnp_Ku_c3929_7189422
DEAP_moulon_NE_7177_42	wsnp_Ku_c55961_59662821
DEAP_moulon_NE_7533_44	wsnp_Ra_c1020_2062200
DEAP_moulon_NE_7547_45	wsnp_Ra_c107797_91270622
DEAP_moulon_NE_987_48	wsnp_CAP12_c7952_3403722
DEAP_moulon_NE_1644_49	wsnp_Ex_c1255_2411550
DEAP_moulon_NE_4662_51	wsnp_Ex_c7362_12622736
DEAP_moulon_NE_5666_52	wsnp_Ex_rep_c70036_68988728
DEAP_moulon_NE_4465_58	wsnp_Ex_c62701_62229607
DEAP_moulon_NE_5912_59	wsnp_JD_c20555_18262260
DEAP_moulon_NE_5228_60	wsnp_Ex_rep_c66389_64588992
DEAP_moulon_NE_7135_62	wsnp_Ku_c51039_56457361
DEAP_moulon_NE_7471_66	wsnp_Ku_rep_c70220_69775367
DEAP_moulon_NE_7507_69	wsnp_Ku_rep_c72211_71920520
DEAP_moulon_NE_8377_75	wsnp_RFL_Contig2729_2446041
DEAP_moulon_NE_5071_86	wsnp_Ex_rep_c102707_87814407
DEAP_moulon_NE_7519_89	wsnp_Ku_rep_c73198_72796386

Chapitre 3

Développement d'un algorithme de cartographie à partir de données de déséquilibre de liaison

3.1 Introduction

3.1.1 Détection de QTL

Si la génétique s'est construite dans ses débuts grâce à l'analyse de la ségrégation de gènes ayant un effet majeur sur le phénotype (couleur, texture, pilosité), l'essentiel des caractères d'intérêt adaptatif, agronomique ou médical présentent une variation quantitative due à un déterminisme polygénique. L'enjeu est donc d'identifier les différents loci impliqués dans cette variation phénotypique. La détection de ces QTLs (Quantitative Trait Locus) s'effectue généralement grâce à l'association statistique entre la variation du caractère quantitatif et le polymorphisme d'un marqueur moléculaire dans la descendance de parents de génotypes connus [Sax, 1923]. Dans ces approches, disposer d'une cartographie des marqueurs sur le génome s'avère stratégique, car la carte génétique permet de localisation des QTLs (cartographie par intervalle) [Rebai et al., 1995], et le développement d'outils de génotypage plus performants (set optimisé de marqueurs). En effet les méthodes comme la cartographie par intervalle, permettent de localiser plus précisément les QTLs et de mieux estimer leurs effets [Broman, 2001].

3.1.2 Principe de la cartographie génétique

La cartographie génétique ordonne les marqueurs polymorphes d'une population donnée, à partir des données de génotypage des parents et des descendants ainsi que du plan de croisement, ou encore de la généalogie. Elle étudie la co-ségrégation de marqueurs pendant la phase de recombinaison à la méiose et les cartographie en fonction de cette corrélation. La dépendance entre deux marqueurs est quantifiée par le logarithme décimal du ratio entre la probabilité que les deux marqueurs soient liés et la probabilité qu'ils soient indépendants (LOD score). Les marqueurs ségrégeant de manière indépendante (LOD score faible) seront assignés à des groupes de liaison différents. Quand la densité et le marquage sont suffisants, le nombre de groupes de liaison doit correspondre au nombre de chromosomes. Une fois les marqueurs assignés à un groupe de liaison, leur ordre est établi à partir du taux de recombinaison. Le taux de recombinaison est ensuite converti en distance génétique, exprimée

en centimorgan (cM). La distance génétique peut être calculée de deux façons différentes avec les équations de Haldane [Haldane, 1919] et de Kosambi [Kosambi, 1943] suivant la prise en compte ou non de l'interférence entre les recombinaisons.

La construction des cartes génétiques se fait classiquement à l'aide de familles en ségrégation de type populations F2, populations rétro-croisées, lignées haploïdes doublées (HD) ou lignées recombinantes (RILs), issues de croisements bi-parentaux ou plus récemment de croisements multi-parentaux avec quatre ou huit parents croisés de manière pyramidale (MAGIC) [Cavanagh et al., 2013]. Pour les populations plus complexes telles que les populations naturelles, les chercheurs ont développé une méthode de recherche de QTL qui ne nécessite pas de carte génétique : c'est le principe de la génétique d'association. Elle recherche l'association statistique entre un polymorphisme à un marqueur ou un gène et la variation d'un caractère phénotypique dans une population. Cette méthode fait l'hypothèse que le déséquilibre de liaison décroît avec la distance génétique entre un marqueur et le QTL. Toutefois même si la méthode ne nécessite pas une carte génétique, la localisation du marqueur associé est importante pour l'interprétation.

3.1.3 Déséquilibre de liaison et lien avec les distances génétiques

Le déséquilibre de liaison est l'association non aléatoire d'allèles appartenant à deux loci différents [Bennett, 1954]. Il s'exprime de différentes façons prenant plus ou moins en compte les fréquences alléliques des marqueurs comme D, D' et r^2 [Hill and Robertson, 1968] avec :

$$D = p_{AB} - p_A p_B \tag{3.1}$$

$$D' = \frac{D}{D_{max}} \; avec \; D_{max} = \begin{cases} min(p_A p_b ; p_a p_B) \; si \; D > 0 \\ min(p_a p_b ; p_A p_B) \; si \; D < 0 \end{cases} \tag{3.2}$$

$$r^2 = \frac{D}{p_a p_A p_b p_B} \tag{3.3}$$

où p_a, p_A, p_b et p_B sont respectivement les fréquences alléliques des allèles a, A, b et B et p_{AB} la fréquence du génotype [AB].

D'après Hill and Weir [1994], le r^2 est l'estimateur le plus utilisé pour exprimer le DL ; il est moins sensible à la taille de l'échantillon et donc moins « bruité » que le D'. De plus il est plus robuste aux erreurs de génotypage [Akey et al., 2001]. Compris entre 0 et 1, r^2 représente la corrélation entre les allèles à deux loci différents. Il prend en compte les fréquences alléliques, ainsi le r^2 est égal à 1 uniquement si les deux marqueurs ont le même nombre d'allèles, les mêmes fréquences alléliques et s'ils sont strictement associés.

Pour comprendre comment le DL entre marqueurs peut nous renseigner sur leur cartographie génétique, il est important de s'intéresser aux propriétés du DL dans une population théorique. La décroissance du DL le long du chromosome a été modélisée par Hill and Weir [1988] en fonction de la taille efficace d'une population panmictique, du taux de recombinaison entre marqueurs et d'une

taille d'échantillonnage donnée (Equation 3.4) sous les hypothèses d'un équilibre dérive-recombinaison et d'un taux de mutation négligeable. Dans le cas de marqueurs bi-alléliques et sans mutation, une équation plus simple du DL attendu est obtenue (Equation 3.5) [Sved, 1971; Hill and Weir, 1994].

$$r^2 = \frac{10 + C}{(2 + C)(11 + C)} \times \frac{1 + (3 + C)(12 + 12C + C^2)}{n(2 + C)(11 + C)} \tag{3.4}$$

avec $C = 4Nc$ le paramètre de recombinaison de la population, r^2 le DL, N la taille efficace de la population, c le taux de recombinaison et n la taille de la population échantillonnée [Hill and Weir, 1988].

$$E(r^2) = \frac{1}{1 + C} \tag{3.5}$$

En effet, le DL est initialement créé par les mutations, et par la dérive, qui de façon stochastique peut également créer ou renforcer certaines associations d'allèles, sachant que plus la taille efficace de la population est grande et plus l'effet de la dérive sera limité. Une fois le DL créé, il décroit au cours des générations sous l'effet de la recombinaison. Cette décroissance est d'autant plus rapide que la paire de loci considérée est génétiquement distante (Figure 3.1, Equation 3.6).
En espérance dans une population infinie, on attend :

$$D_f = D_{ini} \times (1 - c)^g \tag{3.6}$$

avec D_f et D_{ini} respectivement le DL final et initial (exprimé en r^2, D ou D'), c le taux de recombinaison et g le nombre de générations de panmixie.

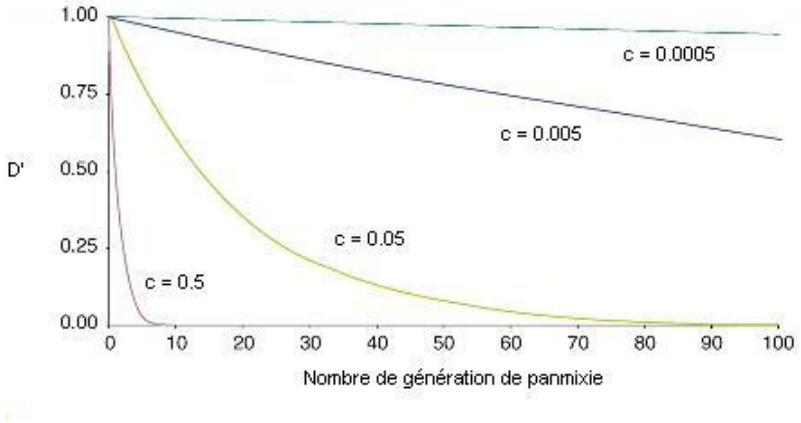


FIGURE 3.1 – Décroissance du DL (D') en fonction du nombre de générations de recombinaison pour des loci avec quatre taux de recombinaison différents (c) : de 0,5 pour les marqueurs non liés jusqu'à 0,0005 pour les marqueurs très liés. Le DL a été calculé à partir de l'équation 3.6 avec un DL initial de 1. [Mackay and Powell, 2007]

3.1.4 La cartographie du DL de Morton

Morton a publié plusieurs papiers concernant la relation entre distance physique et déséquilibre de liaison. Cette carte du DL est le DL entre deux loci (D) [Morton et al., 2001]. Cette association allélique est modélisée à partir du taux de recombinaison, de la taille efficace de la population, des pressions de mutation et de migration [Morton et al., 2001]. Elle suit une décroissance exponentielle en fonction de la distance (Equation 3.7) [Maniatis et al., 2002; Zhang et al., 2002]. Construite à partir du DL entre haplotypes, la carte du DL est exprimée en LDU (LD Unit), avec les marqueurs en DL complet localisés à la même position. Les différentes positions en LDU sont cohérentes avec les points chauds de recombinaisons (Figure 3.2).

$$\rho = (1 - L)Me^{-\varepsilon d} + L \tag{3.7}$$

avec L l'association allélique entre les marqueurs indépendants et d la distance physique ou génétique.

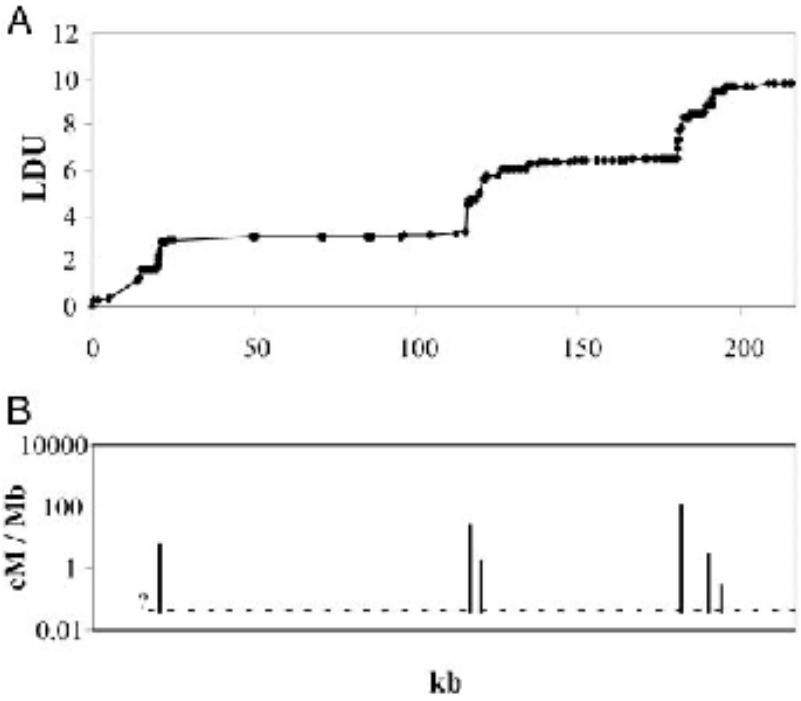


FIGURE 3.2 – Représentation (A) de la carte du DL et (B) des points chauds de recombinaison en fonction de la carte physique d'un segment de 216kb du chromosome 6 (6p21.3) de l'Homme [Zhang et al., 2002].

La méthode de Morton présentée précédemment analyse le DL le long du génome. Pour le moment, aucune méthode de cartographie n'a été publiée pour les

populations pour lesquels le taux de recombinaison ne peut être calculé comme les populations naturelles ou les populations expérimentales avec un plan de croisement complexe et des inter-croisements non contrôlés. A partir de ce constat, nous avons voulu tester si le développement d'un algorithme basé sur le DL pourrait permettre la création de carte génétique ou l'amélioration de cartes déjà publiées.

Sachant que le DL décroît, en espérance, de manière monotone avec la distance génétique, nous avons cherché une méthode qui permettrait d'ordonner les marqueurs de telle façon que la matrice de DL soit "peignée" avec les forts DL proches de la diagonale et les faibles DL à l'extérieur (Figure 3.3).

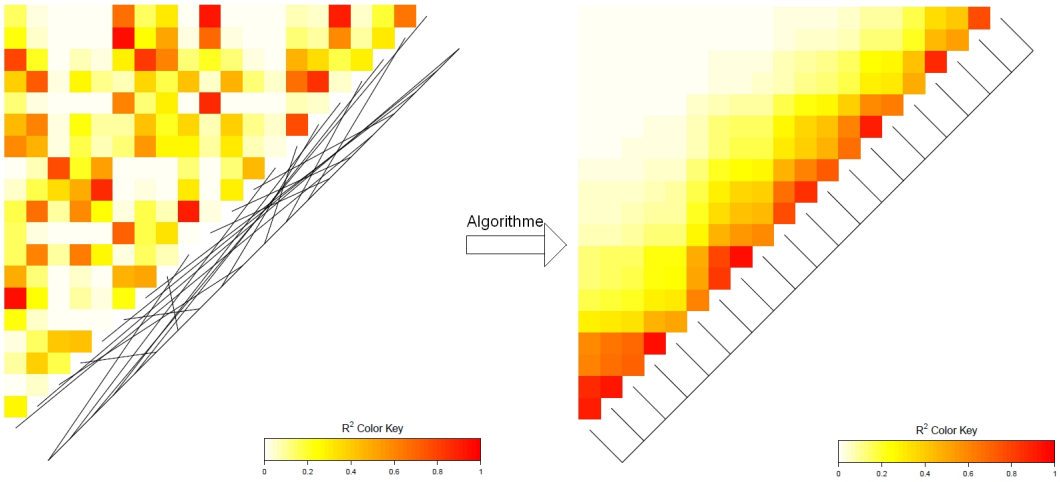


FIGURE 3.3 – Illustration d'une manière d'ordonner des marqueurs à partir du DL. L'abscisse et l'ordonnée représentent des marqueurs ordonnés dans le même ordre. Chaque pixel représentant le DL entre les deux marqueurs, avec à gauche, les données brutes et à droite, le résultat théorique de l'algorithme d'ordonnancement.

Dans cette étude, après avoir vérifié qu'une méthode comme la classification hiérarchique n'ordonnait pas les marqueurs de manière à avoir un DL décroissant avec la distance, nous avons développé notre propre algorithme. Cet algorithme ordonne les marqueurs sur la base du DL avec une méthode inspirée de la celle utilisée dans MapMaker. L'algorithme a été testé dans un premier temps avec quatre populations différentes sur des fenêtres de marqueurs puis avec deux populations MAGIC de blé à l'échelle du chromosome.

3.2 Matériels et Méthodes

3.2.1 Méthodes

3.2.1.1 Classification hiérarchique

Des méthodes comme la classification hiérarchique pourraient être adaptées à l'ordonnancement de marqueurs. La fonction "hclust" du logiciel R qui réalise une classification hiérarchique des marqueurs à partir de la matrice de dissimilarité des données [Maechler et al., 2013] a été testée. Sachant qu'au départ, chaque marqueur forme une classe, le but de l'algorithme est de regrouper les classes qui ont la dissimilarité la plus faible, jusqu'à rassembler toutes les classes. Cette méthode ordonnerait donc les marqueurs en fonction de leur similarité. Elle a été testée sur les données de DL mais aussi directement sur les données de génotypage.

3.2.1.2 Principe du logiciel MapMaker

Cheema and Dicks [2009] ont passé en revue de nombreuses méthodes de cartographie ainsi que les logiciels de cartographie dans lesquels ils sont implémentés. Seule celle incluse dans le logiciel MapMaker sera présentée ici [Lander et al., 1987]. MapMaker est l'un des premiers logiciels de cartographie génétique, qui s'est imposé comme référence lors de l'explosion des marqueurs moléculaires et de l'établissement des cartes génétiques, dans les années 90. Il utilise un algorithme qui ordonne, sur la base du taux de recombinaison, de façon itérative tous les marqueurs à partir de trois fonctions principales. La fonction COMPARE estime le maximum de vraisemblance parmi tous les ordres possibles au sein d'un sous ensemble de marqueurs. La fonction TRY place un marqueur non cartographié dans une carte provisoire, en comparant les vraisemblances des différentes insertions possibles entre les marqueurs déjà ordonnés. La fonction RIFFLE valide un ordre final par permutation des marqueurs sur une fenêtre glissante. Comme le nombre de permutations à tester augmente de manière exponentielle avec la taille de la fenêtre glissante, cette fenêtre est réduite à trois marqueurs (méthode du test trois points).

3.2.1.3 Description de l'algorithme de cartographie basée sur le DL

L'algorithme a été développé sur le principe illustré dans la Figure 3.3. Notons cependant que ce principe correspond au cas idéal d'une population de large taille efficace et ne subissant aucune pression de sélection, ou distorsion de ségrégation, pouvant perturber cette structure monotone du DL. Le but est d'ordonner les marqueurs pour maximiser les DL dans la diagonale de la matrice de DL.

L'algorithme travaille directement sur la matrice de DL. Il a été inspiré de la méthode utilisée dans Ganal et al. [2011]. Un couple de marqueurs liés est aléatoirement tiré parmi les couples de marqueurs avec un DL supérieur à 0,8. Ce couple servira de point de départ. Tous les marqueurs liés aux marqueurs de départ avec un DL supérieur à un seuil sont testés aux positions autour du marqueur auquel ils sont le plus liés (fonction TRY). Pour chaque position, des indices de qualité sont

calculés (paragraphe 3.2.1.4). Le meilleur ordre est celui qui minimise ces indices de qualité. Une fois que la meilleure position de chaque marqueur est trouvée, un test de robustesse (fonction COMPARE) est réalisé par permutation des marqueurs dans une fenêtre d'au maximum cinq marqueurs centrée sur le marqueur testé. Si le meilleur ordre issu des permutations est l'ordre testé, cet ordre est gardé. L'ordre avec le meilleur indice de qualité parmi les ordres robustes sera gardé et considéré comme point de départ pour l'intégration d'un nouveau marqueur. Cette boucle est réalisée tant qu'un marqueur s'intègre. Quand plus aucune intégration de marqueur ne donne d'ordre robuste, l'ensemble des marqueurs sélectionnés définit un "ordre robuste". Un autre ordre, nommé "souple" est ensuite établi sur la base de cet ordre robuste en intégrant tous les marqueurs avec un DL supérieur au seuil avec les marqueurs de l'ordre robuste à leur meilleure position (fonction TRY mais pas de COMPARE). Quand tous ces marqueurs ont été positionnés, l'ordre souple est défini et la chaîne s'arrête. Une nouvelle chaîne démarre avec un nouveau couple de marqueurs liés avec un DL supérieur à 0,8 parmi les marqueurs non ordonnés. Cette nouvelle chaîne représentera un autre groupe de liaison. L'algorithme s'arrête quand parmi les marqueurs restant à ordonner, aucun couple n'est lié avec un DL supérieur à 0,8. Le principe de l'algorithme est récapitulé de manière schématique dans l'annexe H. Le seuil de DL pour déterminer les marqueurs à tester dans l'algorithme a été défini pour n'utiliser que les DL significatifs pour les différentes populations. Il a été fixé à 0,1. Les seuils de 0,8 et de 0,1 sont des seuils ajustables en fonction des populations.

3.2.1.4 Indices de qualité

L'algorithme s'appuie sur deux indices de qualité. Un indice, nommé "sj", est la somme des différences positives de DL. Les différences sont réalisées selon le rang décroissant des cellules (Figure 3.1) , *i.e.* de l'extérieur vers l'intérieur de la matrice de manière verticale et horizontale (Figure 3.4, Equation 3.8).

$$sj = \sum_{i=2}^{n-1} \sum_{j=i-1}^{n} \frac{(r_{ij}^2 - r_{i(j-1)}^2) + abs(r_{ij}^2 - r_{i(j-1)}^2)}{2} + \frac{(r_{ij}^2 - r_{(i-1)j}^2) + abs(r_{ij}^2 - r_{(i-1)j}^2)}{2} \tag{3.8}$$

l'utilisation de la valeur absolue permet de réduire à 0 les valeurs négatives et donc de sommer uniquement les différences positives.

Tableau 3.1 – Rang des cellules dans la matrice. Le rang est la position de la cellule par rapport à la diagonale.

	V1	V2	V3	V4	V5
V1	1	2	3	4	5
V2		1	2	3	4
V3			1	2	3
V4				1	2
V5					1

	V1	V2	V3	V4	V5
V1	1	0,96	0,92	0,8	0,72
V2		1	0,95	0,85	0,83
V3			1	0,9	0,87
V4				1	0,9
V5					1

	V1	V2	V3	V5	V4
V1	1	0,96	0,92	0.72	0,80
V2		1	0,95	0,83	0,85
V3			1	0,87	0,90
V5				1	0,9
V4					1

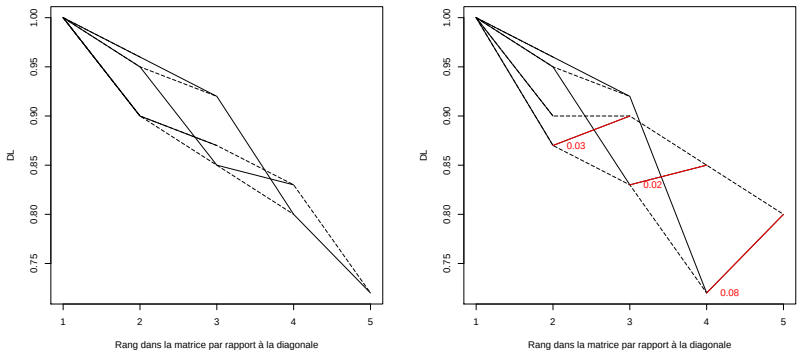


FIGURE 3.4 – Schémas explicatifs de l'indice de qualité "sj" minimisé par l'algorithme. Il est la somme des différences positives verticales et horizontales de l'extérieur vers l'intérieur de la matrice. A gauche, une matrice de DL représentant cinq marqueurs avec $sj = 0$. Tous les DLs diminuent lorsqu'on s'écarte de la diagonale. A droite, une matrice de DL avec $sj = 0,02 + 0,03 + 0,08 = 0,13$. L'inversement de deux marqueurs (V4 et V5) a entraîné une augmentation du DL lorsqu'on s'écarte de la diagonale dans trois cas. Les graphiques situés sous les matrices représentent l'évolution du DL en fonction du rang du DL (Tableau 3.1). Le rang 1 correspond au DL situé dans la diagonale. Le trait plein représente l'évolution horizontale sur la matrice et le trait pointillé l'évolution verticale.

Le deuxième indice à minimiser en cas d'égalité de "sj" est l'AIC [Akaike, 1974]. Il est calculé à partir du modèle linéaire de -log(DL) en fonction du rang de ce DL sur la matrice (Tableau 3.1).

3.2.2 Données

Dans le but de tester l'algorithme dans différents contextes, des données simulées et les données réelles de quatre populations différentes de complexité variable ont été utilisées : population bi-parentale, population MAGIC 8 parents, population MAGIC INRA en considérant d'une part le panel de 56 parents et d'autre part la population de lignées recombinantes. La méthode de classification hiérarchique a été testé sur un chromosome de la population bi-parentale. L'algorithme développé a ensuite été testé sur des fenêtres de marqueurs sur les données simulées ainsi que sur la population bi-parentale et la population MAGIC INRA (le panel de parents et la population de lignées recombinantes). Les populations MAGIC 8 parents de blé (MAGIC NIAB) et MAGIC INRA ont ensuite été utilisées pour tester l'algorithme à l'échelle d'un chromosome. La population MAGIC de drosophile a servi pour la discussion.

3.2.2.1 Jeux de données simulées

Un tirage aléatoire du taux de recombinaison de 50 marqueurs a permis de simuler une carte génétique. Les DL finaux, après 15 générations de panmixie, ont été estimés à partir de l'équation 3.6 avec un DL initial de 1. Ces données ont été bruitées par ajout d'un nombre aléatoire tiré dans une loi uniforme de paramètres [0 ;0,1] ou bien de [0 ;0,2]. La matrice a ensuite été randomisée. Ce processus a été répété 30 fois.

3.2.2.2 Population bi-parentale de maïs

Des données de génotypage d'une population F2 bi-parentale de maïs IBM (croisement entre B73 et MO17) de 302 individus ont été téléchargées depuis le site `http://maizegdb.org/ibm302scores.html`. La carte génétique des marqueurs génotypés était également disponible. La méthode de classification hiérarchique a été testée avec un chromosome pris au hasard (chromosome 9) de cette population. Les données génome-entier ont ensuite été utilisées pour tester l'algorithme sur des fenêtres de marqueurs.

3.2.2.3 Population MAGIC de Drosophile

Les données de génotypage d'une population de drosophile de 487 individus, de type MAGIC 8 parents (population A) avec 20 générations de recombinaison ont été téléchargées depuis le site `flyrils.org` [King et al., 2012]. Seuls les 50 premiers marqueurs sans données manquantes et localisés sur le chromosome 2L ont été utilisés à cause du temps de calcul de l'algorithme.

3.2.2.4 Population MAGIC INRA

Panel de lignées : parents MAGIC INRA Les 56 parents, génotypés avec la puce 9K iSelect [Cavanagh et al., 2013], de la population MAGIC

INRA ont été utilisés et considérés comme un panel de lignées. Ces données de génotypage génome-entier ont été utilisées pour tester l'algorithme sur des fenêtres de marqueurs.

Population MAGIC INRA La population MAGIC INRA est une population issue du croisement aléatoire durant 12 générations de 60 parents permis grâce à l'intégration d'un gène de stérilité mâle. 380 lignées SSD S4 ont été génotypées avec la puce 9K iSelect [Cavanagh et al., 2013]. Ces données de génotypage génome-entier ont été utilisées pour tester l'algorithme sur des fenêtres de marqueurs mais aussi à l'échelle d'un chromosome.

Carte génétique de la puce 9K iSelect Pour ces populations de blé (panel, MAGIC INRA), la carte de référence utilisée est celle de la puce 9K iSelect publiée dans Cavanagh et al. [2013]. Cette carte génétique, consensus de six cartes élaborées avec des populations bi-parentales et d'une population MAGIC 4 parents, cartographie 6 822 marqueurs de manière unique répartis sur 2 742 points génétiques dans 22 groupes de liaison (deux groupes de liaison pour le chromosome 5D). 2,5 marqueurs se trouvent donc en moyenne positionnés sur le même point génétique mais 73% des points génétiques sont spécifiques d'un marqueur, le reste positionnant entre 2 et 100 marqueurs (Figure 3.5).

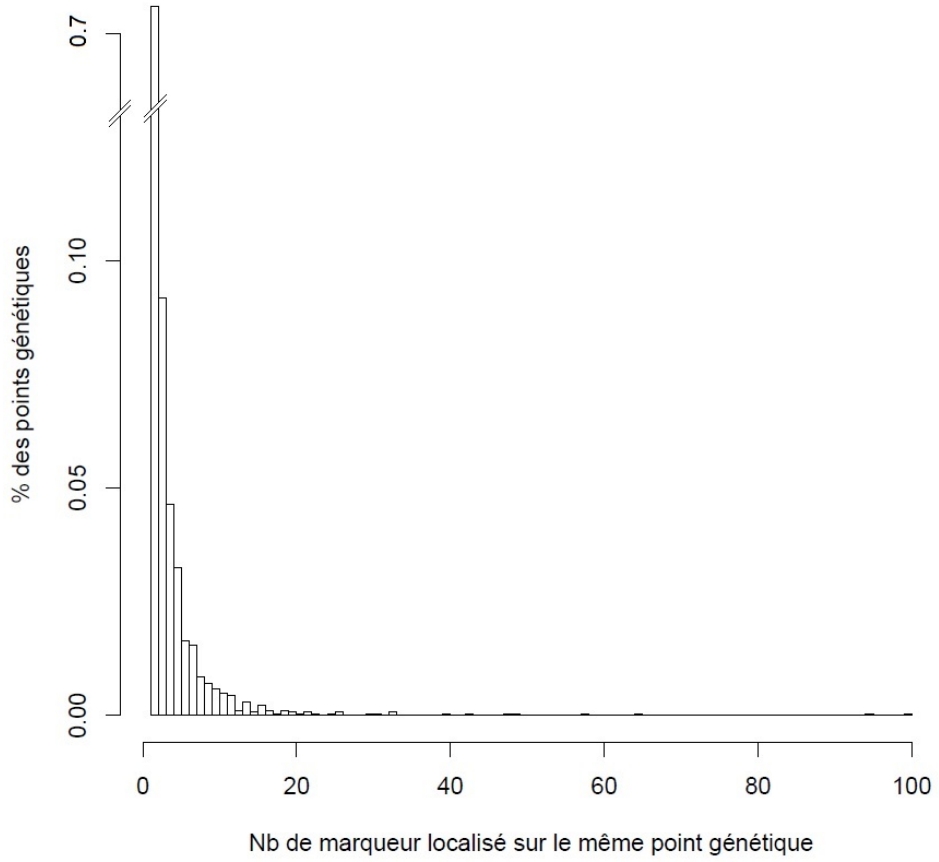


FIGURE 3.5 – Distribution du nombre de marqueurs par point génétique sur la carte génétique disponible de la puce 9K iSelect [Cavanagh et al., 2013].

3.2.2.5 Population MAGIC NIAB

L'équipe de Ian Mackay (NIAB, Cambridge) a développé une population MAGIC, issue du croisement pyramidal de 8 parents. Elle a été génotypée avec la puce 90K, et dans l'objectif de comparer avec l'autre population MAGIC de blé, nous avons réduit ces données aux 157 marqueurs localisés sur le chromosome 3B communs avec le jeu de données de la population MAGIC INRA, la puce 90K incluant la plupart des marqueurs de la puce 9K iSelect.

Cette population a été utilisée pour tester l'algorithme à l'échelle d'un chromosome et de comparer les résultats avec ceux obtenus avec la population MAGIC INRA. Le chromosome étudié est le chromosome 3B car la carte physique a été récemment développée par l'équipe GDEC de Clermont Ferrand (Comm. Pers. F. Choulet, Paux et al. [2008]). Les marqueurs de la puce 9K iSelect génotypés sur les populations MAGIC NIAB et MAGIC INRA ont été blastés par F. Choulet sur la pseudo-molécule d'ADN du chromosome 3B . De ce fait, la position physique des marqueurs a pu être utilisée comme référence.

L'algorithme a été testé sur toutes ces populations qui ont des histoires très différentes, pour évaluer son efficacité dans des cas où la relation entre le DL et la distance génétique est différente (Figure 3.6). En effet, le DL courte distance qui représente le DL parental est plus fort pour la population bi-parentale en comparaison au panel et aux populations MAGIC. Mais à longue distance, le DL du panel reste élevé en comparaison avec celui des populations recombinantes pour lesquelles ce DL a été cassé lors des générations de recombinaison. Le tableau 3.2 récapitule les caractéristiques des populations utilisés.

Tableau 3.2 – Tableau récapitulatif des caractéristiques des différentes populations utilisées. Il indique l'espèce de la population, le nombre de méioses qu'elle a subi, la localisation des marqueurs utilisés soit génome entier soit sur un chromosome en particulier, le nombre d'individus utilisés pour l'estimation du DL, la corrélation entre le DL et la distance génétique avec l'ensemble des marqueurs pour toutes les populations exceptée la MAGIC Droso (distance physique)) et le DL inter-chromosome maximal observé dans 99% des cas.

Population	espèce	nb méioses	localisation	nb d'ind	cor(DL,dist)	DL inter-chr
données simulées		15				
Bi-parentale	maïs	1	génome entier	302	-0,46	0,048
Panel (corrigé par l'apparentement)	blé tendre		génome entier	56	-0,24	0,12
MAGIC INRA	blé tendre	15	génome entier	380	-0,25	0,024
MAGIC NIAB	blé tendre	3	Chr 3B	720	-0,41	NA
MAGIC Droso	Drosophile	20	Chr 2L	487	-0,27	NA

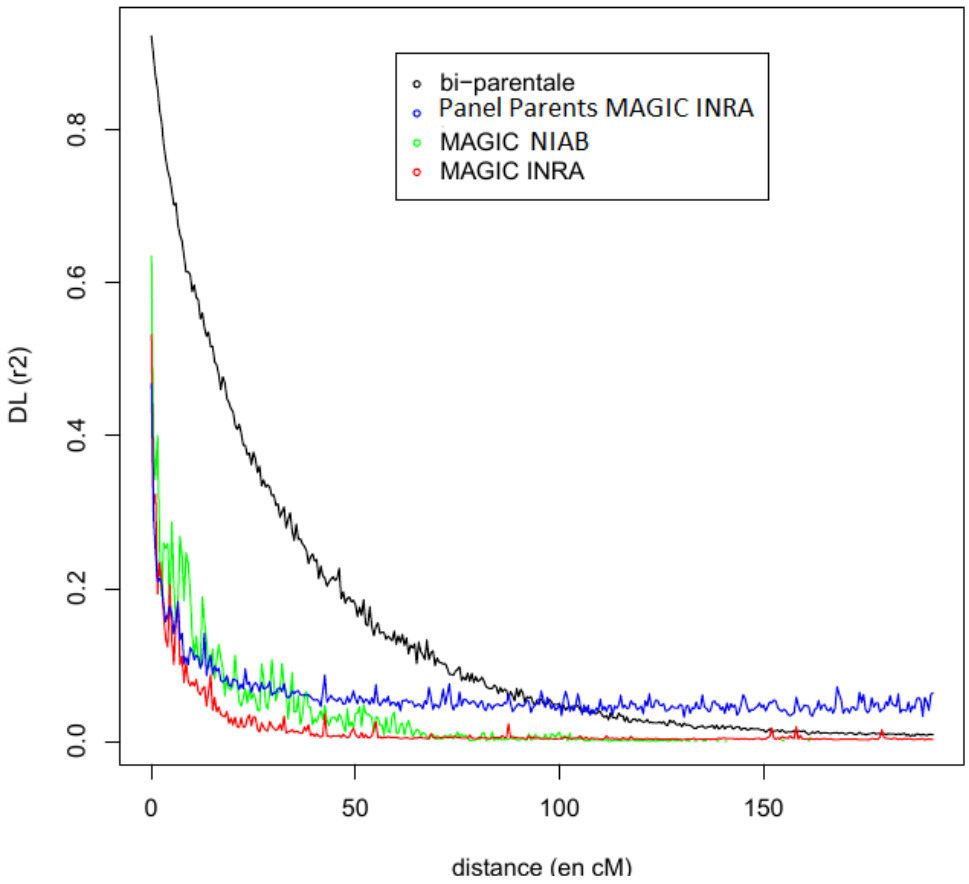


FIGURE 3.6 – Décroissance du DL en fonction de la distance génétique (en cM) pour les différentes populations étudiées avec des cartes génétiques disponibles.

3.2.3 Construction des fenêtres de test

Comme le temps de calcul de l'algorithme est long, la précision de cartographie de l'algorithme a été testée sur différentes fenêtres de marqueurs de quatre jeux de données différents (Tableau 3.3). Les fenêtres de marqueurs à ordonner ont été construites de deux manières différentes :

- des fenêtres de *50 marqueurs successifs* pris aléatoirement sur la carte génétique, dans ce cas la distance entre les marqueurs dépend de la densité de la carte utilisée. Dix fenêtres par population ont été définies de cette façon.
- des fenêtres de marqueurs avec *des distances inter-marqueurs contrôlées*. A partir d'un marqueur tiré aléatoirement, les autres marqueurs ont été tirés successivement de manière à avoir une distance minimum entre les marqueurs qui a pris pour valeurs 0cM, 0,5cM, 1cM, 2cM, 2,5cM, 5cM, 10cM et 15cM. Les fenêtres contenaient au maximum 31 marqueurs. En moyenne une vingtaine de fenêtres a été définie par population et par distance.

Le tableau 3.3 récapitule le nombre de fenêtres ainsi que le nombre de marqueurs par fenêtre en fonction du DL moyen présent dans chacune. Ce tableau montre bien la différence de DL qu'il y a entre les différentes populations étudiées. En effet les fenêtres construites avec la population bi-parentale ont un DL moyen compris entre 0,3 et 0,9, alors que le panel et la population MAGIC NIAB ont un DL compris entre 0 et 0,5. La population MAGIC INRA a le DL moyen maximum des fenêtres le plus faible parmi les populations d'étude avec 0,3.

3.2.4 Etude du chromosome 3B

Les populations MAGIC INRA et MAGIC NIAB ont été comparées à deux échelles sur le chromosome 3B : i) à très faible échelle, à l'échelle du point génétique de la carte disponible ii) à l'échelle chromosomique. Dans cette partie, nous avons pris comme carte de référence la carte physique.

3.2.5 Indices de comparaison

Les statistiques réalisées dans chaque cas sont le pourcentage de marqueurs ordonnés, la corrélation entre l'ordre trouvé par l'algorithme et l'ordre sur la carte de référence, la corrélation entre l'ordre trouvé et la position des marqueurs sur la carte de référence (en cM ou bp). Ces indices ont été calculés pour l'ordre robuste comme pour l'ordre souple.

Tableau 3.3 – Tableau récapitulatif des différentes fenêtres utilisées pour les statistiques, avec le nombre de fenêtre (NF), et le nombre minimum (NMmin), moyen (NMmoy) et maximum (NMmax) de marqueurs composant ces fenêtres, rangées en fonction du DL moyen présent dans chacune.

DL moyen inter-marqueurs	0-0,1				0,1-0,2				0,2-0,3			
Population	NF	NM min	NM moy	NM max	NF	NM min	NM moy	NM max	NF	NM min	NM moy	NM max
Bi-parentale												
Panel (corrigé par l'apparentement)	9	14	27,9	50	46	10	33,7	50	11	14	28,1	31
Panel (non corrigé)	10	16	25,9	50	45	15	34,7	50	18	10	34,1	50
MAGIC INRA	148	9	32,3	50	37	14	35,9	50	7	31	46,8	50
DL moyen inter-marqueurs	0,3-0,4				0,4-0,5				0,5-0,6			
Population	NF	NM min	NM moy	NM max	NF	NM min	NM moy	NM max	NF	NM min	NM moy	NM max
Bi-parentale	7	16	25,3	31	10	29	30,8	31	14	22	30,4	31
Panel (corrigé par l'apparentement)	1	50	50	50								
Panel (non corrigé)					2	50	50	50				
MAGIC INRA												
DL moyen inter-marqueurs	0,6-0,7				0,7-0,8				0,8-0,9			
Population	NF	NM min	NM moy	NM max	NF	NM min	NM moy	NM max	NF	NM min	NM moy	NM max
Bi-parentale	29	17	30,3	31	35	18	34,8	50	3	31	43,7	50
Panel (corrigé par l'apparentement)												
Panel (non corrigé) MAGIC INRA												

3.3 Résultats

3.3.1 Classification hiérarchique

La fonction "hclust" du logiciel R a été testée sur les 90 premiers marqueurs du chromosome 9 génotypés sur la population bi-parentale de maïs (Figure 3.7). La corrélation entre l'ordre attendu de la carte génétique et l'ordre trouvé par la fonction avec la matrice de génotypage et la matrice de DL sont respectivement de 0,56 et 0,48 (Figure 3.8). La fonction "hclust" rassemble les marqueurs similaires qui ont une forte corrélation dans les mêmes "clusters", puis les "clusters" les plus proches pour former un arbre de similarité. A l'échelle locale, l'ordre des marqueurs est cohérent avec la carte génétique. Mais à une échelle plus globale, comme l'ordre de deux marqueurs se trouvant dans un même "cluster" n'est pas important pour l'algorithme, l'ordre et le sens des groupes ne sont pas toujours cohérents avec la carte génétique. La comparaison entre l'ordre des marqueurs trouvé par le "hclust" et la carte génétique montre des inversions d'ordre intra-groupe et des déplacements de blocs. Ce résultat était attendu puisque les clusters ne sont pas ordonnés, il peuvent tourner autour de leur axe et donc être inversés.

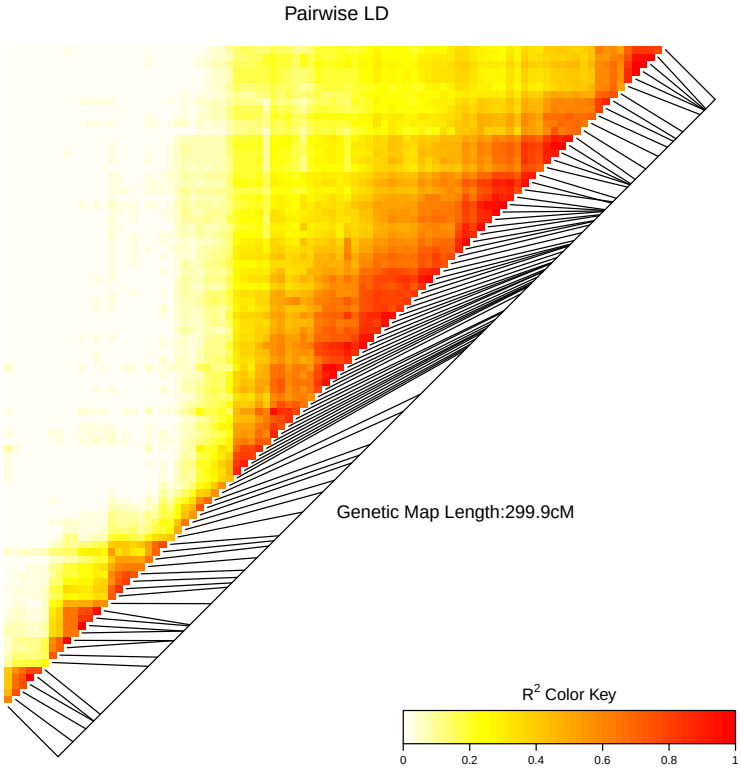


FIGURE 3.7 – Matrice de DL des 90 premiers marqueurs du chromosome 9 ordonnés selon la carte génétique de la population bi-parentale de maïs IBM.

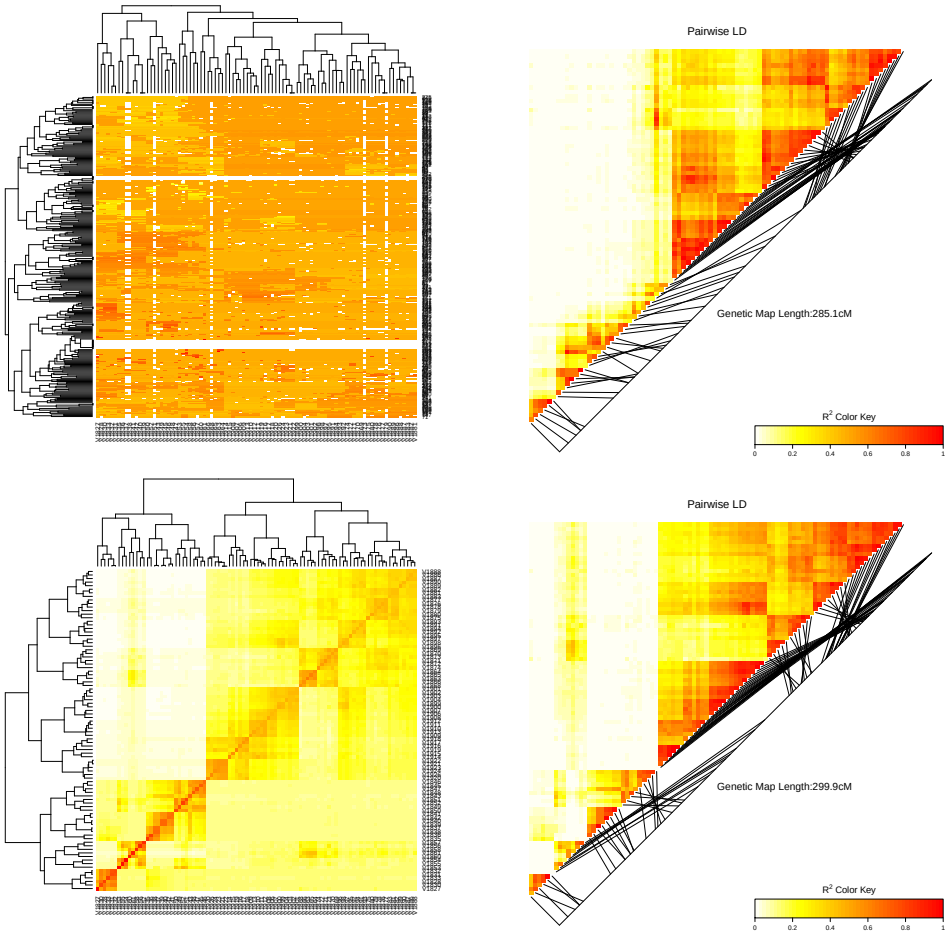


FIGURE 3.8 – Test de la procédure "hclust" avec la matrice de génotypage (en haut) avec les individus en ligne et les marqueurs en colonne et avec la matrice de DL (en bas) des 90 premiers marqueurs du chromosome 9 de la population bi-parentale de maïs. A gauche : dendogramme issu de la fonction "hclust" avec arbre de similarité. A droite : demi-matrice de DL avec les marqueurs ordonnés avec la fonction "hclust" et comparaison de cet ordre avec celui de la carte génétique.

Les procédures disponibles ne nous ayant pas permis d'obtenir un ordonnancement optimal des marqueurs, nous avons développé un algorithme d'ordonnancement des marqueurs basé sur l'unique utilisation de la mesure du DL. Il a été testé sur diverses populations, pour déterminer la fiabilité de l'approche et ses limites.

3.3.2 Validation de l'algorithme d'ordonnancement

3.3.2.1 Données simulées

Sur 30 tests, l'algorithme ordonne parfaitement tous les marqueurs quel que soit le bruit. L'augmentation du bruit dans les données accroît le nombre de groupes de

liaison indépendants, par masquage du signal de DL entre les marqueurs les plus distants (Figure 3.9).

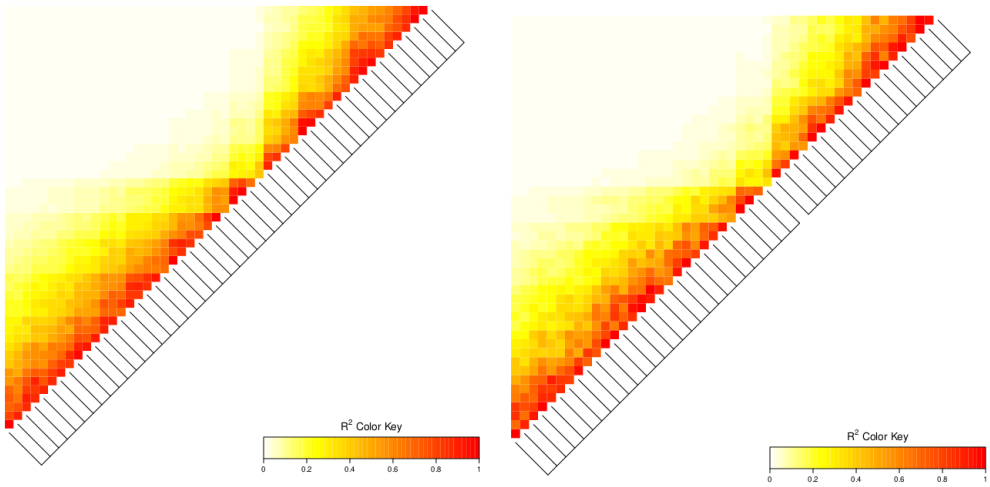


FIGURE 3.9 – Matrice de DL des marqueurs simulés bruités ordonnés par l'algorithme. A gauche : données bruitées par ajout d'un nombre aléatoire tiré dans une loi uniforme [0 ;0,1]. A droite : données bruitées par ajout d'un nombre aléatoire tiré dans une loi uniforme [0 ;0,2]. Le trait blanc vertical présent sur le graphique de droite est une séparation entre deux groupes de liaison indépendants.

3.3.2.2 Population bi-parentale

Le premier test a été réalisé avec les 50 premiers marqueurs du chromosome 9 (Figure 3.7), sur la matrice de DL randomisée de la population bi-parentale de maïs. L'algorithme a ordonné 44 marqueurs dans l'ordre robuste et 49 marqueurs dans l'ordre souple. La corrélation entre l'ordre attendu et l'ordre trouvé est de 0,99 pour les deux ordres. Les erreurs faites par l'algorithme sont majoritairement des inversions locales (Figure 3.10). Sachant que l'indice de qualité "sj" est de 15,03 avec l'ordre de la carte génétique, l'algorithme a trouvé un meilleur ordre que l'ordre de départ selon notre critère, avec un indice "sj" de 7,01 pour l'ordre robuste et de 12,95 pour l'ordre souple.

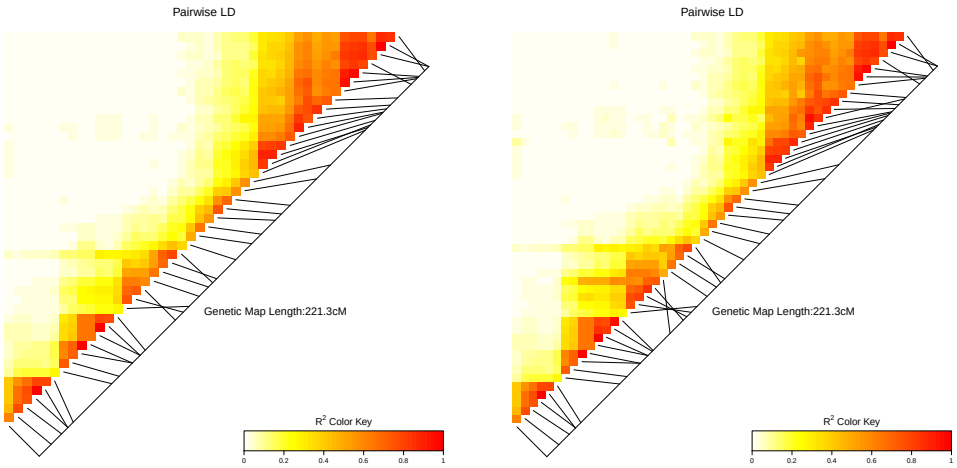


FIGURE 3.10 – Résultat du test sur les 50 premiers marqueurs localisés sur le chromosome 9 avec les données issues d'une population bi-parentale de maïs. Demi-matrice de DL avec comparaison entre ordre trouvé par l'algorithme et la carte génétique de référence. A gauche : ordre robuste ($sj = 7.01$). A droite : ordre souple ($sj = 12.95$)

Après réitération du test, sur environ une centaine de fenêtres comportant entre 16 et 50 marqueurs avec des densités de marqueurs différentes, l'algorithme a positionné en moyenne 72% des marqueurs avec un ordre corrélé à 0,96 avec la carte génétique (pour l'ordre souple 89% des marqueurs sont ordonnés avec une corrélation de 0,95). Les corrélations entre l'ordre trouvé avec la position attendue des marqueurs sur la carte génétique ou avec l'ordre des marqueurs sur cette même carte sont très similaires voire identiques dans le cas de cette population (cor=0,96) (Tableau 3.4). La corrélation globale du DL avec la distance génétique (en cM) est de -0,46 (Tableau 3.2). Le DL inter-chromosome est très faible puisqu'il est inférieur à 0,048 dans 99% des cas (Tableau 3.2).

Tableau 3.4 – Tableau comparatif des résultats d'ordonnancement des marqueurs sur les différentes populations.

	robuste								souple								
	corrélation position			corrélation ordre			% marqueurs positionnés			corrélation position			corrélation ordre			% marqueurs positionnés			nb de groupe de liaison moyen
	min	moy	max	min	moy	max	min	moy	max	min	moy	max	min	moy	max	min	moy	max	
bi-parentale	0,57	0,96	1	0,5	0,96	1	10	72	100	0,15	0,95	0,99	0,18	0,95	1	18	89	100	1,09
MAGIC INRA	0,01	0,64	1	0	0,63	1	6	21	86	0,0001	0,63	1	0	0,62	1	10	51	97	2,4
parents MAGIC INRA (non corrigé)	0,05	0,42	1	0,02	0,42	1	6	57	88	0,01	0,36	1	0,001	0,37	1	16	82	100	1.2
parents MAGIC INRA (corrigé)	0,02	0,53	0,98	0,03	0,52	1	6	36	87	0,003	0,49	0,98	0,02	0,48	1	10	69	100	2.2

3.3.2.3 Panel de lignées : parents MAGIC INRA

Les tests sur les 70 fenêtres comportant entre 10 et 50 marqueurs avec des densités de marqueurs différentes montrent que l'algorithme ordonne 36% des marqueurs avec une corrélation entre les ordres de 0,52 (69% des marqueurs avec une corrélation de 0,48 pour l'ordre souple) (Tableau 3.4). La corrélation entre le DL et la distance génétique est de -0,24 (Tableau 3.2). Le DL inter-chromosomique est inférieur à 0,12 dans 99% des cas (Tableau 3.2).

Du fait de la structure génétique présente dans le panel des 56 parents (groupes de lignées fortement apparentées ; voir Chap 4)), certains marqueurs génétiquement indépendants présentent des valeurs de DL fortes. De ce fait, une correction de l'apparentement entre les individus [Mangin et al., 2012] a été effectuée (Figure 3.11). Cette correction a une tendance à diminuer le DL longue distance.

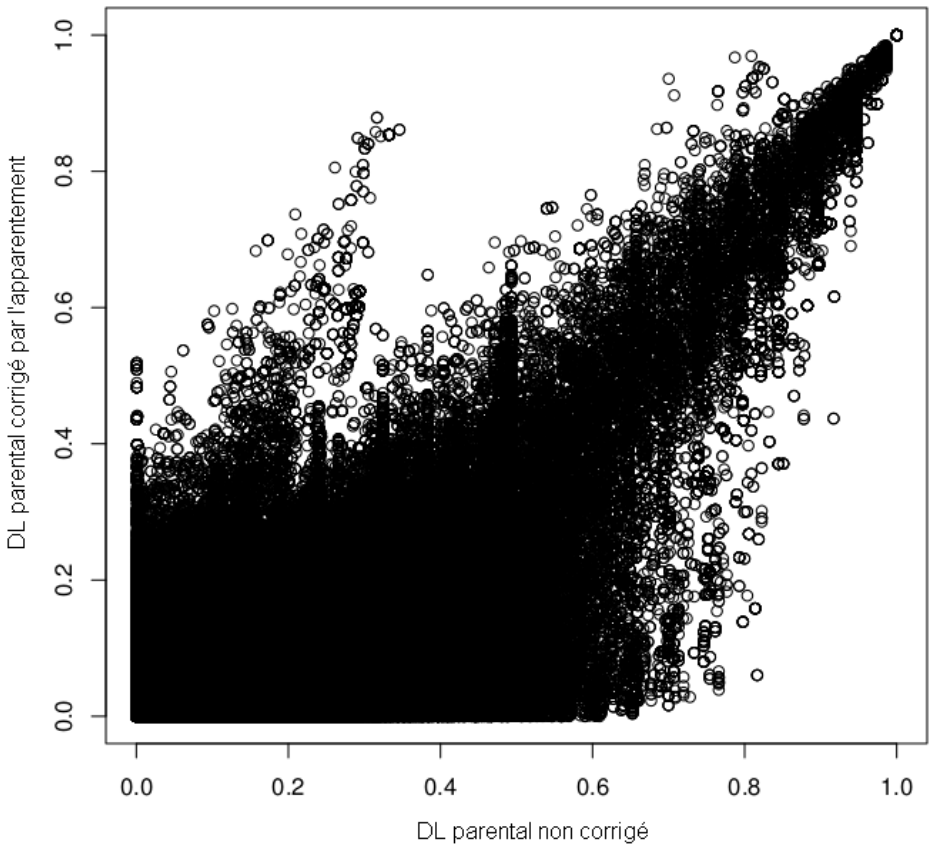


FIGURE 3.11 – DL parental de la population MAGIC INRA corrigé par l'apparentement en fonction du DL parental non corrigé.

Le nombre de marqueurs intégrés dans la cartographie est plus faible après correction. Avec les données corrigées, l'algorithme fait cependant plus de groupes de liaison indépendants, et la corrélation entre les ordres est meilleure de 0,1 pour l'ordre robuste (et de 0,11 pour l'ordre souple) (Tableau 3.4).

3.3.2.4 Population MAGIC INRA

Avec la population très diversifiée et hautement recombinante, les tests sur les 190 fenêtres contenant entre 9 et 50 marqueurs ordonnent 21% des marqueurs avec une corrélation entre les ordres de 0,63 (Tableau 3.4) (51% des marqueurs avec une corrélation de 0,62 pour l'ordre souple). La corrélation entre le DL et la distance génétique est de -0,25 (Tableau 3.2). Le DL inter-chromosome est inférieur à 0,024 dans 99% des cas (Tableau 3.2).

La population MAGIC INRA génotypée avec la puce 9K iSelect fournit 6 449 marqueurs polymorphes et de bonne qualité. Sur ces 6 449 marqueurs, 5 313 sont uniques *i.e.* ne sont pas en DL complet. Les marqueurs de la puce 9K iSelect ont été cartographiés en utilisant sept populations [Cavanagh et al., 2013] sur 2 742 points génétiques. Dans la population MAGIC INRA, de nombreux marqueurs polymorphes uniques sont donc localisés sur le même point génétique (Figure 3.5).

Les différentes populations étudiées n'apportent pas les mêmes niveaux de précision et d'information, car elles ont des niveaux de DL et des densités de marquage assez disparates (Figure 3.6). Comme nous avons des données de génotypage sur des marqueurs communs dans les populations MAGIC NIAB et MAGIC INRA, ces deux populations ont été étudiées de manière plus précise sur le chromosome 3B.

3.4 Résultats sur un chromosome : exemple du chromosome 3B du blé tendre

3.4.1 A l'échelle locale

Une approche de cartographie locale a permis de tester l'efficacité de l'algorithme pour ordonner des marqueurs à très faible échelle. Effet dans la population MAGIC INRA, plusieurs marqueurs sont cartographiés sur le même point génétique dans la carte de référence disponible [Cavanagh et al., 2013]. L'ensemble de marqueurs communs présents sur la carte génétique du chromosome 3B, la carte physique, le génotypage de la population MAGIC NIAB et de la population MAGIC INRA représente 157 marqueurs répartis sur 89 points génétiques dont 12 points génétiques avec trois marqueurs ou plus (paragraphe 3.2.2.4, Figure 3.5). La corrélation entre l'ordre attendu et l'ordre observé est toujours supérieure de 0,06 pour la population MAGIC INRA par rapport à la population MAGIC NIAB avec une corrélation entre les ordres de 0,70 *vs* 0,64 (Tableau 3.5).

Tableau 3.5 – Comparaison des corrélations entre l'ordre ou la position des marqueurs sur la carte physique et l'ordre trouvé par l'algorithme pour l'ordonnancement de marqueurs localisés sur le même point génétique dans la carte disponible.

	corrélation position	corrélation ordre
MAGIC NIAB	0,68	0,64
MAGIC INRA	0,74	0,70

3.4.2 A l'échelle globale

L'ordonnancement de tous les marqueurs du chromosome 3B à partir des données de DL de la population MAGIC NIAB par l'algorithme a formé un unique groupe de liaison (Figure 3.12). Certaines régions sont inversées entre l'ordre trouvé par l'algorithme et les positions de la carte physique. La corrélation est seulement de 0,12 même si la figure laisse deviner des zones présentant une meilleure corrélation.

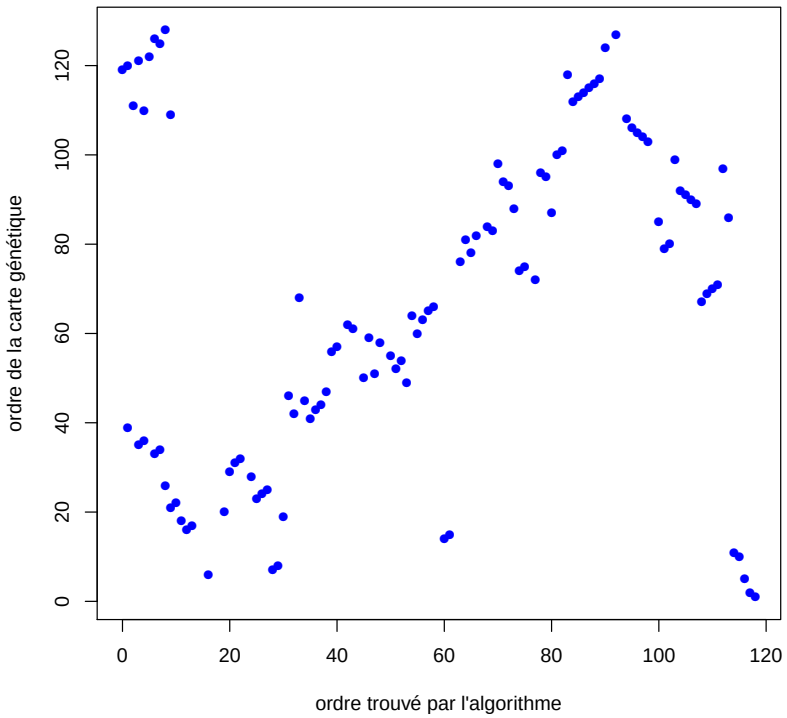


FIGURE 3.12 – Corrélation pour la population MAGIC NIAB entre l'ordre trouvé par l'algorithme et la carte génétique, les couleurs représentent des groupes de liaison indépendants.

Avec la population MAGIC INRA, l'algorithme a crée six groupes de liaison avec tous les marqueurs localisés sur le chromosome 3B (Figure 3.13). Parmi ces six groupes, deux groupes contiennent des marqueurs situés sur le même point génétique, de plus parmi les autres groupes, des marqueurs localisés sur le même point de la carte génétique ont été ordonnés. En moyenne dans les trois autres groupes avec plus de quatre marqueurs, l'ordre trouvé par l'algorithme et l'ordre sur la carte génétique ont une corrélation de 0,42. Si on veut comparer avec la population MAGIC NIAB, la corrélation moyenne de groupes de même taille est de 0,51.

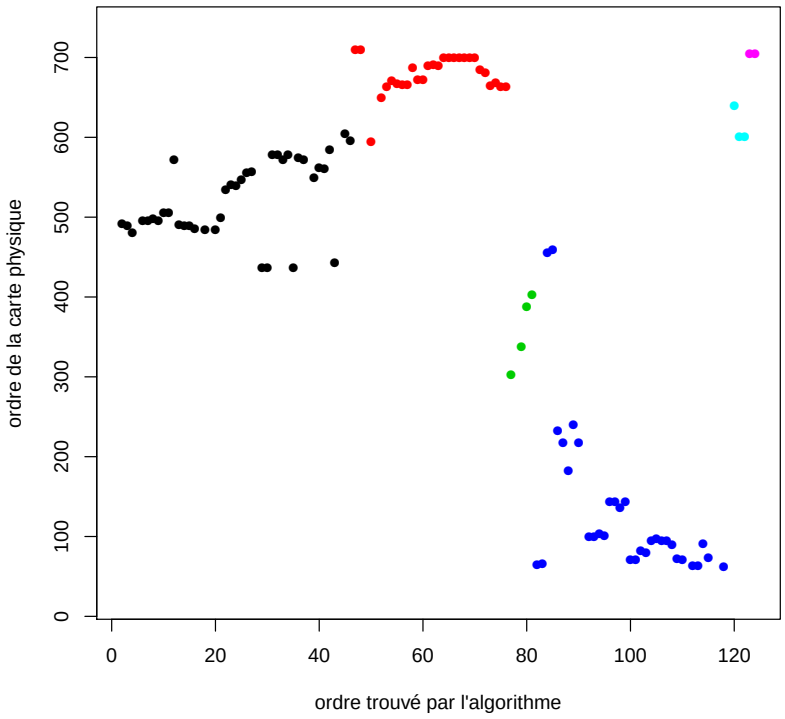


FIGURE 3.13 – Corrélation entre l'ordre trouvé par l'algorithme et la carte physique avec un ordre aléatoire des groupes de liaison, les couleurs représentent des groupes de liaison indépendants.

3.5 Discussion

A partir des résultats sur les différentes populations, les possibilités d'amélioration de l'algorithme développé ont été discutées. Le DL exploité par l'algorithme est la résultante de i) un niveau de DL initial (dépendant notamment du nombre de parents fondateurs) et ii) de l'évolution de ce DL lors de la création de la population (type de croisements effectués, nombre de générations

de recombinaison, taille efficace,...). Les modalités d'utilisation de cet algorithme avec des populations naturelles ont ensuite été définies.

Les étapes d'ordonnancement de l'algorithme sont proches de celles incluses dans le logiciel MapMaker [Lander et al., 1987] (TRY, COMPARE). Au lieu de calculer des ratios de vraisemblance comme dans MapMaker, l'algorithme minimise un indice comme dans le logiciel MadMapper (`http://cgpdb.ucdavis.edu/XLinkage/MadMapper/`).

L'algorithme que nous avons développé s'avère performant, car il ordonne correctement les marqueurs afin que le DL soit décroissant monotone sur la matrice de DL. Cependant, il intègre entre 21% et 72% des marqueurs dans l'ordre robuste suivant les populations considérées. Ce nombre limité de marqueurs positionnés pourrait être augmenté de différentes façons mais avec potentiellement une baisse d'efficacité de l'algorithme.

Premièrement nous avons établi un ordre sans test de robustesse, cet ordre souple contient logiquement plus de marqueurs que l'ordre robuste mais avec une diminution de la corrélation entre l'ordre obtenu et l'ordre attendu. Cette baisse de corrélation n'est pas toujours très importante (comprise entre 0,01 et 0,05) vu le pourcentage de marqueurs supplémentaires ordonnés (entre 17% et 33%).

Une deuxième façon serait d'avoir un indice à minimiser moins sélectif, en effet seuls les marqueurs avec un "pattern" de DL très "propre" peuvent être ordonnés. L'indice utilisé pourrait être assoupli en pondérant la différence de DL cumulée en fonction de la distance entre les marqueurs afin de limiter le poids du bruit longue distance. L'estimateur du DL est un point clé de la méthode. L'utilisation du r^2 nous a semblé approprié puisqu'il prend en compte les fréquences alléliques et qu'il est robuste pour des données avec au maximum 2% d'erreur de génotypage [Akey et al., 2001]. Dans les populations de lignées recombinantes étudiées, le DL inter-chromosomique est très faible, ce qui valide son utilisation possible pour l'ordonnancement des marqueurs ; des marqueurs localisés sur des chromosomes différents ne devraient pas se retrouver dans un même groupe de liaison.

Le dernier point serait de jouer sur les valeurs seuils. Le seuil de liaison que nous avons utilisé pour démarrer une chaîne est très élevé (DL=0,8), afin qu'il soit utilisable pour toutes les populations. Suivant le DL de la population utilisée, il pourrait être abaissé. De même le seuil de liaison (DL=0,1) pour définir les marqueurs à tester pour les intégrer à l'ordre, pourrait être optimisé en fonction de la population. Il faut s'assurer de n'exploiter que des DL significatifs. Ces différents ajustements possibles montrent bien que l'algorithme est adaptable à la structure de la population étudiée.

Les résultats ont été obtenus majoritairement sur des fenêtres de petite taille (30 à 50 marqueurs) car le temps de calcul est assez long. Dans toutes les méthodes de cartographie, il faut trouver un équilibre entre la complexité de l'algorithme et le temps de calcul [Cheema and Dicks, 2009]. Il existe différentes méthodes de cartographie comme la méthode de classification hiérarchique (hclust) ou d'analyse de réseaux. La construction d'une carte fiable pour une population donnée pourrait

se faire grâce à une carte consensus à partir des cartes crées avec différentes méthodes [Cheema and Dicks, 2009].

Une limite de l'étude de l'efficacité de la méthode réside dans la nécessité de s'appuyer sur une carte de référence. Nous avons bâti une partie de notre raisonnement sur l'utilisation d'une carte physique du chromosome 3B. Cependant, si cette carte est la meilleure disponible à ce jour, elle présente certainement un niveau d'erreur, qui n'est pas encore connu. De plus, par construction, la carte physique utilise les données physiques uniquement à très faible échelle, (l'échelle du "scaffold"), ensuite l'ordonnancement des "scaffolds" est réalisé à l'aide de carte génétique et finalement les combinaisons de "scaffold" sont ordonnées à l'aide de l'information du DL (d'un panel et d'une population bi-parentale, Comm. Pers. F. Balfourier). Même si le DL est calculé avec d'autres populations, cette carte de référence n'est pas complètement indépendante de la méthode testée. Cette carte physique est corrélée à 0,87 en ordre et à 0,76 en position avec la carte génétique. La figure 3.14 montre en effet quelques incohérences entre ces deux cartes. Étonnamment l'analyse de quelques marqueurs qui présentent une forte incohérence entre la carte physique et la carte génétique montre que les cartographies par DL des populations MAGIC INRA et MAGIC NIAB sont en plus fortes cohérences avec la carte génétique, ce qui repose la question de la validité de la carte physique sur ces points. De manière plus générale, la corrélation entre l'ordre trouvé par l'algorithme et les positions sur la carte génétique est supérieure d'environ 0,15 par rapport à celle avec les positions de la carte physique. Aucune carte n'est vraiment fiable, mais les cartes construites sur la base du DL n'ont pas la même utilité que les cartes physiques ou génétiques ; il a été montré que les cartes basées sur le DL sont plus efficaces que les cartes physiques pour la génétique d'association [Maniatis et al., 2004].

Si l'on considère que l'ordre de la carte physique est le "bon" ordre, le DL à moyenne distance est trop structuré, i.e. ne présente pas une décroissance monotone en fonction de la distance (Figure 3.15) pour que l'algorithme puisse retrouver cet ordre. Nous avons regardé des données d'une population de drosophile de type MAGIC 8 parents avec 20 générations de recombinaison. Avec la carte physique associée, ces données montrent là encore que le DL est très structuré (Figure 3.16) à moyenne distance donc notre algorithme ne pourrait ordonner les marqueurs avec ces données brutes.

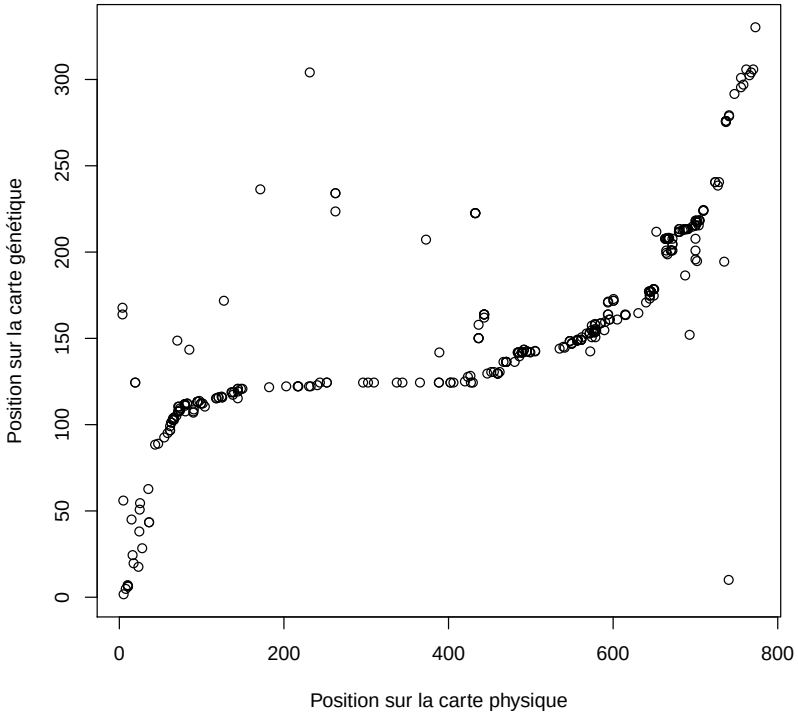


FIGURE 3.14 – Corrélation entre la carte physique [Paux et al., 2008] et la carte génétique australienne [Cavanagh et al., 2013] pour les marqueurs du chromosome 3B.

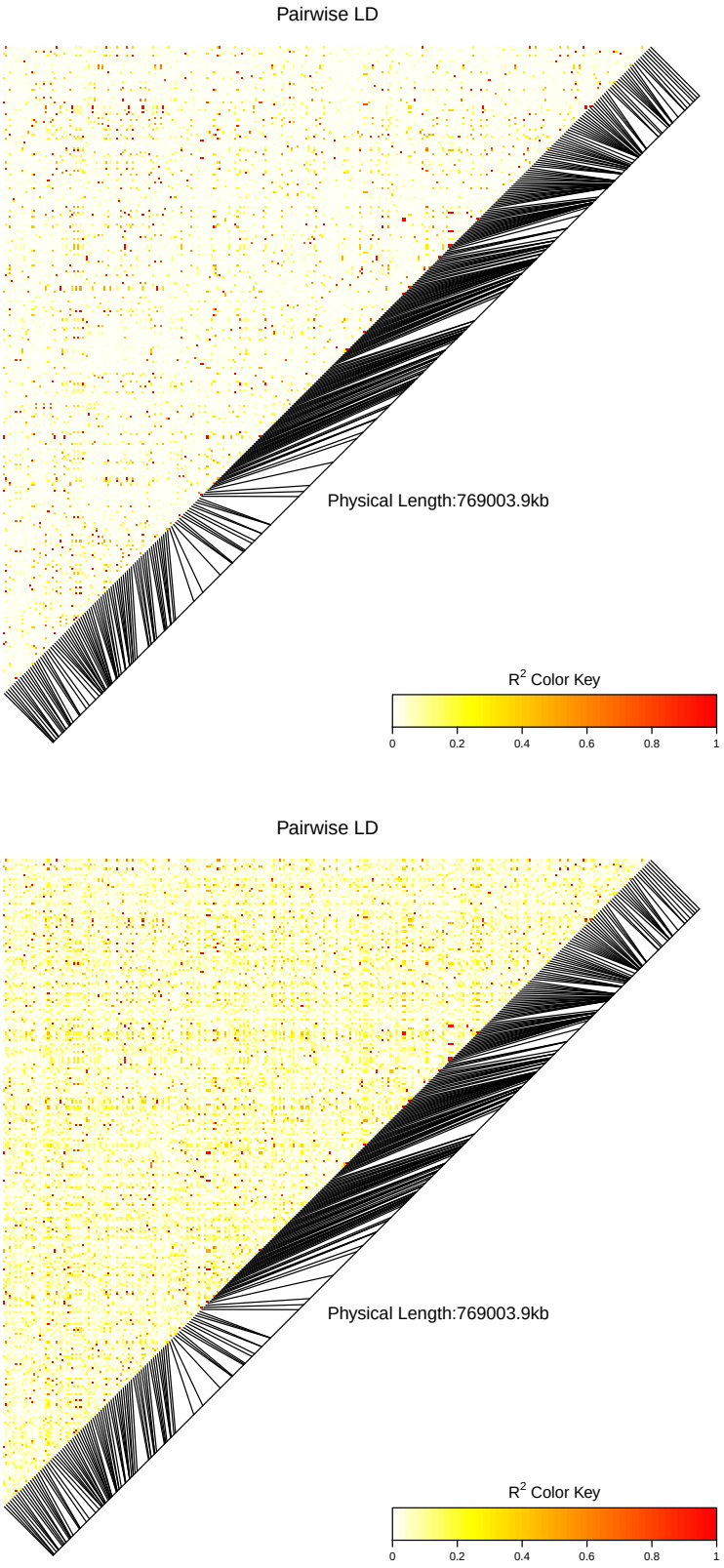


FIGURE 3.15 – Matrice de DL du chromosome 3B de la population MAGIC INRA. Les marqueurs sont ordonnés suivant la carte physique. En haut : chez les descendants. En bas : chez les parents

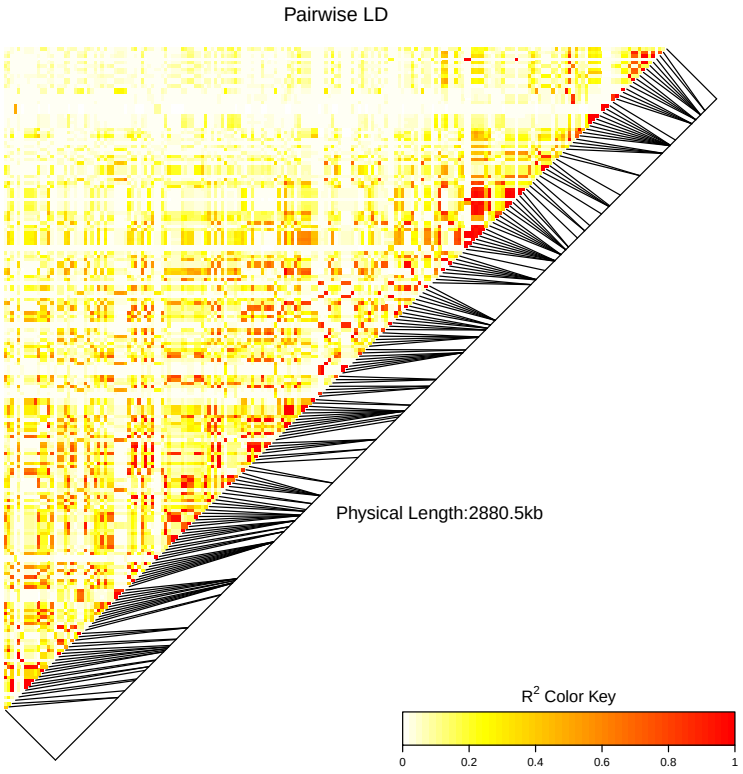


FIGURE 3.16 – Matrice de DL du chromosome 2L de la drosophile. Les marqueurs sont ordonnés suivant la carte physique [King et al., 2012].

L'absence de corrélation entre DL et distance génétique dans la population expérimentale de drosophile, et le fait que la population bi-parentale ait donné les meilleurs résultats, suivie dans l'ordre de la population MAGIC NIAB, puis la population MAGIC INRA (que ce soit les parents ou les descendants), montrent que la performance de la cartographie à partir du DL sera directement dépendante de la structure du DL (quantifiée par la corrélation entre le DL et la position sur la carte génétique de référence) dans la population.

Les différentes populations étudiées ont des histoires différentes qui impactent la structure du DL (Figure 3.6). Premièrement, le nombre de parents dans la population diminue le DL initial, tout en le structurant en fonction des relations phylogénétiques entre parents, représentatives des recombinaisons ancestrales. Ce DL est directement affecté par l'apparentement des individus et/ou par la structure de la population. Comme ce DL initial peut brouiller le signal de générations de recombinaison plus récentes, nous nous sommes demandés si une correction du DL final de la population à partir du DL initial pourrait améliorer la corrélation entre le DL et la carte génétique. Un seul test de correction du DL des descendants de la population MAGIC INRA avec le DL parental a été réalisé par simple multiplication du DL final par le DL initial. Le résultat était une forte augmentation du nombre de

groupes de liaison (18 groupes) avec très peu de marqueurs dans chaque groupe (en moyenne 6 marqueurs par groupe). Mais la corrélation moyenne pour les groupes de liaison avec au moins trois marqueurs entre l'ordre des marqueurs trouvé par l'algorithme et celui de la carte physique est de 0,83 (soit une amélioration de 0,2). Il était attendu que des groupes avec un nombre de marqueurs plus faible aient une meilleure corrélation.

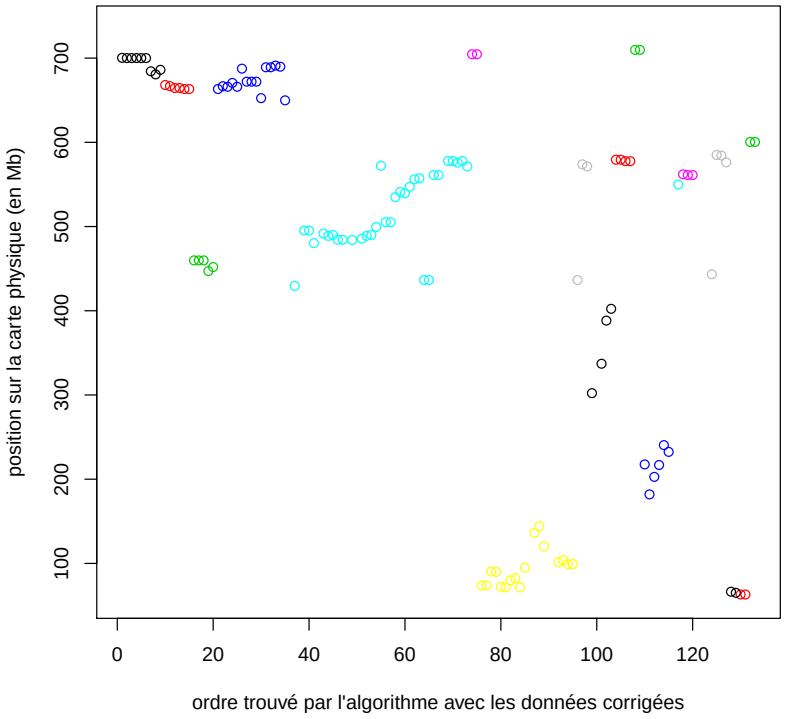


FIGURE 3.17 – Comparaison entre l'ordre de la carte génétique et l'ordre construit par l'algorithme avec les données de DL corrigées par le DL parental. Les couleurs représentent les groupes de liaison indépendants.

Cette correction est assez drastique puisqu'elle ne garde que les DL forts dans la population initiale et dans la population finale. Les groupes ainsi formés ont une très bonne cohérence mais ils sont trop nombreux, en effet l'augmentation du nombre de groupes correspond à une perte d'information même si la corrélation est plus forte. Il serait donc intéressant de tester d'autres pondérations du DL initial, et de combiner les analyses du DL brut et corrigé pour améliorer la cartographie. D'autres approches pourraient être testées, comme l'utilisation du génotypage multi-locus à partir de la reconstruction d'haplotypes parentaux. L'analyse des haplotypes parentaux permettrait de définir les recombinaisons ancestrales, et donc de détecter les nouvelles recombinaisons ayant eu lieu pendant la création de la

population, repérables sur le génotypage des descendants, et à partir desquelles baser la construction de la carte génétique. Une autre façon d'utiliser les haplotypes ancestraux seraient de les considérer comme des marqueurs multi-alléliques et d'ordonner les blocs haplotypiques sur la base de leur DL.

Deuxièmement, ce DL initial évolue notamment en fonction du nombre de générations de recombinaison. Plus les loci sont indépendants et plus le DL est cassé avec très peu de générations de recombinaison (Figure 3.1). Il varie donc en fonction du temps, de la distance génétique et du mode de reproduction de la population [Weir and Hill, 1980]. La population bi-parentale et la population MAGIC NIAB sont le résultat de respectivement 1 et 2 générations de recombinaison (Tableau 3.2) qui ont surtout cassées le DL longue distance (marqueurs non liés). Pour la population MAGIC INRA, les individus ont subi un grand nombre de générations d'inter-croisement. Dans ce cas, le DL longue distance est donc très fortement diminué ($r^2 = 0,0037$). Cependant la comparaison du DL des descendants avec le DL parental montre que dans certains cas, sa valeur s'est accrue (Figure 3.18), avec comme probable explication la dérive (et la sélection). La dérive entraîne une variation stochastique du DL, qui crée d'autant plus de DL que la taille efficace de la population est petite (goulot d'étranglement) [Sved, 1971]. Ce biais peut-être corrigé par un facteur inversement proportionnel à la taille efficace. Il faut donc trouver un équilibre dans le nombre de recombinaisons pour que le DL longue distance soit le plus faible possible tout en limitant les biais comme la dérive.

Le DL entre les marqueurs est cassé à chaque recombinaison. La densité de marquage optimal est donc liée au DL moyen de la population. Dans le cas de la MAGIC INRA il y a un manque d'adéquation entre le très faible DL de la population et le nombre de marqueurs étudiés. Un plus grand nombre de marqueurs bien répartis amènerait l'algorithme à ne former qu'un seul groupe de liaison. Pour avoir un seul groupe de liaison, il faudrait avoir un couple de marqueurs liés à 0,8 puis tous les autres marqueurs avec un DL supérieur à 0,1 avec au moins un marqueur. Sachant que la distance médiane entre les marqueurs avec un DL supérieur ou égal à 0,1 est de 0,89cM, et qu'on considère que le génome du blé fait environ 3 060cM (sur la base d'en moyenne 0,18cM/Mb sur le chromosome 3B [Saintenac et al., 2009]), il suffirait de 4 438 marqueurs bien répartis sur le génome. Le nombre de marqueurs utilisés est proche de ce chiffre mais les marqueurs ne sont pas répartis de manière homogène sur tout le génome.

En conclusion, l'utilisation du DL dans les populations naturelles, très intéressantes pour leur diversité génétique, ou chez les espèces pour lesquelles aucune information de cartographie n'est disponible, pourrait donner une idée de l'ordre d'un certain nombre de marqueurs si une correction fiable est trouvée. En effet les populations naturelles sont le résultat de plusieurs générations de brassage de nombreux fondateurs au cours desquelles la sélection naturelle a créé de la structure. Une correction par l'apparentement est donc fortement conseillée. Une correction supplémentaire serait nécessaire par exemple avec le DL initial quand il est connu.

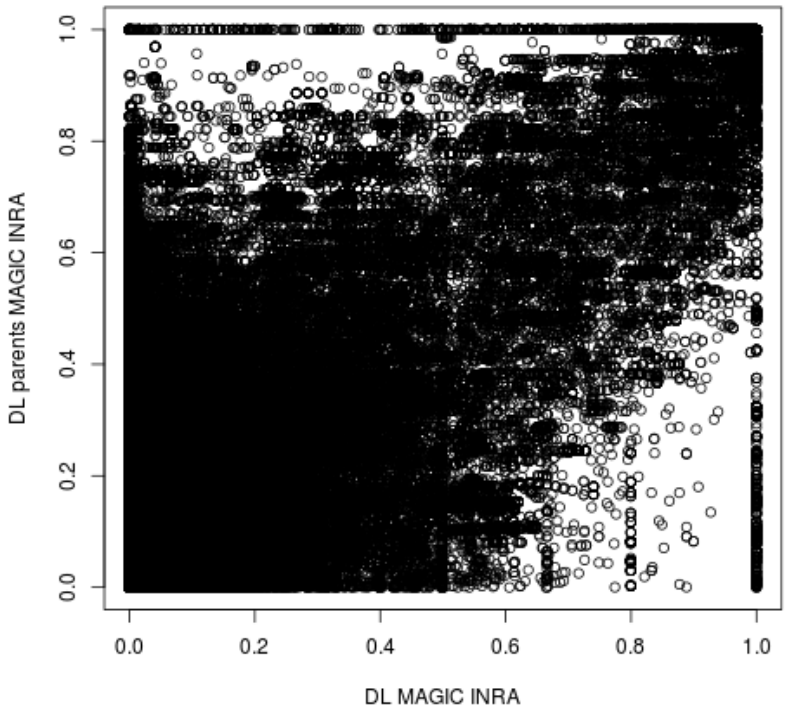


FIGURE 3.18 – Relation entre le DL des parents et le DL des descendants de la MAGIC INRA.

Pour avoir une idée de la qualité des données pour construire une carte fiable, nous pensons qu'un indice ("Pred") dérivé de l'indice de qualité des matrices de DL utilisé dans l'algorithme ("sj", Equation 3.8) corrigé par le nombre de marqueurs présents dans cette matrice (Equation 3.9) pourrait être utilisé comme prédicteur de la pertinence des données (Figure 3.19).

D'autre part dans certains cas, même si une carte génétique est disponible, les populations naturelles ont souvent plus de marqueurs polymorphes que les populations utilisées pour construire les cartes génétiques [Skøt et al., 2005], certains marqueurs ne seront donc pas cartographiés. L'approche pourrait donc être utile pour rajouter des marqueurs sur une carte connue ou ordonner des marqueurs positionnés sur le même point génétique.

$$Pred = \frac{sj}{\frac{n \times (n-1)}{2}} \tag{3.9}$$

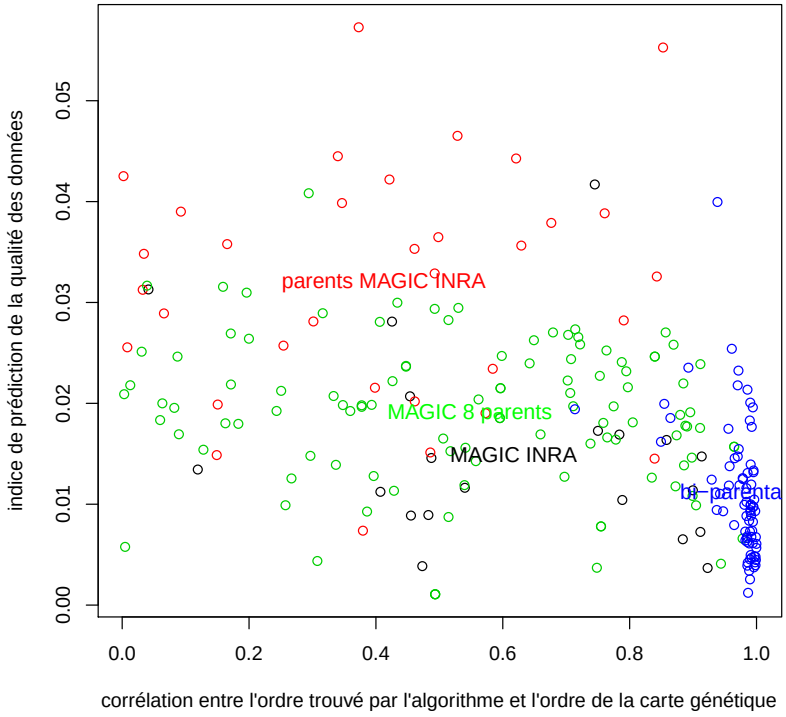


FIGURE 3.19 – Relation entre l'indice de qualité "Pred" (Equation 3.9) et la corrélation entre l'ordre trouvé par l'algorithme et l'ordre de la carte génétique de référence, pour les différentes fenêtres utilisées dans le chapitre "Résultats" (Tableau 3.3) avec un nombre de marqueurs minimum de 10. Les couleurs représentent les types de population. Le nom des populations montre le point moyen de corrélation et de l'indice de prédiction pour cette population.

Chapitre 4

Etude de l'évolution de la population durant les douze générations de panmixie

Ce chapitre décrit la structure génétique de la population MAGIC INRA, en s'intéressant particulièrement à son évolution au cours des 12 générations de panmixie. L'analyse porte sur le déséquilibre de liaison, la structure génétique, les zones génomiques soumises à sélection ainsi que l'évolution phénotypique de la précocité de floraison. Il est composé de deux parties : un manuscrit soumis au journal "Molecular and Biology Evolution" décrivant finement l'évolution à l'échelle moléculaire et l'évolution de la précocité de floraison avec un semis d'automne, et un manuscrit intégré dans les actes de l'"International wheat genetics symposium" 2013, décrivant l'évolution de la proportion de génotypes de type printemps dans la population. Le texte ci-après résume les principaux résultats de ces deux manuscrits.

L'étude a été réalisée sur une population de 436 individus (56 parents + 380 lignées SSD (Chap 2.1.2)). Un phénotypage de la précocité de floraison a été réalisé sur deux ans ainsi qu'un génotypage avec la puce 9K iSelect [Cavanagh et al., 2013] complétées par 14 marqueurs localisés dans des gènes candidats (Tableau 2.3). A partir de la puce 9K iSelect, qui comporte 8 632 marqueurs SNPs, 7 270 ont présenté un signal exploitable, correspondant à une ségrégation bi-allélique. Parmi ces marqueurs, 89% étaient polymorphes, dont 77% apportaient une information spécifique, car 12% des SNPs avaient une ségrégation totalement redondante avec d'autres marqueurs. L'étude a donc utilisé 5 635 marqueurs polymorphes et ne présentant pas mutuellement de déséquilibre de liaison (DL) total (5 621 marqueurs de la puce 9K iSelect + 14 marqueurs localisés dans des gènes candidats).

Dans le but d'étudier l'évolution de la population lors des 12 générations de panmixie, et en l'absence de connaissances sur la population initiale, nous avons dans un premier temps estimé la contribution réelle des 56 parents dans la population évoluée (Figure 2.1), grâce au développement d'une méthode Bayésienne basée sur les fréquences alléliques. La population initiale que nous avons considérée est donc une population composée des 56 parents pondérés par leur contribution estimée dans la population évoluée. Cette contribution est autour de 33% pour Probus, le donneur de stérilité mâle, et comprise entre 0,008% et 3,8% pour les autres parents. Cette estimation est cohérente avec nos connaissances du plan de croisements (Chap 2 et Annexe A).

La diversité génétique estimée dans la population MAGIC INRA est apparue

assez élevée étant donné le nombre de marqueurs polymorphes mais la forte proportion de marqueurs présentant un allèle rare (fréquence allélique de l'allèle mineur inférieure à 0,05) diminuait fortement cette diversité globale. La présence d'allèle rare était attendu connaissant le plan de croisements de la population. Tous les allèles spécifiques d'un parent excepté Probus ont une fréquence initiale d'environ 1%. La diversité génétique présente dans la population initiale et dans la population évoluée sont très similaires (MAF=0,18 vs 0,17 ; He=0,25 vs 0,24), ce qui reflète également l'étape d'inférence de la population initiale. Cette diversité initiale est inférieure à la diversité du pool parental non pondéré (MAF=0,27 ; He=0,34), essentiellement à cause de Probus qui compte pour un tiers dans la population initiale.

Les études de la structure de la population et du DL dans les deux populations ont montré que les 12 générations de panmixie ont cassé la structure présente initialement dans les lignées parentales et que le DL a très fortement diminué à longue comme à moyenne distance (inférieur à 10cM).

Le développement d'un algorithme pour détecter les marqueurs soumis à sélection, a permis la mise en évidence de 57 marqueurs qui présentaient une évolution peu probable sous l'unique effet de la seule dérive. Ces 57 marqueurs représentent 26 régions génomiques indépendantes localisées sur 15 des 21 chromosomes du blé (1A, 1B, 2A, 2B, 2D, 3A, 3B, 3D, 4A, 4B, 5A, 5B, 5D1, 6A et 6B). Parmi ces régions, une zone soumise à sélection est localisée sur le bras court du chromosome 4B, localisation du gène de stérilité *ms1* [Driscoll, 1975] utilisé pour transformer la population de blé naturellement autogame en une population allogame. Dans cette région un marqueur en particulier semble être localisé à proximité de ce gène, l'allèle présent chez Probus ayant une fréquence quasiment nulle dans la population évoluée, ce qui est cohérent avec la forte contre-sélection de la région génomique en déséquilibre de liaison avec le gène *ms1* lors de la fixation des lignées SSD (l'autofécondation d'un génotype *ms1b/ms1b* ne produisant aucun descendant).

Une partie importante des analyses s'est ensuite attachée à décrire les relations entre certaines régions génomiques soumises à sélection et la précocité de floraison. En effet, à l'échelle phénotypique, une évolution significative de la précocité de floraison en condition de semis d'automne et de la proportion de génotypes de type printemps a été observée : la population évoluée a fleuri en moyenne 128dj plus tôt que la population initiale et a une proportion de génotypes de type printemps qui a augmenté de 20% à 47%.

Pour relier ces évolutions phénotypiques aux zones soumises à sélection, nous avons testé l'association entre marqueurs dans les zones identifiées comme soumises à sélection et phénotypes dans la population MAGIC INRA. Cinq zones génomiques se sont avérées associées à la précocité de floraison avec un semis d'automne. Parmi elles se trouve le marqueur localisé dans le gène candidat *Ppd-D1*, gène majeur impliqué dans la sensibilité à la photopériode [Beales et al., 2007]. Ce marqueur explique 56% de la variabilité phénotypique de la population MAGIC INRA et 53% de l'évolution phénotypique observée entre la population initiale et la

population évoluée. Tous les autres marqueurs détectés sous sélection et associés à ce caractère expliquent seulement quelques pourcents de l'évolution. Une deuxième zone a été repérée sur le bras long du chromosome 4B, et pourrait correspondre à la zone contenant le gène *Vrn2B* [Yan et al., 2004b] impliqué à la fois dans le contrôle du besoin en vernalisation et de la sensibilité à la photopériode. Pour le caractère "type printemps", seulement trois zones ont été trouvées associées. Elles sont localisées sur les chromosomes 4A et 5D et expliquent 2,8% de la variation phénotypique. Le chromosome 4A n'est pas connu pour être impliqué dans la régulation du caractère printemps/hiver [Le Gouis et al., 2011] et les marqueurs soumis à sélection du chromosome 5D ne sont pas en DL avec le gène majeur *Vrn1D* localisé sur ce chromosome [Fu et al., 2005]. Etonnamment aucun des gènes majeurs de la famille *Vrn1* n'a été détecté sous sélection alors qu'ils sont très fortement associés et que les haplotypes aux différents marqueurs localisés dans ces gènes ($Vrn1A_{Prom}, Vrn1A_{ex7}, Vrn1D$) expliquent 30% de la variation phénotypique.

Ces deux caractères ne présentent pas le même « pattern » d'évolution à l'échelle génomique. Dans le premier cas, un gène majeur a été détecté sous sélection avec quatre autres zones ayant des effets plus faibles, alors que pour le caractère printemps/hiver seulement quelques marqueurs ont été détectés et expliquent très faiblement l'évolution phénotypique observée. Dans les deux cas, l'évolution phénotypique résulte du changement de fréquence sur différents marqueurs : soit de fortes variations alléliques, soit éventuellement des variations faibles, mais pour des marqueurs ayant des effets majeurs sur le caractère, comme dans le cas de *Vrn1* [Le Corre and Kremer, 2012]. Le fait que *Vrn1* n'ait pas été détecté sous sélection dans la population MAGIC INRA, alors que son rôle dans l'évolution d'autres populations autogames avait été clairement démontré [Rhoné et al., 2008, 2010], soulève la question du rôle de l'allogamie, qui limite l'emprise de la sélection sur les combinaisons alléliques épistatiques, comme celles présentes dans les haplotypes de *Vrn1*, du fait du brassage génétique à chaque génération.

L'étude des évolutions phénotypiques et moléculaires de la population MAGIC INRA, issue de 12 générations de panmixie, nous a permis de détecter des zones génomiques déjà décrites [Le Gouis et al., 2011] mais aussi de nouvelles zones génomiques impliquées dans le contrôle de la précocité de floraison (besoin en vernalisation ou précocité post vernalisation). Cette étude montre la complémentarité entre analyse des évolutions temporelles et étude par génétique d'association. L'absence de structure génétique dans la population MAGIC INRA, le niveau élevé de polymorphisme et le très faible DL font de cette population une ressource originale pour la détection de QTLs.

4.1 A pamictic experimental wheat population to detect markers under selection associated with earliness

Molecular Biology and Evolution

A panmictic experimental wheat population to detect markers under selection associated with earliness

Journal:	*Molecular Biology and Evolution*
Manuscript ID:	Draft
Manuscript Type:	Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Thépot, Stéphanie; UMR 0320/ UMR 8120 Génétique Végétale, Goldringer, Isabelle; UMR 0320/ UMR 8120 Génétique Végétale, Restoux, Gwendal; Unité d'Ecologie, Systématique et Evolution - CNRS UMR8079, Université Paris-Sud, Hospital, Frédéric; INRA, GABI Gouache, David; Arvalis, Institut du Végétal, Mackay, Ian; NIAB, Research Group Enjalbert, Jérôme; UMR 0320/ UMR 8120 Génétique Végétale,
Key Words:	QTL detection, temporal evolution, parental contribution, recombinant population, selection detection, experimental evolution

Article, Methods

A panmictic experimental wheat population to detect markers under selection associated with earliness

Stéphanie Thépot*[1], Isabelle Goldringer[2], Gwendal Restoux[3], Frédéric Hospital[4], David Gouache[5], Ian Mackay[6], Jérôme Enjalbert[2]

[1] Univ Paris-Sud, UMR 0320 / UMR 8120 Génétique Végétale, F-91190 Gif-sur-Yvette, France
[2] INRA, UMR 0320 / UMR 8120 Génétique Végétale, F-91190 Gif-sur-Yvette, France
[3]Unité d'Ecologie, Systématique et Evolution - CNRS UMR8079, Université Paris-Sud, Orsay, France
[4] INRA, UMR 1313 Génétique Animale et Biologie Intégrative, F-78352 Jouy en Josas, France
[5] Arvalis, Institut du Végétal. Station Expérimentale, F-91720 Boigneville, France
[6] NIAB, Huntingdon Road, Cambridge CB3 0LE, UK

Email: Stéphanie Thépot - stephanie.thepot@gmail.com,

Keywords: QTL detection, temporal evolution, parental contribution, recombinant population, selection detection, experimental evolution

Abstract

Advanced intercross populations are new tools used for QTLs discovery, potentially allowing finer mapping of a higher number of loci. Here we explored the interest of an evolved wheat population, recombined during 12 generations using nuclear male sterility (*ms1b*), derived from a population built through crosses between 60 parents. Available parents and a subset of 380 SSD lines were phenotyped for earliness and genotyped with the 9K i-select SNP array.

We showed that the 12 panmictic generations reduced LD to a very low level even at short distance and broke the population structure exhibited among the parents. We developed a Bayesian method based on allelic frequency to estimate the contribution of each parent in the evolved population. To detect loci under selection and estimate selective pressure, we developed a new method using shifts in allelic frequency between the initial and the evolved populations, and considering the effect of genetic drift. This evolutionary approach allowed us to identify 26 genomic areas under selection. One of these areas matched the location of the male sterility gene, while another corresponded to *Ppd-D1*, a major gene involved in the photoperiod sensitivity. This significant *Ppd-D1* allelic frequency shift is consistent with the observed shift in earliness, with the evolved population flowering earlier than the initial population. Using association tests between earliness and polymorphisms within the selected areas, we found four additional regions that appeared to carry earliness QTLs.

The interest of this new outcrossing population, mixing numerous initial parental lines through multiple generations of panmixia, is discussed regarding the power to detect new genes under selection, as well as association mapping.

Introduction

Adaptation is crucial for population survival, as a population interacts with a biotic community and an abiotic environment generally subject to rapid changes. Assuming that short-term adaptation mostly relies on preexisting variability at fitness traits, the analysis of the genetic architecture of these underlying traits is a matter of intense research. One key approach is the use of experimental evolution (Bataillon et al. 2010)where a population (or a set of populations) is grown in a given environment to study a specific response to natural or artificial selection (David 1992; Brachi et al. 2010). In crops many experimental populations are built by breeders and geneticists, mostly for the purpose of QTL (Quantitative Trait Locus) mapping, such as RILs from a bi-parental cross, NAM and MAGIC populations(Cavanagh et al. 2008). These populations cannot be considered as experimental evolution material, because they are built in order to avoid any selection. One notable exception is the dynamic management of genetic resources populations, which aim at the conservation of adaptive potential of crops through permanent cultivation of numerous populations in contrasted locations.

The theoretical basis of dynamic management relies on the metapopulation theory (Olivieri et al. 1990), which explains how adaptive diversity is maintained by evolutionary forces in connected populations. Many practical questions arise when applying this concept to the management of crop diversity, starting with the amount of initial genetic diversity to input in such network in order to allow sufficient adaptability to evolve. Rapid evolution is possible only in populations with initial genetic diversity. The number of founders and genetic distance among them give an idea of the initial genetic diversity, fuel for evolution over generations in response to selection and genetic drift. The size of the population at each generation is one of the main factors which acts on the effect of genetic drift relative to selection pressure: the smaller the population, the higher the genetic drift, thus requiring a higher selective pressure to fix low frequency positive alleles(Robertson 1960). The number of populations in the network and gene flow between them are also important factors affecting the maintenance of adaptive diversity (Gilpin 1991; Wang and Caballero 1999; Wang and Whitlock 2003; van Heerwaarden et al. 2009). Finally the mating system also affects the dynamic of adaptation (Bürger 1999). Indeed selfing enhances the efficiency of selection through increasing parent/offspring correlation. However it jointly increases inbreeding of populations leading to reduce effective size making them more prone to suffer from genetic drift (Charlesworth and Charlesworth 1995).

Very few applications of the concept of dynamic management have been developed so far: They all demonstrated phenotypic and molecular evolution in response to local selective pressures, coupled with a good preservation of diversity (Porcher et al. 2004; Enjalbert et al. 2011), as soon as population size is not critically small (Lavigne et al. 2001). Experiments on wheat populations exemplified both

the potential benefits and drawbacks of such dynamic management of genetic resources. In particular, plant competition quickly led to a general increase in plant height (Raquin et al. 2008; Enjalbert et al. 2011) and adaptation to local climatic conditions led to a divergence in earliness between populations cultivated in contrasting environments (Goldringer et al. 2006; Rhoné et al. 2008). Despite the large genetic diversity and census size, these studies stressed a low effective size due to selection that eroded genetic diversity (Enjalbert et al. 1999; Goldringer et al. 2001).

The different adaptive response of dynamic management populations to natural/human selective pressures provides opportunities to detect genomic areas involved in local adaptation (Allard 1988; Goldringer et al. 2001; Porcher et al. 2004). The main approaches used to detect selection assumed that it acts on restricted genomic regions. These methods, called neutrality tests, are thus aiming to detect selective sweeps (Maynard Smith and Haigh 1974), revealed by a large allelic frequency shift of some genes relative to the whole genome (Nielsen 2005). Shifts can be estimated using *Fst* computed in a sliding-window along the genome comparing populations evolving independently in time or space (i.e. different age or locations). Presence of strong initial linkage disequilibrium (LD), LD hereafter, enhances the capacity to detect selection (Lavigne et al. 2001; Rhoné et al. 2007; Raquin et al. 2008). Molecular detection of selection is often combined with phenotypic data using association genetics (Mackay 2001). These studies are common to detect local adaptation or acquisition of disease resistance by comparing populations in time and/or space (Allard 1988; Rogers and Bernatchez 2005; Rhoné et al. 2008, 2010).

The purpose of this study was to describe the specific genetic properties of an evolved wheat dynamic management population (*Triticum aestivum*), and to demonstrate how it can be a valuable tool for gene discovery, using flowering earliness as example. The studied population had three specific features: i) a high number of founders (60 lines), ii) a mating system modified from predominantly selfing to strict outcrossing based on genetic male sterility, and iii) a long term evolution with 12 generations of open pollination. First we described the evolution of its genetic diversity through the description of parental and population genetic structure using the 9K i-select SNP array. Then, due to the partially uncontrolled crossing scheme, we developed a Bayesian method to infer the contributions of parents to the resulting population. Third we estimated effective size of the population, analyzed the level of LD and tested for possible selected loci using a large set of SNPs after development of a new test for selection. We then investigated the evolution of flowering time, using both phenotypic and genomic variations to detect loci under selection. Finally we discussed the use of such panmictic multiparental populations to study the genetic bases of local adaptation.

Results

The INRA MAGIC population

The studied population is derived from random crossing of 60 wheat breeding lines (Trottet 1988, Table S1). This random crossings were facilitated by the integration of a nuclear sterility allele (*ms1b*, McIntosh 1988) present in the Probus line (Fossati and Ingold 1970) (Figure 1). During 12 generations male-sterile individuals (*ms1b/ms1b*) were tagged at flowering time, naturally wind pollinated by fertile plants then harvested at maturity. At the end of the 12th generation, the population was sampled for fixation through the production of 1 000 'S4' lines by single seed descent (SSD) (Figure 1).

In this study, we analyzed 56 parental lines including Probus (4 were no longer available (Table S1)), as well as a subset of 380 of the 'S4' SSD (Single Seed Descent) lines chosen to represent the phenotypic population diversity, called the evolved population hereafter.

Molecular diversity

The genotyping of the 436 wheat lines (56 parents + 380 SSD lines) with the 9K iselect SNP assay provided a dataset of 8 632 SNPs. After removing SNPs failing to generate clear genotype clustering, 7 270 SNPs (84.2%) with high quality genotype were kept. Among these SNPs, 6 476 (namely 74.7%) were polymorphic. This polymorphism rate was reasonably high, compared to the 100% rate observed with a worldwide panel on the same SNP array (Cavanagh et al. 2013). After removing monomorphic markers and pairs of markers which displayed exactly the same segregation over the 436 wheat lines, the final dataset consisted in 5 621 unique polymorphic SNPs for the parental and the evolved populations.

Structure analyses

The first two axes of the Principal Component Analyses (PCA) on parental population explained 17.1% of genetic variance. The K-means clustering on PCA axes revealed that the optimal genetic structure of parental lines consisted in two groups. According to the DAPC, these two groups mainly separated European wheat lines (France, Bulgaria, Germany, Great Britain, Netherlands Poland and Switzerland), from non European ones (United States, South America, URSS, Australia, Japan and Brazil) (Figure 2). With 81% of non European in one group and 93% of European in the other, the χ^2 test was highly significant (p-value = $3.5.10^{-6}$) and confirmed this hypothesis. This is in agreement with the clear division found between European and Asian origin lines in the known diversity in worldwide wheat varieties (Balfourier et al. 2007). AMOVA on K-means groups showed that these two groups explained 15% of the parental genetic diversity. Increasing the number of groups, led to discriminating six groups with known related pedigrees, in particular one group with all US lines, and

another group with lines related to the prebreeding line VPM. The first two axes of the PCA explained 17.1% of genetic diversity (Figure 2).

Both AMOVA and PCA results showed a very low structure in the evolved population. The K-means clustering on PCA axes also revealed two groups that explained 4.2% of the variation according an AMOVA. The first two axes of PCA explained 5.09% of the genetic variance.

These results showed that the structure exhibited among parental population has been broken after the 12 panmictic generations leading to a very low structure in the evolved population.

Parental contributions estimation

Because of open-pollination during population creation, we do not know with certainty the real contribution of each parent to the genome in G0. Thus we developed a new Bayesian method to estimate them. Analyses were performed at the genome-wide and at the chromosome levels (See Materials and Methods).

The genome-wide estimates were peaked having their 95% credible interval including in the interval mean +/- 0.1% for each parent, but varied widely over parents. Probus (n°38) had the highest contribution (35%), followed by Talent (n°43) and TJB240 (n°45) with contributions about 6%. The other parents had much lower contributions, ranging from 0.008% to 3.8% (Figure 3).

The contribution of Probus in the initial cross was estimated at 35% using a Bayesian approach. This high value is due to its over-representation during *ms1b* sterility introduction. This estimation is compatible with the expectations under a balanced contribution between selfed and back-crossed F1 in initial crosses (Figure 1).

We observed strong variations of parental contributions among the 55 other parents. Differences between parental contributions may be due to differences in reproductive success during initial crosses and/or during the following cycles of population evolution. However, uncertainty in estimates of some parental contributions can also be due to a lack of specific alleles for these lines. It is to be noted that five genotypes had a very low estimated contribution ($<10^{-4}$, n°11, 30, 40, 41 and 48), indicating that these lines were mainly incompatible with alleles present in the evolved population.

Discrepancies between contributions estimated at single chromosome level or genome-wide were low, with on average a difference of 1.4% in mean estimates. However, about 40% of the comparisons were significantly different (Bonferroni threshold), showing slight variations in parental contributions from given chromosomes (Table S4). D genome chromosomes exhibited a lower number of significant tests (chromosomal mean of comparisons 18%, Table S4) despite their higher absolute differences. This reflects lower power due to a lower number of markers on the D genome leading to flatter

distributions and thus larger credible intervals. In contrast among chromosome with a large number of markers, most parental contributions with 1B, 2A, 2B, 3A and 4B chromosomes were significantly different from the genome-wide contribution (chromosomal average comparisons respectively 57%, 57%, 57%, 66% and 58%), which could be due to the effect of selection at the chromosomal level. Among parents, Probus and TJB 240 varied the most with more than 15 chromosomes significantly different from their genome-wide estimates, while Lagoa Vermelha and Condor had only one (Table S4). This difference in significant comparisons between parents could mean that Lagoa Vermelha and Condor have more neutral alleles regarding the selection applied on the INRA MAGIC population. Hereafter, only results obtained with the genome-wide Bayesian method are presented.

Evolution between the initial and the evolved populations

Due to the Bayesian inference step, the diversity of the inferred initial population (MAF: 0.18, *He*: 0.25) and of the evolved population (MAF: 0.17, *He*: 0.24) were similar, with however 160 markers being polymorphic in one population, and monomorphic in the other. Overall, 129 markers turned to monomorphic in the evolved population, this loss being either due to non-detection in the sample, or to fixation due to genetic drift/selection, especially when initial MAF was low. Reversely, 31 markers with new alleles were detected in the evolved population, and could be explained by i) the four lost parents missing from our parental panel, ii) potential evolution in some of the initial founders (advance breeding material) presenting an initial residual heterozygosity, then lost during the management procedure in genebanks (regeneration through selfing cycles) (Esquinas-Alcazar 2005), iii) contamination, despite the isolation of plots (Hucl and Matus-Cadiz 2001) and iiii) mutation, as found for SSR markers in another wheat experimental population (Raquin et al. 2008). The observed heterozygosity in the evolved population was 3.2%. This is significantly higher than predicted after four selfing generations starting from an initial *He* of 0.24 (1.6%). A possible explanation to this difference could be a fitness advantage of heterozygote individuals also known as heterosis. The *Ne* estimated from *Fc* was of 310.97, much lower than the demographic population size (Ne_d = 7 500).

Linkage disequilibrium (LD) decay as a function of genetic distance was compared between the initial and the evolved populations. The LD between independent markers, or between chromosomes, was much lower in the evolved population, due to the number of recombinations (r^2=0.05 vs 0.003). The short-distance LD was less impacted: for markers at 2.5cM, it decreased from 0.32 in the initial population to 0.28 in the evolved population (Figure 4). The plateau is reached at a distance of about 50cM both for the initial and the evolved populations. However this plateau reached a lower value in the evolved than in the initial population, indicating a higher relatedness in the latter one whereas it was almost null in the former one.

Detection of trace of selection

Among 5 621 markers, 57 markers representing 26 independent genomic areas were detected as significantly under selection (i.e., behaving non-neutrally), with a minimum likelihood ratio of 12.45 (Table S3). The selection coefficients ranged from 0.07 to 0.7 (Table S3). The selected markers were located on different chromosomes (1A, 1B, 2A, 2B, 2D, 3A, 3B, 3D, 4A, 4B, 5A, 5B, 5D1, 6A, 6B), but mainly on 2B, 4B, 5A, and 6B (Figure 5). Only one of the studied candidate gene for flowering time was found under significant selection (Ppd-D1: Figure 5).

Heading date was highly heritable, with heritability of 0.95 and of 0.96 in the two years of experimentation. The average heading date of the initial population was 128dd later than that of the evolved population (Figure 6). Heading date variability was larger in the initial population than in the evolved one.

The markers detected as under significant selection were tested for their association with heading date. Two markers located on chromosome 2D were strongly associated with heading date: *Ppd-D1* (p-value = 5.10^{-70}), and "wsnp_CAP11_c3842_1829821" (p-value = $3.5.10^{-11}$) (Figure 5 & 6). *Ppd-D1* explained 56% of phenotypic variation in a single marker model (Figure 6). The frequency of its allele associated with earlier flowering at this locus (insensitive photoperiod allele) increased by 0.58 between the initial and the evolved populations. With a 117dd difference between the two Ppd-D1 alleles (Figure 6), this allelic frequency variation accounted for 53% of the total phenotypic evolution ($\frac{0.58\times117dd}{128dd}$). "wsnp_CAP11_c3842_1829821" explained 12% of phenotypic variation and its insensitive photoperiod allele frequency increased by 0.21 between the initial and the evolved populations. The difference between the effects of the two alleles being 103dd, allelic frequency variation at this marker accounted for 17% of the total phenotypic evolution ($\frac{0.21\times103dd}{128dd}$). The LD between these two loci was low but significant (r^2=0.18, p-value=0.002).

Discussion

We analyzed an evolved population derived from an experimental population with a very broad genetic basis (60 diverse parents, Table S1). To our knowledge, this is the first long term use of nuclear male sterility to modify a plant reproductive biology, i.e. turning wheat from selfing to outcrossing during twelve generations (except in recurrent selection population (Mackay and Gibson 1993; Kannenberg and Falk 1995). The specific characteristics of the population and of its history led us to develop different methods to describe its evolution, and the genetic basis of this evolution.

Consistently with previous studies (Allard 1988; Le Boulc'h et al. 1994; Goldringer et al. 2006), we found that earliness is a major trait submitted to selection in the dynamic management populations. This fast response to selection is in strait agreement with the large initial genetic variability and high

heritability of earliness. The evolved population flowered earlier which might be explained by at least three reasons. First, the viability of pollen might decrease in late flowering plants (higher temperature, Welsh and Klatt 1971). Second, earlier male flowering plants have a selective advantage as few female are already fertilized (Gérard et al. 2006), leading to less competition and higher reproductive success of early flowering plants. Third, there might be a bias during the male sterile plants tagging phase, with a higher tagging intensity in the beginning of the season as sterile spikes are easier to recognize. Such clear phenologic evolution was used as a model trait when analyzing genomic areas under positive selection.

The temporal variations in allelic frequencies gave two types of evidence that the evolved population was under the influence of selection during the twelve generations of intercrosses. Firstly, the effective size of the population (*Ne*=311) was very low compared to its demographic size (Ne_d=7 500). The ratio Ne/Ne_d = 0.04 was close but still higher than the one estimated by Goldringer et al. (2001) for selfing populations in dynamic management (Ne/Ne_d = 0.03, with *Ne* ranging from 40 to 150; (Enjalbert et al. 1999). Such higher *Ne* for an outcrossing population is expected, with a theoretical two fold difference due to the uncorrelated segregation of the two alleles present in each open-pollinated individual. Despite the remaining uncertainty about initial allelic frequencies, the striking gap observed between genetic and demographic sizes, is an indication of a strong selective pressure on dynamic management populations even though they are managed to avoid bottlenecks. However, other factors can explain the low Ne/N ratio, for example the variance in output of male and female gametes per plant, is probably higher than the theoretical Poisson distribution (Caballero 1994). Note that our method effectively estimates parental contributions as those fitting the allele frequencies in the evolved population best, and therefore minimizes the genetic divergence between the initial and the evolved populations. This should minimize the apparent impact of selection on allele frequencies, hence be conservative with respect to the detection of selection.

As a second source of evidence of selection, our test for non neutral temporal variations in SNP allele frequencies detected 26 genomic areas, mainly located on chromosomes 2B, 4B, 5A and 6B (Figure 5). Note that chromosomes 2B and 4B were among those for which many parent contributions estimated at the chromosome level diverged from their genome-wide estimation (Table S3), which can be seen as another indirect evidence of selection. Two markers under selection: *Ppd-D1* and "wsnp_CAP11_c3842_1829821", both located on chromosome 2D, were highly associated with earliness. However, a global model including both markers showed that the effect of "wsnp_CAP11_c3842_1829821" was only due to its small but significant LD with *Ppd-D1* (LD=0.18, p-value=0.002). Hence, *Ppd-D1*, which is a known candidate gene involved in photoperiod sensitivity, is probably the only selected gene in this area.

Besides *Ppd-D1*, five other markers exhibited a low association with heading date (two markers with p-value<0.05 and three markers with p-value<0.1, Table S3). Most were located on chromosome 5A. Assuming that allelic effects are additive, these five markers explained only 1% of the phenotypic shift in heading date. Note that two markers had an opposite effect in comparison to the phenotypic shift. Hence, the heading date evolution is mainly explained by one gene (*Ppd-D1)* and the remainder is due to numerous loci either with a strong effect on earliness and a low allelic frequency shift or with a significant allelic frequency evolution and a low effect (Kremer and Le Corre 2012). One additional area was detected as under selection at the end of the chromosome 4B (Figure 5), with a low but significant association with heading date (p-value<0.1, Table S3). This area could match with the location of *Vrn2B,* a known gene involved in an integrative pathway of vernalization requirement and photoperiod sensitivity but not mapped in our available map.

In the selfing populations of the dynamic management experiment (Rhoné et al. 2008, 2010), not only *Ppd-D1* showed a strong effect on the evolution of earliness, but a significant association was also found for additional candidate genes with major effects on flowering time, such as *Vrn1*, which were not detected here. *Vrn1* is known for its epistatic control of vernalization requirements (Rousset et al. 2011) and in these selfing populations, different combinations of alleles born by each of the three copies of *Vrn1* (*Vrn-A1*, *Vrn-B1* and *Vrn-D1*) have been selected. One hypothesis for this difference might be that, in outcrossing populations, selection is not efficient on epistatic combinations, because they are broken by recombination.

At the phenotypic level, we only studied earliness and found it associated with 24% of the genomic areas detected under selection (6 different areas, Table S3). The other markers found under selection and not associated with earliness could still be anyway involved in its regulation, as they could have reached near fixation in evolved population in response to selection, and therefore escape to association test (Luo 1998). Alternatively they could be involved in the regulation of other adaptive traits like disease resistance, height, etc...

Another area under selection, on the short arm of 4B chromosome, corresponds to the location of the sterility gene *ms1b* (Driscoll 1975). In this area, one marker exhibited an allelic frequency change from 0.37 in the initial population (allele present in 5 parents including Probus which contributed to 34% of the initial population) to 0.02 in the evolved population. In this 4B chromosome area, the Probus allelic frequency decreased dramatically for four markers; in contrast with the other markers in adjacent regions (mean of allelic frequency differences between the initial and the evolved population: 0.37 vs 0.04, data not shown). This is in full agreement with the expected evolution of male sterility gene *ms1b,* which was under stabilizing selection during the population evolution (75% frequency of

sterile allele, Doggett and Eberhart 1968), and then counter-selected during the SSD line fixation (sterility of *ms1b/ms1b* genotypes).

The introduction of the nuclear male-sterility gene, *ms1b*, transformed the wheat self-pollinating habit to outcrossing. The 12 generations of panmixia broke both the initial genetic structure and long distance LD, resulting in a very low level of LD in the evolved population (r^2=0.003 at 100cM). This is to be compared to a LD of 0.04 between independent marker in a 19-parent MAGIC population of *Arabidopsis thaliana* (Kover et al. 2009) or of 0.004 in a 4-parent MAGIC population of wheat (Huang et al. 2012). The LD in the INRA MAGIC experimental population is the lowest described so far for a wheat population, with r^2= 0.28 at a 2.5cM distance (Figure 3). As a comparison, r^2 decayed to 0.8 at 5cM for a 4-parent MAGIC population (Huang et al. 2012), or between 0.25 and 0.5 at 5cM for an association mapping panel (Cavanagh et al. 2013). The low value of short distance LD in the MAGIC INRA population is mostly due to the low initial LD within the set of 56 parents (R^2=0.3 at 2.5cM). This weight of parental diversity at very short distance LD (close to 0cM) is well illustrated in the 4-parent and 19-parent MAGIC populations, with LD values of 1.0 and 0.2 respectively (Kover et al. 2009; Huang et al. 2012). Thus the overall very low LD of the INRA MAGIC population is the result of both a low initial LD due to the high number of parents (and their relatively low relatedness) and a high number of recombinations breaking LD at medium-long genetic distances. This low LD, combined with the absence of structure and a large diversity, make this population of great value for fine mapping.

The MAGIC INRA population is similar in concept to other advanced intercross populations such as the mouse Collaborative Cross (Churchill et al. 2004), the Arabidopsis 19-parent MAGIC lines (Kover et al. 2009), the Arabidopsis 8 parent MAGIC lines (Huang et al. 2011), the Drosophila Synthetic Population Resource (DSPR) (King, Macdonald, et al. 2012), the rice 8-parent MAGIC lines (Bandillo et al. 2013) and the yeast 4-parent MAGIC lines (Cubillos et al. 2013). All of these populations result from some generations of intercrossing which allowed more precise detection of QTLs. They were mainly analyzed using traditional association mapping methods directly on the phenotype and genotype of lines (Huang et al. 2011; Cubillos et al. 2013) or on reconstructed parental haplotypes (Kover et al. 2009; King, Merkes, et al. 2012). Only Cubillos et al. (2013) complemented standard association mapping with an evolutionary approach allowing them to detect and map precisely a gene involved in heat stress resistance.

With an evolutionary approach, both at phenotypic and genetic levels, we successfully identified several genomic areas under selection, and many of them were associated with earliness. As the population exhibited a quite fast evolution in one environment, growing such diversified gene pools in different divergent environments could give access to more genetic areas associated with local

adaptation. In dynamic management, experiments have been so far performed without gene flow between populations, although gene flow would add genetic variability and could strengthen the detection of selected areas, as it is expected to homogenize neutral genomic areas (if properly tuned). In addition, the genetic basis of local adaptation could be analyzed through association genetics since the number of panmictic generations guarantees a very low LD and an absence of structure. Using more individuals (1 000 SSD lines are available (seed samples are available on request)) with a higher density of markers should confirm that these evolutionary mapping populations are an invaluable platform for trait discovery and validation in the future.

Materials and Methods

Biological material: the INRA MAGIC population

The population studied is derived from a very diverse composite population created between 1976 and 1980 (Trottet 1988), by crossing 60 European and worldwide wheat breeding lines selected for their resistance to diseases and their good agronomic values (Table S1). Each of the 59 parents was first crossed with the 60th, male-sterile variety Probus (Fossati and Ingold 1970), which was homozygous for the recessive nuclear sterility gene *ms1b* (McIntosh 1988). Among the F1 plants (*Ms1b/ms1b* = fertile), some were selfed (F2) while others were back-crossed (BC1) with the 59 parents to reduce the proportion of "Probus" genome in the population. The bulked seeds of the two progenies (F2 & BC1) were then sown in a mixed row design, in an isolated field surrounded by rye (*Secale cereal*) (Figure 1). Male-sterile individuals (*ms1b/ms1b*) were tagged at flowering time, naturally wind pollinated by fertile plants then harvested at maturity. Due to this open-pollination, the relative contribution of F2 vs BC1 plants was unknown, leading to uncertainty regarding parental contributions. The progeny of this first out-crossing cycle (G0 population) was re-sown in isolation and male-sterile plants harvested again. Such management stabilizes the proportion of male sterile plants to 50% as a result of the cross of male sterile (*ms1b/ms1b*) plants and male fertile (*Ms1b/ms1b*) plants.

This initial population has then been managed in Le Moulon (48.4°N, 21°E) for 12 generations, as part of the French dynamic management project (Henry et al. 1991). The population was grown in isolation without artificial selection. Each year, at least 10 000 seeds were sown and about 3 000 spikes from tagged male sterile plants were harvested and threshed together. At the 12th generation, the population was sampled for fixation, and 1 000 'S4' lines were derived by single seed descent (SSD) (Figure 1). Note that in total, these lines underwent 15 outcrossing cycles (3 for the creation of the original population, plus 12 at Le Moulon).

In this study, we analyzed 56 parental lines including Probus (4 were no longer available (Table S1)), as well as a subset of 380 of the 'S4' SSD lines chosen to represent a broad phenotypic diversity of the

population on the basis of a Principal Component Analysis (PCA) made on all phenotypic traits scored the first year. Extreme individuals and a random sample of in-between individuals were chosen. This subset of 380 lines is called the evolved population.

Phenotyping

Flowering time was assessed in field trials at Le Moulon over two seasons (2010-2011 and 2011-2012), with a November sowing. Each genotype (S4 SSD lines + parents) was observed in two replicates with 20 seeds per genotype sown on single row plots. For each line, the heading date was scored when 50% of the plants had half of the main ear emerged from the flag leaf. The heading date was transformed into sum of the mean temperatures per day (in degree-days (dd)) from sowing to heading, based on data recorded by the meteorological station located at Le Moulon.

Molecular analysis

Total DNA of each of the 436 lines (380 SSD lines + 56 parents) was extracted from 500mg of leaves of one plant. Extractions were performed using a modified procedure of Dellaporta et al. (1983) including a carbohydrate precipitation described in Michaels et al. (1994). Genotyping was performed by MJ Hayden's team at DPI Victoria in Bundoora, Australia, using a 9K i-select SNP array (Cavanagh et al. 2013). SNP allele clustering was performed using GenomeStudio software (http://www.illumina.com/software/genomestudio_software.ilmn). Errors in allele assignation by GenomeStudio were detected by visual inspection of SNP allele clusters and manually corrected. Only SNPs that could be unambiguously scored as biallelic were kept .i.e SNPs exhibiting three clusters or less.

Fourteen additional polymorphisms located in candidate genes involved in the earliness pathway (Table S2) were genotyped using the KASPar SNP genotyping system developed by KBioscience (http://www.kbioscience.co.uk/).

Population Structure analysis

Population genetic structure was analyzed independently for the parental and the evolved populations with a discriminant analysis of principal components (DAPC; Jombart et al. 2010). We determined the optimal number of genetic clusters using a K-means method applied on a Principal Component Analysis and inferred the most likely genetic clusters. The DAPC computation was made using the package adegenet ver. 1.3–6 (Jombart and Ahmed 2011) and the statistical program R ver. 2.15.3-1 (R Development Core Team 2013). Euclidian distance among genotypes was calculated using the procedure dist.gene in package "ape" (Paradis 2010). An Analysis of Molecular Variance (Excoffier et al. 1992) was performed to estimate the percentage of genetic variance explained by the structure determined by the DAPC analysis with the "pegas" R-package (Paradis 2010).

Estimation of parental contributions

As the contribution of parental lines to the evolved population was unknown due to open-pollination generations, we developed a Bayesian approach to estimate it. For each of the 56 available parental wheat lines, we computed the posterior probability of their relative contribution to the evolved population using the SNP allelic frequencies. The four unavailable parental lines were ignored since no genetic information was available; the 56 remaining parental lines are thus considered as the only contributors of the evolved population. We considered the SSD lines to be a finite sample of a theoretical infinite population (i.e. no genetic drift) resulting from many cycles of free recombination of a set of 56 parental lines without selection pressure, mutation, migration or linkage disequilibrium among markers.

The observations, n_{il}, were defined as the number of alleles A (arbitrarily chosen among the two A/B alleles of bi-allelic SNPs) observed in the SSD lines for each individual i at each locus l. The random variable N_{il}, defined as the number of alleles A of the individual i at the locus l can take three possible values: 0 or 2 when homozygote for allele A or B respectively and 1 if heterozygote. We considered no linkage disequilibrium; consequently the number of alleles A at each locus for each individual is independent. The likelihood function is thus defined as:

$$L = \prod_{i}^{I} \prod_{l}^{L} P(N_{il} = n_{il}),$$

with I the total number of individuals (i.e. the 380 SSD lines here) and L the total number of loci considered (i.e the number of markers). The probability function of N_{il} follows a binomial distribution leading to:

$$L = \prod_{i}^{I} \prod_{l}^{L} \binom{2}{n_{il}} p_l^{n_{ij}} (1 - p_l)^{2 - n_{ij}},$$

with p_l the probability for an individual to get the allele A at the locus l (i.e. the expected frequency of allele A at the locus l in the evolved population). Individuals and loci being independent in the model the above equation is exactly equivalent to:

$$L = \prod_{l}^{L} \binom{380 \times 2}{n_{.l}} p_l^{n_{.l}} (1 - p_l)^{380 - n_{.l}},$$

with $n_{.l}$ the total number of alleles A at locus l over all individuals. In others terms this model only relied on allele frequencies in the evolved population and not on their associations (i.e. no LD), it is thus capable to cope with the numerous selfing events necessary to obtain SSD lines. The probability

p_j depends on the expected contribution of each parental line k, $P(Par = k)$, and on the frequency of allele A in each parental line k, $f(A)_k$. We can thus formulate p_j as follows:

$$p_j = \sum_{k}^{K} f(A)_k P(Par = k),$$

with K the total number of parental lines (i.e. K=56 in this study).Genetic information is available for all putative parental lines K thus:

$$\sum_{k}^{K} P(Par = k) = 1.$$

Furthermore $P(Par = k)$ is the same for all the loci considered during estimation. Finally because no previous information was available on the virtual contribution of each putative parental line k, we considered a flat Dirichlet prior distribution,

$$P(Par = 1 \dots K) \sim Dir(\alpha),$$

with $\alpha_k = 1$ for each of the K elements of the vector α. We computed genome-wide estimations of the posterior distributions of $P(Par = k)$, considering the whole set of markers (5590 markers, polymorphic in parental and the evolved populations), $P_{gw}(Par = k)$, or at chromosomal level considering subsets of markers for each chromosome c, $P_c(Par = k)$ (5056 markers mapped).

The posterior distributions were estimated with a MCMC method using the Gibbs sampler (i.e. a particular case of the Metropolis-Hastings algorithm) implemented in JAGS (Plummer 2003). Three independent Markov chains ran for 5 millions iterations for genome-wide estimates (56 parental lines x 3 chains x 5.10^6 iterations) or 1 million iterations for estimates at the chromosome level (3x8=22 chromosomes x 3 chains x 10^6 iterations). To avoid strong temporal correlation between successive elements within the Markov chain only 1 element over 5 was kept. Chains convergence was checked using a Gelman and Rubin test (1992). Analyses were conducted using R (R Development Core Team 2013) and the CODA package (Plummer et al. 2006).

To validate this estimation method we used simulated genotypes of 50 individuals based on 30 bi-allelic SNP markers. These individuals resulted from a generation of random mating with complete recombination and genetic drift, with initial known contributions of 6 parental genotypes. In all cases the simulated contributions were always close to the mean and to the mode of the posterior distribution of $P(Par = k)$ and all included in its highest posterior density interval (HPD). To test the robustness of this method to the assumption of linkage-equilibrium, we compared estimates performed with all markers on the chromosome 1A (442 markers) with estimates performed after removing markers in

linkage disequilibrium i.e by removing markers with a pair correlation above a 0.06 threshold (130 independent markers). Estimates of parental contributions with the two sets of markers only differed by a mean difference of 0.004 (max=0.02) showing the low impact of linked loci on estimates. We also analyzed other scenarios with different contributions of Probus to the initial population (25%, 33% or 50%) corresponding to increasing numbers of back-crossed F1 at the expense of selfed F1, always with balanced contributions of the 55 other parents. These scenarios confirmed the main conclusions of our study (data not shown).

Comparison between chromosomal and genome-wide contribution estimates

The contributions of parental lines to the evolved population were computed genome-wide, $P_{gw}(Par = k)$, and at the chromosome level, $P_c(Par = k)$. In absence of evolutionary forces, contributions of parental line *k* at both levels should be similar. An upward or downward deviation of the contribution of a parental line at the chromosome level relative to the genome-wide estimate, indicating that this contributor was favored or depreciated over the others, can result from selection acting on genes located within this chromosome. We defined a test based on the following statistic:

$$\Delta_{kc} = P_c(Par = k) - P_{gw}(Par = k).$$

In absence of selection the whole genome must behave similarly leading to $\Delta_{kc} = 0$ (i.e. null hypothesis), whereas it should differ significantly from 0 (positively or negatively) if parental line *k* is respectively over- or under-represented in the chromosome *c* relative to its genome wide contribution (i.e. alternative hypothesis). $P_c(Par = k)$ and $P_{gw}(Par = k)$ differed in their statistical power to estimate the contribution of parental lines due to the number of markers considered. Furthermore the number of iterations used to reach the convergence of Markov chains also differed. Thus to test ifΔ_{kc} differed from 0 we estimated its posterior distribution with 10 000 values of Δ_{kc}computed by randomly drawing within posterior distributions of $P_c(Par = k)$ and $P_{gw}(Par = k)$respectively. This allowed for coping with non-symmetric and/or multimodal posterior distributions of $P_c(Par = k)$ and $P_{gw}(Par = k)$, contrarily to simple mean difference test based on summary statistics (e.g. Student test). The significance was tested by computing $IC_\alpha(\Delta_{kc})$,the credibility interval of the posterior distribution of Δ_{kc}at the $(1-\alpha)$ level. If $0 \in IC_\alpha(\Delta_{kc})$ the null hypothesis was conserved whereas if $0 \notin IC_\alpha(\Delta_{kc})$ we rejected the null hypothesis with a risk α and thus considered Δ_{kc} to differ significantly from 0. We computed this test for all combinations of each parental line *k* and each chromosome *c*. Because of multiple comparisons to keep an overall $\alpha = 5$ level we adjusted the levelα_{kc} for each single test with a Bonferroni correction resulting in $\alpha_{kc} = \frac{\alpha}{K \times C} = \frac{0.05}{56 \times 22} = 4.0610^{-5}$.

The initial population for further analysis will then correspond to a population composed of the 56 parental lines weighted according to their contribution estimated with the genome-wide Bayesian method.

Temporal genetic evolution

To monitor the evolution of the genetic diversity, we estimated the average minor allele frequencies (MAF), the average expected heterozygosity (*He*, Nei diversity) and the observed heterozygosity for both the initial and the evolved populations.

We checked the linkage disequilibrium evolution by computing the Pearson correlation (r^2) between all pairs of loci (R Development Core Team 2013). The decay of LD with genetic distance was compared between the initial and the evolved populations using the map of the 9K iselect SNPs (Cavanagh et al. 2013).

Temporal variance of allelic frequencies (*Fc*) was computed by the standardized variance at each bi-allelic locus (Nei and Tajima 1981):

$$Fc_l = \frac{(Fi_l - Fe_l)^2}{\frac{Fi_l + Fe_l}{2} - Fi_l * Fe_l}$$

where Fi_l and Fe_l are the frequency at locus l in the initial and the evolved populations respectively.

The multilocus *Fc* was calculated as the average of the single locus Fc_l estimated over all markers of the 9K chip. Then the genetic effective size (*Ne*), assumed constant over time, was estimated following the temporal Waples (1989) method:

$$Ne = \frac{\Delta t}{2Fc - \frac{1}{2S_0} - \frac{1}{2S_t}}$$

where Δt is the number of generations of recombination between the initial and the evolved populations, S_0 the sample size of the initial population and S_t the sample size of the population at generation t, respectively 15, 56 and 380 in our case. Due to the inbreeding level, numbers of independent alleles sampled ($2S_0$ and $2S_t$) were approximated by the number of individuals (S_0 and S_t).

The demographic size (Ne_d) was estimated as the harmonic mean of the minimal true number of plants of each gender grown in the population (above 5 000 flowered male plants and 3 000 harvested female plants per generation) (Charlesworth 2009).

Detection of trace of selection

In order to determine whether the SNP allele frequencies were possibly driven by selection and drift (as opposed to drift alone) we used a maximum likelihood approach. Considering one of the alleles at a given SNP, let X_0 be the number of copies of this allele in the population at generation 0, and X_{15} be the number of allele copies at generation 15. Let Ne be the effective population size. Here, Ne was assumed to be known. Finally, let s be the coefficient of selection of the allele, such as the fitness of the genotypes "aa", "aA", and "AA" is 1, 1+s, and 1+2s, respectively (with $s \geq 0$). Here, "A" is the positively selected allele at the SNPs$s \geq 0$. Assuming these parameters are known, one can compute $Pr(X_{15} \vee X_0, s, Ne)$, the probability that the allele copy number at generation 15 is X_{15}, given that the allele copy number at generation 0 was X_0, with a selection coefficient s and an effective population size Ne. This probability was computed numerically by iterating a standard formulation of a Wright-Fisher model with genetic drift and selection, as described for example in Neuhauser (2001).

The effect of sampling of genotyped individuals was taken into account by setting Ne=n in the last generation when iterating the Wright-Fisher model, where n is the size of the sample of individuals genotyped at generation 15 (n=380 for the INRA MAGIC population). A similar sampling effect could also affect the estimate of initial allelic frequencies, but in our case we considered that the founders were extensively representing the initial population.

Now, for a given locus, we know (X_0, X_{15}, Ne), so the only parameter of the model is s, and $L(s) = Pr(X_{15} \vee X_0, s, Ne)$ gives the likelihood of the model. In practice, L(s) was maximized numerically for positives values of s ranging from 0 to 1, and for 'negative' values by symmetry. Negative selection was then implemented as:

$$\Pr(X_{15}|X_0, s, Ne) = \Pr(2N - X_{15}|2Ne - X_0, -s, Ne) for\ s < 0$$

Now, for given $\{X_0, X_{15}\}$, let s* be the value of s that maximizes L(s). A Likelihood Ratio Test (LRT) was computed as:

$$LRT(X_0, X_{15}) = -2\ln\left[\frac{L(s=0)}{L(s=s*)}\right]$$

The significance of the test was assessed by assuming that the LRT follows a chi squared distribution with one degree of freedom (Wilks' theorem) under the null hypothesis of absence of selection (i.e if s=s*=0). As we performed multiple tests, estimated p-values were transformed into q-values, which are measures of significance in terms of false discovery rate (FDR) rather than the false positive rate (Storey and Tibshirani 2003), using the "qvalue" R package (Dabney and Storey 2004). Markers were tested with a FDR threshold of 0.05.

Phenotypic analysis, differentiation and association tests

Broad sense heritability of earliness traits was assessed for each year using a first ANOVA model only including a replicate and a genotype effect. Experiment factor effects were then tested using the second following linear model:

$$Y_{ijk} = \mu + y_j + r(y)_{jk} + G_i + y_j \times G_i + \varepsilon_{ijk}$$

where G_i is the effect of the genotype i, y_j is the effect of the year j (2010-2011 or 2011-2012), $r(y)_{jk}$ is effect of the replicate k in year j, y_j x G_i is the interaction between year and genotype and ε_{ijk} is the residual error term. Genotype and genotype by year interaction effects were both declared as random effects. We estimated adjusted means over the replicate and year effects for each genotype for the evolved population. Markers were tested for association with flowering time phenotypes on individuals of the evolved population. The association was tested independently on heading date score for each marker.

Acknowledgements

The authors thank S. Pin, N. Galic and V. Sanchez and all trainers (Romain Angeleri, Caroline Biton, Katia DeBray, Julien Ratet and Josephine Thomas) for their technical help in the field experimentation and V. Combes for their help in DNA extraction. This work was financially supported by Arvalis institut du végétale for the phenotyping. This work is a part of S. Thépot's PhD supported by the Ministère de l'enseignement supérieur et de la recherche.

References

Allard RW. 1988. Genetic Changes Associated with the Evolution of Adaptedness in Cultivated Plants and Their Wild Progenitors. Journal of Heredity 79:225 –238.

Balfourier F, Roussel V, Strelchenko P, et al. 2007. A worldwide bread wheat core collection arrayed in a 384-well plate. TAG 114:1265–1275.

Bandillo N, Raghavan C, Muyco PA, et al. 2013. Multi-parent advanced generation inter-cross (MAGIC) populations in rice: progress and potential for genetics research and breeding. Rice 6:1–15.

Bataillon T, Dillmann C, Gaba S, et al. 2010. Evolution expérimentale. Biologie Evolutive:617–646.

Le Boulc'h V, David J, Brabant P, de Vallavieille-Pope C. 1994. Dynamic conservation of variability: responses of wheat populations to different selective forces including powdery mildew. Genet Sel Evol 26:S221–240.

Brachi B, Faure N, Horton M, Flahauw E, Vazquez A, Nordborg M, Bergelson J, Cuguen J, Roux F. 2010. Linkage and Association Mapping of Arabidopsis thaliana Flowering Time in Nature. PLoS Genet 6:e1000940.

Bürger R. 1999. Evolution of genetic variability and the advantage of sex and recombination in changing environments. Genetics 153:1055–1069.

Caballero A. 1994. Developments in the prediction of effective population size. Heredity 73:657.

Cavanagh C, Morell M, Mackay I, Powell W. 2008. From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. Current Opinion in Plant Biology 11:215–221.

Cavanagh CR, Chao S, Wang S, et al. 2013. Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. Proc. Natl. Acad. Sci. U.S.A. 110:8057–8062.

Charlesworth B. 2009. Effective population size and patterns of molecular evolution and variation. Nature Reviews Genetics 10:195–205.

Charlesworth D, Charlesworth B. 1995. Quantitative genetics in plants: the effect of the breeding system on genetic variability. Evolution:911–920.

Churchill GA, Airey DC, Allayee H, et al. 2004. The Collaborative Cross, a community resource for the genetic analysis of complex traits. Nat Genet 36:1133–1137.

Cubillos FA, Parts L, Salinas F, et al. 2013. High-Resolution Mapping of Complex Traits with a Four-Parent Advanced Intercross Yeast Population. Genetics 195:1141–1155.

Dabney A, Storey JD. 2004. Q-value estimation for false discovery rate control. Medicine 344:539–548.

David J. 1992. Approche méthodologique d'une gestion dynamique des ressources génétiques chez le blé tendre (Triticum aestivum L).

Dellaporta SL, Wood J, Hicks JB. 1983. A plant DNA minipreparation: version II. Plant molecular biology reporter 1:19–21.

Doggett H, Eberhart SA. 1968. Recurrent selection in sorghum. Crop Science 8:119–121.

Driscoll C. 1975. Cytogenetic Analysis of Two Chromosomal Male-sterility Mutants in Hexaploid Wheat. Aust. Jnl. Of Bio. Sci. 28:413–416.

Enjalbert J, Dawson JC, Paillard S, Rhoné B, Rousselle Y, Thomas M, Goldringer I. 2011. Dynamic management of crop diversity: From an experimental approach to on-farm conservation. Comptes Rendus Biologies 334:458–468.

Enjalbert J, Goldringer I, Paillard S, Brabant P. 1999. Molecular markers to study genetic drift and selection in wheat populations. J. Exp. Bot. 50:283–290.

Esquinas-Alcazar J. 2005. Protecting crop genetic diversity for food security: political, ethical and technical challenges. Nat Rev Genet 6:946–953.

Excoffier L, Smouse PE, Quattro JM. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479–491.

Fossati A, Ingold M. 1970. A male sterile mutant in Triticum aestivum. Wheat Information Service:8–10.

Gelman A, Rubin DB. 1992. Inference from iterative simulation using multiple sequences. Statistical science:457–472.

Gérard PR, Klein EK, Austerlitz F, Fernández-Manjarrés JF, Frascaria-Lacoste N. 2006. Assortative mating and differential male mating success in an ash hybrid zone population. BMC Evolutionary Biology 6:96.

Gilpin M. 1991. The genetic effective size of a metapopulation. Biological Journal of the Linnean Society 42:165–175.

Goldringer I, Enjalbert J, David J, Paillard S, Pham JL, Brabant P. 2001. Dynamic Management of Genetic Resources : a 13-year Experiment on Wheat. In: Broadening the genetic base of crop production.

Goldringer I, Prouin C, Rousset M, Galic N, Bonnin I. 2006. Rapid Differentiation of Experimental Populations of Wheat for Heading Time in Response to Local Climatic Conditions. Annals of Botany 98:805–817.

Van Heerwaarden J, van Eeuwijk FA, Ross-Ibarra J. 2009. Genetic diversity in a crop metapopulation. Heredity 104:28–39.

Henry JP, Pontis C, David J, Gouyon PH. 1991. An experiment on dynamic conservation of genetic resources with metapopulation. In Species conservation : a population biological approach.

Huang BE, George AW, Forrest KL, Kilian A, Hayden MJ, Morell MK, Cavanagh CR. 2012. A multiparent advanced generation inter-cross population for genetic analysis in wheat. Plant Biotechnology Journal 10:826–839.

Huang X, Paulo M-J, Boer M, Effgen S, Keizer P, Koornneef M, van Eeuwijk FA. 2011. Analysis of natural allelic variation in Arabidopsis using a multiparent recombinant inbred line population. Proceedings of the National Academy of Sciences 108:4488 –4493.

Hucl P, Matus-Cadiz M. 2001. Isolation distances for minimizing out-crossing in spring wheat. Crop Science 41:1348–1351.

Jombart T, Ahmed I. 2011. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics 27:3070–3071.

Jombart T, Devillard S, Balloux F. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC genetics 11:94.

Kannenberg LW, Falk DE. 1995. Models for activation of plant genetic resources for crop breeding programs. Canadian Journal of Plant Science 75:45–53.

King EG, Macdonald SJ, Long AD. 2012. Properties and power of the Drosophila Synthetic Population Resource for the routine dissection of complex traits. Genetics 191:935–949.

King EG, Merkes CM, McNeil CL, Hoofer SR, Sen S, Broman KW, Long AD, Macdonald SJ. 2012. Genetic dissection of a model complex trait using the Drosophila Synthetic Population Resource. Genome research 22:1558–1566.

Kover PX, Valdar W, Trakalo J, Scarcelli N, Ehrenreich IM, Purugganan MD, Durrant C, Mott R. 2009. A Multiparent Advanced Generation Inter-Cross to Fine-Map Quantitative Traits in Arabidopsis thaliana. PLoS Genet 5:e1000551.

Kremer A, Le Corre V. 2012. Decoupling of differentiation between traits and their underlying genes in response to divergent selection. Heredity 108:375–385.

Lavigne C, Reboud X, Lefranc M, Porcher E, Roux F, Olivieri I, Godelle B. 2001. Evolution of genetic diversity in metapopulations: Arabidopsis thaliana as an experimental model. Genetic Selection and Evolution 33:S399–S423.

Luo ZW. 1998. Detecting linkage disequilibrium between a polymorphic marker locus and a trait locus in natural populations. Heredity 80:198–208.

Mackay IJ, Gibson JP. 1993. The effect of gametic-phase disequilibrium on the prediction of response to recurrent selection in plants. Theoretical and Applied Genetics 87:152–160.

Mackay TFC. 2001. The genetic architecture of quantitative traits. Annu. Rev. Genet. 35:303–339.

Maynard Smith J, Haigh J. 1974. Hitch-hiking effect of a favorable gene. Genetical Research 23:23–35.

McIntosh RA. 1988. Catalogue of gene symbols for wheat. Available from: http://agris.fao.org/agris-search/search/display.do?f=1991/GB/GB91012.xml;GB9012195

Michaels SD, John MC, Amasino RM. 1994. Removal of polysaccharides from plant DNA by ethanol precipitation. Biotechniques 17:274–276.

Nei M, Tajima F. 1981. Genetic drift and estimation of effective population size. Genetics 98:625–640.

Neuhauser C. 2001. Mathematical models in population genetics. Handbook of statistical genetics [Internet]. Available from: http://onlinelibrary.wiley.com.gate1.inist.fr/doi/10.1002/0470022620.bbc20/full

Nielsen R. 2005. Molecular signatures of natural selection. Annu. Rev. Genet. 39:197–218.

Olivieri I, Couvet D, Gouyon PH. 1990. The genetics of transient populations: Research at the metapopulation level. Trends in Ecology & Evolution 5:207–210.

Paradis E. 2010. pegas: an R package for population genetics with an integrated–modular approach. Bioinformatics 26:419–420.

Plummer M, Best N, Cowles K, Vines K. 2006. CODA: Convergence diagnosis and output analysis for MCMC. R news 6:7–11.

Plummer M. 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In: Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003). March. p. 20–22. Available from: http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Drafts/Plummer.pdf

Porcher E, Giraud T, Goldringer I, Lavigne C. 2004. Experimental Demonstration of a Causal Relationship Between Heterogeneity of Selection and Genetic Differentiation in Quantitative Traits. Evolution 58:1434–1445.

R Development Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria Available from: http://www.R-project.org.

Raquin AL, Brabant P, Rhoné B, Balfourier F, Leroy P, Goldringer I. 2008. Soft selective sweep near a gene that increases plant height in wheat. Molecular ecology 17:741–756.

Raquin A-L, Depaulis F, Lambert A, Galic N, Brabant P, Goldringer I. 2008. Experimental Estimation of Mutation Rates in a Wheat Population With a Gene Genealogy Approach. Genetics 179:2195–2211.

Rhoné B, Raquin A-L, Goldringer I. 2007. Strong linkage disequilibrium near the selected Yr17 resistance gene in a wheat experimental population. Theoretical and Applied Genetics 114:787–802.

Rhoné B, Remoué C, Galic N, Goldringer I, Bonnin I. 2008. Insight into the genetic bases of climatic adaptation in experimentally evolving wheat populations. Molecular Ecology 17:930–943.

Rhoné B, Vitalis R, Goldringer I, Bonnin I. 2010. Evolution of Flowering Time in Experimental Wheat Populations: A Comprehensive Approach to Detect Genetic Signatures of Natural Selection. Evolution 64:2110–2125.

699 Robertson A. 1960. A Theory of Limits in Artificial Selection. Proc. R. Soc. Lond. B 153:234–249.

700 Rogers SM, Bernatchez L. 2005. FAST-TRACK: Integrating QTL mapping and genome scans
701 towards the characterization of candidate loci under parallel selection in the lake whitefish
702 (Coregonus clupeaformis). Molecular Ecology 14:351–361.

703 Rousset M, Bonnin I, Remoué C, et al. 2011. Deciphering the genetics of flowering time by an
704 association study on candidate genes in bread wheat (Triticum aestivum L.). Theor Appl
705 Genet 123:907–926.

706 Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. PNAS 100:9440–9445.

707 Trottet M. 1988. Use of genic male sterility for breeding wheat lines resistant to Leptosphaeria
708 nodorum Muller: Results of a first cycle and prospect. In: Proc. Seventh Intnatl Wheat
709 Genetics Symp. Cambridge, UK pp. p. 1199–1202.

710 Wang J, Caballero A. 1999. Developments in predicting the effective size of subdivided populations.
711 Heredity 82:212–226.

712 Wang J, Whitlock MC. 2003. Estimating Effective Population Size and Migration Rates From Genetic
713 Samples Over Space and Time. Genetics 163:429–446.

714 Waples RS. 1989. A generalized approach for estimating effective population size from temporal
715 changes in allele frequency. Genetics 2:379–391.

716 Welsh JR, Klatt AR. 1971. Effects of temperature and photoperiod on spring wheat pollen viability.
717 Crop Science 11:864–865.

718

Figure Captions

Figure 1 : INRA MAGIC population creation scheme. Grey intensity represents Probus proportion in the population.

Figure 2 : Individuals projections on discriminant axes with groups represented with different colors: two groups (European and non European lines) (on the left) and six groups for parental lines (pedigree related) (on the right).

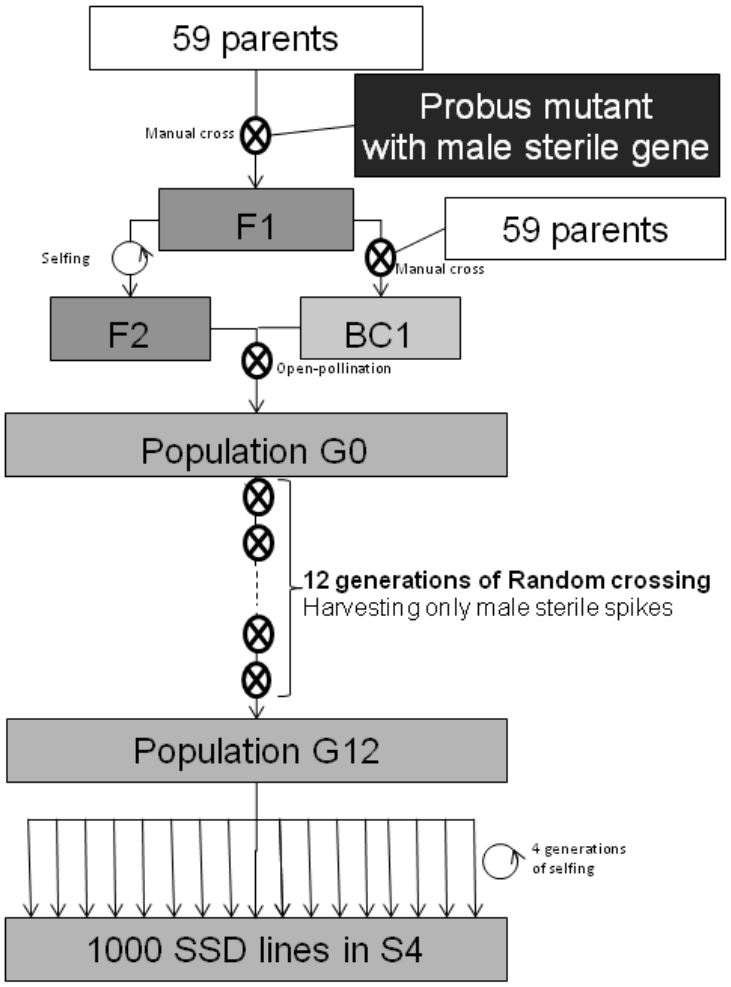
Figure 3 : Distribution of contribution estimated by chromosome for each parent. Black stars are parental contributions estimated genome-wide with credible interval (max=0.002). Parents which have the highest contribution are in red and them with the lowest in green.

Figure 4 : LD decay (mean and standard deviation every 5cM) as a function of the genetic distance in the initial population (in blue circle) and in the evolved population (in green triangle). Inter-chromosome mean LD is at an arbitrary distance of 320cM.

Figure 5 : Manhattan plot of the -log(p-value) of selection tests. The horizontal line represents the significance threshold. Markers under significant selection and associated to the heading date are circled in black, dark grey and grey representing respectively a threshold of 0.001, 0.05 and 0.1. Ppd-D1 is located on the chromosome 2D and is circled in black.

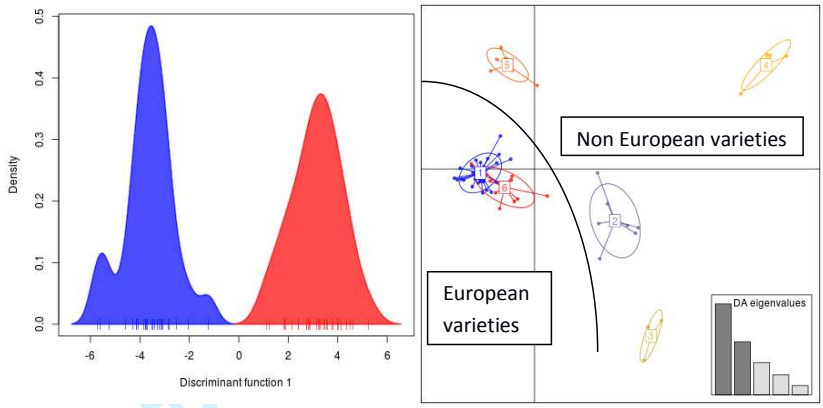
Figure 6 : Heading date (in dd) distributions for initial population (above) estimated with Bayesian method and evolved population (SSD) (below). Colors represent the proportion in each class of individuals with the photoperiod insensibility allele (Ppd-D1a) in grey and with the photoperiod sensibility allele (Ppd-D1b) in white.

740

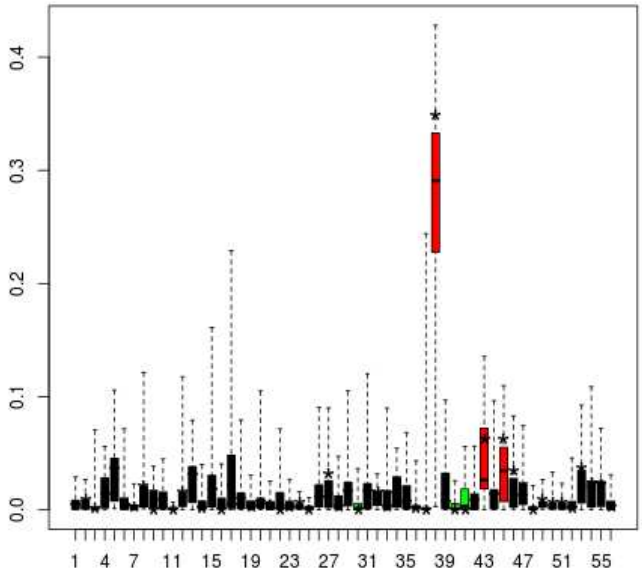


741

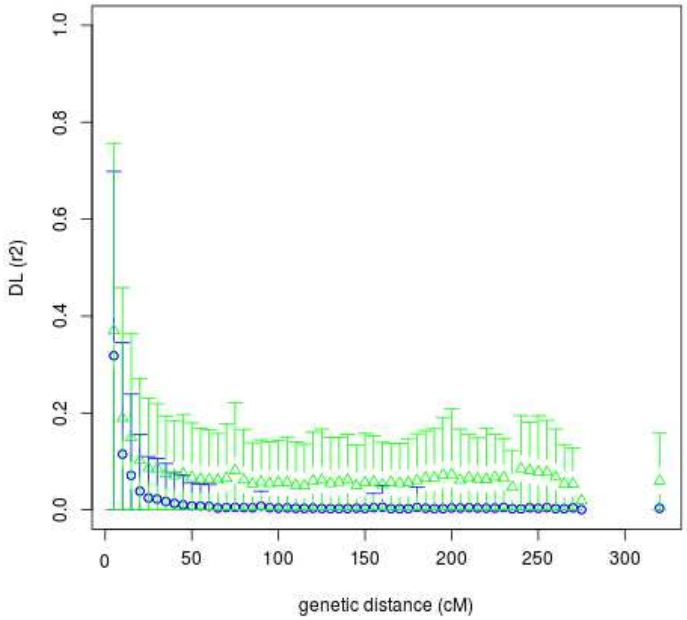
742 Figure 1



743 Figure 2

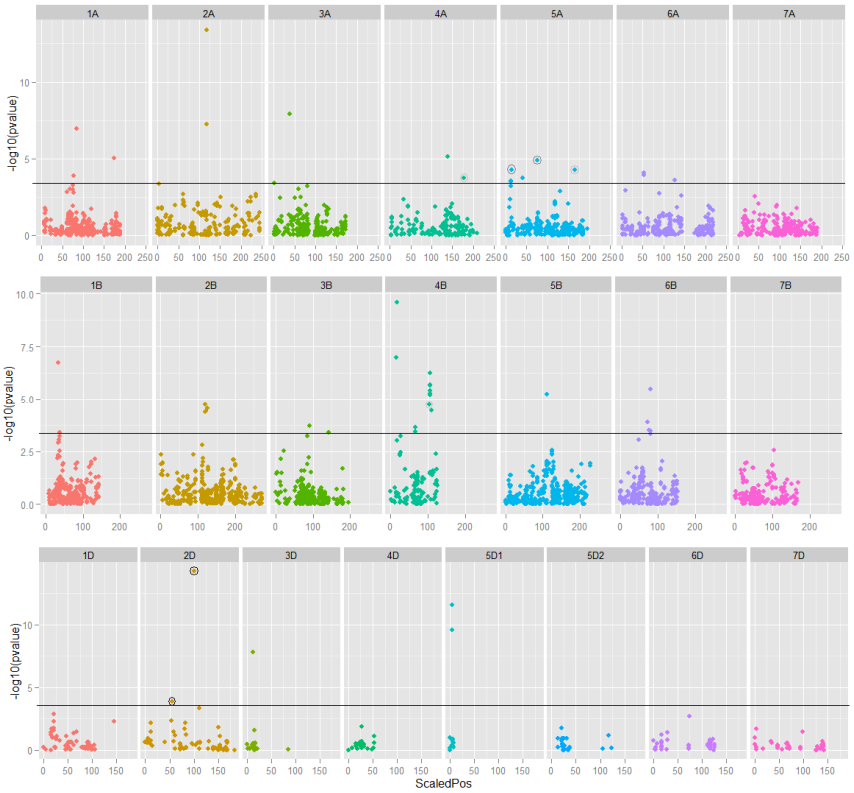


744 Figure 3



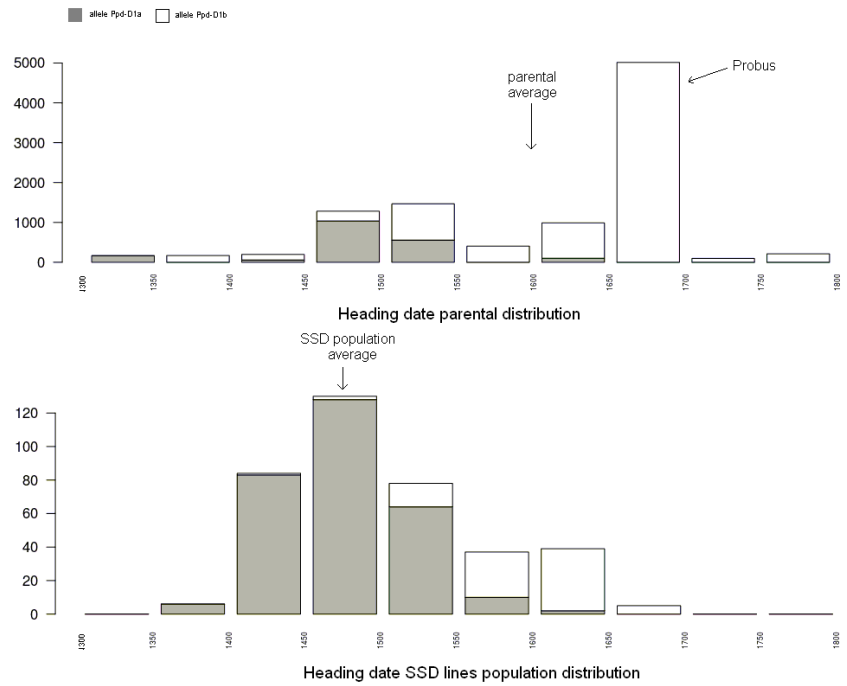
745 Figure 4

746



747

748 Figure 5



749 Figure 6

4.2 Tableaux supplémentaires du manuscrit "A pamictic experimental wheat population to detect markers under selection associated with earliness"

Supplementary Tables and Figures

Table S1: List of parental lines with their pedigree and their origin (from Pontis 1992, & Trottet)

	Genotype	Pedigree	Origin
1	3596-58		Bulgaria
2	ANDES 56		South Am
3	ATLAS 66	Frondosa / Redhart3-Noll28	USA
4	C 56-5	(2-7 x Cappelle) x 16-4-2) 56-5 (with 16-4-2 sister line of Courtôt)	France
5	CA68	(Champlein x Aronde) 68	France
6	CARALA	selection in Alabama bluestem	USA
7	CARDENAL	Mengavi / 8156 // Jar / 3 / 8157	Australia
8	CHALK	Carsten VIII / H4255	GB
9	CLEMENT	Hope / Timstein // 3x Heines VII / Riebesel 57-41 / 2x Heines VII / 4 / Cléo	NL
10	COMTAL	*T.durum* / Champlein // Dakota-51	France
11	CONDOR	Lerma Rojo // Norin10 / Brevor14 / 4 / Y54 // Norin10 / Brevor14 / 3 / 3 x Andes	Australia
12	CORIN	ST 102 x Tadorna	GB
13	COURTOT	Mexique-50 / Versailles-B21	France
14	DARKAN	Eureka2 / Kenya C6041 // ?	Australia
15	DOMUS		Germany
16	DUCAT	Prieur / 3 / SW // 90 / Etoile de Choisy	France
17	FERMO Mutant 9-14		Switzerland
18	GRANA	Etoile de Choisy x Wysokolitewka Sztywnosloma x Dankowskabiola	Poland
19	HADDEN		USA
20	HARUHIKARI		Japan
21	IAS 20	Iassul =Colonias // Kenya58 / Frontana	Brazil
	IAS 63		**Brazil**
22	KAVKAZ	Lutenscens 314 H147 / Bezostaya 1	URSS
23	L707	Wakeland x Blueboy	USA
24	LAFRON		South Am
25	LAGOA VERMELHA	landrace	Brazil
26	LAPIS		Germany
27	LUTIN	Champlein / Cappelle // Versailles-B21	France
28	MARIS HOBBIT	(Professeur Marchal x (Marne x VG-9144)) x TJB 16	GB
29	MARIS HUNTSMAN	Ci 12633 / 5 x Cappelle // Hybride46 / Cappelle / 3 / 2 x Professeur Marchal	GB
30	MARQUILLO	Marquis / Uimillo	USA
31	MINISTRE NAIN	Benoist40 / Professeur Delos // ? x line RHT3	Belgium
32	MIRONOVSKAIA 808	selection Artemovka	URSS
33	mutant du 81-12	[(68-2 x Yga) 8-2 x (90-2 x Etoile de Choisy) 1-8]	France
34	NAUTICA	Mildress / Manella	NL
35	OASIS	Arthur / 5 / Arthur x 3 / 3 / Ribo // Riley x 2 / Riley67	USA
36	ORLANDO	Descendent line from rye with Neuzucht	Germany
37	OXLEY	Penjamo62 /4x Gabos56 // TPP / Nainari60 / 4 / 2 x Lerma Rojo // Norin10 / Brevor14 / 3 / 3 x Andes	Australia
38	mutant of PROBUS	Trubillo / Platahof (line given the male sterility)	France
39	R 5-1	[(VPMM x Moisson) 9 x (US(60)43 x Prieur) 61] 5-1	France
40	REDHART	selection in Southern Flint or Red May	USA
41	REDON M4	From population from Brittany	France

	Genotype	Pedigree	Origin
42	SAPPO	Ci 12633 / 5 xRing // Els / 6 x Ring	GB
43	TALENT	Champlein / 3 / Thatcher / Vilmorin27 // Fortunato	France
44	TJB 155 = KINSMAN	[(Ci 12633 x Cappelle) x (Hybride45 x Cappelle)] x (Professeur Marchal x Maris Ranger)	GB
45	TJB 240	Maris Envoy / TL365-A25	GB
	TJB 251		**GB**
46	TJB 636		GB
	TL 25-11		**GB**
47	TL 365 A34	TJB 16-18 / 3 / Cappelle // Vilmorin 29 / VG8058	GB
48	TOROPI	Petitblanco8 // Frontana 1971-37 / Quaderna	Brazil
49	US 113		USA
50	US 117		USA
51	US 123		USA
52	US 125		USA
53	V2D11	(VPMM 1.1.2.4 x D65.5) 11-3	France
	V3D8	**[(VPMM 1-1-2-4 x D65-5) 8-4**	**France**
54	(VPM x Moisson4) 3lines		France
55	VPMM1-1-1-2 R4		France
56	WEINSTEPHAN 1007-53	Heines Bart / seigle 4x // ? x bread wheat	Germany

Genotypes in bold are no longer available.
15896 = descendent line from TL25-11 ; 15958 = descendent line from TJB251 ; VG = vogel ;
TJB 16 = Ci 12633 / 4 x Cappelle // Heine 10 / Cappelle / 3 / Nord Desprez

Reference :

Pontis C. 1992. Utilisation de marqueurs génétiques pour le suivi de la variabilité de 3 composites de blé tendre d'hiver (Triticum aestivum L.) menées en gestion dynamique.

Table S2 : List of genotyped markers located in candidate genes

Gene	Polymorphism	References
CO-B	SNP	Rhone et al 2008
FTA	SSR	Bonnin et al. 2008
FTD	1bp indel	Bonnin et al. 2009
LDDB	SNP	Rhone et al 2008
Ppd-D1	2kb indel	Beales et al. 2007
Vrn1A	4bp indel	Rhone et al. 2008
Ppd-A1	1117bp indel	Bentley et al. 2011
Ppd-A1	305bp indel	S. Griffiths, pers. comm.
TaGW2	SNP	Su et al. 2011
RHT-B1	SNP	S. Griffiths, pers. comm.
Ppd-B1	SNP	S. Griffiths, pers. Comm.
Ppd-B1	SNP	S. Griffiths, pers. comm.
Vrn1A	SNP	Sherman et al. 2004
Vrn1D	4bp indel	Fu et al. 2005

References :

Beales J, Turner A, Griffiths S, Snape J, Laurie D. 2007. A Pseudo-Response Regulator is misexpressed in the photoperiod insensitive Ppd-D1a mutant of wheat (Triticum aestivum L.). Theoretical and Applied Genetics 115:721–733.

Bentley AR, Turner AS, Gosman N, Leigh FJ, Maccaferri M, Dreisigacker S, Greenland A, Laurie DA. 2011. Frequency of photoperiod-insensitive Ppd-A1a alleles in tetraploid, hexaploid and synthetic hexaploid wheat germplasm. Plant Breeding 130:10–15.

Bonnin I, Rousset M, Madur D, Sourdille P, Dupuits C, Brunel D, Goldringer I. 2008. FT genome A and D polymorphisms are associated with the variation of earliness components in hexaploid wheat. Theoretical and Applied Genetics 116:383–394.

Fu D, Szűcs P, Yan L, Helguera M, Skinner JS, von Zitzewitz J, Hayes PM, Dubcovsky J. 2005. Large deletions within the first intron in VRN-1 are associated with spring growth habit in barley and wheat. Molecular Genetics and Genomics 273:54–65.

Rhoné B. 2008. Etude de mécanismes génétiques impliqués dans l'adaptation climatique de populations expérimentales de blé tendre.

Sherman JD, Yan L, Talbert L, Dubcovsky J. 2004. A PCR Marker for Growth Habit in Common Wheat Based on Allelic Variation at the VRN-A1 Gene. Crop Sci 44:1832–1838.

Su Z, Hao C, Wang L, Dong Y, Zhang X. 2011. Identification and development of a functional marker of TaGW2 associated with grain weight in bread wheat (Triticum aestivum L.). Theoretical and Applied Genetics 122:211–223.

Table S3: Summary table of selection tests and association tests results with the location of markers and allelic frequency in initial and evolved populations. Only results of markers significantly under selection are presented.

Marker name	Selection coefficient	Likelihood ratio	p-value selection test	Fc	Parental freq	q-value
wsnp_Ex_rep_c66846_65240088	-0.10	14.68	1.27E-04	0.12	0.88	2.13E-02
wsnp_Ra_rep_c110367_93018635	-0.16	28.18	1.10E-07	0.06	0.94	5.63E-05
wsnp_Ex_c15377_23637176	0.10	19.58	9.64E-06	0.79	0.21	2.57E-03
wsnp_Ex_c14896_23029875	-0.13	27.21	1.82E-07	0.12	0.88	8.60E-05
wsnp_Ku_c28591_38511783	-0.09	12.45	4.17E-04	0.59	0.41	4.65E-02
wsnp_Ex_rep_c67264_65795361	-0.13	29.44	5.77E-08	0.19	0.81	3.53E-05
wsnp_JD_c37979_27684951	-0.22	57.03	4.27E-14	0.05	0.95	8.73E-11
wsnp_Ex_c2560_4764402	-0.18	18.32	1.87E-05	0.02	0.98	4.56E-03
wsnp_Ex_c5943_10422207	0.15	16.83	4.08E-05	0.97	0.03	8.63E-03
wsnp_CAP11_c3842_1829821	-0.14	14.67	1.28E-04	0.73	0.27	2.13E-02
Ppd.D1prom.2kb.indel.	0.20	61.02	5.66E-15	1.01	0.19	1.73E-11
wsnp_Ex_c10014_16477392	0.16	12.58	3.88E-04	0.98	0.02	4.49E-02
wsnp_Ex_c21950_31124594	0.21	32.52	1.18E-08	0.98	0.02	9.01E-06
wsnp_RFL_Contig383_4163148	-0.09	13.79	2.04E-04	0.17	0.83	3.12E-02
wsnp_Ex_c13217_20858366	-0.08	12.52	4.01E-04	0.31	0.69	4.55E-02
wsnp_Ex_c8316_14023638	-0.15	31.97	1.56E-08	0.09	0.91	1.06E-05
wsnp_Ex_rep_c104859_89444355	-0.10	20.01	7.68E-06	0.19	0.81	2.14E-03
wsnp_Ex_rep_c67145_65628860	-0.09	13.97	1.85E-04	0.20	0.80	2.91E-02
wsnp_Ku_c7153_12360198	-0.12	28.31	1.03E-07	0.29	0.71	5.63E-05
wsnp_Ex_c16389_24884851	-0.28	40.03	2.49E-10	0.63	0.37	2.29E-07
wsnp_Ra_rep_c71114_69138821	0.28	40.03	2.49E-10	0.37	0.63	2.29E-07
wsnp_Ex_c57212_59019379	-0.08	12.64	3.76E-04	0.19	0.81	4.43E-02
wsnp_Ex_c22740_31947788	0.09	13.49	2.40E-04	0.81	0.19	3.49E-02
wsnp_Ex_c39876_47057394	0.11	18.25	1.93E-05	0.86	0.14	4.56E-03
wsnp_Ku_c12557_20249122	0.10	20.70	5.36E-06	0.71	0.29	1.73E-03
wsnp_Ex_c21293_30421496	-0.11	22.37	2.25E-06	0.31	0.69	8.60E-04
wsnp_Ex_rep_c67510_66116823	0.11	22.54	2.06E-06	0.69	0.31	8.41E-04
wsnp_Ex_c47536_52716088	0.10	21.34	3.84E-06	0.66	0.34	1.31E-03
wsnp_CAP7_c599_312057	0.10	20.39	6.31E-06	0.65	0.35	1.84E-03
wsnp_Ex_c15490_23776560	0.09	17.01	3.71E-05	0.60	0.40	8.12E-03
wsnp_Ex_c23795_33034037	0.09	12.95	3.19E-04	0.45	0.55	4.02E-02
wsnp_Ku_rep_c71232_70948744	-0.09	13.08	2.98E-04	0.55	0.45	3.88E-02
wsnp_Ex_c2332_4371926	-0.11	16.33	5.32E-05	0.61	0.39	1.03E-02
wsnp_CAP11_c299_251533	-0.08	14.09	1.74E-04	0.26	0.74	2.80E-02
wsnp_Ra_c10915_17838202	-0.10	19.02	1.29E-05	0.42	0.58	3.29E-03
wsnp_Ex_c31830_40573624	-0.09	16.31	5.37E-05	0.23	0.77	1.03E-02
wsnp_JD_c9613_10432955	-0.14	20.56	5.77E-06	0.64	0.36	1.77E-03
wsnp_Ex_c24659_33912464	-0.21	48.83	2.78E-12	0.05	0.95	4.26E-09
wsnp_Ku_c5228_9318604	0.19	39.94	2.61E-10	0.94	0.06	2.29E-07
wsnp_Ex_c31149_39975724	-0.10	15.46	8.40E-05	0.14	0.86	1.56E-02
wsnp_CAP8_c5350_2554478	-0.09	13.45	2.45E-04	0.16	0.84	3.49E-02
wsnp_Ex_c23474_32717535	-0.10	14.66	1.28E-04	0.14	0.86	2.13E-02
wsnp_Ex_c1398_2676484	0.09	12.93	3.22E-04	0.83	0.17	4.02E-02
wsnp_Ex_c1398_2677882	-0.09	12.86	3.35E-04	0.17	0.83	4.07E-02
wsnp_BQ159615B_Ta_2_1	0.12	21.65	3.27E-06	0.90	0.10	1.18E-03
wsnp_CAP12_c2118_1040487	-2.00	13.74	2.09E-04	0.91	0.09	3.12E-02

Marker name	freq ssd	Chr	ScaledPos	number of marker in total LD	pvalue association with earliness
wsnp_Ex_rep_c66846_65240088	0.63	1A	76.49		1.33E-01
wsnp_Ra_rep_c110367_93018635	0.62	1A	84.56		7.63E-01
wsnp_Ex_c15377_23637176	0.54	1A	174.32		7.54E-01
wsnp_Ex_c14896_23029875	0.52	1B	32.84		1.25E-01
wsnp_Ku_c28591_38511783	0.16	1B	37.19		8.48E-01
wsnp_Ex_rep_c67264_65795361	0.40	2A	117.57		1.31E-01
wsnp_JD_c37979_27684951	0.46	2A	117.57		1.99E-01
wsnp_Ex_c2560_4764402	0.79	2B	119.56		6.58E-01
wsnp_Ex_c5943_10422207	0.23	2B	121.97	2	2.01E-01
wsnp_CAP11_c3842_1829821	0.06	2D	55.84		3.53E-11
Ppd.D1prom.2kb.indel.	0.77	2D	NA		4.85E-70
wsnp_Ex_c10014_16477392	0.17	3A	2.48		7.86E-01
wsnp_Ex_c21950_31124594	0.32	3A	37.70		3.90E-01
wsnp_RFL_Contig383_4163148	0.57	3B	90.65		2.19E-01
wsnp_Ex_c13217_20858366	0.42	3B	142.75		8.01E-01
wsnp_Ex_c8316_14023638	0.53	3D	11.92		7.64E-01
wsnp_Ex_rep_c104859_89444355	0.48	4A	139.05		3.92E-01
wsnp_Ex_rep_c67145_65628860	0.53	4A	176.91		6.50E-02
wsnp_Ku_c7153_12360198	0.30	4B	15.76		7.77E-01
wsnp_Ex_c16389_24884851	0.02	4B	16.37		8.85E-01
wsnp_Ra_rep_c71114_69138821	0.98	4B	16.37		8.85E-01
wsnp_Ex_c57212_59019379	0.55	4B	65.82		8.46E-01
wsnp_Ex_c22740_31947788	0.46	4B	65.82		8.08E-01
wsnp_Ex_c39876_47057394	0.44	4B	104.66		8.60E-02
wsnp_Ku_c12557_20249122	0.64	4B	105.57		8.16E-01
wsnp_Ex_c21293_30421496	0.32	4B	105.57		2.31E-01
wsnp_Ex_rep_c67510_66116823	0.68	4B	105.57		1.41E-01
wsnp_Ex_c47536_52716088	0.70	4B	106.03	1	1.73E-01
wsnp_CAP7_c599_312057	0.70	4B	106.45		1.56E-01
wsnp_Ex_c15490_23776560	0.72	4B	110.59		2.39E-01
wsnp_Ex_c23795_33034037	0.81	5A	12.72	3	8.93E-01
wsnp_Ku_rep_c71232_70948744	0.18	5A	12.72		8.78E-01
wsnp_Ex_c2332_4371926	0.12	5A	13.27		3.41E-02
wsnp_CAP11_c299_251533	0.45	5A	39.74		7.75E-01
wsnp_Ra_c10915_17838202	0.25	5A	76.39		1.78E-02
wsnp_Ex_c31830_40573624	0.46	5A	164.67		6.32E-02
wsnp_JD_c9613_10432955	0.08	5B	111.56		9.55E-01
wsnp_Ex_c24659_33912464	0.50	5D1	4.57		2.77E-01
wsnp_Ku_c5228_9318604	0.45	5D1	4.79		5.27E-01
wsnp_Ex_c31149_39975724	0.59	6A	52.58	1	5.19E-01
wsnp_CAP8_c5350_2554478	0.58	6A	126.45		3.62E-01
wsnp_Ex_c23474_32717535	0.59	6B	73.70		5.82E-01
wsnp_Ex_c1398_2676484	0.42	6B	75.80		3.74E-01
wsnp_Ex_c1398_2677882	0.58	6B	78.37	1	2.98E-01
wsnp_BQ159615B_Ta_2_1	0.40	6B	80.00		5.51E-01
wsnp_CAP12_c2118_1040487	0.00	NA	NA		NA

Table S4 : Results of t-tests between contribution estimated with one chromosome and genome-wide by chromosome and by parent.

Name	1A	1B	1D	2A	2B	2D	3A	3B	3D	4A	4B	4D	5A	5B	5D1	5D2	6A	6B	6D	7A	7B	7D	number of significant chromosomes
CONDOR	*	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	1
LAGOA.VERMELHA	NS	NS	NS	NS	NS	NS	NS	NS	NS	*	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	1
OXLEY	NS	NS	NS	NS	NS	NS	NS	*	NS	*	NS	NS	NS	NS	*	NS	NS	NS	NS	NS	NS	NS	3
ATLAS.66	NS	NS	NS	*	NS	NS	*	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	*	*	NS	4
CLEMENT	NS	NS	NS	NS	NS	*	*	*	NS	NS	NS	NS	*	NS	NS	NS	NS	NS	NS	NS	NS	NS	4
DARKAN	NS	*	NS	NS	*	NS	NS	NS	NS	NS	NS	NS	NS	NS	*	NS	NS	NS	NS	*	NS	NS	4
ORLANDO	NS	NS	NS	*	*	NS	*	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	*	NS	NS	NS	NS	4
REDHART	NS	NS	NS	NS	NS	*	*	NS	NS	NS	NS	NS	NS	*	NS	NS	*	NS	NS	NS	NS	NS	4
TOROPI	*	*	NS	NS	*	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	*	NS	NS	NS	NS	4
DUCAT	*	*	NS	NS	NS	NS	NS	NS	NS	NS	*	NS	NS	NS	NS	NS	NS	*	NS	NS	*	NS	5
KAVKAZ	*	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	*	NS	*	*	NS	*	5
MARQUILLO	*	NS	*	NS	NS	*	NS	NS	*	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	*	NS	NS	5
NAUTICA	*	NS	NS	*	*	NS	NS	NS	NS	NS	*	NS	NS	*	NS	NS	NS	NS	NS	NS	NS	NS	5
REDON.M4	NS	*	NS	NS	NS	NS	*	NS	NS	NS	*	NS	NS	NS	*	NS	NS	NS	NS	NS	*	NS	5
CARDENAL	*	*	NS	NS	NS	NS	*	*	NS	NS	*	NS	NS	NS	NS	NS	NS	NS	NS	NS	*	NS	6
MARIS.HOBBIT	NS	NS	*	*	NS	*	*	NS	NS	NS	NS	NS	NS	*	NS	NS	NS	NS	NS	NS	*	NS	6
US.113	NS	NS	NS	NS	NS	*	*	*	NS	NS	NS	NS	NS	*	NS	*	*	NS	NS	NS	NS	NS	6
US.125	NS	NS	*	NS	NS	*	*	NS	NS	*	NS	NS	*	NS	NS	NS	NS	NS	NS	*	NS	NS	6
WEINSTEPHAN.1007.53	*	*	NS	*	*	NS	NS	NS	NS	NS	*	NS	NS	NS	NS	NS	NS	NS	NS	NS	*	NS	6
3596.58	*	NS	NS	NS	NS	NS	NS	*	NS	NS	*	NS	*	*	NS	NS	NS	*	NS	NS	*	NS	7
L707	*	NS	NS	NS	*	NS	NS	*	NS	NS	*	NS	NS	NS	NS	NS	*	*	NS	*	NS	NS	7
CARALA	*	NS	*	*	*	NS	*	*	NS	NS	NS	NS	*	NS	NS	NS	*	NS	NS	NS	NS	NS	8
HADDEN	NS	*	NS	*	NS	NS	*	NS	NS	*	NS	NS	*	*	NS	NS	NS	*	NS	*	NS	NS	8
LAPIS	NS	NS	NS	*	*	NS	*	NS	NS	*	NS	NS	*	*	NS	NS	*	NS	NS	NS	*	NS	8
VPMxMoisson4.3lignees	NS	NS	NS	*	NS	NS	NS	NS	NS	*	NS	NS	*	*	NS	NS	*	NS	NS	*	*	*	8
GRANA	*	*	NS	*	NS	*	*	NS	NS	NS	NS	NS	*	NS	NS	NS	NS	*	NS	*	*	NS	9
HARUHIKARI	*	*	NS	*	NS	NS	*	*	NS	*	*	NS	*	NS	*	NS	NS	NS	NS	NS	NS	NS	9
LAFRON	NS	*	*	*	NS	*	*	*	NS	NS	*	NS	NS	*	NS	NS	NS	NS	NS	*	NS	NS	9
MIRONOVSKAIA.808	NS	*	*	*	NS	NS	*	NS	NS	NS	NS	NS	NS	*	*	NS	*	NS	NS	*	NS	*	9
TL365A34	*	*	NS	NS	*	*	*	NS	NS	*	*	NS	NS	*	NS	NS	NS	NS	NS	NS	*	NS	9
US.123	NS	*	NS	*	*	NS	*	NS	NS	NS	*	NS	*	*	NS	NS	NS	*	NS	*	NS	NS	9
VPMM1.1.1.2.R4	*	NS	NS	NS	*	*	*	NS	NS	NS	*	NS	NS	NS	NS	NS	*	*	NS	NS	*	*	9
CORIN	NS	*	NS	NS	*	NS	*	*	*	*	*	NS	*	*	NS	NS	NS	NS	NS	*	NS	NS	10
IAS.20	*	NS	NS	*	*	*	*	NS	NS	NS	*	NS	NS	NS	NS	*	NS	*	NS	*	*	NS	10
MARIS.HUNTSMAN	NS	*	NS	*	NS	*	*	NS	NS	*	*	NS	*	*	NS	NS	NS	NS	NS	*	*	NS	10
MINISTRE.NAIN	*	NS	*	*	*	NS	*	NS	NS	*	*	NS	NS	*	NS	NS	NS	*	NS	NS	*	NS	10
US.117	*	*	NS	*	*	*	*	*	NS	NS	*	NS	NS	*	NS	NS	*	NS	NS	NS	NS	NS	10
CA68	*	*	NS	NS	*	*	NS	*	NS	NS	NS	NS	*	*	*	NS	*	*	NS	NS	NS	*	11
SAPPO	NS	NS	*	*	*	*	*	*	NS	NS	*	NS	NS	*	NS	NS	*	NS	NS	*	*	NS	11
C.56.5	NS	*	NS	*	*	*	NS	*	NS	*	*	NS	*	*	NS	NS	*	*	NS	NS	*	NS	12
COMTAL	NS	*	NS	*	*	*	*	NS	NS	*	*	NS	*	*	NS	NS	*	NS	NS	*	NS	*	12
DOMUS	NS	*	*	*	*	NS	NS	NS	*	NS	*	NS	*	*	*	NS	*	*	NS	NS	*	NS	12
FERMO.Mutant.9.14	*	NS	NS	*	*	*	*	*	*	*	NS	NS	*	NS	NS	NS	*	*	NS	*	NS	NS	12
OASIS	NS	*	*	*	NS	NS	NS	NS	NS	*	*	*	*	*	NS	NS	*	*	NS	*	*	NS	12

Name	1A	1B	1D	2A	2B	2D	3A	3B	3D	4A	4B	4D	5A	5B	5D1	5D2	6A	6B	6D	7A	7B	7D	number of significant chromosomes
TJB155.KINSMAN	NS	*	NS	*	*	*	*	*	NS	*	*	NS	*	*	NS	NS	NS	*	NS	NS	*	NS	12
TJB636	*	*	NS	NS	*	*	*	*	*	NS	*	NS	NS	*	NS	NS	*	*	NS	*	NS	NS	12
V2D11	*	*	*	NS	*	NS	*	*	NS	*	NS	NS	*	*	NS	NS	*	*	NS	*	NS	NS	12
ANDES.56	NS	*	*	NS	*	*	*	NS	NS	*	*	NS	*	NS	NS	*	*	*	NS	*	*	NS	13
CHALK	NS	*	*	*	*	*	NS	*	NS	*	*	NS	*	*	*	NS	*	*	NS	NS	NS	NS	13
LUTIN	NS	*	NS	NS	*	*	*	*	NS	*	*	NS	NS	*	*	NS	*	*	NS	*	NS	*	13
NON.mutant.du.81.12	*	*	NS	*	NS	*	*	*	NS	*	*	NS	*	*	NS	NS	*	NS	NS	*	*	NS	13
R.5.1	NS	*	NS	*	*	NS	*	NS	NS	*	*	*	*	*	NS	NS	*	*	NS	*	*	NS	13
COURTOT	*	*	NS	*	*	*	*	*	NS	*	*	NS	*	NS	NS	NS	*	*	*	NS	*	NS	14
TALENT	*	*	NS	*	*	*	NS	*	NS	*	*	NS	NS	NS	*	*	*	NS	*	*	*	*	15
TJB240.Sportsman	*	NS	*	*	*	*	*	*	*	*	*	*	*	NS	NS	*	NS	*	*	*	NS	NS	16
PROBUS..male.fertile.	*	*	*	*	*	*	*	*	*	*	*	*	NS	NS	*	NS	*	*	NS	NS	*	*	17
number of significant parent	27	32	15	32	32	28	37	25	7	26	33	4	26	29	11	5	27	26	4	27	26	9	
number of markers	562	402	119	405	708	126	417	473	34	392	191	34	470	568	33	45	462	463	71	466	330	51	

Total number of significant tests and the number of markers located are given in table margins.

4.3 Interest of a multiparental and outcrossing wheat population for fine mapping

Interest of a multiparental and outcrossing wheat population for fine mapping.

Stéphanie Thépot*[1], Gwendal Restoux[2], Frédéric Hospital[3], David Gouache[4], Ian Mackay[5], Isabelle Goldringer[1], Jérôme Enjalbert[1]

[1] INRA, UMR 0320 / UMR 8120 Génétique Végétale, Gif-sur-Yvette, France

[2]Unite d'Ecologie, Systématique et Evolution - CNRS UMR8079, Université Paris-Sud, Orsay, France

[3] INRA, Génétique Animale et Biologie Intégrative, Jouy en Josas, France

[4] Arvalis, Institut du Végétal. Station Expérimentale, Boigneville, France

[5] NIAB, Huntingdon Road, Cambridge CB3 0LE, UK

Email: Stéphanie Thépot - stephanie.thepot@moulon.inra.fr,

Abstract

The use of multiparental populations for QTL discovery has been recently highlighted by different theoretical and experimental developments. Here, we explored the interest of French populations using heterogeneous genetic stocks of cultivated wheat, maintained *in situ* over 12 sites since 1984 with an outcrossing mating system. We studied one of these populations (Le Moulon, 48.4°N, 21°E), derived from twelve cycles of random crosses between 60 founders, selected to maximize genetic diversity. Outcrossing was allowed by the integration of a nuclear male sterility allele (ms1b, Probus donor) in the population. We analyzed 1000 Single Seed Descent lines (SSD) derived from the 12th generation of cultivation. This population was genotyped using the 9k i-select SNPs (Single Nucleotide Polymorphisms) array, covering the whole genome. Polymorphism and quality checks resulted in the selection of around 6500 SNPs.

First, the evolution of genetic diversity was explored through the comparison of SSD lines and the inferred initial population. The low population structure and the strong decay in linkage disequilibrium between SSD lines and the inferred initial population confirmed the efficiency of the twelve cycles of the random outcrossing in producing a highly diverse and recombined population. Two years of observations of population earliness under different environments were used to show the complementarity of association genetics, which allowed the detection of already known *Vrn* major genes, and evolutionary approach, which, lead to the discovery of two new minor effect QTLs.

Key words: evolution approach, recombinant population, dynamic management, wheat

2

Introduction:

Dynamic management (DM) aims at maintaining crop genetic diversity through *in situ* conservation of genetic resources. Genetically diverse populations are grown year after year, in various sites, differing for climate conditions, pathogen pressures and/or agricultural practices (Allard 1988; Henry et al. 1991; Porcher et al. 2004). In France, dynamic management has been experimented on bread wheat (*Triticum aestivum*) since 1984 (Henry et al. 1991), using three gene pools: two selfing populations (based on a pyramidal cross of 16 parents) and one outcrossing population. Samples of each of the three initial populations were sent to 7 to 12 sites in France and cultivated year after year in the same sites under the same conditions. Thus, these three "meta-populations" evolved over 10 to more than 20 generations without migration or conscious human selection (Enjalbert et al. 2011).

Studies on the selfing populations showed a good maintenance of global diversity at the network level (Raquin et al. 2008), both at phenotypic and molecular levels. A fast evolution of flowering time was observed, both over time and space (Rhoné et al. 2008): all populations flowered later than the initial population; and populations from Northern French sites flowered much later than Southern ones. Association genetics and spatio-temporal shifts in allelic frequencies revealed polymorphisms located in major genes controlling vernalization requirement or photoperiod sensitivity, partially explained climatic adaptation (Rhoné et al. 2008). Therefore, in addition to genetic resource preservation, DM populations can be an appropriate material to detect genes involved in local adaptation (Goldringer et al. 2001).

In the present study, we analyzed one outcrossing DM population, which is characterized by a high number of parents (60 lines), and numerous panmictic generations. In this population, wheat natural selfing habit was turned to an outcrossing mating system, using a recessive male sterility gene (*ms1b*), and harvesting solely open pollinated male-sterile plants. Taking the parental lines as reference, we studied an evolved population (12^{th} generation) and tested for possible markers selection, trying to link detected markers to the observed evolution in vernalization requirement.

Materials and Methods:

The population studied is derived from the cross of Probus, a mutant carrying a male sterile allele (*ms1b*), with 59 lines covering a large genetic diversity. Resulting F1 progenies were alternatively selfed (F2) or back-crossed (BC1) with the 59 parents to reduce the Probus genome contribution to the population. These two progenies (F2 & BC1) were sown together in an isolated field surrounded by rye. Male sterile spikes were tagged during flowering and harvested at maturity. Then, and over twelve generations, a random sample of harvested seeds on male sterile

plants was drawn each year, and resown in fall, in order to reach between 5 000 and 10 000 adult plants (2 500-5 000 male-sterile plants). After 12 generations we derived 1 000 SSD lines (F5, Fig. 1).

First we inferred allelic frequencies in the initial population on the basis of the 56 parental lines (including Probus) still available in seed banks (4 missing ones), estimating their contributions to the global pool using a Bayesian method (Thépot et al. in prep.). The evolved population was studied through a subset of 380 SSD lines, representative of the phenotypic diversity of the 1 000 lines.

Vernalization requirement was assessed in field trial at Le Moulon over two seasons (2010-2011 and 2011-2012), with a spring sowing (April), on a single row of 20 seeds per genotype. For each row, the heading date was scored when half of the plants had half of the main ear emerged from the flag leaf. The heading date was transformed into sums of degree-days (dd) (sums of the mean temperature per day) from sowing to heading. On the basis of the bimodal distribution of the heading date (Fig. 2), SSD lines were classified as spring type (heading before 2 000dd), or winter type for the others. Genotypes with inconsistent behavior between both years were discarded (8 SSD lines).

Genotyping was performed using the 9k i-select SNP array. Only SNPs unambiguously scored as biallelic after a visual inspection using Genome Studio software, were kept. Using KASPar SNP genotyping system (KBioscience), 14 additional polymorphisms located in candidate genes (earliness pathway such as PPD or VRN families) were genotyped.

The diversity detected by SNPs in both populations (parental lines and SSD lines) was compared using Minor Allele Frequency (MAF) and expected heterozygosity (He, Nei diversity). Evolution of growth habit was tested through a comparison of spring/winter ratio in the initial and the evolved populations (Chi square test). Strong shifts in allelic frequencies were used to detect markers under selection, using a new method (Thépot et al. in prep.). Q-values were estimated to cope with the multiple tests (Storey and Tibshirani 2003). Each marker under selection was also tested for association with growth habit using a Logit model.

Results and Discussion:

The genotyping of 436 lines (56 parents + 380 F5 lines) with the 9k i-select SNP assay resulted in 7 270 SNPs with high scoring quality. Among these SNPs, 88.2% were polymorphic in the initial population and 86.9% in the evolved population. This slightly higher diversity in the initial population was also observed on allelic frequencies of polymorphic SNPs (mean MAF: 0.18 *vs* 0.17 and He: 0.25 *vs* 0.24). The MAF distribution (Fig. 3) showed a globally high frequency of SNPs with a MAF inferior to 0.05, rare alleles being more frequent in the evolved population. This distribution contrasts with the one observed on a worldwide panel, using the same SNP array, which demonstrates a deficit of SNPs with a low MAF (Cavanagh et al. 2013). This deficit might be due to the fact that lines were chosen

4

to maximize the genetic diversity, and SNPs were intentionally selected to favor common allele in a panel of 26 cultivars from mainly USA and Australia (Cavanagh et al. 2013).

Among the 6459 polymorphic markers, 54 were detected under selection, representing 25 independent genomic areas. When assessing phenotypic evolution for flowering time, we observed a significant shift from 20% of spring type in the initial population to 47% in the evolved population. Among markers under selection, three were associated with the growth habit (p-value < 0.05). These markers, located on the 5D and 4A chromosomes, only explained a rather limited part of the phenotypic variation (2.8% with a global model included the three markers). Yet, they all have, experienced an increase of the spring allele frequency in the evolved population which explained a raise of spring type ranging from 3.5% to 5.4%. The 5D markers are located on the same chromosome as *VrnD1* but did not present linkage disequilibrium with the marker located in this gene (r^2<0.009). Surprisingly polymorphisms in candidate genes like *Vrn* families (*VrnA1-Prom*, *VrnA1-ex7* and *VrnD1*), although strongly associated to the growth habit (p-value<10^{-10}, r^2=28%) have not been detected as targeted by selection. To take into account the complexity of interaction between these three markers, we assumed that spring alleles are both dominant and epistatic (Rousset et al. 2011). Thus as soon as there is at least one *Spring* allele, the haplotype was classified as *Spring* type or Winter type otherwise. For parental lines, these *Vrn* haplotypes are almost completely explaining phenotype (97.5% of correspondence, r^2=0.85). However for the evolved population 30% of SSD lines with winter *Vrn* haplotype exhibited a spring phenotype (Table 1) (r^2=0.3). This evolution might be explained by i) high level of recombination that broke the initial full linkage disequilibrium between causal mutations and the three SNPs genotyped, and/or ii) the increase of *Spring* alleles at one (or several) non-genotyped *Vrn* genes, such as *VrnB1*. As SNPs, *Vrn* haplotypes did not present a significant shift (p-value=0.29), although a 6% increase of *Spring* haplotypes was observed between the initial and the evolved population (Table 1). One hypothesis to explain this absence of significant shift at these candidate genes could be their strong effect: a little variation in frequency at these major genes may have a strong effect on growth habit.

Association genetics and evolutionary approach provided complementary results. The first method detected QTLs with major effects while the second detected QTLs with lower effect but contributing to the evolution of phenotypes. Joint study of phenotypic and genetic evolutions allowed to detect new markers involved in the control of the growth habit on the 5D and 4A chromosomes.

With its high diversity, absence of structure and low LD (Thépot et al. in prep.), this population appears as a new QTL mapping resource, allowing the discovery of original genomic regions controlling traits of interest. Ongoing studies will better explore the potential of this population for detection, using the 1 000 SSD lines.

Allard RW (1988) Genetic Changes Associated with the Evolution of Adaptedness in Cultivated Plants and Their Wild Progenitors. J Hered 79:225 –238.

Cavanagh CR, Chao S, Wang S, et al. (2013) Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. Proc Natl Acad Sci 110:8057-8062.

Enjalbert J, Dawson JC, Paillard S, et al. (2011) Dynamic management of crop diversity: From an experimental approach to on-farm conservation. C R Biol 334:458–468.

Goldringer I, Enjalbert J, Raquin A-L, Brabant P (2001) Strong selection in wheat populations during ten generations of dynamic management. Genet Sel Evol 33:S441–S463.

Henry JP, Pontis C, David J, Gouyon PH (1991) An experiment on dynamic conservation of genetic resources with metapopulation. Species Conserv. Popul. Biol. Approach

Porcher E, Gouyon P-H, Lavigne C (2004) Dynamic management of genetic resources: maintenance of outcrossing in experimental metapopulations of a predominantly inbreeding species. Conserv Genet 5:259–269.

Raquin A-L, Depaulis F, Lambert A, et al. (2008) Experimental Estimation of Mutation Rates in a Wheat Population With a Gene Genealogy Approach. Genetics 179:2195–2211.

Rhoné B, Remoué C, Galic N, et al. (2008) Insight into the genetic bases of climatic adaptation in experimentally evolving wheat populations. Mol Ecol 17:930–943.

Rousset M, Bonnin I, Remoué C, et al. (2011) Deciphering the genetics of flowering time by an association study on candidate genes in bread wheat (Triticum aestivum L.). Theor Appl Genet 123:907-926.

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci 100:9440–9445.

Table 1: Summary results of Vrn haplotypes evolution between the initial population and the evolved population and their association to the winter/spring phenotypes. Genotypes are coded with S, H, W, with S for the homozygote spring allele, H for heterozygote and W for homozygote winter allele, assuming that spring allele are dominant and epistatic.

Genotypes			Haplotype	Freq initial	Freq SSD
Vrn1A.Prom	Vrn1A.Ex7	Vrn1D	Type	population	lines
S	S	W	Spring	19.6% (94% S ; 6% W)	25.3% (93% S ; 7% W)
S	H	W			
S	H	NA			
W	S	S			

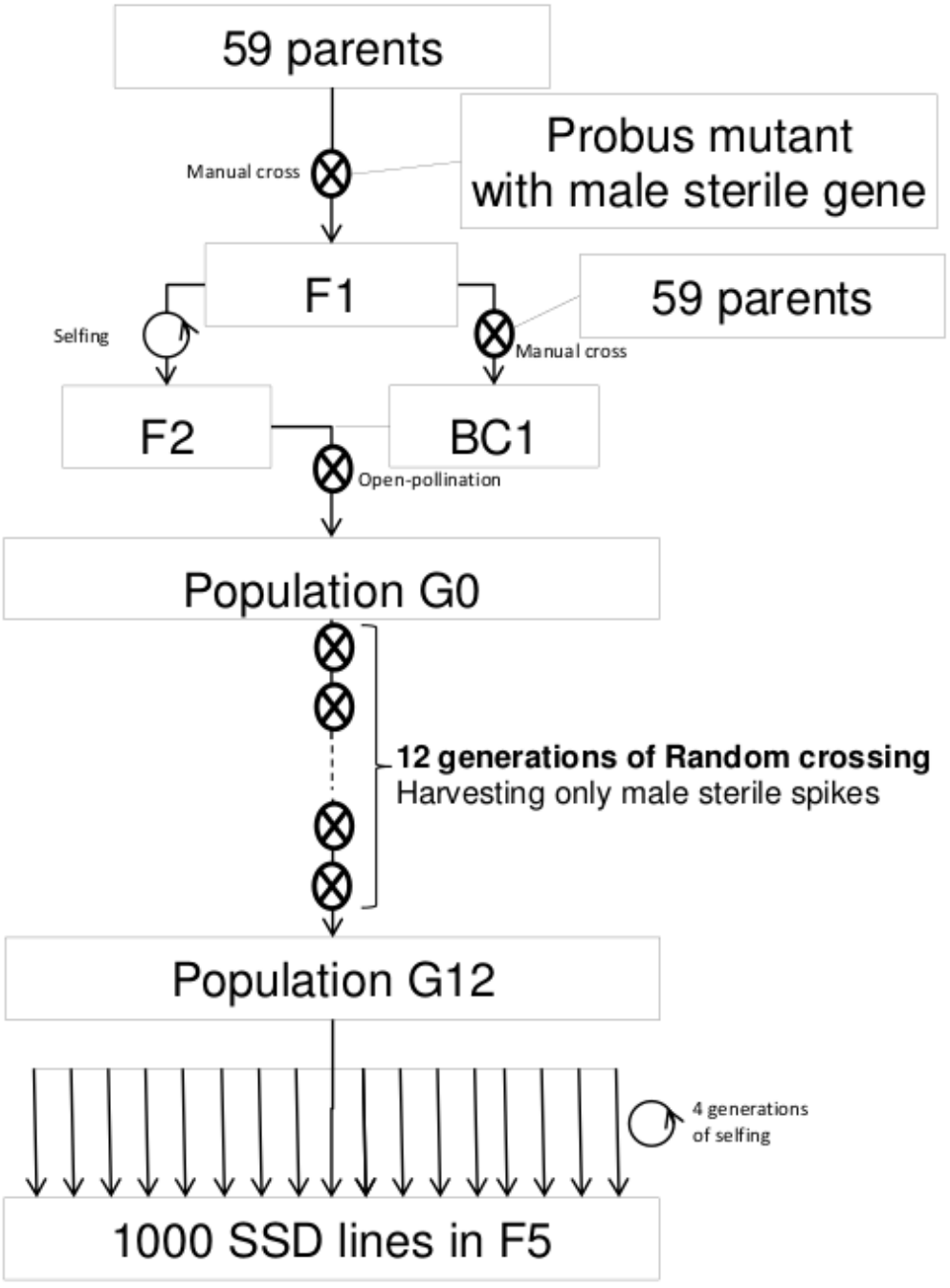
6

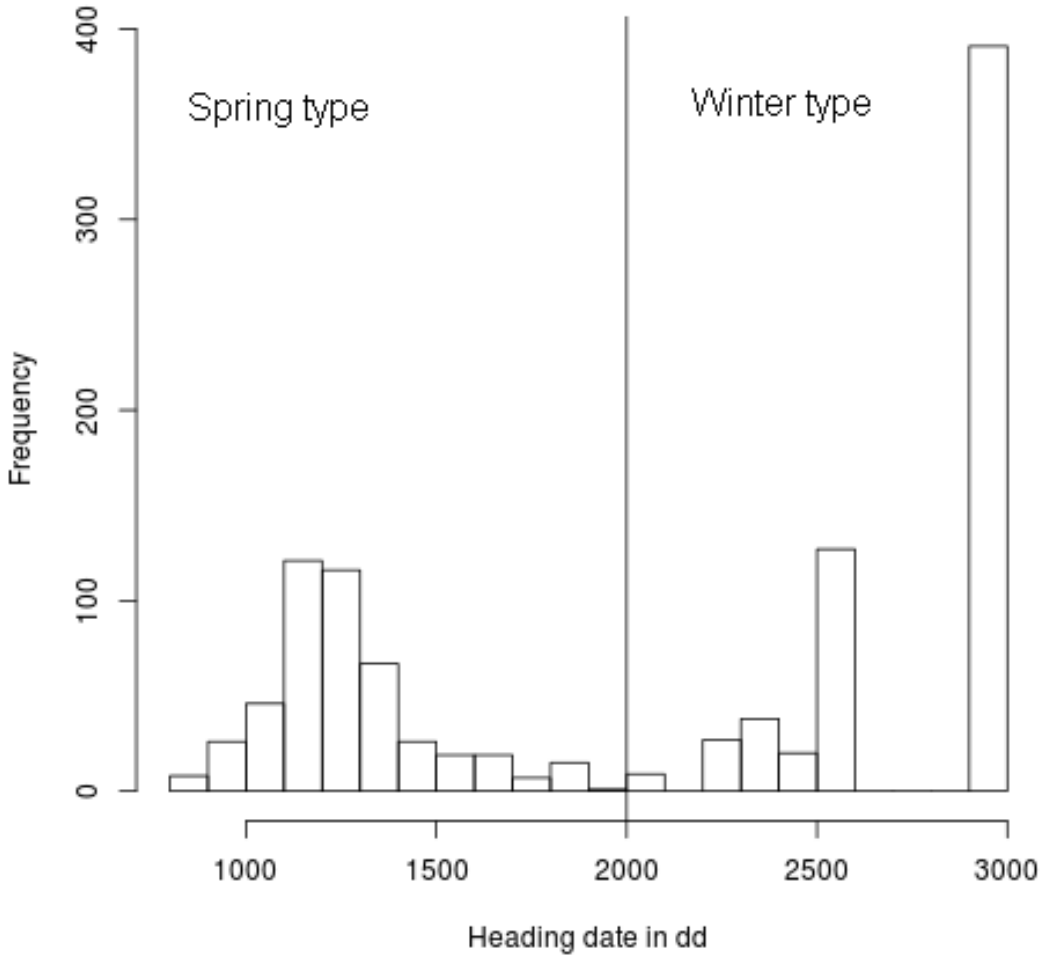
W	S	W			
W	S	NA			
W	H	S			
W	W	S			
NA	S	W			
W	W	W	Winter	79.3% (0.8% S ; 99.2% W)	69.7% (30% S ; 70% W)
W	W	NA			
W	NA	W	NA	1.1%	5%
W	H	W			

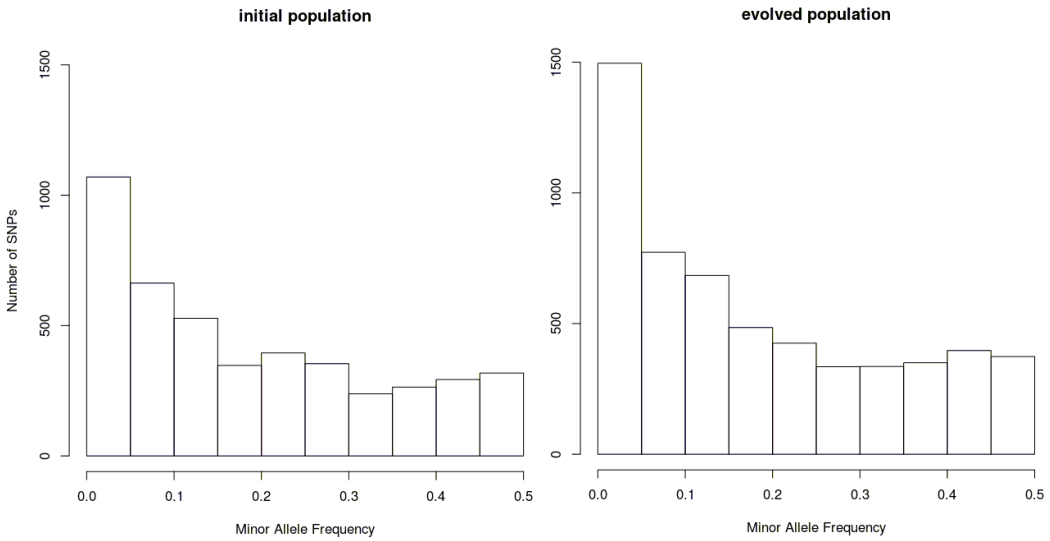
Figure 1: Population creation scheme.

Figure 2: Heading date distribution of all genotypes for spring sowing in 2011, the condition with the least vernalization to determine winter and spring type.

Figure 3: Distribution of minor allele frequency in the initial and evolved populations


59 parents
Manual cross
Probus mutant with male sterile gene
F1
59 parents
Selfing
Manual cross
F2
BC1
Open-pollination
Population G0
12 generations of Random crossing
Harvesting only male sterile spikes
Population G12
4 generations of selfing
1000 SSD lines in F5


Spring type
Winter type
Frequency
0
100
200
300
400
1000
1500
2000
2500
3000
Heading date in dd


initial population
evolved population
Number of SNPs
0
500
1000
1500
0.0
0.1
0.2
0.3
0.4
0.5
Minor Allele Frequency

Chapitre 5

Etude de l'architecture génétique de la précocité de floraison par génétique d'association

5.1 Analyse des données phénotypiques

Ce chapitre est divisé en deux parties. La première partie concerne l'analyse des données de phénotypage mesurées en pépinière au Moulon pendant la thèse obtenues sur les 56 parents et les 1 026 lignées SSD de la population MAGIC INRA. Cette population a en effet été phénotypée dans sept modalités durant ces deux années : un semis de novembre en 2010 et en 2011 avec éclairage naturel, un semis de novembre en 2011 en jours longs, un semis de mars en 2011 et en 2012 et un semis d'avril en 2011 et en 2012 (Chap 2). Après avoir décrit les conditions climatiques de ces différentes expérimentations, la variabilité génétique observée sur l'ensemble des caractères liés à la précocité de floraison mesurés en pépinière sera analysée. Cette partie permettra d'identifier les caractères utilisés dans une analyse de génétique d'association présentée dans la deuxième partie.

5.1.1 Description des données climatiques couvrant la période d'expérimentation

Le phénotypage a été réalisé sur deux années consécutives : 2010-2011 et 2011-2012. Ces deux années ont été très différentes d'un point de vue climatique : l'automne a été plus doux et le printemps plus froid en 2011-2012, qu'en 2010-2011 (Figure 5.1).

Pour étudier les besoins en vernalisation, les semis ont été échelonnés sur trois dates : novembre, mars et avril. Le besoin en vernalisation d'un génotype est défini en nombre de jours vernalisants nécessaires à la levée de l'inhibition de l'initiation florale. L'efficacité de vernalisation (V_{eff}) varie en fonction de la température, et peut être approchée à partir de la température journalière moyenne par une fonction en plateau définie par quatre températures critiques $T1, T2, T3$ et $T4$ respectivement fixées à -4°C , 3°C , 10°C et 17°C par Weir et al. [1984] (Figure 5.2).

FIGURE 5.1 – Evolution des températures moyennes au cours des campagnes 2010-2011 (rouge) et 2011-2012 (vert) et de la longueur du jour (noir).

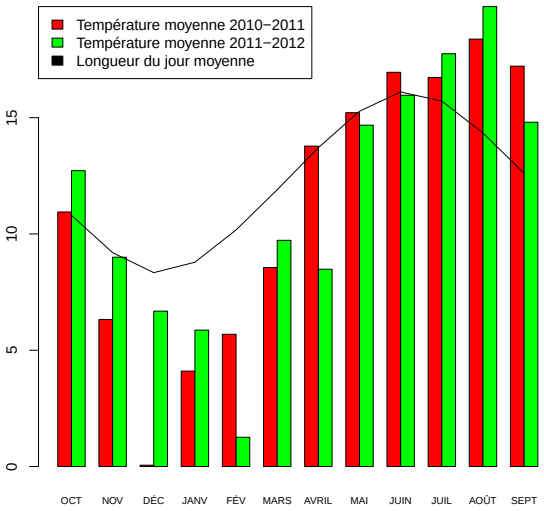
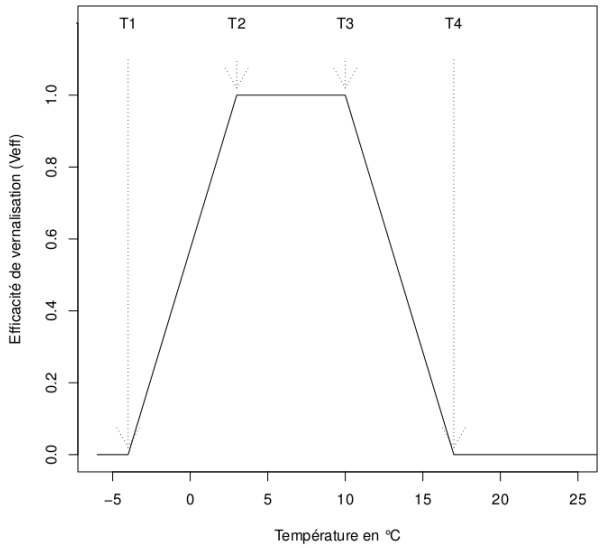


FIGURE 5.2 – Efficacité de vernalisation en fonction de la température moyenne journalière en degré-celsius (°C).



Cette fonction permet de calculer le cumul de vernalisation à partir du semis sur une séquence climatique donnée. Ce cumul montre clairement l'important différentiel de vernalisation entre dates de semis : les plantes semées en hiver ont reçu en

moyenne 36 jours de vernalisation de plus que les plantes semées en mars, qui elles-mêmes ont reçu 19 jours de plus que les plantes semées en avril. Par contre la comparaison des deux années toutes dates confondues montre que les plantes semées en 2010-2011 ont reçues en moyenne 29 jours de vernalisation supplémentaires que celle semées en 2011-2012. En nombre de jours vernalisants, le semis de mars 2011 est plus proche du semis d'avril 2012 (avec 0,5j de différence) que du semis de mars 2012 (Figure 5.3 & Tableau 5.1).

FIGURE 5.3 – Somme cumulée des jours vernalisants pour chaque semis (V_{eff}) depuis le semis. Les semis de "novembre" ont été semés le 27 et 28 octobre.

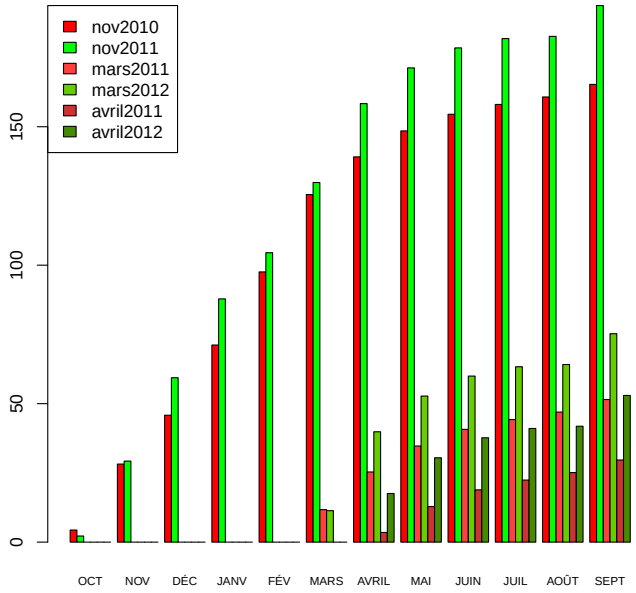


Tableau 5.1 – Tableau de comparaison des nombres de jours vernalisants reçus par les plantes, lors des différents semis. Ces données ont été cumulées sur quatre mois depuis la date de semis.

Dates de semis	Nb de jours vernalisants
nov 2011	100,5
nov 2010	74,0
mars 2012	61,5
mars 2011	41,7
avril 2012	41,2
avril 2011	23,2

En plus de la vernalisation, les différentes dates de semis entraînent des différences significatives de photopériode pendant la croissance des plantes. A la levée les semis d'hiver sont en condition de jours courts ($\sim 8h$) alors que les semis de printemps sont en condition de jours longs ($\sim 14h$), ce qui peut entraîner de fortes différences de phénologie chez les génotypes sensibles à la photopériode (induction par les jours longs). En 2011, pour essayer de différentier les effets de la vernalisation et de la photopériode, une condition supplémentaire a été réalisée, avec un semis de novembre sous éclairage artificiel, pour simuler des jours longs (16h) (Annexe C).

5.1.2 Caractères mesurés

Dans chacune des conditions de cultures, une cinétique de cinq stades de développement a été notée (Protocoles en Annexe E). Ces cinq stades sont :

- Le stade "dernière feuille ligulée" (DFE) : stade noté à la sortie de la ligule de la dernière feuille. Ce stade est important dans les modèles écophysiologiques, car la dernière feuille a un statut particulier, elle a notamment un rôle clé dans le remplissage du grain : c'est la dernière feuille à entrer en sénescence, elle participe donc de façon tardive à la photosynthèse.
- Le stade "épiaison" : stade noté de manière fiable, car l'émergence de l'épi de sa gaine se repère aisément, et elle s'effectue en quelques jours. A ce stade, de bonnes conditions de croissance sont déterminantes, car les stress (lumière, température, azote) peuvent entraîner l'avortement d'une partie des fleurs initiées.
- Le stade "floraison" : stade délicat à noter puisqu'il est caractérisé par la sortie des étamines des fleurs fertiles. L'extrusion des étamines peut se faire en quelques heures en cas de chaleur. Des notations quotidiennes seraient nécessaires pour le noter précisément. Sa forte corrélation avec le stade épiaison fait que le stade épiaison est plus souvent utilisé dans l'étude de précocité de floraison.
- Le "stade 5" et le "stade 7" : ces deux stades permettent d'approcher la durée de remplissage du grain à partir de la sénescence de la dernière feuille et de la couleur de l'épi (Annexe D). Cette notation est délicate car une sénescence précoce peut être entraîner par des maladies. Le stade 7 représente la fin du remplissage du grain. La durée de remplissage est importante pour les sélectionneurs, car une stratégie face au réchauffement climatique serait d'obtenir des variétés qui fleurissent rapidement mais qui ont une longue durée de remplissage du grain. Pendant le remplissage, des conditions climatiques trop chaudes ou trop humides ont de gros impacts sur le rendement notamment le poids de mille grains ou la qualité du grain (teneur en protéines).

Cette cinétique de stades a été mesurée sur deux années, sous des conditions contrastées de température et photopériode, afin de révéler la réponse à différents niveaux de vernalisation et la réponse aux jours longs. Pour les semis de printemps, les plantes les plus sensibles à la vernalisation ne fleurissent pas, et afin d'éviter

les données manquantes, nous avons fixé arbitrairement leur floraison à 3100dj (valeur supérieure aux dernières dates de floraisons observées). Le semis d'avril 2011 présente le moins de jours de vernalisation (Figure 5.3 & Tableau 5.1) : la distribution bimodale bien marquée des caractères de précocité correspond aux génotypes hiver et printemps (Figure 5.4). Les génotypes "printemps" ayant de faibles besoins en vernalisation, atteignent le stade épiaison bien avant les génotypes "hiver". Le stade épiaison mesuré sur ce semis a donc été utilisé pour déterminer le caractère "printemps" ou "hiver" des génotypes, les types "hiver" épiant après le seuil de 2000dj (ou pas du tout). La cohérence entre les deux années a été vérifiée, et 15 génotypes incohérents (ou intermédiaires) ont été notés avec une donnée manquante.

5.1.3 Méthodes d'analyses statistiques

5.1.3.1 Description des parents et des lignées SSD

Une comparaison des différents traits mesurés sur les parents et les lignées SSD a été réalisée sur la base des distributions et à l'aide d'une comparaison de moyenne utilisant le modèle 5.1.

$$Y_{ijkl} = \mu + Geno(Pop)_{il} + Rep(An)_{jk} + An_k + Pop_{il} + \varepsilon_{ijkl} \tag{5.1}$$

$$\varepsilon \sim N(0, \sigma^2)$$

avec Y_{ijkl} le phénotype de l'individu i mesuré sur la ligne de pépinière de la répétition j et de l'année k, μ la moyenne globale, $Geno_i$ l'effet génotype, An_k l'effet année 2010-2011 et 2011-2012, $Rep(An)_{jk}$ l'effet répétition dans chaque année, Pop_l l'effet population parents ou lignées SSD et ε l'erreur. L'effet génotype a été déclaré en aléatoire.

Pour les caractères mesurés en semis d'avril, le modèle utilisé était le modèle 5.1 sans effet "Rep". Pour ceux en semis de novembre en jour long, le modèle a dû être simplifié avec seulement un effet "Pop".

5.1.3.2 Estimation de l'héritabilité et de l'effet année

L'héritabilité a été calculée à partir de l'équation 5.2 avec σ_g la variance génétique estimée à partir du modèle 5.3. Les héritabilités des parents d'une part et des lignées SSD d'autre part ont été calculées par caractère et par année, pour les conditions qui présentent deux répétitions : les semis de novembre et de mars.

$$h^2 = \frac{\sigma_g}{\sigma_g + \sigma_e} \tag{5.2}$$

avec σ_g la variance génétique et σ_e la variance environnementale.

$$Y_{ij} = \mu + Geno_i + Rep_j + \varepsilon_{ij} \tag{5.3}$$

$$\varepsilon \sim N(0, \sigma^2)$$

avec Y_{ij} le phénotype de l'individu i mesuré sur la ligne de pépinière de la répétition j et de l'année k, μ la moyenne globale, $Geno_i$ l'effet génotype, Rep_j l'effet répétition de chaque année et ε l'erreur. L'effet génotype a été déclaré en aléatoire.

L'effet année des caractères a été testé par une analyse de variance (Equation 5.4) pour les différentes modalités d'expérimentation.

$$Y_{ijk} = \mu + Geno_i + Rep(An)_{jk} + An_k + (An \times Geno)_{ik} + \varepsilon_{ijk} \tag{5.4}$$

$$\varepsilon \sim N(0, \sigma^2)$$

avec Y_{ijk} le phénotype de l'individu i mesuré sur la ligne de pépinière de la répétition j et de l'année k, μ la moyenne globale, $Geno_i$ l'effet génotype, An_k l'effet année 2010-2011 et 2011-2012, $Rep(An)_{jk}$ l'effet répétition de chaque année, $(An \times Geno)_{ij}$ est l'interaction entre le génotype et l'année et ε l'erreur. L'effet génotype a été déclaré en aléatoire.

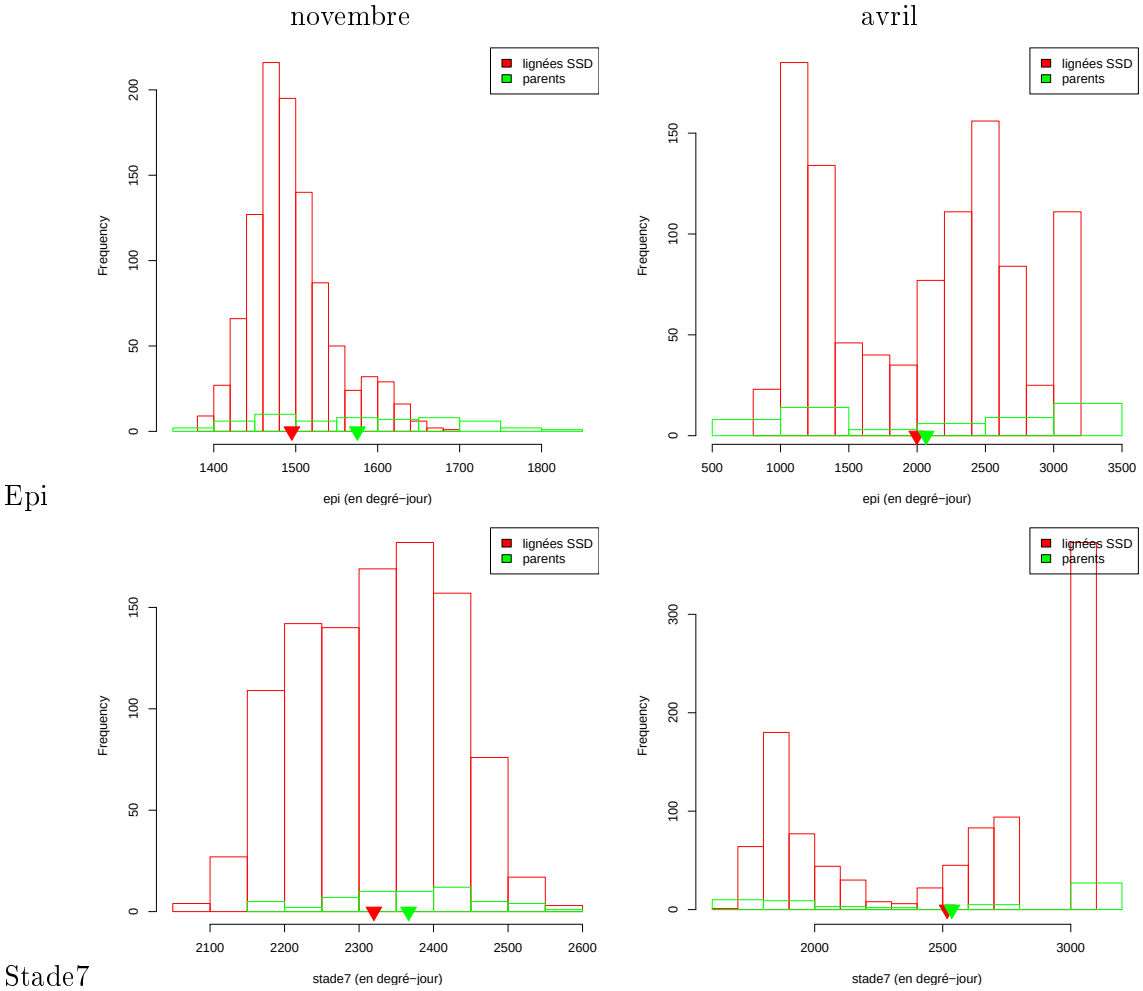
5.1.4 Résultats

5.1.4.1 Description des parents et des lignées SSD

Les distributions des caractères phénotypés montrent des formes plutôt unimodales pour les semis de novembre (lumière naturelle ou jours longs) et de plus en plus bimodales pour les semis de printemps (Figure 5.4 & Figure 5.13). Une tri-modalité peut aussi apparaître artificiellement à cause des génotypes qui n'ont pas atteint le stade observé en fin de notation.

Les moyennes des parents et des lignées SSD ont été comparées pour toutes les combinaisons caractères × conditions. Quel que soit le caractère × condition, les parents sont toujours en moyenne plus tardifs (entre 17 et 156dj) et avec une variance supérieure aux lignées SSD (Tableau 5.3). L'analyse de variance (Equation 5.1) a mis en évidence une différence très significative ($p - value < 10^-3$) pour 75% des caractères × conditions (Tableau 5.3). Les caractères × conditions pour lesquels il n'y a pas de différence sont uniquement les caractères mesurés avec le semis de mars.

FIGURE 5.4 – Distribution des moyennes ajustées sur deux années et deux répétitions (si disponibles) des stades *Epiaison* et *stade 7* sur les semis de novembre et d'avril, mesurés sur les 1 026 lignées SSD (rouge) et les 56 parents (vert). Les triangles représentent les moyennes des lignées SSD et des parents. Les autres graphiques sont disponibles dans la Figure 5.13 dans le chapitre 5.1.7.

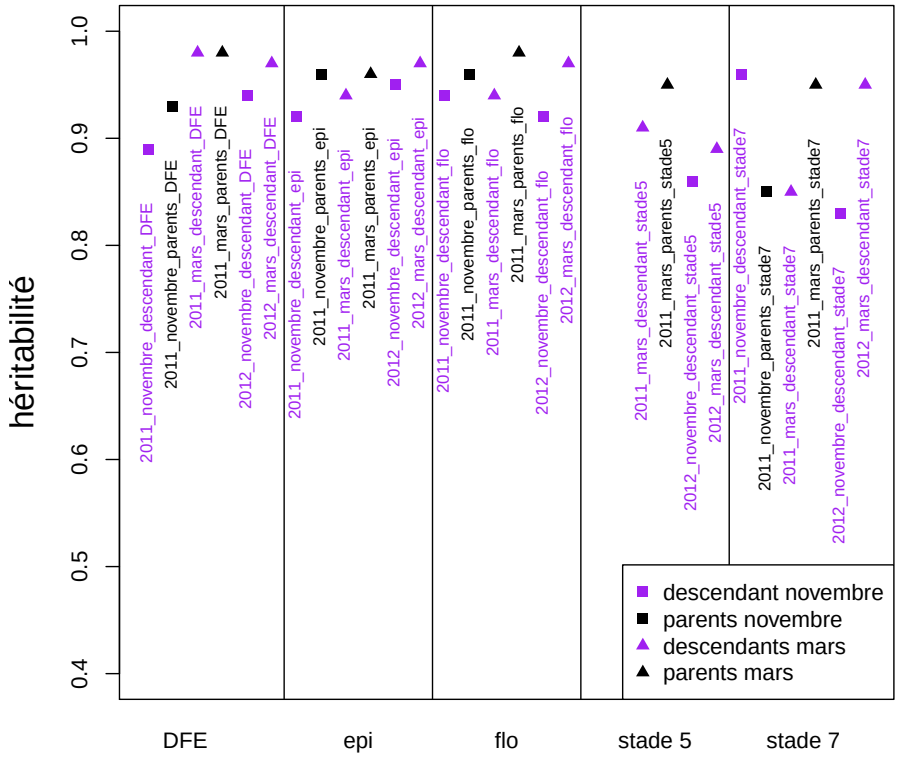


5.1.4.2 Estimation de l'héritabilité et de l'effet année

Une héritabilité a été calculée par caractère et par modalité pour les parents et les lignées SSD séparémment. Tous les caractères, excepté le stade 5 mesuré sur le semis de novembre 2010 (non représenté sur la figure 5.5), ont une héritabilité supérieure ou égale à 0,83 aussi bien pour les parents que pour les descendants (Figure 5.5, Equation 5.2). L'héritabilité négative observée sur le stade 5 est due à une très forte verse sur le semis de novembre 2010, qui a entravé les notations pour un grand nombre de génotypes (cette notation a été écartée des analyses). Après avoir redressé toutes les plantes, nous avons par contre pu mesurer le stade 7 pour la majorité des génotypes. La forte héritabilité des autres caractères montre que le

dispositif expérimental est robuste et que les caractères étudiés ont une forte variance phénotypique d'origine génétique. La comparaison des héritabilités a montré un léger effet de la date de semis sur l'héritabilité, avec une héritabilité des caractères plus forte avec un semis de mars qu'avec un semis de novembre (+0,03). Ces résultats sont en accord avec la plus forte variance des phénotypes en semis de printemps, distribués sur une plus large gamme. L'héritabilité estimée avec les parents est souvent plus élevée que celles estimées avec les lignées SSD, ce qui signifie qu'en variance les parents présentent une plus grande variabilité.

FIGURE 5.5 – Héritabilité des caractères par année et par population (parents/descendants). L'héritabilité n'a pu être calculée que pour les caractères mesurés avec les semis de novembre et de mars (2 répétitions). L'héritabilité du stade 5 pour le semis de novembre 2010 n'a pas été représentée pour une meilleure lisibilité (valeur proche de 0).



L'analyse de variance montre également qu'il existe un effet d'interaction génotype × année dans toutes les conditions et pour tous les stades. Cette interaction est plus marquée pour les caractères mesurés avec un semis de printemps avec des p-values variant entre $8,78.10^{-253}$ et 9.42^{-103} qu'avec le semis de novembre avec des

p-values comprises entre $6,54.10^{-69}$ et $3,98.10^{-5}$ (Tableau 5.2).

Tableau 5.2 – Résultats (p-value) des analyses de variance testant l'effet année, génotype (aléatoire), répétition et interaction année x génotype pour les différents caractères mesurés dans les conditions répétées deux années.

stade	semis	année	génotype	répétition	génotype x année
DFE	novembre	0	0	8,22e-7	1,67e-33
Epiaison	novembre	0	0	2,59e-17	2,95e-42
Floraison	novembre	0	0	2,63e-26	6,51e-69
Stade 5	novembre	1,82e-220	7,95e-93	4,73e-6	7,55e-17
Stade 7	novembre	0	0	1,12e-21	3,98e-5
DFE	mars	0	1,62e-104	0,023	8,78e-253
Epiaison	mars	0	1,14e-192	3,43e-12	2,52e-171
Floraison	mars	0	7,77e-171	2,47e-15	2,42e-171
Stade 5	mars	0	3,94e-146	1,07e-55	4,41e-132
Stade 7	mars	6,53e-21	9,53e-6	6,61e-5	9,42e-103
DFE	avril	NA	NA	NA	NA
Epiaison	avril	2,27e-29	5,84e-93	NA	NA
Floraison	avril	1,09e-6	1,87e-76	NA	NA
Stade 5	avril	1,26e-6	5,67e-50	NA	NA
Stade 7	avril	1,46e-12	3,66e-48	NA	NA

5.1.4.3 Corrélations entre les caractères

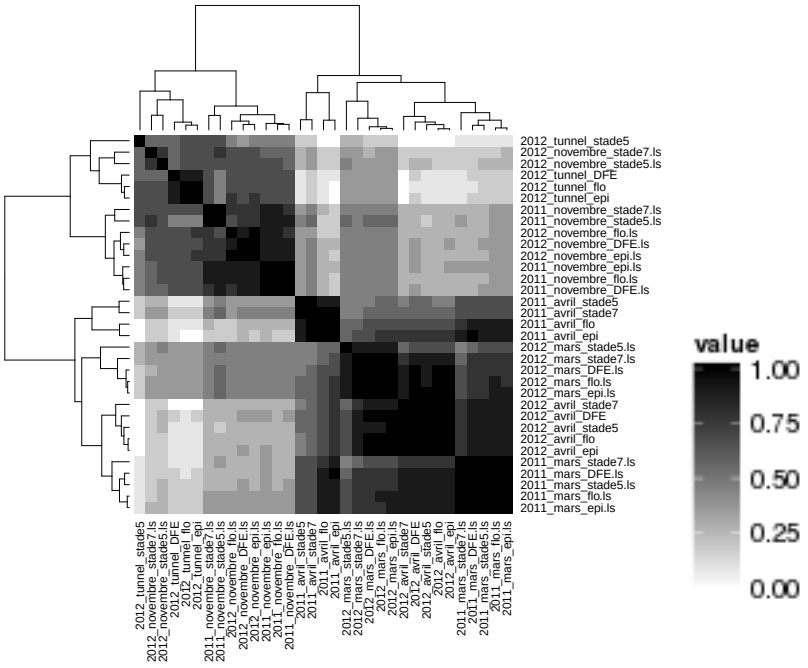
Comme les traits étudiés constituent une série chronologique, mesurés successivement sur les mêmes plantes, ou que ces mesures ont été répétées sur deux années dans des conditions environnementales différentes, ils présentent nécessairement des corrélations. La similarité entre les caractères × conditions × années est identifiable avec un clustering sur les corrélations (Figure 5.6). Deux groupes se distinguent, un groupe rassemblant les stades mesurés sur les semis d'automne (jours naturels et jours longs (tunnel)), et un autre groupe mesuré sur les semis de printemps. Pour les semis de printemps, les caractères sont d'autant plus corrélés qu'ils ont été mesurés sur une même condition × année. Au contraire, pour les stades de semis d'automne, les caractères sont moins clairement regroupés. Cela valide le fait que les semis d'automne sont plus proches entre eux que les semis de printemps entre eux.

Dans la suite des analyses, deux caractères seront gardés pour réaliser l'analyse de génétique d'association : l'épiaison et un caractère de remplissage du grain correspondant à la différence entre l'épiaison et le score 7. Les deux années pour les conditions de semis de printemps ont été prises indépendemment les unes des autres car elles sont trop différentes pour être considérées comme l'observation d'un même caractère. A l'inverse, les deux années d'observations en semis d'automne (condition de jour naturel) ont été moyennées (moyenne ajustée par répétition et

année) pour ne former qu'une seule variable.

Nous garderons les deux caractères mesurés dans six modalités : les semis de novembre avec éclairage naturel, le semis de novembre en 2011 en jours longs, le semis de mars en 2011 et en 2012 et le semis d'avril en 2011 et en 2012.

FIGURE 5.6 – Corrélations entre tous les caractères phénotypés dans les différentes conditions. L'arbre de similarité a ordonné les variables afin de rassembler les plus corrélées.



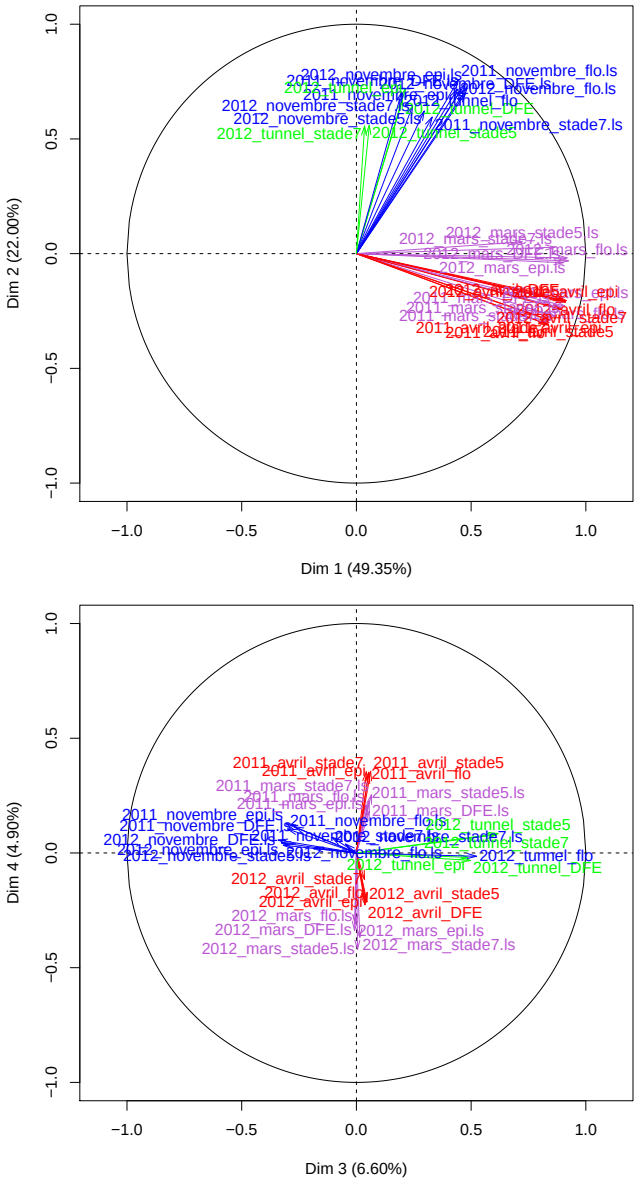
Je présente ensuite deux méthodes d'extraction d'informations à partir d'un grand nombre de variables de ce jeu de données phénotypiques : une analyse en composantes principales et un modèle écophysiologique.

5.1.5 Analyse multi-variées

Vu le nombre de caractères phénotypés, le nombre de conditions (5×6) et leur corrélation, les variables ne peuvent pas toutes être étudiées indépendamment. Une analyse en composantes principales (ACP) permet de créer des nouvelles variables indépendantes (axes de l'ACP), combinaisons linéaires des variables initiales. Les premiers axes de l'ACP maximisent la part de variance globale expliquée, et synthétisent ainsi l'information apportée par toutes les variables. Une ACP normée sur l'ensemble des caractères phénotypés explique 82,85% de la variation phénotypique sur les quatre premiers axes. Le premier axe de l'ACP, qui explique 49,35% de la variation phénotypique, est composé essentiellement des caractères mesurés sur semis de printemps (Figure 5.7), et représente donc principalement le besoin en vernalisation des individus. Le deuxième axe de l'ACP, qui explique 22%

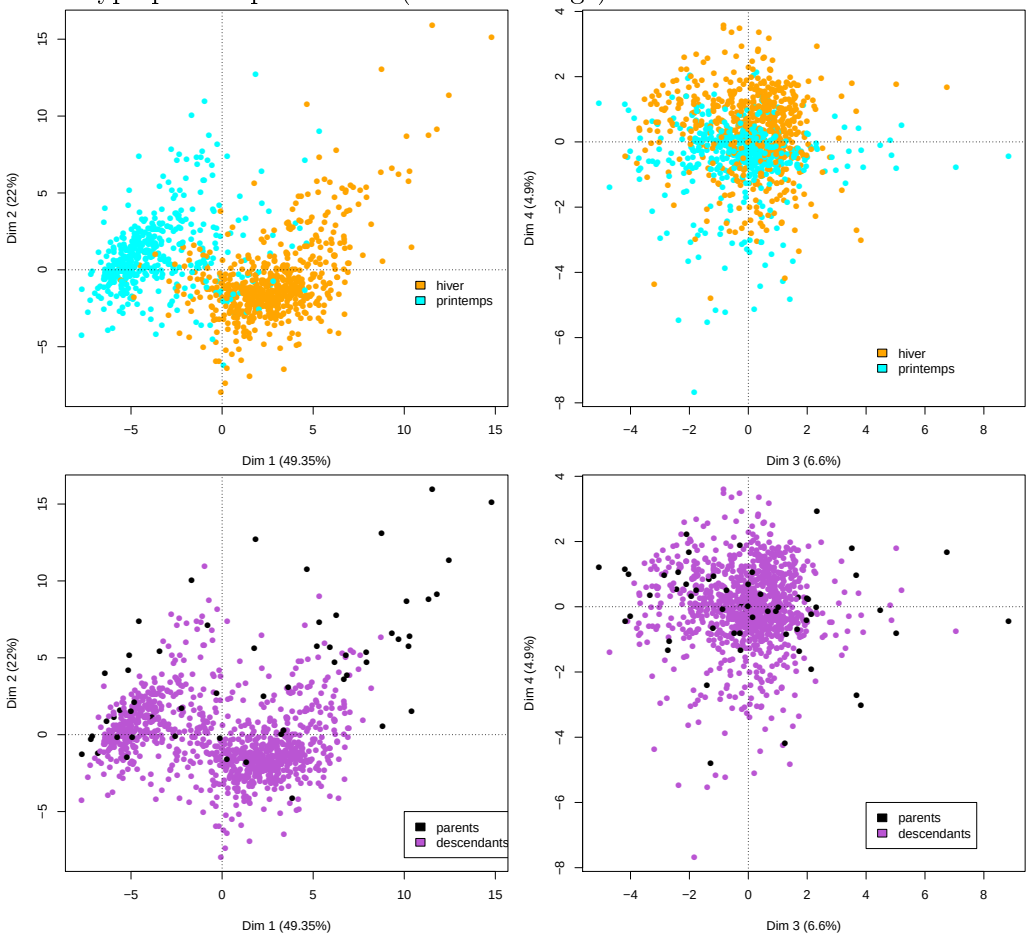
de la variation phénotypique, rassemble les traits des semis d'automne ; il semble représenter la composante de précocité intrinsèque. Le troisième axe oppose les traits des semis d'automne au champs avec une photopériode naturelle et au tunnel avec des conditions de jours longs, il représente donc la sensibilité à la photopériode. Le quatrième axe oppose les traits de semis de printemps 2011 à ceux de 2012, ce qui est plus difficile à interpréter.

FIGURE 5.7 – Projections des variables sur les deux plans définis par les quatre premiers axes de l'ACP avec tous les caractères phénotypés. Les variables en vert représentent les caractères phénotypés en semis de novembre avec éclairage artificiel (jours longs), en bleu ceux en semis de novembre en lumière naturelle, en violet ceux en semis de mars et en rouge ceux en semis d'avril.



La projection des individus selon leur type printemps ou hiver sur ces quatre axes, confirme que l'axe 1 représente le besoin en vernalisation, comme l'indique la séparation assez claire qu'il effectue entre les deux groupes (Figure 5.8). L'axe 4 représente aussi une composante de vernalisation, mais moins marquée (séparation moins claire qu'avec l'axe 1). La projection des individus montre que les parents et les lignées SSD ont une grande variabilité pour les besoins en vernalisation et la précocité intrinsèque (axe 1 et 2), avec une légère différenciation des deux populations selon ces deux axes. La répartition des parents est légèrement moins éclatée sur les axes 3 et 4, même si les descendants conservent une distribution plus agrégée (Figure 5.8).

FIGURE 5.8 – Projections des individus sur les quatre premiers axes de l'ACP, colorés en fonction de leur population parents ou descendants (noir ou violet) ou en fonction de leur type printemps ou hiver (bleu ou orange).



Les quatre premiers axes de l'ACP représentent bien la grande diversité des données phénotypiques (82,85% de la variabilité phénotypique) et semblent correspondre aux différentes composantes de la précocité de floraison. Les coordonnées des individus sur ces quatre axes seront utilisées pour les analyses de

génétique d'association.

5.1.6 Caractérisation de la précocité de floraison des génotypes à l'aide d'un modèle écophysiologique

Nous avons vu dans la section précédente que malgré la forte héritabilité des caractères phénotypés dans une condition particulière, les mesures d'un même caractère évalué sur deux années, ou deux dates de semis différentes, pouvaient être faiblement corrélées. Ceci tient à l'importance des interactions Génotype × Environnement (G × E) pour bon nombre de caractères quantitatifs, et notamment pour la précocité de floraison. Afin de synthétiser l'information contenue dans les données de phénotypage, nous avons voulu utiliser un modèle écophysiologique, apte à intégrer une partie de ces interactions GxE. Par définition, un modèle écophysiologique caractérise la variabilité génétique à travers des valeurs de paramètres qui sont indépendants de l'environnement. C'est le formalisme du modèle qui génère donc des interactions GxE, ces dernières étant des propriétés émergentes du modèle. L'idée dans notre cas est de montrer que des paramètres de modèle, permettant de reproduire fidèlement le comportement d'un génotype dans une gamme d'environnements sont des caractères plus héritables que des caractères mesurés et soumis à des interactions GxE fortes. Pour ce faire il s'agit donc d'optimiser les valeurs de paramètres pour chaque génotype de notre panel dans les environnements pour lesquels nous avons des données de phénotypage. Ces valeurs de paramètres seront ensuite considérées comme des caractères dont nous étudierons le déterminisme génétique. Un premier élément de choix réside dans la sélection d'un formalisme (d'un modèle) adapté à nos données et à nos objectifs. Nous nous sommes orientés vers un modèle inspiré du modèle ARCWHEAT [Weir et al., 1984] qui caractérise la précocité de floraison de chacun des génotypes à l'aide de deux paramètres : le premier quantifie le besoin en vernalisation, et le second la sensibilité à la photopériode. Cette même approche a été précédemment utilisée par M. Bogard et V. Allard (INRA UMR GDEC, Clermont Ferrand, Bogard et al, under review). Pour réaliser l'optimisation de ce modèle seules les données de phénotypage de la première année 2010-2011 et les données de semis de novembre 2011 en condition de jours longs ont été utilisées.

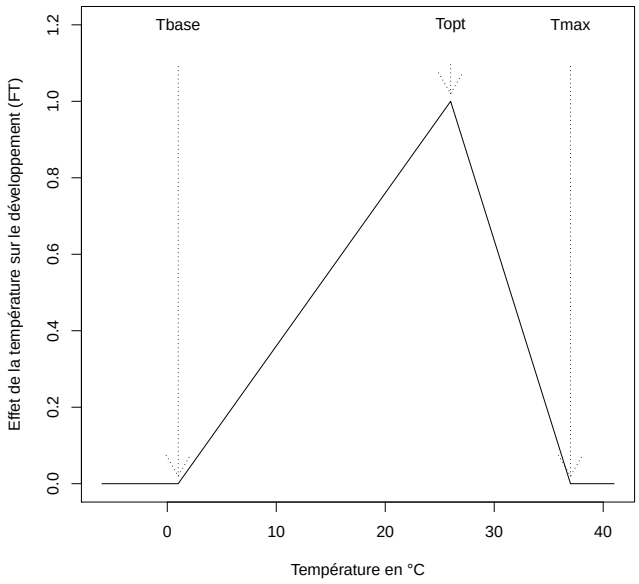
5.1.6.1 Description du modèle

Ce modèle est basé sur l'accumulation de temps thermiques, modulée par l'effet du besoin en vernalisation et de la sensibilité à la photopériode [Weir et al., 1984]. Il utilise les températures moyennes journalières et les durées du jour journalières en variables d'entrée. Les trois paramètres clés du modèle sont le temps thermique (T_t), le besoin en vernalisation (FV) et la sensibilité à la photopériode (FP). La composante temps thermique T_t est calculée à partir de FT (Equation 5.5) défini par une fonction bilinéaire (Figure 5.9) avec trois températures critiques Tbase, Topt et

Tmax respectivement fixées à 1°C, 26°C et 37°C [Weir et al., 1984].

$$T_t = FT \times T_{opt} \tag{5.5}$$

FIGURE 5.9 – Effet de la température sur le développement de la plante en fonction de température en degré-celsius (°C).



La composante de vernalisation FV, comprise entre 0 et 1, est calculée à partir du nombre de jours vernalisants (VDD) (Equation 5.6). VDD est défini comme la somme des jours vernalisants (V_{eff}, Figure 5.2) depuis le semis jusqu'à l'épiaison (Figure 5.3).

$$FV = \frac{VDD - V_{base}}{V_{sat} - V_{base}} \tag{5.6}$$

Le paramètre V_{base} est fixé à 0. V_{sat} est un paramètre génétique à optimiser caractérisant le besoin en vernalisation spécifique de chaque génotype. La composante de photopériode FP, comprise entre 0 et 1, est définie par une fonction de la longueur du jour quotidienne (Ph) et deux paramètres P_{opt} et P_{base} (Equation 5.7). P_{opt} est fixé à 20h alors que P_{base} est le paramètre de sensibilité à la longueur du jour à optimiser.

$$FP = \frac{Ph - P_{base}}{P_{opt} - P_{base}} \tag{5.7}$$

Dans ce modèle le temps thermique cumulé est modifié par la composante de vernalisation et de photopériode comme décrit dans l'équation 5.8. Le modèle prédit une date d'épiaison correspondant à la date où l'accumulation du temps thermique

modifié par le frein de vernalisation et de photopériode (PVT) atteint un seuil, qui a été fixé à 782dj.

$$PVT_t = T_t \times FV \times FP \tag{5.8}$$

5.1.6.2 Optimisation des paramètres

Les deux paramètres V_{sat} et P_{base} ont été optimisés par la méthode exhaustive de calcul d'erreur (RMSE ; Equation 5.9) pour tous les couples de paramètres avec V_{sat} variant de 0 à 130jours avec un pas d'un jour et P_{base} de 0 à 10h avec un pas de 0,1h. Dans le cas où plusieurs couples minimisent l'erreur, celui minimisant aussi le biais est choisi (Equation 5.10).

$$RMSE_i = \sqrt{\frac{\sum_{j=1}^{n}(Obs_{ij} - Pred_{ij})^2}{n}} \tag{5.9}$$

$$Biais_i = \frac{\sum_{j=1}^{n}|Obs_{ij} - Pred_{ij}|}{n} \tag{5.10}$$

avec Obs_{ij} et $Pred_{ij}$ les dates d'épiaison en jours observées et prédites du génotype i pour les conditions j et n le nombre de conditions. Si plusieurs couples minimisent l'erreur de prédiction et le biais, alors un couple médian est choisi. Dans ce cas les différences de V_{sat} et P_{base} entre les couples restants sont très faibles. Cela dénote donc une faible sensibilité du modèle dans cette partie de l'espace des paramètres mais aussi un impact très limité sur la qualité de l'estimation des paramètres. Un problème est posé par la présence de gènes à effet fort dans le panel. En effet la population contient des génotypes de types printemps qui épient avec des semis de printemps et des génotypes de type hiver qui n'épient pas dans ces conditions. Or, la méthode d'optimisation utilisée permet de gérer des variables quantitatives (date d'épiaison) mais pas une variable qualitative du type « absence d'épiaison ». De ce fait, les groupes « hiver » et « printemps » ont été gérés de manières différentes. Pour les types printemps, l'ensemble des conditions ont été prises en compte pour l'étape d'optimisation des paramètres du modèle. Pour les types hiver, seul le semis d'hiver a été pris en compte pour l'optimisation mais un critère de tri qualitatif a été ajouté pour sélectionner les couples de paramètres optimisés : la prédiction du modèle dans le cas d'un semis de printemps devait être postérieure d'une semaine au moins à la dernière date de notation réalisée l'année d'expérimentation considérée.

5.1.6.3 Résultat sur la population MAGIC INRA

Au total, les paramètres V_{sat} et P_{base} ont été optimisés pour 1 113 génotypes. Un unique couple de paramètres a été estimé pour chacun des génotypes. 96 des génotypes ont une erreur de prédiction (RMSE) exorbitante (supérieure à 60 000). Le fait que, les valeurs de P_{base} et V_{sat} « optimales » soient identiques entre ces génotypes avec des valeurs respectives de 8 et 130 démontre clairement une limite des capacités prédictives du modèle et/ou des erreurs de phénotypage. Parmi les autres, les erreurs de prédictions vont de 0 à 40 jours (Figure 5.10). Comme la

population est très diversifiée, les V_{sat} et P_{base} estimés balayent tous les possibles (Figure 5.11).

FIGURE 5.10 – Distribution des erreurs de prédiction pour la population MAGIC INRA (génotypes avec erreur supérieure à 60 000 exclus).

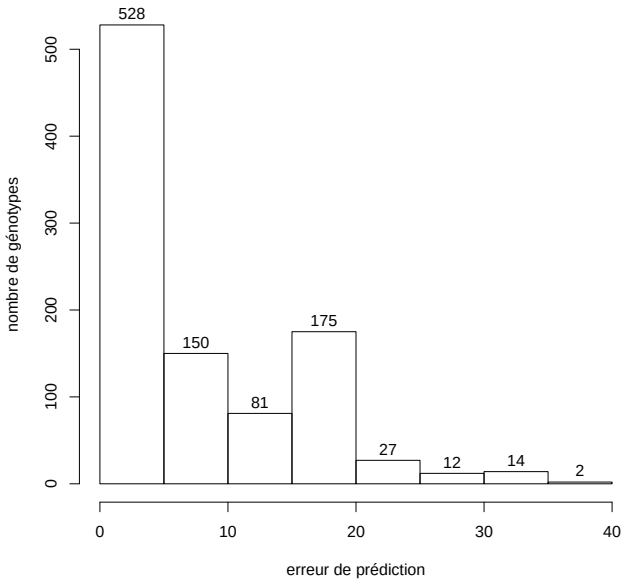
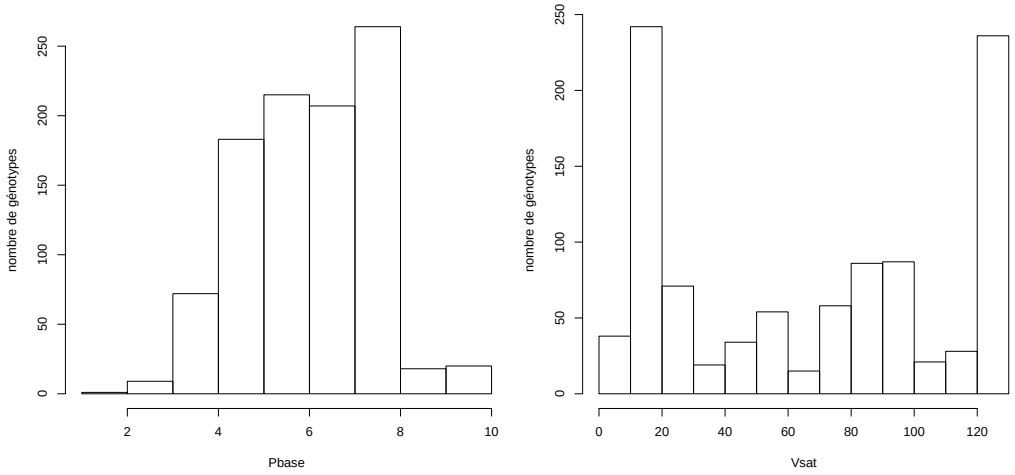


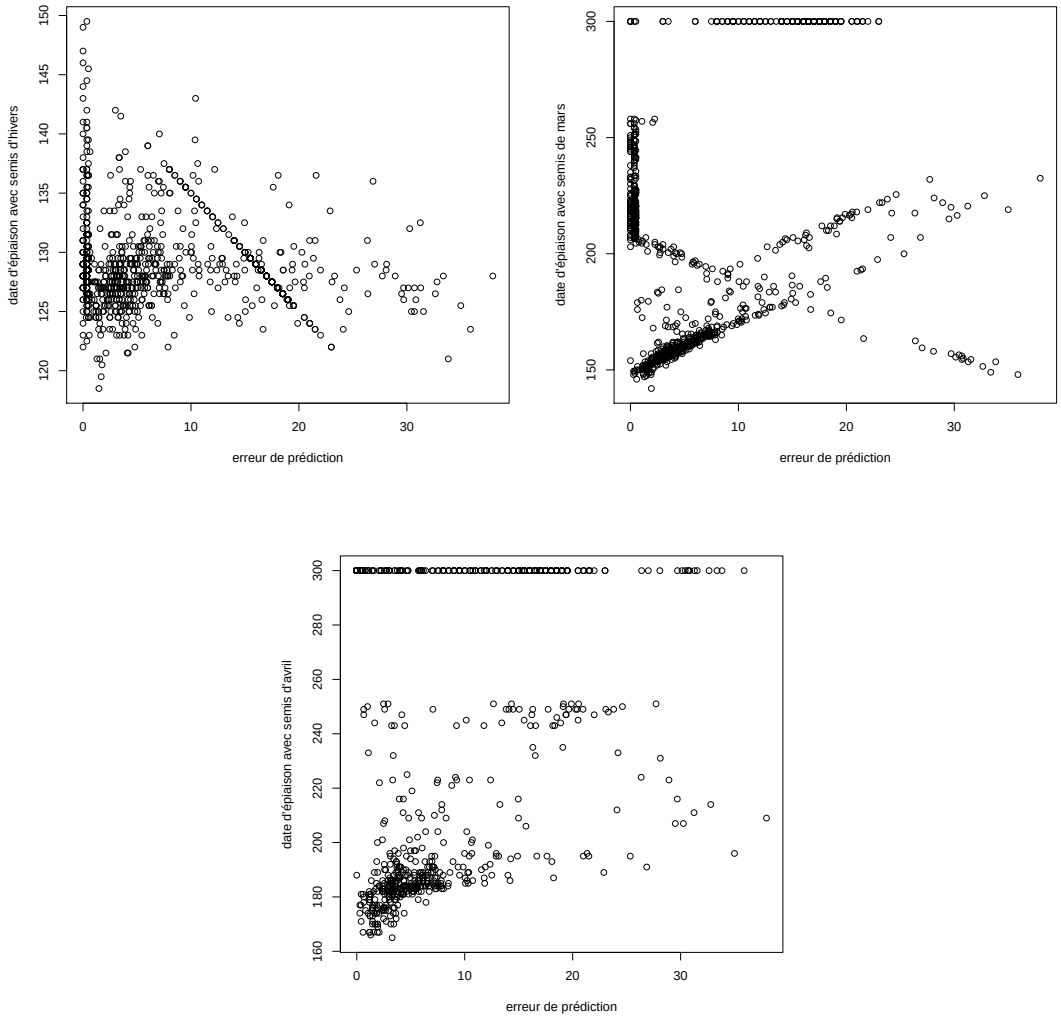
FIGURE 5.11 – Distribution des paramètres P_{base} à gauche et V_{sat} à droite pour la population MAGIC INRA (génotypes avec erreur supérieure à 60 000 exclus).



5.1.6.4 Discussion

En tentant de minimiser l'erreur de prédiction du modèle, l'algorithme a estimé des valeurs de paramètres sur l'ensemble des génotypes, avec des erreurs inacceptables pour 9% des génotypes. Des erreurs même légères lors du phénotypage peuvent conduire à des difficultés massives lors de l'étape d'optimisation. Une étude rapide du lien entre les erreurs du modèle et l'hétérogénéité des lignes de pépinière, à la recherche d'éventuelles erreurs de notation, n'a pas été concluante. Nos résultats montrent cependant clairement que les prédictions ne sont pas biaisées. En effet l'erreur de prédiction est indépendante de la précocité (Figure 5.12). De ce fait, la fixation d'un seuil arbitraire d'erreur considéré comme acceptable pour faire les tests d'association n'aura *a priori* aucun impact sur la diversité génétique représentée dans le panel choisi par rapport à la population de départ. Ce seuil d'erreur a été fixé à 10 jours sur l'ensemble des quatre expérimentations. 678 génotypes (49 parents & 629 lignées SSD soit 61% de la population) ont une erreur de prédiction inférieure à ce seuil et seront donc utilisés pour faire de la génétique d'association.

FIGURE 5.12 – Relation entre l'erreur de prédiction et les dates d'épiaison des semis de novembre, de mars (de gauche à droite) et d'avril (en dessous) (génotypes avec erreur supérieure à 60 000 exclus).



5.1.7 Figures Supplémentaires de l'analyse phénotypique de la population MAGIC INRA

FIGURE 5.13 – Distribution des moyennes ajustées par année et par répétition (si disponible) de tous les stades relevés par date de semis pour les 1 026 lignées SSD (rouge) et les 56 parents (vert). Les triangles représentent les moyennes des lignées SSD et des parents.

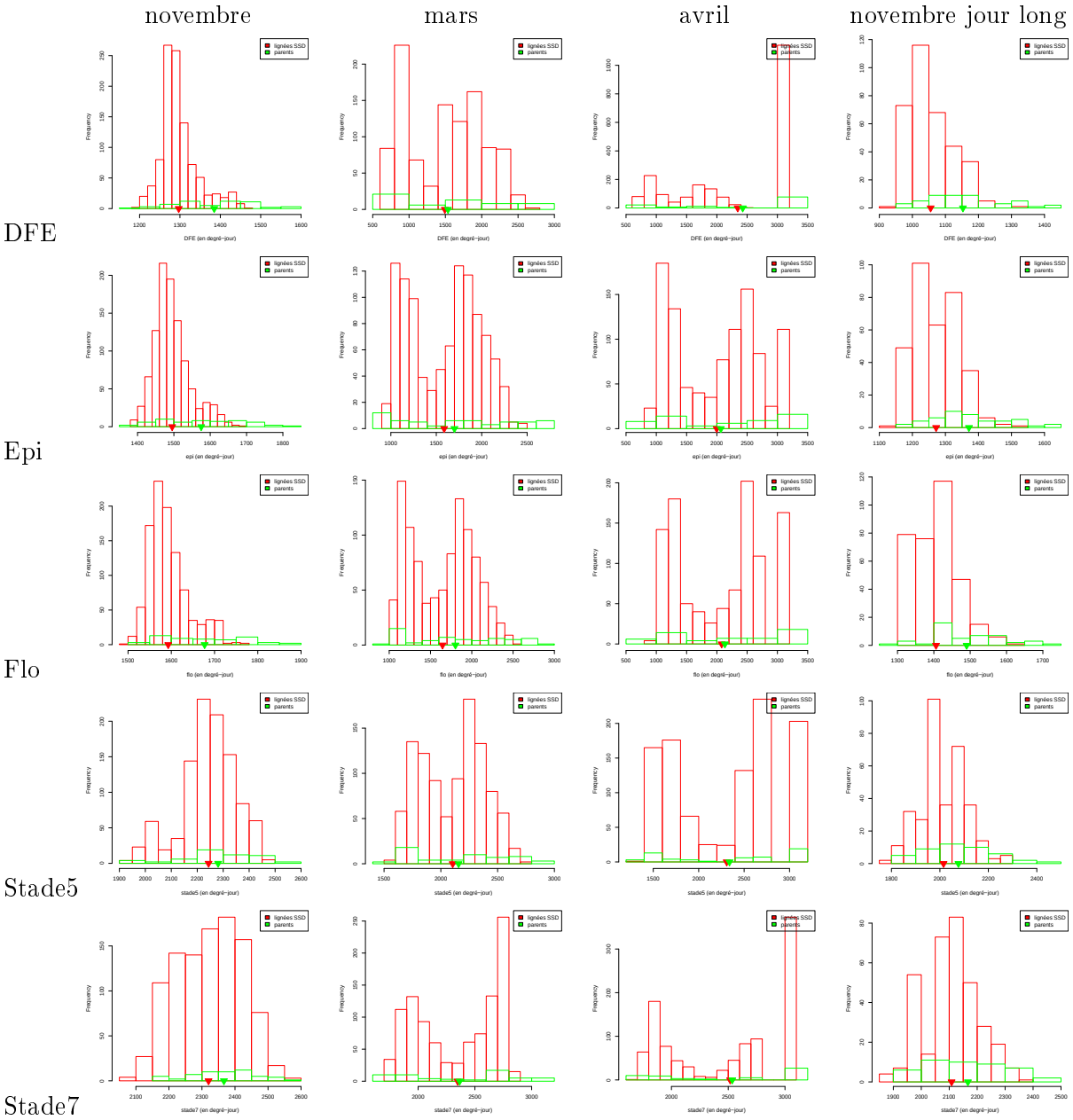
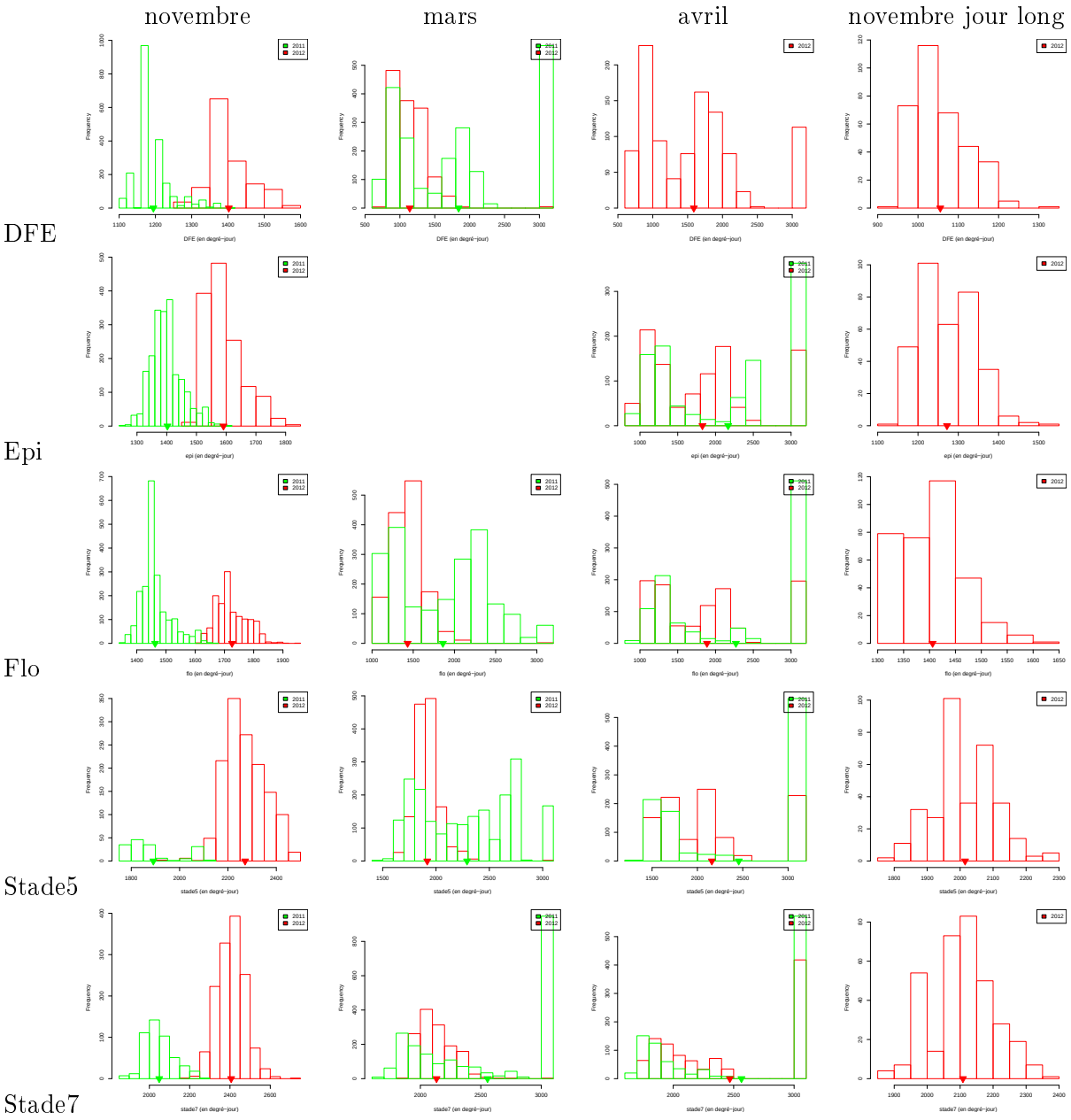


Tableau 5.3 – Tableau répertoriant pour chaque trait et chaque modalité les moyennes et variances observées avec les 56 parents et les 1 026 lignées SSD, ainsi que le résultat de l'analyse de variance correspondante au modèle 5.1.

Modalités	Traits	Moyenne parents	Variance Parents	Moyenne lignées	Variance lignées	P-value
novembre	DFE	1384,84	8850,16	1297,00	2131,79	***
	epi	1575,22	12587,81	1495,24	2495,42	***
	flo	1677,04	9071,59	1592,93	2080,24	***
	stade5	2280,47	21774,66	2243,65	11209,96	***
	stade7	2366,50	10122,74	2319,96	9541,17	***
novembre jour long	DFE	1153,05	1304,88	1055,31	1153,05	***
	epi	1371,22	12759,74	1271,89	1371,22	***
	flo	1490,25	9394,35	1405,99	1490,25	***
	stade5	2076,59	16449,90	2015,03	2076,59	***
	stade7	2166,89	16084,32	2108,51	2166,89	***
mars	DFE	1535,73	551281,98	1492,68	27816,978	NS
	epi	1700,01	382971,43	1586,71	150495,54	.
	flo	1801,93	343543,76	1645,49	144497,22	NS
	stade5	2156,51	191667,07	2104,72	76084,72	NS
	stade7	2365,46	227337,84	2348,32	128447,96	.
avril	DFE	2427,03	887797,40	2344,40	2427,03	***
	epi	2066,30	779262,83	1997,65	488423,24	***
	flo	2132,86	729100,81	2079,44	511545,86	***
	stade5	2338,78	469364,65	2308,63	335808,84	***
	stade7	2535,55	363077,27	2517,26	284117,08	***

NS : non significatif . :$<$0.1 * :$<$0.05 ** :$<$0.01 *** :$<$0.001

FIGURE 5.14 – Distribution de tous les stades relevés par date de semis et par année pour les 1 026 lignées SSD. Pour les conditions dans lesquelles deux répétitions ont été réalisées, la distribution des moyennes ajustées est présentée. Les triangles représentent les moyennes des caractères par année.



5.2 Détection de QTLs par génétique d'association

5.2.1 Introduction

La génétique d'association est une méthode de détection de QTLs, qui repose sur l'identification d'associations statistiques entre les polymorphismes de marqueurs génétiques et la variation d'un caractère quantitatif [Ingvarsson and Street, 2011]. Cette méthode s'appuie sur l'utilisation d'un panel de génotypes puisés dans les ressources génétiques disponibles. Ces génotypes sont choisis pour représenter un pool génétique d'intérêt (pool élite, landrace, zone de culture spécifique) et minimiser l'apparentement entre les génotypes. La génétique d'association peut être utilisée avec deux stratégies différentes : i) une approche locale dans le but de valider un gène candidat ou de valider et préciser la localisation d'un QTL suspecté d'être impliqué dans la variation d'un caractère quantitatif d'intérêt, ii) une approche génome entier, où l'utilisation d'un grand nombre de marqueurs couvrant le génome permet de découvrir des QTLs sans *a priori*.

Contrairement à la cartographie génétique, qui utilise les recombinaisons récentes dans des croisements dédiés, la génétique d'association exploite les recombinaisons ancestrales ayant eu lieu au cours de la différenciation des génotypes étudiés. Ces panels de génotypes bénéficient donc de centaines voire milliers de générations de recombinaisons ce qui permet une grande précision de localisation des QTLs grâce au faible déséquilibre de liaison (DL) présent dans ces populations. Le faible DL présent dans certaines populations naturelles permet même parfois d'identifier un seul gène candidat [Nordborg and Weigel, 2008]. En contrepartie, ce faible DL entraîne des contraintes de génotypage, avec une densité minimale de marquage à atteindre, afin d'avoir de fortes probabilités de trouver certains marqueurs en fort DL avec les loci responsables de la variation du caractère d'intérêt [Rafalski, 2002].

Cependant les panels de génotypes peuvent présenter des fréquences alléliques ou des structures génétiques moins favorables à la détection de QTLs que les populations de cartographie. Ainsi, les apparentements récents qui peuvent exister entre les génotypes étudiés, et/ou l'existence d'une structure génétique plus ancienne dans le panel [Thornsberry et al., 2001; Kang et al., 2008], généralement liée aux différentes origines géographiques représentées [Rousset et al., 2011] ou à des pools génétiques travaillés séparément par les sélectionneurs, entraînent du DL inter chromosome, qui peut générer de fausses associations (faux positifs).

Un avantage, mais également un inconvénient dans l'utilisation de panels de génotypes, provient de leur richesse allélique : à la fois le nombre de QTLs en ségrégation dans le panel sera plus élevé, mais certains allèles de ces QTLs, ainsi que les marqueurs associés, seront présents en faible fréquence (allèle rare : $p < 0{,}1$). L'existence d'allèles rares aux marqueurs/QTLs limite la puissance de détection, l'idéal étant d'avoir des allèles aux marqueurs/QTLs en fréquences alléliques équilibrées ($p=0{,}5$ en biallélique [Clark et al., 2007]).

Les études de génétique d'association ont connu ces dernières décennies un essor important, et de nombreuses méthodes statistiques ont été développées, notamment

pour tenir compte de la structure génétique particulière des panels de lignées. La méthode de détection la plus simple consiste à effectuer une analyse de variance (ANOVA) ou une régression linéaire marqueur par marqueur entre la variation génotypique et phénotypique. Mais lorsque cette analyse est étendue à des données génome entier (Genome wide : GW), composées de plusieurs milliers de marqueurs, elle pose des problèmes de correction des tests multiples réalisés. La correction la plus stricte est celle de Bonferroni [Bland and Altman, 1995] : le seuil de comparaison multiple correspond au seuil de significativité divisé par le nombre de tests. Du fait du grand nombre de marqueurs utilisés dans les approches GW, les marqueurs présentent des liaisons génétiques (DL) et l'hypothèse d'indépendance entre les tests est fausse : le seuil corrigé est trop conservatif. Une alternative est de déterminer ce seuil par la méthode des permutations. Cette méthode détermine le seuil sous l'hypothèse nulle (le marqueur n'est pas associé à un QTL), en fonction du nombre de faux positifs tolérés, et en calculant la distribution des p-values d'association observées sur l'ensemble des marqueurs étudiés, sur un jeu de données où les phénotypes sont aléatoirement redistribués sur les génotypes [Churchill and Doerge, 1994]. Cette méthode est une méthode simple mais qui peut être très coûteuse en temps de calcul. Il existe aussi la méthode du FDR (False Discovery Rate [Storey and Tibshirani, 2003]) qui corrige les p-values sur la base de l'espérance de la proportion de faux positifs parmi tous les positifs détectés. En dehors du problème de seuil, la redondance d'information entre marqueurs (DL) peut être corrigée par l'utilisation de modèles d'ANOVA avec mise en cofacteur des marqueurs les plus associés. Plus généralement, l'utilisation de l'information multi-locus permet de gagner en puissance [Segura et al., 2012] et de décrire les interactions entre QTLs (épistasie) [Lin et al., 2000].

Les jeux de données de phénotypage sont souvent composés de caractères observés sur un même individu ou d'un même caractère dans des conditions différentes, ce qui génère des corrélations entre caractères. Généralement, ces caractères sont étudiés séparément, ce qui i) produit des détections redondantes des mêmes QTLs, et ii) affecte les probabilités de faux positifs, du fait de multiples tests statistiques sur des données très liées. Certaines méthodes comme l'analyse en composantes principales (ACP) sont utilisées pour créer des nouvelles variables indépendantes, résultantes de la combinaison linéaire des traits phénotypés [Stearns et al., 2005; Musani et al., 2006]. En dehors de la redondance des caractères et/ou environnements mesurés, certains QTLs peuvent avoir des effets pléiotropes. Dans le but d'étudier cette pléiotropie des QTLs ou encore les interactions QTL-environnement, des méthodes multi-traits ont été développées. La présence de pléiotropie est souvent mise en évidence avec des modèles qui analysent la covariance entre deux traits [Mangin et al., 1998; Jiang and Zeng, 1995]. Les interactions entre traits et/ou entre traits et environnement peuvent être analysées par réseau Bayésien [Scutari and Nagarajan, 2011] ou à l'aide de modèle mixte [Malosetti et al., 2008]. Ces méthodes multi-traits sont plus puissantes que les méthodes mono-trait et peuvent permettre une meilleure localisation des QTLs [Stearns et al., 2005].

Enfin, comme nous l'avons vu, les panels de lignées peuvent présenter une

structure génétique et/ou un apparentement entre génotypes, qui génèrent de fausses associations. Dans le but de limiter les faux positifs, plusieurs stratégies sont possibles. La méthode de contrôle génomique consiste à quantifier l'effet de la structure sur les tests d'association à partir d'un set de marqueurs aléatoires considérés comme non liés au caractère d'étude [Devlin and Roeder, 1999]. La méthode détermine l'écart éventuel de la distribution de la statistique de test obtenue sur ces marqueurs (supposés) indépendants du caractère à la distribution théorique en l'absence de stratification, et corrige la statistique pour cet écart. Une autre façon de quantifier cette structure peut être à l'aide d'une matrice d'apparentement (K) et/ou par une matrice de structure (Q) [Yu et al., 2006]. La matrice de structure donne le pourcentage d'appartenance de chaque individu à un nombre de groupes déterminés (Pritchard et al. [2000]; Jombart et al. [2010]. Cette matrice est intégrée en cofacteur (effet fixe) dans le modèle d'association. La matrice d'apparentement estime la similarité de paires d'individus. Pour la prendre en compte, l'effet génotype est déclaré en effet aléatoire dans le modèle. Cet effet suit alors une loi normale centrée de covariance K [Astle and Balding, 2009]. Ces matrices, souvent estimées avec le même jeu de marqueurs que ceux utilisés en détection, peuvent conduire à des sur-corrections et à l'élimination de certains QTLs [Zhao et al., 2007; Kang et al., 2008]. Pour éviter cette sur-correction due à l'utilisation des mêmes marqueurs pour l'association et dans les matrices d'apparentement et/ou de structure, Rincent et al. [2014] a proposé de réaliser les tests d'association chromosome par chromosome en éliminant le chromosome dans le calcul de la matrice d'apparentement. La matrice d'apparentement peut aussi être calculée de manière plus fine par la pondération des marqueurs, avec un poids plus faible pour les marqueurs fortement liés à d'autres marqueurs, afin d'équilibrer les contributions de l'ensemble des régions génomiques.

Dans le chapitre précédent, le modèle le plus simple de génétique d'association (ANOVA) nous a permis de détecter quelques QTLs, parmi les marqueurs identifiés par notre test de sélection. Dans ce chapitre, nous avons appliqué une plus large gamme de méthodes de génétique d'association, sur l'ensemble des caractères phénotypés, pour exploiter l'intérêt de la population MAGIC INRA. Nous avons donc dans un premier temps utilisé un modèle mixte monolocus avec prise en compte de la matrice d'apparentement, puis un modèle multi-locus. Une détection de QTLs a ensuite été réalisée sur 12 caractères liés à la précocité de floraison ainsi que sur des variables intégratives de ces caractères pour tester l'approche multi-traits. Nous avons ensuite discuté de l'intérêt de la population MAGIC INRA pour la détection de QTLs, en termes de puissance de détection et de précision de localisation.

5.2.2 Matériels et Méthodes

5.2.2.1 Matériel

Nous avons utilisé 380 lignées recombinantes issues d'une population allogamisée de blé de type MAGIC brassée pendant 12 générations de panmixie. La population

fondatrice a été créée par croisement de 59 parents choisis pour leur large base génétique avec la variété mutante mâle stérile "Probus". 380 lignées (S4) ont été fixées par Single Seed Descent (SSD) (Chap 2).

5.2.2.2 Données génétiques

Les données génétiques utilisées sont les mêmes que celles étudiées dans chapitre 4. Les 380 lignées ont été génotypées avec la puce 9K iSelect ainsi que 14 polymorphismes localisés dans des gènes candidats pour l'adaptation (Figure 2.3) représentant au total 6 301 marqueurs.

5.2.2.3 Données phénotypiques

Les 380 lignées ont été phénotypées au moyen d'essais au champ implantés suivant trois dates de semis (novembre, mars et avril) sur deux années consécutives (2010-2011, 2011-2012). En 2011-2012, une modalité supplémentaire a été réalisée avec une condition de jour long forcé (semis de novembre). Dans chacune des conditions, une cinétique a été établie par le suivi de cinq stades de développement (Chap 2). Chaque année pour les semis de novembre et de mars, deux répétitions ont été réalisées et une seule pour la condition en jour long et les semis d'avril (Chap 2). Tous les caractères phénotypés avaient une héritabilité supérieure à 0,8 (Chap 5.1).

Vu le nombre de caractères et de conditions dans lesquelles les lignées ont été phénotypées et sachant que certains couples caractères × conditions sont très corrélés, nous avons décidé de réduire l'étude à deux caractères clés : i) l'épiaison, qui représente la précocité de floraison (nommée "epi") et possède la plus forte héritabilité, et ii) la différence entre le stade 7 et l'épiaison, qui représente la durée de remplissage du grain (nommée "remplissage"). Des moyennes ajustées sur les répétitions et sur les années ont été calculées pour les semis de novembre. Comme les semis de printemps sont faiblement corrélés entre eux d'une année sur l'autre (Chap 5.1), les moyennes ont été ajustées sur les répétitions, en analysant indépendamment chaque année. La condition de semis en jour long a été utilisée en comparaison avec le semis au champ de la même année pour estimer la sensibilité à la photopériode (par différence, condition nommée "photo").

Dans le but de prendre en compte l'ensemble des informations récoltées, une analyse en composantes principales (ACP) a été réalisée sur l'ensemble des données phénotypiques liées à la précocité des campagnes 2010-2011 et 2011-2012 (cinq caractères × sept conditions ; Chap 5.1). Les coordonnées des génotypes sur les quatre premiers axes de l'ACP représentant 82,85% de la variabilité phénotypique ont été utilisées. Les trois premiers axes représentent respectivement le besoin en vernalisation (axe 1), la précocité intrinsèque (axe 2) et la sensibilité à la photopériode (axe 3), le quatrième axe étant plus complexe, mais aussi représentatif du besoin en vernalisation (axe 4) (Chap 5.1). Une deuxième stratégie de synthèse des informations phénotypiques a été d'utiliser un modèle écophysiologique de prédiction de la date de floraison. Ce modèle formalisé comme le modèle

ARCWHEAT [Weir et al., 1984], a permis d'estimer deux paramètres : un paramètre de besoin en vernalisation ($Vsat$) et un de sensibilité à la photopériode ($Pbase$). L'estimation de ces paramètres a été réalisée à partir des données de floraison sur les trois dates de semis de l'année 2010-2011 ainsi que sur la modalité en jour long (semis de novembre) de 2011-2012. Seuls 238 des 380 individus analysés ont été retenus pour l'analyse d'association, ceux pour lesquels l'optimisation des paramètres prédisait les dates de floraisons des quatre conditions avec une erreur (RMSE) inférieure à 10 jours.

Au total 16 traits répertoriés dans le tableau 5.4 ont été étudiés.

Tableau 5.4 – Tableau récapitulatif des 16 traits utilisés dans les analyses de génétique d'association

Descriptifs	Caractère	Nom de variable
Moyenne pondérée sur les semis de novembre (2 années × 2 répétitions)	Epiaison	nov_epi
	Remplissage	nov_remplissage
Moyenne pondérée sur les semis de mars 2011 (2 répétitions)	Epiaison	mars2011_epi
	Remplissage	mars2011_remplissage
Moyenne pondérée sur les semis de mars 2012 (2 répétitions)	Epiaison	mars2012_epi
	Remplissage	mars2012_remplissage
Observations sur le semis d'avril 2011	Epiaison	avril2011_epi
	Remplissage	avril2011_remplissage
Observations sur le semis d'avril 2012	Epiaison	avril2012_epi
	Remplissage	avril2012_remplissage
Coordonnées des individus de l'ACP réalisée avec l'ensemble des données phénotypiques (5 caractères × 7 conditions)	Axe 1	acp_ax1
	Axe 2	acp_ax2
	Axe 3	acp_ax3
	Axe 4	acp_ax4
Paramètres estimés par le modèle écophysiologique utilisant la date de floraison mesurée sur les modalités de 2010-2011 et celle de semis de novembre 2011 en jour long	Paramètre de besoin en vernalisation	Vsat
	Paramètre de sensibilité à la photopériode	Pbase

5.2.2.4 Inférence des données manquantes

Une première étape dans l'analyse a consisté à inférer les 0,7% de données manquantes présentes dans la matrice de génotypage. En effet, certains logiciels ne peuvent pas être utilisés avec un jeu de données incomplet, et de plus différents auteurs ont montré que cette inférence pouvait légèrement améliorer la qualité des analyses génétiques [Hao et al., 2009; Marchini and Howie, 2010]. Les inférences ont été effectuées à l'aide du logiciel Beagle (v.3.3.2. [Browning and Browning, 2007]), qui ne nécessite ni données généalogiques, ni carte de localisation des marqueurs, et toutes ces inférences ont été acceptées quelle que soit leur probabilité. Pour les analyses d'association, le jeu de 6 301 marqueurs disponible après inférence a été réduit à 4 924 marqueurs "uniques", *i.e.* présentant une information spécifique (chaque marqueur a un DL différent de 1 avec les 4 923 autres marqueurs).

5.2.2.5 Estimation de la matrice d'apparentement

La matrice d'apparentement a été calculée sur la base des 4 924 SNPs, avec l'équation 5.11 [Astle and Balding, 2009] à l'aide du logiciel R [R Development Core Team, 2012].

$$K_{ij} = \frac{1}{p(1-p)} \sum_{l=1}^{L} (G_{il} - p_l)(G_{jl} - p_l) \tag{5.11}$$

avec K_{ij} l'apparentement entre les individus i et j, p la fréquence moyenne dans la population de l'allèle 1, p_l la fréquence de l'allèle 1 du marqueur l et G_{il}, G_{jl} les génotypes de l'individu i et j au marqueur l. Pour mieux comprendre les niveaux de similarité entre les individus, un indice d'identité par état a aussi été calculé correspondant au pourcentage d'allèles communs entre deux individus.

5.2.2.6 Analyse des différents modèles

Du fait de la très faible structure génétique décrite dans la population MAGIC INRA (Chap 4), seule la matrice d'apparentement a été prise en compte. Notons que cette matrice d'apparentement intègre généralement l'effet de la structure génétique de la population [Yu et al., 2006]. L'étude a été réalisée avec seulement deux caractères × conditions très contrastés : nov_epi et avril2011_epi.

Modèle avec ou sans prise en compte de l'apparentement La nécessité dans notre étude de corriger pour l'apparentement a été étudiée par comparaison des résultats d'un modèle linéaire simple (Equation 5.12) avec ceux d'un modèle linéaire mixte avec prise en compte de la matrice d'apparentement (Equation 5.13). Ces modèles ont été testés à l'aide du logiciel R [R Development Core Team, 2012] avec respectivement les fonctions "lm" et "anova" pour le premier et la librairie R "mlmm" pour le second (sans cofacteur) [Segura et al., 2012]. Les p-values issues des deux modèles ont été comparées graphiquement. La q-value des marqueurs a aussi été estimée avec la fonction "p.adjust" dans la librairie R "stats". Les marqueurs avec une q-value inférieure à 10% ont été considérés comme significatifs.

1) un modèle linéaire simple :

$$Y = \mathbb{1}\mu + \beta_m X_m + E \tag{5.12}$$

$$E \sim N(0, I\sigma_e^2)\ iid$$

avec Y le vecteur des phénotypes, $\mathbb{1}\mu$ un vecteur moyenne composé d'éléments égaux à la moyenne générale, X_m le vecteur des génotypes au marqueur m, β_m l'effet additif à estimer du marqueur m et E le vecteur des résidus qui suit une loi normale centrée de variance $I\sigma_e^2$, avec I la matrice identité, tous les vecteurs étant de dimension N (nombre d'individus).

2) un modèle linéaire mixte avec prise en compte de la matrice d'apparentement :

$$Y = \mathbb{1}\mu + \beta_m X_m + U + E \tag{5.13}$$
$$E \sim N(0, I\sigma_e^2)\ iid$$
$$U \sim N(0, K\sigma_{gM}^2)\ iid$$

avec Y le vecteur des phénotypes, $\mathbb{1}\mu$ le vecteur moyenne, X_m le vecteur des génotypes du marqueur m, β_m l'effet additif à estimer du marqueur m, U le vecteur des effets aléatoires génétiques liés à l'apparentement K, et qui suit une loi normale centrée de variance $K\sigma_{gM}^2$ et E le vecteur des résidus qui suit une loi normale centrée de variance $I\sigma_e^2$, avec I la matrice identité, tous les vecteurs étant de dimension N (nombre d'individus).

Modèle d'association multi-locus Afin d'augmenter le pouvoir de détection de marqueurs à effet faible, nous avons utilisé le modèle mixte multi-locus "mlmm" avec prise en compte de l'apparentement [Segura et al., 2012]. Cette méthode de type "stepwise" intègre successivement en cofacteur les marqueurs, par niveaux d'effet décroissants (Equation 5.14) et ayant une p-value supérieure au seuil de 10% corrigé par la méthode de Bonferroni et calcule un indice EBIC (Extended Bayesian Information Criterion [Chen and Chen, 2008]) pour chaque modèle successif. L'intégration des marqueurs en cofacteur s'arrête quand l'héritabilité (Equation 5.2) est proche de 0. Une fois le processus d'intégration des marqueurs en cofacteur fini, un processus de "backward" avec élimination successive des marqueurs en cofacteur les moins significatifs est réalisé. Le modèle retenu correspond au modèle qui minimise l'indice EBIC.

$$Y = \mathbb{1}\mu + \sum_{c=1}^{C} \beta_c X_c + \beta_m X_m + U + E \tag{5.14}$$
$$E \sim N(0, I\sigma_e^2)\ iid$$
$$U \sim N(0, K\sigma_{gM}^2)\ iid$$

avec Y le vecteur des phénotypes, 1μ le vecteur moyenne, C le nombre de cofacteurs incorporés, X_c le vecteur des génotypes au marqueur c mis en cofacteur, β_c l'effet additif du marqueur c mis en cofacteur, X_m le vecteur des génotypes au marqueur m, β_m l'effet additif à estimer du marqueur m, U est le vecteur des effets aléatoires génétiques qui suit une loi normale centrée de variance $K\sigma_{gM}^2$ et E est le vecteur des résidus qui suit une loi normale centrée de variance $I\sigma_e^2$, avec I la matrice identité de dimension N. La librairie R "mlmm" a été utilisée pour réaliser tous les modèles mixtes avec cofacteurs [Segura et al., 2012].

5.2.2.7 Représentation en Manhattan plot

Pour l'établissement des Manhattan plots, nous avons utilisé la carte génétique de la puce 9K iSelect publiée dans Cavanagh et al. [2013]. Sur cette carte, 4 502

des 4 924 marqueurs étaient cartographiés. Les 422 marqueurs non localisés dont les 14 polymorphismes localisés dans les gènes candidats ont été cartographiés approximativement sur la base de leur DL avec les marqueurs cartographiés. Pour chaque marqueur à cartographier, le DL a été calculé à l'aide d'une corrélation de Pearson avec le logiciel R [R Development Core Team, 2012] entre ce marqueur et tous les SNPs cartographiés. Si les SNPs les plus liés au marqueur étaient localisés de façon cohérente sur une même zone chromosomique, alors le marqueur a été cartographié à la même position que le SNP en plus fort DL.

5.2.2.8 Estimation de la couverture du génome

Nous avons voulu estimer le niveau de couverture du génome correspondant à la densité de marquage 9K utilisée sur la population MAGIC INRA. Les marqueurs étant répartis de manière hétérogène le long des chromosomes, le but était de calculer la proportion du génome pour laquelle il n'existe pas de marqueurs génotypés à une distance inférieure à un seuil donné. Cette distance non couverte a été calculée par intégration sur l'ensemble du génome des différences entre la distance entre deux marqueurs adjacents et la distance seuil. La gamme explorée pour cette distance seuil était comprise entre 1cM et 100cM, avec un pas de 0,5. La distance non couverte a ensuite été ramenée à la distance totale de la carte (3 567cM) pour estimer un pourcentage.

Une fois ce pourcentage de génome couvert estimé en fonction d'une distance génétique, cette distance génétique a été transformée en DL à l'aide d'une relation estimée par optimisation statistique entre le DL et la distance génétique dans la population MAGIC INRA (Equation 5.15).

$$r^2 = \frac{0,6}{\frac{d}{2}} - 0,005 \tag{5.15}$$

avec r^2 le DL et d la distance génétique.

5.2.2.9 Estimation de la puissance de détection de la population

La puissance de détection des QTLs dans le dispositif MAGIC INRA a été étudiée au moyen d'un programme R communiqué par R. Rincent. Cette puissance π représente la capacité à rejeter l'hypothèse nulle quand celle-ci est fausse ou en d'autres termes à rejeter un marqueur alors qu'il est effectivement associé au trait d'étude. Pour réaliser cette étude de puissance, nous avons supposé le cas de la ségrégation d'un marqueur M, en DL r^2 avec un QTL ayant un effet β sur le phénotype, et dans une population de variance génétique pour le trait de σ_p^2. Nous avons pris le cas de la variable nov_epi, qui présente une moyenne $\mu = 1\,500,71dj$, et une variance $\sigma_p^2 = 4\,208,11$. Sous ces hypothèses, le test d'association du marqueur au QTL est une statistique de Student décentrée à $N-2$ degrés de liberté (avec N le nombre d'individus) et un coefficient de décentrage λ, avec λ qui dépend de l'effet

β, du DL r^2, de la fréquence allélique p du QTL et de la variance phénotypique σ_p^2 [Guedj, 2007] (Equation 5.16).

$$\lambda = \frac{\beta \times \sqrt{r^2}}{\sqrt{\sigma_p^2(\frac{1}{p \times N} + \frac{1}{(1-p) \times N})}} \tag{5.16}$$

La puissance de détection a été calculée pour une gamme d'effet du QTL, β, compris entre 1 et 200dj avec un pas de 3dj, une variance phénotypique constante correspondant à celle de nov_epi et un seuil de détection constant correspondant au seuil de 10% corrigé ensuite par la méthode de Bonferroni ($\frac{0,1}{nbtests}$) et utilisé dans l'analyse d'association. La puissance de détection a ensuite été calculée en fonction de l'effet du QTL (transformé en pourcentage d'explication de variance phénotypique (R^2 ; Equation 5.17)), avec une p constante fixée à 0,28 correspondant à la MAF moyenne dans la population MAGIC INRA. Le calcul de puissance a été réalisé dans quatre cas répertoriés dans le tableau 5.5, correspondant à deux tailles de population d'étude, et deux densités de marquage. L'effet de la MAF sur la puissance a été analysé sur la base des paramètres de l'étude actuelle (N=380, r^2=0,5) pour une gamme de 0,5 , 0,28 , 0,1 , 0,05 et 0,01 .

$$R^2 = \frac{\beta^2 \times p \times (1-p)}{\sigma_p^2} \tag{5.17}$$

Tableau 5.5 – Tableau des paramètres des quatre cas pour lesquels la puissance a été estimée.

	Nb ind	MAF	DL (r^2)
Paramètres de l'étude MAGIC-INRA	380	0,28	0,5
augmentation du nombre d'individus	1 026	0,28	0,5
augmentation de la densité de marquage	380	0,28	1
augmentation de la densité de marquage et du nombre d'individus	1 026	0,28	1

Afin de mieux approcher la puissance du dispositif MAGIC INRA, nous avons modélisé de façon sommaire la ségrégation de plusieurs QTLs dans la population. Pour cela, nous avons tout d'abord fixé le nombre de QTLs en ségrégation à 100, de façon totalement arbitraire, mais somme toute réaliste face au peu d'informations disponibles sur cette question : Buckler et al. [2009] a détecté avec le dispositif NAM sur maïs une centaine de QTLs alors que Le Gouis et al. [2011] en a détectés 45 sur un panel de blé. Nous nous sommes ensuite appuyés sur la distribution, de type exponentielle, des effets absolus des QTLs détectés dans un dispositif NAM-maïs pour la date de floraison [Buckler et al., 2009]. Nous avons ensuite supposé que les 100 QTLS en ségrégation étaient bialléliques, et avaient une distribution de fréquences alléliques identique à celle des SNPs 9K iSelect (même distribution que

les MAF ; Figure 3 Chap 4.3). A partir de ces hypothèses, nous avons ajusté les valeurs des effets, de façon à obtenir une variance génétique pour le trait d'intérêt identique à celle observée dans la MAGIC INRA. Pour cela, nous avons sommé sur les 100 QTLs les numérateurs de l'équation 5.17, sous l'hypothèse de stricte additivité des QTLs. Nous avons de plus fixé un des QTLs avec un effet majeur (58% de variance) pour mieux tenir compte des données réelles.

Une population F2 de 1 026 individus a été simulée, en fixant la fréquence allélique de tous les QTLs à p=0,5. La puissance de détection des 100 QTLs simulés qui suivent la même distribution d'effet que décrit précédemment a été calculée avec un DL marqueur QTL de 1. On a considéré que les QTLs avec une puissance de détection dans le dispositif F2 inférieure à 1% sont des QTLs indétectables. Sur les 100 QTLs simulés, 50 sont détectables et en ségrégation dans une population F2 avec 1 026 individus. Seuls ces 50 QTLs détectables ont été considérés pour comparer différents dispositifs MAGIC.

Dans la population MAGIC, ces 50 QTLs ont une fréquence allélique distribuée selon la distribution de MAF décrite précédemment. La puissance de détection de chacun de ces QTLs a été calculée en fonction de leur MAF, leur effet, mais aussi d'un nombre d'individus et d'un DL QTL marqueur, comme décrit dans le tableau 5.5.

5.2.2.10 Etude des qualités prédictives

Pour chacun des traits étudiés, nous avons évalué le pouvoir de prédiction des marqueurs détectés par génétique d'association, en effectuant des tests de validation croisée. Pour ce faire, nous avons ajusté nos modèles sur un sous-échantillon aléatoire de 90% des individus, et mesuré leur qualités prédictives sur les 10% restants. Deux modèles statistiques ont été comparés : un modèle additif et un modèle prenant en compte les interactions possibles des marqueurs deux à deux. La qualité de prédiction a été quantifiée par la moyenne et l'écart-type des corrélations de Pearson entre les valeurs prédites par le modèle et les valeurs observées, sur la base de 20 validations croisées (20 tirages aléatoires des individus : 90% +10%). Dans un but de comparaison des résultats, nous avons testé un modèle de sélection génomique ajusté avec la méthode Lasso ([Tibshirani, 1996] implémentée dans le package R "glmnet" [Friedman et al., 2008]) avec tous les marqueurs génotypés (après élimination des marqueurs redondants (DL : $r^2 = 1$)). 20 étapes de validation croisée ont également été réalisées pour les prédictions de sélection génomique.

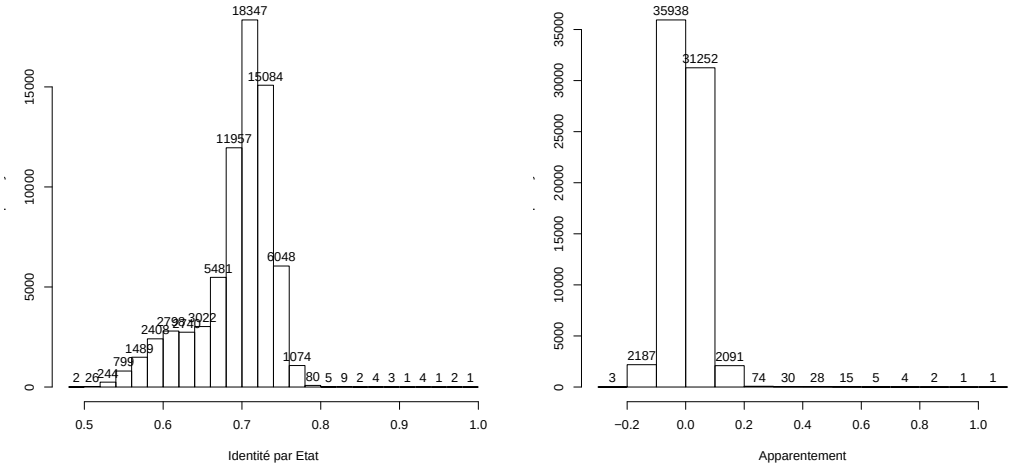
5.2.3 Résultats

5.2.3.1 Modèle avec prise en compte de l'apparentement

Les distributions d'indices d'identité par état (IBS) et d'apparentement montrent une similarité entre les individus assez homogène avec une moyenne d'IBS à 0,69. Ce chiffre est comparable à l'IBS attendu pour deux individus non apparentés dans la population (0,73). 0,04% de ces indices sont supérieurs à 0,5 en apparentement

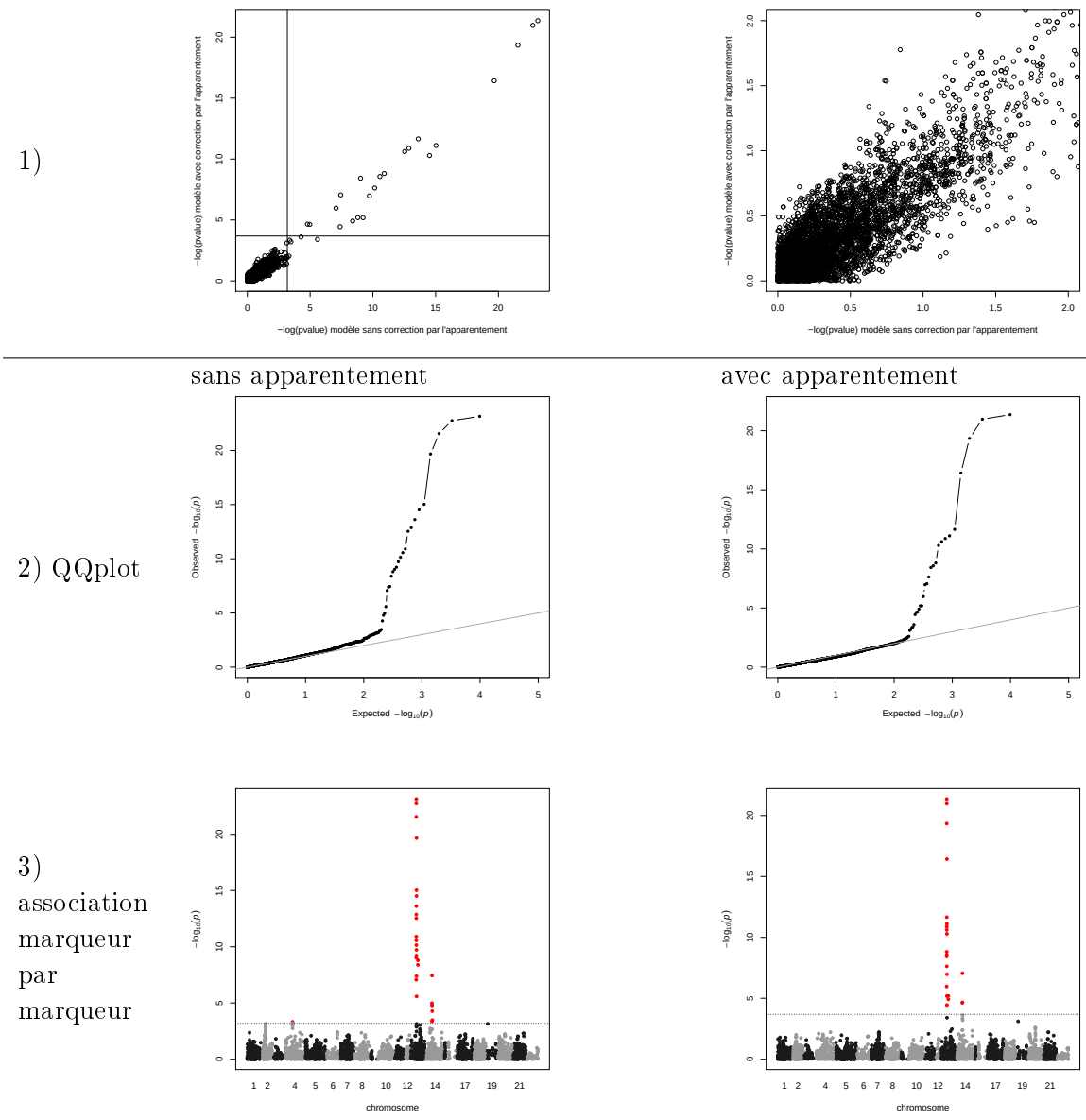
ou à 0,8 en IBS, concernant 38 individus (Figure 5.15). 9 couples d'individus ont un IBS supérieur à 0,9 dont 4 couples avec un IBS supérieur à 0,95.

FIGURE 5.15 – Distribution des indices d'identité par état (à gauche) et d'apparentement (à droite) entre paires de lignées SSD. Les chiffres indiquent le nombre d'éléments de chaque classe.



Les p-values estimées avec le modèle linéaire simple et le modèle linéaire mixte avec prise en compte de la matrice d'apparentement sont corrélées à 0,99 pour le caractère nov_epi (Figure 5.23) et de 0,96 pour le caractère avril2011_epi (Figure 5.16). Les QQplots montrent un décrochement légèrement moins marqué de la relation entre p-values observées et attendues sous l'hypothèse nulle, dans le cas du modèle avec matrice d'apparentement. La prise en compte de l'apparentement a une conséquence nulle ou positive pour les traits nov_epi et avril2011_epi. Ainsi la correction réduit de quatre QTLs à deux QTLs détectés pour nov_epi et de trois QTLs à deux pour avril2011_epi. La prise en compte de l'apparentement a donc des effets variables suivant les caractères, mais toujours positifs, nous avons donc préféré présenter les analyses avec cette correction.

FIGURE 5.16 – Comparaison des résultats issus du modèle linéaire avec et sans prise en compte de l'apparentement pour le trait avril2011_epi : 1) Comparaison des -log(p-values) (à gauche) avec zoom sur les valeurs comprises entre 0 et 2 (à droite). Les lignes représentent les seuils de significativité trouvés par la méthode de FDR (10%), 2) QQplot avec prise en compte de l'apparentement (à droite) et sans (à gauche), 3) Manhattan plot des résultats de détection d'association.

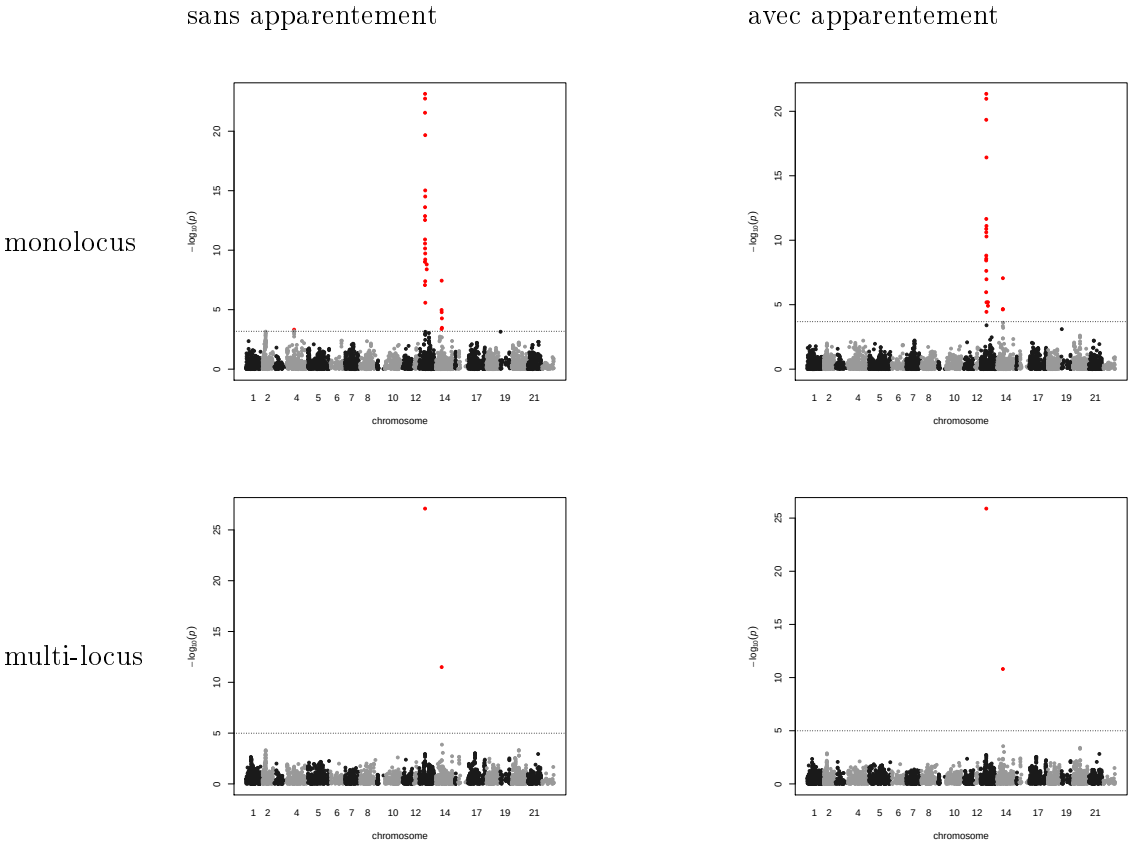


5.2.3.2 Impact d'un modèle avec cofacteur

La mise en cofacteur d'un marqueur associé diminue très fortement les p-values des marqueurs génétiques en fort DL avec ce cofacteur, ce qui i) élimine les

associations des marqueurs redondants et ii) augmente la puissance de détection de QTLs à effet faible. L'effet de correction du DL est évident sur le chromosome 5A (Figure 5.17) , pour lequel à la fois les effets des SNPs cartographiés sur la même zone sont corrigés, mais aussi des "faux positifs" situés à plus longue distance (liés à du DL longue distance, ou des incohérences de cartographie). A partir du critère EBIC, cette méthode a ainsi mis en évidence deux QTLs pour chacun des caractères nov_epi et avril2011_epi, localisés sur les chromosomes 2D et 7D, et 5A et 5B respectivement(Figure 5.17 & 5.24). Ces quatre QTLs sont associés aux marqueurs ayant les effets maximums lors de l'analyse mono-locus avec prise en compte de l'apparentement. Par la suite, nous avons présenté les résultats des analyses avec la méthode multi-locus, avec prise en compte de l'apparentement.

FIGURE 5.17 – Comparaison des résultats d'association avec (à droite) et sans (à gauche) prise en compte de l'apparentement pour le caractère avril2011_epi avec la méthode mono-locus et la méthode multi-locus.



Les quatre QTLs précédemment détectés sont clairement visualisés sur les Manhattan plots des p-values d'association des deux traits étudiés (nov_epi et avril2011_epi). Les régions génomiques identifiées ne montrent pas le même motif : sur le caractère nov_epi, un et trois marqueurs identifient les deux QTLS, alors que 6 et 20 marqueurs identifient les deux QTLS sur le caractère avril2011_epi.

Ceci reflète les hétérogénéités en densité de marqueurs de la carte (Figure 5.18) : certaines zones sont très riches en marqueurs très liés et d'autres sont vides ou très peu représentées (Figure 5.18), avec par exemple les chromosomes du génome D très peu couverts en comparaison aux deux autres génomes.

5.2.3.3 Impact des densités de marquage et de la taille de la population sur la puissance de détection

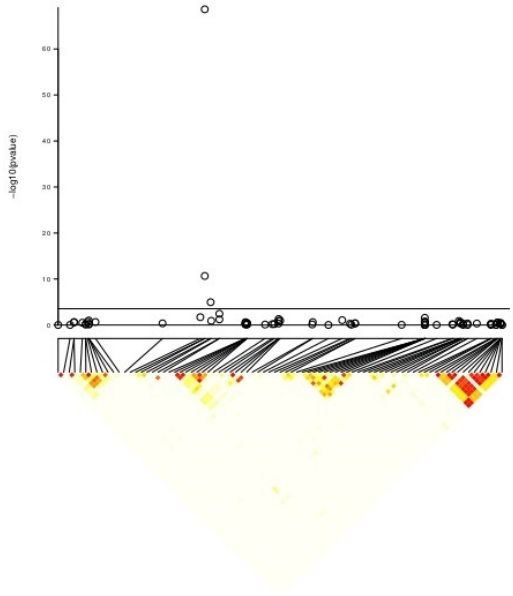
Avec le marquage 9K, 90% du génome est couvert avec un marqueur tous les 10cM ou moins, soit un marqueur à une distance maximale de 5cM d'un potentiel QTL (Figure 5.19).

Le calcul de puissance de détection a montré qu'avec la densité de marquage 9K, et sur la base d'une fréquence au QTL égale à la MAF moyenne p = 0,28, on a une puissance de détection supérieure à 50% pour les QTLs avec un effet d'au moins 9%. Une augmentation de la densité de marquage jusqu'à une densité de marquage qui entraîne un DL de 1 entre un QTL et un marqueur, permettrait de détecter avec une puissance de 10% les QTLs avec un effet minimum de 5% (Figure 5.21). L'augmentation du nombre d'individus de 380 à 1 026 permettrait à puissance égale de détecter des QTLs à partir d'effets de 1,5%, alors que l'augmentation du nombre d'individus et du nombre de marqueurs abaisserait l'effet à 0,8%. L'augmentation du nombre d'individus améliore donc plus la puissance que l'augmentation du nombre de marqueurs. La fréquence allélique du QTL joue également un rôle important sur la puissance en fonction de l'effet du QTL. La MAF moyenne dans la population est de 0,28, et le gain de puissance produit par une fréquence optimale p=0,5, avec les deux allèles au QTL équifréquents, n'est pas très important : pour une puissance de 50%, un QTL devient détectable à partir d'un effet de 43dj au lieu 39dj. Cette puissance chute par contre rapidement pour une MAF inférieure à 10%. Ainsi pour détecter avec une puissance de 50L'analyse théorique de puissance de détection de QTLs dans une population MAGIC a été effectuée en modélisant 100 QTLs en ségrégation, 50 ayant un effet détectable, les 50 autres contribuant à la variation du fond génétique (cf MM). Avec la densité de marquage et la taille de population actuelles, seulement 9 QTLs sur 50 (18%) sont détectables dans la population MAGIC INRA. Le même calcul de puissance, réalisé sur une population F2 avec 380 individus permettrait de détecter 60% des QTLs. Dans la population MAGIC-INRA, l'augmentation du nombre de marqueurs ou du nombre d'individus permettrait la détection de 36% et 46% des QTLs, respectivement. L'augmentation simultanée des deux paramètres amènerait ce chiffre à 70%.

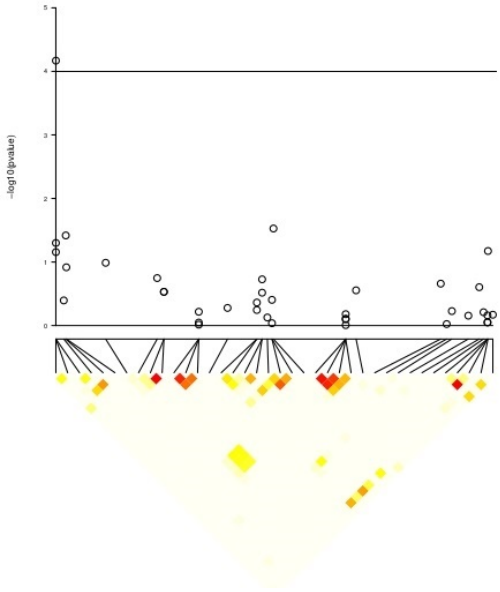
FIGURE 5.18 – Manhattan plots des p-values d'association marqueur par marqueur au voisinage des quatre QTLs détectés avec la méthode multi-locus pour les caractères nov_epi (en haut) et avril2011_epi (en bas). La matrice de DL entre marqueurs est représentée sous l'axe. La droite horizontale représente le seuil de significativité de 10% corrigé avec la méthode de FDR.

Caractère associé : nov_epi

Chromosome 2D (169cM)

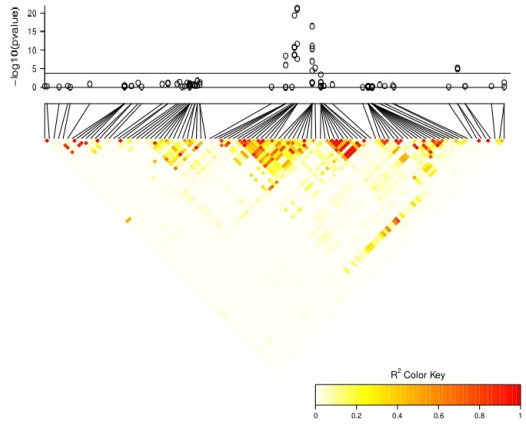


Chromosome 7D (143cM)



Caractère associé : avril2011_epi

région autour du QTL sur le 5A (60cM)



région autour du QTL sur le 5B (100cM)

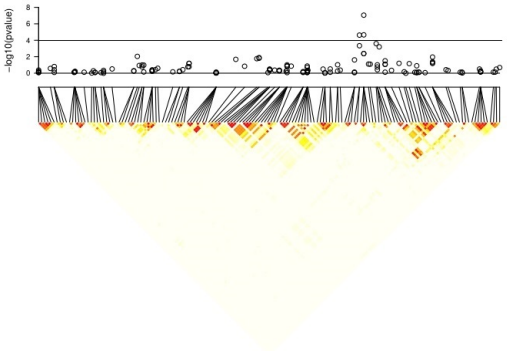


FIGURE 5.19 – Représentation du pourcentage de génome couvert dans la population MAGIC INRA par les 4 924 marqueurs en fonction de la distance (en cM) (en haut) ou du DL (en r^2) (en bas) limite marqueur QTL .

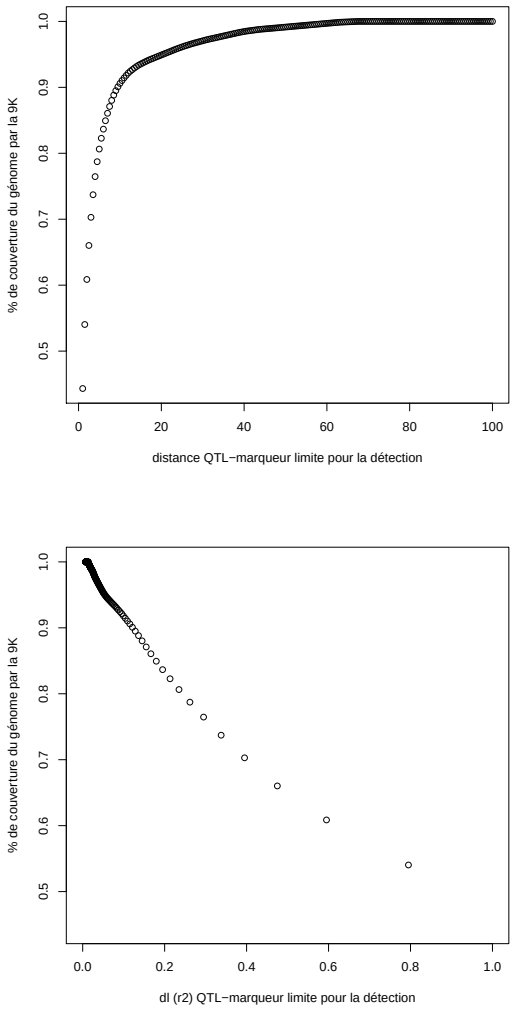


FIGURE 5.20 – Représentation de la puissance de détection des QTLs en fonction de leur effet pour une MAF=0,28 et différents nombres d'individus (380 ou 1 026) et différents r^2 (0,5 et 1)

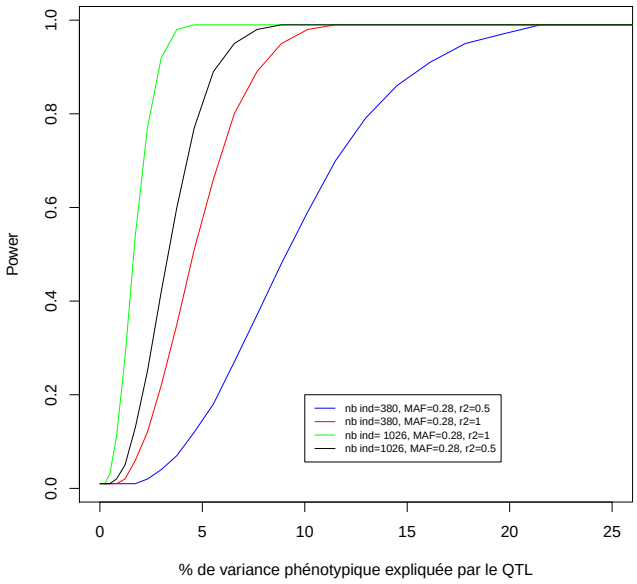
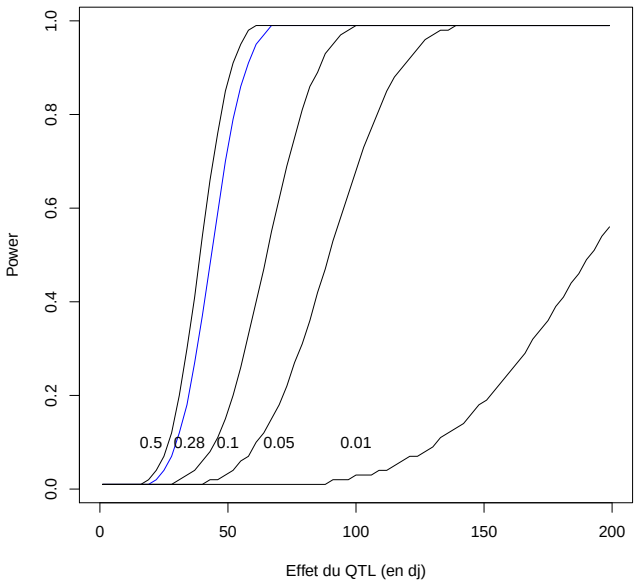


FIGURE 5.21 – Représentation de l'effet de la MAF (0,5 ; 0,28 ; 0,1 ; 0,05 ; 0,01) sur la puissance de détection en fonction de l'effet du QTL avec le r^2 fixé à 0.5 et le nombre d'individus à 380.



5.2.3.4 Résultats d'association

Caractère simple Pour chacun des 12 traits étudiés, une analyse avec le modèle multi-locus "mlmm" avec prise en compte de l'apparentement a été réalisée, avec les 4 924 marqueurs. Au total, cinq QTLs distincts ont été détectés (Tableau 5.6), de nombreux caractères × conditions dépendant des mêmes QTLs. Ainsi du fait de leur corrélation, le remplissage est expliqué par les mêmes QTLs que l'épiaison. De même le fort contraste entre les semis d'automne et les semis de printemps conduit à identifier deux ensembles de QTLs. Deux QTLs expliquent chaque trait, excepté pour nov_remplissage et mars2011_remplissage où un seul QTL a été détecté et aucun pour photo_remplissage. Ces cinq QTLs (identifiés par les SNPs *Ppd-D1prom.2kb.indel., VRN1A.ex8.1bp.indel., wsnp_RFL_Contig2809_2587619, wsnp_Ku_9541_15976096, VRN3D-FTD.G.deletion*) sont localisés sur les chromosomes 2D, 5A, 5B (2 QTLs) et 7D (Figure 5.25 & Figure 5.26). Trois d'entre eux sont associés à des marqueurs localisés dans des gènes candidats : *Ppd-D1, Vrn3D/FTD, Vrn1A*. Ces QTLs expliquent entre 4,5% et 56,3% de la variance génétique suivant les caractères. Les marqueurs localisés dans les gènes Ppd-D1 et Vrn3D ont été détectés en association avec les traits mesurés avec un semis d'automne et de sensibilité à la photopériode, *Vrn3D* a été associé avec un seul caractère nov_epi ; alors que le marqueur localisé dans *Vrn1A* est associé à la précocité mesurée sur les semis de printemps. Les deux autres QTLs sont localisés sur le chromosome 5B, l'un des QTLs étant essentiellement associé aux traits mesurés en semis de printemps, et l'autre, à photo_epi. Dans l'analyse de Le Gouis et al. [2011], il semble en effet y avoir deux QTLs sur le chromosome 5B contrôlant la précocité de floraison mais tous deux expliquant les besoins en vernalisation. La zone génomique la plus large associée à un QTL est celle de *wsnp_RFL_Contig2809_2587619* avec 5,6cM. La localisation des deux marqueurs en très fort DL avec *Vrn1A* et distants de 22,6cM a été attribuée à une erreur de cartographie.

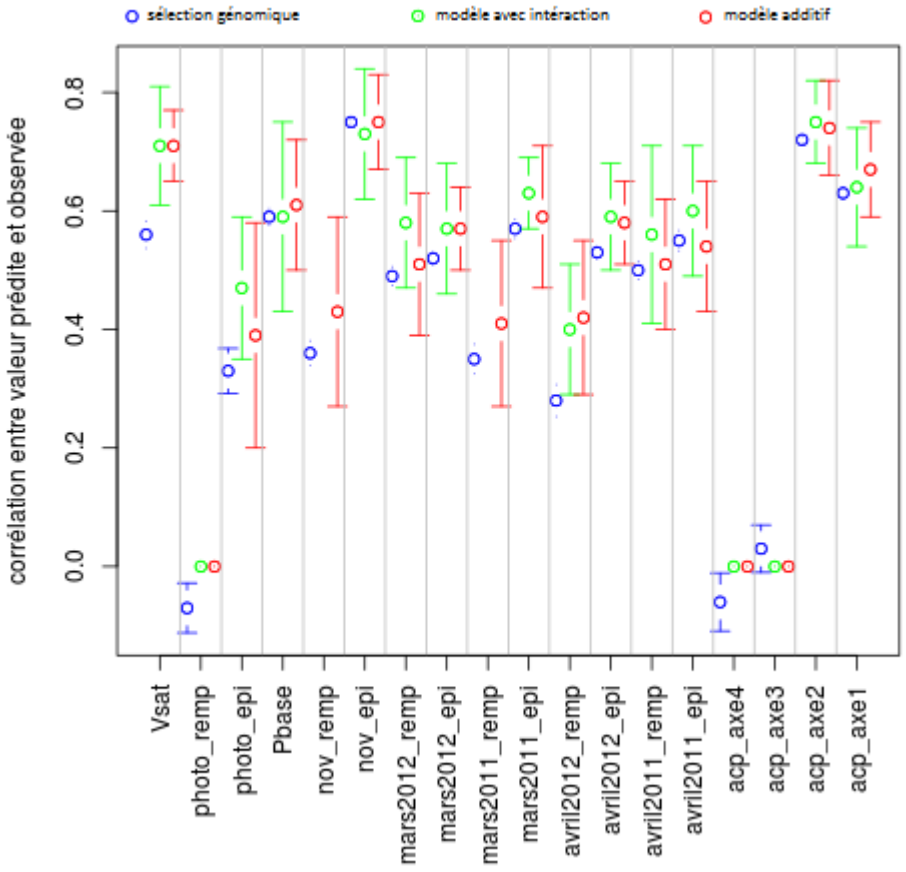
Caractères intégratifs Dans le but d'intégrer les informations complémentaires présentes dans les données de phénotypage, nous avons aussi réalisé des tests d'association sur les coordonnées des individus sur les quatre premiers axes d'une ACP réalisée sur l'ensemble des caractères, ainsi que sur les deux paramètres estimés par le modèle écophysiologique. Au total, respectivement cinq et quatre QTLs ont été détectés (Tableau 5.6) avec la méthode multi-locus, en utilisant un seuil de 10% avec correction de Bonferroni (Figure 5.25 & Figure 5.26). Quatre des cinq QTLs détectés sur l'ACP correspondent à ceux identifiés par les tests trait par trait ou par le modèle écophysiologique, excepté *wsnp_Ku_9541_15976096*. Les coordonnées de l'ACP ont permis de détecter un QTL supplémentaire (effet faible de 1,7%).

5.2.3.5 Pouvoir de prédiction

Le pouvoir de prédiction des marqueurs détectés par association a été évalué, en considérant un modèle additif, et un modèle avec interaction entre les marqueurs

deux à deux. Avec en moyenne une corrélation de Pearson (prédit/observé) de 0,56 (min=0,39 ; max=0,75) pour le modèle additif et de 0,60 (min=0,40 ; max=0,75) pour le modèle avec interaction, les deux modèles ont un pouvoir de prédiction similaire. En comparaison avec ces modèles, la sélection génomique a une précision de prédiction plus faible avec une moyenne de 0,42 (min=0, max=0,75) malgré un plus faible écart type (0,02 *vs* 0,11).

FIGURE 5.22 – Moyenne et écart-type des corrélations entre les valeurs prédites et observées sur l'ensemble des caractères par différents modèles : sélection génomique avec régression sur tous les marqueurs de la puce (LASSO) (bleu), un modèle avec interaction avec les marqueurs détectés par association (vert), un modèle additif avec les marqueurs détectés par association (rouge). nov_remplissage et mars_remplissage n'ont pas de point correspondant à la prédiction avec le modèle avec interaction car un seul QTL est associé avec chacun.



5.2.4 Discussion

Dans cette étude, nous avons mis en oeuvre les principales méthodes de détection de QTLs par génétique d'association pour identifier la méthode la plus adaptée à notre population. L'analyse d'association a été réalisée sur deux caractères (précocité et remplissage du grain) observés dans six conditions de culture différentes, en utilisant un modèle mixte d'analyse de variance, multi-locus, qui prend en compte la matrice d'apparentement. L'analyse trait par trait a permis de détecter cinq QTLs localisés sur les chromosomes 2D, 5A, 5B (2 QTLs) et 7D. Cette analyse a été confrontée à deux analyses d'association réalisées avec des traits synthétisant l'information des 12 traits testés indépendamment : les coordonnées des individus sur les quatre premiers axes d'ACP et les paramètres de modèle écophysiologique (Chap 5.1). Ces analyses nous ont permis de détecter quatre QTLs dans les deux cas mais avec un QTL différent. Tous les QTLs détectés expliquent entre 13,1 (pour mars2011_remplissage) et 59,8% (pour nov_epi) de la variation phénotypique.

Avec le caractère épiaison, quatre QTLs ont été détectés. Parmi ces quatre QTLs, trois l'ont aussi été avec le caractère remplissage. Aucun QTL supplémentaire n'a été détecté avec le caractère remplissage et dans une condition sur deux des QTLs détectés avec le caractère épiaison ne le sont pas avec le caractère remplissage. On pensait s'affranchir de la corrélation entre ces deux caractères en prenant la différence entre deux stades ($cor(epi; score7) = 0,53$), mais la redondance des QTLs détectés montre une corrélation résiduelle ($cor(epi; score7 - epi) = 0,43$), qui est liée à la génétique. Il faudrait donc penser à une autre méthode pour décorréler ces deux caractères. La plus faible héritabilité du caractère remplissage (0,70) par rapport au caractère épiaison (0,95) montre que la variation de la durée de remplissage a une plus grande composante environnementale.

Malgré une très faible structure présente dans la population MAGIC INRA (Chapitre 4), nous avons opté pour une prise en compte de la matrice d'apparentement dans les modèles, qui permettrait une amélioration marginale des détections pour certains caractères. Cette matrice d'apparentement montre effectivement une forte similitude entre neuf couples d'individus (IBS>0,9), ce qui n'était pas attendue sur 380 individus échantillonnés dans une population panmictique de 10 000 individus. Même si elle est faible, il est étonnant que l'effet de cette correction soit caractère dépendant, alors que les individus fortement apparentés ne présentent pas des phénotypes particulièrement similaires sur les différents traits.

La méthode "multi-locus", qui intègre successivement les principaux SNPs en cofacteur dans l'analyse de variance, a été comparée à une méthode locus par locus. Cette méthode a permis d'éliminer les marqueurs redondants (fort DL), et de détecter des QTLs indépendants par rapport à la méthode marqueur par marqueur. L'amélioration de puissance décrite dans de Bakker et al. [2005]; Segura et al. [2012], n'a été visible dans nos analyses que sur l'axe 2 de l'ACP, pour lequel cette méthode a permis de détecter deux QTLs supplémentaires.

La détermination du seuil est une question critique en détection de QTLs :

un seuil trop lâche entraîne la détection de fausses associations alors qu'un seuil trop sélectif diminue fortement la puissance de détection [Bernardo, 2004]. Les seuils de FDR et de Bonferroni sont les seuils généralement utilisés en génétique d'association en correction des tests multiples [Ehrenreich et al., 2009; Cockram et al., 2010]. Une façon alternative de définir le seuil de détection est basée sur le principe des tests exacts, et utilise les permutations pour définir les probabilités d'associations aléatoires [Churchill and Doerge, 1994]. Nous avons testé ces trois approches, la méthode du FDR produit un seuil plus souple tandis que les méthodes de Bonferroni et de permutation ont produit sur nos données des seuils très similaires plus conservatifs. La correction de Bonferroni n'est donc pas particulièrement forte, ce qui est cohérent avec le faible DL présent dans la population, qui entraîne une relative indépendance entre les tests d'association.

L'analyse des nombreux caractères mesurés, et leurs corrélations parfois fortes (stades de développement successifs par exemple), posent des questions méthodologiques, et une analyse intégrée des différents traits est généralement préférable [Stearns et al., 2005]. Les stratégies multi-traits (ACP et modélisation écophysiologique) et trait par trait ont montré des résultats cohérents et complémentaires. En effet quatre QTLs communs ont été identifiés par la méthode trait par trait et les deux méthodes multi-traits, un QTL spécifique étant détecté par l'approche trait par trait, et un second par l'analyse de l'ACP. Les paramètres du modèle écophysiologique Pbase et Vsat correspondant respectivement à la sensibilité à la photopériode et au besoin en vernalisation sont associés respectivement aux mêmes marqueurs que l'épiaison avec un semis de novembre (nov_epi) et l'épiaison avec un semis de printemps (mars2011_epi, mars2012_epi, avril2011_epi et avril2012_epi). Ceci confirme la pertinence du dispositif pour décrire ces deux composantes majeures de la précocité de floraison. L'analyse des axes de l'ACP a produit des résultats surprenants : les mêmes QTLs ont été détectés avec les deux premiers axes de l'ACP, avec un QTL supplémentaire pour l'axe 2, alors que ces axes sont construits de façon à être indépendants. Ce résultat s'explique bien par les projections des variables sur les axes d'ACP, qui montrent que les axes 1 et 2 sont en légère rotation par rapport aux variables semis d'automne/semis de printemps (Figure 5.7). Ces résultats illustrent bien les questions d'interprétation biologique qui se posent à la suite de ce type d'analyses multivariées. La méthode avec l'ACP semble légèrement plus puissante que la méthode trait par trait, puisqu'elle détecte le QTL supplémentaire en comparaison à la méthode trait par trait qui de plus présente le plus petit effet (1,7% de la variance phénotypique expliquée). La détection d'un QTL supplémentaire peut-être due à la méthode multi-traits mais aussi au fait que des caractères supplémentaires ont été utilisés dans l'analyse multi-variées (cinq caractères au lieu de deux). Cette finesse de détection peut être également reliée à la qualité de prédiction des QTLs : la prédiction des coordonnées des deux premiers axes d'ACP ainsi que des paramètres du modèle écophysiologique sont les plus forts, avec nov_epi, ce qui montre certainement la meilleure extraction des effets génétiques produite par les intégrations multi-traits.

La qualité de prédiction a été estimée avec trois modèles dont deux modèles

basés sur les QTLs détectés et un modèle de sélection génomique. La sélection génomique qui prend en compte l'effet de tous les marqueurs explique moins de variance phénotypique que des modèles de types sélection assistée par marqueurs qui sélectionnent que les marqueurs les plus explicatifs. Ceci peut être dû au fait que la précocité de floraison est un caractère qui implique des gènes majeurs. Mais ce résultat est en contradiction avec les résultats d'autres études [Lorenzana and Bernardo, 2009; Heffner et al., 2011] qui trouvent de meilleures prédictions, avec les méthodes de sélection génomique en comparaison avec des méthodes de sélection assistée par marqueurs pour 4 à 12 traits, notamment la date d'épiaison sur sept populations différentes.

Les analyses multi-traits sont de plus en plus développées car il existe maintenant de nombreux jeux de données de phénotypage sur les mêmes plantes ou des phénotypages du même caractère dans des environnements différents. Nous avons testé des méthodes qui créent un nombre de nouvelles variables indépendantes plus restreint à partir d'un plus grand nombre de traits corrélés, mais il existe aussi d'autres méthodes qui analysent les traits deux par deux sur la base de la covariance. Ces méthodes permettent d'étudier la pléiotropie des marqueurs [Mangin et al., 1998; Korte et al., 2012] ou de mieux comprendre les interactions QTLs-environnement [Mathews et al., 2008; van Eeuwijk et al., 2010]. D'autres stratégies d'analyses multi-traits s'attachent plus spécifiquement aux séries temporelles : la cinétique à cinq points que nous avons établie aurait pu être valorisée par modélisation de la courbe puis analyse des paramètres, par des méthodes statistiques dédiées aux séries temporelles, ou par utilisation d'une aire sous la courbe (AUDPC), comme en épidémiologie [Chartrain et al., 2009].

Cette étude, qui cumule une approche génome entier et l'utilisation de marqueurs localisés dans des gènes candidats, nous a permis de détecter six QTLs avec 380 individus. Sur les six QTLs, trois correspondent à des gènes candidats connus pour avoir un rôle majeur dans le contrôle de la précocité de floraison (*Ppd-D1, VrnA1* et *VrnD3 (FTD)*). Les trois autres QTLs sont localisés sur les chromosomes 2B et 5B. Le QTL du chromosome 2B pourrait correspondre à un QTL également identifié dans Le Gouis et al. [2011], associé au besoin en vernalisation et à la sensibilité à la photopériode (Figure 5.26). Sur le chromosome 5B, Le Gouis et al. [2011] a aussi détecté deux zones indépendantes associées à la précocité de floraison, une liée au besoin en vernalisation et une qui pourrait correspondre à *VrnB1*. La comparaison des positions des QTLs détectés dans cette analyse et ceux de l'analyse de Le Gouis et al. [2011] est difficile car l'absence de marqueurs communs empêche la construction d'une carte consensus.

L'analyse a été réalisée avec 4 924 marqueurs, ce qui offre une faible couverture assez hétérogène du génome. Du fait du très faible déséquilibre de liaison de la population MAGIC INRA, cette densité de marquage correspond à un DL moyen entre deux marqueurs consécutifs de 0,5, avec un DL longue distance pratiquement nul (DL=0,003 ; Chapitre 4). Chez l'humain, une densité satisfaisante pour la détection de QTLs a été évaluée à un DL inter-marqueur de $r^2 = 0,8$ [Barrett and Cardon, 2006]. D'après nos calculs, avec notre densité actuelle seulement 55%

du génome est couvert avec un DL supérieur ou égal à ce seuil (Figure 5.19). Nous avons vu que des zones comme celle qui porte *VrnA1* étaient riches en marqueurs, contrairement au voisinage de *VrnD3*, où seul le gène candidat est associé à la précocité de floraison. La faible densité de marquage combinée au très faible DL souligne dans notre cas l'intérêt de l'approche gène candidat, et la nécessité de densifier le marquage pour une approche génome entier (*VrnD3* n'aurait pas été détecté avec la puce 9K iSelect). Ce très faible DL est pénalisant pour la détection de QTLs mais en contrepartie il permet de resserrer l'intervalle de confiance de localisation autour des QTLs identifiés. Ainsi, la distance maximale observée entre des marqueurs associés au même caractère est de 5,6cM, ce qui est bien inférieur aux intervalles de confiance décrits pour des populations bi-parentales de l'ordre de 10cM à 30cM [Cavanagh et al., 2008], ou même de ceux des populations MAGIC 4 parents, de l'ordre de 10cM [Huang et al., 2012b].

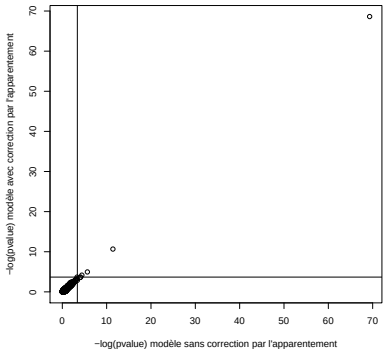
La précocité de floraison est un caractère contrôlé par un grand nombre de QTLs [Buckler et al., 2009; Le Gouis et al., 2011], dont quelques gènes majeurs expliquant un large pourcentage de la variance phénotypique. L'importance de ces gènes majeurs est bien illustrée dans la MAGIC INRA, avec une variance phénotypique expliquée comprise entre 10,7% et 56,3% pour les QTLs les plus forts, leur présence compliquant la détection des autres QTLs. Chez le maïs, Buckler et al. [2009] a trouvé une centaine de QTLs associés à la précocité de floraison alors que chez le blé [Le Gouis et al., 2011] en a trouvé 45. Une fois les quelques QTLs majeurs détectés, les QTLs restants expliqueraient chacun en moyenne entre un et deux pourcents de la variation génétique avec une hypothèse d'additivité. Le dispositif actuel ne permet pas de détecter les QTLs avec un effet si faible, et d'après notre étude de puissance, l'augmentation de densité de marquage de manière à saturer le génome (DL QTL-marqueur = 1) permettrait de détecter les QTLs avec un effet à 2% avec une puissance de 9% mais la détection des QTLs avec un effet à 1% reste toujours impossible. L'augmentation du nombre d'individus est plus bénéfique que l'augmentation du nombre de marqueurs puisqu'il augmenterait la puissance de détection des QTLs avec un effet de 1% et 2% respectivement à 3,5% et 16%. L'augmentation du nombre de marqueurs et du nombre d'individus élèverait ces chiffres à 20% et 40%, ce qui reste toujours très faible. Tout effet confondu, l'étude de puissance a montré que le dispositif actuel détecterait 18% des QTLs détectables en ségrégation dans la population. Sachant que six QTLs ont été détectés, cela signifie que dans cette population 33 QTLs seraient détectables. Ce chiffre de 33 est faible en comparaison avec les études de Buckler et al. [2009] et de Le Gouis et al. [2011]. Ce chiffre pourrait être augmenté en modulant le seuil de détection, mais l'utilisation de plusieurs populations de type MAGIC comme les populations NAM permettrait d'augmenter le nombre de QTLs en ségrégation tout en gardant la finesse de localisation. Sachant que le nombre moyen de QTLs détectés dans des analyses bi-parentales est de neuf [Hanocq et al., 2007], le nombre moyen de QTLs de précocité de floraison en ségrégation dans les populations bi-parentales est donc bien plus faible que le 33 trouvés dans notre population. Même si le pourcentage de détection de QTLs de ce dispositif est pour le moment assez faible, une augmentation

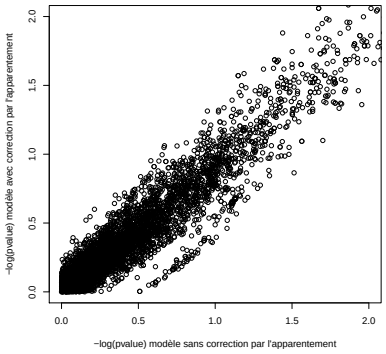
du nombre d'individus ou du nombre de marqueurs permettrait de l'améliorer à respectivement 46% et 36%. Une augmentation du nombre d'individus serait donc préférable à une augmentation du nombre de marqueurs. L'augmentation simultanée des deux paramètres permettrait la détection de 70% des QTLs soit 23 QTLs. Ces chiffres de puissance sont à prendre avec précaution puisque le modèle utilisé est assez simple et pourrait être raffiné, mais ils donnent une idée sur le potentiel de la population et des améliorations possibles.

5.2.5 Figures et tableaux supplémentaires

FIGURE 5.23 – Comparaison des résultats issus du modèle linéaire avec et sans prise en compte de l'apparentement pour le trait nov_epi : 1) Comparaison des -log(p-values) (à gauche) avec agrandissement de la zone avec les p-values comprises entre 0 et 2 (à droite). Les lignes représentent les seuils de significativité trouvés par la méthode de FDR (10%), 2) QQplot avec prise en compte de l'apparentement (à droite) et sans (à gauche), 3) Manhattan plot des résultats de détection d'association.

1)

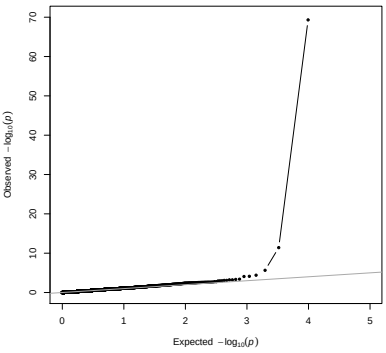


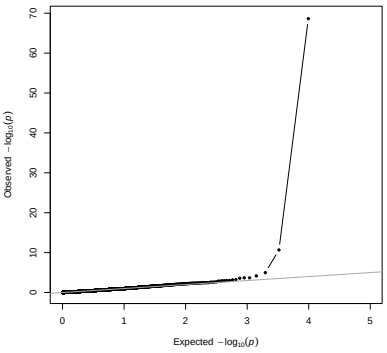


sans apparentement

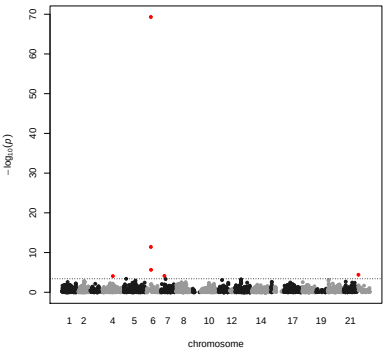
avec apparentement

2) QQplot





3) association marqueur par marqueur



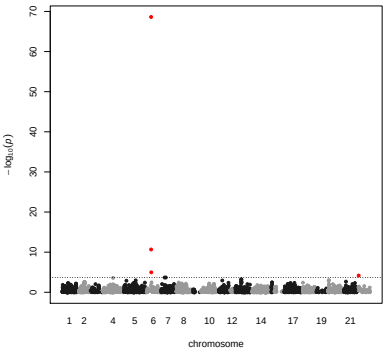


FIGURE 5.24 – Comparaison des résultats d'association du modèle linéaire mixte avec prise en compte de l'apparentement et du modèle linéaire simple pour le caractère nov_epi avec la méthode mono-locus et la méthode multi-locus et avec (à gauche) et sans (à droite) correction par la matrice d'apparentement.

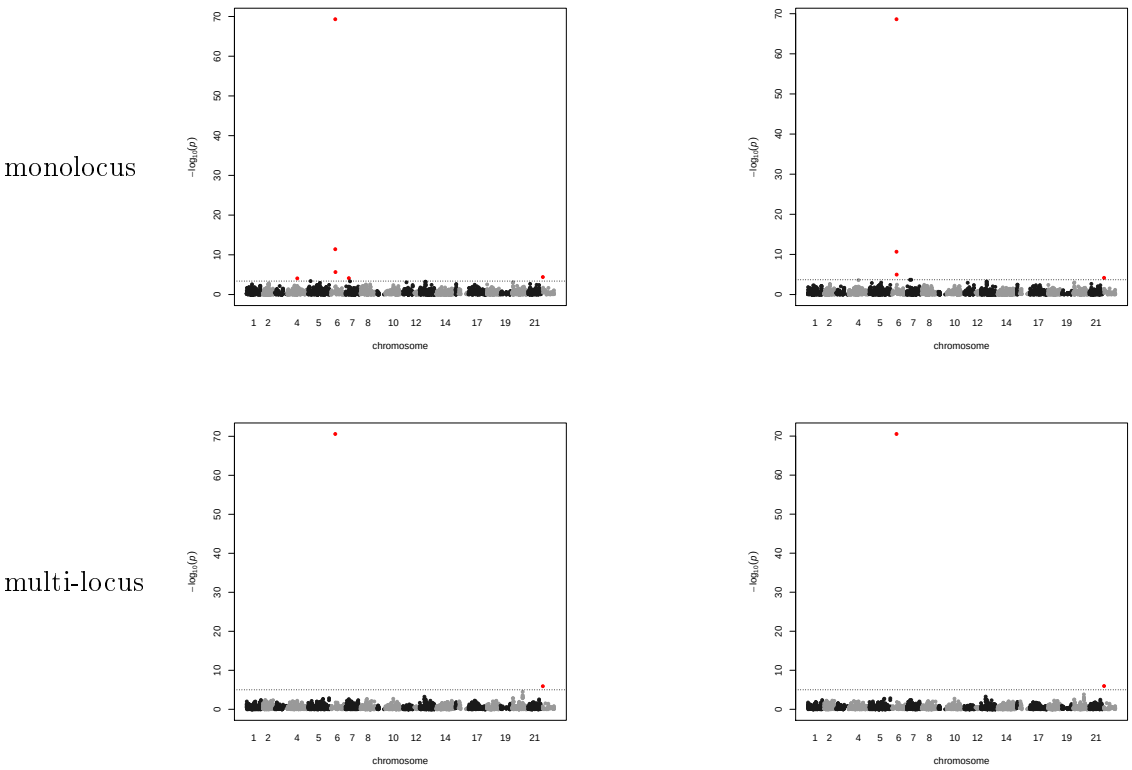
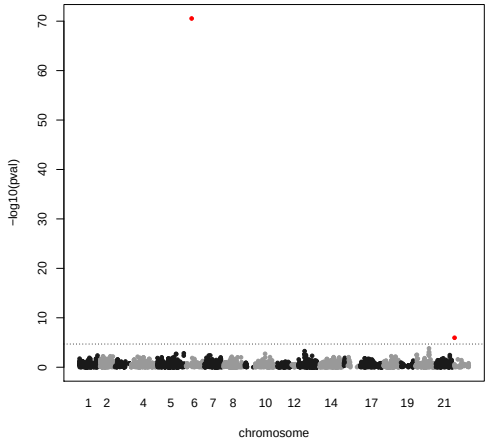


Tableau 5.6 – Résultats des tests d'association (r^2) pour les marqueurs détectés comme associés par la méthode multi-locus : variance phénotypique expliquée par marqueur et par le modèle intégrant tous les QTLs par caractère étudié.

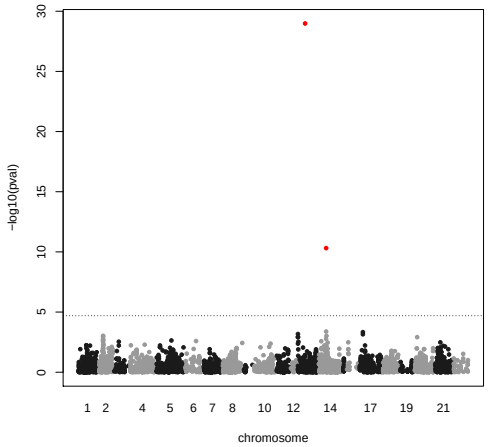
	wsnp_Ex_c32493_41138957	Ppd.D1prom.2kb.indel	VRN1.ex8..1bp.indel.	wsnp_RFL_Contig2809_2587619	wsnp_Ku_c9541_15976096	FTD.G.deletion (VrnD3)	variance globale expliquée par les QTLs détectés
Position (en cM)	188,8	55	72,9	71,1	186	0	
nov_epi		56,3 %				4,1 %	58,9 %
mars2011_epi			26,7 %	6,4 %			34,8 %
mars2012_epi			23,4 %	5,6 %			30,4 %
avril2011_epi			23,4 %	7,5 %			32,5 %
avril2012_epi			25,3 %	6,6 %			33,5 %
photo_epi		12,8 %			8 %		20,3 %
nov_remplissage		16,4 %					16,4 %
mars2011_remplissage			13,1 %				13,1 %
mars2012_remplissage			4 %	20,0 %			25,2 %
avril2011_remplissage			6,7 %	19,9 %			28,0 %
avril2012_remplissage			3,8 %	10,7 %			15,3 %
photo_remplissage							
acp_ax1		6 %	26,8 %	6,4 %			42,4 %
acp_ax2	1,7 %	45,2 %	7,4 %	2,7 %			55,6 %
acp_ax3							
acp_ax4							
Vsat			29,7 %	5,9 %			38,6 %
Pbase		35,6 %				7,9 %	41,8 %

FIGURE 5.25 – Manhattan plot caractère par caractère avec un modèle multi-locus avec prise en compte de l'apparentement, les marqueurs marqués en rouge sont les marqueurs mis en cofacteurs. La ligne en pointillée représente le seuil de 10% corrigé avec la méthode de Bonferroni.

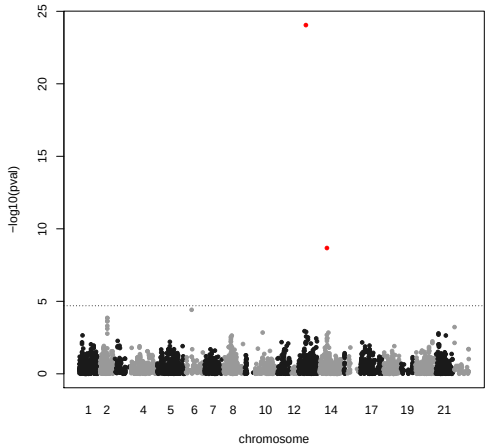
nov_epi



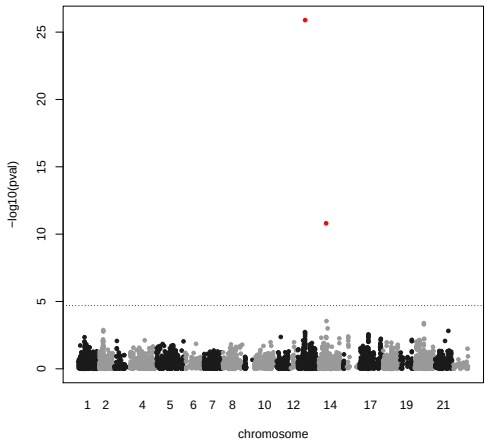
mars2011_epi



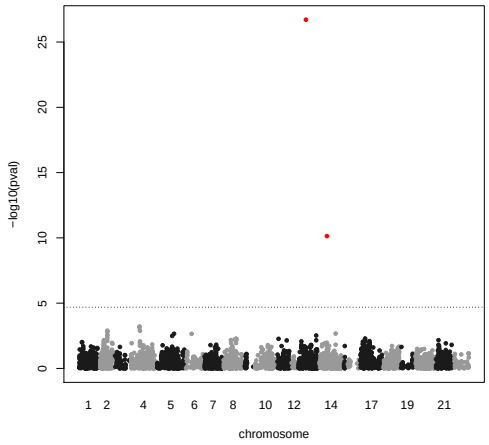
mars2012_epi



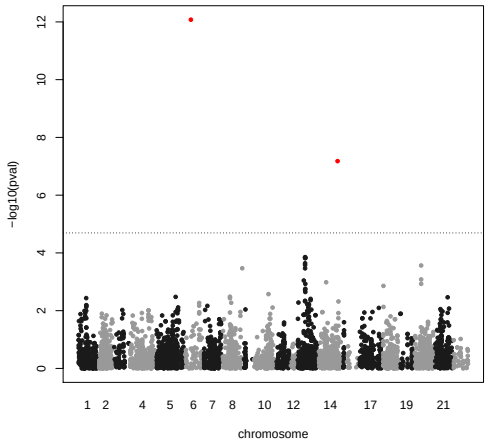
avril2011_epi



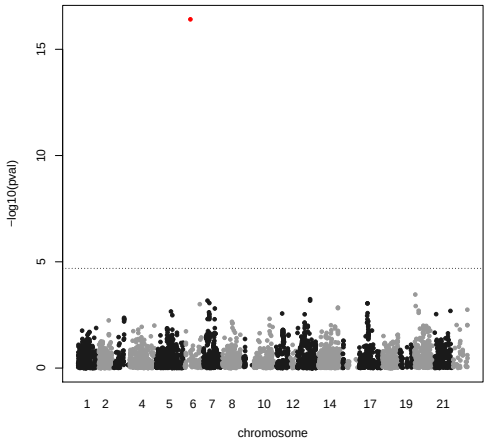
avril2012_epi



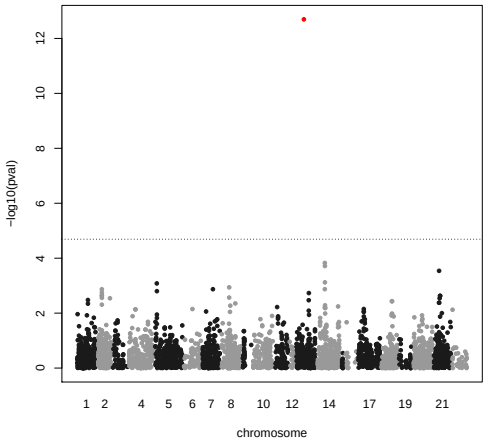
photo_epi



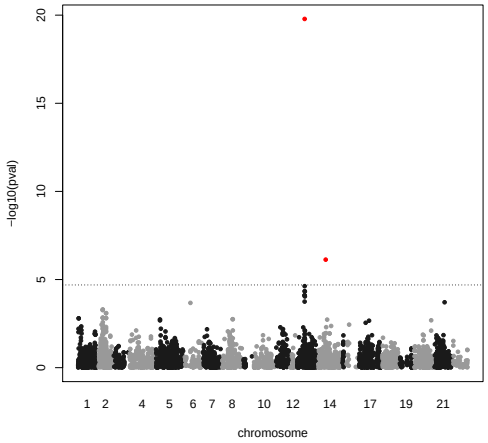
nov_remplissage



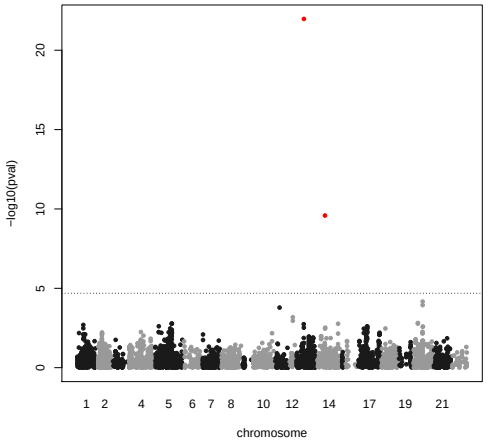
mars2011_remplissage



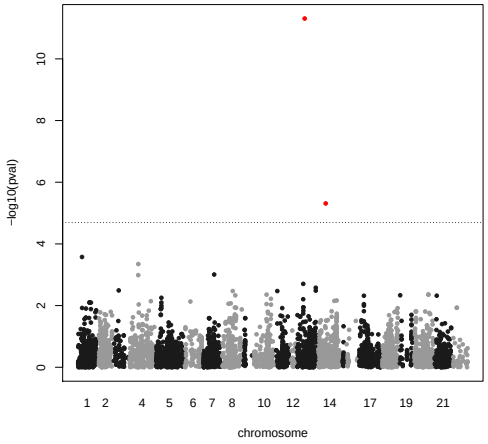
mars2012_remplissage



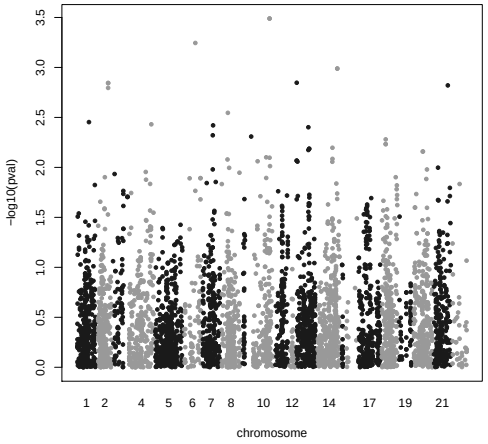
avril2011_remplissage



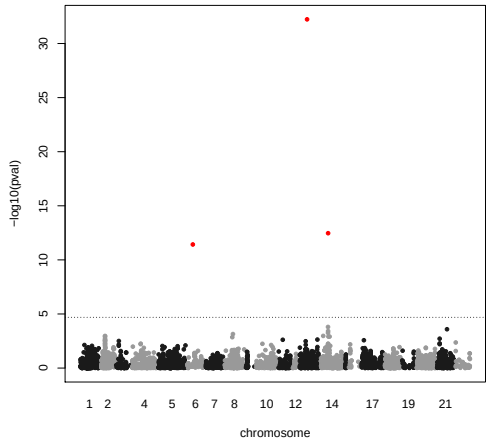
avril2012_remplissage



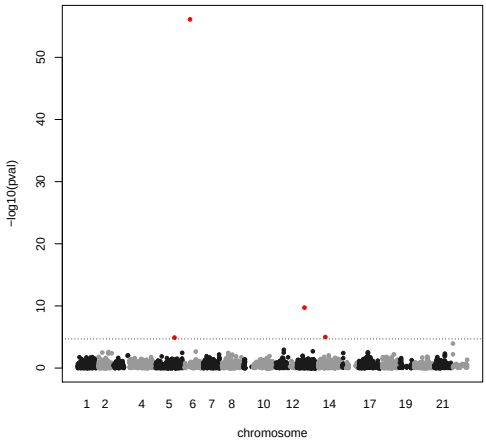
photo_remplissage



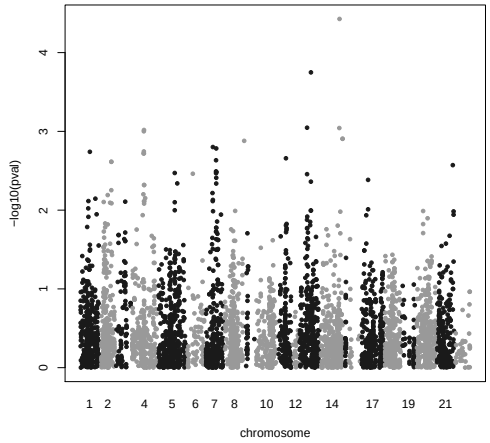
acp_axe1



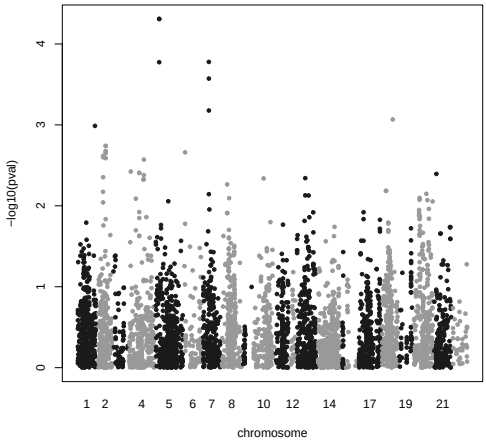
acp_axe2



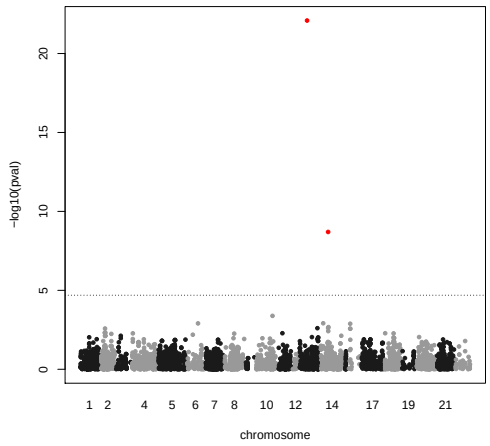
acp_axe3



acp_axe4



Pbase



Vsat

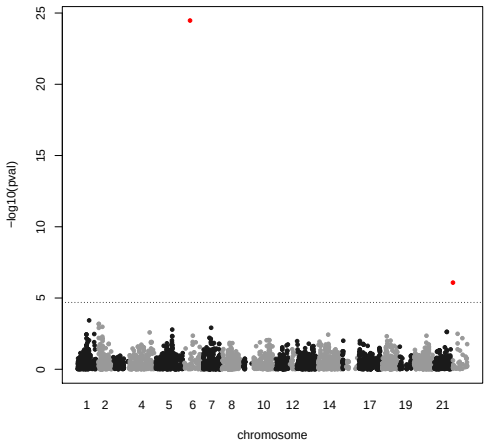
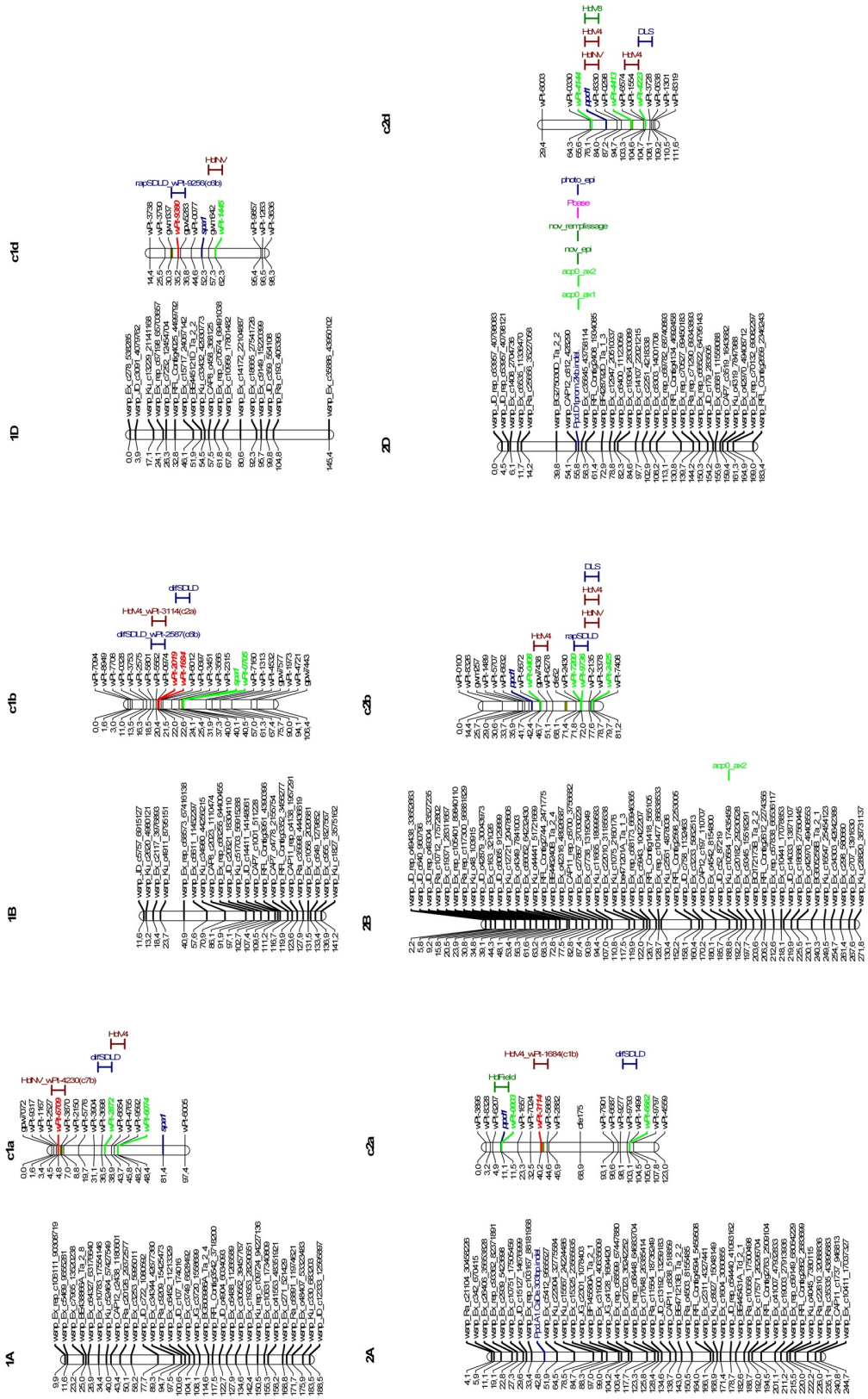
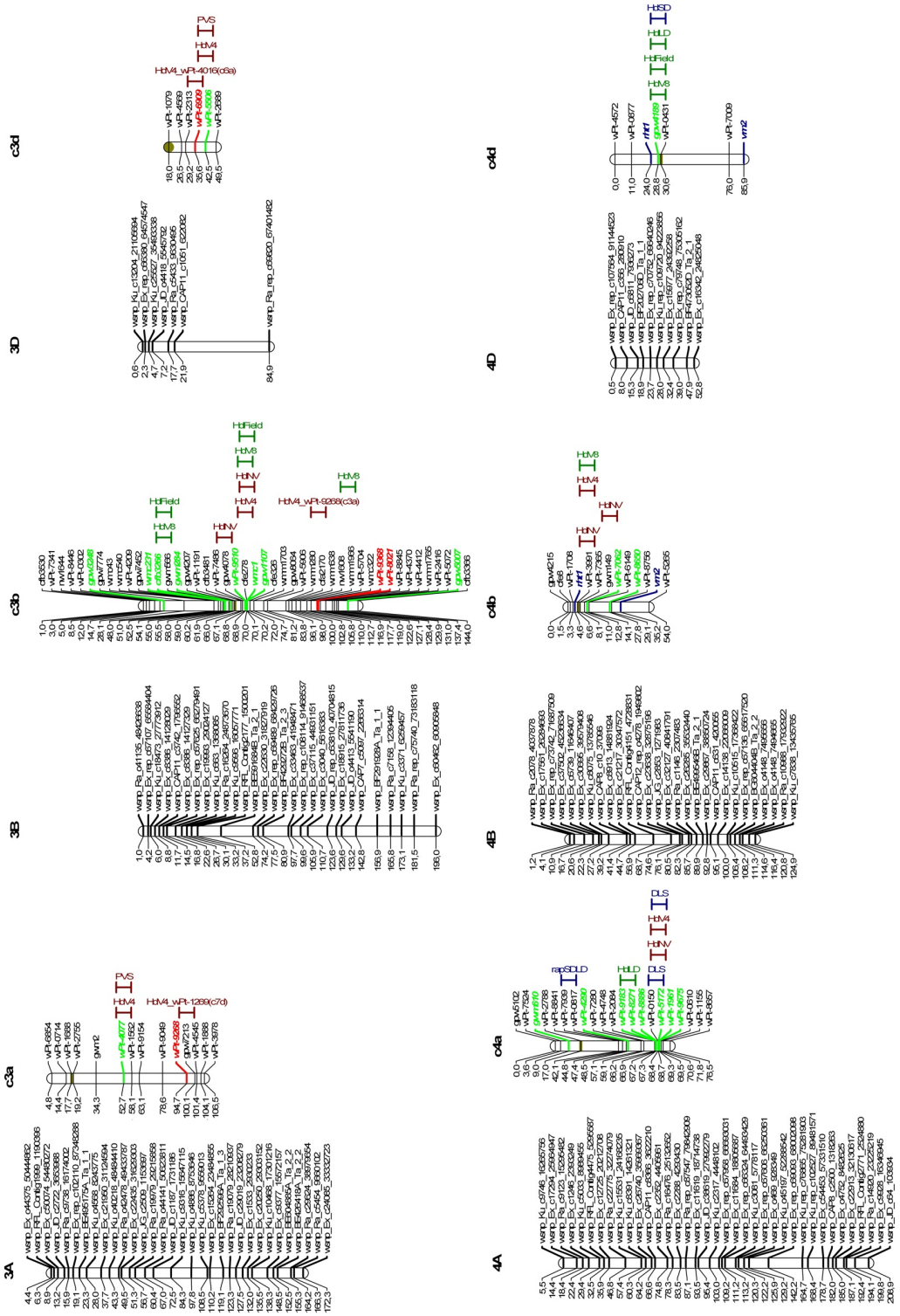
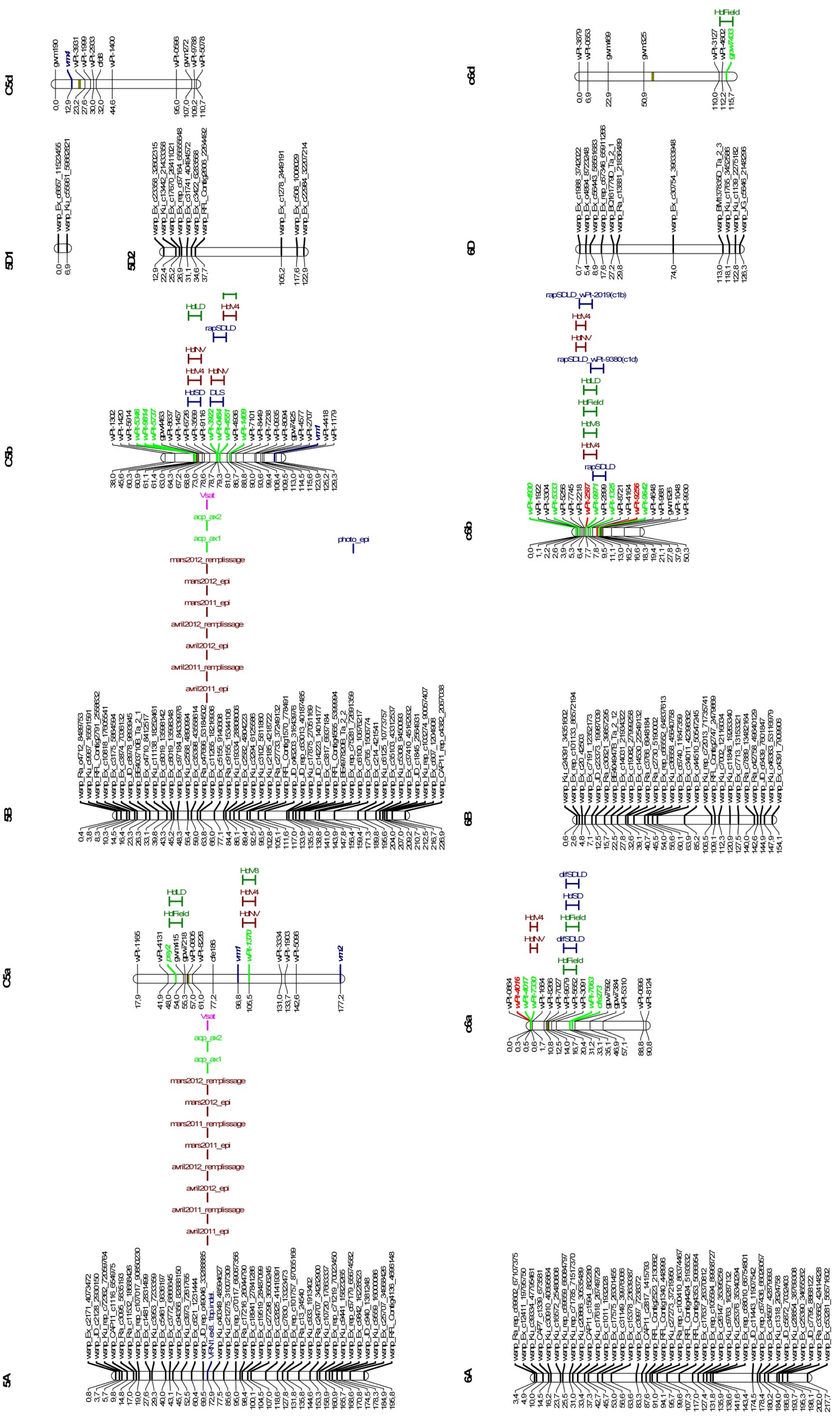
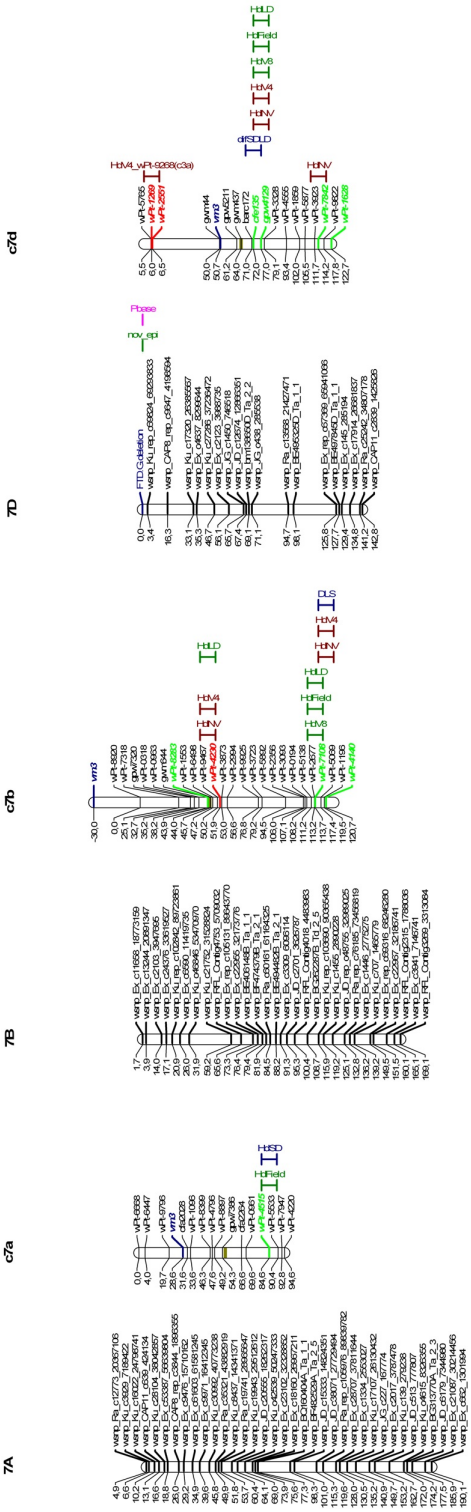


FIGURE 5.26 – Résultats d'association avec la population MAGIC INRA. Les chromosomes avec un nom commençant par "c" sont les résultats d'association de l'étude de Le Gouis et al. [2011].


1A
1B
1D
2A
2B
2D
c1a
c1b
c1d
c2a
c2b
c2d


3A
3B
3D
4A
4B
4D
c3a
c3b
c3d
c4a
c4b
c4d
PVS
HtV4
HtV4_wPt-4016(c3a)
HtV4_wPt-1269(c7d)
HtV4_wPt-9263(c3a)
HtField
HtV8
HtNV
HtLD
HtSD
DLS
rapSDLD


C5d
5D1
5D2
C5b
5B
C5a
5A
c6d
6D
c6b
6B
c6a
6A


7A
7B
7D

Chapitre 6

Conclusion & Perspectives

Les avancées technologiques des outils de génomiques ont formidablement accru la quantité de données moléculaires accessible pour le généticien. Aujourd'hui le phénotypage devient le principal facteur limitant pour la compréhension de l'architecture génétique des caractères complexes. Des plateformes de phénotypage haut débit se développent progressivement [Montes et al., 2007], et permettent des suivis écophysiologiques très précis, retraçant les dynamiques de développement (3D) aériens ou racinaires de centaines de plantes. Il est essentiel de réfléchir aux populations qui permettront de mieux exploiter cette quantité massive d'informations. Ainsi dans les approches classiques de cartographie génétique, l'utilisation de populations biparentales a démontré son intérêt, grâce à une très bonne puissance de détection, dans une situation où les marqueurs génétiques étaient rares et coûteux. Cependant le fort DL de ces populations, avec de larges blocs haplotypiques en ségrégation, entraîne une forte redondance des données de génotypage haut débit, et ne valorise donc pas la quantité de marqueurs disponibles (large intervalle de confiance lors de la cartographie). A l'opposé, les panels d'association disposent d'une puissance de détection plus faible, mais l'importance des brassages génétiques historiques à l'origine du matériel génétique étudié permet une localisation précise des QTLs détectés [Ingvarsson and Street, 2011]. L'étude conjointe de plusieurs populations bi-parentales, reliées par l'utilisation d'un génotype commun (pivot) permet de cumuler les intérêts de ces deux approches (NAM [Yu et al., 2008]). Une autre stratégie réside dans le développement de populations MAGIC, créées par inter-croisement équilibré de nombreux génotypes, dans le but d'avoir une population avec une forte diversité et donc potentiellement un grand nombre de QTLs en ségrégation et une localisation plus précise, grâce à de nombreuses générations d'inter-croisements entre les descendants [Cavanagh et al., 2008]. Le matériel d'étude de cette thèse appartient à cette dernière classe, et peut être considéré comme une population MAGIC "extrême", aussi bien par son nombre de parents (60) que par le nombre de générations d'intercroisement (12). L'objectif principal de cette thèse était d'évaluer l'intérêt de cette population MAGIC INRA pour l'étude de l'architecture génétique d'un caractère complexe, en prenant pour modèle la précocité de floraison. L'étude de la population que ce soit en termes de déséquilibre de liaison (DL), de structure génétique, ou d'évolution temporelle nous a amenés à développer des méthodes spécifiques d'analyse (cartographie génétique sur la base du DL, inférence des contributions parentales, et détection de sélection). Deux méthodes ont été utilisées pour détecter des régions génomiques impliquées dans l'adaptation : l'analyse de régions génomiques soumises à sélection durant les

douze générations de panmixie et la génétique d'association.

6.1 La cartographie par déséquilibre de liaison : intérêts et limites de la méthode

A la réception des premières données de marquage génétique (puce 9K iSelect), la représentation des haplotypes présents dans la population MAGIC INRA et la mise en évidence de groupes de marqueurs cartographiés sur un même point génétique (absence de recombinaison dans les populations utilisées pour la construction de cette carte) a révélé l'existence de recombinaisons potentiellement informatives sur l'ordre des marqueurs. Cette observation, que nous avons attribuée initialement aux 12 générations de recombinaisons panmictiques accumulées dans ce matériel, nous a inspiré le développement d'une méthode de cartographie basée sur l'utilisation du DL. L'objectif de ce développement était d'évaluer l'information contenue dans la matrice de DL pour cartographier les marqueurs de façon à compléter les informations de la carte génétique disponible [Cavanagh et al., 2013] ou de reconstruire une carte génétique.

Les méthodes actuelles de construction de cartes génétiques reposent sur l'écriture des probabilités de recombinaison entre des combinaisons alléliques parentales dans des croisements simples (bi-parentaux, MAGIC à 4/8/16 parents), ou sur des généalogies connues (chez l'homme par exemple [Botstein et al., 1980]). Ces probabilités ne sont pas accessibles pour des populations ayant des pedigrées plus complexes ou non connus, et nous avons donc développé une méthode de cartographie basée sur uniquement la matrice du DL présent entre les marqueurs. Cette méthode, qui ordonne les marqueurs de façon à obtenir un DL décroissant de la façon la plus monotone possible en fonction de la distance, a été testée sur cinq jeux de données différents, allant de la population bi-parentale à la population MAGIC INRA. Sur une population classique de cartographie F2 biparentale, l'algorithme ordonne parfaitement les marqueurs (même ordre que la carte génétique) ; par contre avec des populations plus complexes comme la population MAGIC INRA les marqueurs ont un ordre cohérent avec la carte génétique à une échelle large (100 cM), mais présentent des inversions et incohérences à une échelle locale. Ces résultats de cartographie par le DL non satisfaisants sur des populations de type MAGIC sont dus à l'hypothèse lourde d'une décroissance continue du déséquilibre de liaison avec la distance génétique. Cette hypothèse est recevable à longue distance mais le DL à courte distance évolue théoriquement de manière assez stochastique. En effet le DL longue distance est aisément cassé par les générations de recombinaison (fonction de la distance entre les marqueurs), alors qu'à courte distance, le DL créé de manière stochastique par les effets de la dérive (fonction de la taille de population) est peu affecté par la rareté des événements de recombinaison. Dans le cas de la MAGIC INRA, ce DL courte distance est issu du DL parental (Chapitre 4). Dans le but de n'utiliser que les recombinaisons les plus récentes, nous avons proposé d'utiliser des corrections du DL récemment développées, notamment par

l'apparentement des individus et/ou la structure de la population [Mangin et al., 2012] ou une correction par le DL initial présent chez les parents. Sur la population MAGIC INRA, l'algorithme produit plusieurs groupes de liaison par chromosome, ce qui montre que la densité de marquage pour cette population est insuffisante. La disponibilité de données de marquage génétique plus denses sur la population MAGIC permettra de tester les limites de la méthode, et apportera également une information sur le lien attendu entre QTLs et marqueurs associés.

6.2 Intérêts et limites de la population MAGIC INRA pour l'étude de l'architecture génétique d'un caractère complexe

Les deux chapitres Chap 4 & 5.2 de cette thèse se sont attachés à décrire la structure génétique de la population MAGIC INRA et tester son intérêt pour l'étude de l'architecture génétique d'un caractère complexe, la précocité de floraison. Deux approches complémentaires ont été développées : l'étude des zones du génome soumises à sélection pendant le développement de la population et une étude par génétique d'association.

6.2.1 Identification de QTLs par génétique évolutive

Dans le chapitre 4, nous nous sommes intéressés à l'évolution de la population MAGIC INRA, gérée pendant douze générations en panmixie sur le site du Moulon, dans un projet de gestion dynamique. La population MAGIC est issue d'une population recombinante créée à la fin des années 80, mais les archives disponibles ne permettent pas de retracer parfaitement les premières étapes de sa création, ni le nombre précis de fondateurs (Annexe A). Cependant approcher la composition de la population initiale (G0) est un préalable indispensable pour discuter de l'évolution de la MAGIC. Nous avons développé une méthode bayésienne pour estimer la contribution de chaque fondateur dans la population d'étude (G12), sur la base des fréquences alléliques en G12 et des génotypes des 56 parents encore disponibles (sur les 60 *a priori* utilisés). La méthode a été validée sur des données simulées, et elle a produit des inférences sur la contribution des 56 parents très cohérentes avec les quelques informations que nous avions sur la création de la population (notamment sur la forte contribution de Probus, donneur de la stérilité mâle). Nous avons donc considéré ces contributions comme un état représentatif de la population initiale (G0), et cette population G0 a été utilisée comme référence pour les analyses d'évolution temporelle de la population. Nous sommes conscients que l'inférence de cette population initiale sous-estime par construction les différences entre la population G0 et la population G12. Cependant cette contribution estimée sur l'ensemble du génome reste représentative, car il est difficile d'imaginer une sélection qui élimine l'essentiel d'une contribution parentale, avec le niveau important de brassage généré très précocement lors de la création

de la population. La comparaison entre cette population initiale et la population G12 a confirmé que les douze générations de panmixie ont effectivement cassé la structure initialement présente chez les parents. Cette structure était liée à la fois aux origines géographiques distinctes des parents, issus de pools génétiques contrastés, et à l'apparentement de certaines lignées issues de croisements interspécifiques telle que la série des VPM par exemple. Pour détecter la présence de traces de sélection sur le génome, nous avons développé un nouveau test de sélection, basé sur l'estimation de l'effectif efficace (Ne), lui même estimé à partir de la variance des fréquences alléliques sur tout le génome entre G0 et G12. Ce test repère les variations temporelles de fréquences alléliques à certains locus qui s'avèrent anormalement élevées (*outliers*), calcule leur probabilité sous l'effet de la dérive seule, et estime un coefficient de sélection directionnelle explicatif. Appliquée à nos données (MAGIC INRA G0 et G12, avec un Ne = 310,97), cette méthode a détecté 25 régions soumises à une sélection directionnelle, dont cinq étaient associées à la précocité de floraison. Parmi ces régions, on retrouve des zones déjà connues pour leur rôle sur le contrôle de la précocité de floraison comme *Ppd-D1* et d'autres n'ont pour le moment pas été répertoriées [Le Gouis et al., 2011]. Les zones soumises à sélection et non associées à la précocité de floraison peuvent être associées à d'autres caractères adaptatifs non étudiés ici comme la hauteur ou la résistance aux maladies comme cela a été observé dans les études d'évolution de populations d'orge et de blé [Ibrahim and Barrett, 1991; Paillard et al., 2000a,b]. Cette approche est encourageante car l'étude réalisée minimise les variations de fréquences alléliques temporelles, du fait de l'inférence de la population initiale G0, et sous-estime certainement le nombre de régions soumises à sélection. L'utilisation de la génération G1 (population la plus ancienne encore disponible dans la banque de semences) pourrait confirmer/corriger l'approche par inférence de la G0, et permettre d'augmenter le nombre de régions détectées. Les évolutions rapides observées sur différentes zones génomiques traduisent à la fois la diversité initiale apportée par les 60 parents d'origine mondiale, et l'importance des pressions de sélection s'exerçant lors du maintien de la population. Les nombreuses études conduites par l'équipe DEAP sur les populations autogames du programme de Gestion Dynamique ont déjà montré une rapide évolution phénotypique et génotypique pour des caractères d'adaptation comme la précocité de floraison, des résistances aux maladies ou la hauteur [Paillard et al., 2000a; Goldringer et al., 2006; Raquin et al., 2008a; Rhoné et al., 2010]. Le régime de reproduction impacte fortement la réponse à la sélection [Bürger, 1999] : en autogamie, la fixation en lignées permet une reproduction des génotypes à l'identique au cours des générations, ce qui peut permettre la sélection de combinaisons alléliques, et d'interactions épistatiques. En allogamie, la redistribution aléatoire des QTLs dans les descendants favorise la sélection sur les effets additifs. Si il est difficile de comparer l'importance des effets épistatiques/additifs dans les deux types de population de GD, nous disposons d'informations de DL dans les deux populations. Nous avons vérifié que le DL longue distance en population autogame (0,08 [Raquin et al., 2008a; Rhoné et al., 2007]) était bien supérieur au DL dans la population MAGIC INRA (0,003), et il serait intéressant de comparer cette différence aux attendues sous le seul effet de la dérive.

Pour les zones impliquées dans la précocité de floraison, la sélection d'haplotypes contrastés dans les populations autogames [Rhoné et al., 2010] confirme l'impact des interactions épistatiques entre gènes VRN sur le DL inter-chromosomique. L'existence d'*outliers* sur le DL inter-chromosomique pourrait donc servir de test pour détecter des zones fortement épistatiques [Rohlfs et al., 2010]. Notons que pour effectuer ces comparaisons de structure du DL et d'évolution temporelle, nous disposons d'une autre population autogame et d'une population allogamisée par *Ms1b* ayant évolué à partir de la même population, sur le même nombre de générations, sur le même site et avec la même conduite.

6.2.2 Par l'approche de génétique d'association

L'analyse d'association a été réalisée sur deux traits, la précocité de floraison et la durée de remplissage du grain, mesurés dans six conditions de culture différentes. La comparaison de différents modèles de génétique d'association a montré qu'une correction pour l'apparentement des individus améliorait légèrement la puissance de détection pour certains caractères, alors qu'une analyse multi-locus à l'aide du modèle "mlmm" [Segura et al., 2012] permettait de détecter des QTLs supplémentaires, tout en éliminant les informations redondantes (SNPs en fort DL). L'utilisation du modèle "mlmm" avec prise en compte de l'apparentement a permis de détecter cinq QTLs au total, dont trois localisés dans des gènes candidats. Il faut noter que ces analyses trait-par-trait sont fortement redondantes : en semis de printemps, toutes les analyses détectent les deux mêmes QTLs de précocité de floraison (sur le chromosome 5A et 5B). L'analyse des coordonnées des individus sur des axes d'ACP ou des paramètres de modèle écophysiologique, deux approches intégratives des multiples mesures effectuées, ont aussi détecté cinq QTLs, dont quatre communs avec l'analyse trait-par-trait. Une analyse de la puissance de détection de la population MAGIC INRA avec la densité de marquage actuelle a montré qu'elle était plutôt faible (10%) par rapport à celle d'une population F2 avec le même nombre d'individus (50%). Dans ce calcul de puissance de détection des QTLs, les paramètres clés sont le DL marqueur QTL, l'effet du QTL, la taille de la population de détection, la fréquence allélique du QTL, et enfin la distribution de la variation phénotypique globale présente dans la population pour le caractère étudié. Du fait du faible DL présent dans la population, une densité de marquage élevée (100K au lieu de la puce 9K) doublerait la puissance de détection. La précision de localisation des QTLs bénéficie de ce faible DL, les QTLs étant localisés sur des intervalles de quelques centiMorgans.

Ces analyses d'associations mériteraient d'être étayées par d'autres méthodes, non abordées dans ce chapitre. En complément de la création de variables synthétiques ou reliées à des paramètres biologiques (ACP, modélisation éco-physiologique), des méthodes d'analyses simultanées de plusieurs traits pourraient être utilisées. Par exemple, Malosetti et al. [2008] propose une méthode qui intègre dans un modèle mixte une analyse multi-traits et multi-environnements, basée sur la matrice de variance-covariance entre les caractères et les environnements.

Cette méthode permet donc d'étudier les interactions QTL-environnement, et pourrait mieux exploiter notre plan d'expérimentation, couvrant de nombreux traits mesurés sur deux années climatiques très contrastées. Cette approche permettrait par exemple de mieux comprendre les relations et interactions entre les deux traits étudiés (épiaison et remplissage du grain) et les six environnements différents. Une autre méthode multi-trait propose des solutions graphiques pour identifier les QTLS [Scutari and Nagarajan, 2011] : chaque trait d'étude et chaque marqueur constitue un noeud et les liens représentent la corrélation entre les noeuds. Cette méthode aurait pu révéler les groupes de traits associés, et les groupes de marqueurs liés à ces traits.

6.3 Utilisation des données phénotypiques

Sur cette population INRA MAGIC, un effort considérable de phénotypage a été réalisé durant ces deux années de thèse. Dans le but d'avoir des données les plus fiables possibles, et sachant que différents notateurs pouvaient travailler simultanément sur le dispositif (souvent trois personnes mobilisées), nous avons développé différents outils : i) la mise en forme de protocoles simples et illustrés (annexe E) ii) la mise en place d'un didacticiel sur la base de photographies pour apprendre, s'entraîner et s'évaluer sur la notation des différents stades de développement (annexe F) iii) l'élaboration d'un script de fusion de fichiers de notations avec des filtres pour des éventuels erreurs de frappes, incohérences de notations ou conflits entre deux fichiers de notations (annexe G).

Etant donnée la quantité d'informations récoltées durant cette thèse que ce soit au niveau phénotypique ou génotypique, nous avons également réfléchi à l'intégration de toutes ces données dans une base adaptée à nos besoins sur la base de Thalia-DB, une base de données créée pour gérer des données utilisées pour des analyses de génétique d'association. L'intégration de toutes ces données dans une base de données est un travail considérable qui permettrait de mieux les valoriser sur le long terme.

En effet cette thèse n'aborde qu'une partie de la valorisation de toutes ces données. Je n'ai étudié que les données de phénotypage de pépinière liées à la précocité de floraison et au remplissage du grain, mais des données des mêmes caractères sont disponibles avec des micro-parcelles en France et en Angleterre, ainsi que des données de rendements et de hauteur.

6.4 Perspectives

Différentes analyses se sont heurtées à la trop faible densité de marquage disponible actuellement sur la MAGIC INRA, en raison du très faible DL qu'elle présente. Dans le projet ANR Breedwheat, un génotypage de cette population (380 lignées SSD + 56 parents) est prévu avec une puce de 420 000 SNPs. Cette densité de marquage devrait impacter toutes les analyses réalisées dans cette thèse : à la fois les

approches de cartographie par DL qui pourront être améliorées par la disponibilité de marqueurs liés et possédant des MAF différentes, l'inférence des contributions parentales bénéficiera du plus grand nombre de polymorphismes spécifiques à une lignée parentale, et l'augmentation de densité conduira également à un meilleur DL QTL-marqueur, et donc une meilleure puissance de détection. Par contre des adaptations seront nécessaires sur les algorithmes (parallélisation, travail sur des sous-jeux de marqueurs), pour répondre à une explosion des temps de calcul. L'ajout de SNPs devrait donc permettre la détection de QTLs additionnels à effet plus faible. Avec cette densité, une analyse basée sur l'inférence d'haplotypes parentaux et ancestraux pourrait être envisagée ce qui augmenterait encore la puissance de détection [Meuwissen and Goddard, 2004]. En plus de l'augmentation de la couverture du génome qui apporterait un gain de puissance de 9%, une augmentation du nombre d'individus pourrait être envisageable puisque cette population contient 1 026 lignées SSD. L'étude de puissance a estimée que l'ajout de ces 646 individus supplémentaires permettrait un gain supplémentaire de puissance de 18%.

La population MAGIC INRA a été créée sur une trentaine d'années, ce qui représente un investissement aussi bien financier qu'humain, et demande une continuité de moyens pour mener des projets de recherche à l'échelle de plus d'une génération de chercheur. Par comparaison, les populations de cartographie bi-parentales, lorsqu'elles bénéficient d'une fixation de lignées au moyen d'haploïdes doublés, peuvent être générées relativement rapidement (4-6 années). Il est donc important de bien évaluer l'intérêt de la population MAGIC INRA, au vu des forces et faiblesse décrites dans les chapitres précédents. La population MAGIC INRA se rapproche dans sa structure d'un panel d'association, du fait de sa grande diversité (60 parents fondateurs) et du faible DL lié aux nombreuses générations de recombinaison et aux recombinaisons ancestrales portées par les fondateurs. Elle s'affranchit cependant des problèmes de structure liée à l'histoire évolutive des lignées constitutives d'un panel. Son utilisation en détection de QTLs se heurte principalement aux conséquences d'une population à base génétique large : il existe effectivement un compromis entre le nombre de parents utilisés et la puissance de détection de la population. Une perte de puissance découle du grand nombre de QTLs en ségrégation et de la présence d'allèles rares aux QTLs. Ces QTLs en ségrégation dans la population se divisent le pourcentage de variance phénotypique expliquée, et apportent donc chacun une contribution faible à la variabilité phénotypique. La détection des QTLs à effet faible demande une population avec un grand nombre d'individus [Ersoz et al., 2007], avec les coûts de phénotypage inhérents. A ce nombre élevé de QTLs s'ajoute la question de la distribution de leurs effets. Dans le cas de caractères comme la précocité de floraison, un ou deux gènes ont un effet majeur sur le phénotype, écrasant la variance génétique répartie sur les autres QTLs. Les QTLs mineurs sont difficiles à détecter à cause de leur faible association au phénotype dû à une faible taille de population mais aussi à cause de relations épistatiques ou d'interaction génotype-environnement ou à cause d'une confusion d'effet avec un QTL majeur. Mais ces QTLs sont importants puisque cumulés ils peuvent avoir un effet aussi important qu'un QTL majeur [Yang

et al., 2013].Plusieurs méthodes existent pour étudier spécifiquement les QTLs à effet faible. La création de populations ou panels monomorphes pour le ou les gènes majeurs permettrait de faciliter la détection de QTLs à effet plus faible [Sinha et al., 2008].

Une fois que le pool parental a été défini, quelle méthode de brassage utiliser ? Pour les populations autogames, les croisements manuels entre un grand nombre de parents sont coûteux et difficiles à organiser. La solution retenue pour la MAGIC a été de modifier son régime de reproduction, et de la rendre allogame grâce à l'intégration d'un gène de stérilité mâle. Pour ensuite maintenir cette stérilité mâle au cours des générations, et ainsi garantir le brassage génétique, un marquage rigoureux des mâles stériles est indispensable. En lien avec les contraintes de gestion, ce marquage peut entraîner des biais de sélection dans la population. Ce biais pourrait provenir d'un marquage préférentiel des plantes i) les plus précoces, plus facile à repérer (premiers épis dans la parcelle) et à identifier (bâillement des glumelles uniquement dû à la stérilité et non au gonflement du grain à des stades plus avancés), ii) les plus grandes, et donc facile à repérer/marquer, iii) les plus productive et plus résistante, l'oeil pouvant être attiré par les épis les plus beaux et les plantes les plus belles... Certains de ces biais peuvent expliquer l'évolution vers une plus grande précocité de la population. Une des difficultés du marquage tient à la difficulté de repérage des mâles stériles (Figure 2.1), le bâillement n'étant visible que trois à quatre jours, ce qui demande un passage régulier dans la parcelle (2 à 3 fois par semaine). La pléiotropie de certains gènes de stérilité [Rao et al., 1990] pourrait être utile pour éviter ce marquage fastidieux. En effet si la stérilité était liée à un phénotype visible tout au long du cycle de développement de la plante, le marquage de tous les mâles-stériles pourrait être réalisé en un seul passage. L'identification de la présence de la stérilité sur grain éliminerait même cette étape de marquage. En effet, chez le maïs un gène de stérilité *ms1* est lié à la couleur du grain (jaune ou blanc), chez l'orge aussi la présence de l'allèle de stérilité de msg_{19} ou de msg_6 est visible sur les graines, mais elles sont chétives, ce qui peut avoir un impact négatif au niveau agronomique (fitness). La possibilité de reconnaître les génotypes stériles dès le stade grain permet cependant des dispositifs d'une grande efficacité : il serait par exemple possible de semer des bandes de plantes stériles et d'autres fertiles, pour ne récolter que les bandes stériles, ce qui allègerait considérablement la gestion.

Un des intérêts des populations de GD réside dans leur capacité à révéler les régions génomiques impliquées dans une adaptation locale ou un autre avantage sélectif, grâce à l'analyse des évolutions temporelles des fréquences alléliques. Cependant ces évolutions temporelles peuvent s'avérer négatives pour les approches de génétique d'association. Tout d'abord, la sélection (directionnelle ou stabilisatrice) amène généralement la population vers un optimum, ce qui peut éliminer une partie de la variance phénotypique initiale. Ainsi la phénologie de la MAGIC INRA a évolué vers une plus grande précocité, tout en réduisant la variance phénotypique majoritairement due aux pressions de sélection durant le maintien de la population par comparaison aux parents. Ensuite cette réduction de variance peut être liée à la fixation de certains QTLs, ce qui également diminue la puissance de

détection ces QTLs. Dans quelques cas, un allèle rare à un QTL dans la population initiale pourra voir sa fréquence allélique augmentée, ce qui facilitera sa détection [Mackinnon and Georges, 1992].

Face à ces conflits entre les approches par génétique évolutive et par génétique d'association, il est possible de concevoir une stratégie en deux étapes. Dans un réseau de gestion dynamique, ou de sélection participative, une population en ségrégation est distribuée largement, et cultivée sous des environnements (pratiques × milieu) très contrastés. Ce réseau constitue une métapopulation. Les pressions de sélection spécifiques à certains environnements conduisent les sous-populations à s'adapter localement, diverger, se structurer. Pour explorer l'adaptation à un environnement donné, une première étape pourrait consister à faire une analyse d'évolutions temporelles (ou de structuration inter-populations), ciblée sur des populations soumises à des stress particuliers (sècheresse, maladie, etc...) afin de détecter par des analyses moléculaires les zones soumises à sélection, comme cela a pu être réalisées précédemment avec des populations autogames [Rhoné, 2008]. Avec les progrès des technologies de génotypage/séquençage, nous pouvons anticiper que le coût d'une telle étude deviendra extrêmement faible, et que d'ici quelques années il sera aisé d'acquérir des données moléculaires avec une couverture génomique, sur des milliers d'individus. Une fois qu'une ou plusieurs zones génomiques auront été repérées comme pouvant être impliquées dans une adaptation locale, une nouvelle sélection d'individus, puisés dans les populations adaptées et non adaptées, voire les parents, pourra être effectuée à partir des données moléculaires de la population initiale. Le but sera alors de créer un panel diversifié et présentant des fréquences alléliques équilibrées dans les zones d'intérêt, afin de maximiser la puissance de détection de leurs effets. La méthode pour définir le panel pourrait s'inspirer des méthodes déjà développées pour diminuer le nombre d'individus à phénotyper tout en optimisant la puissance de détection et la résolution de localisation de QTLs [Huang et al., 2012a]. Ce panel serait ensuite phénotypé pour réaliser une analyse d'association avec une stratégie gène-candidat.

L'étude de cette population sous différents angles (la cartographie, la détection de QTLs par étude évolutive et par génétique d'association) a montré un potentiel de cette population pour mieux comprendre l'architecture génétique de caractère complexe grâce au développement d'outils adaptés. Le principal avantage de cette population est le faible DL induit par les douze générations de recombinaison qui permet une localisation fine des QTLs détectés. Cependant le dispositif actuel présente un certain nombre de limites notamment dues à la densité de marquage et à la taille de la population qui est pourtant relativement élevée en comparaison aux populations F2 ou à des panels. Sa complexité relative à sa grande diversité et son faible DL demande un grand investissement pour explorer sa richesse. La stratégie d'étude proposée pourrait être mise en place au sein des réseaux déjà existants comme le réseau de stations de recherche de l'INRA ou d'Arvalis institut du végétal ou un réseau de fermes en sélection participative. Ce dispositif aurait l'avantage de fournir un matériel adapté aux conditions de culture à des sélectionneurs ou à des paysans et de fournir une ressource d'étude intéressante pour comprendre les

mécanismes d'évolution et de contrôle génétique des caractères adaptatifs.

Bibliographie

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6) :716–723. (Cité en page 42.)

Akey, J. M., Zhang, K., Xiong, M., Doris, P., and Jin, L. (2001). The effect that genotyping errors have on the robustness of common linkage-disequilibrium measures. *The American Journal of Human Genetics*, 68(6) :1447–1456. (Cité en pages 36 et 59.)

Akhunova, A. R., Matniyazov, R. T., Liang, H., and Akhunov, E. D. (2010). Homoeolog-specific transcriptional bias in allopolyploid wheat. *BMC Genomics*, 11(1) :505. (Cité en page 5.)

Allard, R. W. (1988). Genetic changes associated with the evolution of adaptedness in cultivated plants and their wild progenitors. *Journal of Heredity*, 79(4) :225–238. (Cité en pages 4, 5, 6, 7 et 13.)

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3) :403–410. (Cité en page 31.)

Alvarez, N., Garine, E., Khasah, C., Dounias, E., Hossaert-McKey, M., and McKey, D. (2005). Farmers' practices, metapopulation dynamics, and conservation of agricultural biodiversity on-farm : a case study of sorghum among the duupa in sub-sahelian cameroon. *Biological Conservation*, 121(4) :533–543. (Cité en page 4.)

Andeden, E. E., Yediay, F. E., Baloch, F. S., Shaaf, S., Kilian, B., Nachit, M., and Özkan, H. (2011). Distribution of vernalization and photoperiod genes (*Vrn-A1, Vrn-B1, Vrn-D1, Vrn-B3, Ppd-D1)* in turkish bread wheat cultivars and landraces. *Cereal Research Communications*, 39 :352–364. (Cité en pages 18, 19 et 23.)

Arumuganathan, K. and Earle, E. D. (1991). Nuclear DNA content of some important plant species. *Plant molecular biology reporter*, 9(3) :208–218. (Cité en page 14.)

Astle, W. and Balding, D. J. (2009). Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, 24(4) :451–471. (Cité en pages 144 et 147.)

Barnes, C. and Bugbee, B. (1991). Morphological responses of wheat to changes in phytochrome photoequilibrium. *Plant physiology*, 97(1) :359–365. (Cité en page xi.)

Barrett, J. C. and Cardon, L. R. (2006). Evaluating coverage of genome-wide association studies. *Nature Genetics*, 38(6) :659–662. (Cité en page 163.)

Barrett, R. D. and Schluter, D. (2008). Adaptation from standing genetic variation. *Trends in Ecology & Evolution*, 23(1) :38–44. (Cité en pages 2 et 6.)

Beales, J., Turner, A., Griffiths, S., Snape, J., and Laurie, D. (2007). A pseudo-response regulator is misexpressed in the photoperiod insensitive ppd-d1a mutant of wheat (triticum aestivum l.). *Theoretical and Applied Genetics*, 115(5) :721–733. (Cité en pages 16, 32 et 70.)

Beddington, J. (2010). Food security : contributions from science to a new and greener revolution. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 365(1537) :61–71. (Cité en page 1.)

Bennett, J. H. (1954). On the theory of random mating. *Annals of eugenics*, 17(1) :311–317. (Cité en page 36.)

Bentley, A. R., Turner, A. S., Gosman, N., Leigh, F. J., Maccaferri, M., Dreisigacker, S., Greenland, A., and Laurie, D. A. (2011). Frequency of photoperiod-insensitive ppd-a1a alleles in tetraploid, hexaploid and synthetic hexaploid wheat germplasm. *Plant Breeding*, 130(1) :10–15. (Cité en pages 17 et 32.)

Bernardo, R. (2004). What proportion of declared QTL in plants are false ? *Theoretical and Applied Genetics*, 109(2) :419–424. (Cité en page 162.)

Bland, J. M. and Altman, D. G. (1995). Multiple significance tests : the bonferroni method. *BMJ : British Medical Journal*, 310(6973) :170. (Cité en page 143.)

Bonneuil, C. and Thomas, F. (2009). *Gènes, pouvoirs et profits : Recherche publique et régimes de production des savoirs de Mendel aux OGM*. Éditions Quae. (Cité en page 2.)

Bonnin, I., Rousset, M., Madur, D., Sourdille, P., Dupuits, C., Brunel, D., and Goldringer, I. (2008). FT genome a and d polymorphisms are associated with the variation of earliness components in hexaploid wheat. *Theoretical and Applied Genetics*, 116(3) :383–394. (Cité en pages 16 et 32.)

Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American journal of human genetics*, 32(3) :314. (Cité en page 178.)

Brenchley, R., Spannagl, M., Pfeifer, M., Barker, G. L. A., D'Amore, R., Allen, A. M., McKenzie, N., Kramer, M., Kerhornou, A., Bolser, D., Kay, S., Waite, D., Trick, M., Bancroft, I., Gu, Y., Huo, N., Luo, M.-C., Sehgal, S., Gill, B., Kianian, S., Anderson, O., Kersey, P., Dvorak, J., McCombie, W. R., Hall, A., Mayer, K. F. X., Edwards, K. J., Bevan, M. W., and Hall, N. (2012). Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, 491(7426) :705–710. PMID : 23192148. (Cité en page 14.)

Bretting, P. K. and Duvick, D. N. (1997). Dynamic conservation of plant genetic resources. *Advances in Agronomy*, 61 :1–51. (Cité en pages 3 et 4.)

Brisson, N., Gate, P., Gouache, D., Charmet, G., Oury, F.-X., and Huard, F. (2010). Why are wheat yields stagnating in europe ? a comprehensive data analysis for france. *Field Crops Research*, 119(1) :201–212. (Cité en page 1.)

Broman, K. W. (2001). Review of statistical methods for QTL mapping in experimental crosses. *Lab Animal*, 30(7). (Cité en page 35.)

Browning, S. R. and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, 81(5) :1084–1097. (Cité en page 146.)

Brush, S. B. (2000). The issues of in situ conservation of crop genetic resources. *Genes in the Field. On-Farm Conservation of Crop Diversity, IPGRI, IDRC, Lewis Publishers*, page 3–26. (Cité en page 4.)

Buckler, E. S., Holland, J. B., Bradbury, P. J., Acharya, C. B., Brown, P. J., Browne, C., Ersoz, E., Flint-Garcia, S., Garcia, A., Glaubitz, J. C., Goodman, M. M., Harjes, C., Guill, K., Kroon, D. E., Larsson, S., Lepak, N. K., Li, H., Mitchell, S. E., Pressoir, G., Peiffer, J. A., Rosas, M. O., Rocheford, T. R., Romay, M. C., Romero, S., Salvo, S., Villeda, H. S., Sofia da Silva, H., Sun, Q., Tian, F., Upadyayula, N., Ware, D., Yates, H., Yu, J., Zhang, Z., Kresovich, S., and McMullen, M. D. (2009). The genetic architecture of maize flowering time. *Science*, 325(5941) :714–718. (Cité en pages 150 et 164.)

Bullrich, L., Appendino, M., Tranquilli, G., Lewis, S., and Dubcovsky, J. (2002). Mapping of a thermo-sensitive earliness per se gene on triticum monococcum chromosome 1Am. *Theoretical and Applied Genetics*, 105(4) :585–593. (Cité en page 17.)

Butault, J. P., Dedryver, C. A., Gary, C., Guichard, L., Jacquet, F., Meynard, J. M., Nicot, P., Pitrat, M., Reau, R., and Sauphanor, B. (2010). Ecophyto R&D quelles voies pour réduire l'usage des pesticides. *Synthèse du rapport d'étude. France : INRA éditeur.* (Cité en page 2.)

Bürger, R. (1999). Evolution of genetic variability and the advantage of sex and recombination in changing environments. *Genetics*, 153(2) :1055–1069. PMID : 10511578. (Cité en pages 8 et 180.)

Carbone, M. A., Jordan, K. W., Lyman, R. F., Harbison, S. T., Leips, J., Morgan, T. J., DeLuca, M., Awadalla, P., and Mackay, T. F. (2006). Phenotypic variation and natural selection at catsup, a pleiotropic quantitative trait gene in drosophila. *Current Biology*, 16(9) :912–919. (Cité en pages 9 et 10.)

Cardinale, B. J., Duffy, J. E., Gonzalez, A., Hooper, D. U., Perrings, C., Venail, P., Narwani, A., Mace, G. M., Tilman, D., and Wardle, D. A. (2012). Biodiversity loss and its impact on humanity. *Nature*, 486(7401) :59–67. (Cité en page 2.)

Cavanagh, C., Morell, M., Mackay, I., and Powell, W. (2008). From mutations to MAGIC : resources for gene discovery, validation and delivery in crop plants. *Current Opinion in Plant Biology*, 11(2) :215–221. (Cité en pages 164 et 177.)

Cavanagh, C. R., Chao, S., Wang, S., Huang, B. E., Stephen, S., Kiani, S., Forrest, K., Saintenac, C., Brown-Guedira, G. L., Akhunova, A., See, D., Bai, G., Pumphrey, M., Tomar, L., Wong, D., Kong, S., Reynolds, M., Silva, M. L. d., Bockelman, H., Talbert, L., Anderson, J. A., Dreisigacker, S., Baenziger, S., Carter, A., Korzun, V., Morrell, P. L., Dubcovsky, J., Morell, M. K., Sorrells, M. E., Hayden, M. J., and Akhunov, E. (2013). Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *Proceedings of the National Academy of Sciences*, 110(20) :8057–8062. PMID : 23630259. (Cité en pages 11, 31, 36, 43, 44, 45, 56, 61, 69, 148 et 178.)

Ceccarelli, S. and Grando, S. (2007). Decentralized-participatory plant breeding : an example of demand driven research. *Euphytica*, 155(3) :349–360. (Cité en page 4.)

Chao, S., Dubcovsky, J., Dvorak, J., Luo, M.-C., Baenziger, S., Matnyazov, R., Clark, D., Talbert, L., Anderson, J., Dreisigacker, S., Glover, K., Chen, J., Campbell, K., Bruckner, P., Rudd, J., Haley, S., Carver, B., Perry, S., Sorrells, M., and Akhunov, E. (2010). Population- and genome-specific patterns of linkage disequilibrium and SNP variation in spring and winter wheat (triticum aestivum l.). *BMC Genomics*, 11(1) :727. (Cité en page 31.)

Charmet, G. (2011). Wheat domestication : Lessons for the future. *Comptes Rendus Biologies*, 334(3) :212–220. (Cité en page 14.)

Chartrain, L., Sourdille, P., Bernard, M., and Brown, J. K. M. (2009). Identification and location of stb9, a gene for resistance to septoria tritici blotch in wheat cultivars courtot and tonic. *Plant Pathology*, 58(3) :547–555. (Cité en page 163.)

Chaumet, J. M., Delpeuch, F., Dorin, B., Ghersi, G., Hubert, B., Le Cotty, T., Paillard, S., Petit, M., Rastoin, J. L., and Ronzon, T. (2009). Agrimonde : Agricultures et alimentations du monde en 2050 : scénarios et défis pour un développement durable. *CIRAD/INRA*. (Cité en page 2.)

Cheema, J. and Dicks, J. (2009). Computational approaches and software tools for genetic linkage map estimation in plants. *Briefings in Bioinformatics*, 10(6) :595–608. PMID : 19933208. (Cité en pages 40, 59 et 60.)

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3) :759–771. (Cité en page 148.)

Churchill, G. A., Airey, D. C., Allayee, H., Angel, J. M., Attie, A. D., Beatty, J., Beavis, W. D., Belknap, J. K., Bennett, B., and Berrettini, W. (2004). The collaborative cross, a community resource for the genetic analysis of complex traits. *Nature genetics*, 36(11) :1133–1137. (Cité en page 11.)

Churchill, G. A. and Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics*, 138(3) :963–971. PMID : 7851788. (Cité en pages 143 et 162.)

Clark, R. M., Schweikert, G., Toomajian, C., Ossowski, S., Zeller, G., Shinn, P., Warthmann, N., Hu, T. T., Fu, G., and Hinds, D. A. (2007). Common sequence polymorphisms shaping genetic diversity in arabidopsis thaliana. *Science*, 317(5836) :338–342. (Cité en page 142.)

Cockram, J., Jones, H., Leigh, F. J., O'Sullivan, D., Powell, W., Laurie, D. A., and Greenland, A. J. (2007). Control of flowering time in temperate cereals : genes, domestication, and sustainable productivity. *Journal of experimental botany*, 58(6) :1231–1244. (Cité en pages 22 et 23.)

Cockram, J., White, J., Zuluaga, D. L., Smith, D., Comadran, J., Macaulay, M., Luo, Z., Kearsey, M. J., Werner, P., and Harrap, D. (2010). Genome-wide association mapping to candidate polymorphism resolution in the unsequenced barley genome. *Proceedings of the National Academy of Sciences*, 107(50) :21611–21616. (Cité en page 162.)

Conway, G. (1998). *The doubly green revolution : food for all in the twenty-first century*. Cornell University Press. (Cité en page 2.)

Darvasi, A. and Soller, M. (1995). Advanced intercross lines, an experimental population for fine genetic mapping. *Genetics*, 141(3) :1199–1207. (Cité en page 10.)

David, J. (1992). Approche méthodologique d'une gestion dynamique des ressources génétiques chez le blé tendre (triticum aestivum l). (Cité en pages 5, 6 et iii.)

Dawson, J. C., Rivière, P., Berthellot, J.-F., Mercier, F., Kochko, P. d., Galic, N., Pin, S., Serpolay, E., Thomas, M., and Giuliano, S. (2011). Collaborative plant breeding for organic agricultural systems in developed countries. *Sustainability*, 3(8) :1206–1223. (Cité en page 4.)

de Bakker, P. I. W., Yelensky, R., Pe'er, I., Gabriel, S. B., Daly, M. J., and Altshuler, D. (2005). Efficiency and power in genetic association studies. *Nature Genetics*, 37(11) :1217–1223. (Cité en page 161.)

De Koeyer, D. L., Phillips, R. L., and Stuthman, D. D. (2001). Allelic shifts and quantitative trait loci in a recurrent selection population of oat. *Crop Science*, 41(4) :1228. (Cité en page 13.)

Dellaporta, S. L., Wood, J., and Hicks, J. B. (1983). A plant DNA minipreparation : version II. *Plant molecular biology reporter*, 1(4) :19–21. (Cité en page 31.)

Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometrics*, 55(4) :997–1004. (Cité en page 144.)

Driscoll, C. (1975). Cytogenetic analysis of two chromosomal male-sterility mutants in hexaploid wheat. *Australian Journal of Biological Sciences*, 28(4) :413–416. (Cité en page 70.)

Dubcovsky, J. and Dvorak, J. (2007). Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science*, 316(5833) :1862–1866. PMID : 17600208. (Cité en page 14.)

Dubcovsky, J., Loukoianov, A., Fu, D., Valarik, M., Sanchez, A., and Yan, L. (2006). Effect of photoperiod on the regulation of wheat vernalization genes VRN1 and VRN2. *Plant molecular biology*, 60(4) :469–480. (Cité en page 17.)

Dvorak, J. and Akhunov, E. D. (2005). Tempos of gene locus deletions and duplications and their relationship to recombination rate during diploid and polyploid evolution in the aegilops-triticum alliance. *Genetics*, 171(1) :323–332. (Cité en page 14.)

Ehrenreich, I. M., Hanzawa, Y., Chou, L., Roe, J. L., Kover, P. X., and Purugganan, M. D. (2009). Candidate gene association mapping of arabidopsis flowering time. *Genetics*, 183(1) :325–335. (Cité en page 162.)

Elias, M., McKey, D., Panaud, O., Anstett, M., and Robert, T. (2001). Traditional management of cassava morphological and genetic diversity by the makushi amerindians (guyana, south america) : Perspectives for on-farm conservation of crop genetic resources. *Euphytica*, 120(1) :143–157. (Cité en page 4.)

Enjalbert, J., Dawson, J. C., Paillard, S., Rhoné, B., Rousselle, Y., Thomas, M., and Goldringer, I. (2011). Dynamic management of crop diversity : From an experimental approach to on-farm conservation. *Comptes Rendus Biologies*, 334(5–6) :458–468. (Cité en pages 4 et 7.)

Ersoz, E. S., Yu, J., and Buckler, E. S. (2007). Applications of linkage disequilibrium and association mapping in crop plants. In Varshney, R. K. and Tuberosa, R., editors, *Genomics-Assisted Crop Improvement*, pages 97–119. Springer Netherlands. (Cité en page 183.)

Esquinas-Alcazar, J. (2005). Protecting crop genetic diversity for food security : political, ethical and technical challenges. *Nat Rev Genet*, 6(12) :946–953. (Cité en page 3.)

Evans, L. T. (1998). *Feeding the ten billion : plants and population growth*. Cambridge University Press. (Cité en page 1.)

FAO (2010). The second report on the state of the world's plant genetic resources for food and agriculture. Technical report. (Cité en page 3.)

FAO (2013). FAO statistical yearbook 2013 world food and agriculture. Technical report. (Cité en page 1.)

Ficher, R. A. and Maurer, R. (1978). Drought resistance in spring wheat cultivars. i. grain yield response. *Aust. J. Agric. Res*, 29 :897–912. (Cité en page 15.)

Finger, R. (2010). Evidence of slowing yield growth–the example of swiss cereal yields. *Food Policy*, 35(2) :175–182. (Cité en page 1.)

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford. (Cité en page 7.)

Flint, J. and Mackay, T. F. (2009). Genetic architecture of quantitative traits in mice, flies, and humans. *Genome research*, 19(5) :723–733. (Cité en page 9.)

Foolad, M. R., Stoltz, T., Dervinis, C., Rodriguez, R. L., and Jones, R. A. (1997). Mapping QTLs conferring salt tolerance during germination in tomato by selective genotyping. *Molecular Breeding*, 3(4) :269–277. (Cité en page 13.)

Fossati, A. and Ingold, M. (1970). A male sterile mutant in triticum aestivum. *Wheat Information Service*, (30) :8–10. (Cité en page 25.)

Frankel, O. H. (1995). *The conservation of plant biodiversity*. Cambridge University Press. (Cité en pages 2 et 3.)

Frankham, R. (1995). Effective population size/adult population size ratios in wildlife : a review. *Genetical Research*, 66(02) :95–107. (Cité en page 7.)

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3) :432–441. (Cité en page 151.)

Fu, D., Szűcs, P., Yan, L., Helguera, M., Skinner, J. S., von Zitzewitz, J., Hayes, P. M., and Dubcovsky, J. (2005). Large deletions within the first intron in VRN-1 are associated with spring growth habit in barley and wheat. *Molecular Genetics and Genomics*, 273(1) :54–65. (Cité en pages 16, 32 et 71.)

Gabriel, D., Sait, S. M., Kunin, W. E., and Benton, T. G. (2013). Food production vs. biodiversity : comparing organic and conventional agriculture. *Journal of Applied Ecology*, 50(2) :355–364. (Cité en page 2.)

Ganal, M. W., Durstewitz, G., Polley, A., Bérard, A., Buckler, E. S., Charcosset, A., Clarke, J. D., Graner, E.-M., Hansen, M., Joets, J., Le Paslier, M.-C., McMullen, M. D., Montalent, P., Rose, M., Schön, C.-C., Sun, Q., Walter, H., Martin, O. C., and Falque, M. (2011). A large maize (zea mays l.) SNP genotyping array : Development and germplasm genotyping, and genetic mapping to compare with the b73 reference genome. *PLoS ONE*, 6(12) :e28334. (Cité en page 40.)

Gawroński, P. and Schnurbusch, T. (2012). High-density mapping of the earliness per se-3Am (eps-3A m) locus in diploid einkorn wheat and its relation to the syntenic regions in rice and brachypodium distachyon l. *Molecular Breeding*, 30(2) :1097–1108. (Cité en page 17.)

Gepts, P. (2006). Plant genetic resources conservation and utilization. *Crop Science*, 46(5) :2278–2292. (Cité en page 3.)

Ghiglione, H. O., Gonzalez, F. G., Serrago, R., Maldonado, S. B., Chilcott, C., Curá, J. A., Miralles, D. J., Zhu, T., and Casal, J. J. (2008). Autophagy regulated by day length determines the number of fertile florets in wheat. *The Plant Journal*, 55(6) :1010–1024. (Cité en page xi.)

Gill, B. S., Appels, R., Botha-Oberholster, A.-M., Buell, C. R., Bennetzen, J. L., Chalhoub, B., Chumley, F., Dvorak, J., Iwanaga, M., Keller, B., Li, W., McCombie, W. R., Ogihara, Y., Quetier, F., and Sasaki, T. (2004). A workshop report on wheat genome sequencing : International genome research on wheat consortium. *Genetics*, 168(2) :1087–1096. (Cité en pages 14 et 18.)

Godfray, H. C. J., Beddington, J. R., Crute, I. R., Haddad, L., Lawrence, D., Muir, J. F., Pretty, J., Robinson, S., Thomas, S. M., and Toulmin, C. (2010). Food security : The challenge of feeding 9 billion people. *Science*, 327(5967) :812–818. PMID : 20110467. (Cité en page 1.)

Goldringer, I., Enjalbert, J., Raquin, A.-L., and Brabant, P. (2001). Strong selection in wheat populations during ten generations of dynamic management. *Genetics selection evolution*, 33 :S441–S463. (Cité en pages 4, 7 et 13.)

Goldringer, I., Prouin, C., Rousset, M., Galic, N., and Bonnin, I. (2006). Rapid differentiation of experimental populations of wheat for heading time in response to local climatic conditions. *Annals of Botany*, 98(4) :805–817. (Cité en pages 7 et 180.)

González, F. G., Slafer, G. A., and Miralles, D. J. (2003). Grain and floret number in response to photoperiod during stem elongation in fully and slightly vernalized wheats. *Field Crops Research*, 81(1) :17–27. (Cité en pages 28 et xi.)

Gouache, D., Le Bris, X., Bogard, M., Deudon, O., Pagé, C., and Gate, P. (2012). Evaluating agronomic adaptation options to increasing heat stress under climate change during wheat grain filling in france. *European Journal of Agronomy*, 39 :62–70. (Cité en page 15.)

Guedj, M. (2007). *Méthodes statistiques pour l'analyse des données génétiques d'association à grande échelle*. PhD thesis, Evry val d'essonne, Evry. (Cité en page 150.)

Gunderson, L. H. (2000). Ecological resilience–in theory and application. *Annual review of ecology and systematics*, page 425–439. (Cité en page 2.)

Guo, Z., Song, Y., Zhou, R., Ren, Z., and Jia, J. (2010). Discovery, evaluation and distribution of haplotypes of the wheat "Ppd-D1" gene. *New Phytologist*, 185(3) :841–851. (Cité en pages 18, 19 et 23.)

Haldane, J. B. S. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. *J Genet*, 8(29) :299–309. (Cité en page 36.)

Haldane, J. B. S. (1924). A mathematical theory of natural and artificial selection. part II. the influence of partial self-fertilization, inbreeding, assortative mating, and selective fertilization on the composition of mendelian populations, and on natural selection. In *Proc. Cambridge Philos. Soc*, volume 1, page 158–163. (Cité en page 8.)

Hamrick, J. L. and Godt, M. J. W. (1996). Effects of life history traits on genetic diversity in plant species. *Philosophical Transactions of the Royal Society of London. Series B : Biological Sciences*, 351(1345) :1291–1298. (Cité en page 8.)

Hanocq, E., Laperche, A., Jaminon, O., Lainé, A. L., and Le Gouis, J. (2007). Most significant genome regions involved in the control of earliness traits in bread wheat, as revealed by QTL meta-analysis. *Theoretical and Applied Genetics*, 114(3) :569–584. (Cité en page 164.)

Hao, K., Chudin, E., McElwee, J., and Schadt, E. E. (2009). Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies. *BMC Genetics*, 10(1) :27. (Cité en page 146.)

Haseman, J. K. and Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics*, 2(1) :3–19. (Cité en page 10.)

Heffner, E., Sorrells, J., and Mark, E. (2011). Genomic selection accuracy using multifamily prediction models in a wheat breeding program. *The Plant Genome*, 4(1) :65. (Cité en page 163.)

Henry, J. P., Pontis, C., David, J., Gouyon, P. H., Seitz, A., and Loeschcke, V. (1991). An experiment on dynamic conservation of genetic resources with metapopulations. *Species conservation : a population-biological approach.*, page 185–198. (Cité en page 4.)

Herndl, M., White, J. W., Hunt, L. A., Graeff, S., and Claupein, W. (2008). Field-based evaluation of vernalization requirement, photoperiod response and earliness per se in bread wheat (triticum aestivum l.). *Field Crops Research*, 105 :193–201. (Cité en pages 18 et 19.)

Higgins, J. A., Bailey, P. C., and Laurie, D. A. (2010). Comparative genomics of flowering time pathways using brachypodium distachyon as a model for the temperate grasses. *PLoS ONE*, 5(4) :e10065. (Cité en page 31.)

Hill, W. and Weir, B. (1988). Variances and covariances of squared linkage disequilibria in finite populations. *Theoretical Population Biology*, (33) :54–78. (Cité en pages 36 et 37.)

Hill, W. G. and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, 38(6) :226–231. (Cité en page 36.)

Hill, W. G. and Weir, B. S. (1994). Maximum-likelihood estimation of gene location by linkage disequilibrium. *American journal of human genetics*, 54(4) :705. (Cité en pages 36 et 37.)

Holmes, M. G. and Smith, H. (1977). The function of phytochrome in the natural environment—I. characterization of daylight for studies in photomorphogenesis and photoperiodism. *Photochemistry and Photobiology*, 25(6) :533–538. (Cité en page xi.)

Huang, B. E., Clifford, D., and Cavanagh, C. (2012a). Selecting subsets of genotyped experimental populations for phenotyping to maximize genetic diversity. *TAG Theoretical and Applied Genetics*, pages 1–10. (Cité en page 185.)

Huang, B. E., George, A. W., Forrest, K. L., Kilian, A., Hayden, M. J., Morell, M. K., and Cavanagh, C. R. (2012b). A multiparent advanced generation inter-cross population for genetic analysis in wheat. *Plant Biotechnology Journal*, 10(7) :826–839. (Cité en page 164.)

Ibrahim, K. and Barrett, J. (1991). Evolution of mildew resistance in a hybrid bulk population of barley. *Heredity*, 67(2) :247–256. (Cité en pages 7 et 180.)

Ingvarsson, P. K. and Street, N. R. (2011). Association genetics of complex traits in plants. *New Phytologist*, 189(4) :909–922. (Cité en pages 142 et 177.)

Iqbal, M., Shahzad, A., and Ahmed, I. (2011). Allelic variation at the vrn-a1, vrn-b1, vrn-d1, vrn-b3 and ppd-d1a loci of pakistani spring wheat cultivars. *Electronic Journal of Biotechnology*, 14(1). (Cité en pages 18 et 19.)

Iwaki, K., Haruna, S., Niwa, T., and Kato, K. (2001). Adaptation and ecological differentiation in wheat with special reference to geographical variation of growth habit and vrn genotype. *Plant Breeding*, 120(2) :107–114. (Cité en pages 18, 19 et 22.)

Iwaki, K., Nakagawa, K., Kuno, H., and Kato, K. (2000). Ecogeographical differentiation in east asian wheat, revealed from the geographical variation of growth habit and vrn genotype. *Euphytica*, 111(2) :137–143. (Cité en pages 18, 19 et 22.)

Jiang, C. and Zeng, Z.-B. (1995). Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics*, 140(3) :1111–1127. (Cité en page 143.)

Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant analysis of principal components : a new method for the analysis of genetically structured populations. *BMC genetics*, 11(1) :94. (Cité en pages 11 et 144.)

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, 178(3) :1709–1723. (Cité en pages 142 et 144.)

Kato, K., Miura, H., and Sawada, S. (2000). Mapping QTLs controlling grain yield and its components on chromosome 5A of wheat. *TAG Theoretical and Applied Genetics*, 101(7) :1114–1121. (Cité en page 15.)

Kato, K. and Yokoyama, H. (1992). Geographical variation in heading characters among wheat landraces, triticum estivum l., and its implication for their adaptability. *Theoretical and Applied Genetics*, 84(3-4) :259–265. (Cité en page 18.)

King, E. G., Macdonald, S. J., and Long, A. D. (2012). Properties and power of the drosophila synthetic population resource for the routine dissection of complex traits. *Genetics*, 191(3) :935–949. (Cité en pages 43 et 63.)

Kippes, N., Zhu, J., Chen, A., Vanzetti, L., Lukaszewski, A., Nishida, H., Kato, K., Dvorak, J., and Dubcovsky, J. (2013). Fine mapping and epistatic interactions of the vernalization gene VRN-D4 in hexaploid wheat. *Molecular Genetics and Genomics*, page 1–16. (Cité en page 16.)

Kolev, S., Ganeva, G., Christov, N., Belchev, I., Kostov, K., Tsenov, N., Rachovska, G., Landgeva, S., Ivanov, M., Abu-Mhadi, N., and Todorovska, E. (2010). Allele variation in loci for adaptive response and plant height and its effect on grain yield in wheat. *agriculture and environmental biotechnology*, 24(2) :1807–1813. (Cité en pages 18, 19 et 23.)

Korte, A., Vilhjálmsson, B. J., Segura, V., Platt, A., Long, Q., and Nordborg, M. (2012). A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature Genetics*, 44(9) :1066–1071. (Cité en page 163.)

Kosambi, D. D. (1943). The estimation of map distances from recombination values. *Annals of Eugenics*, 12(1) :172–175. (Cité en page 36.)

Kover, P. X., Valdar, W., Trakalo, J., Scarcelli, N., Ehrenreich, I. M., Purugganan, M. D., Durrant, C., and Mott, R. (2009). A multiparent advanced generation inter-cross to fine-map quantitative traits in arabidopsis thaliana. *PLoS Genet*, 5(7) :e1000551. (Cité en page 13.)

Ladha, J. K., Dawe, D., Pathak, H., Padre, A. T., Yadav, R. L., Singh, B., Singh, Y., Singh, Y., Singh, P., and Kundu, A. L. (2003). How extensive are yield declines in long-term rice–wheat experiments in asia ? *Field Crops Research*, 81(2) :159–180. (Cité en page 1.)

Lander, E. S., Green, P., Abrahamson, J., Barlow, A., Daly, M. J., Lincoln, S. E., and Newburg, L. (1987). MAPMAKER : an interactive computer package

for constructing primary genetic linkage maps of experimental and natural populations. *Genomics*, 1(2) :174–181. (Cité en pages 40 et 59.)

Lavigne, C., REBOUD, X., Lefranc, M., Porcher, E., Roux, F., Olivieri, I., and GODELLE, B. (2001). Evolution of genetic diversity in metapopulations : Arabidopsis thaliana as an experimental model. *Genetics selection evolution*, 33 :S399–S423. (Cité en pages 4, 5, 6 et 7.)

Law, C. and Worland, A. J. (1997). Genetic analysis of some flowering time and adaptive traits in wheat. *New Phytologist*, 137 :19–28. (Cité en pages 16 et 22.)

Law, C. N., Worland, A. J., and Giorgi, B. (1976). The genetic control of ear-emergence time by chromosomes 5A and 5D of wheat. *Heredity*, 36(1) :49–58. (Cité en page 16.)

Le Boulc'h, V. (1994). *Evolution de la résistance à l'oïdium (Erysiphe graminis f. sp. Tritici) dans des populations composites de blé tendre (Triticum aestivum L.) menées en gestion dynamique*. PhD thesis. (Cité en pages 25, iii et iv.)

Le Corre, V. and Kremer, A. (2012). The genetic differentiation at quantitative trait loci under local adaptation. *Molecular Ecology*, 21(7) :1548–1566. (Cité en page 71.)

Le Gouis, J., Bordes, J., Ravel, C., Heumez, E., Faure, S., Praud, S., Galic, N., Remoué, C., Balfourier, F., Allard, V., and Rousset, M. (2011). Genome-wide association analysis to identify chromosomal regions determining components of earliness in wheat. *Theoretical and Applied Genetics*. (Cité en pages 71, 150, 159, 163, 164, 172 et 180.)

Li, W., Zhang, P., Fellers, J. P., Friebe, B., and Gill, B. S. (2004). Sequence composition, organization, and evolution of the core triticeae genome. *The Plant Journal*, 40(4) :500–511. (Cité en page 14.)

Li, X., Yan, W., Agrama, H., Jia, L., Shen, X., Jackson, A., Moldenhauer, K., Yeater, K., McClung, A., and Wu, D. (2011). Mapping QTLs for improving grain yield using the USDA rice mini-core collection. *Planta*, 234(2) :347–361. (Cité en page 9.)

Lin, H. X., Yamamoto, T., Sasaki, T., and Yano, M. (2000). Characterization and detection of epistatic interactions of 3 QTLs, hd1, hd2, and hd3, controlling heading date in rice using nearly isogenic lines. *Theoretical and Applied Genetics*, 101(7) :1021–1028. (Cité en page 143.)

Lobell, D. B., Schlenker, W., and Costa-Roberts, J. (2011). Climate trends and global crop production since 1980. *Science*, 333(6042) :616–620. (Cité en page 15.)

Lorenzana, R. E. and Bernardo, R. (2009). Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theoretical and Applied Genetics*, 120(1) :151–161. (Cité en page 163.)

Louette, D. (2000). Chapter 5. traditional management of seed and genetic diversity : what is a landrace ? (Cité en page 4.)

Mackay, I. and Powell, W. (2007). Methods for linkage disequilibrium mapping in crops. *Trends in Plant Science*, 12(2) :57–63. (Cité en page 37.)

Mackay, T. F. C., Stone, E. A., and Ayroles, J. F. (2009). The genetics of quantitative traits : challenges and prospects. *Nature Reviews Genetics*, 10(8) :565–577. (Cité en pages 9 et 12.)

Mackinnon, M. J. and Georges, M. A. (1992). The effects of selection on linkage analysis for quantitative traits. *Genetics*, 132(4) :1177–1185. PMID : 1459434. (Cité en page 185.)

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2013). cluster : Cluster analysis basics and extensions. (Cité en page 40.)

Malosetti, M., Ribaut, J. M., Vargas, M., Crossa, J., and Eeuwijk, F. A. v. (2008). A multi-trait multi-environment QTL mixed model with an application to drought and nitrogen stress trials in maize (zea mays l.). *Euphytica*, 161(1-2) :241–257. (Cité en pages 143 et 181.)

Mangin, B., Siberchicot, A., Nicolas, S., Doligez, A., This, P., and Cierco-Ayrolles, C. (2012). Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity*, 108(3) :285–291. (Cité en pages 55 et 179.)

Mangin, B., Thoquet, P., and Grimsley, N. (1998). Pleiotropic QTL analysis. *Biometrics*, page 88–99. (Cité en pages 143 et 163.)

Maniatis, N., Collins, A., Gibson, J., Zhang, W., Tapper, W., and Morton, N. E. (2004). Positional cloning by linkage disequilibrium. *The American Journal of Human Genetics*, 74(5) :846–855. (Cité en page 60.)

Maniatis, N., Collins, A., Xu, C.-F., McCarthy, L. C., Hewett, D. R., Tapper, W., Ennis, S., Ke, X., and Morton, N. E. (2002). The first linkage disequilibrium (LD) maps : Delineation of hot and cold blocks by diplotype analysis. *Proceedings of the National Academy of Sciences*, 99(4) :2228–2233. (Cité en page 38.)

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F. C., McCarroll, S. A., and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265) :747–753. (Cité en page 9.)

Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7) :499–511. (Cité en page 146.)

Mathews, K. L., Malosetti, M., Chapman, S., McIntyre, L., Reynolds, M., Shorter, R., and van Eeuwijk, F. (2008). Multi-environment QTL mixed models for drought stress adaptation in wheat. *Theoretical and Applied Genetics*, 117(7) :1077–1091. (Cité en page 163.)

Matsuoka, Y., Takumi, S., and Kawahara, T. (2008). Flowering time diversification and dispersal in central eurasian wild wheat aegilops tauschii coss. : Genealogical and ecological framework. *PLoS ONE*, 3(9) :e3138. (Cité en page 18.)

Maxted, N., Ford-Lloyd, B. V., and Hawkes, J. G. (1997). *Plant genetic conservation : the in situ approach*. Springer. (Cité en page 3.)

Maxted, N., Guarino, L., Myer, L., and Chiwona, E. A. (2002). Towards a methodology for on-farm conservation of plant genetic resources. *Genetic Resources and Crop Evolution*, 49(1) :31–46. (Cité en page 4.)

McIntosh, R. A. (1988). Catalogue of gene symbols for wheat. In *Proceeding of the Seventh International Wheat Genetics Symposium*, volume 2, Cambridge, UK. (Cité en page 25.)

Meuwissen, T. H. E. and Goddard, M. E. (2004). Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. *Genet. Sel. Evol*, 36 :261–279. (Cité en page 183.)

Meuwissen, T. H. E., Karlsen, A., Lien, S., Olsaker, I., and Goddard, M. E. (2002). Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. *Genetics*, 161(1) :373–379. (Cité en page 13.)

Michael, T. P. and Jackson, S. (2013). The first 50 plant genomes. *The Plant Genome*, 6(2). (Cité en page 14.)

Michaels, S. D., John, M. C., and Amasino, R. M. (1994). Removal of polysaccharides from plant DNA by ethanol precipitation. *Biotechniques*, 17(2) :274–276. (Cité en page 31.)

Michéli, E., Schad, P., Spaargaren, O., Dent, D., and Nachtergaele, F. (2006). *World reference base for soil resources : 2006 : a framework for international classification, correlation and communication*. FAO. (Cité en page 1.)

Moiseeva, E. and Goncharov, N. P. (2007). Genetic control of the spring growth habit in old local cultivars and landraces of common wheat from siberia. *Russian journal of genetics*, 43(4) :369–375. (Cité en pages 18 et 19.)

Monfreda, C., Ramankutty, N., and Foley, J. A. (2008). Farming the planet : 2. geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000. *Global Biogeochemical Cycles*, 22(1). (Cité en page 14.)

Montes, J. M., Melchinger, A. E., and Reif, J. C. (2007). Novel throughput phenotyping platforms in plant genetic studies. *Trends in Plant Science*, 12(10) :433–436. (Cité en page 177.)

Morton, N. E., Zhang, W., Taillon-Miller, P., Ennis, S., Kwok, P.-Y., and Collins, A. (2001). The optimal measure of allelic association. *Proceedings of the National Academy of Sciences*, 98(9) :5217–5221. (Cité en page 38.)

Musani, S. K., Zhang, H.-G., Hsu, H.-C., Yi, N., S. Gorman, B., B. Allison, D., and D. Mountz, J. (2006). Principal component analysis of quantitative trait loci for immune response to adenovirus in mice. *Hereditas*, 143(2006) :189–197. (Cité en page 143.)

Nakasako, M., Wada, M., Tokutomi, S., Yamamoto, K. T., Sakai, J., Kataoka, M., Tokunaga, F., and Furuya, M. (1990). Quaternary structure of pea phytochrome i dimer studied with small-angle x-ray scattering and rotary-shadowing electron microscopy. *Photochemistry and Photobiology*, 52(1) :3–12. (Cité en page xi.)

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences*, 70(12) :3321–3323. PMID : 4519626. (Cité en page 6.)

Nei, M. and Tajima, F. (1981). Genetic drift and estimation of effective population size. *Genetics*, 98(3) :625–640. (Cité en page 7.)

Nordborg, M. and Weigel, D. (2008). Next-generation genetics in plants. *Nature*, 456(7223) :720–723. (Cité en page 142.)

Olivieri, I., Couvet, D., and Gouyon, P. (1990). The genetics of transient populations : Research at the metapopulation level. *Trends in Ecology & Evolution*, 5(7) :207–210. (Cité en page 4.)

Paillard, S., Goldringer, I., Enjalbert, J., Doussinault, G., Vallavieille-Pope, C. d., and Brabant, P. (2000a). Evolution of resistance against powdery mildew in winter wheat populations conducted under dynamic management. i – is specific seedling resistance selected ? *Theoretical and Applied Genetics*, 101(3) :449–456. (Cité en pages 6, 7 et 180.)

Paillard, S., Goldringer, I., Enjalbert, J., Trottet, M., David, J., de Vallavieille-Pope, C., and Brabant, P. (2000b). Evolution of resistance against powdery mildew in winter wheat populations conducted under dynamic management. II. adult plant resistance. *Theoretical and Applied Genetics*, 101(3) :457–462. (Cité en pages 7 et 180.)

Parker, J. (2011). *The 9 Billion-People Question : A Special Report on Feeding the World*. Economist Newspaper. (Cité en page 1.)

Parzies, H. K., Spoor, W., and Ennos, R. A. (2000). Genetic diversity of barley landrace accessions (hordeum vulgare ssp. vulgare) conserved for different lengths of time in ex situ gene banks. *Heredity*, 84(4) :476–486. (Cité en page 3.)

Paux, E., Sourdille, P., Salse, J., Saintenac, C., Choulet, F., Leroy, P., Korol, A., Michalak, M., Kianian, S., Spielmeyer, W., Lagudah, E., Somers, D., Kilian, A., Alaux, M., Vautrin, S., Berges, H., Eversole, K., Appels, R., Safar, J., Simkova, H., Dolezel, J., Bernard, M., and Feuillet, C. (2008). A physical map of the 1-gigabase bread wheat chromosome 3B. *Science*, 322(5898) :101–104. (Cité en pages 46 et 61.)

Plucknett, D. L., Smith, N. J., Williams, J. T., and Anishetty, N. (1987). *Gene banks and the world's food*. Princeton University Press. (Cité en page 3.)

Pontis, C. (1992). *Utilisation de marqueurs génétiques pour le suivi de la variabilité de 3 composites de blé tendre d'hiver (Triticum aestivum L.) menées en gestion dynamique*. PhD thesis. (Cité en page iii.)

Porcher, E., Giraud, T., Goldringer, I., and Lavigne, C. (2004). Experimental demonstration of a causal relationship between heterogeneity of selection and genetic differentiation in quantitative traits. *Evolution*, 58(7) :1434–1445. (Cité en pages 6 et 13.)

Potts, S. G., Biesmeijer, J. C., Kremen, C., Neumann, P., Schweiger, O., and Kunin, W. E. (2010). Global pollinator declines : trends, impacts and drivers. *Trends in ecology & evolution*, 25(6) :345–353. (Cité en page 2.)

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2) :945. (Cité en pages 11 et 144.)

R Development Core Team (2012). R : A language and environment for statistical computing. r foundation for statistical computing. (Cité en pages 147 et 149.)

Rafalski, J. A. (2002). Novel genetic mapping tools in plants : SNPs and LD-based approaches. *Plant Science*, 162(3) :329–333. (Cité en page 142.)

Rao, M. K., Devi, K. U., and Arundhati, A. (1990). Applications of genie male sterility in plant breeding. *Plant Breeding*, 105(1) :1–25. (Cité en pages 8 et 184.)

Raquin, A.-L., Brabant, P., Rhoné, B., Balfourier, F., Leroy, P., and Goldringer, I. (2008a). Soft selective sweep near a gene that increases plant height in wheat. *Molecular ecology*, 17(3) :741–756. (Cité en pages 6, 13 et 180.)

Raquin, A.-L., Depaulis, F., Lambert, A., Galic, N., Brabant, P., and Goldringer, I. (2008b). Experimental estimation of mutation rates in a wheat population with a gene genealogy approach. *Genetics*, 179(4) :2195–2211. (Cité en page 7.)

Ratet, J. (2011). Elaboration d'un dispositif d'éclairage pour le blé de l'expérience 2011/2012. (Cité en page 28.)

Rebai, A., Goffinet, B., and Mangin, B. (1995). Comparing power of different methods for QTL detection. *Biometrics*, page 87–99. (Cité en page 35.)

Reganold, J. P. (1988). Comparison of soil properties as influenced by organic and conventional farming systems. *American Journal of Alternative Agriculture*, 3(04) :144–155. (Cité en page 2.)

Rhoné, B. (2008). *Etude de mécanismes génétiques impliqués dans l'adaptation climatique de populations expérimentales de blé tendre*. Thèse de doctorat. (Cité en pages 32 et 185.)

Rhoné, B., Raquin, A.-L., and Goldringer, I. (2007). Strong linkage disequilibrium near the selected yr17 resistance gene in a wheat experimental population. *Theoretical and Applied Genetics*, 114(5) :787–802. (Cité en pages 13 et 180.)

Rhoné, B., Remoué, C., Galic, N., Goldringer, I., and Bonnin, I. (2008). Insight into the genetic bases of climatic adaptation in experimentally evolving wheat populations. *Molecular Ecology*, 17(3) :930–943. (Cité en pages 6, 7, 32 et 71.)

Rhoné, B., Vitalis, R., Goldringer, I., and Bonnin, I. (2010). Evolution of flowering time in experimental wheat populations : A comprehensive approach to detect genetic signatures of natural selection. *Evolution*, 64(7) :2110–2125. (Cité en pages 7, 71, 180 et 181.)

Rincent, R., Moreau, L., Monod, H., Kuhn, E., Melchinger, A., Malvar, R., Moreno-Gonzalez, J., Nicolas, S., Madur, D., Combes, V., Dumas, F., Altmann, T., Brunel, D., Ouzunova, M., Flament, P., Dubreuil, P., Charcosset, A., and Mary-Huard, T. (2014). Recovering power in association mapping panels with variable levels of linkage disequilibrium. *submitted*. (Cité en page 144.)

Rivière, P. (2014). *Méthodologie de la sélection décentralisée et participative : un exemple sur le blé tendre*. PhD thesis, Paris Sud. (Cité en page 4.)

Robertson, A. (1960). A theory of limits in artificial selection. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 153(951) :234–249. (Cité en page 7.)

Rohlfs, R. V., Swanson, W. J., and Weir, B. S. (2010). Detecting coevolution through allelic association between physically unlinked loci. *The American Journal of Human Genetics*, 86(5) :674–685. (Cité en page 181.)

Roussel, V., Koenig, J., Beckert, M., and Balfourier, F. (2004). Molecular diversity in french bread wheat accessions related to temporal trends and breeding programmes. *Theoretical and Applied Genetics*, 108(5) :920–930. (Cité en page 2.)

Roussel, V., Leisova, L., Exbrayat, F., Stehno, Z., and Balfourier, F. (2005). SSR allelic diversity changes in 480 european bread wheat varieties released from 1840 to 2000. *Theoretical and Applied Genetics*, 111(1) :162–170. (Cité en page 2.)

Rousset, M., Bonnin, I., Remoué, C., Falque, M., Rhoné, B., Veyrieras, J.-B., Madur, D., Murigneux, A., Balfourier, F., Gouis, J., Santoni, S., and Goldringer, I. (2011). Deciphering the genetics of flowering time by an association study on candidate genes in bread wheat (triticum aestivum l.). *Theoretical and Applied Genetics*. (Cité en pages 9, 11 et 142.)

Saintenac, C., Falque, M., Martin, O. C., Paux, E., Feuillet, C., and Sourdille, P. (2009). Detailed recombination studies along chromosome 3B provide new insights on crossover distribution in wheat (triticum aestivum l.). *Genetics*, 181(2) :393–403. (Cité en page 65.)

Sax, K. (1923). The association of size differences with seed-coat pattern and pigmentation in phaseolus vulgaris. *Genetics*, 8(6) :552. (Cité en page 35.)

Scutari, M. and Nagarajan, R. (2011). On identifying significant edges in graphical models. *arXiv preprint arXiv :1104.0896*. (Cité en pages 143 et 182.)

Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, [U+FFFD], Long, Q., and Nordborg, M. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genetics*, 44(7) :825–830. (Cité en pages 143, 147, 148, 161 et 181.)

Shaw, L. M., Turner, A. S., and Laurie, D. A. (2012). The impact of photoperiod insensitive ppd-1a mutations on the photoperiod pathway across the three genomes of hexaploid wheat (triticum aestivum). *The Plant Journal*, 71(1) :71–84. (Cité en page 17.)

Sherman, J. D., Yan, L., Talbert, L., and Dubcovsky, J. (2004). A PCR marker for growth habit in common wheat based on allelic variation at the VRN-A1 gene. *Crop Sci*, 44(5) :1832–1838. (Cité en page 32.)

Shiferaw, B., Smale, M., Braun, H.-J., Duveiller, E., Reynolds, M., and Muricho, G. (2013). Crops that feed the world 10. past successes and future challenges to the role played by wheat in global food security. *Food Security*, page 1–27. (Cité en pages 1 et 14.)

Shimada, S., Ogawa, T., Kitagawa, S., Suzuki, T., Ikari, C., Shitsukawa, N., Abe, T., Kawahigashi, H., Kikuchi, R., and Handa, H. (2009). A genetic network of flowering-time genes in wheat leaves, in which an APETALA1/FRUITFULL-like gene, VRN1, is upstream of FLOWERING LOCUS t. *The Plant Journal*, 58(4) :668–681. (Cité en page 17.)

Sineshchekov, V., Koppel, L., Shlumukov, L., Barro, F., Barcelo, P., Lazzeri, P., and Smith, H. (2001). Fluorescence and photochemical properties of

phytochromes in wild-type wheat and a transgenic line overexpressing an oat phytochrome a (PHYA) gene : functional implications. *Plant, Cell & Environment*, 24(12) :1289–1297. (Cité en page xi.)

Sinha, H., David, L., Pascon, R. C., Clauder-Münster, S., Krishnakumar, S., Nguyen, M., Shi, G., Dean, J., Davis, R. W., Oefner, P. J., McCusker, J. H., and Steinmetz, L. M. (2008). Sequential elimination of major-effect contributors identifies additional quantitative trait loci conditioning high-temperature growth in yeast. *Genetics*, 180(3) :1661–1670. PMID : 18780730. (Cité en page 184.)

Skøt, L., Humphreys, M. O., Armstead, I., Heywood, S., Skøt, K. P., Sanderson, R., Thomas, I. D., Chorlton, K. H., and Hamilton, N. R. S. (2005). An association mapping approach to identify flowering time genes in natural populations of lolium perenne (l.). *Molecular Breeding*, 15(3) :233–245. (Cité en page 66.)

Slafer, G. A. and Rawson, H. M. (1995). Intrinsic earliness and basic development rate assessed for their response to temperature in wheat. *Euphytica*, 83(3) :175–183. (Cité en page 17.)

Slate, J. (2005). Quantitative trait locus mapping in natural populations : progress, caveats and future directions. *Molecular Ecology*, 14(2) :363–379. (Cité en page 11.)

Smil, V. (2002). Eating meat : Evolution, patterns, and consequences. *Population and development review*, 28(4) :599–639. (Cité en page 1.)

Snape, J. W., Butterworth, K., Whitechurch, E., and Worland, A. J. (2001). Waiting for fine times : genetics of flowering time in wheat. *Euphytica*, 119(1) :185–190. (Cité en pages 16 et 17.)

Soengas, P., Cartea, E., Lema, M., and Velasco, P. (2009). Effect of regeneration procedures on the genetic integrity of brassica oleracea accessions. *Molecular Breeding*, 23(3) :389–395. (Cité en page 3.)

Stearns, T. M., Beever, J. E., Southey, B. R., Ellis, M., McKeith, F. K., and Rodriguez-Zas, S. L. (2005). Evaluation of approaches to detect quantitative trait loci for growth, carcass, and meat quality on swine chromosomes 2, 6, 13, and 18. II. multivariate and principal component analyses. *Journal of Animal Science*, 83(11) :2471–2481. PMID : 16230643. (Cité en pages 143 et 162.)

Stebbins Jr, C. L. (1950). Variation and evolution in plants. *Variation and evolution in plants*. (Cité en page 14.)

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16) :9440–9445. PMID : 12883005. (Cité en page 143.)

Su, Z., Hao, C., Wang, L., Dong, Y., and Zhang, X. (2011). Identification and development of a functional marker of TaGW2 associated with grain weight in bread wheat (triticum aestivum l.). *Theoretical and Applied Genetics*, 122 :211–223. (Cité en page 32.)

Sun, Q., Zhou, R., Gao, L., and Jia, J. (2009). The characterization and geographical distribution of the genes responsible for vernalization requirement in chinese bread wheat. *Journal of integrative plant biology*, 51(4) :423–432. (Cité en pages 18 et 19.)

Sved, J. (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology*, (2) :125–141. (Cité en pages 37 et 65.)

Tanksley, S. D. and McCouch, S. R. (1997). Seed banks and molecular maps : unlocking genetic potential from the wild. *Science*, 277(5329) :1063–1066. (Cité en page 8.)

Thabuis, A., Lefebvre, V., Bernard, G., Daubèze, A. M., Phaly, T., Pochard, E., and Palloix, A. (2004). Phenotypic and molecular evaluation of a recurrent selection program for a polygenic resistance to phytophthora capsici in pepper. *Theoretical and Applied Genetics*, 109(2) :342–351. (Cité en page 13.)

Thomas, M., Dawson, J. C., Goldringer, I., and Bonneuil, C. (2011). Seed exchanges, a key to analyze crop diversity dynamics in farmer-led on-farm conservation. *Genetic Resources and Crop Evolution*, 58(3) :321–338. (Cité en page 4.)

Thornsberry, J. M., Goodman, M. M., Doebley, J., Kresovich, S., Nielsen, D., and Buckler, E. S. (2001). Dwarf8 polymorphisms associate with variation in flowering time. *Nature genetics*, 28(3) :286–289. (Cité en page 142.)

Thuillet, A.-C., Bru, D., David, J., Roumet, P., Santoni, S., Sourdille, P., and Bataillon, T. (2002). Direct estimation of mutation rate for 10 microsatellite loci in durum wheat, triticum turgidum (l.) thell. ssp durum desf. *Molecular Biology and Evolution*, 19(1) :122–125. PMID : 11752198. (Cité en page 5.)

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, page 267–288. (Cité en page 151.)

Trethowan, R. M., Reynolds, M. P., Ortiz-Monasterio, J. I., and Ortiz, R. (2007). The genetic basis of the green revolution in wheat production. *Plant Breeding Reviews*, 28 :39. (Cité en page 23.)

Trottet, M. (1988). Use of genic male sterility for breeding wheat lines resistant to leptosphaeria nodorum muller : Results of a first cycle and prospect. In *Proceeding of the Seventh International Wheat Genetics Symposium*, page 1199–1202, Cambridge, UK. (Cité en pages 25 et iii.)

Valdar, W., Flint, J., and Mott, R. (2006). Simulating the collaborative cross : Power of quantitative trait loci detection and mapping resolution in large sets of recombinant inbred strains of mice. *Genetics*, 172(3) :1783–1797. (Cité en page 11.)

van Beem, J., Mohler, V., Lukman, R., van Ginkel, M., William, M., Crossa, J., and Worland, A. J. (2005). Analysis of genetic factors influencing the developmental rate of globally important CIMMYT wheat cultivars. *Crop Sci*, 45(5) :2113–2119. (Cité en pages 18 et 19.)

van Eeuwijk, F. A., Bink, M. C., Chenu, K., and Chapman, S. C. (2010). Detection and use of QTL for complex traits in multiple environments. *Current Opinion in Plant Biology*, 13(2) :193–205. (Cité en page 163.)

Wang, W. Y. S., Barratt, B. J., Clayton, D. G., and Todd, J. A. (2005). Genome-wide association studies : theoretical and practical concerns. *Nature Reviews Genetics*, 6(2) :109–118. (Cité en page 10.)

Warner, R. M. (2006). Supplemental lighting on bedding plants–Making it work for you. *OFA Bull*, (899). (Cité en page xi.)

Waterlow, J. C., Armstrong, D. G., Fowden, L., and Riley, R. (1998). *Feeding a world population of more than eight billion people : a challenge to science.* Oxford University Press. (Cité en page 1.)

Weir, A. H., Bragg, P. L., Porter, J. R., and Rayner, J. H. (1984). A winter wheat crop simulation model without water or nutrient limitations. *Journal of Agricultural Science*, 102(2) :371–382. (Cité en pages 121, 133, 134 et 146.)

Weir, B. S. and Hill, W. G. (1980). Effect of mating structure on variation in linkage disequilibrium. *Genetics*, 95(2) :477–488. PMID : 7203003. (Cité en page 65.)

Welsh, J., Keim, D., Pirasteh, B., and Richards, R. (1973). Genetic control of photoperiod response in wheat. In *Proceed- ings of the 4th International Wheat Genetics Symposium*, page 879– 884, University of Missouri, Columbia, Mo. E.R. Sears and L.M.S. Sears. Missouri Agricultural Experiment Station, University of Missouri, Columbia. (Cité en page 16.)

White, J. W., Herndl, M., Hunt, L. A., Payne, T. S., and Hoogenboom, G. (2008). Simulation-based analysis of effects of vrn and ppd loci on flowering in wheat. *Crop Sci*, 48(2) :678–687. (Cité en pages 18 et 19.)

Wilhelm, E., Turner, A., and Laurie, D. (2009). Photoperiod insensitive ppd-a1a mutations in tetraploid wheat (triticum durum desf.). *TAG Theoretical and Applied Genetics*, 118(2) :285–294. (Cité en page 16.)

Worland, A. J. (1996). The influence of flowering time genes on environmental adaptability in european wheats. *Euphytica*, 89(1) :49–57. (Cité en pages 15 et 22.)

Worland, A. J., Appendino, M. L., and Sayers, E. J. (1994). The distribution, in european winter wheats, of genes that influence ecoclimatic adaptability whilst determining photoperiodic insensitivity and plant height. *euphytica*, 80(3) :219–228. (Cité en pages 18 et 19.)

Wright, S. (1931). Evolution in mendelian populations. *Genetics*, 16(2) :97. (Cité en page 7.)

Wright, S. (1969). *Evolution and the genetics of populations : Vol. 2. The theory of gene frequencies*. (Cité en page 7.)

Yan, L. (2003). Positional cloning of the wheat vernalization gene VRN1. *Proceedings of the National Academy of Sciences*, 100(10) :6263–6268. (Cité en page 16.)

Yan, L., Fu, D., Li, C., Blechl, A., Tranquilli, G., Bonafede, M., Sanchez, A., Valarik, M., Yasuda, S., and Dubcovsky, J. (2006). The wheat and barley vernalization gene VRN3 is an orthologue of FT. *Proceedings of the National Academy of Sciences*, 103(51) :19581–19586. (Cité en page 16.)

Yan, L., Helguera, M., Kato, K., Fukuyama, S., Sherman, J., and Dubcovsky, J. (2004a). Allelic variation at the VRN-1 promoter region in polyploid wheat. *Theoretical and Applied Genetics*, 109(8) :1677–1686. (Cité en pages 16 et 32.)

Yan, L., Loukoianov, A., Blechl, A., Tranquilli, G., Ramakrishna, W., SanMiguel, P., Bennetzen, J. L., Echenique, V., and Dubcovsky, J. (2004b). The wheat VRN2 gene is a flowering repressor down-regulated by vernalization. *Science*, 303(5664) :1640–1644. (Cité en pages 16 et 71.)

Yang, F. P., Zhang, X. K., Xia, X. C., Laurie, D. A., Yang, W. X., and He, Z. H. (2009). Distribution of the photoperiod insensitive ppd-d1a allele in chinese wheat cultivars. *Euphytica*, 165(3) :445–452. (Cité en pages 18, 19 et 23.)

Yang, Y., Foulquié-Moreno, M. R., Clement, L., Erdei, [U+FFFD], Tanghe, A., Schaerlaekens, K., Dumortier, F., and Thevelein, J. M. (2013). QTL analysis of high thermotolerance with superior and downgraded parental yeast strains reveals new minor QTLs and converges on novel causative alleles involved in RNA processing. *PLoS genetics*, 9(8) :e1003693. (Cité en page 183.)

Yoshida, T., Nishida, H., Zhu, J., Nitcher, R., Distelfeld, A., Akashi, Y., Kato, K., and Dubcovsky, J. (2010). Vrn-d4 is a vernalization gene located on the centromeric region of chromosome 5D in hexaploid wheat. *Theoretical and Applied Genetics*, 120(3) :543–552. (Cité en page 16.)

Yu, J., Holland, J. B., McMullen, M. D., and Buckler, E. S. (2008). Genetic design and statistical power of nested association mapping in maize. *Genetics*, 178(1) :539–551. (Cité en pages 13 et 177.)

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., and Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2) :203–208. (Cité en pages 11, 13, 144 et 147.)

Zamir, D. (2001). Improving plant breeding with exotic genetic libraries. *Nature reviews genetics*, 2(12) :983–989. (Cité en page 8.)

Zhang, W., Collins, A., Maniatis, N., Tapper, W., and Morton, N. E. (2002). Properties of linkage disequilibrium (LD) maps. *Proceedings of the National Academy of Sciences*, 99(26) :17004–17007. (Cité en page 38.)

Zhang, X. K., Xiao, Y. G., Zhang, Y., Xia, X. C., Dubcovsky, J., and He, Z. H. (2008). Allelic variation at the vernalization genes vrn-a1, vrn-B1,Vrn-D1, vrn-b3 and in chinese wheat cultivars and their association with growth habit. *Crop Science*, 48(2) :458–470. (Cité en pages 18, 19 et 22.)

Zhao, K., Aranzana, M. J., Kim, S., Lister, C., Shindo, C., Tang, C., Toomajian, C., Zheng, H., Dean, C., and Marjoram, P. (2007). An arabidopsis example of association mapping in structured samples. *PLoS Genetics*, 3(1) :e4. (Cité en page 144.)

Zheng, B., Biddulph, B., Li, D., Kuchel, H., and Chapman, S. (2013). Quantification of the effects of VRN1 and ppd-d1 to predict spring wheat (triticum aestivum) heading time across diverse environments. *Journal of experimental botany*, 64(12) :3747–3761. (Cité en page 17.)

Chapitre 7

Annexes

Annexe A

Bilan sur la création de la population MAGIC INRA

Les premiers croisements de la création de la population MAGIC INRA ont été réalisés en 1976 par Maxime Trottet à l'INRA de Rennes. A cette époque, cette population était nommée "pool PS". Le premier document officiel retrouvé est le recueil de résumés de l'"international wheat genetics symposium" (IWGS) de Cambridge en 1988 [Trottet, 1988]. M. Trottet explique dans celui-ci que 50 lignées ont été choisies pour leur résistance à *Septoria nodorum* ou pour leur courte paille et leur bonne valeur agronomique. Ces lignées ont été croisées avec Probus (mâle stérile). Un recroisement avec les parents a suivi pour réduire la proportion de Probus dans la population. Les générations suivantes ont été isolées et des croisements aléatoires ont eu lieu avec récolte uniquement des épis mâles stériles.

En complément, trois thèses soutenues en 1992 et 1994 au Moulon évoquent la création de cette population. C. Pontis [Pontis, 1992] liste 54 parents de cette population, Probus exclu. La même année J. David [David, 1992] a écrit que 50 lignées ont été croisées avec Probus. Les F1 ont ensuite été semés en lignes entourées d'une ligne de mélange des 49 autres F1 et de quelques parents. La population PS aurait été obtenue par mélange équilibré des descendants des mâles stériles issus de chaque F2. En 1994, V. Le Boul'ch [Le Boulc'h, 1994] a écrit que la population PS est issue du croisement de 60 lignées avec Probus. Après autofécondation des plantes F1, un croisement entre les F2 aurait été réalisé. Une nouvelle autofécondation a été faite sur les descendants des mâles stériles et la population serait le mélange équilibré des grains issus de mâles stériles après croisement des F2.

Suite à la lecture de ces informations discordantes, une recherche dans les archives s'imposait. Peu de documents explicites ont été retrouvés : une lettre de M. Trottet datant de 1983 puis plusieurs listes de lignées difficilement identifiables comme liste de parents d'origine ou liste de lignées utilisées en rétrocroisement.
Dans cette lettre, M. Trottet décrit la population comme :

> "le croisement des plantes ms de Probus avec 50 géniteurs[...]. Une partie des plantes F1 fertiles a été recroisée par ces mêmes géniteurs. Les 50 F2 ont été semées en isolement, chaque F2 entourée du mélange de toutes les F2 et après la récolte on a reconstitué un mélange en équilibrant l'apport de chacune des F2 donc de chacun des géniteurs."

Le tableau A.1 récapitule les listes trouvées dans les thèses et dans les archives. Aucune conclusion n'a pu être émise sur la contribution réelle de chacun de ces génotypes, même après discussion avec M. Trottet. Notre liste de départ est celle de

60 parents (Tableau 2.1) publiée dans la thèse de V. Le Boul'ch [Le Boulc'h, 1994]. Les différents textes officiels ainsi que la lettre de M. Trottet nous ont amené à deux hypothèses de schéma de croisement représentés dans la figure A.1.

Tableau A.1 – Tableau récapitulatif des différentes listes trouvées de génotypes pouvant contribuer à la création de la population MAGIC INRA.

	Génotypes	Liste 1 : Croisements initiaux M. Trottet (1976?)	Liste 2 : second croisements? "F1 Septo": 1976	Listes 3 : ajout au crayon (ajout tardif de lignées?)	Bilan des trois listes
Liste de référence Pontis, 1992 & Trottet	3596-58		x		x
	ANDES 56	x	x	x	x
	ATLAS 66	x	x	x	x
	C 56-5	x	x		x
	CA68	x	x	x	x
	CARALA	x	x	x	x
	CARDENAL			x	x
	CHALK		x		x
	CLEMENT	x			x
	COMTAL	x	x		x
	CONDOR			x	x
	CORIN				
	COURTOT	x	x		x
	DARKAN				
	DOMUS	x	x		x
	DUCAT	x		x	x
	FERMO Mutant 9-14				
	GRANA				
	HADDEN	x	x		x
	HARUHIKARI	x	x		x
	IAS 20				
	IAS 63			x	x
	KAVKAZ	x	x		x
	L707			x	x
	LAFRON			x	x
	LAGOA VERMELHA				
	LAPIS		x		x
	LUTIN	x	x		x
	MARIS HOBBIT	x	x		x
	MARIS HUNTSMAN	x			x
	MARQUILLO			x	x
	MINISTRE NAIN			x	x
	MIRONOVSKAIA 808	x	x		x
	mutant du 81-12		x		x
	mutant of PROBUS	x			x
	NAUTICA				
	OASIS		x	x	x
	ORLANDO				
	OXLEY				
	R 5-1	x			x
	REDHART			x	x
	REDON M4	x	x		x
	SAPPO		x		x
	TALENT	x			x
	TJB 155 = KINSMAN				
	TJB 240 = Sportman		x		x
	TJB 251		x		x
	TJB 636		x	x	x
	TL 25-11		x		x
	TL 365 A34				
	TOROPI				
	US 113		x		x
	US 117		x		x
	US 123		x		x
	US 125		x		x

	V2D11		x		x
	V3D8		x		x
	VPM x Moisson4 - 3lines	4.1.2.2.11 4.1.2.3.3.11 1.5	x		x
	VPMM1-1-1-2 R4	x	x	x	x
	WEINSTEPHAN 1007-53			x	x
Nombre de génotypes:		23	33	17	

Individu supplémentaire Liste 1	Mutants de Bronnimann	x			x
	US60(43)xPrieur 61.5.3.2	R5-1			
	104-83	x			x
	Aurora	x			x
Nombre de génotypes:		4			

Individu supplémentaire Liste 2	RPB		x		x
	H75 HD 1				
	H75 HD 2				
	H75 HD 3				
	H75 HD 4				
Nombre de génotypes:			1		

Individu supplémentaire Liste 3	Maris Freeman			x	
	76HMH30			x	
	bonza63			x	
	lignée INRA 465			x	
	Tenmarq			x	
	Knox			x	
	Manon			x	
	Yacora			x	
	Cotipora			x	
	Ci13406			x	
Nombre de génotypes:				10	

Nombre total:		27	34		0

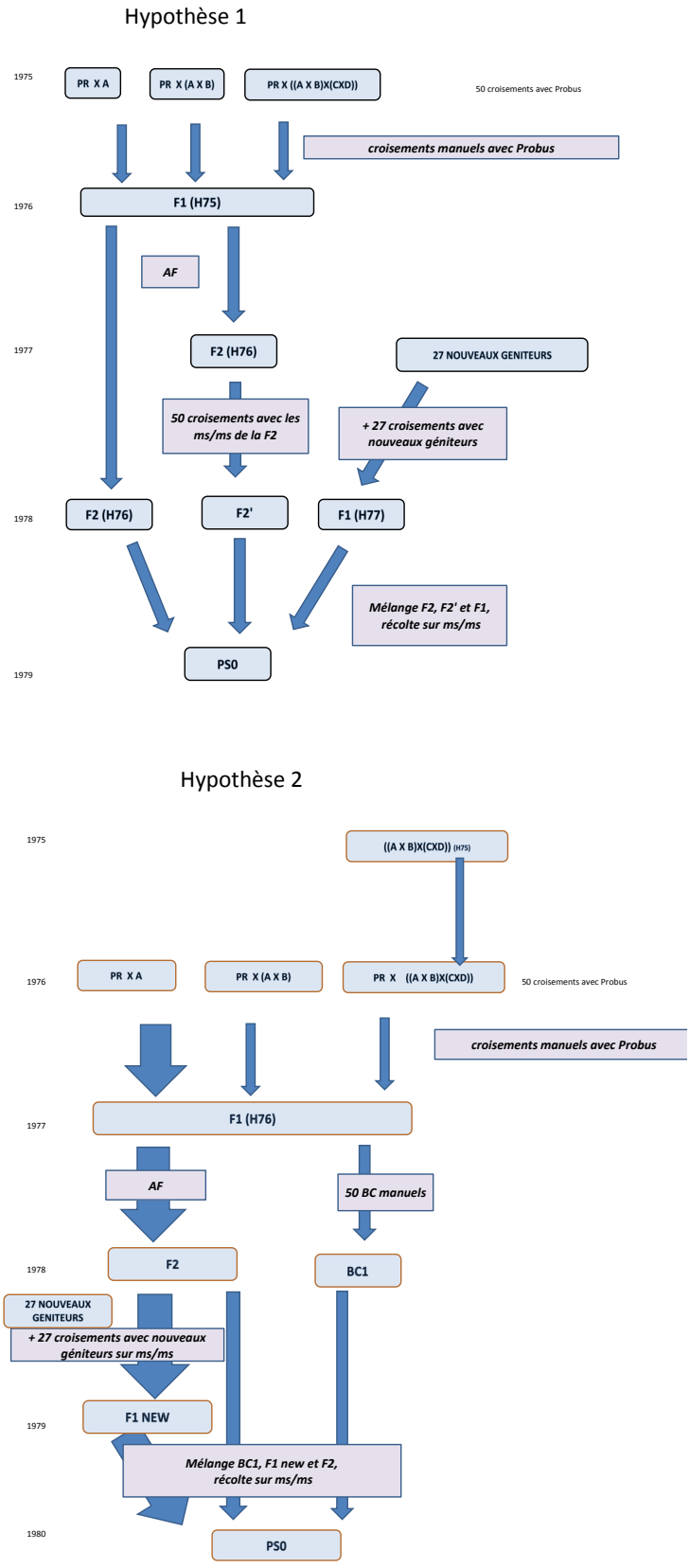


FIGURE A.1 – Schémas récapitulant les deux hypothèses de croisements lors de la création de la population MAGIC INRA.

Annexe B

Utilisation des témoins pour le phénotypage

Lors des expérimentations pour le phénotypage, 24 témoins de type variété moderne ont aussi été implantés parmis les génotypes étudiés. 23 témoins (Tableau B.1) sur les 24 ont été utilisés pour analyser les effets année, effet site et effet milieu pour l'expérimentation réalisée en Angleterre (Tableau B.1).

Tableau B.1 – Liste des témoins utilisés lors du phénotypage

ALCEDO	BRIGADIER	RITMO
ALCHEMY	CAPHORN	ROBIGUS
ALTIGO	CEZANNE	SOISSONS
APACHE	CLAIRE	SPARK
AUBUSSON	CUBIS	TAMBO
AUTAN	HEREWARD	TRANSIT
AZTEC	ORVANTIS	XI-19
BATIS	RECITAL	

Le dernier est la variété Goélant, elle a été semée de manière régulière dans les expérimentations réalisées au Moulon afin d'étudier l'hétérogénité intra-bloc.

Annexe C

L'expérimentation au champs en jour long

Le blé tendre est sensible au jour long ce qui signifie que la floraison est initiée quand la durée du jour est plus longue que la durée de la nuit. Pour étudier précisement la composante de précocité intrinsèque et la sensibilité à la photopériode, nous avons mis en place en 2011-2012 une expérimentation en plein champ avec un semis d'hiver en condition de jour long. Après une petite description de la biologie de la plante sur les signaux qui stimulent la sensibilité à la photopériode, une présentation technique de l'expérimentation sera présentée. Cette étude a fait l'objet du stage de Master 1 de J. Ratet.

La perception de cette longueur du jour est possible grâce à un phytochrome [Sineshchekov et al., 2001]. Cette molécule est composée de deux sous-unités : un chromophore photosensible et une chaîne polypestiptidique nommé apoprotéine [Nakasako et al., 1990]. Sous l'effet de la lumière rouge (R, 650-680nm), le chromophore est isomérisé de façon réversible grâce à l'action de la lumière rouge lointain (FR, 710-740nm). La lumière naturelle est composée d'un mélange de ces deux lumières ce qui équilibre le ratio des deux formes du phytochrome. Ce ratio R/FRvariant de 1,15 en pleine journée à 0,07 au coucher du soleil [Holmes and Smith, 1977] permet à la plante de percevoir l'ombre ou la lumière [Barnes and Bugbee, 1991]. Pour stimuler cette molécule une quantité minimale de lumière comprise entre 2 et 4 $\mu Einstein.m^2$ est nécessaire [González et al., 2003; Warner, 2006; Ghiglione et al., 2008]. Il est important de noter que la photosynthèse est initiée à partir de $200\mu Einstein.m^2$, limite à ne pas dépasser pour pouvoir comparer cette condition de croissance en jour long avec celle en condition de jour naturel.

A partir de ces informations bibliographiques, la distance optimale entre la source lumineuse et les plantes, la puissance nécessaire de la source lumineuse et l'espacement entre les sources pour avoir une lumière la plus homogène possible ont dû être estimés. Après modélisation, il a été décidé d'installer des guirlandes d'ampoules de 20 mètres de long avec 16 ampoules chacune de 40W à une hauteur de 1,50m (Figure C.1). Ces guirlandes ont été allumées en relais avec le soleil pour avoir des journées de 16h (de 6h à 22h), et cela depuis la levée des plantes.

FIGURE C.1 – Photo de l'expérimentation en condition de jour long.

Annexe D

Protocole visuel d'évaluation du remplissage du grain

Ce protocole a été défini par l'équipe d'Ian Mackay du NIAB (Cambridge, UK) dans le but d'avoir une mesure non destructive de la durée de remplissage du grain. Dans les expérimentations de phénotypage de la population MAGIC INRA, les stades 5 et 7 ont été observés.

Suggested protocol for assessing the end of grain-fill

I've been keeping an eye on material in the glasshouse and have come up with the following scale (1-9, where 1= late and 9 = early). As with flowering/heading scores, these should be a rough average for the whole plot.

Canvassing opinion, it seems that grain fill continues right up until the end of senescence so I propose that we call a score of 5 below "mid grain-fill" (grain at maximum size, starting to change colour) and a score of 8 "end of grain fill" (no green colour remaining, but grain not yet dried down).

Scale:

1. Ear, flag leaf, peduncle and grain all very green, no senescence visible. Surface of grain still dull in appearance. ~Zadoks 73
2. Grain swollen but still green, with shiny appearance. Flag leaf just starting to senesce (<1/3 senescent), ear and peduncle still green.
3. Intermediate between 2 and 4
4. Flag leaf approximately halfway through senescence. Grain starting to change colour. Still no real colour change to glumes or peduncle.
5. Flag leaf senescence almost complete (<1/3 still green). Peduncle and glumes beginning to bleach. Grain appears more yellow than green, but still swollen and yields some liquid when squashed. ~Zadoks 77
6. Either peduncle or flag leaf, but not both, still has some colour. Grain starting to dry down. ~Zadoks 83
7. Peduncle and flag leaf completely yellow. Some colour still present on glumes. Grain yellow, but can still be deformed. ~Zadoks 87
8. No green colour to canopy, peduncle or glumes, but ear still feels like it has some moisture to it. ~Zadoks 91.
9. Completely dried – ear feels crisp to the touch, grain dried right down. ~Zadoks 93

Score 1 

Score 2  

Score 4  

Score 5 

Score 7  

Score 9 

Annexe E

Protocoles de notations des différentes expérimentations réalisées au Moulon

Objectifs :

Réaliser des notations sur la population MAGIC INRA de *Triticum aestivum L.*

Principes généraux des notations :

- Réalisées tous les 2 jours (lundi - mercredi - vendredi).
- Flexibilité de la note : de J-2 à J+2.
- Prendre en compte l'effet bordure (zone tampon et concurrence moindre).
- Prendre en compte l'effet phénotypique de chaque variété (toutes les variétés sont différentes).

Les notations effectuées permettront d'apporter des informations sur :

I. La dernière feuille ligulée.
II. L'épiaison.
III. La floraison.
IV. La fin de remplissage du grain.
V. La densité de plantes au stade végétatif.
VI. La hauteur moyenne de parcelle.
VII. La stérilité mâle.

I. La dernière feuille ligulée.

Matériel :
Station de terrain.

Déroulement de la manipulation :
Observer les maîtres brins de la parcelle et identifier la présence des dernières feuilles ligulées.
Attention à vérifier la présence d'une feuille pointante.

Notation :
Début : 3-4 feuilles drapeau ligulées sur la parcelle (Z39).
Fin : 90% des plantes de la parcelle ont une feuille drapeau ligulée (Z39).

II. L'épiaison.

Matériel :
Station de terrain.

Déroulement de la manipulation :
Observer les maîtres brins ligulés de la parcelle et identifier la présence d'épis sortis à 50% de la gaine.

Notations :
Début : 3-4 talles sur la parcelle ayant un épi sortant à 50% de la gaine (Z54).
50% : 50% des épis de la parcelle sortant à 50% de la gaine minimum (Z55).
Fin : 90% des épis de la parcelle sortant à 50% de la gaine minimum (Z59).

III. La floraison.

Matériel :
Station de terrain.

Déroulement de la manipulation :
Observer les épis de la parcelle et identifier la présence d'étamines.

Notations :
Début : 3-4 épis de la parcelle en fleurs (étamines sortant des glumes) (Z61).
50% : 50% des épis de la parcelle en fleurs (Z65).
Fin : 90% des épis de la parcelle en fleurs (Z69).

IV. La fin de remplissage du grain.

1) Score 5

Matériel :
Station de terrain.

Déroulement de la manipulation :
Observer la sénescence des feuilles drapeau de la parcelle (flag leaf).

Notations :
Début : 3-4 talles sur la parcelle ayant la feuille drapeau au moins sénescente aux 2/3 (Z85).
50% : Environ 50% des feuilles drapeau de la parcelle au moins sénescentes aux 2/3 (Z86).
Fin : 90% des feuilles drapeau de la parcelle au moins sénescentes aux 2/3 (Z87).

2) Score 7

Matériel :
Station de terrain.

Déroulement de la manipulation :
Observer la sénescence des feuilles drapeau de la parcelle (flag leaf).

Notations :
Début : 3-4 talles sur la parcelle ayant la paille et la feuille drapeau complètement sèche, l'épi et le rachis sec à 90% (Z90).
50% : Environ 50% des talles de la parcelle à avoir la paille et la feuille drapeau complètement sèches, l'épi et le rachis sec à 90% (Z91).
Fin : 90% des talles de la parcelle à avoir la paille et la feuille drapeau complètement sèche, l'épi et le rachis sec à 90% (Z92).

V. Densité de plantes au stade végétatif.

Matériel :
Station de terrain.

Déroulement de la manipulation :
Observer la densité post semis et les espacements inter-plantes.

Notations :

Densité au stade végétatif :

Compter le nombre de plantes de blé par parcelle
(densité de semis = 20 sauf exceptions).

Espacement :

Seuls les espacements supérieurs ou égaux à 20cm sont relevés.
Les valeurs sont arrondies à la dizaine.
Sur le fichier.xls, seul le chiffre de la dizaine sera noté (ex : 2 pour 20, 3 pour 30…)
En cas d'espace vide sur la bordure, il sera noté ''b''.

VI. Hauteur moyenne.

Matériel :

Station de terrain, mètre de maçon.

Déroulement de la manipulation :

Observer la hauteur moyenne des plantes de la parcelle.

Notations :

On mesure la taille moyenne des plantes sur la parcelle.
La plante est sélectionnée en milieu de parcelle.
La plante doit être représentative de la parcelle.
La valeur observée est arrondie à un chiffre pair.

VII. Stérilité mâle.

Matériel :

Station de terrain, laine rouge, Tapner.

Déroulement de la manipulation :

Observer et marquer les épis ayant le caractère mâle stérile dominant.

Notations :

Compter le nombre de mâles stériles observés par parcelle.

Caractéristiques physiques :

Glumes bâillantes.
Étamines grises et rachitiques (absence de pollen).
Hypertrophie des styles plumeux.

Risques d'erreur

Lors des notations, il est nécessaire de vérifier différentes choses :

- Homogénéité/ hétérogénéité de la parcelle (la ségrégation génétique ou la vernalisation des individus pouvant ne pas être complètement effectuée).
- Les bordures peuvent être plus précoces/ tardives que le reste de la parcelle.
- Prendre en compte l'effet phénotypique de la variété (chaque variété est différente des autres).
- Faire attention à ne pas prendre en compte les tardillons/ rejets de la plante.
- Les plantes extrêmement différentes des autres seront classées ''hors type'', c'est-à-dire qu'ils ne seront pas pris en compte lors des notations.
  Exemple : une plante n'ayant pas épié alors que le reste de la parcelle termine la dessiccation du grain sera classée ''hors type''.

Protocole de notations du blé tendre

(Triticum aestivum L.)

Stades

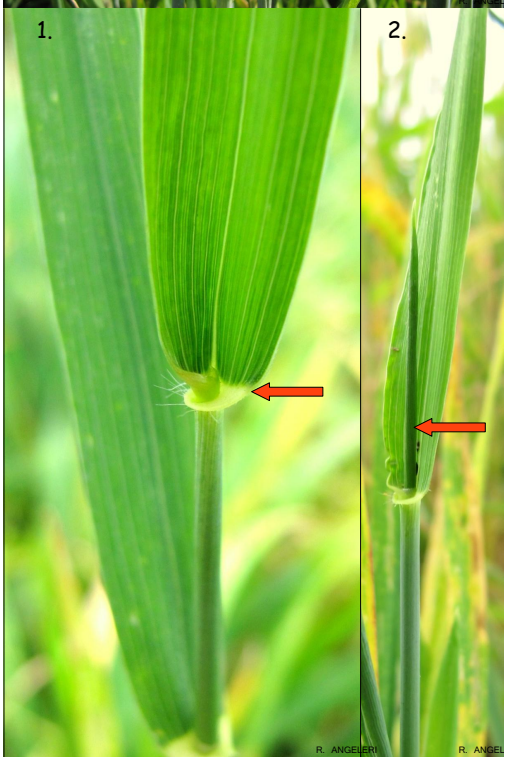
- Dernière feuille ligulée
- Épiaison
- Floraison
- Fin de remplissage du grain
- Densité
- Hauteur
- Stérilité mâle

Attention, lors des notations, il faut toujours prendre en compte les bordures et l'effet phénotypique de chaque variété.

Dernière feuille ligulée

(cf. photo n° 1)

- Vérifier qu'il n'y ait pas de feuille pointante avant de valider le stade. *(cf. Photo n° 2)*



Floraison (ou anthèse)

Début :

2-3 épis de la parcelle en fleurs (étamines sortant des glumes) (Z61).

50% :

50% des épis de la parcelle en fleurs (Z65).

Fin :

90% des épis de la parcelle en fleurs (Z69).

Fin de remplissage du grain

Différents stades de maturation du grain

1. Élongation du grain (Z70)
   Le grain occupe en longueur environ 20% de la taille des glumes.

2. Grain laiteux (Z75)
   Taille potentielle du grain atteinte.
   Le grain est de couleur verte.

3. Grain pâteux (Z85)
   Fin de migration des réserves
   Le grain commence à jaunir.

4. Grain dur (Z92)
   Le grain est jaune et sec.
   Il casse sous la dent.

Score 5

Début :

2-3 talles sur la parcelle ayant la feuille drapeau au moins sénescente aux 2/3 (Z85).

50% :

Environ 50% des feuilles drapeau de la parcelle au moins sénescentes aux 2/3 (Z86).

Fin :

100% des feuilles drapeau de la parcelle au moins sénescentes aux 2/3 (Z87).

Score 7

Début :

2-3 talles sur la parcelle ayant la paille et la feuille drapeau complètement sèches, l'épi et le rachis sec à 90% (Z90).

50% :

Environ 50% des talles de la parcelle à avoir la paille et la feuille drapeau complètement sèches, l'épi et le rachis sec à 90% (Z91).

Fin :

100% des talles de la parcelle à avoir la paille et la feuille drapeau complètement sèches, l'épi et le rachis sec à 90% (Z92).

Densité

- Comptage du nombre de plants de blé par parcelle.
- Comptage des espaces sans plantes (>20cm) de la parcelle.

- Sur la photo :
- - La densité est de 7 plantes.
- - L'espace vide est de 6 (60cm) : b2 + 4.

Hauteur moyenne

- La plante sélectionnée doit être représentative de la parcelle.
- On estime visuellement la taille moyenne de la plante sélectionnée à l'aide d'un mètre.
- La valeur observée est arrondie à un chiffre pair.
- Sur la photo, la hauteur moyenne est de 92cm.

Stérilité mâle

Caractéristiques physiques :

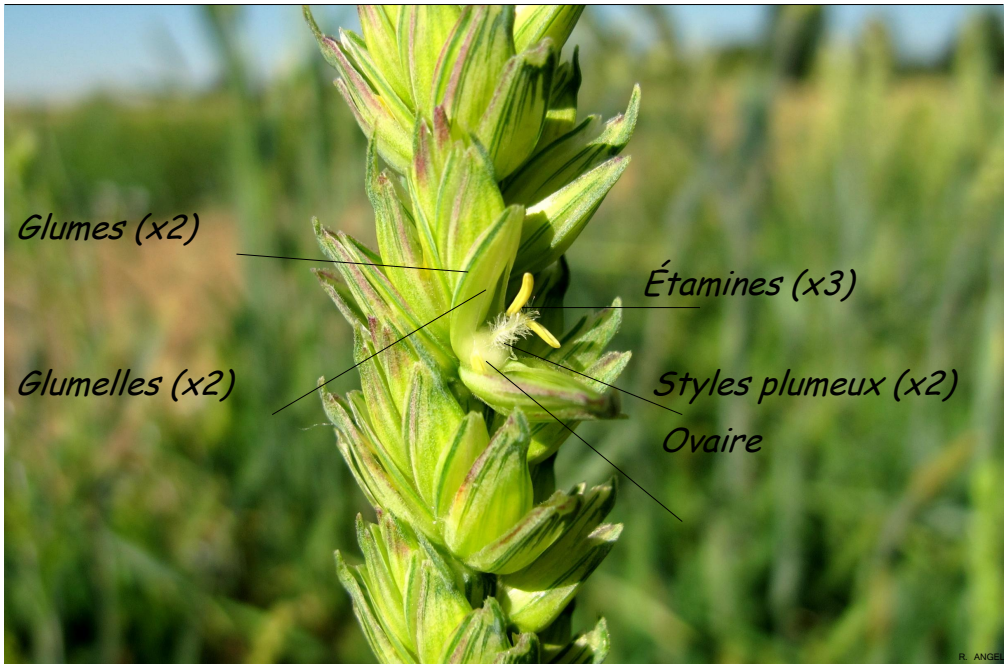
- Glumes baillantes.

- Étamines grises et rachitiques (absence de pollen).

- Hypertrophie des styles plumeux.

Se vérifie sur l'épi

Stérilité mâle : Anatomie d'un épillet



Comparaison entre 2 épillets

Protocole de notations et mesures pour l'expérimentation MAGIC en parcelle agronomique 2011/2012

Gérée par Didier
Interlocuteur scientifique : Stéphanie et Jérôme

Nombre *de parcelles :*

400 génotypes * 2 répétitions + 4 témoins inter blocs

Itinéraire technique :

- Densité de semis : 130 grains/m2
- Pas de régulateur de croissance
- Apport en azote minimum (à ajuster en fonction des besoins de préférence à la floraison) 40U après floraison
- Protection phytosanitaire raisonnée : 3 passages fongicide

Notations :

- Nombre de plantes début hivers
- Nombre de plantes sortie hivers (si très différents de celui début hivers)
- Date d'épiaison : 50% de la parcelle avec un épi à 50% sorti
- Date de floraison : 50% des épis de la parcelle fleuris
- Date de maturation du grain : changement de couleur du collet
- Note de verse
- Note de maladie (si besoin car parcelles traitées)
- Hauteur
- Notations d'aristation (NB, T, B) et de couleur d'épi (blanc, rose, rouge)

Composantes du rendement :

- Nombre d'épis totaux
- Poids de grains totaux
- PMG à 15% d'humidité
- Humidité
- Garder un échantillon d'environ 300g pour future expérimentation

Protocole de notations et mesures pour l'expérimentation MAGIC en pépinière 2011/2012

Nombre *de lignes :*

Semis	Rep	Nb génotypes	Semences	
oct	Rep 1	1114	AF plante marquée	randomisé
	Rep 2	346	FL plante marquée	dans l'ordre
mars	Rep 3	1114	FL plante marquée	randomisé
	Rep 4	346	FL plante marquée	randomisé
avril	Rep 5	1114	FL plante marquée	randomisé

Si il n'y a pas assez de grains en AF plantes marquées, prendre l'autre plantes
Si il n'y a pas assez de grains en FL associé à l'AF (de la rep 1), vérifier si l'autre plante est fertile pour prendre les FL. Si l'autre plante est stérile prendre le reste de la ligne de la rep1 (2010/2011)

Itinéraire technique :

- Densité de semis : 20 grains/lignes
- Pas de régulateur de croissance
- Apport en azote minimum (à ajuster en fonction des besoins)
- Protection phytosanitaire raisonnée

Notations :

- Nombre de plantes levées
- Date de début de dernière feuille ligulée : 2-3 talles avec leur dernière feuille ligulée
- Date de dernière feuille ligulée : 50% des talles avec leur dernière feuille ligulée
- Date de début d'épiaison : 2-3 plantes avec un épi à 50% sorti
- Date d'épiaison : 50% des plantes de la ligne avec un épi à 50% sorti
- Date de début de floraison : 2-3 épis fleuris
- Date de floraison : 50% des épis de la ligne fleuris
- Date de début de maturation du grain (score5 et score 7): 2-3 plantes avec leur feuille drapeau entièrement sénescente. (le score 5 a été noté uniquement si il n'y avait pas eu d'attaque de maladie sur la feuille drapeau)
- Date de maturation du grain (score5 et score 7): 50% des plantes avec leur feuille drapeau entièrement sénescente
- Hauteur dominante ou moyenne si pas de dominance + hétérogénéité (% de plantes à cette hauteur)
- Notations du pourcentage de plantes avec les 3 classes d'aristation (NB, T, B) (5 classes de pourcentage : 1, 25,50,75,100)

Pour les notations de stade, il faut se référer au powerpoint de romain

Composantes du rendement (sur les 400 génotypes répétés):

- Récolte des AF de la rep1 : battage + pesée
- Nombre d'épis totaux, manquant et vert (petits tardillons et grands)
- Poids de grains totaux
- PMG
- sur 3 épis (à récolter avant le reste de la ligne)
    - Nombre d'épillets totaux et stériles par épis
    - Longueur d'épi
    - Poids de chaque épi

Protocole de notations et mesures pour l'expérimentation MAGIC en pépinière 2011/2012 avec éclairage contrôlé

Nombre *de lignes :*

- 400 génotypes
- Semences provenant de la FL de la plante marquée

Conditions contrôlées :

- Eclairage à partir de la levée des plantes
- Eclairage continue avant le levée et après le coucher du soleil pour avoir une journée continue de 6h à 22h (16h) avec une guirlande d'ampoule de 40W espacée d'1m et à 1,5m de hauteur

Itinéraire technique :

- Densité de semis : 10 grains par demi-ligne
- Pas de régulateur de croissance
- Apport en azote minimum (selon les besoins)
- Traitement curatif seulement

Notations :

- Date de dernière feuille ligulée : 50% des talles avec leur dernière feuille ligulée
- Date d'épiaison : 50% des plantes de la ligne avec un épi à 50% sorti
- Date de floraison : 50% des épis de la ligne fleuris
- Date de maturation du grain (score 5 et score7): 50% des plantes avec leur feuille drapeau entièrement sénescente
- Notation du nombre de feuille sur plantes par demi-ligne
- Nombre de plantes par demi-ligne
- Nombre d'épi par plante dans l'ordre de l'extérieur vers l'intérieur de la bande

Pour les notations de stade, il faut se référer au powerpoint de romain

Composantes du rendement :

- Pas de composantes du rendement

Annexe F

Didacticiel d'entraînement à la notation de stades

La réalisation du didacticiel a été le sujet de stage de Romain Angelery en collaboration avec l'INRA Dijon. Cet outil permet s'entraîner et s'évaluer à la notation des stades sur la base de photos dans le but d'harmoniser les notations des différents notateurs.

Il est composé de deux parties : une partie d'apprentissage avec une notice explicative et des photos de chacun des stades et d'une partie de test sous la forme d'un quizz. Pour chacun des stades, quatre dates de notations sont proposées («Non noté», «J-1», «J0», «J+1» et «J+2») avec une banque de trois images par date. A la fin du test, un score est donné, avec une analyse des résultats pour voir les points à améliorer.

Le didacticiel est accessible en ligne à l'adresse suivante : `http://194.94.61.13/quantipest/quantipest/quantipest_utilisateur/index_appli.php?portail=Agrescience&produit=quantipest&main=63&ssrub1=328&ssrub2=329&id_fiche=129&id_fiche=129&theme=178`

Annexe G

Protocoles de fusion des fichiers de notations au champ

PROTOCOLE : FUSION DES FICHIERS DE NOTATIONS

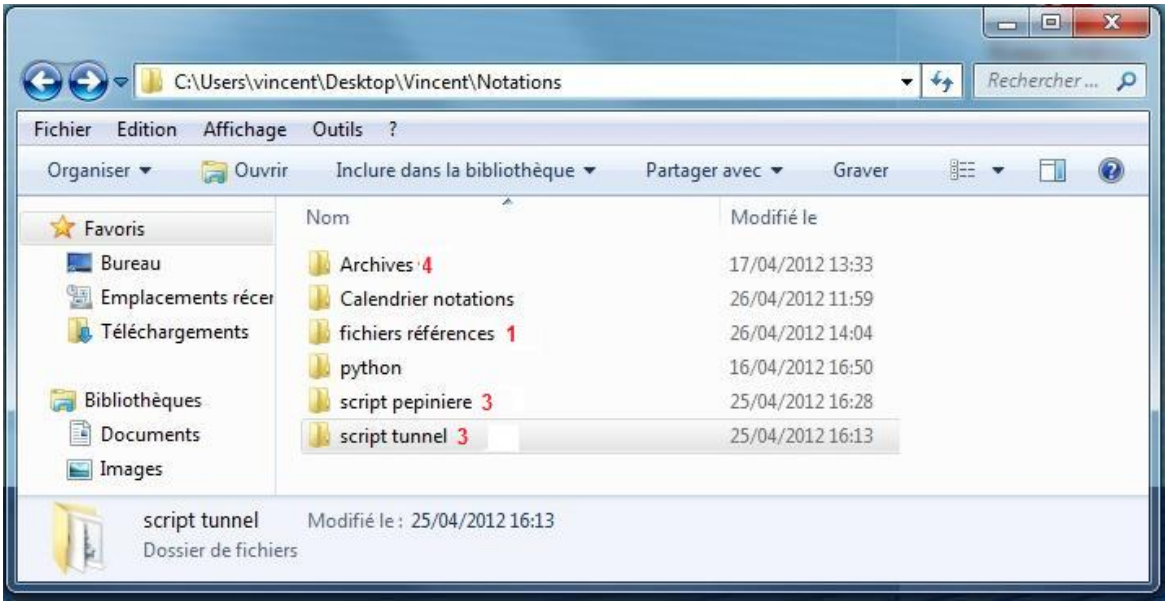
Matériels :

- 3 Psions nommés A, B et C
- 1 Base par Psion
- 1 Ordinateur

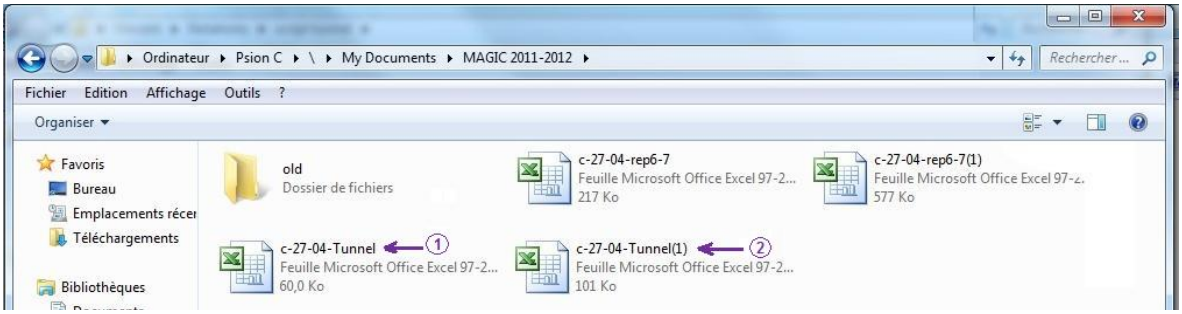
Chemins d'accès aux fichiers :

Dans un but de claireté, nous appellerons « xp » les différentes expérimentations dans lesquelles nous faisons des notations. Les expérimentations sont à considérer de manière indépendante même si les manipulations de fichiers sont similaires. Pour l'année 2011-2012, « xp » corresponds soit à « pepiniere » soit à « tunnel »

- Vincent → Notations → Fichiers références (1) → Archives (2)
  → Script *xp* (3) (*ex : pepiniere ou tunnel*)
  → Archives (4) → mois → xp

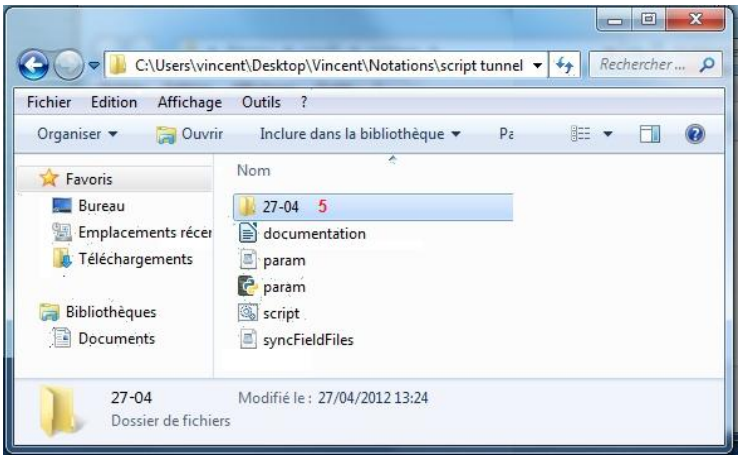


- Psion (A, B ou C) → \ → My Documents → Magic 2011-2012 → old



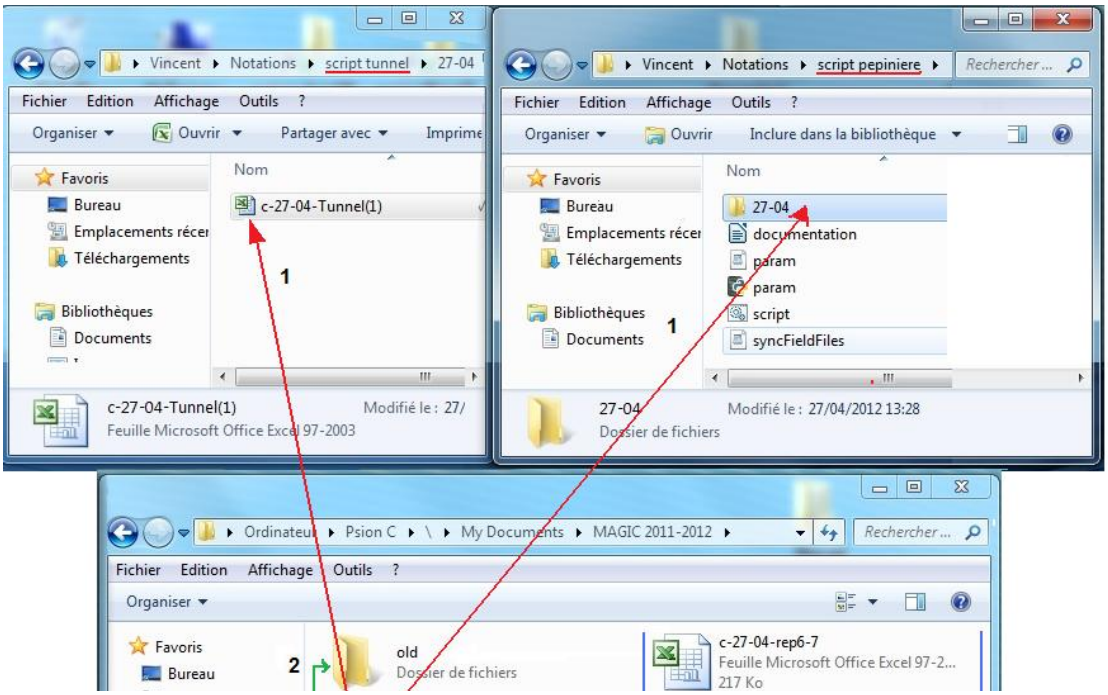
Décharger les fichiers de notations :

- Créer dans « script xp » (**3**) un nouveau dossier (**5**) portant la date de la notation jj-mm (*ex : 27-04*).



**/ ! \ Respecter la nomenclature : jj-mm / ! **

- Brancher tous les Psions en même temps sur leur base respective, (un seul se connecte à l'ordinateur).
- Pour transférer le fichier de notations du Psion, suivre le chemin d'accès : « Ordinateur→ Psion (A, B ou C)→ \ → My Documents → Magic 2011-2012 ».
- Copier le fichier de notation complété du jour (exemple pour le psion C : « *c-27-04-Tunnel (1)* ») et le coller dans le nouveau dossier créé (**5**).
- Faire de même pour les autres expériences.
- Dans le Psion, déplacer les fichiers copiés dans « old » afin de les archiver. (ex : « *c-27-04-Tunnel (1)* »).
- Supprimer les fichiers ayant servi de bases pour les notations du jour (*ex : c-27-04-Tunnel).*



- Débrancher le Psion venant d'être vidé (l'ordinateur se connecte au Psion suivant).
- Transférer les autres fichiers de notations dans le dossier créé (**5**).

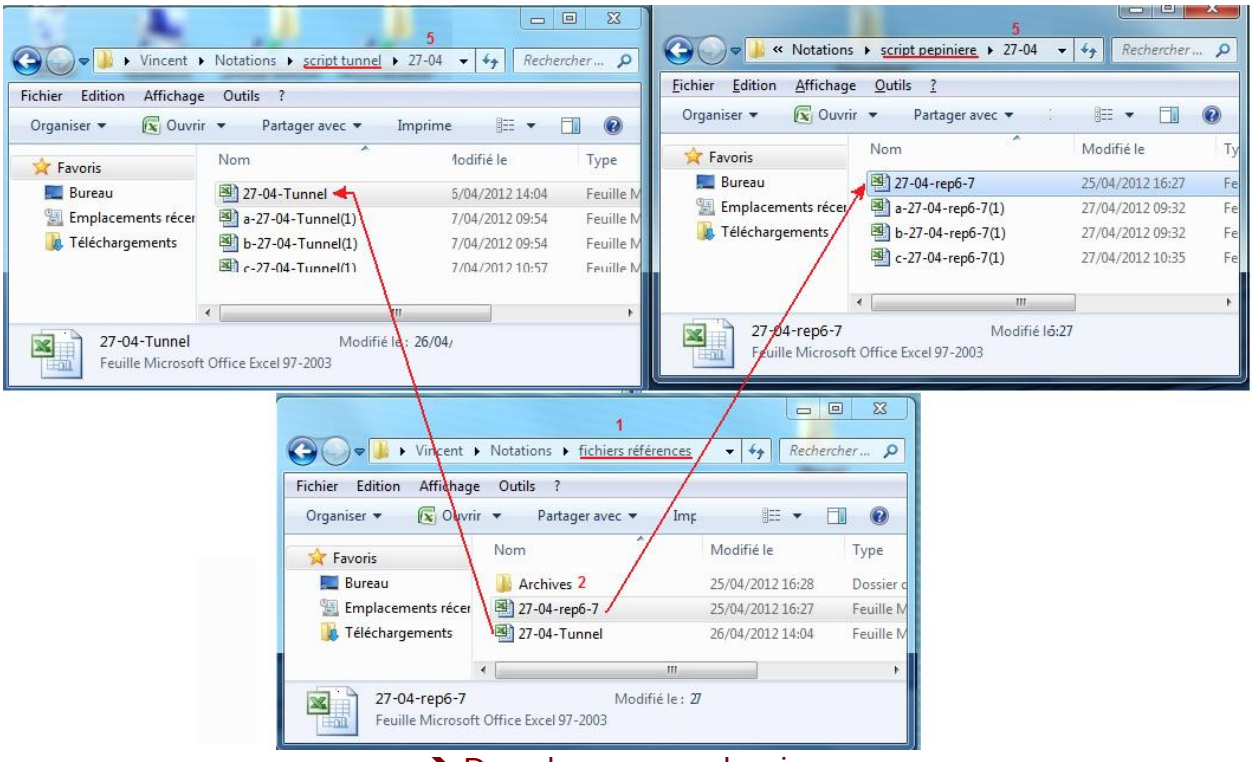
Remarques :

Si le Psion ne se connecte pas :

- attendre 30 secondes
- nettoyer l'interface de connexion
- essayer sur une autre base
- Copier le fichier de notations du jour sur une clé usb (le port usb est sur la base du Psion)

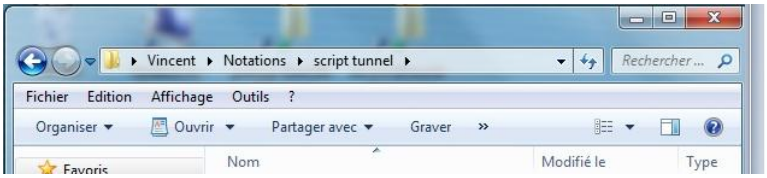
Fusionner les fichiers de notations :

- Dans « fichiers références » (**1**) copier le fichier de référence de chaque xp ayant servi à faire les notations (*ex : 27-04-Tunnel).*
- Coller le fichier de référence dans le nouveau dossier créé (**5**).



➔ Dans le nouveau dossier :
le fichier de référence + les fichiers de notations des Psions.

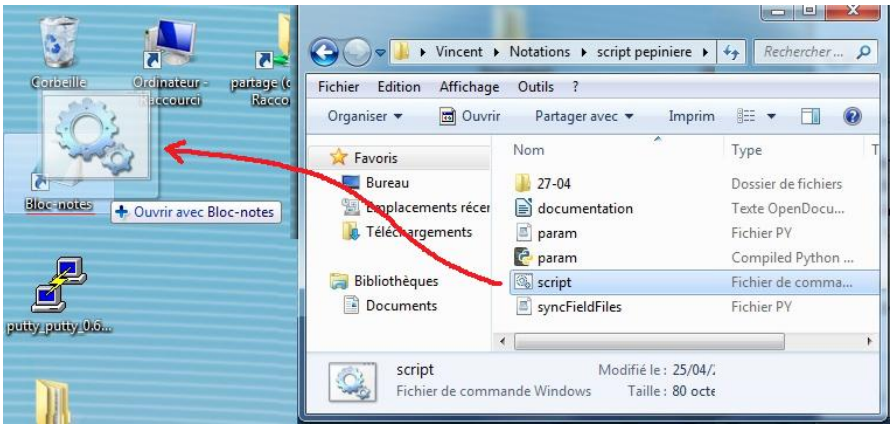
- Dans « script xp » (**3**) ouvrir le document « param » pour régler les paramètres du script.



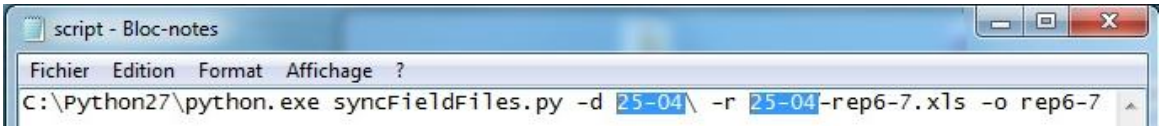

```
#encoding: utf-8

DATE_COLS = [7,8,9,10,11,12]
NUM_COLS = [13,14,]
TEXT_COLS = [1,2,3,4,5,6,15]

DATE_FORMAT = 'DD/MM'
```

➔Le « GAP » correspond au nombre de jours de différence que tolère le script sans marquer de message d'erreur entre la notation du fichier de référence et la notation ajoutée par le notateur.

- Choisir un « GAP » de **2** (jours) si les notations sont en semaine, **3** (jours) si c'est un retour de week-end.
- Enregistrer.
- Faire **Glisser** le fichier « script » dans « bloc note ».

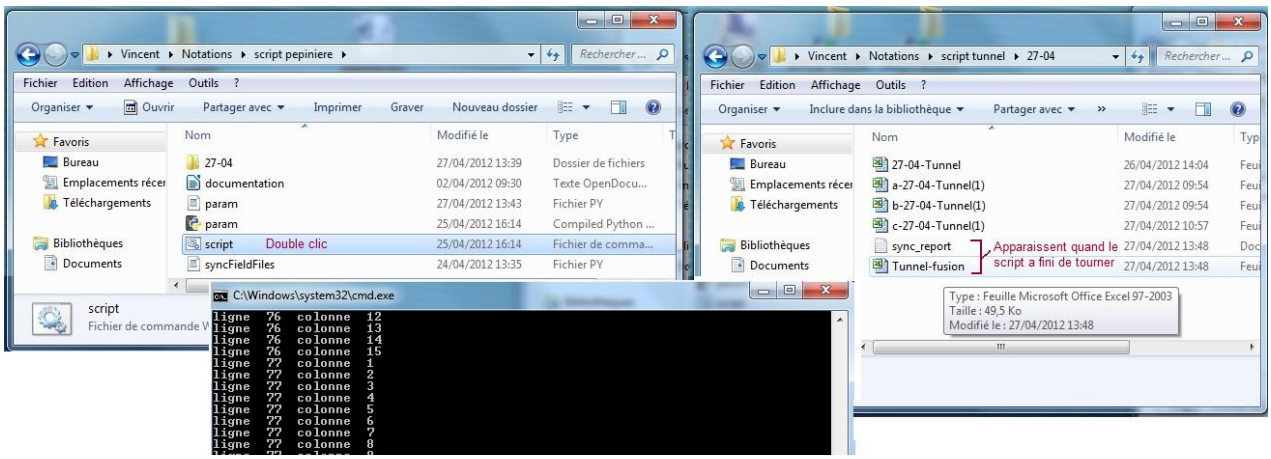


- Changer les deux dates du document par la date du nouveau dossier (**5**) (la dernière date de notations).



- Enregistrer les modifications.
- Quitter le document.
- **Double-cliquer** sur le fichier « script ».

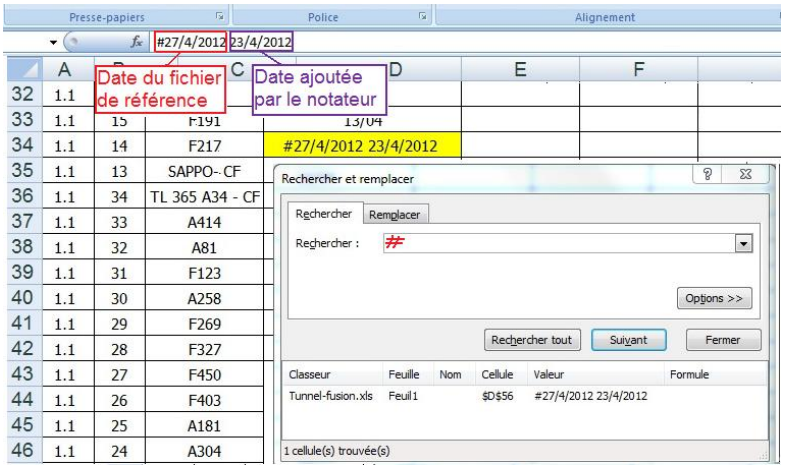
➔ Une nouvelle fenêtre s'ouvre et fait tourner le script.
➔ Un fichier xp-fusion ainsi qu'un fichier sync_report apparaissent dans le nouveau dossier (**5**) contenant les fichiers de notations et le fichier de référence.



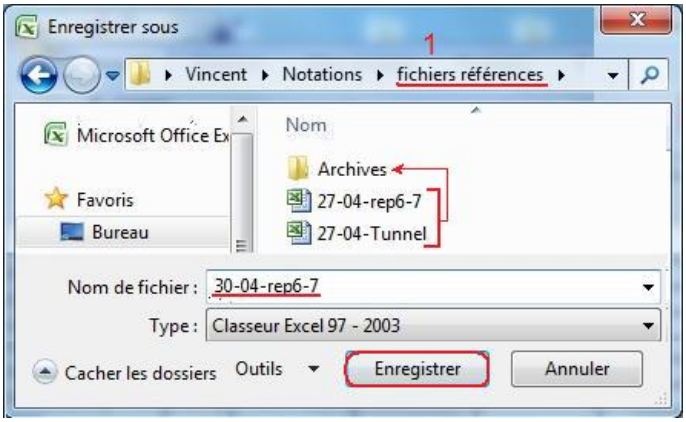
notations extraits des Psion s'il n'y a pas autre chose que des dates dans les colonnes de notation de stade. Vérifier aussi que les noms des colonnes n'ont pas été changés.

- Ouvrir le fichier « xp-fusion » du nouveau dossier (**5**) (ex : Tunnel-fusion).
- Rechercher via la fonction rechercher (Edition → Rechercher ou Ctrl + F) les « # ».

Ces « # » correspondent à une différence de jours entre la date de notation du fichier référence et celle du notateur supérieure au « GAP » prédéfini dans le fichier « param » (2 ou 3 jours).

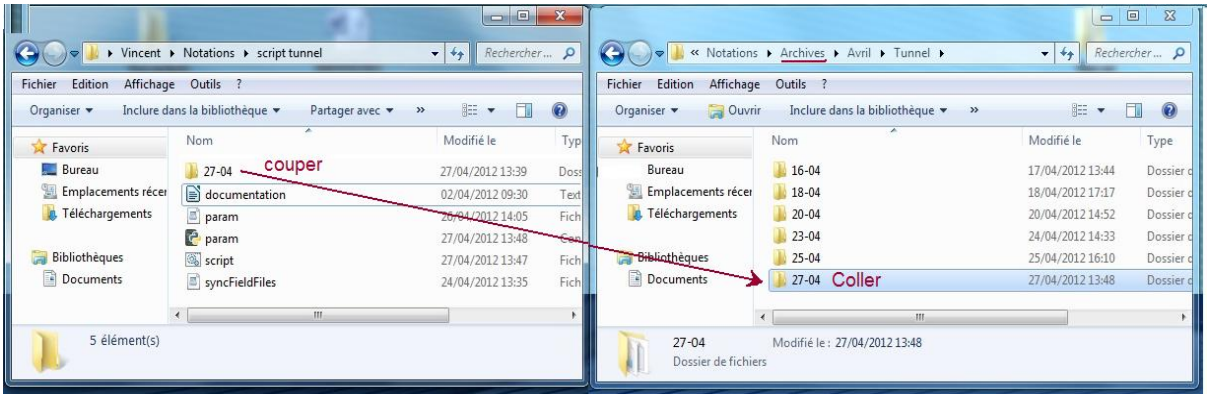


- Corriger les cases contenant des « # » par la date corrigé par le notateur.
- Enregistrer le fichier fusion corrigé dans « Fichiers références » (**1**) au format « date-xp » à la date de la prochaine notation. (ex : *30-04-Tunnel*)
- Déplacer les anciens fichiers de référence dans « Archives » pour les archiver.



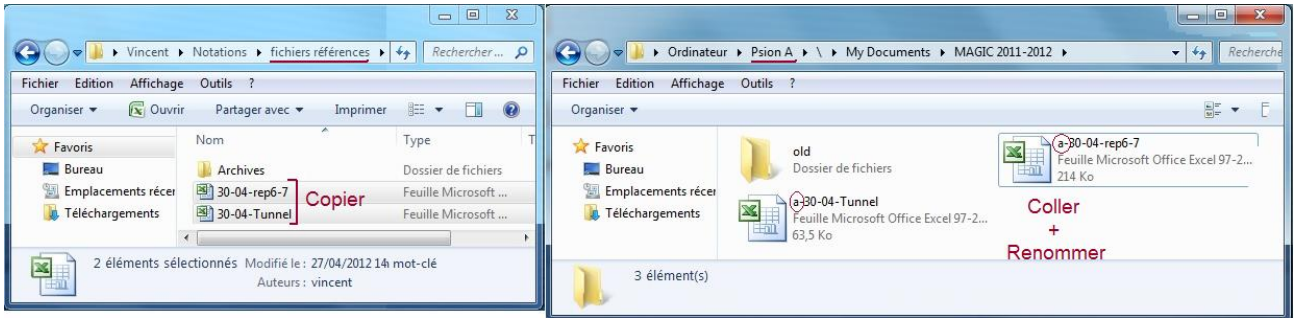
**/ ! ** Conserver le format du fichier : Excel 97-2003 (.xls)
**/ ! ** Ne pas changer la forme du titre du fichier : jj-mm-xp
(ex : 30-04-Tunnel / 30-04-rep6-7)

- Déplacer le nouveau dossier (**5**) dans « Archives » (**2**) au mois correspondant à la date de notation afin de l'archiver.


Vincent ▸ Notations ▸ script tunnel
Rechercher...
Fichier Edition Affichage Outils ?
Organiser Inclure dans la bibliothèque Partager avec
Favoris
Bureau
Emplacements récer
Téléchargements
Bibliothèques
Documents
Nom
Modifié le
Typ
27-04
couper
27/04/2012 13:39
documentation
02/04/2012 09:30
param
param
27/04/2012 13:48
script
27/04/2012 13:47
syncFieldFiles
24/04/2012 13:35
5 élément(s)
« Notations ▸ Archives ▸ Avril ▸ Tunnel ▸
Rechercher...
Fichier Edition Affichage Outils ?
Organiser Ouvrir Inclure dans la bibliothèque
Favoris
Bureau
Emplacements récer
Téléchargements
Bibliothèques
Documents
Nom
Modifié le
Type
16-04
17/04/2012 13:44
18-04
18/04/2012 17:17
20-04
20/04/2012 14:52
23-04
24/04/2012 14:33
25-04
25/04/2012 16:10
27-04
Coller
27/04/2012 13:48
27-04
Modifié le : 27/04/2012 13:48
Dossier de fichiers

Recharger les Psions avec les nouveaux fichiers de notations :

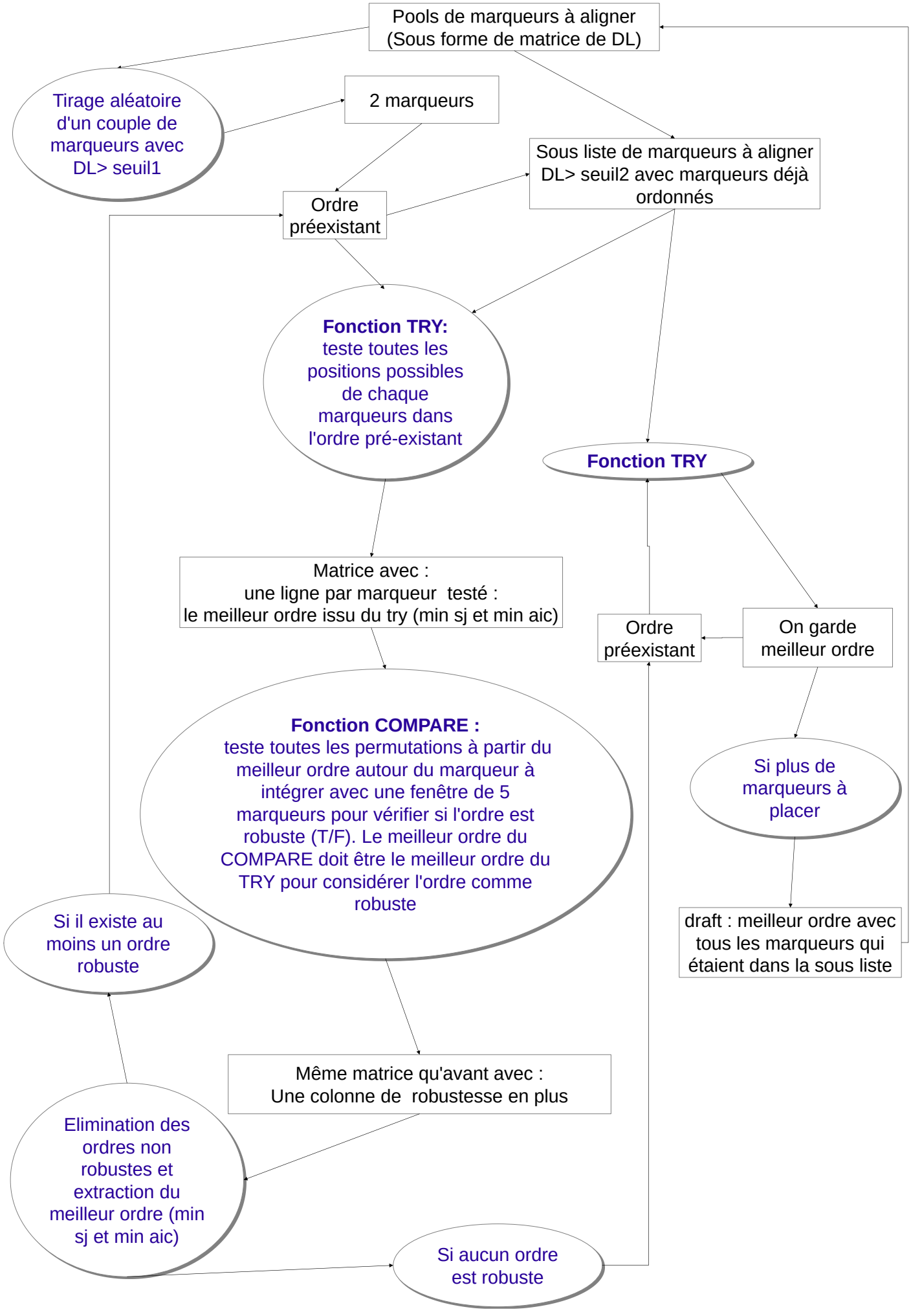
- Copier les 2 nouveaux fichiers de référence pour la prochaine date de notations à partir de « Fichiers référence » (**1**)
- Brancher les Psion sur leur base respective.
- Coller les 2 fichiers dans le Psion détecté par l'ordinateur :
« Ordinateur→ Psion (A, B ou C)→ \ → My Documents → MAGIC 2011-2012 ».
- Renommer les 2 fichiers de référence avec le nom du Psion connecté (ex : *a-25-04-Tunnel pour le Psion A*)



- Enlever le Psion et recommencer pour les autres.

Annexe H

Schéma explicatif de l'algorithme d'ordonnancement des marqueurs


Pools de marqueurs à aligner
(Sous forme de matrice de DL)
Tirage aléatoire d'un couple de marqueurs avec DL> seuil1
2 marqueurs
Sous liste de marqueurs à aligner DL> seuil2 avec marqueurs déjà ordonnés
Ordre préexistant
Fonction TRY:
teste toutes les positions possibles de chaque marqueurs dans l'ordre pré-existant
Fonction TRY
Matrice avec :
une ligne par marqueur testé :
le meilleur ordre issu du try (min sj et min aic)
Ordre préexistant
On garde meilleur ordre
Fonction COMPARE :
teste toutes les permutations à partir du meilleur ordre autour du marqueur à intégrer avec une fenêtre de 5 marqueurs pour vérifier si l'ordre est robuste (T/F). Le meilleur ordre du COMPARE doit être le meilleur ordre du TRY pour considérer l'ordre comme robuste
Si plus de marqueurs à placer
Si il existe au moins un ordre robuste
draft : meilleur ordre avec tous les marqueurs qui étaient dans la sous liste
Même matrice qu'avant avec :
Une colonne de robustesse en plus
Elimination des ordres non robustes et extraction du meilleur ordre (min sj et min aic)
Si aucun ordre est robuste