Utilisation de la sélection génomique dans un programme de pre-breeding : application chez la betterave sucrière

Prune Pegot-Espagnet

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *04/06/2020* **par :**

PRUNE PEGOT-ESPAGNET

Utilisation de la sélection génomique dans un programme de pre-breeding : application chez la betterave sucrière

---

JURY

JACQUES DAVID	Professeur Institut Agro	Rapporteur
CHARLES-ÉRIC DUREL	Directeur de Recherche	Rapporteur
GWENNAELE FICHANT	Professeur d'Université	Examinatrice
JUDITH BURSTIN	Directrice de Recherche	Examinatrice
FABRICE ROUX	Directeur de Recherche	Directeur de thèse
ELLEN GOUDEMAND	Ingénieure agronome	Invitée
BRIGITTE MANGIN	Directrice de Recherche	Invitée

École doctorale et spécialité :
SEVAB : Interactions plantes-microorganismes

Unité de Recherche :
Laboratoire des Interactions Plantes-Microorganismes (UMR CNRS/INRA 2594/441)

Directeur(s) de Thèse :
Fabrice ROUX et *Brigitte MANGIN (retraitée)*

Rapporteurs :
Jacques DAVID et *Charles-Éric DUREL*

Remerciements

« Essentially all models are wrong, but some are useful. » George Box

Ce doctorat n'aurait pas été possible sans mes directrices de thèse, Brigitte MANGIN et Ellen GOUDEMAND, que je tiens à remercier très chaleureusement. Elles ont su me guider avec bienveillance et efficacité tout au long de cette thèse, à Toulouse comme à Lille. Merci tout particulièrement à Brigitte de m'avoir suivie jusqu'au bout de ce travail malgré son départ à la retraite.

Je remercie également Bruno DESPREZ et Karine HENRY, qui m'ont apporté leur expertise sur les questions biologiques et qui se sont toujours impliqués lors des réunions de suivi. Plus largement, je tiens à remercier toutes les personnes de Florimond Desprez Veuve & Fils qui ont pris le temps de m'expliquer et de me montrer leurs travaux sur les betteraves sucrières, en champ comme en serre, et de m'avoir fait découvrir la plante derrière toutes les données numériques.

Je tiens ensuite à remercier tous les membres de mon jury qui ont accepté d'évaluer l'aboutissement de ces trois années de thèse, Charles-Éric DUREL et Jacques DAVID en tant que rapporteurs, Judith BURSTIN et Gwennaele FICHANT en tant qu'examinatrices. Malgré le contexte sanitaire particulier de cette année 2020 et les contraintes matérielles qui en découlent, ils ont tous répondu présent pour que je puisse soutenir dans les meilleurs délais possibles et je les en remercie vivement ! Un grand merci aussi à Fabrice ROUX d'avoir accepté de devenir mon directeur de thèse suite au départ à la retraite de Brigitte, ce qui m'a permis d'aller au bout de cette aventure.

Mes remerciements vont également aux membres de mon comité de thèse, Hendrik TSCHOEP et Renaud RINCENT, qui ont su apporter un regard extérieur pertinent sur le travail effectué.

Merci à tous les membres de l'équipe Tournesol au sens large pour ces nombreux moments de partage autour de gâteaux divers et variés, pour parier sur qui du diabète ou du cholestérol aura notre peau ! Je remercie en particulier mon chef d'équipe, Nicolas LANGLADE, qui a pris soin de m'associer aux différentes activités de l'équipe, mais aussi Alexandra L, Camille T, Marie-Claude B, Mireille C, Nicolas B, Nicolas P, Olivier C. et Stéphane M. pour leurs encouragements.

Merci aux différentes personnes du bureau des non-perm pour leur soutien sans faille et leur humour ~~quelquefois~~ très limite ! Merci notamment à Harold DURUFLÉ pour ses conseils et sa super voiture, à Pierre MARTIN toujours prêt à trouver de bons plans resto, à Alex DUHNEN, la bonne conscience du bureau, et enfin à Jean LECONTE et Camille PUBERT, la future génération de thésards qui seront chargés de transmettre « l'esprit du bureau ».

Je me dois aussi de remercier Caroline BAROUKH qui partage nos déjeuners ainsi que son expérience, aussi bien sur le plan scientifique que culinaire en passant par le bricolage et la couture.
Merci également à ceux qui m'ont accompagnée dans une grande partie de cette aventure et qui sont partis découvrir de nouveaux horizons, comme Pauline DURIEZ à Vancouver, Louise GODY à Paris, Caroline LAFFRAY à Bordeaux, ou Fanny BONNAFOUS dans l'appartement juste au dessous du mien.

Un merci tout particulier à Florie GOSSEAU pour avoir dessiné « Madame Better » qui m'a accompagnée pendant le concours Ma Thèse en 180 secondes et pour son grain de folie qui met des paillettes dans nos vies !

D'un point de vue plus personnel, je tiens aussi à remercier tous mes amis de Toulouse qui me permettent de me changer les idées tous les week-ends depuis plus de 10 ans maintenant. Ce groupe très soudé regorge de personnalités très différentes mais toutes formidables, toujours partantes pour tester de nouvelles activités et voyager, ou débattant pour choisir LE film de la soirée. Merci à tous, vous êtes géniaux !

Pour finir, je tiens à remercier l'ensemble de ma famille qui m'a toujours soutenue (et qui a même fait semblant de comprendre mon sujet de thèse), et plus particulièrement ma mère qui est devenue une super cuisinière de peur que je meure de faim pendant les dernières semaines de rédaction, mon père qui a dessiné de superbes betteraves pour mes présentations, et enfin Maxime qui a su se rendre indispensable et illuminer ma vie depuis plus de 7 ans.

Merci à tous !

Table des matières

Résumé

L'objectif de la sélection variétale est de produire de nouvelles variétés à partir de la diversité existante. L'accumulation des caractères prérequis à la commercialisation des variétés de betterave sucrière tels que le rendement en sucre par hectare, la résistance à des maladies de plus en plus nombreuses sur le cahier des charges, ou encore la diminution des intrants comme les intrants azotés, a eu pour conséquence de réduire considérablement la variabilité génétique disponible dans les programmes de pre-breeding. Le projet AKER a été mis en place pour permettre d'élargir cette diversité génétique grâce à une approche originale d'utilisation de ressources génétiques exotiques. 16 accessions exotiques représentant l'ensemble de la diversité allélique qui n'est pas déjà présente au sein des betteraves sucrières élites ont ainsi été identifiées à partir de l'analyse d'une collection de gènes en provenance de ressources du monde entier. L'objectif de cette thèse est de favoriser l'introgression des ressources génétiques exotiques découvertes dans le cadre d'AKER dans un programme de pre-breeding en utilisant la sélection génomique. Pour ce faire différents schémas de pre-breeding doivent pouvoir être simulés et comparés afin de guider la production d'une population de pre-breeding comprenant des fragments d'accessions exotiques, population qui constituera un réservoir de diversité génétique utile dans lequel les sélectionneurs pourront puiser pour créer de nouvelles variétés de betteraves sucrières. Le postulat fait dans le programme AKER selon lequel l'introduction de régions exotiques dans un programme de pre-breeding peut permettre d'apporter une diversité génétique utile a tout d'abord été vérifié par la comparaison de l'architecture génétique de plusieurs caractères dans deux populations : une descendance (élite x exotique), et un panel élite. Cette architecture génétique qui correspond au nombre de régions génomiques impliquées dans l'expression du caractère, leur localisation sur le génome, et la proportion du caractère qu'elles expliquent, a été déterminée grâce à une étude de QTLs. L'architecture génétique du rendement racinaire a ensuite été étudiée dans des populations appelées « populations AKER » , issues de l'introgression de chaque accession exotique au sein d'un germoplasme élite. Cette étude a permis d'évaluer l'effet de chaque fragment exotique sur le rendement racinaire. Un simulateur a alors été développé grâce auquel, à partir des populations AKER, différents schémas de pre-breeding utilisant la sélection génomique ont été simulés et comparés. Ces simulations ont permis d'étudier l'impact de plusieurs paramètres sur l'évolution du rendement racinaire et sur la diversité génétique présente au sein de la population de pre-breeding finale.

Mots clés : sélection génomique, schéma de pre-breeding, simulations, betterave sucrière.

Summary

The goal of varietal selection is to produce new varieties from existing diversity. For marketable sugar beet varieties, the accumulation of prerequisite traits such as sugar yield per hectare, multiple disease resistance or reduced nitrogen inputs, dramatically reduced the genetic variability available in breeding programs. One of the AKER project's main goal was to enrich this genetic diversity with an original approach making use of exotic genetic resources. An analysis of a collection of genes from resources around the world allowed to identify 16 exotic accessions representing all the allelic diversity absent from elite sugar beets. The purpose of this PhD is to promote the introgression of exotic genetic resources discovered in the AKER project in a pre-breeding program using genomic selection. Different pre-breeding schemes must be able to be simulated and compared in order to guide the production of a pre-breeding population comprising fragments of exotic accessions, a population which will constitute a useful and diverse gene pool from which breeders can draw to create new varieties of sugar beets. The hypothesis made in the AKER program that the introduction of exotic regions into a pre-breeding program can provide useful genetic diversity was first verified by comparing the genetic architecture of several characters in two populations : an (elite x exotic) progeny and an elite panel. This genetic architecture, which corresponds to the number of genomic regions involved in the expression of the trait, their location on the genome, and the proportion of the trait that they explain, was determined thanks to a QTL study. The genetic architecture of root yield was then studied in populations called "AKER populations", each population created from the introgression of a single exotic accession within an elite germplasm. This study assessed the effect of every exotic fragment on root yield. With a specially developed simulator, different pre-breeding schemes all starting with AKER populations were simulated and compared. These simulations made it possible to study the impact of several parameters on the evolution of root yield and on the genetic diversity present in the final pre-breeding population.

Keywords : genomic selection, pre-breeding schemes, simulations, sugar beet

Abréviations

ACP : Analyse en Composantes Principales

ADN : Acide DésoxyriboNucléique

BLUP : Meilleur prédicteur linéaire sans biais (Best Linear Unbiased Predictor)

CMS : Stérilité mâle cytoplasmique (Cytoplasmic Male Sterility)

DL : Déséquilibre de Liaison

EMMA : Efficient Mixed-Model Association

GBLUP : Genomic Best Linear Unbiased Predictor

GS : sélection génomique (Genomic Selection)

GV : sélection sur la vraie valeur génétique

GWAS : Genome-Wide Association Studies

K : Taux de potassium

LG : Chromosome / Groupe de liaison (Linkage Group)

LIPM : Laboratoire des Interactions Plantes Micro-organismes

MAF : fréquence de l'allèle minoritaire (Minor Allele Frequency)

MAS : Sélection assistée par marqueur (Marker Assisted Selection)

MLMM : Multi-Locus Mixed Model

N : Taux d'acide alpha-aminé

Na : Taux de sodium

PS : sélection phénotypique (Phenotypic Selection)

QTL : locus d'un caractère quantitatif (Quantitative Trait Locus)

RY : Rendement racinaire (Root Yield)

S : Taux de sucre (Sugar)

SELKIT : KIT de SELection génomique

SF : gène d'auto fertilité

SNP : Single Nucleotide Polymorphism

WS : Taux de sucre blanc (White Sugar)

WSY : Rendement en sucre blanc (White Sugar Yield)

Chapitre 1

Contexte de l'étude

La betterave sucrière est une plante cultivée principalement pour sa racine charnue riche en sucre. Elle est produite dans plus de 50 pays et représente 30% de la production de sucre mondiale (Iqbal and Saleem 2015), derrière le sucre de canne dont le coût de production est plus faible. La France est le premier pays producteur de betterave sucrière au niveau européen, et le second au niveau mondial. En 2019, 450 000 hectares de betteraves ont été cultivés en France, produisant 38 millions de tonnes de betteraves dont est extrait du sucre blanc. Le sucre de betterave est présent sur le marché sous différentes formes : 11% est consommé comme sucre de bouche, 58% est destiné aux industries alimentaires, 12% est utilisé par les industries chimiques et pharmaceutiques, et 19% est transformé en alcool et éthanol. La betterave sucrière est donc une plante importante pour l'économie française. C'est pourquoi les sélectionneurs cherchent à améliorer cette plante afin d'obtenir des variétés plus concentrées en sucre, tout en diminuant les coûts de production. Cependant, à force de sélections, la diversité génétique utile présente au sein du matériel de sélection s'est amenuisée.

L'objectif de cette thèse est de simuler et de comparer différents schémas de pre-breeding afin de guider la création d'une nouvelle population de pre-breeding intégrant des accessions exotiques croisées avec du matériel élite. Cette population de pre-breeding constituera un réservoir de diversité génétique utile dans lequel les sélectionneurs pourront puiser afin de créer de nouvelles variétés de betteraves sucrières plus performantes. Ces variétés pourront être intégrées au réseau des ressources génétiques des Beta (Beta Genetic Ressources Network) auquel participent la plupart des sélectionneurs de betteraves sucrières avec l'appui du Conseil International pour les Ressources Génétiques des plantes (IPGRI).

Dans cette introduction, les informations biologiques importantes concernant la betterave sucrière, l'histoire de sa domestication (notamment en France) et les objectifs guidant sa sélection sont présentés. Les différentes méthodes de sélection phénotypiques et génomiques, utilisées pour son amélioration, sont ensuite définies. Enfin, le projet AKER dans lequel s'inscrit cette thèse est présenté avant d'aborder la problématique ainsi que le matériel expérimental et les approches qui ont été menées au cours de ce doctorat.

1.1 La betterave sucrière

1.1.1 Classification botanique et informations biologiques

La betterave sucrière (*Beta vulgaris* L. subsp. *vulgaris*.) appartient au genre *Beta* de la famille des Chénopodiacées. Ce genre est divisé en quatre sous-genres : *Beta*, *Corollinae*, *Nanae* et *Procumbentes*. Les betteraves cultivées appartiennent au sous-genre *Beta*, tout comme d'autres bettes sauvages. Tous les membres du sous-genre *Beta* sont diploïdes avec 2n = 18 chromosomes mis à part la *Beta macrocarpa* qui peut être diploïde ou tétraploïde. Les espèces regroupées dans ce sous-genre peuvent être croisées et donnent une descendance féconde. La betterave sucrière a été séquencée pour la première fois en 2014, et son génome a une taille estimée entre 714 et 758 Mbp (Dohm et al. 2014). La betterave sucrière est une plante principalement bisannuelle. Elle développe une racine charnue riche en saccharose pendant la première année, ou stade végétatif (Figure 1.1), et une tige portant les futures graines la seconde année, ou stade reproductif (Figure 1.2). Une période de vernalisation d'environ trois mois est nécessaire pour passer du stade végétatif au stade reproductif, stade où la plante monte en graines. Il faut alors compter environ cinq semaines jusqu'à la floraison (Figure 1.3). Les fleurs ne sont fertiles qu'une dizaine de jours, il est donc indispensable que les parents du futur croisement fleurissent en même temps. Il est possible de couper le sommet de la hampe florale pour ralentir la croissance de plantes trop précoces.

FIGURE 1.1 – Betterave sucrière au stade végétatif (stockage de sucre dans la racine)

FIGURE 1.2 – Betteraves sucrières au stade reproductif (montée en graines)

(a) (b) (c)

FIGURE 1.3 – (a) Fleurs de betterave sucrière mâle-fertile ; (b) Fleurs de betterave sucrière mâle stérile ; (c) Graines de betterave sucrière

Actuellement la betterave sucrière est majoritairement commercialisée sous forme d'hybride diploïde monogerme (Figure 1.4). A l'origine multigerme, il était nécessaire pour l'agriculteur d'effectuer manuellement un démariage au champ afin de ne garder que les pousses les plus vigoureuses. La découverte d'une source génétique monogerme en 1950 (Savitsky et al. 1950) a permis d'éviter cette opération coûteuse et qui pouvait blesser la plante. Des lignées mâles stériles monogermes sont ainsi croisées avec des pollinisateurs multigermes, diploïdes ou tétraploïdes, pour donner les semences hybrides monogermes actuelles. La création de telles variétés hybrides a été rendue possible grâce à la découverte de la stérilité mâle nucléo-cytoplasmique (SMC) par Owen (1945). Cette SMC est le résultat d'une interaction entre des gènes nucléaires et de modifications du génome mitochondrial. Pour obtenir une descendance de plantes mâles stériles, les plantes à SMC doivent être pollinisées par des plantes dites « mainteneuses de stérilité » . Ces « mainteneuses » apportent les allèles appropriés pour la stérilité mâle, mais sont fertiles car possédent un génotype mitochondrial « normal » . Pour obtenir des graines d'hybride 3-voies, une lignée maternelle monogerme SMC est tout d'abord croisée avec une lignée « mainteneuse » appelée O-type en référence à Owen afin d'assurer un rendement élevé lors de la production de semences hybrides. La descendance de ce croisement, également monogerme et SMC, est alors croisée avec un pollinisateur multigerme. Il en résulte des graines hybrides monogermes qui seront commercialisées.

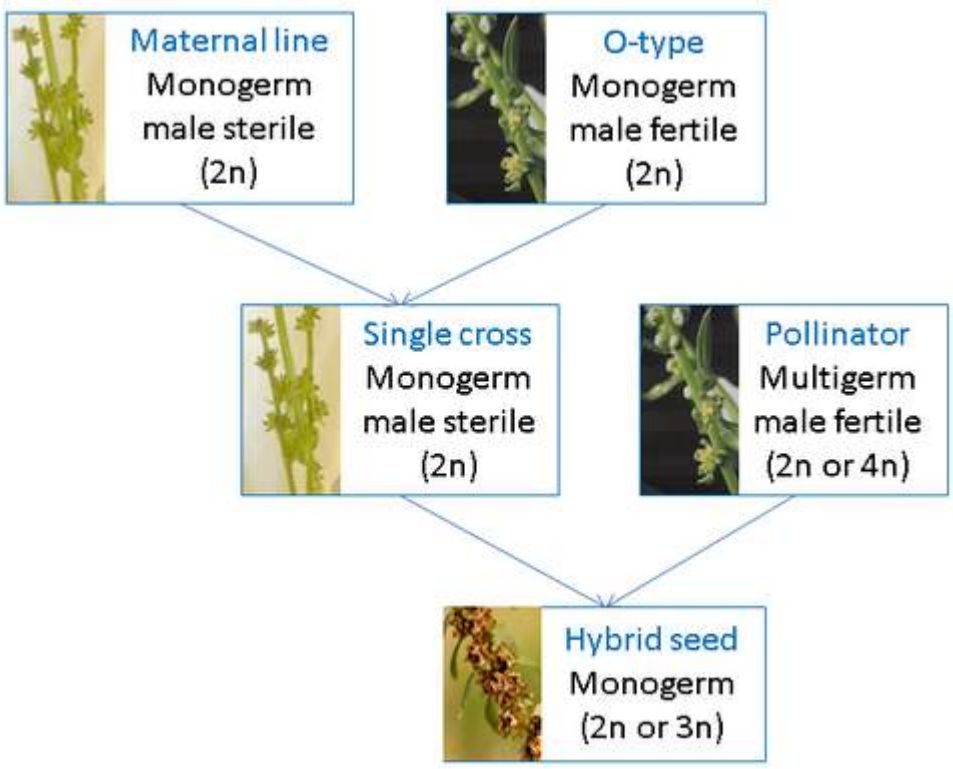


FIGURE 1.4 – Schéma de l'obtention de graines hybrides 3-voies, inspiré de la Figure 3 de l'article (Märländer et al. 2011). n représente le niveau de ploïdie de la plante.

Un allèle de stérilité mâle nucléaire récessif a également été découvert sur le chromosome 1 par Owen (1952), mais il ne permet pas d'obtenir une population composée à 100% de plantes mâles stériles ce qui est nécessaire pour produire des graines hybrides à l'échelle industrielle. Cet allèle est en revanche utilisé pour faciliter les croisements dans du matériel autofertile.

La betterave sucrière est généralement fortement incompatible. Isolée, elle ne produit que très peu, voire aucune graine. Cette auto-incompatibilité est due à une série allélique au locus appelé S ((Larsen 1977)) : lorsqu'un allèle S porté par le pollen peut être apparié avec un allèle identique dans le pistil, la graine ne se développe pas. Il existe cependant un gène dominant de fertilité spéciale (Sf) découvert à nouveau par Owen ((Owen and Ryser 1942)) qui induit une auto-fertilité obligatoire. Les grains de pollen porteurs de cet allèle Sf ne sont en effet pas reconnus par le pistil et surtout germent plus vite que les autres grains de pollens compatibles. Lorsqu'une plante est fertile et possède cet allèle Sf dominant, la totalité des graines produites sont issues d'autofécondations même en présence d'autres betteraves sucrières fertiles à proximité.

Le pollen de la betterave sucrière est majoritairement véhiculé par le vent (plante anémophile), les insectes ne jouant qu'un rôle mineur (très légèrement entomophile). Il est donc important d'isoler les champs de sélection ou les champs de production commerciale de semences pour éviter les mélanges de matériel génétique, le pollen pouvant être transporté sur une grande distance.

1.1.2 Histoire de la sélection de la betterave sucrière

La betterave sucrière est une culture très récente vis-à-vis de l'histoire de l'agriculture. En effet, alors que la canne à sucre est cultivée en Nouvelle-Guinée dès 8 000 av. J.-C. (Deerr et al. 1950), il faut attendre la fin du XVIème siècle pour que l'agronome français Olivier de Serres remarque une similarité entre le jus de betterave cuit et le sirop issu de la canne à sucre. Cette découverte n'est exploitée qu'en 1747, lorsque le chimiste allemand Andreas Sigismund Marggraf parvient pour la première fois à cristalliser du sucre de betterave en laboratoire. Son élève François Charles Achard s'emploie alors à l'industrialisation de ce processus et finit par créer la toute première fabrique de sucre de betterave en 1801. Il se consacre alors à la sélection de betteraves propres à la production de sucre, et réussit à obtenir une variété ayant une teneur en sucre de 7%, la « Blanche de Silésie » (Bosemark 1979). Il s'agit de l'ancêtre des betteraves sucrières cultivées de nos jours en France et en Europe.

L'essor de la betterave sucrière en France est profondément lié à l'Histoire française. En 1806, Napoléon Ier instaure un blocus continental afin d'empêcher le Royaume-Uni de commercer avec le reste de l'Europe (Ganiere 1971). L'import de sucre de canne en France depuis les colonies antillaises en est fortement impacté et le sucre devient une denrée rare au prix exorbitant. Pour pallier ce problème, l'empereur ordonne le 25 mars 1811 que 32 000 hectares du nord de la France soient réservés à la culture de la betterave et promet un million de francs à toute entreprise qui parviendrait à produire du sucre de betterave à un prix raisonnable (Viel and Brançon 1999). Le 2 janvier 1812, Napoléon Ier apprend que deux pains de sucre blanc semblables à ceux issus de la canne ont été produits dans les ateliers du jeune botaniste et industriel français Benjamin Delessert. Il décore aussitôt l'entrepreneur de la légion d'honneur et publie un nouveau décret ordonnant la culture de 100 000 hectares de betteraves sucrières, la création de cinq écoles de sucrerie et accordant 500 licences pour établir de nouvelles fabriques : l'industrie de la betterave sucrière française est née. Le marché étant assuré, les efforts pour améliorer la plante ainsi que ses techniques culturales et industrielles ont continué après la fin du blocus. Les résultats de cet investissement sont spectaculaires : la teneur en sucre est passée de 7% de la matière fraîche de la racine au début du XIXème siècle à plus de 17% en 1992 et d'une production par hectare de 700 kg à 15.000 kg de sucre blanc soit 20 fois plus en 200 ans d'amélioration.

1.1.3 Importance économique et objectifs de sélection

En 2017, la France est le second pays producteur de sucre de betterave au niveau mondial, après la Russie (Figure 1.5).

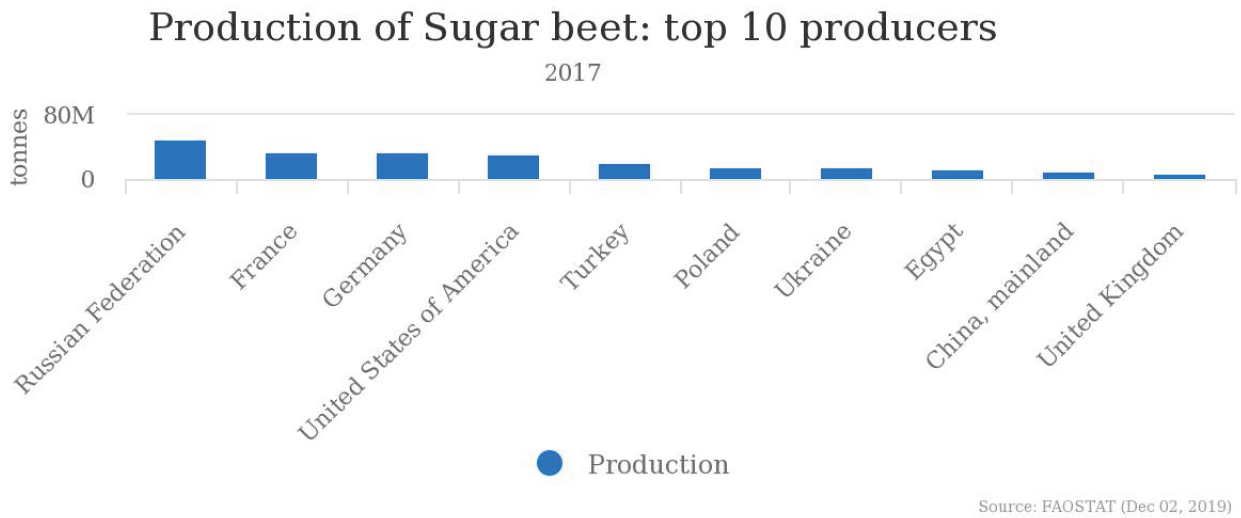


FIGURE 1.5 – Production en tonnes des 10 meilleurs pays producteurs de betterave sucrière en 2017 (source : FAOSTAT)

Au niveau européen, la France est le premier producteur de sucre de betterave. Environ 5.1Mt de sucre blanc ont été produit en 2018/2019, soit 30% de la production européenne. 2.9Mt ont été vendus en France, 2 100Mt ont été commercialisés dans l'Union Européenne, et 0.7Mt ont été exporté hors de l'Union Européenne. Lors de l'année civile 2018, l'industrie sucrière a ainsi rapporté 963 M€. Il est donc important de continuer à améliorer cette plante, dont la commercialisation constitue une contribution essentielle à la balance commerciale de la France. L'amélioration de la production de sucre de betterave au cours des deux derniers siècles est due à plusieurs grandes découvertes (Doré and Varoquaux 2006). La création de variétés de betteraves sucrières monogermes représente une avancée majeure puisqu'elle permet de s'affranchir du démariage des pousses, qui entraînait un stress pour les plantes ainsi qu'un coût important. Le rendement a également pu être amélioré par le passage aux variétés hybrides ainsi que par l'introduction de la résistance à la montaison qui permet de semer plus précocement et ainsi de laisser plus de temps à la racine pour stocker le saccharose. Enfin, le développement de tolérances ou de résistances à certaines maladies et parasites a contribué à limiter les pertes de rendement. Il est ainsi possible de citer l'introduction de la résistance au virus de la frisolée américaine (appelée *curly top*), et surtout au virus de la rhizomanie (maladie répandue mondialement, véritable *Phyloxera* de la betterave, virus transmis par un champignon du sol *Polymyxa betae*), des résistances à plusieurs maladies fongiques comme le rhizoctone brun et la *Cercosporiose*, ou encore la limitation de la multiplication du nématode à kyste (Desplanque 1999).
De nos jours, l'objectif de la sélection des betteraves sucrières est triple. Il s'agit de développer des variétés stables donnant le rendement le plus élevé en sucre blanc par hectare, tout en réduisant les coûts de production et en s'inscrivant dans une agriculture agroécologique. La réduction des coûts de production est un critère d'autant plus important que la demande en sucre subit actuellement une baisse importante, entraînant une baisse du prix de vente. Les caractères soumis à la sélection peuvent être répartis en trois catégories (Bosemark 1979) :

— les caractères morphologiques, qui impactent la récolte ou la transformation industrielle comme la taille, la forme, l'enterrage, la fibrosité de la racine ou encore la quantité de terre attenante.

— les caractères physiologiques, qui représentent la résistance à la montée à graine, la capacité de germination et de levée au champ, la résistance aux maladies et aux parasites et la résistance ou encore la tolérance à différentes conditions de stress abiotiques.

— les caractères chimiques, qui ont une influence sur l'extractibilité du sucre blanc, comme la teneur en sucre brut et la teneur en impuretés, à savoir la concentration

en sodium, en potassium, ou encore la teneur en azote alpha-aminé liée à la teneur en protéines solubles (à éliminer lors de l'extraction grâce à la chaux).

En 2019, 385 variétés sont inscrites au Catalogue officiel des espèces et variétés de plantes cultivées en France, maintenu par le GEVES.

1.2 Méthodes et outils pour la sélection variétale

La sélection variétale, ou amélioration des plantes, consiste à créer de nouvelles variétés répondant aux besoins de l'Homme tout en respectant les contraintes liées à l'environnement. Selon l'Union internationale pour la Protection des Obtentions Végétales (UPOV), une variété est une population sélectionnée par l'Homme qui doit « être reconnaissable à ses caractères, différer notablement de toute autre variété et demeurer inchangée au cours du processus de reproduction ou de multiplication » . L'objectif des sélectionneurs est donc de créer de nouvelles variétés regroupant le maximum d'allèles favorables pour les caractères recherchés. Afin de favoriser les combinaisons favorables d'allèles il est possible d'améliorer la population de pre-breeding d'où sont extraits les géniteurs utilisés dans les programmes de création variétale (Bouquet et al. 1981). Le but est ici d'augmenter le niveau moyen de la population pour le ou les caractères étudiés tout en préservant la variabilité génétique de départ. Afin de créer ces nouvelles variétés, les sélectionneurs peuvent s'appuyer sur différentes méthodes.

1.2.1 Principales méthodes de sélection phénotypique utilisées pour l'amélioration de la betterave sucrière

Sélection massale

La sélection massale est la première méthode de sélection utilisée dans l'histoire de l'Agriculture, à l'origine de la domestication des plantes. Les plantes les plus performantes vis-à-vis des caractères d'intérêt sont sélectionnées, et leurs semences sont utilisées pour produire la génération suivante. Répéter cette sélection sur plusieurs générations permet ainsi d'augmenter la fréquence des allèles favorables et donc d'améliorer peu à peu la valeur moyenne de la population (Figure 1.6). La sélection massale ne fait appel à aucune notion de génétique.

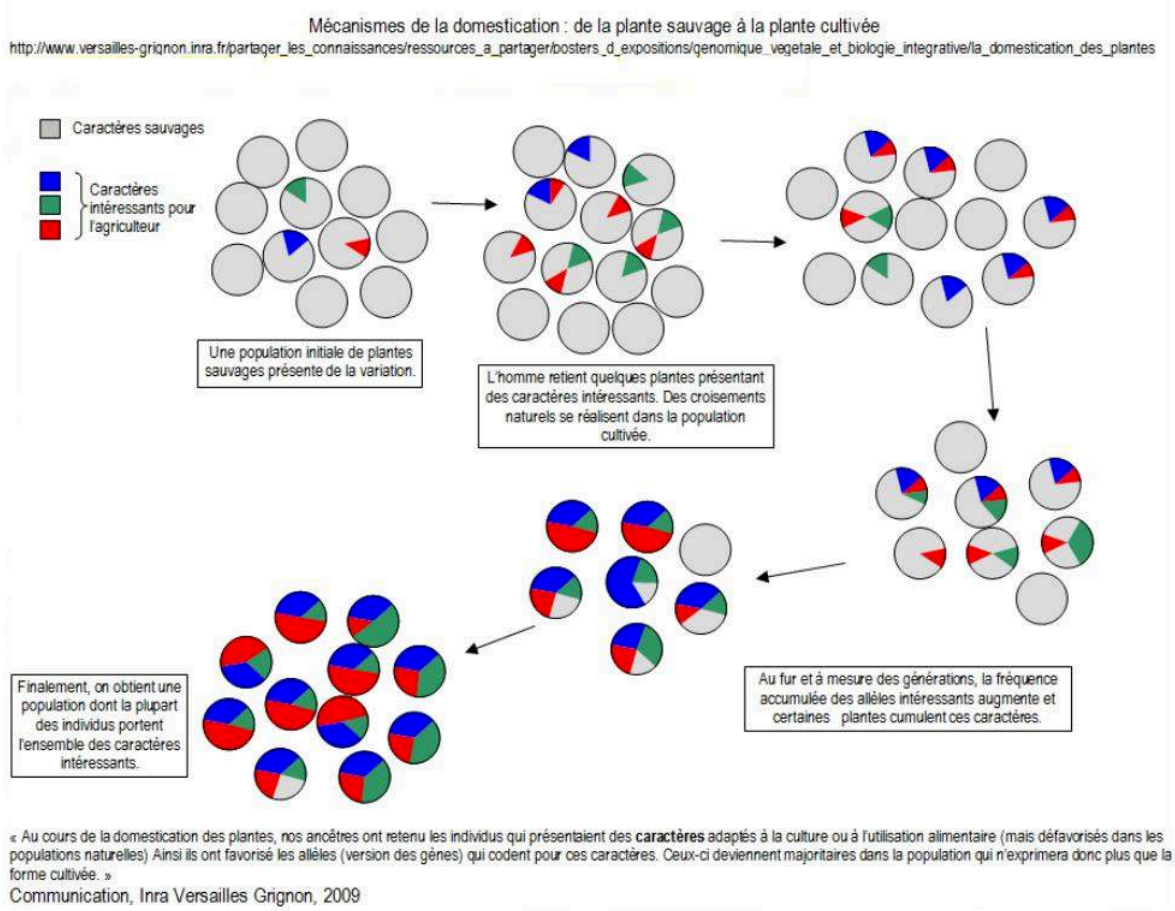


FIGURE 1.6 – Mécanisme sous-jacent de la domestication (Source : INRA Versailles Grignon (2009))

La sélection massale a été utilisée pour améliorer la teneur en sucre des betteraves dès 1786, ce qui a conduit à l'émergence de la betterave sucrière et à la construction de la première sucrerie. Cette sélection a d'abord été basée sur des caractères morphologiques, tels que la couleur blanche, l'enterrage et le sillon sacharifère, avant d'être axée sur la densité de la racine des betteraves, caractère très lié avec la teneur en sucre. La densité a d'abord été mesurée par immersion dans différents bains de sel, mesure qui permettait de déterminer le prix de vente de la betterave, puis par réduction avec des sels de cuivre. La betterave sucrière a ainsi été la première plante à être soumise à une sélection chimique systématique (Desprez and Desprez 1993).

Sélection généalogique

La sélection généalogique a pour objectif de créer des variétés en sélectionnant les meilleurs individus issus d'un croisement tout en identifiant les filiations correspondantes. Cette méthode a été développée par Pierre-Louis Lévêque de Vilmorin pour améliorer la productivité de la betterave sucrière, et ce dès 1856 (Vilmorin 1856). Ce sélectionneur a été le premier à établir que pour apprécier la valeur d'une plante, il fallait en étudier la

descendance. Il établit ainsi les bases de la sélection généalogique près de 10 ans avant les travaux de Mendel (1865).

Sélection hybride et hétérosis

Pour les plantes allogames comme la betterave sucrière, la consanguinité induit une forte perte de vigueur (Darwin 1877). Pour améliorer les performances des populations allogames, il est utile de créer des variétés hybrides afin de bénéficier de l'effet d'hétérosis (Shull 1914). La vigueur hybride, ou hétérosis, peut être définie comme la supériorité du phénotype d'un hybride par rapport à la moyenne phénotypique de ses deux parents, ou par rapport au meilleur de ses deux parents (Goldman 1999). En betterave, les hybrides sont produits par le croisement de deux parents : un parent femelle (mâle stérile), et un parent pollinisateur (hermaphrodite et utilisé comme mâle). Il existe donc deux pools hétérotiques distincts : un pool contenant les plantes mâles stériles ainsi que leurs mainteneurs, appelé pool femelle, et un pool comprenant les pollinisateurs, nommé pool mâle. Deux résultats complémentaires permettent de choisir les meilleurs parents à croiser : l'aptitude générale à la combinaison (AGC), définie comme le comportement moyen du parent dans toutes les combinaisons hybrides générées (Le Cochec and Soreau 1982), et l'aptitude spécifique à la combinaison (ASC) qui représente, pour un croisement donné, la déviation entre la valeur observée d'un hybride et sa valeur prédite en fonction de l'AGC de ses parents. Le Cochec and Soreau (1982) expliquent que l'AGC est le résultat des effets additifs des gènes et de leur interaction, tandis que l'ASC dépend des effets non additifs des gènes. Plusieurs études répertoriées par Le Cochec and Soreau (1982) démontrent que la variance de l'AGC est plus importante que celle de l'ASC dans l'expression du poids des racines et de la teneur en sucre des betteraves sucrières, c'est donc sur cette valeur que seront choisis les parents des hybrides. Avant la découverte de la stérilité mâle cytoplasmique chez la betterave sucrière par Owen (1945) le contrôle de l'hybridation n'était pas possible à grande échelle, la castration de cette plante demandant une grande minutie. La découverte de cette stérilité a donc révolutionné la façon de sélectionner la betterave sucrière, première espèce pour laquelle la stérilité mâle a été utilisée à grande échelle (Desprez and Desprez 1993).

Sélection récurrente

La sélection récurrente reprend le principe de la sélection massale dans le sens où la valeur moyenne d'une population vis-à-vis d'un caractère est améliorée par la sélection des meilleurs individus qui deviennent les parents de la génération suivante. Cependant

le critère de sélection n'est pas l'expression du caractère étudié, mais l'aptitude à la combinaison des individus. Les plantes d'une population S0 de départ sont d'abord autofécondées, générant une descendance S1 (Figure 1.7). Un test d'aptitude à la combinaison est effectué en croisant les individus S1 avec un testeur. Les meilleurs S1 sélectionnés sont croisés pour donner une nouvelle population de départ. Le nombre de cycles de sélection est variable, il est possible à tout moment d'extraire des individus pour produire des lignées ou des hybrides, et donc de créer des variétés, ou au contraire d'ajouter de nouvelles plantes à la population de départ pour réintroduire de la variabilité génétique. La capacité de limiter la perte de variabilité génétique au cours du processus de sélection fait de cette méthode un bon choix pour travailler sur des caractères au déterminisme génétique complexe.

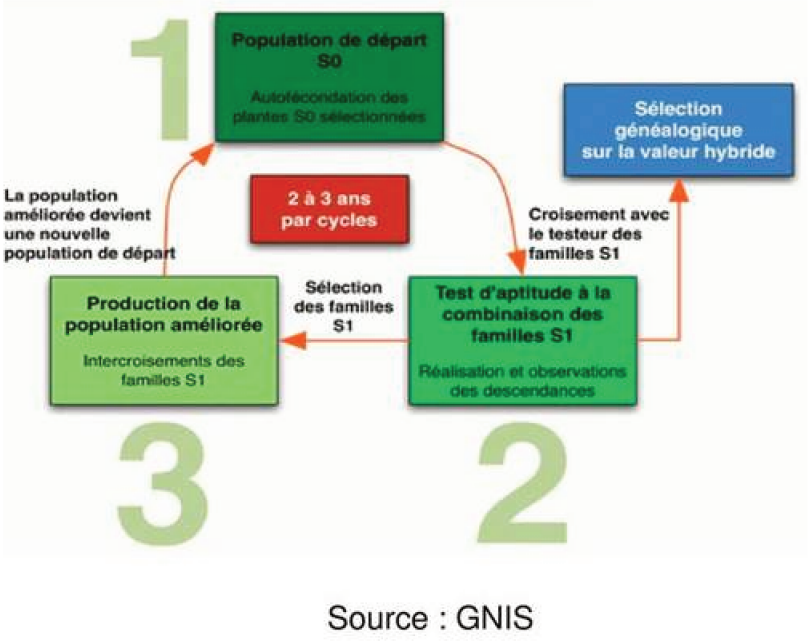


FIGURE 1.7 – Principe de la sélection récurrente classique (Source : GNIS)

Ce type de schéma de sélection, basé sur l'alternance des phases de sélection et de recombinaison, nécessite de pouvoir générer des populations issues de pollinisations aléatoires. Or l'introduction de l'allèle d'autofertilité (SF) découvert par Owen (Owen and Ryser 1942) a rendu la betterave sucrière autofertile, les lignées ne se croisent alors que difficilement, sous l'action d'un sélectionneur. Doggett and Eberhart (1968) ont l'idée d'utiliser la stérilité mâle cytoplasmique récessive du sorgho, espèce également autofertile, afin d'appliquer la sélection récurrente aux plantes autogames. En créant une population à l'équilibre, c'est-à-dire à moitié mâle stérile et moitié mâle fertile (Figure 1.8), et en ne récoltant les graines que sur les individus mâles stériles, il s'assure de générer des populations issues de pollinisations aléatoires en évitant le recours à une pollinisation contrôlée par les sélectionneurs, beaucoup plus laborieuse. Les probabilités

de recombinaison augmentent au fur et à mesure que les populations se croisent de façon aléatoire, cycle après cycle (Figure 1.9). Owen ayant découvert un gène nucléaire de stérilité mâle récessif sur la betterave sucrière (Owen 1952), un schéma de type Doggett (Doggett and Eberhart 1968) peut ainsi également être utilisé sur cette plante.

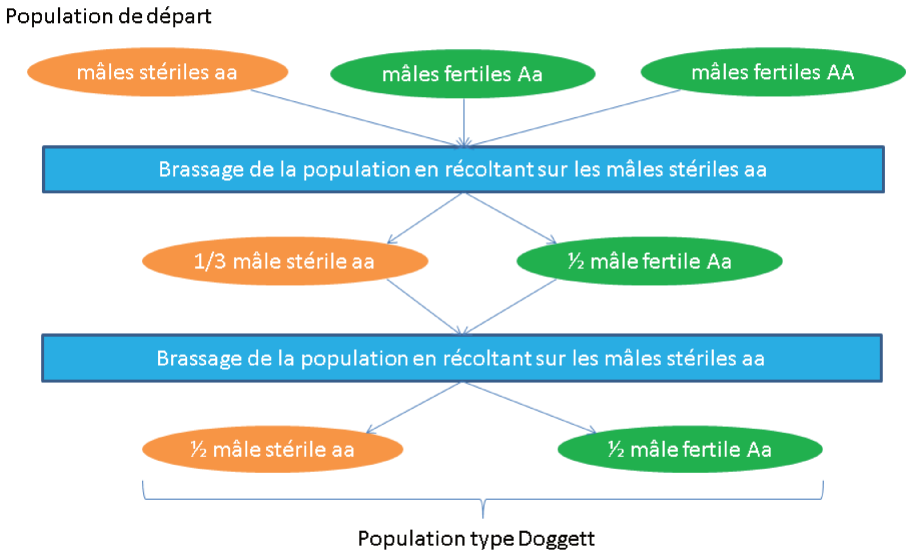


FIGURE 1.8 – Création d'une population de type « Doggett »

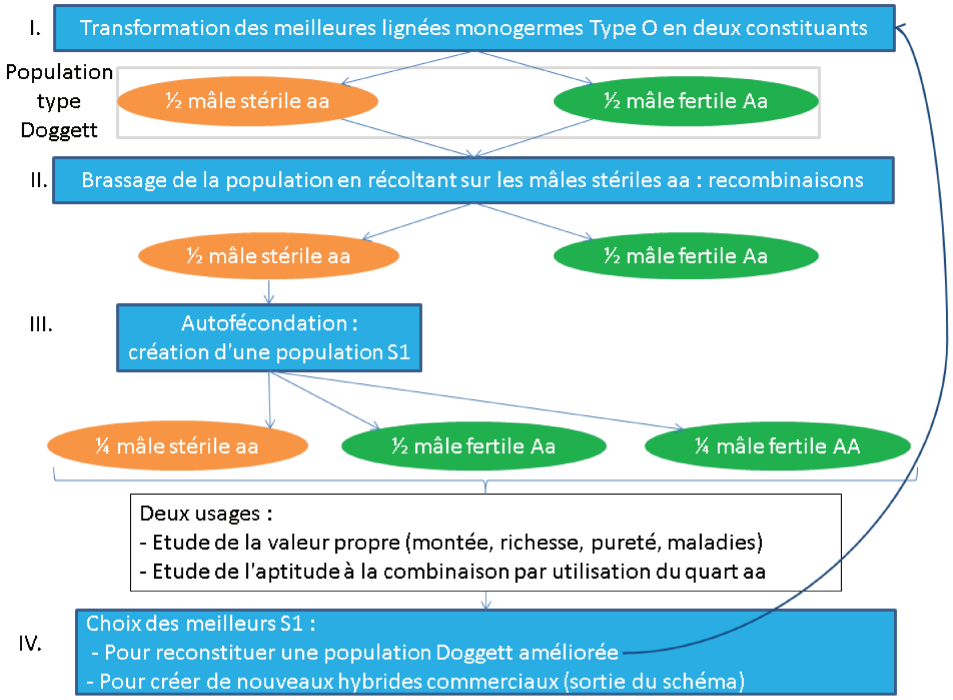


FIGURE 1.9 – Sélection récurrente selon la méthode de Doggett (Inspiré de Desprez and Desprez (1993))

1.2.2 Méthodes de sélection basées sur les marqueurs moléculaires

Jusque dans les années 1970, la sélection reposait uniquement sur l'appréciation de la valeur génétique des plantes au travers de leurs performances, les régions géniques impliquées dans les variations phénotypiques observées n'étant pas connues (Avise 2012). L'avènement des biotechnologies et notamment des marqueurs moléculaires a doté les sélectionneurs d'outils permettant l'étude des bases génétiques des caractères et ainsi de développer de nouvelles méthodes de sélection permettant d'accélérer les schémas de sélection. Les marqueurs moléculaires sont des séquences d'ADN qu'il est possible d'identifier spécifiquement et qui permettent de connaître partiellement le génotype d'un individu. Des marqueurs moléculaires basés sur le changement ponctuel d'une seule base dans une séquence donnée ont été développés. Appelés SNPs (Single Nucleotid Polymorphism), ces marqueurs généralement bi-alléliques permettent d'appréhender des variations génomiques impliquées dans l'expression d'un caractère étudié. Il est aujourd'hui possible de génotyper rapidement plusieurs dizaines de milliers de SNPs d'un individu, et ce pour un coût modéré. Grâce à cette densité de données génomiques disponibles, différentes approches ont été développées afin d'identifier les régions génomiques impliquées dans l'expression d'un caractère par le biais de la détection de QTL, ou de s'affranchir d'une partie du phénotypage en sélectionnant les individus sur leur valeur génétique prédite à partir des données génétiques.

Analyses d'association

L'une des approches permettant de détecter les régions génomiques liées au caractère d'intérêt, appelées QTL (Quantitative Trait Loci), est l'analyse d'association. Cette méthode a tout d'abord été utilisée en génétique humaine (Corder et al. 1993), avant d'être appliquée sur des données génétiques animales et végétales (Yu et al. 2006), (Zhou et al. 2012), (Kang et al. 2008), (Wang et al. 2016). A partir des données génotypiques et phénotypiques d'un panel d'individus d'une même espèce, l'analyse d'association permet d'évaluer la corrélation statistique entre les allèles à un locus donné et les valeurs phénotypiques observées ou mesurées. Plusieurs contraintes sont à prendre en compte dans le test statistique, notamment la présence d'une structure dans la population étudiée. Cette structure peut avoir une origine ancienne, découlant de phénomènes de mutations, de dérive génétique ou de migration qui n'affectent pas toute la population étudiée et qui génèrent ainsi des sous-groupes. A certains locus, les fréquences alléliques sont déséquilibrées entre les différents sous-groupes. Si le caractère

d'intérêt est principalement exprimé dans l'un des sous-groupes de la population, alors le test statistique indiquera une association statistiquement significative entre la variabilité du caractère et les locus présentant un polymorphisme entre les différents sous-groupes : il s'agit d'une confusion entre les effets des allèles aux marqueurs et l'appartenance aux sous-populations. Ces locus détectés comme associés au caractère étudié sont en réalité de faux positifs puisqu'ils ne sont pas (ou pas forcément) liés au locus causal expliquant la variabilité du caractère. Une autre contrainte à intégrer au test statistique est l'apparentement entre les individus. En effet, des individus apparentés présentent une partie de leur génome identique par descendance et forment ainsi un sous-groupe. Cela peut à nouveau engendrer une confusion entre les effets des allèles aux marqueurs et l'appartenance à un sous-groupe. Pour éviter les faux positifs décrits ci-dessus il est nécessaire de contrôler la structure et l'apparentement entre les individus dans les analyses d'association (Pritchard et al. 2000), et donc de ne pas considérer que les individus sont indépendants. Un modèle linéaire mixte est utilisé pour introduire une structure de covariance entre les variables aléatoires associées à chacun des individus (Yu et al. 2006) :

$$y_i = \mu + c_i + x_i^l \theta_a^l + u_i + e_i$$

où y_i est le phénotype y de l'individu i, c_i est le sous-groupe auquel appartient l'individu i, x_i^l est le génotype de l'individu i au locus l, θ_a^l est le vecteur des effets additifs au locus l, u_i est l'effet aléatoire du fond génétique de l'individu i avec $\boldsymbol{u} \sim \mathcal{N}(0, \sigma_u^2 \boldsymbol{K})$, $\boldsymbol{K}$ étant la matrice d'apparentement entre les individus et σ_u^2 la variance génétique et e_i est l'erreur résiduelle avec $\boldsymbol{e} \sim \mathcal{N}(0, \sigma_e^2 \boldsymbol{Id})$, $\boldsymbol{Id}$ étant la matrice d'identité et σ_e^2 la variance résiduelle. Les marqueurs moléculaires sont rarement positionnés exactement sur le locus causal du caractère étudié. Le principe des études d'association est donc de détecter les marqueurs en déséquilibre de liaison (DL) avec ce locus causal. Le déséquilibre de liaison représente la relation entre deux locus différents. Ce déséquilibre est dit complet si le fait de connaître le génotype au premier locus permet de déduire le génotype au second locus. A l'inverse, il n'existe pas de déséquilibre de liaison entre deux locus si aucun lien génomique ne peut être fait entre eux (Rafalski 2002). L'existence et l'étendue du DL résultent des phénomènes survenus pendant l'évolution de la population étudiée, tels que les mutations ou les recombinaisons génétiques successives. Le génotype au marqueur détecté grâce aux études d'association est ainsi lié au génotype responsable de la variation du caractère d'intérêt. L'étendue du DL est un indicateur de la densité de marqueurs nécessaires pour détecter les marqueurs en DL avec le locus causal. Comme représenté sur la Figure 1.10, l'analyse d'association en cas de DL très étendu permet de détecter des marqueurs associés au caractère même si ces marqueurs sont éloignés du locus causal. En revanche, un DL

peu étendu implique d'avoir une densité de marquage plus forte pour pouvoir détecter un marqueur en DL avec le locus causal. La position du locus causal est de ce fait mieux estimée en cas de DL à faible étendue.

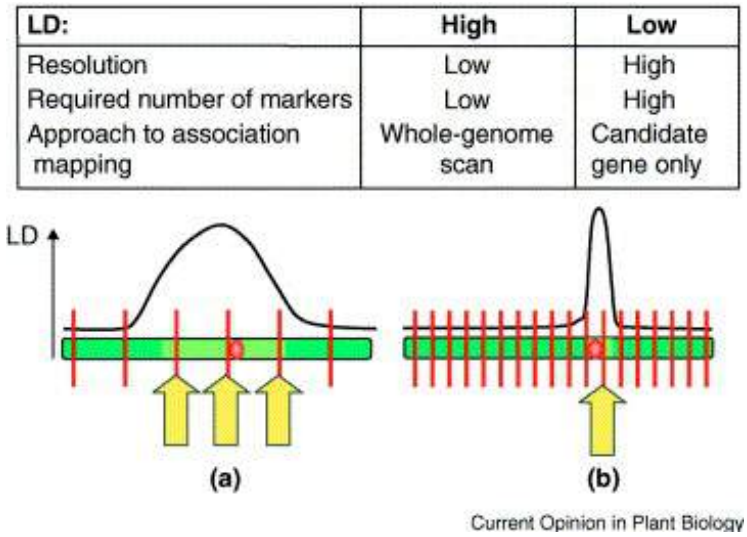


FIGURE 1.10 – Relation entre l'étendue du DL et la résolution des études d'association, d'après (Rafalski 2002). En (a) le DL est étendu autour du locus causal responsable du phénotype (ovale rouge) sur un chromosome. Dans ce cas, même une faible densité de marqueurs (représentés par des barres verticales rouges) est suffisante pour identifier des marqueurs associés (flèches jaunes) au locus causal. En (b), le DL diminue très rapidement autour du locus causal, et une densité beaucoup plus grande de marqueurs est nécessaire pour identifier un marqueur associé (flèche jaune).

Sélection Assistée par Marqueur

La Sélection Assistée par Marqueurs (SAM) consiste à utiliser les marqueurs moléculaires identifiés comme étant liés à un caractère d'intérêt, c'est-à-dire proches d'un gène d'intérêt pour ce caractère, pour détecter la présence de l'allèle favorable dans les plantes à sélectionner et ce à un stade précoce. Cette méthode permet d'accélérer les schémas de sélection car il n'est plus nécessaire d'attendre que la plante exprime le caractère pour pouvoir l'évaluer, un petit fragment de feuille est suffisant pour la génotyper. Cette méthode a principalement été utilisée pour introduire des résistances à une maladie, résistances majoritairement codées par un gène majeur. La résistance à la rhizomanie a ainsi été introduite dans les variétés de betterave sucrière cultivées (Barzen et al. 1997). La SAM est cependant souvent inefficace pour la sélection de caractères complexes, contrôlés par un grand nombre de gènes.

Sélection Génomique

La méthode de sélection génomique consiste à prédire la valeur génétique des individus d'une population à partir de leurs informations génomiques, et à sélectionner les individus en fonction de cette valeur. Contrairement à la sélection assistée par marqueurs qui utilise uniquement l'effet des marqueurs associés au caractère étudié, la sélection génomique prend en compte l'effet de tous les marqueurs disponibles. Si les marqueurs couvrent l'ensemble du génome, tous les QTLs sont en DL avec au moins un marqueur. Considérer l'ensemble des marqueurs permet ainsi de capturer l'effet de tous les QTLs, même ceux dont l'effet est faible (Schulz-Streeck et al. 2012). Afin d'estimer l'effet de chacun des marqueurs, il est nécessaire de connaître la relation entre le génotype et le phénotype des individus. Un modèle de prédiction peut être établi par apprentissage sur une population d'entraînement pour laquelle le génotype et le phénotype de chacun des individus est connu. Meuwissen et al. (2001) ont été les premiers à proposer de faire une régression des phénotypes sur les marqueurs génétiques en utilisant un modèle de régression paramétrique simple. Il s'agit des débuts de la sélection génomique. Le modèle linéaire mixte utilisé, appelé RR-BLUP pour *Ridge Regression - Best Linear Unbiaised Predictor*, suppose que chacune des régions génomiques est représentée par un seul marqueur et que les effets de l'ensemble de ces régions suivent une même distribution Gaussienne, expliquant la variabilité du phénotype dans la population. L'effet des marqueurs est alors considéré comme aléatoire (Whittaker et al. 2000). Depuis, de nombreux modèles ont été proposés dans la littérature mais le plus robuste semble être le GBLUP (*Genomic Best Linear Unbiaised Predictor*) proposé par VanRaden (2008). Le GBLUP est également un modèle linéaire mixte dans lequel la variabilité des phénotypes dans une population est cette fois expliquée par la covariance attendue entre les individus pour les effets génétiques additifs. Cette covariance est supposée être proportionnelle à la matrice d'apparentement entre individus, elle même estimée à partir de l'ensemble des marqueurs du génome sur l'ensemble des individus. L'effet des individus est ici considéré comme aléatoire. Malgré cette différence dans la modélisation, les modèles RR-BLUP et GBLUP sont en réalité équivalents puisque les effets des marqueurs dans le modèle RR-BLUP capturent l'apparentement génomique utilisé dans le GBLUP (Hayes et al. 2009). Goddard (2009) démontre que ces deux modèles sont semblables aux modèles mixtes classiques de pedigree si l'apparentement entre les individus est estimée à partir des marqueurs moléculaires. Ce modèle linéaire standard considère que la réponse phénotypique de l'individu i (Y_i) est expliquée par un facteur commun à tous les individus (μ), par un facteur génétique spécifique à l'individu i (u_i) qui peut être décrit par la somme des effets des marqueurs moléculaires et par une résiduelle comprenant tous les autres effets non génétiques

(e_i) comme les effets de l'environnement ou encore du design expérimental. Ce modèle génétique peut s'écrire de la façon suivante :

$$y_i = \mu + u_i + e_i$$

Würschum et al. (2013) ont étudié l'efficacité de la sélection génomique sur un panel de betteraves sucrières et ont montré que cette méthode était un outil intéressant à utiliser dans les programmes de sélection de betterave sucrière.

1.3 Projet AKER et enjeux de la thèse

1.3.1 AKER

L'accumulation des caractères prérequis à la commercialisation des variétés de betterave à sucre tels que le rendement en sucre par hectare, la résistance à des maladies de plus en plus nombreuses sur le cahier des charges, ou encore la diminution des intrants comme les intrants azotés, a eu pour conséquence de réduire considérablement la variabilité génétique disponible dans les programmes de pre-breeding. Or c'est cette diversité génétique utile qui permet de créer de nouvelles variétés toujours plus performantes vis-à-vis de tous ces critères. Le programme AKER (`http://www.aker-betterave.fr/en/`) a pour objectif d'améliorer la compétitivité de la betterave sucrière française en élargissant la diversité génétique du germoplasme élite grâce à une approche originale d'utilisation des ressources génétiques exotiques, permettant l'enrichissement du matériel de sélection par de nouveaux allèles d'intérêt. Il s'agit d'un Programme d'Investissement d'Avenir, initié par l'Etat français dans le cadre de l'Agence Nationale de la Recherche. Mis en place en 2012 pour une durée de huit ans, il est porté par onze partenaires publics et privés de la filière betterave-sucre-alcool française. Le premier objectif du projet est d'élargir la variabilité génétique de la betterave sucrière en constituant une collection de gènes en provenance de ressources du monde entier. Pour ce faire, une cinquantaine de banques de gènes possédant des accessions de betteraves cultivées ou sauvages ont été analysées à l'aide d'outils de marquage moléculaire. 10 000 accessions différentes ont ainsi été référencées. Sur la base de données passeport répertoriant le type de betterave ainsi que la latitude et la longitude de leur lieu d'origine, 3000 accessions ont été géographiquement retenues pour représenter le maximum d'origines possibles. Pour chacune de ces accessions, l'ADN d'une plante a été étudié à l'aide de très nombreux marqueurs moléculaires : plus de 30 millions de données moléculaires (marqueurs SNP et DArT) ont été analysées. Après analyse de la diversité allélique des marqueurs, 40 accessions se sont révélées

suffisantes pour représenter l'ensemble de la variabilité disponible chez les betteraves. De plus, seulement 16 d'entre elles sont suffisantes pour représenter l'ensemble de la diversité allélique qui n'est pas déjà présente au sein des betteraves sucrières cultivées. La carte en figure 1.11 donne la localisation de ces différentes accessions.

FIGURE 1.11 – Localisation des différentes accessions exotiques identifiées dans le projet AKER comme représentant le maximum de la diversité génétique non présente dans le matériel de breeding

FIGURE 1.12 – Accessions exotiques utilisées dans le programme AKER

Les accessions exotiques sont les suivantes (Figure 1.12) :
- *Beta macrocarpa* (Espagne, Maroc)
- *Beta vulgaris maritima* (France × 3, Portugal, Royaume-Uni, Danemark, Irlande, Grèce, Italie × 2)
- *Beta vulgaris vulgaris leaf beet* (Grèce)
- *Beta adanensis* (Israël)
- *Beta vulgaris sugar beet* (Chine, USA)
- *Beta vulgaris fodder beet* (France)

Chacune des 16 accessions a été séquencée. Une descendance d'environ 200 individus, issue du croisement entre une accession exotique et une betterave sucrière élite et suivi de deux rétrocroisements successifs ainsi que d'une autofécondation, a été générée pour chaque accession exotique. Les plantes obtenues (théoriquement 16 × 200) possèdent ainsi une mosaïque de fragments d'origine exotique dans un contexte élite. Ces plantes ont été observées et phénotypées (évaluées) lors des deux dernières années du programme AKER (2018 et 2019) sur plusieurs lieux d'essais.
L'objectif suivant est d'évaluer chacun de ces fragments exotiques pour leur attribuer une valeur positive, négative ou neutre selon le caractère étudié. A terme, les 16 populations de référence complètes seront donc disponibles et utilisables rapidement dans les schémas de sélection comme sources de diversité. Leurs génomes seront connus et déjà caractérisés pour bon nombre de caractères utiles aux objectifs de sélection actuels.

1.3.2 SELKIT

Cette thèse constitue le projet SELKIT, pour KIT de SELection génomique, un projet connexe au programme AKER. Le but est de favoriser l'introgression des ressources génétiques exotiques découvertes dans le cadre d'AKER dans un programme de pre-breeding en utilisant la sélection génomique. L'objectif est d'anticiper la fin du projet AKER, en réfléchissant au meilleur schéma d'utilisation des résultats qui en ressortiront.
A l'heure où commence ce projet, cette problématique d'introgression de nouvelles ressources génétiques dans du matériel élite a déjà été étudiée dans la littérature. L'état de l'art réalisé pour constituer le dossier de thèse CIFRE met en exergue les éléments suivants. « *La solution principalement proposée a surtout fait ses preuves dans des cas réels pour l'introgression de peu de gènes/QTLs dans un génome récurrent. Des schémas d'introgression de gène unique ont été entrepris dans de nombreuses espèces comme l'orge (Jefferies et al. 2003), le blé (Vishwakarma et al. 2014), ou encore le riz (Xu and Crouch 2008) et leur intérêt a été démontré. Jefferies et al. (2003) ont démontré que la sélection*

assistée par marqueurs peut être utilisée efficacement pour introgresser la résistance à BYDV grâce au gène Yd2 dans un programme de rétrocroisement chez l'orge. La résistance obtenue s'exprime en une réduction significative des symptômes foliaires et des pertes de rendement en milieu infecté, comparée au génotype élite initial. De leur côté, Vishwakarma et al. (2014) ont effectué le transfert du gène Gpc-B1, lié à la concentration en protéines dans le grain, ainsi que la reconstitution du génome élite récurrent sur un laps de temps de 2 ans et demi (5 cycles de sélection), démontrant l'intérêt du rétrocroisement assisté par marqueurs pour développer des lignées élites à fortes teneurs en protéines dans le grain. Des schémas d'introgression multi-gènes plus complexes ont été suggérés pour des QTLs préalablement identifiés (Hospital et al. 2000), (Servin et al. 2004), (Piyasatian et al. 2008). Servin et al. (2004) proposent notamment une optimisation des schémas de sélection pour cumuler plusieurs gènes identifiés dans différents parents au sein d'un même génotype. Il existe 2 schémas majeurs de rétrocroisements assistés par marqueurs : le premier est idéal pour l'introgression de gènes majeurs et nécessite entre 2 et 10 marqueurs pour chaque gène cible. Le second permet d'introgresser des gènes majeurs tout en contrôlant le fond génétique et nécessite entre 2 et 10 marqueurs par gène introgressé et environ 200 marqueurs pour le retour au fond génétique élite (Xu et al. 2012). Cependant, ces méthodes d'introgression montrent certaines limites dès que plusieurs QTLs à effets faibles tentent d'être intégrés dans un même génotype élite (Hospital 2005). En effet, c'est essentiellement la détection de QTLs préalable à l'introgression qui est limitante puisqu'elle ne détecte pas ou peu de QTLs à effets faibles. Or, capturer seulement une portion de la variance génétique peut aboutir à une surestimation des effets des marqueurs (Lande and Thompson 1990), (Beavis 1998), qui peuvent ne pas être valables dans d'autres fonds génétiques une fois transférés, dans d'autres environnements ou après plusieurs cycles de sélection (Podlich et al. 2004). En betterave sucrière comme dans d'autres espèces, les caractères qui intéressent le plus les sélectionneurs (rendement en sucre blanc chez la betterave, rendement en grains chez les céréales...) sont des caractères polygéniques gouvernés par une multitude de fragments à effets faibles, et les méthodes présentées précédemment ne peuvent donc pas être utilisées efficacement pour intégrer une multitude de fragments génomiques d'intérêt. La sélection génomique offre une méthode alternative dans laquelle les variations génétiques favorables peuvent être sélectionnées à travers l'ensemble des génomes et les variations délétères contre-sélectionnées, sans se concentrer sur un nombre fini de régions génomiques, ce qui est particulièrement intéressant pour les caractères génétiquement complexes. Les programmes d'introgression « classiques » vont avoir pour but d'introgresser un ou plusieurs allèles d'intérêt provenant de lignées donneuses, tout en réduisant simultanément la quantité d'ADN donneur au minimum. Ici, la sélection génomique ne diminuera pas nécessairement la quantité d'ADN issue du

donneur, mais devrait aboutir à une augmentation de la fréquence des allèles favorables, sans tenir compte de leur origine (donneur ou récurrent). La sélection génomique est une méthodologie récente, développée au début des années 2000 chez les animaux (Meuwissen et al. 2001), et mise en pratique quelques années plus tard chez les plantes (Bernardo and Yu 2007). Quelques travaux d'utilisation de la sélection génomique en pre-breeding pour intégrer de la variabilité génétique au matériel élite ont déjà été publiés sur des données simulées (Bernardo 2009), (Ødegård et al. 2009b), (Ødegård et al. 2009a) ; (Gorjanc et al. 2016), mais ce n'est que très récemment que cette technique a été appliquée sur des données réelles (Combs and Bernardo 2013). Plus l'héritabilité du caractère à transférer est faible, plus la sélection génomique va permettre un gain génétique par rapport aux méthodes classiques d'introgression. Combs and Bernardo (2013) montrent, par exemple, que la sélection génomique permet d'obtenir de meilleures performances tout en conservant une proportion plus importante de génome exotique introduit, comparée au schéma de rétrocroisement phénotypique. Gorjanc et al. (2016) proposent, pour leur part, de mettre en place un programme de pre-breeding à partir d'un panel de diversité exotique maïs. Leurs travaux suggèrent qu'un programme de ce type doit être directement initié à partir des lignées exotiques et non pas à partir de croisements élite/exotique, pour éviter un retour trop rapide à l'élite. Dans le contexte d'un programme de pre-breeding, la sélection génomique peut donc être utilisée efficacement pour enrichir une population de départ avec des variations polygéniques favorables. Une telle population, enrichie et proche au niveau des performances agronomiques des population élites, peut ensuite être utilisée facilement comme source de croisements par les sélectionneurs. »

Différents schémas de pre-breeding utilisant la sélection génomique vont ainsi être simulés et comparés dans le projet SELKIT dans l'objectif de guider la production d'une population de pre-breeding. Cette population constituera un réservoir de diversité génétique utile dans lequel les sélectionneurs pourront puiser pour créer de nouvelles variétés de betteraves sucrières. Afin de simuler ces programmes de pre-breeding intégrant la sélection génomique, il est nécessaire de connaître l'architecture génétique des caractères agronomiques à simuler. L'architecture génétique d'un caractère correspond au nombre de régions génomiques impliquées dans l'expression du caractère, leur localisation sur le génome, et la proportion du caractère qu'elles expliquent. Ces informations peuvent être obtenues par le biais d'une détection de QTL, c'est-à-dire en recherchant les régions du génome qui sont liées au caractère agronomique étudié. Pour réaliser cette détection de QTL, il faut donc des données de génotype et de phénotype sur une population. Les travaux réalisés pendant cette thèse s'appuient sur un matériel génétique original développé au sein du projet AKER. Chacune des 16 accessions exotiques a été croisée avec un pollinisateur élite. Après deux backcrosses et une autofécondation, des individus

appelés BC2S1 sont obtenus (environ 200 BC2S1). Ces BC2S1 représentent le génome de l'accession exotique fragmenté au sein du germoplasme élite. Ces 16 descendances ont été génotypées, puis croisés avec une lignée élite du pool femelle pour créer des hybrides évalués dans six champs en 2018. Les premières données de phénotype ont été mises à disposition en janvier 2019. Ces données des populations AKER n'étaient donc pas disponibles au début de cette thèse (avril 2017). Les premiers travaux ont alors porté sur deux autres populations, afin de préparer les analyses qui seront par la suite appliquées aux données issues du programme AKER.

Ce manuscrit de thèse est divisé en cinq chapitres, dont le premier a présenté des généralités à propos de la betterave sucrière, les différentes méthodes de sélection utilisées pour améliorer cette plante, et l'objectif de cette thèse connexe au projet AKER. Dans le deuxième chapitre, le postulat fait dans le programme AKER selon lequel l'introduction de régions exotiques dans un programme de pre-breeding peut permettre d'apporter une diversité génétique utile est vérifié par la comparaison de l'architecture génétique de plusieurs caractères dans deux populations : une descendance (élite x exotique) similaire aux populations AKER, et un panel élite. Dans le troisième chapitre, l'architecture génétique du rendement racinaire, nécessaire à la simulation ultérieure de ce caractère dans différents schémas de pre-breeding, est étudiée dans les différentes descendances (élite $\times$ exotique) issues du programme AKER et les fragments exotiques apportant de la variabilité génétique utile sont identifiés. Le quatrième chapitre présente le schéma de pre-breeding tel qu'imaginé par les sélectionneurs de l'entreprise Florimond Desprez Veuve & Fils, sa traduction en langage informatique et les différents scénarios à simuler. Dans le chapitre 5, les différents indicateurs permettant d'évaluer et de comparer les différents scénarios de pre-breeding simulés sont détaillés, et leur évolution au cours des simulations est interprétée. Enfin, ce manuscrit s'achève par une conclusion globale sur ce travail de thèse.

Un aperçu de l'histoire de la betterave sucrière ainsi que des objectifs de cette thèse ont été présentés lors du concours de vulgarisation scientifique « Ma Thèse en 180 secondes » en janvier 2018 (`https://www.youtube.com/watch?v=bBDHRoK6BZU`).

Chapitre 2

Découverte de nouvelle diversité génétique par comparaison d'une descendance (élite × exotique) avec un panel élite

L'objectif de ce chapitre est de vérifier le postulat fait dans le programme AKER, c'est-à-dire déterminer si l'introduction de régions exotiques dans une population de betterave sucrière permet bien d'apporter de la variabilité génétique utile vis-à-vis des caractères à sélectionner pour améliorer la betterave sucrière. Pour ce faire, deux populations sont étudiées : la descendance d'une accession exotique croisée avec une betterave sucrière élite, ressemblant aux descendances générées dans le programme AKER, et un panel élite. Les données de ces deux populations seront tout d'abord présentées, puis la recherche de l'architecture génétique de sept caractères agronomiques pour chacune de ces deux populations sera détaillée, et les résultats seront comparés.

2.1 Données disponibles

2.1.1 Descendance (élite × exotique)

Deux populations sont ici étudiées : une descendance (élite × exotique), similaire aux descendances AKER, et un panel élite. L'obtention de ces populations est détaillée ci-dessous.

Matériel végétal

La descendance (élite × exotique) est composée de 187 individus qui proviennent de croisements entre une accession exotique *Beta maritima*, originaire du Danemark, et une betterave sucrière élite provenant de l'entreprise Florimond Desprez Veuve & Fils. L'accession exotique est tout d'abord croisée avec un pollinisateur élite non fixé, produisant la génération F1. Un premier backcross est réalisé entre les F1 et le pollinisateur élite, créant la génération BC1. Un second backcross est ensuite réalisé entre la génération BC1 et un second pollinisateur élite non fixé, générant la population BC2. Les BC2 sont alors autofécondés pour produire la génération BC2S1, aussi appelée descendance (élite × exotique). Ces croisements successifs permettent de fragmenter le génome issu de l'accession exotique de façon à ce que la descendance (élite × exotique) soit composée d'individus élites intégrant un fragment de génome exotique, tel qu'illustré dans la Figure 2.1. Lors des deux croisements back-cross successifs, seuls les individus mâle stériles sont choisis pour être fécondés par le pollinisateur élite. De plus, seuls les individus hétérozygotes au marqueur de stérilité mâle produisent la génération BC2S1. Ces sélections sont visibles sur le chromosome 1 où se trouve le marqueur de stérilité. Les génotypes homozygotes élites, homozygotes exotiques et hétérozygotes représentent respectivement 89,1%, 3,7% et 7,2% du génome de la descendance (élite × exotique).

Données génétiques

Une puce Axiom®35K a été développée en partenariat avec Affymétrix (`http://www.affymetrix.com`) pour génotyper 33 621 SNPs de haute qualité. Les détails concernant la constitution de cette puce sont présentés dans l'article Pegot-Espagnet et al. (2019), présenté dans la partie 2.2.1 de ce chapitre. Les marqueurs sur la puce sont conservés quand l'origine parentale de l'allèle peut-être déterminée sans ambiguïté, c'est-à-dire quand il est possible de distinguer les allèles provenant de l'accession exotique et les allèles apportés par les pollinisateur élites. Une carte génétique a été générée par Olivier Guillaume, ingénieur d'études en CDD sur le projet AKER. La création de cette carte est également détaillée dans l'article Pegot-Espagnet et al. (2019).

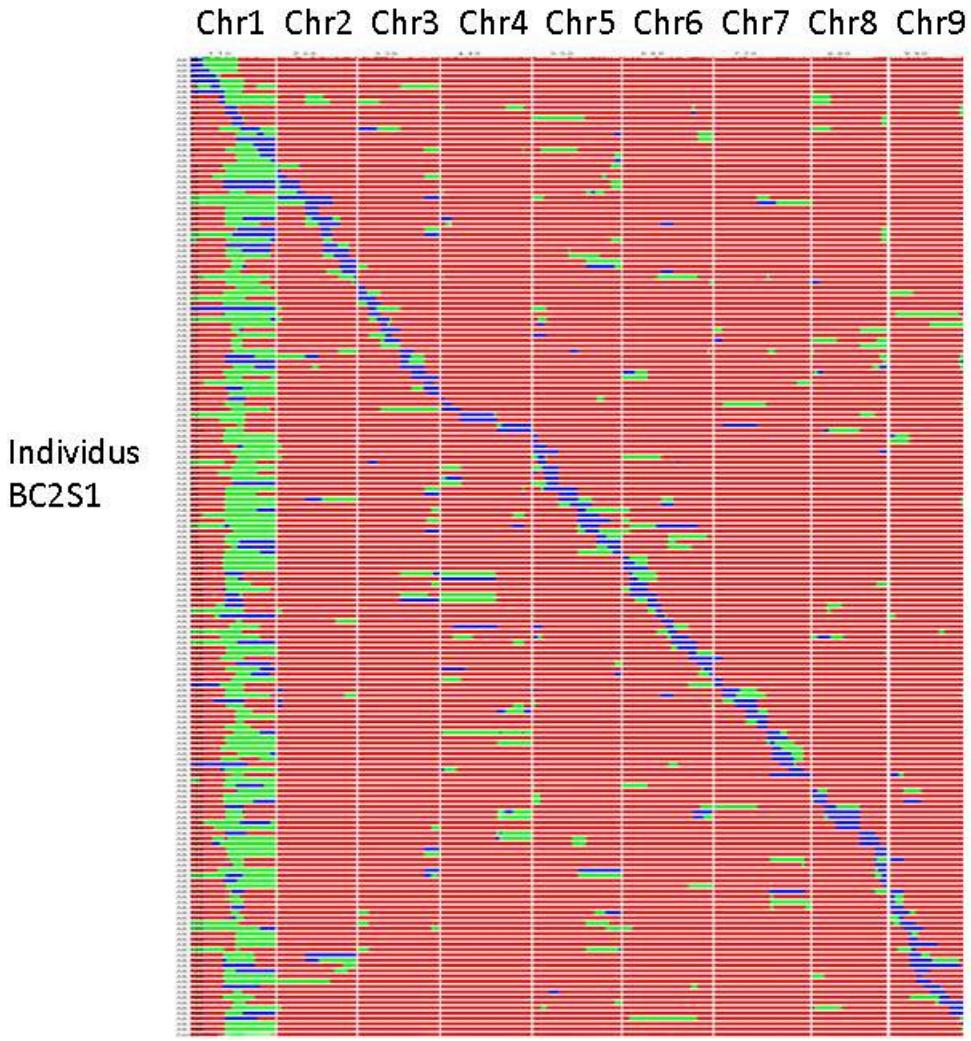


FIGURE 2.1 – Représentation du génome des individus de la descendance (élite × exotique) sur chacun des neuf chromosomes de la betterave sucrière (en colonne). Le génotype homozygote élite est représenté en rouge, le génotype exotique est représenté en bleu, et le génotype hétérozygote est représenté en vert.

Données phénotypiques

Les 187 descendants du croisement (élite × exotique) ont été évalués en test-cross dans neuf environnements en 2016 : sept champs au nord de la France, un en Belgique, et un en Angleterre. Chaque environnement comporte deux répétitions, composées des 187 génotypes ainsi que cinq témoins répétés plusieurs fois dans chaque essai. Une micro-parcelle correspond à un génotype. 80 à 90 betteraves de ce génotype sont plantées sur trois rangs, espacées de 20cm.

Sept caractères agronomiques sont évalués. Trois d'entre eux sont liés à des impuretés contenues dans les betteraves sucrières : le contenu en potassium (K, mesuré en meq/100 g de matière fraîche avec un photomètre de flamme), le contenu en sodium (Na, mesuré en meq/100 g de matière fraîche avec un photomètre de flamme) et le contenu en azote alpha-aminé (N, meq/100 g mesuré par colorimétrie). Ces impuretés sont étudiées car elles ont un impact négatif sur le rendement en sucre en gênant l'extraction de sucre

de betterave (Hoffmann 2010). Les quatre autres caractères sont liés au rendement. Le rendement racinaire (RY, mesuré en tonnes/hectare) et la teneur en sucre (S, mesurée par réfractométrie) sont des caractères mesurés. Les deux derniers caractères, le pourcentage en sucre blanc (WS) et le rendement en sucre blanc (WSY), sont calculés à partir des formules suivantes : $\mathrm{WS} = \mathrm{S} - (0.14 \times (\mathrm{K} + \mathrm{Na}) + 0.25 \times \mathrm{N} + 0.5)$ et $\mathrm{WSY} = (\mathrm{RY} \times \mathrm{WS})/100)$

2.1.2 Panel élite

Matériel végétal

La seconde population étudiée est constituée de 2 101 individus élites, représentatifs de la diversité génétique de l'ensemble des betteraves sucrières cultivées dans le monde. Ce panel élite a précédemment été étudiée dans l'article de Mangin et al. (2019). Les auteurs ont montré que le panel était structuré en deux sous panels, appelés panel A et panel B. Le panel A est composé de 676 accessions, et le panel B des 1425 accessions restantes. Cette structuration hiérarchique est prise en compte dans les études suivantes, bien que son origine ne soit pas identifiée à ce jour.

Données génétiques

Le panel élite a été génotypé à l'aide de 836 SNPs. Les détails concernant la conception de ces marqueurs sont présentés dans les données supplémentaires de l'article Mangin et al. (2019). La position génétique de ces SNPs a été déterminée selon la méthode décrite par Adetunji et al. (2014).

Données phénotypiques

Le panel élite a été évalué dans un dispositif test-cross en blocs incomplets sur plusieurs environnements en 2009, 2010 et 2011. Le détail de ce dispositif est décrit dans les données supplémentaires de l'article Mangin et al. (2019). Les sept caractères évalués pour la descendance (élite × exotique) sont également évalués pour ce panel élite. Les auteurs ont ajusté les données phénotypiques de chaque caractère en deux étapes : un ajustement des effets champ intra-environnement a tout d'abord été réalisé, puis le phénotype moyen sur l'ensemble des environnements a été calculé pour chaque génotype. Au début de cette étude, le phénotype moyen de chaque génotype pour chaque caractère est donc disponible.

2.2 Découverte de nouvelle diversité génétique intéressante en comparant une descendance (élite × exotique) et un panel élite

2.2.1 Etude et comparaison de l'architecture génétique des trois caractères d'impureté sur la descendance (élite × exotique) et sur le panel élite

L'article suivant, publié dans le journal TAG en juillet 2019, présente l'étude de l'architecture génétique de chacun des trois caractères d'impuretés : le contenu en potassium (K, meq/100 g), le contenu en sodium (Na, meq/100 g) et le contenu en azote alpha-aminé (N, meq/100 g) et ce sur chacune des deux populations : la descendance (élite × exotique) et le panel élite. L'architecture génétique d'un caractère est définie comme le nombre de régions génomiques impliquées dans l'expression du caractère, leur position sur le génome, et la part de variance génétique qu'elles permettent d'expliquer. Plusieurs étapes ont permis de déterminer cette architecture génétique pour chacune des deux populations. Ces étapes sont décrites dans l'article qui suit, mais peuvent être résumées de la façon suivante :

- Les données de génotype sont traitées de façon à imputer les données manquantes et à regrouper les SNPs redondants (c'est-à-dire portant la même information génétique) sous un seul SNP appelé « référent » .
- Le phénotype est analysé en suivant une méthode *two-step* : l'effet spatial de chacun des environnements sur le phénotype est d'abord ajusté, puis le phénotype moyen du caractère pour chacune des populations est calculé.
- Les SNPs liés au caractère étudié sont détectés grâce à une analyse d'association en utilisant notamment un modèle additif, puis en appliquant un critère de parcimonie permettant de ne retenir que les SNPs significativement associés à la variation du phénotype du caractère.
- Ces SNPs détectés sont regroupés en QTLs suite à une étude du déséquilibre de liaison (DL).

La comparaison de l'architecture génétique de ces caractères déterminée dans la descendance (élite × exotique) avec celle identifiée dans le panel élite permet de mettre en évidence les éventuelles régions génomiques apportées par l'accession exotique, et de déterminer si l'allèle exotique apporte un effet favorable. Les caractères étudiés étant liés aux impuretés, un effet favorable de l'allèle exotique peut ici être caractérisé comme une diminution de la teneur en impureté en présence de cet allèle.

Discovery of interesting new polymorphisms in a sugar beet (elite × exotic) progeny by comparison with an elite panel

Prune Pegot-Espagnet[1,2] · Olivier Guillaume[1] · Bruno Desprez[2] · Brigitte Devaux[2] · Pierre Devaux[2] · Karine Henry[2] · Nicolas Henry[2] · Glenda Willems[3] · Ellen Goudemand [2] · Brigitte Mangin[1]

Received : 21 December 2018/ Accepted : 24 July 2019

Abstract

Key message **The comparison of QTL detection performed on an elite panel and an (elite × exotic) progeny shows that introducing exotic germplasm into breeding programs can bring new interesting allelic diversity.** Selection of stable varieties producing the highest amount of extractable sugar per hectare (ha), resistant to diseases, and respecting environmental criteria is undoubtedly the main target for sugar beet breeding. As sodium, potassium, and α-amino nitrogen in sugar beets are the impurities that have the biggest negative impact on white sugar extraction, it is interesting to reduce their concentration in further varieties. However, domestication history and strong selection pressures have affected the genetic diversity needed to achieve this goal. In this study, quantitative trait locus (QTL) detection was performed on two populations, an (elite × exotic) sugar beet progeny and an elite panel, to find potentially new interesting regions brought by the exotic accession. The three traits linked with impurities content were studied. Some QTLs were detected in both populations, the majority in the elite panel because of most statistical power. Some of the QTLs were colocated and had favorable effect in the progeny since the exotic allele was linked with a decrease in the impurity content. A few number of favorable QTLs were detected in the progeny, only. Consequently, introgressing exotic genetic material into sugar beet breeding programs can allow the incorporation of new interesting alleles.

Keywords Sugar beet · Exotic accession · Genetic diversity · QTL detection

Brigitte MANGIN

E-mail : brigitte.mangin@inra.fr

[1] LIPM, Université de Toulouse, INRA, CNRS, Castanet-Tolosan, France

[2] Florimond Desprez Veuve & Fils SAS, BP41, 3, Rue Florimond Desprez, Capelle-en-Pévèle 59242, France

[3] SESVanderHave, Industriepark Soldatenplein Zone 2/Nr 15, 3300 Tienen, Belgium

Introduction

The sugar beet (*Beta vulgaris* L. ssp. vulgaris) is an important European crop for sugar production ; it is also used as a source for bioethanol and animal feed. It is one of the youngest domesticated crops and probably originated from a relatively limited range of fodder beet types approximately 200 years ago (Fischer 1989). The selection of varieties in the first half of the nineteenth century was a mass selection, and sugar beets were grouped together according to their similarities (Desprez and Desprez 2015). Shortly before the 1860s, in addition to the visual and weight aspects, beets were classified according to their sugar content. In 1856, Louis de Vilmorin set up genealogical selection by taking into account the pedigree and value of the offspring. This type of selection is more accurate than mass selection because it is less affected by environmental effects ; consequently, this has allowed for great progress, particularly for complex characters such as yield. Many interesting characteristics, such as monogermy, maintenance of cytoplasmic male sterility, or resistance to the beet necrotic yellow vein virus, have been gradually introgressed into cultivated lines, leading to an annual increase of sugar yield of 2% per ha : 15t/ha produced in 2015, whereas only 700 kg/ha was produced in 1802 (Desprez and Desprez 2015). Currently, the sugar demand has increased with agroethanol development and the increase in worldwide sugar consumption. Therefore, the French sugar beet industry has to be more competitive. An investment program for the future, called AKER (2012-2020), aims to double the annual rate of progress through genetic improvement (`http://www.aker-betterave.fr/en/`). AKER proposes to increase the genetic variability of sugar beets by searching for interesting new alleles from exotic resources around the world. The introgression of these new alleles in elite material will produce new varieties with a high potential for use by the industry. Wild crop relatives are a source of potentially adaptive genetic diversity for crop breeding programs (**?**). The wild and cultivated relatives of sugar beets are included in *Beta* section *Beta* (Andrello et al. 2016). In the AKER project, a core collection of 16 accessions has been selected from among the 10,000 accessions maintained by public genebanks worldwide and composed of a wide variety of wild accessions, using passport data, geographic origins, pedigrees, and genotyping data. The first step was maximizing geographic diversity by removing duplicates, going from 10,000 to 2000 exotic accessions. Then, the complementary genetic diversity of elite lines was maximized and the number went from 2000 to 16 exotic accessions. These 16 exotic plants have therefore been selected worldwide as representing the maximum of the genetic variability not present in cultivated lines. One of these accessions was crossed several times with an elite to create a population called (elite × exotic) progeny. The goal of our study was to determine whether new and interesting allele diversity could be found in this (elite × exotic) progeny compared to an elite panel. Therefore, the sodium, potassium, and α-amino nitrogen contents were phenotyped in the progeny and in an elite panel. Indeed, the recovery of crystalline sugar in the factory depends on the composition of the sugar beet root, and sodium, potassium, and α-amino nitrogen are the major melassigenic substances. They increase the solubility of sucrose and thereby reduce the crystallization (Hoffmann 2010), such that the sugar beet quality decreases. It

is interesting to search alleles associated with a decrease in these impurities to improve white sugar extraction. Quantitative trait loci (QTL) detections were performed using the progeny and an elite panel for these impurity traits. The comparison of QTLs found in both populations and their positive or negative effects on impurities will allow us to determine whether the progeny has interesting alleles not present in the elite panel. Looking for exotic QTLs that improve traits of interest has already been done successfully in other species. For example, Nedelkou et al. (2017) created a tri-parental population in wheat, a progeny from two cultivated lines and one exotic donor accession, and demonstrated that two detected exotic QTLs had a substantial favorable effect on the studied traits. In (Schnaithmann and Pillen 2013), favorable exotic QTLs were found in a barley introgression line. These studies confirm the potential of exotic germplasm to induce interesting traits in cultivated lines of different species.

Materials and methods

(elite × exotic) progeny

Plant materials

An exotic accession of *Beta vulgaris maritima* from Denmark was crossed with a sugar beet elite pollinator (*Beta vulgaris* L.) from the Florimond Desprez company. Two successive backcrosses with another elite pollinator also from the same company were then completed, leading to 187 individuals that constituted the (elite × exotic) progeny.

Phenotypic data

The (elite × exotic) progeny was evaluated in 2016. A total of 187 individuals were evaluated in combination with a tester MSF1 and compared to four commercial hybrids and elite accessions as checks. The entire progeny was evaluated in a lattice design with two replicates for productivity and impurity traits in nine locations : AVE607, BAR601, BEL601, BER601, DOM601, MEM601, and PIE601 in France, DAW601 in Great Britain, and UPI601 in Belgium. The three measured impurity traits were the sodium content (Na, meq/100 g) and potassium content (K, meq/100 g) measured by a flame photometer, and the α-amino nitrogen content (N, meq/100 g) measured by colorimetry. The measured traits linked with productivity were the root yield (RY, tons/ha) and the sucrose content (S, %) measured by refractometry. Other traits linked with productivity were calculated according to the impurities : the white sugar (WS) as WS = S - (0.14 * (K + Na) + 0.25 * N + 0.5), and the white sugar yield (WSY) as WSY = (RY * WS) / 100). Data linked with the productivity are not publicly available at this moment.

Phenotypic data analysis

Spatial effects on each of the nine progeny environments were adjusted with the R package SpATS Rodríguez-Álvarez et al. (2017), available from CRAN (`https://CRAN.R-project.org/package=SpATS`). The SpATS method allows us to consider the local trends, thanks to a smooth bivariate function $f(u, v)$ represented by 2D P splines, where u is the numeric vector of rows and v is the numeric vector of columns. For this experiment, additional terms were included in the SpATS model to account for other sources of environmental variation and genotype effects. We assumed a model including random independent factors for rows ($\boldsymbol{c}_r$), columns ($\boldsymbol{c}_c$), and genotypes ($\boldsymbol{c}_g$), and also fixed factors for genetic checks ($\boldsymbol{\beta}_t$), repetitions ($\boldsymbol{\beta}_n$), and the spatially independent error term $\boldsymbol{e} \sim \mathcal{N}(0, \sigma_e^2 \boldsymbol{I})$. Adapting the formulation in Rodríguez-Álvarez et al. (2016) in order to analyze a single trial and to include repetitions, the SpATS mixed model for each trial is :

$$y = f(\boldsymbol{u}, \boldsymbol{v}) + \boldsymbol{Z}_r \boldsymbol{c}_r + \boldsymbol{Z}_c \boldsymbol{c}_c + \boldsymbol{Z}_g \boldsymbol{c}_g + \boldsymbol{X}_t \boldsymbol{\beta}_t + \boldsymbol{X}_n \boldsymbol{\beta}_n + \boldsymbol{\epsilon}$$

where y is the adjusted phenotype vector, and $\boldsymbol{Z}_r$, $\boldsymbol{Z}_c$, $\boldsymbol{Z}_g$, $\boldsymbol{X}_t$ and $\boldsymbol{X}_n$ are the design matrices associated with rows, columns, genotypes, genetic checks, and repetitions, respectively.

The generalized heritability was also computed with the SpATS package.
Before computing the mean phenotype, we looked at the distribution of adjusted traits in each environment (see boxplots in supplementary material, Figs. S1-S3). We can notice that the potassium content was low in BAR601 and MEM601, the sodium content was particularly high in UPI601, and the α-amino nitrogen content was high in MEM601 and particularly high in UPI601. UPI601 was in Belgium, where nitrogen needs were greater than that in France and Great Britain. The role of nitrogen is of high importance as it affects N and Na concentrations in sugar beet roots (Tsialtas and Maslaris 2005). Phenotype values were therefore impacted by the technical itinerary in this environment. A principal component analysis (PCA) of the nine environments according to all the traits evaluated in the progeny (productivity, impurities) shows that the three above environments were far from the six others (see in supplementary material Fig. S4). We wanted to detect QTLs on the mean phenotype representative of a mean stressed environment. UPI601, which had a particular itinerary, and BAR601 and MEM601 which seem extreme were therefore removed from the study. Only the six consistent environments were kept for further analyses, and the mean phenotype for each trait was then calculated as the mean of the trait value in these six adjusted environments. This mean phenotype was used to find generalist SNPs.

Genotyping

A proprietary 35K Axiom® beet genotyping array was developed along with Affymetrix (http://www.affymetrix.com), CA (USA). This array carried 33,621 high-quality SNPs from which 88% were recently generated through next-generation sequencing of the 16 beet accessions, selected into the AKER project. The SNPs put on the chip were chosen in respect to the technical constraints required by Affymetrix and to be distributed homogeneously along the genome. In brief, the NGS reads (2 × 100 bp paired-end) were mapped onto the sugar beet reference genome (Dohm et al. 2014). SNP calling was performed using a classical pipeline of Burrows-Wheeler Aligner (BWA), samtools, mpileup, varscan, and perl scripts. For genotyping, approximately 30 mg of fresh leaf tissue of each individual was sampled in a 96-deep well, immediately placed at -80°C for at least 24 h and then lyophilized for 48 h. The freeze-dried leaves were subsequently ground using a MM400 Retsch grinder (http://www.retsch.com) for 150 s at 30 frequencies per s Magnetic bead DNA extraction was performed on a robotized platform. All individuals were genotyped using the GeneTitan® microarray automated scanner following the manufacturer's recommendations (http://www.affymetrix.com/support/technical/byproduct.affx?product=genetitan). Genotyping crude data were analyzed with the Axiom® 1.1. analysis suite software package (http://www.affymetrix.com/support/technical/byproduct.affx?product=axiomanalysissuite). For further analyses, only the highest quality SNPs corresponding to 'Poly High Resolution' and 'No Minor Homozygote' categories were used.

Only SNPs whose parental allele provenance was known without ambiguity were kept. The missing genomic data were then imputed for each linkage group using Beagle software (Browning and Browning 2009), leading to 1638 SNPs. After this imputation, some SNPs had exactly the same genetic information. The redundancy of information was not useful for further GWAS analyses ; on the contrary it increased the computational burden and could skew the calculation of the relatedness between hybrids, decreasing the power in regions with many redundant markers (Rincent 2014). Thus, redundant SNPs were discarded. Then, a minor allele frequency (MAF) filter was used to remove SNPs with MAF less than 0.03. Finally, only SNPs with three genotypic classes were retained, coded as 0, 1, or 2 for homozygous for the elite allele, heterozygous, or homozygous for the exotic allele, respectively. All the above filtration steps lead to 604 SNPs retained for subsequent analyses.

Elite panel

Plant materials

A panel of 2101 elite lines of diploid sugar beets (*Beta vulgaris* L.), resulting from many different crosses in Florimond Desprez's breeding program, was analyzed in this study. This population was already studied in Mangin et al. (2019).

Phenotypic data

This panel was evaluated in testcrosses in company multi-environment trials (MET) in 2009, 2010, and 2011 as described in supplementary material of Mangin et al. (2019). Testcross progenies were produced by crossing each elite line to the same single-cross hybrid as a tester. The evaluated traits were the same for the (elite × exotic) progeny : the sodium content (Na, meq/100 g) measured by a flame photometer, the potassium content (K, meq/100 g) measured by a flame photometer, and the α-amino nitrogen content (N, meq/100 g) measured by colorimetry for impurity traits, and others traits linked with productivity.

Phenotypic data analysis

Phenotypes were first adjusted per environment, and then the mean phenotype was calculated as described in supplementary material of Mangin et al. (2019). These mean phenotypes were used as the observed phenotypes for further analyses.
The structure of the elite panel was also analyzed. The optimal cluster number of the hierarchical clustering on principle components analysis was set to two. Clusters contained 676 (Panel A) and 1425 individuals (Panel B).
Boxplots of potassium content, α-amino nitrogen content and sodium content in each panel and in the entire population were plotted in supplementary material (see in supplementary material Fig. S5 to Fig. S7). Phenotypic variabilities were similar in both panels and in the entire population.

Genotyping

The genotyping of this population was completed as described in Mangin et al. (2019). A 0.05 MAF filter was applied, and only SNPs with three genotypic classes were retained, coded 0, 1, and 2 for homozygous for one allele, heterozygous, or homozygous for the other allele, respectively. A total of 626 non-redundant and polymorphic SNPs were retained in the Panel B cluster, and a total of 619 SNPs were retained in the Panel A cluster.

Genetic map

In the AKER project (`http://www.aker-betterave.fr/en/`), 16 exotic accessions have been identified as representing the maximum of the genetic diversity not already present in elite lines. Four of these exotic accessions were crossed with an elite line, and two successive backcrosses were realized with the same elite line, leading to four progenies. Each of these progenies was genotyped with the same 33K Axiom® beet genotyping array as the studied (elite × exotic) progeny.

Data cleaning

The 33,621 genotyped SNPs were filtered to remove erroneous data. This filter was applied separately for each population, by the following criteria. First, the SNPs were categorized based on their genotyping quality using the "Ps_Classification" function of the Affymetrix®'s SNPolisher R package. We discarded the SNPs that were not in the category "PolyHighResolution" for F1S1 populations or either "PolyHighResolution" or "NoMinor" for BC1 populations. Second, markers with more than 5% missing genotypes or with missing or heterozygous elite genotypes were discarded. Third, the SNPs showing segregation distortion were removed. They were detected using a chi-square test with Mendelian segregation as the null hypothesis and a p value threshold of 0.05. Fourth, the SNPs for which wild and elite alleles were inverted were discarded. These SNPs were detected using the "checkAlleles" function of the R package qtl (Broman et al. 2003) with default parameters. This was necessary because it could result in the creation of two linkage groups per chromosome. We could have inverted these SNPs back instead of discarding them, but we did not judge it necessary because there were few : from 0 to 5 for each population.

Genetic map building

Linkage groups were created by transitively grouping markers if the estimated recombination frequency between them was less than 0.35 and if the LOD score was greater than 6. Then, the SNPs that were not grouped and small linkage groups of less than five markers were discarded. This lead to 8 or 9 linkage groups for each population. Each linkage group was attributed to the chromosome on which was located most of its SNPs in a preexisting physic map (Dohm et al. 2014) using a different but overlapping set of markers. For any given population, no chromosome was attributed to more than one linkage group. Linkage groups were very consistent between populations. Only 12 markers among those that were included in a linkage group in at least two different populations were not placed on the same chromosome. These 12 SNPs were removed. Chromosome maps were built using the CarthaGène software (De Givry et al. 2005). The datasets of the four populations were merged using the "dsmergor" command. The redundant markers were merged with "mrkdouble" and "mrkmerges". Then, the maps were built using the command line "buildfw 0 0 1". See CarthaGène documentation for more information. This produced a map per chromosome and per population. Marker order was common across populations but distances were distinct. Consensus distances were calculated to simplify the use of the maps. For each map, absent markers were projected on the map. Then, the distances between markers were averaged across the four populations to produce the consensus distances. Three aberrant individuals appeared to have tens of recombinations per chromosome on the maps. We assumed that these individuals did not belong to the populations to which they were assigned. These individuals were discarded, and the maps were rebuilt afterwards. Although the distance between subsequent SNPs rarely exceeded a few cM in the produced maps, we unusually observed large distances between SNPs on one extremity

of chromosome 3 ; indeed 6 SNPs spanned across an interval of 65 cM. We assumed that this was an artifact of the map-building algorithm that could not determine a correct position for some markers. Therefore, this part of the map was manually removed. Another internal map propriety of the Florimond Desprez company allowed for the positioning of 91 more SNPs on the consensus map. Tables 3 and 4 present a summary of the produced maps. The density of the consensus map and the density of the genotyped markers in each population are represented in the supplementary material (see in supplementary material Fig. S8a and Fig. S8b, respectively).

QTL detection

Association mapping were performed for the QTL detection in both populations. In the (elite × exotic) progeny, QTL detection could have been performed using linkage analysis, which allows for the inference of QTL genotypes using all the informative markers. We preferred to use an association mapping method instead of linkage analysis, neglecting the intervals between markers, because we had a dense map with few missing values at each marker both methods give comparable power Rebai et al. (1995). Moreover, this choice allowed us to perform QTL detection using the same method in the two populations : the (elite × exotic) progeny and the elite panel. Each of the three traits was studied using the adjusted phenotype of the six environments and the mean phenotype for the progeny, and only the mean phenotype for the elite panel. A multi-locus approach with forward selection of SNPs (Segura et al. 2012) was used. At each step of the forward method, a mixed model as proposed by Yu et al. (2006) was evaluated. The variance components of polygenic effects and the residuals were estimated once and a Wald test at each SNP was calculated. The SNP with the smallest p-value was included in the model as a fixed regressor for the next step. The variance attributed to the random polygenic terms decreased when fixed regressors were added to the model ; therefore, the forward selection stops when the remaining variances were close to zero. Two models were used : an additive mixed model as proposed by Yu et al. (2006), and an additive and dominance mixed model as proposed by Bonnafous et al. (2018). For the entire elite panel, the two panel clusters were modeled in the structure fixed term but for progeny and for the GWAS within each cluster no structure was added in the model. Genome-wide association studies (GWAS) were conducted using the R package mlmm.gwas available from CRAN (`https://CRAN.R-project.org/package=mlmm.gwas`).

The additive model

The additive model, proposed by Yu et al. (2006), can be written as below : Let y_i denote the adjusted phenotype of the individual i. Then, the additive model is

$$y_i = \mu + x_i^l \theta_a^l + u_i + e_i \qquad \text{(A model)}$$

x_i^l is the centered genotype of the ith individual at the lth marker locus. In the (elite $\times$ exotic) progeny, the genotype is coded 0, 1, or 2 for homozygous for the elite allele, heterozygous, or homozygous for the exotic allele, respectively. In the elite panel, the genotype is coded 0, 1, or 2 for homozygous for one allele, heterozygous, or homozygous for the other allele, respectively. θ_a^l is the additive effect of the lth locus ; u_i denotes the random additive polygenic effect ; and e_i is the residual error. Let $\boldsymbol{u}$ and $\boldsymbol{e}$ be vectors $(u_i,\ i = 1, \cdots, n)$ and $(e_i,\ i = 1, \cdots, n)$, respectively. Then $\boldsymbol{u} \sim \mathcal{N}(0, \sigma_u^2 \boldsymbol{K_a})$, $\boldsymbol{e} \sim \mathcal{N}(0, \sigma_e^2 \boldsymbol{Id})$, where $\boldsymbol{K_a}$ is a matrix of relative kinship coefficients that define the degree of genetic covariance between a pair of individuals, and σ_u^2 and σ_e^2 are polygenic and residual variances, respectively. The relationship matrix is equivalent to the unscaled kinship matrix described by VanRaden (2008) :

$$\boldsymbol{K_a} = \boldsymbol{X}\boldsymbol{X'}$$

where $\boldsymbol{X} = \left[x_i^l\right]_{\substack{l=1,\cdots,L \\ i=1,\cdots,n}}$ is the centered matrix of the genotypes.

For the elite panel the structure of the population is also considered, so the model used is :

$$y_i = \mu + c_i + x_i^l \theta_a^l + u_i + e_i \qquad \text{(A model with structure)}$$

where c_i is the cluster to which the ith hybrid belongs.

The additive and dominance model

A model including additive and dominance effects of SNPs as proposed in Bonnafous et al. (2018) was also used. The additive and dominance model is

$$y_i = \mu + x_i^l \theta_a^l + w_i^l \theta_d^l + A_i + D_i + e_i \qquad \text{(AD model)}$$

x_i^l is the centered genotype of the ith individual at the lth marker locus ; θ_a^l is the additive effect of the lth locus ; w_i^l is defined later ; θ_d^l is the dominance effect of the lth locus ; A_i is the random additive effect i ; D_i is the random dominant effect i ; and e_i denotes error.

Let $\boldsymbol{A}$, $\boldsymbol{D}$, and $\boldsymbol{e}$ denote vectors $(A_i,\ i = 1, \cdots, n)$, $(D_i,\ i = 1, \cdots, n)$, and $(e_i,\ i = 1, \cdots, n)$, respectively, with n denotes the number of individuals.

Then $\boldsymbol{A} \sim \mathcal{N}(0, \sigma_a^2 \boldsymbol{K_a})$, $\boldsymbol{D} \sim \mathcal{N}(0, \sigma_d^2 \boldsymbol{K_d})$, $\boldsymbol{e} \sim \mathcal{N}(0, \sigma_e^2 \boldsymbol{Id})$, where $\boldsymbol{K_a}$ is the additive kinship matrix ; $\boldsymbol{K_d}$ is the dominance kinship matrix ; and σ_a^2, σ_d^2 and σ_e^2 are additive, dominance and residual variances, respectively. $\boldsymbol{K_a}$ has been defined in the additive model, and $\boldsymbol{K_d} = \boldsymbol{W}\boldsymbol{W'}$ where $\boldsymbol{W} = \left[w_i^l\right]_{\substack{l=1,\cdots,L \\ i=1,\cdots,n}}$ with L the number of loci ; and

$$w_i^l = \begin{cases} -p_1^l p_0^l & \text{if the } i\text{th individual is homozygote for the elite allele in} \\ & \text{the (elite } \times \text{ exotic) progeny, or for one allele in the elite panel at locus } l \\ 2p_2^l p_0^l & \text{if the } i\text{th individual is heterozygote at locus } l \\ -p_2^l p_1^l & \text{if the } i\text{th individual is homozyogote for the exotic allele in} \\ & \text{the (elite } \times \text{ exotic) progeny, or for the other allele in the elite panel at locus } l \end{cases} \quad (1)$$

where p_0^l, p_1^l, and p_2^l are the genotypic frequencies for the genotypes 0, 1, and 2 at the locus l. It is equivalent to the unscaled formula described in Vitezica et al. (2017) using the NOIA model (Álvarez-Castro and Carlborg 2007).

For the elite panel the structure of the population is also considered, so the model used is :

$$y_i = \mu + c_i + x_i^l \theta_a^l + w_i^l \theta_d^l + A_i + D_i + e_i \qquad \text{(AD model with structure)}$$

where c_i is the cluster to which the ith individual belongs.

Model selection and SNP estimation

The more integrated the regressors in the models, the lower the remaining trait variance to explain. However, the last SNPs added into the model may have a very small effect, whereas the purpose of GWAS analysis is to find SNPs with strong effects. That is why a parsimony criterion is used to select the best model, where the fewest SNPs explain most of the trait variability. The BIC Bayesian Information Criterion (BIC) is used to find the best model in the elite panel GWAS (2,101 individuals for almost 624 SNPs). However, this criterion is not strict enough for model selection in large model space (Chen and Chen 2008) as in the progeny (only 186 individuals for 604 SNPs). Accordingly, the extended Bayesian Information Criterion (eBIC) (Chen and Chen 2008) was used. It penalizes the BIC calculation according to the number of possible models for a given number of regressors using a mathematical combination. The effects of SNPs selected by eBIC were computed in the AD model, the most complete model, at the best step. Tukey's test of mean comparison was then performed to analyze the significance of the difference among the three genotypic classes (00 homozygous, 01 or 10 heterozygous, 11 homozygous).

QTLs merging

To compare GWAS results in the progeny and in the elite panel, all detected SNPs in a population were merged into QTLs. A QTL was defined as a group of SNPs associated with traits of interest, located on the same chromosome with a maximum of 5 cM between two consecutive SNPs, and with linkage disequilibrium greater than the significance threshold. The significance level of linkage disequilibrium was studied independently for both populations. It was obtained by taking a random sample of 10,000 pairs of markers belonging to different chromosomes, calculating the squared Pearson's correlation r^2 corrected by kinship for the progeny and by the structure for the elite panel (Mangin et al.

2012) between each pair, and taking the 99% quantile of the 10,000 distribution as the threshold. We therefore obtained thresholds of 0.33 and 0.12 for the progeny and for the elite panel, respectively.

Results

Phenotypic data analysis

(elite × exotic) progeny

Figure 1 shows correlations between the six environments of the (elite × exotic) progeny for each of the three impurity traits : potassium content, α-amino nitrogen content, and sodium content. Larger and darker squares between two environments indicate a greater positive correlation between the two environments. Larger and redder square between two environments indicates a greater negative correlation between the two environments. All these environments were located in the north of France except DAW601, located in the south of Great Britain. Five of the six environments, AVE607, BEL601, BER601, DOM601, and PIE601, were well correlated for potassium content with correlations values from 0.67 to 0.81 (Fig. 1a). Correlations between these environments and DAW601 were slightly lower but still positive (from 0.49 to 0.64). For the α-amino nitrogen (Fig. 1b), the six environments were correlated, but with lower correlation values (from 0.43 to 0.71). AVE607, BER601, DOM601, and PIE601 were well correlated for sodium content with correlation values from 0.66 to 0.74 (Fig. 1c). Correlations between these environments and BEL601 were a little bit lower (from 0.58 to 0.64), and correlations between all these environments and DAW601 were really lower (from 0.28 to 0.36).

Table 1 provides the part of the phenotype variance explained by the genotype for each of the three impurity traits, also called the heritability of the trait. Heritability ranged from 0.44 to 0.85 for the potassium quantity, 0.37 to 0.80 for the sodium quantity, and from 0.45 to 0.75 for the α-amino nitrogen. The heritability of the three traits was lower in DAW601 than in other environments. This environment was in Great Britain, whereas all the others were in France ; thus, its soil could be rather different from that of the others. The heritability of sodium content and α-amino nitrogen content were also quite low in BEL601.

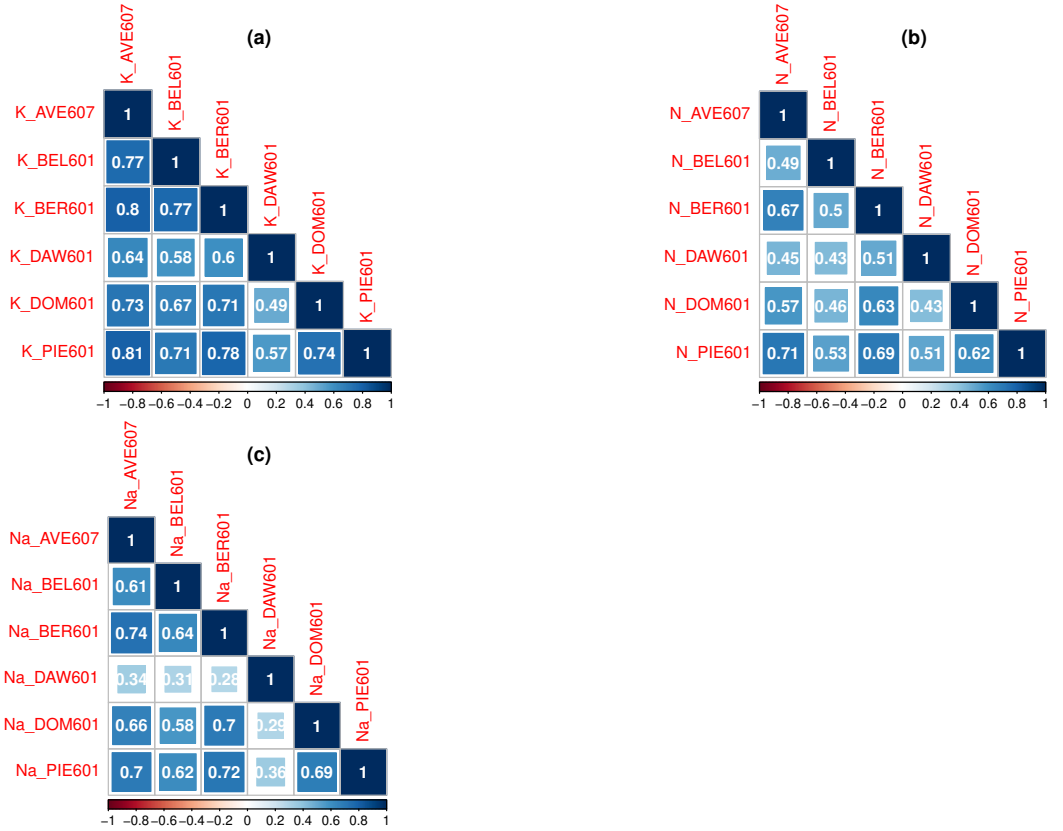


FIGURE 1 – Correlation of potassium content (K ; meq/100g), α-amino nitrogen content (N ; meq/100g) and sodium content (Na ; meq/100g) bewteen six consistent environments AVE607, BEL601, BER601, DAW601, DOM601 and PIE601 of (elite × exotic) progeny. (a) Correlation of potassium content (K ; meq/100g) bewteen six consistent environments of the (elite × exotic) progeny (b) Correlation of α-amino nitrogen content (N ; meq/100g) between six consistent environments of the (elite × exotic) progeny (c) Correlation of sodium content (Na ; meq/100g) between six consistent environments of the (elite × exotic) progeny

TABLE 1 – Heritabilities (h^2) in each of the six consistent environments of the (elite × exotic) progeny (AVE607, BEL601, BER601, DAW601, DOM601 and PIE601) for potassium content (K ; meq/100g), sodium content (Na ; meq/100g), and α-amino nitrogen content (N ; meq/100g)

	AVE607	BEL601	BER601	DAW601	DOM601	PIE601
h^2_K	0.82	0.75	0.87	0.44	0.74	0.85
h^2_{Na}	0.76	0.59	0.77	0.37	0.75	0.80
h^2_N	0.68	0.49	0.73	0.45	0.63	0.75

Elite panel

Table 2 shows heritabilities of each of the three impurity traits in the elite panel, and for each of its clusters. Heritability calculated for each trait in the entire panel was similar to the higher heritability found in the progeny for the corresponding traits.

TABLE 2 – Heritabilities (h^2) in each panel and in the entire population of the elite panel for potassium content (K ; meq/100g), sodium content (Na ; meq/100g), and α-amino nitrogen content (N ; meq/100g)

	Entire panel	Panel A	Panel B
h^2_K	0.88	0.68	0.84
h^2_{Na}	0.77	0.42	0.73
h^2_N	0.70	0.42	0.62

Mapping

There were 604 distinct SNPs for the progeny ; 448 were mapped. There were 626 distinct SNPs for the elite panel ; 322 were mapped. The information regarding the genetical maps of population used to create the consensus map and the genetical consensus map is given in Tables 3 and 4, respectively. The standard nomenclature of the nine chromosomes of sugar beet (Butterfass 1964) is used.

TABLE 3 – Number of SNPs and length of each chromosome of the four genetic maps used to create the consensus map (cM : Haldane). The four genetic maps were created from four (elite × exotic) populations generated in the AKER project

	Population 804		Population 805		Population 809		Population 813	
	SNPs	length (cM)	SNPs	length (cM)	SNPs	length (cM)	SNPs	length (cM)
Chromosome 1	934	70.9	863	79.4	846	66.5	314	83.2
Chromosome 2	1184	76.8	819	77.1	683	77.4	624	86.4
Chromosome 3	1525	69.8	1115	90.6	1005	94.5	1356	103.4
Chromosome 4	1124	75.7	972	77.2	998	99.3	1097	114.2
Chromosome 5	97	10.9	1101	73.7	883	72.6	-	-
Chromosome 6	-	-	902	97.2	587	84.1	850	108.5
Chromosome 7	13	1.3	508	75.8	817	94.6	102	50.6
Chromosome 8	1411	91.2	1302	87.6	1191	113.1	-	-
Chromosome 9	80	25.4	1005	90.3	964	113.4	900	139.1
Total	6368	422	8587	748.9	7974	815.5	6445	828.4

Table 4 – Number of SNPs and length of each chromosome of the genetic consensus map (cM : Haldane) 91 SNPs were then added in the consensus map, from the propriety map. From 0 to 31 SNPs were added to each chromosome, with a mean of 13

	SNPs	length (cM)
Chromosome 1	1054	76
Chromosome 2	1205	84.2
Chromosome 3	1637	89.7
Chromosome 4	1290	91.6
Chromosome 5	1130	73.7
Chromosome 6	941	96.6
Chromosome 7	844	73.1
Chromosome 8	1597	123.4
Chromosome 9	1140	116.2
Total	10838	824.5

QTL detection results

(elite × exotic) progeny

Association studies with the A model and with the AD model were performed on the six environments of the (elite × exotic) progeny and on the mean phenotype for each of the three impurity traits : potassium content, α-amino nitrogen content, and sodium content. As the mean phenotype is certainly the most interesting trait for breeding, we first present results on this mean phenotype. After filtration with the eBIC criterion, 16 distinct SNPs were detected. All were detected with the A model, except one which is only detected with the AD model. Another one was detected with both models. As the studied traits were impurities, we can say that the exotic allele of a detected SNP had a favorable effect if the trait value decreased in the presence of this exotic allele. Table 5 provides details about all SNPs detected with the mean phenotypes.

In Fig. 2 QTL detection results are illustrated with Manhattan plots using the A model of the first GWAS forward step, which is the usual GWAS analysis, and the GWAS forward step selected by eBIC for the mean phenotype of each impurity traits in the (elite × exotic) progeny. Note that the two steps could be the same. These Manhattan plots on all environments of the (elite × exotic) progeny are given in supplementary material (see in supplementary material Fig. S9-S14).

We wanted to see wheter detected SNPs in the mean phenotype were also detected in the six environments. After filtration with the eBIC criterion, 33 distinct SNPs were detected in total on the six environments and the mean phenotype. All were detected with the A model, and three were also detected with the AD model. Figure 3 shows all of them in the (elite × exotic) progeny.

TABLE 5 – SNPs associated with potassium content (K ; meq/100g), sodium content (Na ;meq/100g), and α-amino nitrogen content (N ; meq/100g) for the mean phenotype of (elite $\times$ exotic) progeny. These SNPs are detected in association studies with an additive model (A) and an additive and dominance model (AD), and selected with the eBIC criterion. Their position on chromosome, the proportion of variance they explained in the multi SNPs model selected by eBIC (%var), and information about the favorable or unfavorable effect of the exotic allele are also given

SNP	Trait	model	Chr	Position	%var	favorable.exotic
SNP_10753	Na	A	9	105.93	0.11	yes
SNP_06641	Na	AD	6	31.43	0.13	no
SNP_07975	K	A	7	66.69	0.18	no
SNP_07975	K	AD	7	66.69	0.18	no
SNP_06319	N	A	5	64.98	0.52	no
SNP_00322	Na	A	1	46.05	0.20	no
SNP_06273	Na	A	5	56.20	0.05	no
SNP_01689	Na	A	2	22.23	0.02	yes
SNP_05508	Na	A	5	30.14	0.07	no
SNP_00116	Na	A	1	21.86	0.08	yes
SNP_02804	Na	A	3	50.78	0.04	yes
SNP_00350	Na	A	1	51.00	0.04	yes
SNP_09633	Na	A	8	95.03	0.06	yes
SNP_09271	Na	A	8	67.17	0.04	no
SNP_09818	Na	A	9	10.12	0.06	yes
SNP_09973	Na	A	9	36.71	0.05	yes
SNP_06344	N	A	5	68.13	0.18	yes

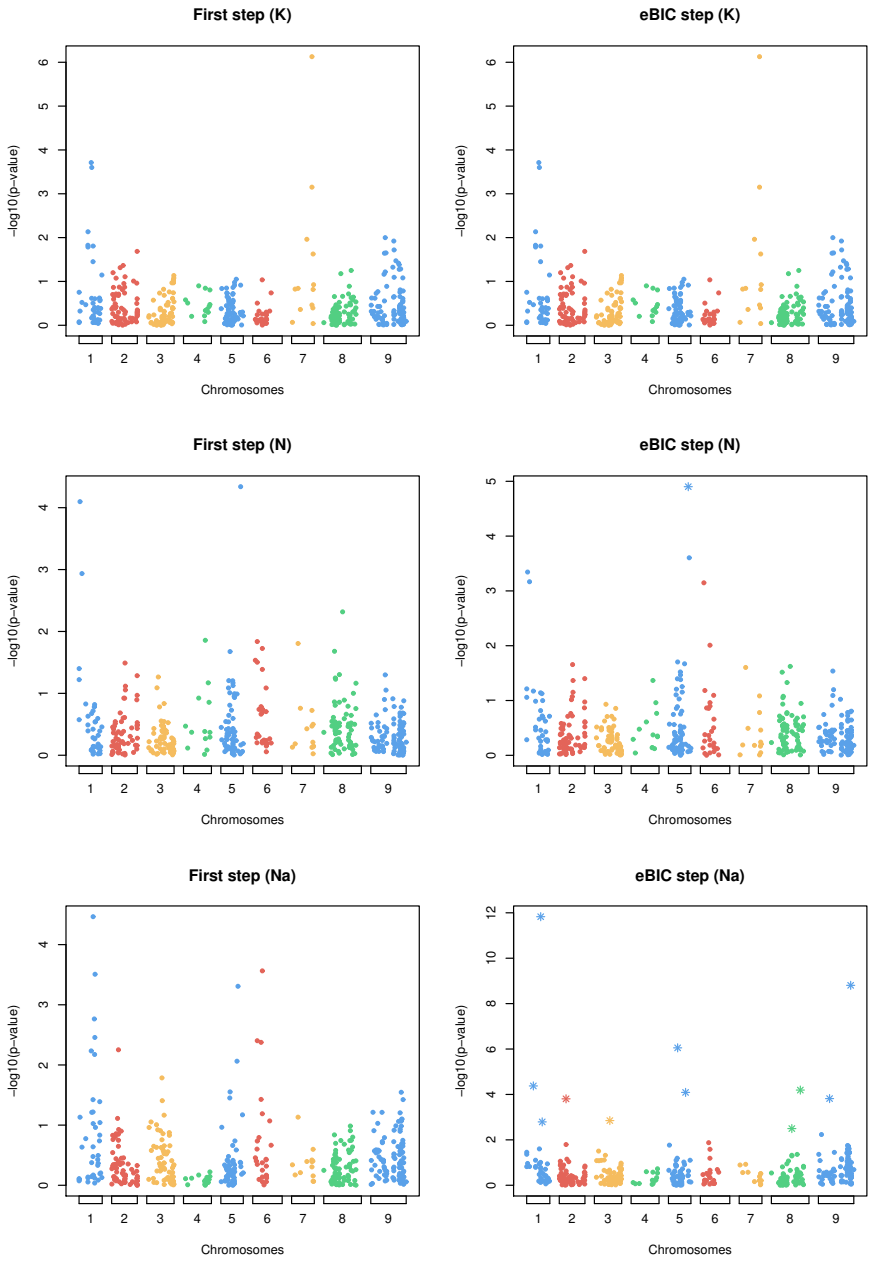


FIGURE 2 – Manhattan plots in the (elite × exotic) progeny using the additive model of the first step of GWAS on the left, and the step selected by eBIC on the right for the mean phenotype of potassium content on the first row, for the mean phenotype of α-amino nitrogen content on the second row and for the mean phenotype of sodium content on the third row. Note that the two steps can be the same. Stars in the step selected by eBIC represent SNPs detected and added into the model in previous steps

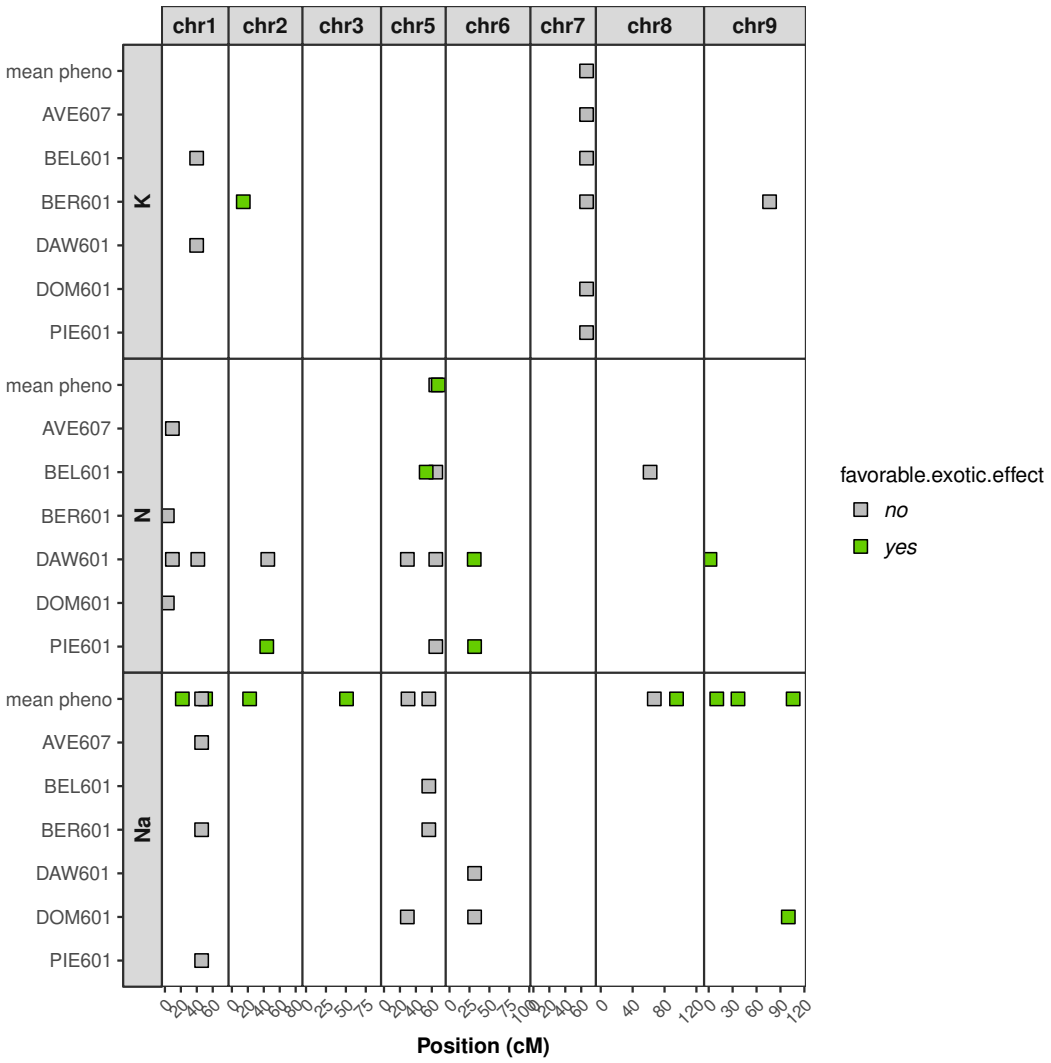


FIGURE 3 – SNPs detected in the (elite × exotic) progeny for potassium content (K, meq/100g), sodium content (Na ; meq/100g) and α-amino nitrogen (N ; meq/100g) with the mean phenotype and in each environment.

Four SNPs associated with the potassium content were detected. One SNP on chromosome 7 was found in all environments, except in DAW601, whereas the others were detected in one or two environments. This SNP was also found with the AD model in four environments. Another SNP found on the chromosome 2 in BER601 was the only to have a favorable exotic allele effect. A total of 14 SNPs associated with α-amino nitrogen content were detected by the A model, and one SNP_00018, was also found with the AD model. Three of these detected SNPs were detected in two or more environments, including SNP_00018. Six SNPs only detected in one environment had a favorable exotic allele. One SNP was not mapped. A total of 16 SNPs were detected for the sodium content with the A model, most of them for the mean phenotype. One of them was also found with the AD model. Nine had a favorable exotic allele. One SNP was not mapped. There was one SNP in common for α-amino nitrogen content and sodium content on chromosome 6 (see Venn diagram in supplementary material Fig. S15) not detected in the same environment for both traits.

Detected SNPs were then merged into QTLs. A QTL grouped together SNPs that were on the same chromosome, with no more than 5 cM between two consecutive SNPs, and with linkage disequilibrium greater than the predefined significance threshold (0.33 for the progeny). If a detected SNP cannot be merged with another, it alone represented a QTL. For the 33 distinct detected SNPs, 31 were mapped on the consensus map ; one SNP associated with the α-amino nitrogen content and one linked with the sodium content were not mapped. In the progeny, 30 QTLs have been defined. Each SNP not mapped constituted a QTL. Others were also composed of only one SNP, except one which merged 2 SNPs on chromosome 5. The list of all SNPs associated with potassium content, α-amino nitrogen, and sodium content, with the name of the QTL to which they belong, the environment in which they were detected, the model used, the chromosome on which they were located, their position on this chromosome, the part of the phenotype variance they explained, and whether they had a favorable effect of the exotic allele are presented in the supplementary material (see in supplementary material Tables S1, S2, S3).

Elite panel

Association studies were performed for each of the three impurity traits with the A model and with the AD model on the elite panel, and on each of its two clusters (Panel A with 676 individuals, and Panel B with 1425 individuals). Table 6 lists the number of detected SNPs detected only with the A model, only with the AD model or with both models for each of the three impurity traits after the selection by eBIC criterion.

TABLE 6 – Number of detected and SNPs for potassium content (K ; meq/100g), sodium content (Na ; meq/100g), and α-amino nitrogen content (N ; meq/100g) in elite panel, with the additive model (A) and the additive and dominance model (AD) or in both models

Trait	A model only	AD model only	Both models
K	60	0	5
N	40	2	0
Na	75	3	9

The majority of SNPs were detected using the A model. The AD model added only five new SNPs. Fourteen were found by both models. Several SNPs were found in common between two or three traits (see the Venn diagram in supplementary material Fig. S16), but none were detected by the AD model only. Finally, a total of 177 distinct SNPs were detected for all traits. Association study results on the entire panel, the panel A and the panel B were illustrated with Manhattan plots (see in supplementary material Figs. S18 to S20).

Detected SNPs were then merged into QTLs. A QTL groups together SNPs that are on the same chromosome, with no more than 5 cM between two consecutive SNPs, and

with linkage disequilibrium greater than the predefined significance threshold (0.12 for the elite panel). If a detected SNP cannot be merged with another, it alone represented a QTL. Of the 177 distinct detected SNPs, 171 were mapped on the consensus map. One SNP detected associated with the potassium content, one linked with α-amino nitrogen content, and three linked with sodium content were not mapped. In the elite panel, 97 QTLs were defined. Each unmapped SNP constituted a QTL, and 35 other QTLs were also composed by only one SNP, and there were therefore 40 QTLs with only one SNP. Fifty-seven QTLs merged two SNPs or more and the three largest were composed of eight SNPs. Two of them were located on chromosome 1, and the other was on chromosome 7. The list of all SNPs detected that were associated with potassium content, α-amino nitrogen, and sodium content, with the name of the QTL to which they belong, the environment in which they were detected, the model used, the chromosome on which they were located, their position on this chromosome, the part of the phenotype variance they explained, and whether they had a favorable effect of the exotic allele are presented in the supplementary material (see in supplementary material Tables S4, S5, S6).

QTL mapping

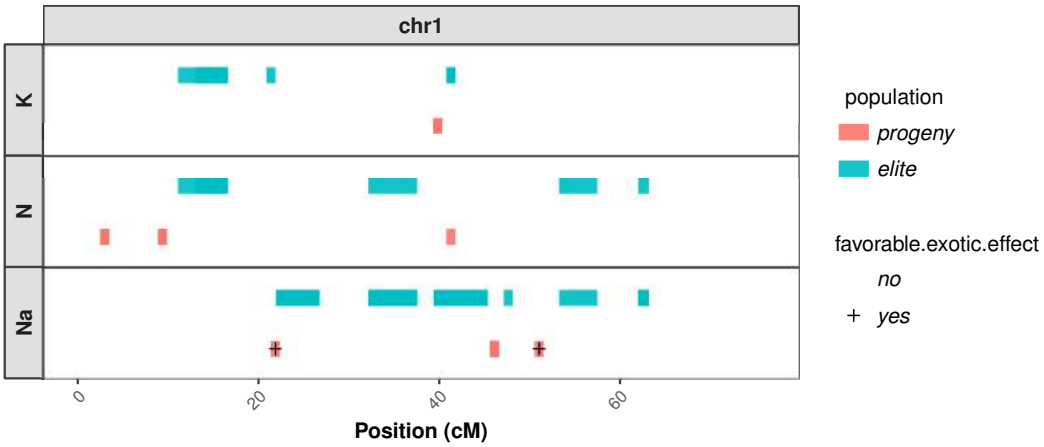


FIGURE 4 – QTLs detected in (elite × exotic) progeny and elite panel for potassium content (K ; meq/100g), sodium content (Na ; meq/100g), and α-amino nitrogen content (N ; meq/100g) mapped on chromosome 1 of the consensus map. The favorable effect of an exotic allele is indicated by the '+' sign

Figure 4 shows QTLs detected in the mean phenotype on chromosome 1 in (elite × exotic) progeny and in elite panel populations for each impurity trait. As the studied traits were impurities, the effects of the exotic allele was considered favorable when it was associated with a decrease in the amount of the impurity. We found more QTLs in the elite panel than in the (elite × exotic) progeny. Some QTLs found in the (elite × exotic) progeny were very close or collocated with those of the elite panel and could have a favorable effect on the exotic allele, as the first QTLs detected that were associated with sodium

content, or an unfavorable effect of the exotic allele, as the second QTL detected that had an association with sodium content and the QTL detected that were associated with potassium content. Other QTLs of the (elite × exotic) progeny were not collocated with QTLs detected in the elite panel. These QTLs could also have a favorable effect on the exotic allele, as the third QTL detected that was associated with sodium content, or an unfavorable effect of the exotic allele, as the three QTLs detected that were associated with α-amino nitrogen content. However, if we look at the coverage of the chromosome with the elite SNPs (see in the supplementary material Fig. S8b), we can see that there was no elite SNP in the region with the first two QTLs associated with α-amino nitrogen content in (elite × exotic) progeny. Thus, with elite SNPs covering this region, we would have also detected it in the elite panel. Results on each chromosome are given in supplementary material (see in supplementary material Fig. S17). The Table 7 gives all QTLs detected in the mean phenotype in the (elite × exotic) progeny that had a favorable exotic effect.

TABLE 7 – SNPs associated with the mean phenotype for potassium content (K ; meq/100g), sodium content (Na ; meq/100g), and α-amino nitrogen content (N ; meq/100g) of (elite × exotic) progeny, which have favorable effect of the exotic allele. These SNPs are detected in association studies with an additive model (A) and an additive and dominance model (AD), selected with the eBIC criterion and merged into QTLs. Their position on chromosome and the proportion of variance they explained in the multi SNPs model selected by eBIC (%var) are also given

SNP	QTL	Trait	model	Chr	Position	%var	favorable exotic
SNP_00116	QTL_19_1	Na	A	1	21.86	0.08	yes
SNP_00350	QTL_09_1	Na	A	1	51.00	0.04	yes
SNP_01689	QTL_18_1	Na	A	2	22.23	0.02	yes
SNP_02804	QTL_07_1	Na	A	3	50.78	0.04	yes
SNP_06344	QTL_43_1	N	A	5	68.13	0.18	yes
SNP_09633	QTL_12_1	Na	A	8	95.03	0.06	yes
SNP_09818	QTL_30_1	Na	A	9	10.12	0.06	yes
SNP_09973	QTL_35_1	Na	A	9	36.71	0.05	yes
SNP_10753	QTL_29_1	Na	A	9	105.93	0.11	yes

Discussion

In this work, we compared the detected QTLs found for three impurities traits in two different populations, an (elite × exotic) progeny composed of 187 individuals and an elite panel composed of 2101 individuals.

The elite panel was divided into two panels by hierarchical clustering. Panel A and panel B contained 676 and 1425 individuals, respectively, so one-third for panel A and two-third for panel B. The cause of this structure was not known, but it had an impact on heritabilities calculated for the three impurity traits in these two panels : heritabilities were one-third lower in panel A. This population was already studied for genomic prediction in Mangin et al. (2019) where authors showed that the accuracy of prediction in panel B

decreases when individuals from panel A are added, whereas a decrease in the accuracy of prediction for panel A is not observed. This is probably due to the larger size of the panel A. Both heritabilities and genomic prediction suggested that the two panels were different, so QTL detection was performed on each of these panels in addition to QTL detection on the full elite panel. Moreover, even we do not have the precise pedigree we know that the elite panel is composed by many biparental populations. Maybe founders of theses populations are very different between the two panels.

Because the individuals from the progeny were related, we could have used a linkage mapping analysis rather than association mapping for QTL detection in this population. To understand why we can use these two methods interchangeably, it is essential to understand what distinguishes them. Linkage analysis and association analysis are both based on linear models, with a statistical test at each position of a putative QTL that gives a scan on the whole genome. The main difference is that linkage analysis makes a test not only at the markers but also between the markers. Another difference is the use of fixed effect model for linkage analysis and mixed model for association analysis. However, mixed model have already been proposed for linkage analysis. They were first used in human pedigrees by Pratt et al. (2000) who assumed that the QTL effect was random. Then, for simple pedigree Pérez-Enciso and Varona (2000) proposed a mixed model of the QTL effect with a fixed part tested at each locus of the chromosome and a random part that considers the relationship between individuals. This model for linkage analysis is identical to that of association analysis. The test performed on the markers by the two methods is also similar if the polygenic variance included in the association model is very low. This follows from the analytical formula of the Wald test used for the association method. This decrease in polygenic variance is obtained by the forward selection procedure of MLMM (Segura et al. 2012). Indeed, at each step of MLMM, the marker that is the most associated with the trait is added in the association model as a regressor and the polygenic variance decreases. Finally, there is no detectable QTL in the polygenic effect of the association model ; they are all fixed effects in the model, similar to the case of linkage analysis with cofactors. Moreover, the choice of cofactors in the linkage analysis is treated in a similar way, because a similar forward procedure limited to markers has been proposed by Jourjon et al. (2005) to decide which cofactors to include in the linkage analysis model. These two methods are thus a response to a model with several QTLs. This is also discussed in Bonnafous et al. (2018), who argues that because the location of multiple QTLs is unknown the association analysis model assumes that each marker is a QTL. The addition of a law on the effects of markers then results in a polygenic effect with a matrix of relatedness, also called kinship. This kinship should be different for each locus, which would increase the power of the association model (Rincent 2014), but for computational cost reasons, generally the same matrix is used for the entire genome. As we have just seen, the tests of the two methods are similar with regard to the markers, but in the linkage analysis pseudo-markers are also used. Rebai et al. (1995) have shown that these pseudo markers provide only 5% more power compared to a marker per marker analysis when the distance between two markers is less than 20 cM. In the (elite $\times$ exotic)

progeny, the two most distant markers were spaced apart by 11.26 cM (markers located on chromosome 8). Testing the QTLs only on markers does not therefore lose a lot of detection power. Finally, because the identity by state (IBS) and the identity by descent (IBD) are identical in a progeny, an association model using an IBS type kinship uses the Mendelian allelic transmission.

Two models were used to do the QTL detection : an additive model and additive and dominance model. Note that the above models are statistical not genetical. For the A model and a testcross, the descendant mean of pollinisators that are heterozygous at a marker is assumed to be the average of the descendant mean of the two homozygous pollinisators. For the AD model, this assumption is not made leading to one more fixed parameter and one more variance component to estimate. These additional parameters cause a decrease in power compared to the A model if the actual dominant effect is not substantial. Most GWAS models consider only the additive effect of markers. However, several studies have shown that the non-additive effects constitute a major part of the variation of complex traits. Bonnafous et al. (2018) showed, by studying the phenotypic hybrid value, that the AD model makes it possible to find SNPs with a dominant effect. Indeed, with hybrids, the AD model corresponds to the model with a genetically dominance. However as we have testcross values, we cannot conclude on the genetically dominance of detected SNPs. Moreover the statistical dominance modeling in the progeny can be proved to be linked to regions with segregation distortion by using the usual decomposition with genetically additive and dominance effects. Besides in the (elite $\times$ exotic) progeny none SNPs were detected with the AD model only, so we could only have used the A model for the QTL detection.

Many more QTLs have been detected in the elite panel than in the progeny (87 and 33, respectively). On several chromosomes, the only QTLs detected were found in the elite panel, and when QTLs were found in the (elite $\times$ exotic) progeny, they were often also found in the panel. This may be because genetic diversity was much more important in the elite panel than in the progeny, since the (elite $\times$ exotic) progeny was a biparental population whereas the elite panel was composed by a lot of small biparental populations. Moreover, the elite panel was composed of 2101 individuals, whereas the (elite $\times$ exotic) progeny was only composed of 187 individuals. Because of this difference in size between the two populations, there was more statistical power in the elite panel for the QTL detection than in the (elite $\times$ exotic) progeny.

On the 33 QTLs detected in the (elite $\times$ exotic) progeny for all impurity traits, only 16 had a favorable exotic allele. The majority of favorable alleles were therefore also present in the elite parent. This was expected as the elite parent was the result of an artificial selection by breeders in the Florimond Desprez company. The sodium content, the α-amino nitrogen content and the potassium content were well-mesured traits with high heritability, so the selection process has probably already fixed the favorable QTL alleles. In addition, the selection of sugar beet is based on the white sugar yield and, the

calculation of the white sugar yield being a function of the impurity content, this choice led to the recruitment of alleles favorable to the reduction of impurities, although selective scanning is currently lacking in sugar beet Adetunji et al. (2014).

Because of the sugar beet breeding history, we know that all chromosomes have not been subjected to the same selection intensity. For example, the heatmap of the progeny shows a strong LD between chromosome 3 and chromosome 9 (Fig. 5a), which is possibly the result of selection for rhizomania resistance. The $Rz1$ on chromosome 3 comes from a single source and its introgression into elite sugar beet breeding lines occurred very rapidly, such that it might have created a genetic bottleneck leading to high LD on other regions of the genome, such as the strong LD on top of chromosome 9 (Adetunji et al. (2014)). We also observed a strong LD between chromosome 3 and chromosome 1, which may also be caused by rhizomania resistance selection. Therefore, SNPs on chromosomes with interchromosomal LD are not independent. The LD threshold in the progeny was probably overestimated. The heatmap of the elite panel did not show interchromosomal LD (Fig. 5b), and there were enough recombination events to break the long range LD. However, we observed several LD intrachromosomes, in particular on chromosome 3, 8, and 9. This high interchromosomal LD could be due to bad mapping of some markers and could cause very long QTLs in the elite panel. To avoid excessively long QTLs, we decided to cut the QTLs into 5 cM intervals in both populations.

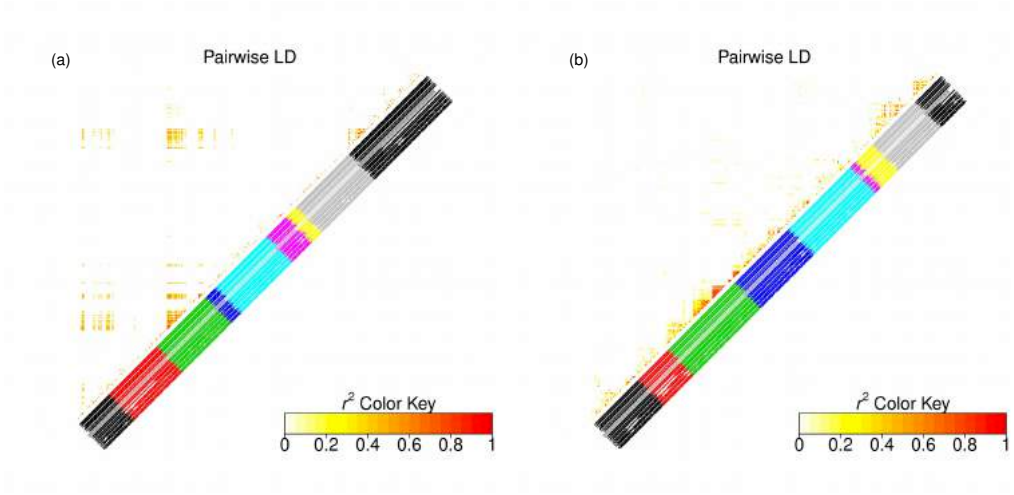


FIGURE 5 – Heatmap of r^2 values between all possible pairs of mapped SNPs in both populations.
(a) LD in the progeny, corrected for genetic relatedness
(b) LD in the elite panel, corrected for the structure in two clusters

Regarding the results, the comparison of the detected QTLs in both populations allowed us to highlight interesting regions in the (elite × exotic) progeny genome. Indeed, some detected QTLs in the (elite × exotic) progeny had a favorable exotic allele on different chromosomes because they were linked with a decrease in the impurity content. Some of them were collocated with detected QTLs in the elite panel, others were found in new regions. To go further in the comparison of the progeny and the elite panel, it would have been interesting to included the elite parent of progeny in the elite panel. Moreover, it could have been interesting to study a progeny generated from the same exotic accession but crossed with another elite line. The genetic background could indeed influence allelic fitness (Ungerer et al. 2003) and therefore hide some interesting regions in the (elite × exotic) progeny. Furthermore, a better genome coverage could have allowed the detection of other QTLs in both populations.

To conclude, the comparison of the detected QTLs in an (elite × exotic) progeny and an elite panel allows the detection of new favorable alleles and genomic regions brought by the exotic accession. Their introgression in a sugar beet elite germplasm is therefore an interesting approach to increasing the genetic diversity that is useful in breeding programs.

Bibliographie

Ibraheem Adetunji, Glenda Willems, Hendrik Tschoep, Alexandra Bürkholz, Steve Barnes, Martin Boer, Marcos Molosetti, Stefaan Horemans, and Fred Van Eeuwijk. Genetic diversity and linkage disequilibrium analysis in elite sugar beet breeding lines and wild beet accessions. *Theoretical and Applied Genetics*, 127(3) :559–571, Mar 2014. ISSN 0040-5752. doi : 10.1007/s00122-013-2239-x.

José M. Álvarez-Castro and Örjan Carlborg. A Unified Model for Functional and Statistical Epistasis and Its Application in Quantitative Trait Loci Analysis. *Genetics*, 176 (2) :1151–1167, June 2007.

Marco Andrello, Karine Henry, Pierre Devaux, Bruno Desprez, and Stéphanie Manel. Taxonomic, spatial and adaptive genetic variation of *Beta* section *Beta*. *Theoretical and Applied Genetics*, 129(2) :257–271, February 2016. ISSN 1432-2242. doi : 10.1007/s00122-015-2625-7.

Fanny Bonnafous, Ghislain Fievet, Nicolas Blanchet, Marie-Claude Boniface, Sébastien Carrère, Jérôme Gouzy, Ludovic Legrand, Gwenola Marage, Emmanuelle Bret-Mestries, Stéphane Munos, Nicolas Pouilly, Patrick Vincourt, Nicolas Langlade, and Brigitte Mangin. Comparison of GWAS models to identify non-additive genetic control of flowering time in sunflower hybrids. *Theoretical and Applied Genetics*, 131(2) :319–332, February 2018. ISSN 1432-2242. doi : 10.1007/s00122-017-3003-4.

Karl W. Broman, Hao Wu, Śaunak Sen, and Gary A. Churchill. R/qtl : QTL mapping in experimental crosses. *Bioinformatics*, 19(7) :889–890, 2003. doi : 10.1093/bioinformatics/btg112.

Brian L. Browning and Sharon R. Browning. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *The American Journal of Human Genetics*, 84(2) :210 – 223, 2009. ISSN 0002-9297. doi : https://doi.org/10.1016/j.ajhg.2009.01.005.

Th Butterfass. Die chloroplastenzahlen in verschiedenartigen zellen trisomer zuckerrüben (*Beta vulgaris* l.). *Z Bot*, 52 :46–77, 1964.

Jiahua Chen and Zehua Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3) :759–771, 2008.

Simon De Givry, Martin Bouchez, Patrick Chabrier, Denis Milan, and Thomas Schiex. Cartha gene : multipopulation integrated genetic and radiation hybrid mapping. *Bioinformatics*, 21(8) :1703–1704, 2005. doi : 10.1093/bioinformatics/bti222.

Michel Desprez and Bruno Desprez. Évolution des méthodes de sélection de Louis de Vilmorin à aujourd'hui. de la sélection phénotypique à la sélection génotypique : l'exemple de la betterave. In Yvette Dattée, editor, *Les Vilmorin, des graines et des hommes (Colloque)*, pages 31–35, 2015. ISBN 978-2-913793-14-9.

Juliane C. Dohm, André E. Minoche, Daniela Holtgräwe, Salvador Capella-Gutiérrez, Falk Zakrzewski, Hakim Tafer, Oliver Rupp, Thomas Rosleff Sörensen, Ralf Stracke, Richard Reinhardt, Alexander Goesmann, Thomas Kraft, Britta Schulz, Peter F. Stadler, Thomas Schmidt, Toni Gabaldón, Hans Lehrach, Bernd Weisshaar, and Heinz Himmelbauer. The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). *Nature*, 505(7484) :546–549, January 2014. ISSN 1476-4687. doi : 10.1038/nature12817.

Hans Eberhard Fischer. Origin of the 'Weisse Schlesische Rübe' (white Silesian beet) and resynthesis of sugar beet. *Euphytica*, 41(1) :75–80, April 1989. ISSN 1573-5060. doi : 10.1007/BF00022414.

Christa M. Hoffmann. Root quality of sugarbeet. *Sugar Tech*, 12(3) :276–287, Dec 2010. doi : 10.1007/s12355-010-0040-6.

Marie-Françoise Jourjon, Sylvain Jasson, Jacques Marcel, Baba Ngom, and Brigitte Mangin. Mcqtl : multi-allelic qtl mapping in multi-cross design. *Bioinformatics*, 21(1) : 128–130, 2005. doi : 10.1093/bioinformatics/bth481.

B Mangin, A Siberchicot, S Nicolas, A Doligez, P This, and C Cierco-Ayrolles. Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity*, 108(3) :285–291, March 2012. ISSN 0018-067X. doi : 10.1038/hdy.2011.73.

Brigitte Mangin, Renaud Rincent, Charles-Elie Rabier, Laurence Moreau, and Ellen Goudemand-Dugue. Training set optimization of genomic prediction by means of ethacc. *PLOS ONE*, 14(2) :1–21, 02 2019. doi : 10.1371/journal.pone.0205629. URL `https://doi.org/10.1371/journal.pone.0205629`.

Ioanna-Pavlina Nedelkou, Andreas Maurer, Anne Schubert, Jens Léon, and Klaus Pillen. Exotic QTL improve grain quality in the tri-parental wheat population SW84. *PLOS ONE*, 12(7) :e0179851, 2017. ISSN 1932-6203. doi : 10.1371/journal.pone.0179851.

Stephen C. Pratt, Mark J. Daly, and Leonid Kruglyak. Exact Multipoint Quantitative-Trait Linkage Analysis in Pedigrees by Variance Components. *The American Journal of Human Genetics*, 66 :1153–1157, March 2000. ISSN 0002-9297. doi : 10.1086/302830.

Miguel Pérez-Enciso and Luis Varona. Quantitative Trait Loci Mapping in F2 Crosses Between Outbred Lines. *Genetics*, 155(1) :391–405, May 2000. ISSN 0016-6731, 1943-2631.

A. Rebai, B. Goffinet, and B. Mangin. Comparing power of different methods for QTL detection. *Biometrics*, 51(1) :87–99, March 1995. ISSN 0006-341X.

Renaud Rincent. *Optimization of association genetics and genomic selection strategies for populations of different diversity levels : Application in maize (Zea mays L.)*. Theses, AgroParisTech, April 2014.

María Xosé Rodríguez-Álvarez, Martin P. Boer, Fred A. van Eeuwijk, and Paul H.C. Eilers. Correcting for spatial heterogeneity in plant breeding experiments with p-splines. *Spatial Statistics*, 23 :52 – 71, 2017.

María Xosé Rodríguez-Álvarez, Martin P. Boer, Fred A. van Eeuwijk, and Paul H. C. Eilers. Spatial Models for Field Trials. *arXiv :1607.08255 [stat]*, July 2016. arXiv : 1607.08255.

Florian Schnaithmann and Klaus Pillen. Detection of exotic QTLs controlling nitrogen stress tolerance among wild barley introgression lines. *Euphytica*, 189(1) :67–88, January 2013. ISSN 1573-5060. doi : 10.1007/s10681-012-0711-3.

Vincent Segura, Bjarni J Vilhjálmsson, Alexander Platt, Arthur Korte, Ümit Seren, Quan Long, and Magnus Nordborg. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature genetics*, 44(7) : 825–830, 2012.

J. T. Tsialtas and N. Maslaris. Effect of N Fertilization Rate on Sugar Yield and Non-Sugar Impurities of Sugar Beets (Beta vulgaris) Grown Under Mediterranean Conditions. *Journal of Agronomy and Crop Science*, 191(5) :330–339, 2005. ISSN 1439-037X. doi : 10.1111/j.1439-037X.2005.00161.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1439-037X.2005.00161.x`.

Mark C. Ungerer, C. Randal Linder, and Loren H. Rieseberg. Effects of Genetic Background on Response to Selection in Experimental Populations of Arabidopsis thaliana. *Genetics*, 163(1) :277–286, January 2003. ISSN 0016-6731, 1943-2631.

PM VanRaden. Efficient methods to compute genomic predictions. *Journal of dairy science*, 91(11) :4414–4423, 2008.

Zulma G. Vitezica, Andrés Legarra, Miguel A. Toro, and Luis Varona. Orthogonal Estimates of Variances for Additive, Dominance, and Epistatic Effects in Populations. *Genetics*, 206(3) :1297–1307, July 2017. ISSN 0016-6731, 1943-2631. doi : 10.1534/genetics.116.199406.

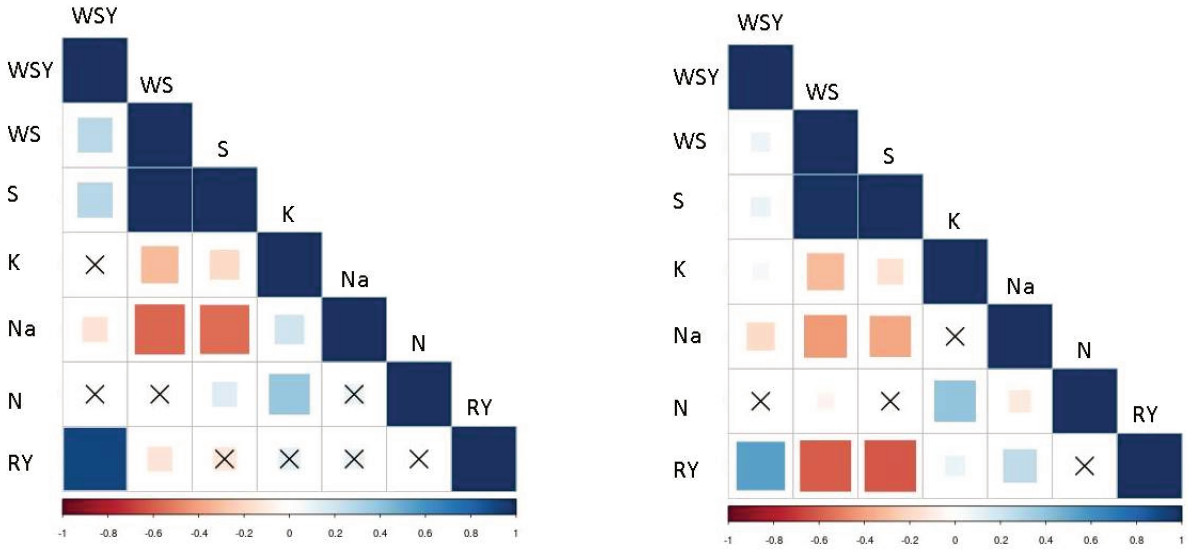
Jianming Yu, Gael Pressoir, William H Briggs, Irie Vroh Bi, Masanori Yamasaki, John F Doebley, Michael D McMullen, Brandon S Gaut, Dahlia M Nielsen, James B Holland, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics*, 38(2) :203–208, 2006.

Des QTLs favorables apportés par l'exotique ont été mis en évidence dans cet article, montrant ainsi qu'une accession exotique peut être source d'une nouvelle diversité génétique favorable vis-à-vis du matériel élite de betterave sucrière. Le postulat d'AKER est ainsi vérifié. Cependant cette étude porte uniquement sur des caractères liés aux impuretés, il est intéressant de vérifier s'il est également possible de trouver des fragments favorables apportés par l'exotique par rapport aux caractères liés au rendement. Ces caractères liés au rendement n'ont pas été traités dans l'article pour des raisons de confidentialité.

2.2.2 Etude de l'architecture génétique sur l'ensemble des caractères

La même méthodologie est utilisée afin de définir l'architecture génétique des caractères liés au rendement : le rendement racinaire (RY), la teneur en sucre (S), le pourcentage en sucre blanc (WS) et le rendement en sucre blanc (WSY). Seuls les résultats seront présentés dans cette partie.

Les corrélations de l'ensemble des caractères ont été évaluées dans chacune la descendance (élite × exotique) et dans le panel élite (Figure 2.2). Les résultats obtenus peuvent être divisés en deux catégories : les corrélations dues aux propriétés agronomiques et les corrélations attendues du fait des formules de calcul permettant d'obtenir les caractères WS et WSY.



(a) (b)

FIGURE 2.2 – Coefficient de corrélation de Pearson (a) sur la descendance (élite × exotique) (b) sur le panel élite entre le phénotype moyen des 7 caractères étudiés : le rendement en sucre blanc (WSY, %), la teneur en sucre blanc (WS, %), la teneur en sucre (S, %), la teneur en potassium (K, meq/100g), la teneur en sodium (Na, meq/100g), la teneur en azote alpha-aminé (N, meq/100g) et le rendement racinaire (RY, t/ha). La taille est la couleur de chacun des carrés est proportionnelle à la valeur de la corrélation entre deux caractères : plus le carré est grand et foncé plus la valeur de la corrélation est élevée. Une corrélation positive est représentée en bleue et une corrélation négative est représentée en rouge. Les croix représentent une corrélation non significative entre deux caractères, le risque α étant fixé à 5%.

Les trois caractères d'impuretés sont étudiés car le potassium, le sodium et l'azote alpha-aminé sont des substances mélassigènes majeures : ils augmentent la solubilité du sucrose et en réduisent l'extractibilité (Hoffmann 2010). Ils affectent ainsi négativement la teneur en sucre. La figure 2.2 montre en effet une corrélation négative entre le caractère S et les caractères d'impuretés K et Na, dans la descendance (élite × exotique) tout comme dans le panel élite. Une faible corrélation positive est en revanche observée entre S et N exclusivement dans la descendance (élite × exotique) : dans une certaine mesure l'azote permet une meilleure croissance de la plante et lui permet ainsi de stocker plus de sucrose. Une forte corrélation négative existe entre RY et S dans les deux populations, corrélation induite par le fait que les grandes cellules du parenchyme de la betterave sucrière stockent plus d'eau et de substances non sucrées aux dépens du sucre que les petites cellules (Milford 1973). Cela explique également qu'une corrélation positive puisse être observée entre RY et les caractères liés aux impuretés.

Les autres corrélations observées peuvent être expliquées par les formules utilisées pour calculer WS et WSY. La formule utilisée pour calculer WS, $WS = S - (0.14 \times (K + Na) + 0.25 \times N + 0.5)$, induit que les caractères WS et S sont très corrélés, tandis que la

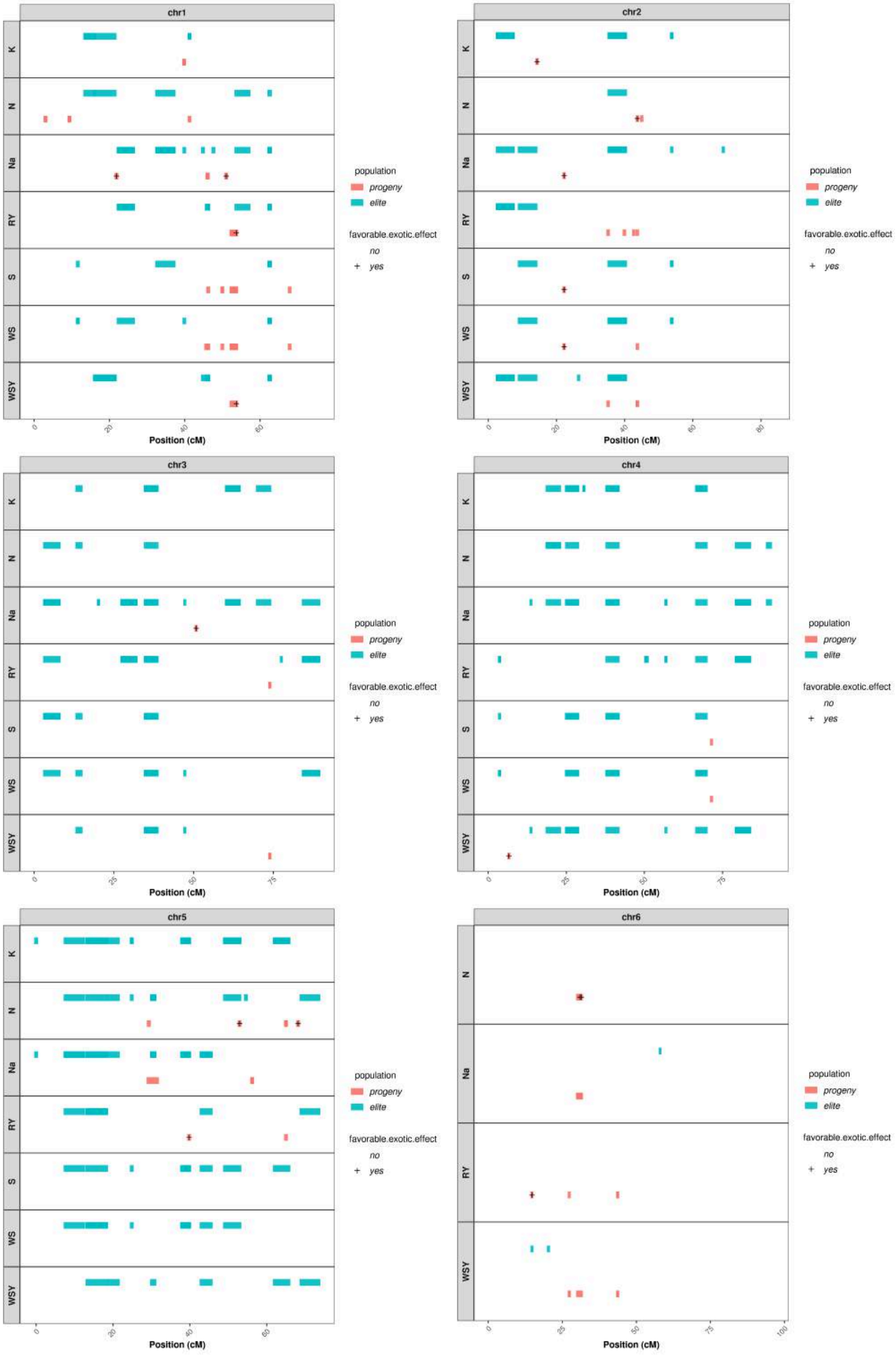
corrélation entre WS et les caractères d'impureté est négative (ou non significative). La formule appliquée pour calculer WSY, WSY = (RY × WS)/100), implique une corrélation positive entre WSY et RY, ainsi qu'une corrélation positive bien que plus faible entre WSY et WS. Puisque WS et S sont très corrélés, cette formule explique également la corrélation positive entre WSY et S et une corrélation négative (ou non significative) entre WSY et les caractères d'impureté.

L'héritabilité de chacun des 7 caractères a été estimée dans chacun des 6 environnements cohérents analysés pour la descendance (élite x exotique) (2.2.1). Les résultats sont présentés dans le Tableau 2.1.

TABLE 2.1 – Héritabilité (H^2) dans les 6 environnements analysés pour la descendance (élite × exotique) (AVE607, BEL601, BER601, DAW601, DOM601 and PIE601) pour la teneur en potassium (K ; meq/100g), en sodium (Na ; meq/100g), en azote α-aminé (N ; meq/100g), en sucre (S ;%), en sucre blanc (WS ;%), le rendement racinaire (RY ; t/ha) et le rendement racinaire en sucre blanc (WSY ; t/ha)

	AVE607	BEL601	BER601	DAW601	DOM601	PIE601
H^2_K	0.82	0.75	0.87	0.44	0.74	0.85
H^2_Na	0.76	0.59	0.77	0.37	0.75	0.80
H^2_N	0.68	0.49	0.73	0.45	0.63	0.75
H^2_S	0.82	0.61	0.51	0.49	0.60	0.73
H^2_WS	0.82	0.62	0.53	0.49	0.60	0.74
H^2_RY	0.71	0.44	0.76	0.57	0.63	0.69
H^2_WSY	0.64	0.41	0.73	0.44	0.59	0.68

La figure 2.3 représente les QTLs détectés pour les sept caractères dans la descendance exotique et dans le panel élite dans l'ensemble des environnements pour chacun des neuf chromosomes de la betterave sucrière. Pour rappel, les SNPs détectés lors de l'analyse d'association sont regroupés en QTL s'ils sont en déséquilibre de liaison, se trouvent sur le même chromosome, et sont distants de moins de 5cM.


chr1
chr2
chr3
chr4
chr5
chr6
K
N
Na
RY
S
WS
WSY
Position (cM)
population
progeny
elite
favorable.exotic.effect
no
yes

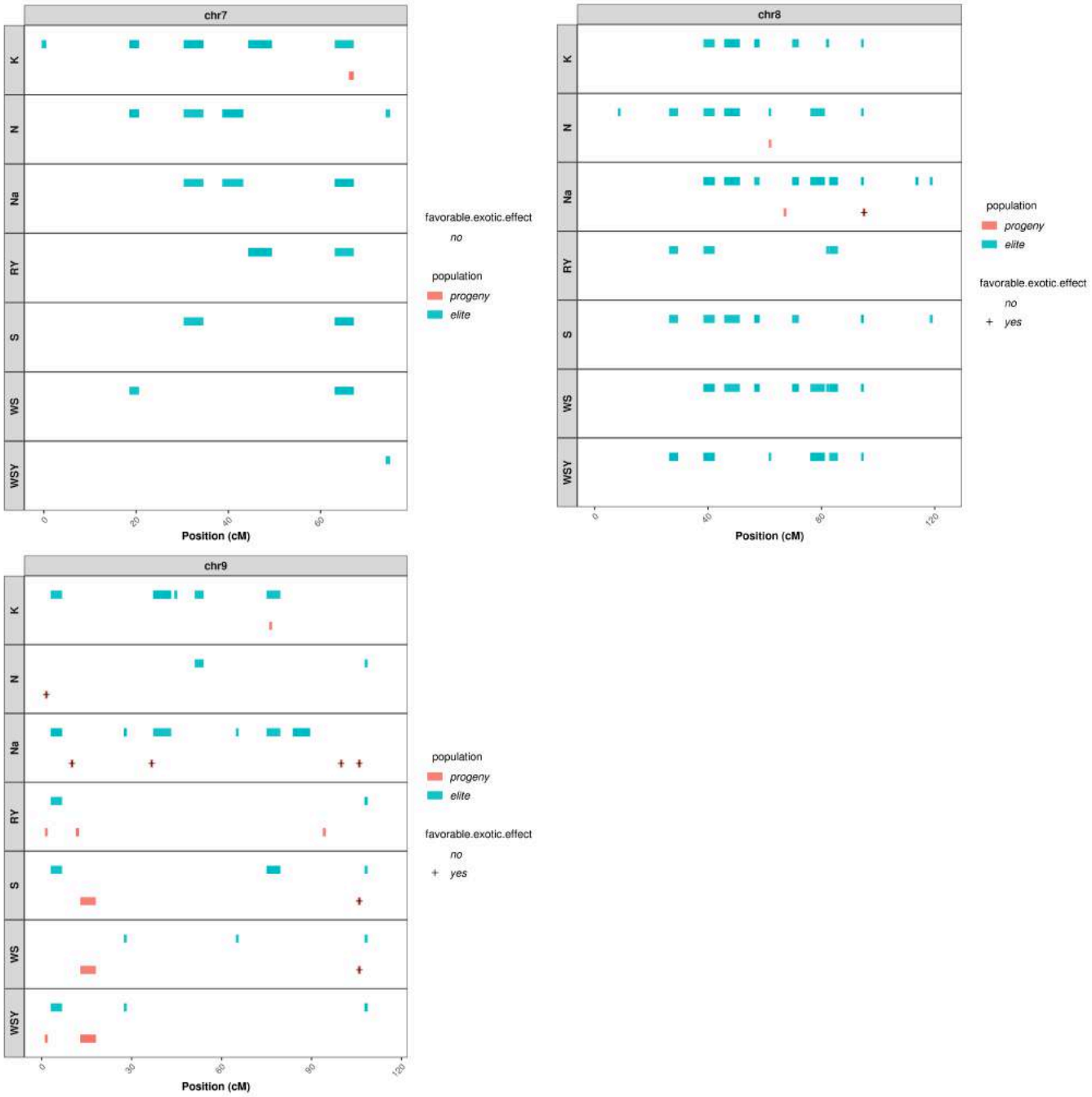


FIGURE 2.3 – QTL détectés chacun des 9 chromosomes dans la descendance (élite × exotique) en rouge et dans le panel élite en bleu pour 7 caractères agronomiques : la teneur en potassium (K), la teneur en azote alpha-aminé (N), la teneur en sodium (Na), le rendement racinaire (RY), la teneur en sucre (S), la teneur en sucre blanc (WS) et le rendement en sucre blanc (WSY). Les QTLs détectés dans la descendance (élite × exotique) dont l'effet de l'allèle exotique est favorable au caractère sont indiqués par le symbole « + ».

Le tableau ci-dessous (Tableau 2.2) rassemble les SNPs détectés pour les 7 caractères étudiés présentant un effet favorable de l'allèle exotique.

SNP	QTL	Trait	Model	Chr	Position	% var
SNP_00116	QTL_19_1	Na	A	1	21.86	0.08
SNP_00350	QTL_09_1	Na	A	1	51.00	0.04
SNP_00375	QTL_02_1	WSY	A	1	53.68	0.04
SNP_00375	QTL_02_1	RY	A	1	53.68	0.20
SNP_01273	QTL_42_1	K	A	2	14.29	0.07
SNP_01689	QTL_18_1	WS	A	2	22.23	0.08
SNP_01689	QTL_18_1	Na	A	2	22.23	0.02
SNP_01689	QTL_18_1	S	A	2	22.23	0.08
SNP_02037	QTL_08_1	N	A	2	43.73	0.09
SNP_02804	QTL_07_1	Na	A	3	50.78	0.04
SNP_03960	QTL_06_1	WSY	A	4	6.63	0.19
SNP_05988	QTL_38_1	RY	A	5	39.79	0.07
SNP_06223	QTL_01_2	N	A	5	52.91	0.10
SNP_06344	QTL_43_1	N	A	5	68.13	0.18
SNP_06428	QTL_14_1	RY	A	6	14.76	0.05
SNP_06609	QTL_13_1	N	A	6	30.98	0.05
SNP_06641	QTL_01_1	N	A	6	31.43	0.37
SNP_09633	QTL_12_1	Na	A	8	95.03	0.06
SNP_09789	QTL_33_1	N	A	9	1.57	0.10
SNP_09818	QTL_30_1	Na	A	9	10.12	0.06
SNP_09973	QTL_35_1	Na	A	9	36.71	0.05
SNP_10511	QTL_31_1	Na	A	9	99.86	0.12
SNP_10753	QTL_29_1	WS	A	9	105.93	0.09
SNP_10753	QTL_29_1	S	A	9	105.93	0.10
SNP_10753	QTL_29_1	Na	A	9	105.93	0.11

TABLE 2.2 – SNPs détectés dans la descendance (élite × exotique) comme associés à l'un des caractères étudiés et présentant un effet favorable de l'allèle exotique

Tout comme les résultats présentés dans l'article dans la partie 2.2.1, un nombre plus important de QTLs est détectés dans le panel élite par rapport à la descendance (élite × exotique), du fait de la puissance de détection plus importante dans le panel élite (plus d'individus). Plusieurs QTLs ont toutefois été détectés dans la descendance (élite × exotique), et plusieurs présentent un effet de l'allèle exotique favorable pour certains caractères. Un effet favorable vis-à-vis des caractères liés au rendement (S, WS, WSY ou encore RY) induit une augmentation de la valeur du phénotype, tandis qu'un effet favorable par rapport à un caractère d'impureté (K, N, Na) entraîne une diminution

de la valeur du phénotype. Certains QTLs présentent un effet de l'allèle exotique favorable pour plusieurs caractères. Ainsi, le QTL 02_1 est à la fois favorable au caractère RY et WSY. Cela peut être expliqué par le fait que ces deux caractères sont très corrélés, WSY étant un caractère calculé à partir du caractère mesuré RY. L'allèle exotique du QTL 18_1 est quant à lui favorable pour les caractères Na et S, et celui du QTL 29_1 est favorable aux caractères WS, S et Na. En effet, la diminution des impuretés telles que le sodium au sein de la betterave sucrière lui permet de stocker plus de sucre.

Conclusion du chapitre

L'étude présentée dans ce chapitre a permis de mettre en place la méthodologie nécessaire pour identifier l'architecture génétique de caractères agronomiques d'intérêt. La comparaison de l'architecture de ces caractères déterminée sur la descendance (élite × exotique) et sur le panel élite met en évidence l'apport favorable d'une accession exotique par rapport à un panel élite pour plusieurs caractères. Le postulat fait dans le programme AKER est donc vérifié : l'introduction de régions exotiques dans une population de betterave sucrière permet bien d'apporter de la variabilité génétique utile.

Chapitre 3

Architecture génétique du rendement racinaire des descendances AKER

L'objectif de cette thèse est de simuler et comparer différents schémas de croisements qui permettront de produire une population de pre-breeding intégrant les fragments intéressants apportés par les accessions exotiques identifiées dans le programme AKER vis-à-vis du rendement racinaire. Afin de simuler l'évolution d'une population de pre-breeding par rapport au rendement racinaire (RY), il est nécessaire de connaître l'architecture génétique de ce caractère. Le dispositif expérimental permettant d'évaluer les descendances AKER sera d'abord présenté, deux méthodes d'ajustement des phénotypes vis-à-vis des effets environnementaux seront ensuite comparés, puis la méthodologie appliquée sur la descendance (élite × exotique) au chapitre précédent sera utilisée pour identifier les QTLs liés au caractère RY de chacune des descendances AKER.

3.1 Données disponibles

Matériel végétal

L'étude porte sur les descendances issues du programme AKER. 13 des 16 descendances ont été évaluées en test-cross dans six environnements en octobre 2018, les trois autres populations n'ayant pas donné assez de graines, elles seront testées pour la première fois en 2019. Les 13 descendances ont à nouveau été évaluées en test-cross avec les trois autres populations à l'automne 2019. Seules les données issues de la campagne de phénotypage de l'automne 2018 étaient disponibles au moment de l'étude, qui a donc porté sur 13 descendances. Le nom, les effectifs composant ces descendances, le nombre de SNPs génotypés, ainsi que la sous-espèce et la provenance géographique de leur parent exotique

sont donnés dans le Tableau 3.1, et l'emplacement des six champs expérimentaux est donné dans la Figure 3.1.

descendance	Sous espèce	Origine géographique	nombre d'individus	nombre de SNPs	nombre de SNPs non redondants
801	*maritima*	Grèce	220	6908	628
802	*maritima*	France	204	7298	551
803	*vulgaris leaf beet*	France	202	6667	505
807	*adanensis*	Israël	197	7113	615
808	*vulgaris sugar beet*	Chine	198	5239	458
809	*maritima*	Danemark	195	6728	544
810	*maritima*	France	199	6682	657
811	*maritima*	Irlande	178	7396	732
812	*maritima*	Italie	201	6551	500
813	*maritima*	France	198	7420	589
815	*vulgaris sugar beet*	Etats-Unis	189	6336	539
823	*vulgaris fodder beet*	France	193	5611	471
830	*maritima*	Royaume-Uni	175	7207	505

TABLE 3.1 – Informations à propos des 13 descendances AKER étudiées

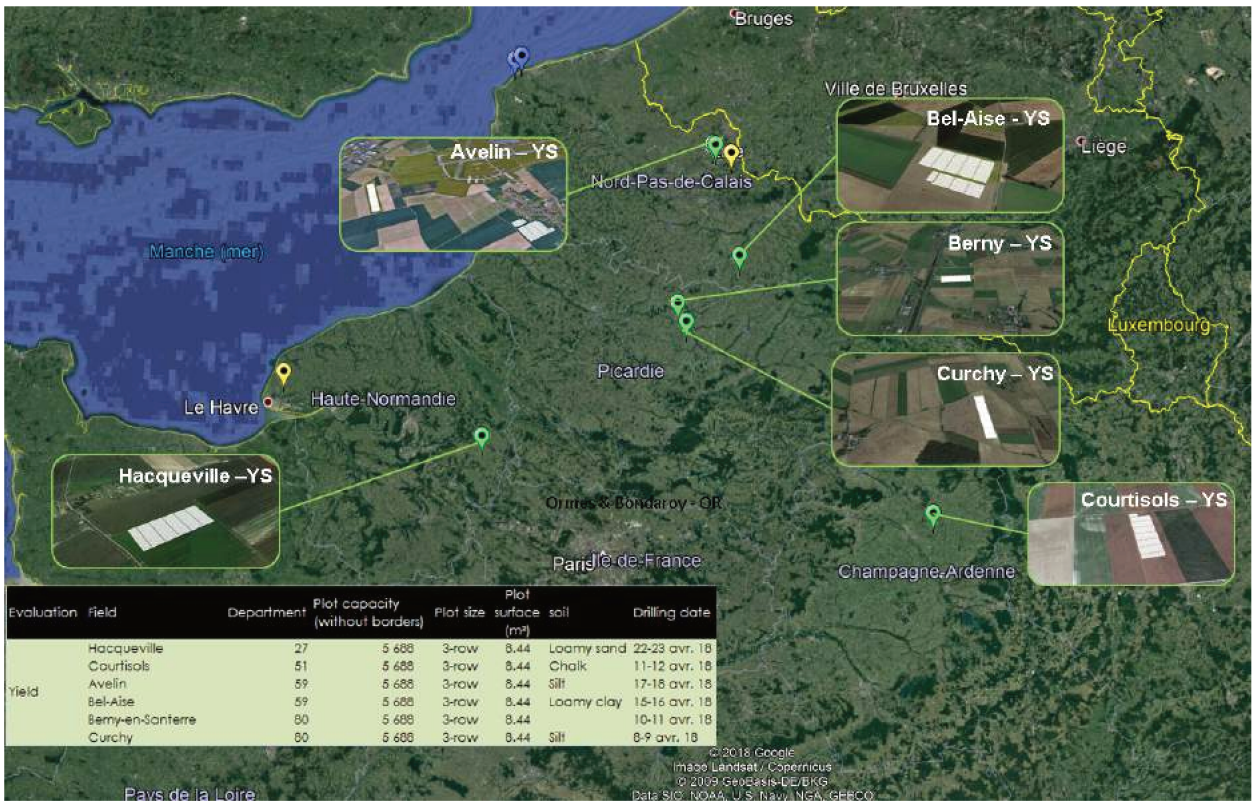


FIGURE 3.1 – Localisation des 6 champs dans lesquels les 13 descendances AKER ont été phénotypées en 2018

Données génétiques

Chacune des 13 descendances étudiées a été génotypée avec plus de 5000 SNPs. Le tableau 3.1 donne le nombre de SNPs génotypés sur chaque chromosome dans chaque descendance. Pour chacune des descendances les données de génotype manquantes sont imputées sur chacun des neuf chromosomes avec le logiciel Beagle (Browning and Browning 2009). Cette imputation n'augmente pas artificiellement la fréquence de l'allèle exotique au sein d'une descendance. Une carte génétique, propriété de Florimond Desprez Veuve & Fils, permet de connaître la position de chacun des marqueurs sur le génome de la betterave sucrière.

Données phénotypiques

Les descendances ont été réparties par « série » par le sélectionneur dans chacun des six champs d'expérimentation en 2018. Une série correspond à un ensemble de parcelles consécutives dans un champ travaillées simultanément : mêmes dates de labour, de semis, de traitements phytosanitaires et de récolte. En fonction de la place disponible dans les champs, une série peut être :

— « Complète » , c'est-à-dire composée de deux répétitions de tous les individus composant une descendance ainsi que des témoins commerciaux et génétiques (les parents à l'origine de la descendance).

— « Incomplète » , lorsqu'elle est composée de deux répétitions d'une descendance pour laquelle certains individus sont placés dans une autre série à cause de la configuration du champ. Les témoins sont en revanche toujours présents dans toutes les séries.

— « Complète + » , composée de deux répétitions d'une descendance complète avec les témoins, auxquels sont ajoutés les individus manquants d'une descendance incomplète.

La figure 3.2 représente la répartition des séries, des témoins, des répétitions et des descendances dans l'un des environnements étudiés. Les autres environnements sont représentés en annexes (Figures A.1, A.2, A.3, A.4, A.5)

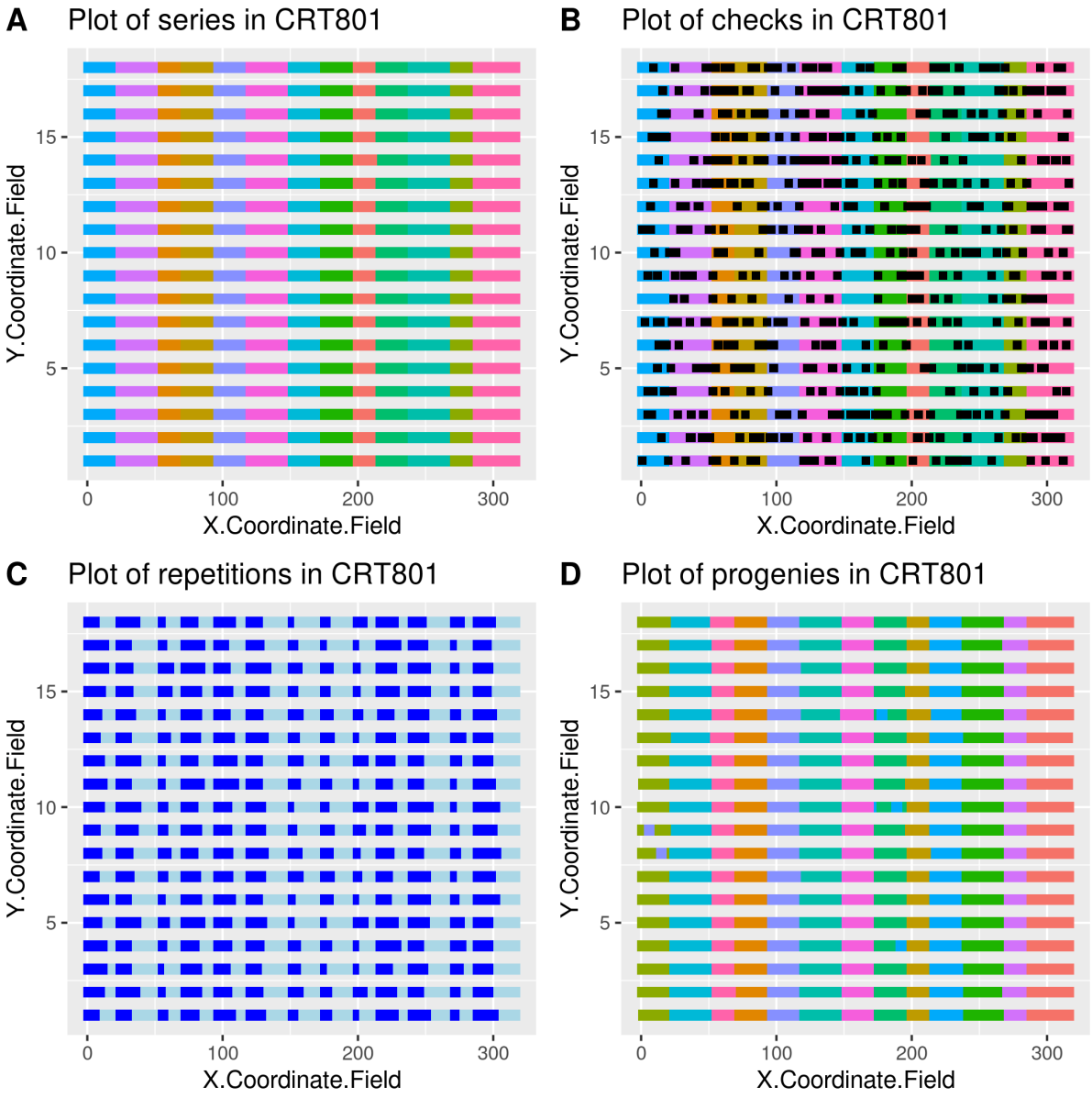


FIGURE 3.2 – Dispositif expérimental de l'environnement CRT801 : A. Répartition des individus dans le champ, chaque couleur correspondant à une série ; B. Position des témoins (en noir) dans les séries ; C. Répartition des répétitions par série dans le champ (bleu foncé pour la répétition 1, bleu clair pour la répétition 2) ; D. répartition des individus des 13 descendances dans le champ, chaque couleur correspondant à une descendance

Le rendement racinaire (RY, mesuré en tonnes/ha) a été évalué pour tous les individus de ce dispositif expérimental.

3.2 Ajustement des phénotypes vis à vis des effets terrain

Concernant la mise au point du modèle d'ajustement des effets terrain sur le phénotype, plusieurs questions peuvent se poser. Tout d'abord, l'effet des génotypes doit-il être fixe

ou aléatoire ? Si cet effet est considéré comme aléatoire, la variabilité des phénotypes ajustés sera écrasée par l'ajustement. Or la puissance de la détection de QTL repose sur l'écart des phénotypes entre deux classes de génotypes : considérer un effet aléatoire des génotypes nous ferait ainsi perdre en puissance de détection. En revanche, si l'on considère l'effet des génotypes comme étant fixe, chaque génotype a sa propre moyenne puisqu'il est répété deux fois dans le champ. La variabilité des phénotypes sera donc plus importante qu'avec un modèle aléatoire, ce qui fournira une meilleure puissance pour la détection de QTLs. Il a donc été décidé de considérer les génotypes en effet fixe.
On peut ensuite se demander s'il est nécessaire d'inclure l'effet des descendances dans le modèle. Cependant, puisque l'effet des génotypes est fixe, on dispose déjà de la moyenne de chaque descendance qui est égale à la moyenne des génotypes la composant. Inclure l'effet descendance en effet fixe dans le modèle ne conduit qu'à surparamétrer le modèle, avec un effet qui n'a pas d'intérêt en soi.
De la même façon, on peut se demander s'il faut ajouter l'effet des témoins. Les témoins étant compris dans les génotypes, utiliser l'effet des génotypes en fixe permet ainsi d'obtenir la moyenne propre à chacun des témoins et de n'avoir aucune nécessité à rajouter l'effet témoin dans le modèle.
L'effet des répétitions est conservé et classiquement considéré comme fixe.
Les champs étant très étendus, deux autres variables sont à prendre en compte dans notre modèle : les effets lignes et les effets colonnes. Deux modèles considérant respectivement ces effets en fixe ou en aléatoire ont été comparés. Les ajustements obtenus sont très corrélés, on peut donc donc choisir indifféremment l'un ou l'autre.
Le package R SpATS proposant la méthode d'ajustement spatiale SpATS (Spatial Analysis of field Trials with Splines) a aussi été utilisé `https://CRAN.R-project.org/package=SpATS`. Cette méthode permet d'effectuer un ajustement plus fin des phénotypes par l'usage local de splines, fonctions mathématiques définies par morceau grâce à des polynômes qui permettent de lisser les données. Afin de visualiser l'impact des splines sur l'ajustement, ce modèle est comparé à un modèle plus simple, sans splines.
Les deux modèles comparés sont les suivants :

$$y = f(\boldsymbol{u}, \boldsymbol{v}) + \boldsymbol{X}_g\boldsymbol{\beta}_g + \boldsymbol{X}_n\boldsymbol{\beta}_n + \boldsymbol{Z}_r\boldsymbol{c}_r + \boldsymbol{Z}_c\boldsymbol{c}_c + \boldsymbol{\epsilon} \text{ (SpATS)}$$

et :

$$y = \boldsymbol{X}_g\boldsymbol{\beta}_g + \boldsymbol{X}_n\boldsymbol{\beta}_n + \boldsymbol{Z}_r\boldsymbol{c}_r + \boldsymbol{Z}_c\boldsymbol{c}_c + \boldsymbol{\epsilon} \text{ (lme4)}$$

avec :

- y le vecteur des phénotypes
- $\boldsymbol{\beta}_g$ le facteur fixe des génotypes
- $f(\boldsymbol{u}, \boldsymbol{v})$ la fonction de lissage bivariée utilisée par SpATS pour calculer les splines, avec u le vecteur numérique des lignes, et v le vecteur numérique des colonnes
- $\boldsymbol{\beta}_n$ le facteur fixe des répétitions
- $\boldsymbol{c}_r$ le facteur aléatoire des lignes, avec $\boldsymbol{c}_r \sim \mathcal{N}(0, \sigma_r^2 \boldsymbol{Id})$
- $\boldsymbol{c}_c$ le facteur aléatoire des colonnes, avec $\boldsymbol{c}_c \sim \mathcal{N}(0, \sigma_c^2 \boldsymbol{Id})$
- $\boldsymbol{\epsilon}$ l'erreur résiduelle, avec $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_\epsilon^2 \boldsymbol{Id})$
- $\boldsymbol{X}_g$, $\boldsymbol{X}_n$, $\boldsymbol{Z}_r$ et $\boldsymbol{X}_c$ les matrices de design associées respectivement aux génotypes, aux répétitions, aux lignes et aux colonnes

Ces deux modèles permettent d'obtenir la moyenne phénotypique ajustée de chaque individu de l'ensemble des 13 descendances sur chacun des 6 environnements, définie comme $\hat{y}_i = \hat{\mu} + \hat{g}_i$ avec $\hat{\mu}$ la moyenne globale estimée et $\hat{g}_i$ la valeur génétique estimée de l'individu i. La figure 3.3 représente les corrélations des moyennes phénotypiques ajustées par les modèles SpATS et lme4 pour chacune des 13 descendances dans chacun des 6 environnements.

FIGURE 3.3 – Corrélation des moyennes phénotypiques ajustées par la méthode SpATS et par la méthode lme4 pour chacune des 13 descendances dans chacun des 6 environnements

Les ajustements réalisés par ces deux modèles sont corrélés positivement, les valeurs des corrélations allant de 0.5 pour la descendance 808 dans l'environnement CRT801 à 1 pour la descendance 807 dans l'environnement CUR801. De façon générale, les corrélations les plus faibles sont trouvées dans l'environnement CRT801 et les plus élevées dans l'environnement CUR801. Il existe donc une différence d'ajustement entre les deux modèles en fonction de l'environnement analysé. Une étude est alors réalisée pour décider du modèle à retenir pour la suite de l'analyse. Les deux modèles sont comparés sur la descendance 808 dans l'environnement CRT801, soit dans les conditions où leur corrélation est la plus faible.

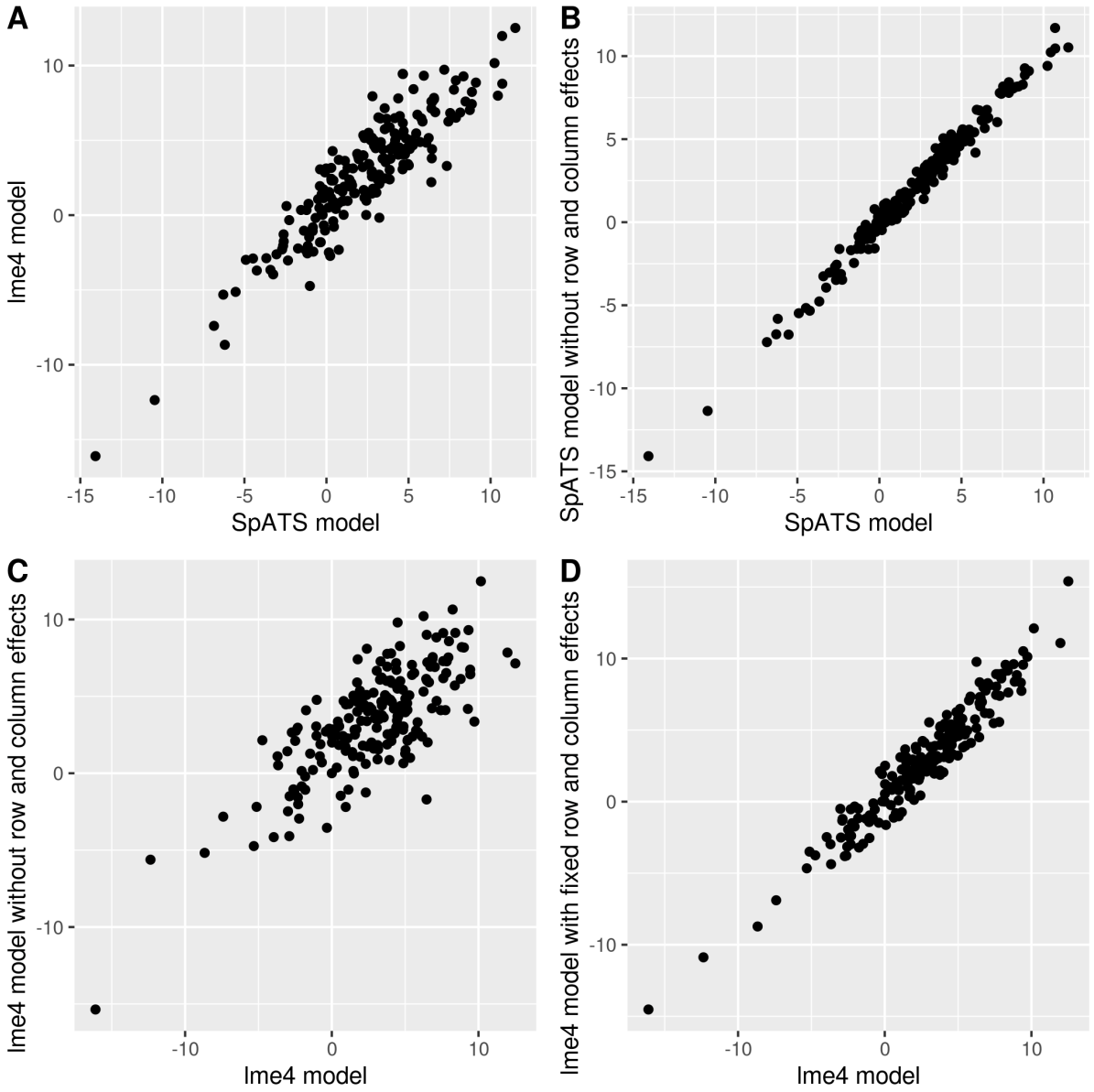


FIGURE 3.4 – Comparaison de l'ajustement de la descendance 808 sur l'environnement CRT801 par différents modèles : A) modèle (SpATS) et modèle (lme4) ; B) modèle (SpATS) et modèle (SpATS) sans effet ligne ni colonne ; C) modèle (lme4) et modèle (lme4) sans effet ligne ni colonne ; D) modèle (lme4) et modèle (lme4) avec effets ligne et colonne fixes

Il apparaît dans la Figure 3.4 que la méthode SpATS est moins sensible que le modèle lme4 au changement de modélisation des facteurs ligne et colonne : la variabilité due à l'effet terrain est capturée par les splines. SpATS étant plus robuste que lme4, c'est ce modèle qui est retenu pour l'ajustement des données phénotypiques.

Afin d'obtenir la valeur génotypique moyenne sur l'ensemble des environnements, un modèle avec l'effet des génotypes aléatoire est utilisé à la suite de l'ajustement réalisé avec SpATS. La valeur génotypique moyenne de chaque génotype est donc calculée sur

l'ensemble des environnements à partir du modèle suivant :

$$\hat{y} = \boldsymbol{Z}_g \boldsymbol{G} + \boldsymbol{X}_f \boldsymbol{\beta}_f + \boldsymbol{\epsilon}$$

avec :

- $\hat{y}$ le vecteur des moyennes ajustées
- $\boldsymbol{G}$ le facteur aléatoire des génotypes avec $\boldsymbol{c}_g \sim \mathcal{N}(0, \sigma_g^2 \boldsymbol{Id})$
- $\boldsymbol{\beta}_f$ le facteur fixe des champs
- $\boldsymbol{\epsilon}$ l'erreur résiduelle, avec $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_\epsilon^2 \boldsymbol{Id})$
- $\boldsymbol{Z}_g$ et $\boldsymbol{X}_f$ les matrices de design associées respectivement aux génotypes et aux champs

L'effet génotype est ici en aléatoire afin de rendre le modèle plus robuste au déséquilibre. En effet tous les génotypes ne sont pas forcément présents le même nombre de fois dans l'ensemble des champs. L'effet champ quant à lui est choisi comme fixe.

Pour obtenir l'estimateur de la valeur génotypique moyenne un ajustement en deux étapes a donc été effectué : un premier ajustement intra-environnement a tout d'abord été réalisé avec la méthode SpATS, puis la valeur génotypique moyenne a été estimée sur l'ensemble des environnements. Il n'était pas possible d'utiliser la méthode SpATS pour faire un ajustement en une seule étape. En effet cette méthode ne permet pas d'attribuer une variance spécifique à chaque environnement. Un postulat fort selon lequel la variabilité est la même dans tous les environnements aurait dû tre posé, ce qui est très probablement faux. De plus, comme justifié dans le paragraphe précédent, l'effet moyen des génotypes est choisi comme aléatoire pour éviter un biais de déséquilibre. Dans le cadre d'un ajustement en une étape, cela aurait impliqué que les effets des génotypes intra-environnement soient également considérés comme aléatoires. La variabilité de ces effets génotype aurait pu être modélisée en fonction de la descendance mais en tant qu'effet aléatoire une réduction de la variabilité des phénotypes ajustés et donc une diminution de la puissance de détection de QTLs auraient été créés. Un paramètre permettant de prendre en compte l'effet de chaque descendance aurait également dû être ajouté, cet effet ne pouvant plus être déduit de la moyenne des génotypes composant chaque descendance. Au vu de tous ces éléments, effectuer un ajustement des effets environnementaux en deux étapes plutôt qu'en une seule permet d'obtenir une valeur génotypique moyenne plus précise, tout en minimisant la perte d'information utile pour détecter des QTLs liés à des caractères d'intérêt.

3.3 Architecture génétique

Etude de l'héritabilité du caractère RY

Suite à l'ajustement des effets terrains par le package R SpATS, l'héritabilité du caractère RY a été calculée à partir des paramètres estimés dans le modèle permettant de calculer la valeur génétique moyenne :

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_e^2}{n_r}}$$

Avec H^2 l'héritabilité, σ_g^2 la variance génétique, σ_e^2 la variance résiduelle et n_r le nombre moyen de répétitions.

Les résultats sont présentés dans le Tableau 3.2

descendance	h^2
801	0.66
802	0.67
803	0.62
807	0.94
808	0.29
809	0.78
810	0.43
811	0.55
812	0.89
813	0.64
815	0.72
823	0.14
830	0.46

TABLE 3.2 – Héritabilité du caractère RY dans les 13 descendances AKER

L'héritabilité du caractère est très différente en fonction de la descendance étudiée, allant de 0.14 dans la descendance 823 à 0.94 dans la descendance 807. Un nombre très différent de QTLs détectés est donc attendu pour les différentes descendances. Ces différentes héritabilités peuvent être expliquées par la provenance des accessions exotiques à l'origine des descendances AKER. L'espèce *Beta vulgaris* est subdivisée en trois sous-espèces, les *Beta vulgaris subsp. maritima*, qui sont des bettes sauvages trouvées en bord de mer, les *Beta vulgaris subsp. vulgaris*, qui regroupent les variétés cultivées, et les *Beta vulgaris subsp. adanensis*, qui sont des bettes sauvages présentes entre la Grèce et la Syrie. D'après

le Tableau 3.1, parmi les 13 accessions exotiques à l'origine des descendances AKER étudiées, huit font partie de la sous-espèce *maritima*. Dans ce sous-groupe l'héritabilité varie entre 0.43 et 0.89. Ces huit accessions proviennent de 6 pays européens, le Danemark, la France, la Grèce, l'Italie, l'Irlande et le Royaume-Uni, ce qui peut expliquer la variabilité de l'hérédité du rendement racinaire calculé dans les descendances issues de ces différentes accessions. Quatre accessions exotiques appartiennent quant à elles à la sous-espèce *vulgaris*, ayant produit des descendances où une héritabilité de 0.14 à 0.72 est observée. Deux d'entre elles sont des betteraves sucrières, une autre est une une bette, et la dernière est une betterave fourragère à l'origine de la descendance 823, qui présente l'héritabilité la plus faible du rendement racinaire toutes descendances confondues. L'origine géographique de ces quatre accessions est très disparate, regroupant la Chine, les Etats-Unis, la France et la Grèce. L'éloignement géographique ainsi que la différence de sous-espèce peut justifier les écarts d'héritabilité observés. Enfin, la dernière accession exotique appartient à la sous-espèce *adanensis*. Il s'agit de l'accession dont est issue la descendance 807, qui présente l'héritabilité la plus élevée toutes descendances confondues (0.94). Il s'agit de la seule descendance provenant de cette sous-espèce, elle est par conséquent très différente des autres descendances, ce qui peut expliquer que l'héritabilité du rendement racinaire qui y est calculé soit supérieure à celle calculée dans les autres descendances. En moyenne, l'héritabilité du rendement racinaire de ce caractère est de 0.60 sur l'ensemble des descendances. C'est cette valeur moyenne qui sera utilisée pour simuler le caractère dans les schémas de pre-breeding (chapitre 4).

Analyses d'association

La méthodologie décrite dans le chapitre 2 a été utilisée pour identifier l'architecture génétique du rendement racinaire sur chacune des 13 descendances issues d'AKER. La détection de SNP a été réalisée grâce à une méthode similaire à celle présentée par Segura et al. (2012), implémentée dans le package mlmm.gwas (`https://CRAN.R-project.org/package=mlmm.gwas`). Cette méthode est basée sur une analyse forward. A chaque étape le SNP ayant la p-valeur la plus faible est ajouté comme régresseur dans le modèle contenant l'effet polygénique aléatoire, ici le modèle additif de Yu et al. (2006). L'effet de chaque SNP est ainsi recalculé à chaque étape en fonction des SNPs présents comme régresseurs. Cette analyse forward de sélection des SNPs s'interrompt lorsque la part de variance expliquée par l'effet polygénique est proche de zéro. Un critère de sélection de modèles, ici le critère eBIC, permet ensuite de retenir le modèle le plus parcimonieux, c'est-à-dire celui expliquant la plus grande part de variance avec le moins de SNPs. Les effets de chacun des SNPs sont estimés dans ce modèle. Les SNPs détectés comme associés au caractère RY et leurs effets sont donnés dans le Tableau 3.3.

descendance	SNP	lg	pos	effet exotique (t/ha)
801	AX_124324125	2	54.93	-1.39
802	AX_124326121	3	46.91	-1.66
802	AX_124331965	7	45.99	-1.90
803	AX_124324480	2	60.32	-5.08
803	AX_124328171	5	19.47	1.42
803	AX_124329055	5	68.11	0.76
803	AX_124329646	6	32.39	1.42
803	AX_124332240	8	12.65	1.71
803	AX_124341522	3	49.02	-1.45
807	AX_124323553	1	5.92	-0.55
807	AX_124324024	2	51.69	-2.10
807	AX_124325629	3	52.19	0.83
807	AX_124328249	5	25.43	-0.59
807	AX_124329106	5	38.00	-1.67
807	AX_124336063	2	58.21	-3.55
807	AX_124341093	2	53.45	1.49
808	AX_124333010	8	53.17	-2.61
810	AX_124333118	8	64.03	-2.45
811	AX_124326245	3	50.78	-2.11
812	AX_124313705	1	33.06	0.46
812	AX_124315511	5	11.31	-1.01
812	AX_124323948	2	47.26	-0.77
812	AX_124324167	2	58.21	-3.53
812	AX_124328086	5	10.00	0.60
812	AX_124328447	5	32.78	-0.13
812	AX_124328455	5	32.78	0.00
812	AX_124332257	8	16.53	0.74
812	AX_124341386	3	52.17	-0.38
813	AX_124323777	2	31.40	-2.52
813	AX_124324046	2	53.45	-1.77
813	AX_124326112	3	48.04	-1.63
813	AX_124329747	6	42.31	-4.34
813	AX_124332257	8	16.53	1.18
815	AX_124313575	4	11.72	-2.89
830	AX_124316288	2	41.9	-1.25
830	AX_124326218	3	49.02	-2.20

TABLE 3.3 – SNPs détectés dans les 13 descendances AKER. Ces SNPs sont localisés sur le groupe de liaison donné dans la colonne « LG », à la position donnée dans la colonne « pos » (en cM), et l'effet de l'allèle exotique au SNP sur le caractère RY est donné dans la colonne « effet exotique (t/ha) »

Le nombre de SNPs détectés comme étant associé au caractère RY varie fortement en fonction de la descendance étudiée. Aucun SNP n'est détecté dans les descendances 809 et 823. Si ce résultat pouvait être attendu dans le cas de la descendance 823, pour laquelle l'héritabilité du caractère RY est la plus faible (0.14), il est plus surprenant dans le cas de la descendance 809 pour laquelle l'héritabilité du caractère RY est de 0.78. Après contrôle des fichiers de sortie de l'analyse d'association, il s'avère qu'un SNP avait bel et bien été détecté dans la population 809 mais n'a pas bien été intégré dans les étapes de regroupement de SNPs ultérieures. Ce SNP est localisé sur le chromosome 2 en position 60.32cM, il est donc colocalisé avec le SNP AX_124324480 détecté dans la population 803. L'effet de l'allèle exotique au niveau de ce SNP est de -4.05 t/ha, quand cet effet au niveau du SNP détecté dans la population 803 est de -5.08 t/ha. Les allèles exotiques apportés par les populations 803 et 809 à ce locus ont donc un effet défavorable sur le rendement racinaire. Cependant, compte tenu de sa découverte tardive, le SNP détecté dans la population 809 n'a pas été intégré dans les simulations de schéma de pre-breeding présentées dans le chapitre 4. Puisque l'objectif des simulations est d'étudier l'évolution de la performance et de la diversité génétique de la population au cours de différents schémas de pre-breeding, cette omission n'a pas un impact trop important dans le cadre de cette étude.

Les descendances comptabilisant le plus de SNPs détectés comme étant liés au caractère RY sont les descendances 812 et 807, totalisant respectivement 9 SNPs et 7 SNPs. Ce résultat pouvait être attendu au vu des valeurs élevées de l'héritabilité du caractère RY dans ces descendances, atteignant respectivement 0.89 et 0.94. Le tableau A.1 en annexe présente la position génétique de tous les SNPs détectés dans chaque descendance ainsi que des marqueurs redondants qui y sont associés. La position de l'ensemble de ces SNPs est représentée sur chaque LG dans la Figure A.7 en annexe.
Un total de 36 SNPs détectés est donné dans le Tableau 3.3. Cependant, le SNP AX_124332257 est détecté à la fois dans la descendance 812 et dans la descendance 813 : seuls 35 SNPs distincts sont détectés. Dans ces deux descendances l'effet de l'allèle exotique à ce locus est positif (respectivement +0.74 t/ha et +1.18 t/ha). L'effet de l'allèle exotique à ce locus est donc en moyenne de +0.96 t/ha.
Les SNPs AX_124328447 et AX_124328455 sont tous deux des SNPs génotypés et détectés comme étant associés au caractère RY dans la descendance 812. Ces deux marqueurs non redondants sont pourtant colocalisés et en déséquilibre de liaison. L'effet de l'allèle exotique à chacun de ces marqueur est respectivement -0.13t/ha et 0. La variabilité génétique du rendement dans cette région est donc mieux définie par ces deux SNPs que par un seul. Cependant l'architecture génétique déterminée doit, par la suite, être utilisée

pour simuler le rendement racinaire. Or, le simulateur n'est pas conçu pour prendre en compte des SNPs en déséquilibre de liaison. C'est pourquoi seul le SNP détecté en premier est conservé. Cela porte à 34 le nombre de SNPs détectés comme associés au caractère RY dans l'ensemble des 13 descendances.
Parmi les 34 SNPs détectés, 9 présentent un effet favorable de l'allèle exotique et 25 un effet défavorable de cet allèle. Le rendement racinaire étant un caractère étudié par les sélectionneurs, la plupart des allèles qui lui sont favorables sont déjà intégrés dans les accessions élites, ce qui explique que la majorité des allèles de source exotique impactent négativement ce caractère. Cependant cette étude permet de mettre en évidence 9 nouveau locus intéressants apportés par les accessions exotiques. Parmi ces locus « favorables », 4 ont été détectés dans la descendance 803, originaire de l'accession exotique *Beta vulgaris vulgaris leaf beet* de Grèce, 3 l'ont été dans la descendance 812, originaire de l'accession exotique *Beta vulgaris maritima* d'Italie (dont un est également détecté dans la descendance 813, dont le parent exotique est une *Beta vulgaris maritima* de France), et les 2 derniers sont trouvés dans la descendance 807, originaires de la *Beta vulgaris adanensis* d'Israël. Chacune des trois sous-espèces de *Beta vulgaris* peut donc apporter plusieurs fragments exotiques « favorables » pour le rendement racinaire.

Ordonner le Tableau présentant les SNPs détectés par position sur les chromosomes permet de remarquer que certains SNPs détectés dans différentes descendances sont colocalisés (Tableau 3.4). Le SNP AX_124324167 détecté dans la descendance 812 est colocalisé avec le SNP AX_124336063 détecté dans la descendance 807 (LG2, 58.21cM). Les effets de l'allèle exotique au niveau de ces deux marqueurs sont défavorables pour le caractère RY, atteignant respectivement -3.53 t/ha et -3.55 t/ha, soit une moyenne de -3.54 t/ha. Les descendances 812 et 807 proviennent respectivement des accessions exotiques *Beta vulgaris maritima* d'Italie et *Beta vulgaris adanensis* d'Israël. Bien qu'appartenant à deux sous-espèces différentes et étant éloignées géographiquement, ces deux accessions exotiques ont donc un effet similaire sur le rendement racinaire à ce locus. De la même façon, le SNP AX_124326218 détecté dans la descendance 830 (provenant de la *Beta vulgaris maritima* du Royaume-Uni) et le SNP AX_124341522 détecté dans la descendance 803 (provenant de la *Beta vulgaris vulgaris leaf beet* de Grèce) sont colocalisés (LG3, 49.02cM) et les effets de l'allèle exotique à ces deux marqueurs sont défavorables au caractère RY, atteignant respectivement -2.20 t/ha et -1.45 t/ha soit une moyenne de -1.83 t/ha.

	SNP	descendance	effet exotique	lg	pos
1	AX_124323553	807	-0.55	1	5.92
2	AX_124313705	812	0.46	1	33.06
3	AX_124323777	813	-2.52	2	31.40
4	AX_124316288	830	-1.25	2	41.90
5	AX_124323948	812	-0.77	2	47.26
6	AX_124324024	807	-2.10	2	51.69
7	AX_124324046	813	-1.77	2	53.45
8	AX_124341093	807	1.49	2	53.45
9	AX_124324125	801	-1.39	2	54.93
10	AX_124324167	812	-3.53	2	58.21
11	AX_124336063	807	-3.55	2	58.21
12	AX_124324480	803	-5.08	2	60.32
13	AX_124326121	802	-1.66	3	46.91
14	AX_124326112	813	-1.63	3	48.04
15	AX_124326218	830	-2.20	3	49.02
16	AX_124341522	803	-1.45	3	49.02
17	AX_124326245	811	-2.11	3	50.78
18	AX_124341386	812	-0.38	3	52.17
19	AX_124325629	807	0.83	3	52.19
20	AX_124313575	815	-2.89	4	11.72
21	AX_124328086	812	0.60	5	10.00
22	AX_124315511	812	-1.01	5	11.31
23	AX_124328171	803	1.42	5	19.47
24	AX_124328249	807	-0.59	5	25.43
25	AX_124328447	812	-0.13	5	32.78
26	AX_124329106	807	-1.67	5	38.00
27	AX_124329055	803	0.76	5	68.11
28	AX_124329646	803	1.42	6	32.39
29	AX_124329747	813	-4.34	6	42.31
30	AX_124331965	802	-1.90	7	45.99
31	AX_124332240	803	1.71	8	12.65
32	AX_124332257	812	0.74	8	16.53
33	AX_124332257	813	1.18	8	16.53
34	AX_124333010	808	-2.61	8	53.17
35	AX_124333118	810	-2.45	8	64.03

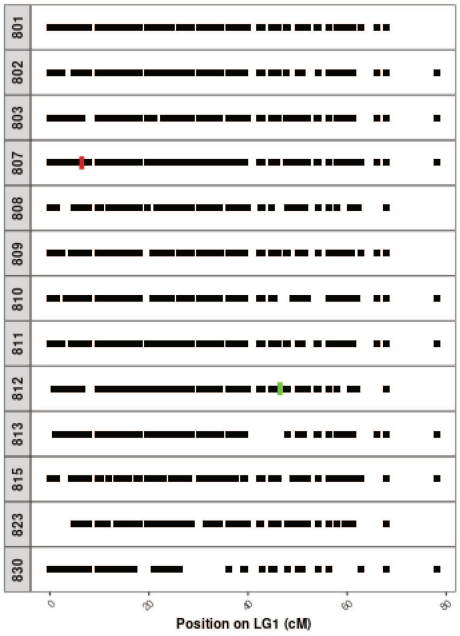
TABLE 3.4 – SNPs détectés comme associés à RY dans les 13 descendances AKER ordonnés par position (cM)

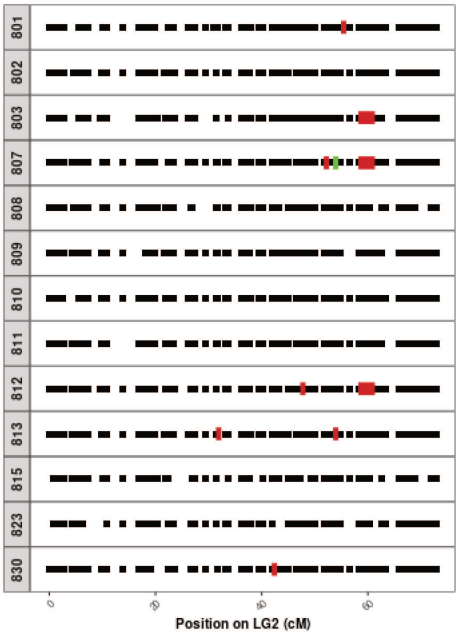
Mais les effets des allèles exotiques au niveau de marqueurs colocalisés peuvent également être antagonistes. En effet le SNP AX_124324046 détecté dans la descendance 813 (provenant d'une *Beta vulgaris maritima* de France) est colocalisé avec le SNP AX_124341093 détecté dans la descendance 807 (provenant de la *Beta vulgaris adanensis* d'Israël) (LG2, 53.45cM), et les effets de l'allèle exotique à ces marqueurs sont opposés, atteignant respectivement de -1.77 tonnes/ha et de +1.49 tonnes/ha. Il est possible que le gène lié à ces QTLs, commun aux deux sous-espèces à l'origine, ait ensuite évolué différemment dans chacune d'elles. L'effet de l'allèle exotique au niveau d'un QTL dépend donc de l'accession exotique ayant fourni l'allèle. Il est donc nécessaire de conserver l'information concernant la structuration des individus en 13 descendances distinctes lors des simulations ultérieures de schémas de pre-breeding.
Pour finir, une étude de DL est réalisée afin de déterminer si les 34 SNPs détectés dans cette analyse représentent des régions distinctes du génome de la betterave sucrière ou s'ils peuvent être regroupés en QTL.

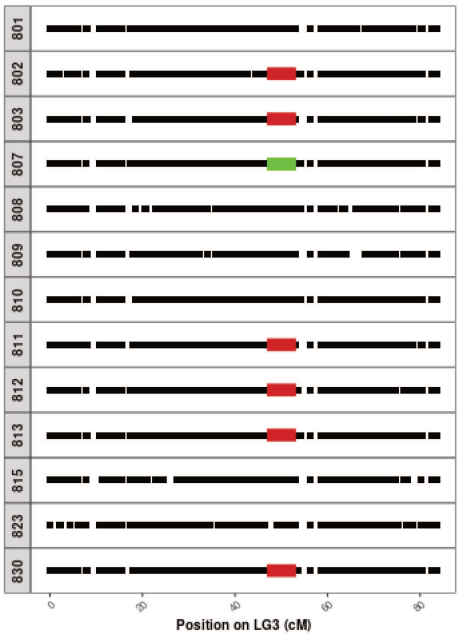
Etude du déséquilibre de liaison

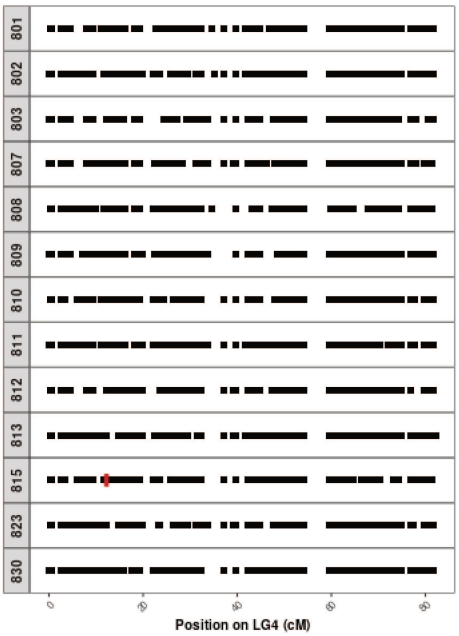
Pour chacune des 13 descendances AKER, seules les données de génotypage des SNPs pour lesquels la descendance présente un polymorphisme sont disponibles. Tous les SNPs, dont les 34 SNPs détectés précédemment, ne sont donc pas génotypés dans toutes les descendances. Afin de disposer des mêmes marqueurs dans toutes les descendances, une nouvelle étape d'imputation des génotypes avec le logiciel Beagle est réalisée. Le nombre de marqueurs non redondants s'élève ainsi à 8924 pour 2549 individus.
Il est alors possible de réaliser une étude de DL sur l'ensemble des descendances en prenant en compte la structure en descendance, selon la méthodologie présentée dans le chapitre 2. Le seuil de l'étendue de DL, c'est-à-dire la valeur au delà de laquelle deux SNPs sont considérés comme étant statistiquement en déséquilibre de liaison et donc faisant partie d'un même QTL, est déterminé en calculant la corrélation de Pearson corrigée par l'apparentement entre 10 000 paires de SNPs indépendants (situés sur deux chromosomes distincts) échantillonnées aléatoirement, et en prenant le 99ème quantile de cette distribution. Ce seuil est de $1,97\text{e}^{-03}$.
Le tableau 3.5 représente les SNPs détectés comme associés avec le caractère RY regroupés en QTLs. Pour rappel, un QTL est défini comme rassemblant un ou plusieurs SNPs détectés comme associés avec le caractère étudié, ces SNPs doivent être situés sur un même chromosome, à moins de 5cM d'intervalle, et leur DL doit être supérieur à un seuil déterminé.
L'analyse de DL a permis de grouper les 34 SNPs détectés dans les différentes descendances en 24 QTLs distincts. 4 QTLs rassemblent de 2 à 7 SNPs, les 20 autres ne comportant

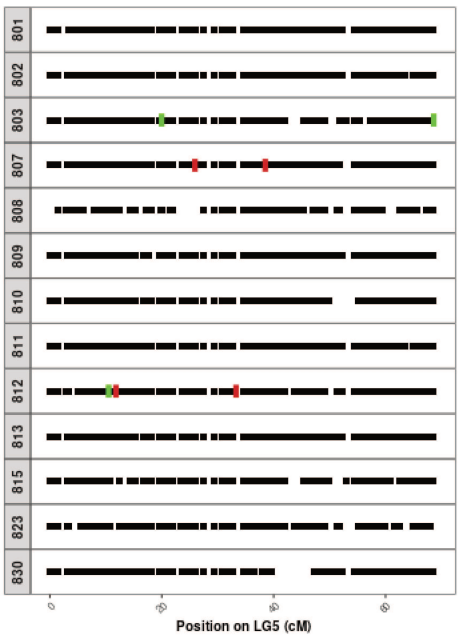
qu'un seul marqueur détecté. Ces 4 QTLs sont rapidement décrits ci-après. Le premier QTL regroupe 7 SNPs détectés entre 46.91cM et 52.19cM sur le chromosome 3 dans 7 descendances différentes. Les effets de l'allèle exotique à ces marqueurs détectés sont tous défavorables vis-à-vis du caractère RY, avec des pertes de rendement racinaires allant de 0,38 tonnes/ha à 2,20 tonnes/ha, excepté pour le SNP détecté dans la descendance 807 pour lequel le rendement racinaire augmente de 0,83 tonnes/ha. La descendance 807 provient de l'accession exotique *Beta vulgaris adanensis*, tandis que 6 des autres descendances dans lesquelles le QTL est trouvé sont issues de *Beta vulgaris maritima* et la dernière de *Beta vulgaris vulgaris leaf beet*. Ces effets opposés mettent en exergue l'importance de connaître la descendance transmettant l'allèle exotique dans les simulations ultérieures des schémas de pre-breeding. La même observation peut être faite vis-à-vis du QTL 3 qui regroupe 2 SNPs colocalisés sur le chromosome 2 à la position 53,45, et dont les effets de l'allèle exotique sont favorables ou défavorables vis-à-vis du caractère RY en fonction de la descendance dont l'allèle exotique provient. En revanche les effets de l'allèle exotique au niveau du QTL 2 qui contiennent 3 SNPs détectés sur le chromosome 2 entre 58,21cM et 60,32cM sont défavorables pour le caractère RY quelle que soit la descendance transmettant les allèles. Enfin, les effets de l'allèle exotique au niveau du QTL 4, composé de 2 SNPs détectés sur le chromosome 8 entre 12,65cM et 16,53cM, sont favorables au caractère RY quelle que soit la descendance d'origine des allèles. Au total, parmi les 24 QTLs, 16 ont un effet défavorable de l'allèle exotique sur le caractère RY, 6 ont un effet favorable, et 2 ont un effet favorable ou défavorable en fonction de la descendance dont sont originaires les allèles. L'effet de ces QTLs est considéré comme nul dans les descendances où ils ne sont pas détectés. Ces QTLs sont représentés graphiquement sur la Figure 3.5.

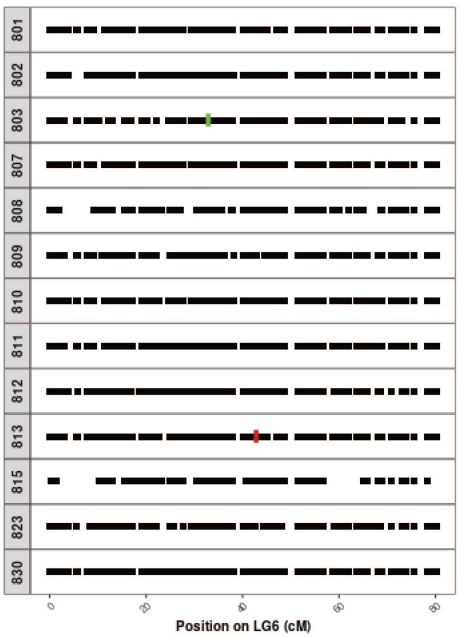

801
802
803
807
808
809
810
811
812
813
815
823
830
0
20
40
60
80
Position on LG1 (cM)


801
802
803
807
808
809
810
811
812
813
815
823
830
0
20
40
60
80
Position on LG2 (cM)


801
802
803
807
808
809
810
811
812
813
815
823
830
0
20
40
60
80
Position on LG3 (cM)


801
802
803
807
808
809
810
811
812
813
815
823
830
0
20
40
60
80
Position on LG4 (cM)


801
802
803
807
808
809
810
811
812
813
815
823
830
0
20
40
60
80
Position on LG5 (cM)


801
802
803
807
808
809
810
811
812
813
815
823
830
0
20
40
60
80
Position on LG6 (cM)

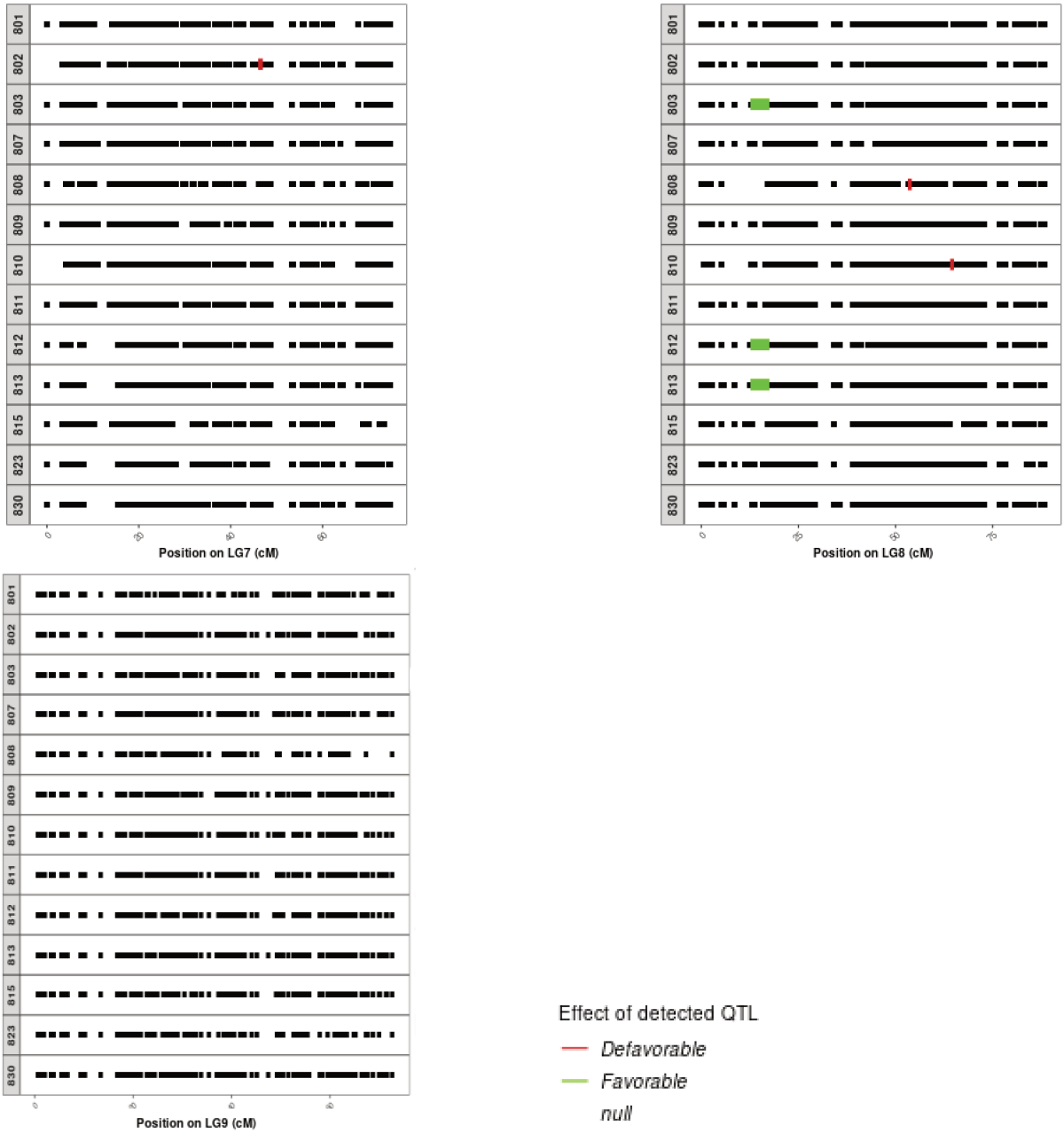


FIGURE 3.5 – QTLs détectés comme associés avec le rendement racinaire dans chacune des 13 descendances AKER sur les 9 chromosomes de la betterave sucrière, en vert lorsque l'effet de l'allèle exotique est favorable vis-à-vis du caractère, en rouge lorsqu'il est défavorable

L'étude de l'architecture génétique du caractère RY présentée dans ce chapitre 3 a donc permis de détecter 24 QTLs, répartis sur huit des neuf chromosomes de la betterave sucrière. Environ 25% de ces QTLs présentent un effet favorable de l'allèle exotique vis-à-vis du caractère RY, mais cet effet est variable en fonction des descendances dont cet allèle est originaire. Cette architecture génétique va être utilisée pour simuler l'évolution du caractère RY en fonction de différents schémas de pre-breeding.

	QTL	SNP	LG	pos	descendance	effet exotique
1	QTL_01	AX_124326121	3	46.91	802	-1.66
2	QTL_01	AX_124326112	3	48.04	813	-1.63
3	QTL_01	AX_124326218	3	49.02	830	-2.20
4	QTL_01	AX_124341522	3	49.02	803	-1.45
5	QTL_01	AX_124326245	3	50.78	811	-2.11
6	QTL_01	AX_124341386	3	52.17	812	-0.38
7	QTL_01	AX_124325629	3	52.19	807	0.83
8	QTL_02	AX_124324167	2	58.21	812	-3.53
9	QTL_02	AX_124336063	2	58.21	807	-3.55
10	QTL_02	AX_124324480	2	60.32	803	-5.08
11	QTL_03	AX_124324046	2	53.45	813	-1.77
12	QTL_03	AX_124341093	2	53.45	807	1.49
13	QTL_04	AX_124332240	8	12.65	803	1.71
14	QTL_04	AX_124332257	8	16.53	812	0.74
15	QTL_04	AX_124332257	8	16.53	813	1.18
16	QTL_05	AX_124313705	1	45.94	812	0.46
17	QTL_06	AX_124323948	2	47.26	812	-0.77
18	QTL_07	AX_124316288	2	41.90	830	-1.25
19	QTL_08	AX_124328249	5	25.43	807	-0.59
20	QTL_09	AX_124329106	5	38.00	807	-1.67
21	QTL_10	AX_124328447	5	32.78	812	-0.13
22	QTL_11	AX_124324125	2	54.93	801	-1.39
23	QTL_12	AX_124331965	7	45.99	802	-1.90
24	QTL_13	AX_124329646	6	32.39	803	1.42
25	QTL_14	AX_124328171	5	19.47	803	1.42
26	QTL_15	AX_124329055	5	68.11	803	0.76
27	QTL_16	AX_124324024	2	51.69	807	-2.10
28	QTL_17	AX_124323553	1	5.92	807	-0.55
29	QTL_18	AX_124333010	8	53.17	808	-2.61
30	QTL_19	AX_124333118	8	64.03	810	-2.45
31	QTL_20	AX_124315511	5	11.31	812	-1.01
32	QTL_21	AX_124328086	5	10.00	812	0.60
33	QTL_22	AX_124329747	6	42.31	813	-4.34
34	QTL_23	AX_124323777	2	31.40	813	-2.52
35	QTL_24	AX_124313575	4	11.72	815	-2.89

TABLE 3.5 – SNPs détectés comme associés à RY regroupés en QTLs

Chapitre 4

Schémas de pre-breeding et simulateur

4.1 Schéma des sélectionneurs

Les 13 descendances issues du projet AKER comportent de la variabilité génétique qui n'est pas présente dans les variétés élites. Le but du pre-breeding est de croiser ces différentes descendances de manière à créer un réservoir de diversité génétique utile, que les sélectionneurs de la filière betterave sucrière pourront utiliser pour produire de nouvelles variétés à haut potentiel de rendement. Plusieurs études dans la littérature montrent que la sélection génomique peut avoir un avantage vis-à-vis de la sélection phénotypique au niveau de la performance de la population de pre-breeding produite (Desta and Ortiz 2014), (Bassi et al. 2016). Le schéma de pre-breeding proposé par les sélectionneurs intègre donc cette méthode de sélection.

4.1.1 Population en entrée du schéma

La population sur laquelle le schéma de pre-breeding est expérimenté est composée des 13 descendances issues du projet AKER pour lesquelles l'architecture génétique du rendement racinaire (RY) a été étudiée dans le chapitre III, soit 2 524 individus BC2S1 génotypés avec 1 435 marqueurs distincts. Les haplotypes de chacun de ces individus ont été déterminés à l'aide du logiciel Beagle (Browning and Browning 2009) à partir des données de génotype de chacune des populations et de la carte génétique. Les individus issus d'AKER possèdent tous le gène de fertilité spéciale découvert par Owen (Owen and Ryser 1942), qui favorise l'autofécondation. Excepté lors des étapes d'autofécondation, la récolte se fait uniquement sur les individus mâles stériles afin d'être certain de récolter des graines provenant du croisement de deux parents distincts. Un marqueur biallélique de la stérilité mâle nucléaire récessive a été identifié sur le chromosome 1, en position 42.82cM (Owen 1952). Les

individus ayant le génotype (AA) ou (Aa) au marqueur sont hermaphrodites, alors que les (aa) sont mâles stériles. Par la suite, les individus mâles stériles seront appelés individus femelles et les individus hermaphrodites seront désignés comme individus mâles. Du fait de la position de ce gène de stérilité mâle nucléaire, des biais de sélection sur le chromosome 1 peuvent apparaître.

4.1.2 Organisation du schéma

Un premier schéma de pre-breeding, représenté en figure 4.1 a été proposé par les sélectionneurs de l'entreprise Florimond Desprez Veuve & Fils. Dans la phase de compréhension de ce schéma, il est apparu que les étapes « Fixation » correspondaient à des autofécondations, et les étapes « Brassage naturel » correspondaient à une pollinisation aléatoire des femelles par des mâles. Ces deux étapes ont respectivement été renommées *Selfing* et *Random pollination*, lors de la traduction de ce schéma en un schéma comportant des actions de croisements génétiques, de phénotypage et de sélection. De plus, le schéma à traduire débute à partir d'individus BC2S2, alors que nous disposons de l'information génétique uniquement sur les BC2S1. Il est donc nécessaire de rajouter une étape de *Selfing* afin de bien décrire les actions de croisements nécessaires. Pour finir, l'étude par simulations a été limitée à la seule la voie utilisant les individus AKER mâles (Aa). Le schéma traduit pour cette voie est organisé par génération dans la Figure 4.2. Dans ce schéma traduit, une génération correspond à une année. En effet, bien que la betterave sucrière soit une plante bisannuelle, donnant une racine charnue contenant le sucre la première année et montant en graines la seconde année après une période de vernalisation, il est courant dans le cadre de programmes de sélection d'accélérer ce processus. Des planchons de betteraves sont ainsi mis en chambre froide afin de simuler la période de vernalisation, et peuvent donc donner des graines dès la première année.

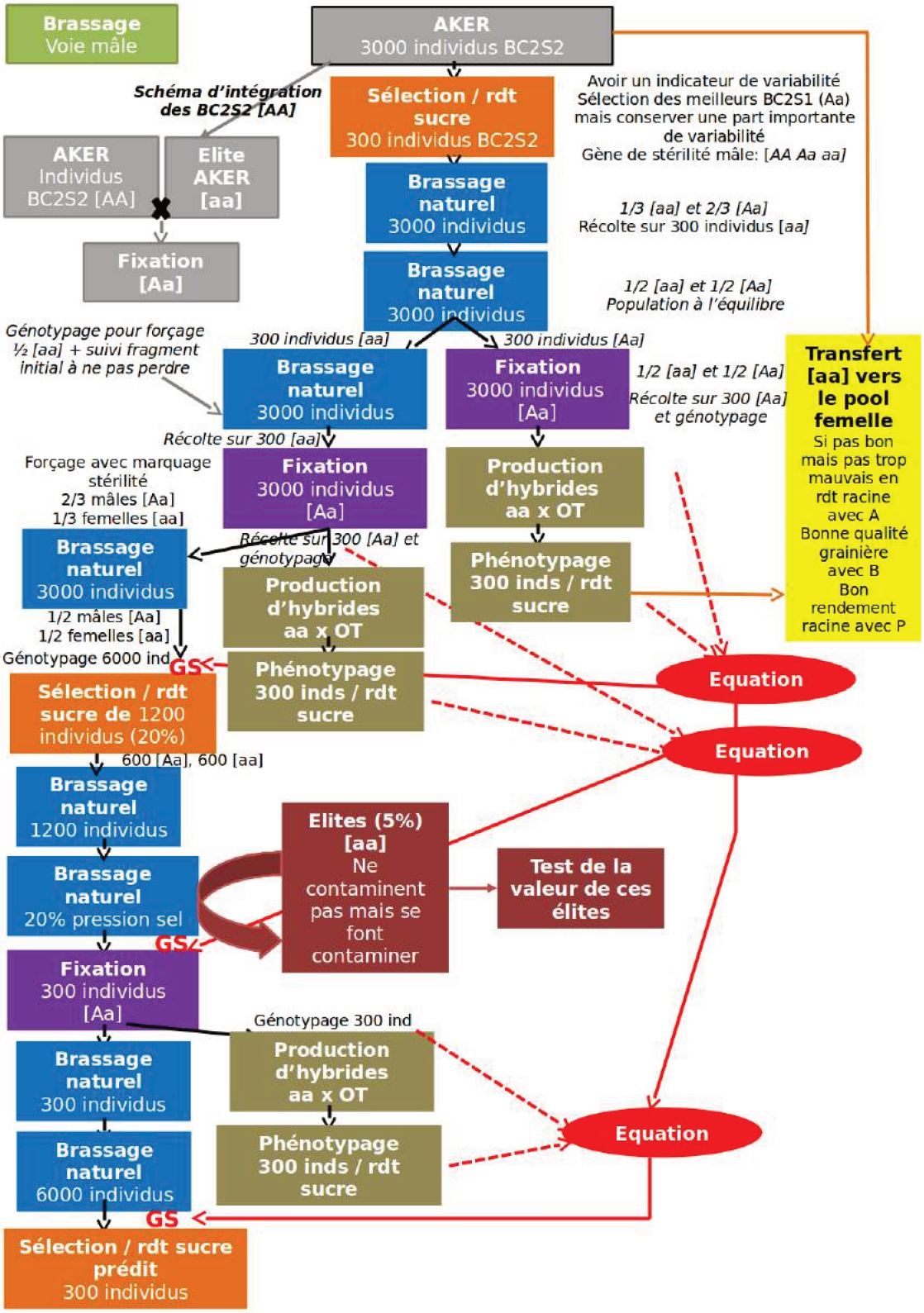


FIGURE 4.1 – Schéma de pre-breeding initial intégrant la sélection génomique, fourni par les sélectionneurs de Florimond Desprez Veuve & Fils

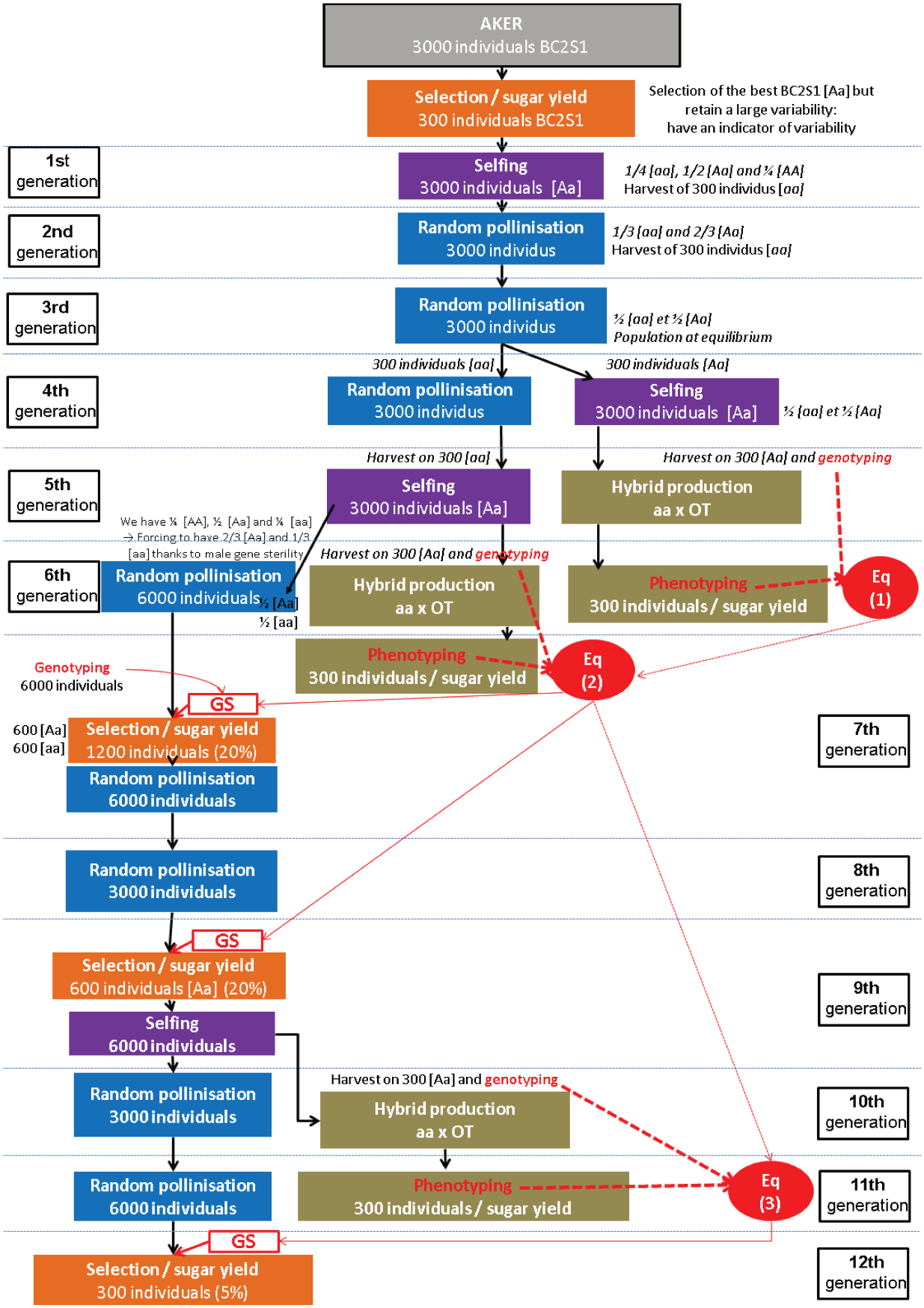


FIGURE 4.2 – Traduction du schéma de pre-breeding initial intégrant la sélection génomique, réorganisé par génération

Le schéma de pre-breeding a pour objectif de créer une population de pre-breeding qui pourra servir de réservoir de diversité génétique utile pour les futurs schémas de sélection. Celui-ci est basé sur la sélection récurrente en utilisant la stérilité mâle nucléaire récessive, comme proposé par Doggett (Doggett and Eberhart 1968) sur le sorgho. Le principe est de créer une population à l'équilibre pour le marqueur de stérilité, c'est-à-dire à moitié mâle et à moitié femelle, en récoltant les graines produites par les individus femelles. Cela favorise l'apparition de combinaisons nouvelles en maximisant le nombre de crossing overs, puisque les graines récoltées sont assurément issues de pollinisations croisées. Les plantes mâles issues de ces graines sont alors autofécondées, ce qui permet d'obtenir des plantes présentant en espérance la moitié de leur génome à l'état homozygote et avec les proportions d'allèles de stérilité suivantes : 1/4 (AA), 1/4 (aa), 1/2 (Aa). Dans un schéma intégrant la sélection génomique tel que proposé par les sélectionneurs, le quart de plante femelle issu d'une autofécondation est génotypé puis croisé avec un testeur élite afin créer une population d'entraînement permettant d'établir une équation de prédiction génomique. Pour retrouver une population à l'équilibre au marqueur de stérilité mâle à partir d'une population issue d'une autofécondation, il suffit de réaliser deux *random pollinations* successives : la première génération sera composée en espérance de de (aa) et de (Aa), et la seconde génération sera bien à l'équilibre. Il est également possible de n'effectuer qu'une seule *random pollination* en génotypant au préalable les plantes au marqueur de stérilité afin de supprimer les individus mâles (AA) avant la *random pollination*. Un schéma de sélection récurrente utilisant la stérilité mâle récessive est donc une succession de cycles composés *a minima* de deux *random pollination* permettant de produire une population à l'équilibre, suivies d'une autofécondation qui permet d'obtenir des individus dont la moitié du génome est à l'état homozygote, puis de la sélection des meilleurs individus qui constituent ainsi la population en entrée du cycle suivant. Le schéma de pre-breeding proposé par les sélectionneurs en figure 4.2 peut ainsi être divisé en trois objectifs : l'obtention de la population à l'équilibre mâles / femelles, la génération des populations d'entraînement nécessaires à l'établissement du modèle de prédiction génomique, et l'utilisation de la sélection génomique dans ce schéma de sélection récurrente. Ces trois objectifs et les étapes du schéma du pre-breeding qui y sont associées sont explicités ci-dessous.

Le premier objectif est d'obtenir une population à l'équilibre où les meilleurs individus auront été brassés. Cette initialisation est réalisée de la génération 0 à la génération 3 inclue. En génération 0 à partir des BC2S1, c'est-à-dire les 13 descendances d'accessions exotiques pour lesquelles on dispose à la fois du génotype et du phénotype et sur lesquelles on a étudié l'architecture génétique du rendement racinaire (RY), une première action de sélection sur le rendement RY ainsi que sur le marqueur de stérilité est réalisée, afin

de retenir les 300 meilleurs individus mâles hétérozygotes pour le marqueur de stérilité (Aa). Ces individus seront appelés par la suite les « fondateurs ». Ce chiffre de 300 a été choisi comme une limite supérieure à partir de laquelle la gestion des individus devient trop difficile pour les sélectionneurs. Les 300 fondateurs sont auto-fécondés en G1, et un sachet de 1 à 70g de graines est récolté sur chaque fondateur. 20 graines provenant de chaque fondateur sont semées, mais seuls 10 des planchons qui en résultent sont conservés en veillant à garder au moins une femelle par fondateur (identifiée par génotypage au marqueur de stérilité mâle). Ces 3000 plantes, 10 pour chacun des 300 fondateurs, constituent la génération G1 avec en espérance, les proportions d'allèles de stérilité suivantes : 1/4 (AA), 1/4 (aa), 1/2 (Aa). Pour faciliter la compréhension de ce schéma, seul le nombre de graines récoltées pour constituer la génération suivante sera indiqué par la suite. En G2 une *random pollination* est mise en place : les femelles vont être pollinisées aléatoirement par les mâles afin de brasser les allèles des individus. 10 graines sont alors récoltées sur 300 femelles choisies aléatoirement. On obtient ainsi une population composée de 3000 individus, avec en espérance les proportions d'allèles au marqueur de stérilité suivantes : 1/3 (aa), 2/3 (Aa). En G3 une nouvelle *random pollinisation* est à nouveau effectuée : les femelles de ces 3000 individus sont pollinisées aléatoirement par les mâles. A nouveau 10 graines sont récoltées sur 300 femelles choisies aléatoirement, ce qui forme une nouvelle population composée de 3000 individus, avec en espérance, les proportions d'allèles au marqueur de stérilité suivantes : 1/2 (aa) et 1/2 (Aa), la population est à l'équilibre.

Un deuxième objectif est d'intégrer dans le schéma plusieurs étapes de sélection génomique, c'est à dire de sélection des meilleurs individus selon leur valeur RY prédite grâce aux informations génomiques. Pour prédire la valeur RY des individus dans une population à partir de leur génotype, il faut au préalable connaître le modèle établissant la relation entre les informations génomiques des individus et leur phénotype. Ce modèle, ou équation de sélection génomique, peut-être appris à partir d'une population d'entraînement, pour laquelle le génotype et le phénotype des individus sont connus. Le deuxième objectif de ce schéma de sélection est donc de produire ces populations d'entraînement, nécessaires pour l'utilisation de la sélection génomique. Pour ce faire, 300 individus parmi les mâles obtenus en G3 (1500 Aa en espérance) sont génotypés afin d'obtenir leurs informations génomiques et autofécondés afin de favoriser l'homozygotie au sein de leurs descendants, comportant 1/2 de (aa) et 1/2 de (Aa). Une femelle (aa) résultant de chacune des autofécondations est ensuite pollinisée par un testeur mâle élite afin de produire des hybrides en G5, phénotypés en G6. Une équation peut alors être établie à partir des informations génomiques de la population en G3 et des phénotypes des hybrides, permettant de faire de la prédiction génomique et donc de la sélection génomique

dès la génération G7. Ces trois mêmes étapes réalisées en G5, G6, et G7 permettent de créer une seconde population d'entraînement. Les deux populations d'entraînement sont utilisées afin de produire une seconde équation de sélection génomique, utilisable dès la génération G8. Enfin, ces trois étapes réalisées en G9, G10 et G11 aboutissent à la création d'une troisième population d'entraînement. Les trois populations servent à établir une troisième équation de sélection génomique, qui peut être utilisée dès la génération G12.
Le dernier objectif consiste à utiliser la sélection génomique dans un contexte de sélection récurrente avec stérilité mâle récessive. Cette phase débute en génération G4 et s'achève à la fin du schéma, soit en génération G12. Comme expliqué précédemment, la génération G3 est à l'équilibre pour le marqueur de stérilité. Cependant le premier modèle de sélection génomique n'est utilisable qu'à partir de la génération G7 : il n'est donc pas possible d'effectuer de sélection génomique dès la génération 4. Une nouvelle *random pollination* est effectuée afin de retarder l'utilisation de la sélection génomique, ce qui permet d'augmenter encore les recombinaisons. 10 graines sont récoltées sur 300 femelles choisies aléatoirement, la nouvelle population composée de 3000 individus est toujours en espérance à l'équilibre au niveau du marqueur de stérilité. En G5, les individus (Aa) de cette population sont auto-fécondés. 10 graines sont récoltées sur 300 d'entre eux, choisis aléatoirement. Grâce au marqueur de stérilité, ces 3000 individus sont choisis pour avoir des proportions au marqueur de stérilité suivantes : 1/3 (aa) et 2/3 (Aa). De cette façon, lorsqu'une nouvelle *random pollination* est réalisée en G6, et que 20 graines sont récoltées sur 300 femelles aléatoires, la population de 6000 individus obtenue est en espérance à l'équilibre au niveau du marqueur de stérilité. En G7, la première équation de sélection génomique est générée. Il est donc possible d'effectuer une pression de sélection définie par les sélectionneurs à 20%. Parmi les 6000 individus, 1200 sont ainsi sélectionnés par sélection génomique en respectant la contrainte d'équilibre soit 600 mâles et 600 femelles. Une *random pollination* est alors effectuée. 10 graines sont récoltées sur les 600 femelles, donnant en espérance une population en équilibre de 6000 individus. En G8, une nouvelle *random pollination* est effectuée sur ces 6000 individus, et 10 graines sont récoltées sur 300 femelles choisies aléatoirement. La population obtenue est composée de 3000 individus, en espérance à l'équilibre vis à vis du marqueur de stérilité mâle. En G9, la seconde équation de sélection génomique est disponible. Une pression de sélection de 20% est appliquée sur les individus (Aa), dans le but de choisir les individus à autoféconder. 600 individus (Aa) sont ainsi sélectionnés parmi les mâles (1500 en espérance) et autofécondés. 10 graines sont récupérées sur les 600 individus autofécondés, avec en espérance les proportions au marqueur de stérilité suivante : 1/4 (aa), 1/2 (Aa) et 1/4 (AA). Afin de revenir à l'équilibre, une *random pollination* est effectuée en G10 puis en G11, et 10 graines sont récoltées à chaque fois sur 300 femelles choisies aléatoirement. En G11, on obtient ainsi une

population de 3000 individus, en espérance à l'équilibre vis-à-vis du marqueur de stérilité. En G12, la dernière équation de GS est connue. Cette équation est utilisée en sélection génomique avec une pression de sélection fixée à 5% pour sélectionner les 150 meilleurs mâles et les 150 meilleures femelles qui composent ainsi la population de pre-breeding finale.

4.2 Traduction du schéma des sélectionneurs en langage informatique

La traduction du schéma de pre-breeding imaginé par les sélectionneurs de l'entreprise Florimond Desprez Veuve & Fils en langage informatique a nécessité plusieurs étapes. Le Cycle en V (Figure 4.3), une méthode d'organisation très connue, tout d'abord appliquée dans l'industrie puis adaptée à l'informatique dans les années 80, a été utilisée de façon flexible. Ce cycle décrit précisément les différentes étapes successives nécessaires au développement d'un produit, ici le simulateur de schémas de pre-breeding. Par définition, les étapes successives sont habituellement effectuées dans un ordre établi. Cependant, certaines idées ayant émergé au cours de l'étude et certaines précisions ayant été explicitées tardivement, des « retours en arrière » dans ce cycle ont été nécessaires. C'est donc avant tout pour clarifier la présentation du travail de développement du simulateur que les étapes sont décrites ci-après selon l'organisation du cycle en V.

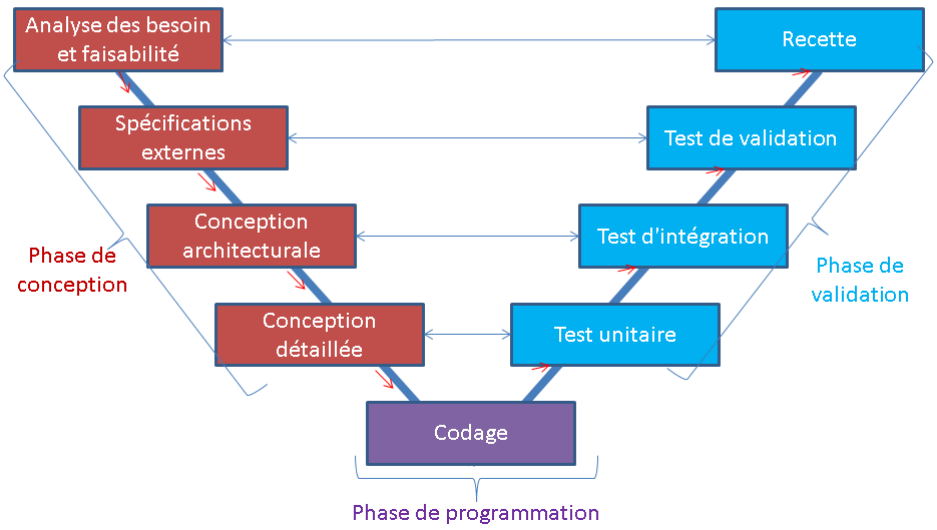


FIGURE 4.3 – Cycle de développement en V

Trois grandes phases peuvent être distinguées au sein de ce cycle : la phase de conception, de programmation et de validation.

4.2.1 Phase de conception

La phase de conception est elle-même divisée en 4 étapes : l'analyse des besoins de l'utilisateur, ici les sélectionneurs, la spécification externe, la spécification architecturale, et la conception détaillée. Chacune de ces étapes est définie et reliée avec notre étude ci-dessous.

4.2.1.1 L'analyse des besoins et les spécifications externes

La phase d'analyse des besoins a pour objectif d'identifier les différentes questions que se posent les sélectionneurs auxquelles le simulateur doit pouvoir apporter des réponses. Elle est suivie de la phase de spécification externe, qui a pour but de définir les fonctionnalités du simulateur permettant de répondre à ces questions. C'est à cette étape que les différents scénarios à simuler vont être déterminés.

Les premières questions portent sur l'effet des différentes méthodes de sélection sur le gain génétique. En effet les schémas de pre-breeding classiques font intervenir la sélection phénotypique, mais la sélection génomique peut être un avantage dans les schémas de sélection pour des caractères complexes (Combs and Bernardo 2013). Le simulateur doit donc permettre de comparer deux scénarios, **un scénario basé sur la sélection phénotypique et un scénario basé sur la sélection génomique** afin de déterminer si l'introduction de la sélection génomique dans le schéma de pre-breeding est une bonne stratégie en ce qui concerne le gain génétique. Par ailleurs, la vraie valeur génétique de chaque individu étant connue au sein du simulateur, il est aussi possible de comparer la performance des populations obtenues par sélection génomique et celle des populations obtenues par sélection sur la vraie valeur génétique. Cette comparaison, possible uniquement in silico, permet de quantifier l'efficacité de la sélection génomique par rapport à la sélection optimale sur la valeur génétique. Un troisième **scénario intégrant la sélection sur la vraie valeur génétique** est donc établi.

Les questions suivantes portent sur l'impact de l'effectif des populations sur l'évolution de la diversité génétique utile. Le schéma de sélection classique imaginé par les sélectionneurs vise tout d'abord à sélectionner les 300 individus hétérozygotes au marqueur de stérilité et ayant le meilleur rendement racinaire parmi les 13 descendances AKER étudiées, en prenant au moins un individu de chacune des 13 descendances. Ces 300 individus sont appelés les « fondateurs » , ce sont les individus en entrée du schéma de sélection. Une première question apparaît : est-il utile de sélectionner un grand nombre de fondateurs, ou

au contraire est-ce contre-productif, « polluant » le schéma avec des morceaux de génome ne possédant pas de variabilité génétique utile ? Restreindre le nombre de fondateurs, ce qui permet à coup sûr de minimiser les coûts du schéma de sélection, doit donc être mis en rapport à la perte de diversité utile que cette diminution des fondateurs va entraîner dans la population de pre-breeding produite en sortie du schéma. Ce schéma de sélection a aussi pour finalité de produire une population de pre-breeding comportant 300 individus, une question similaire se pose donc : diminuer le nombre d'individus retenus à la fin du schéma a-t-il un impact sur la diversité génétique utile au sein de la population de pre-breeding finale ? Pour répondre à ces questions, trois scénarios sont imaginés : un scénario débutant avec **300 fondateurs et produisant une population de pre-breeding avec 300 individus** ; un scénario partant de **30 fondateurs et générant une population de pre-breeding composée de 30 individus** ; et un scénario avec **30 fondateurs aboutissant à une population de pre-breeding de 300 individus**.

Enfin une dernière question concernant l'évolution de la diversité génétique intéresse les sélectionneurs. Lors de croisements entre une plante femelle (mâle stérile) et une plante mâle fertile, les graines composant la génération suivante vont toujours être récoltées sur le parent femelle. Il est donc possible de suivre le lignage maternel des individus. Au sein de chaque génération, les graines à récolter peuvent donc être choisie de façon à toujours avoir au moins un descendant de chacun des fondateurs. Deux scénarios, l'un permettant de **suivre le lignage maternel** et l'autre simulant une **récolte indépendante du lignage maternel** peuvent donc être simulés et comparés pour étudier leur effet sur l'évolution de la diversité génétique utile.

Cependant une contrainte est imposée par les sélectionneurs. Dans un schéma utilisant la sélection phénotypique, l'effectif maximum des individus à sélectionner doit être limité à 1200 individus. L'intensité de sélection appliquée étant de 20% lors des deux premières actions de sélection phénotypiques et de 5% lors de la sélection de la population de pre-breeding, seuls 240 individus et 60 individus seront respectivement sélectionnés. Dans un scénario $300 \rightarrow 300$ il n'est donc pas possible de suivre chacun des 300 fondateurs. Par conséquent, les scénarios intégrant la sélection phénotypique ne peuvent pas intégrer le suivi du lignage maternel des fondateurs.

La phase d'analyse des besoins fait ainsi apparaître que 12 scénarios doivent être simulés : 6 scénarios intégrant la sélection génomique (3 scénarios comparant l'impact des effectifs des populations $\times$ 2 scénarios comparant l'impact du suivi des fondateurs par voie maternelle), 3 scénarios utilisant la sélection phénotypique et 3 scénarios basés sur la sélection de la vraie valeur génétique (scénarios comparant l'impact des effectifs des populations). Ces différents scénarios sont résumés dans le Tableau 4.1.

	GS		PS	GV
	suivi du	sans suivi	sans suivi	sans suivi
Effectifs	lignage	du lignage	du lignage	du lignage
30→30	scenario 1	scenario 4	scenario 7	scenario 10
30→300	scenario 2	scenario 5	scenario 8	scenario 11
300→300	scenario 3	scenario 6	scenario 9	scenario 12

TABLE 4.1 – Récapitulatif des 12 scénarios du schéma de pre-breeding à simuler.

4.2.1.2 La conception architecturale

La phase de conception architecturale est l'étape dans laquelle tous les composants nécessaires à la mise en place des différents éléments du simulateur sont identifiés.
Le schéma de pre-breeding commence à partir de la génération de départ, composée des individus provenant des 13 descendances d'AKER. Un élément *generation* est donc requis. Un enchaînement d'actions permet ensuite de composer les différents scénarios du schéma de pre-breeding. Un élément *generation* est créé à la fin de chaque action. Un nouvel élément regroupant toutes les générations créées, appelé *list generations* semble alors nécessaire pour pouvoir par la suite accéder à chaque génération. Concernant les actions, on peut distinguer deux catégories : les actions liées aux croisements à simuler et les actions liées aux différentes sélections à simuler.

Les actions découlant des croisements à simuler

Simuler un schéma de pre-breeding implique de simuler des croisements entre individus. Trois croisements différents sont envisagés dans les schémas de sélection : l'autofécondation, la pollinisation aléatoire, où les femelles sont pollinisées par des mâles aléatoirement tirés dans un ensemble de plusieurs mâles, et la production d'hybrides où les femelles sont pollinisées par un pollinisateur élite connu. Une action pour chacun de ces croisements doit donc être prévue dans le simulateur : les actions *Selfing*, *Random Pollination* et *Hybrid Production*.

Les actions induites par les différentes sélections à simuler

Les scénarios définis lors de la phase de spécification externe font état de trois sortes de sélection : la sélection phénotypique, la sélection génomique, et la sélection sur la vraie valeur génétique. Une action pour chacun des croisements est définie pour le simulateur : les actions *PS*, *GS* et *GV*.

Autres éléments pris en compte dans les actions

Excepté à la fin du schéma de pre-breeding, une action de sélection est systématiquement suivie d'une action de croisement. Ces actions nécessitent d'avoir en entrée une génération composée d'individus ayant des génotypes adaptés au marqueur de stérilité. Par exemple, effectuer une action *Selfing* ou *Random pollination* est impossible si la population en entrée n'est composée que de femelles. Le génotype au marqueur de stérilité doit donc pouvoir être connu pour chaque individu, et les actions de sélection doivent pouvoir intégrer le fait de sélectionner des individus ayant le génotype au marqueur de stérilité désiré. L'élément *sterility marker*, vecteur indiquant le génotype au marqueur de stérilité de chaque individu, est donc nécessaire au bon fonctionnement du simulateur.

Les scénarios définis précédemment incluent la possibilité de suivre ou non le lignage maternel pour sélectionner au moins un descendant de chaque fondateur à la génération suivante. Il est donc nécessaire de créer l'élément *lineage pedigree* contenant l'information sur le lignage maternel de chacun des individus de chaque génération. En fonction du scénario simulé, les actions de sélection prendront ou non cet élément en compte.

La sélection des fondateurs est un cas particulier de l'action de sélection phénotypique. En effet, les fondateurs sont les individus en entrée du schéma de pre-breeding, hétérozygotes au marqueur de stérilité et sélectionnés comme ayant les meilleurs rendements racinaires parmi les individus des 13 descendances AKER étudiées, en prenant au moins un individu de chacune de ces 13 descendances. Le choix des fondateurs dépend donc à la fois du phénotype, du marqueur de stérilité, mais aussi de la structure en 13 descendances. Une action *founder selection* doit donc être ajoutée.

La conception architecturale a donc permis d'identifier plusieurs éléments divisés en deux catégories : les objets et les actions. Les objets comportent 4 éléments : *generation*, *list generation*, *sterility marker*, et *lineage pedigree*. Les actions sont composées de six éléments, trois actions de croisements : *Selfing*, *Random Pollination*, et *Hybrid Production*, ainsi que quatre actions de sélection : *PS*, *GS*, *GV* et *founder selection*.

4.2.1.3 La conception détaillée

La phase de conception détaillée a pour objectif de décrire la conception de chaque élément identifié lors de l'étape de conception architecturale. Cette conception va s'appuyer sur le package AlphaSimR, package R disponible sur le CRAN permettant de simuler des programmes de sélection (`https://CRAN.R-project.org/package=AlphaSimR`).

Objets

L'élément *generation* est un objet regroupant toutes les informations nécessaires pour décrire une population à une génération donnée. C'est un objet de type Pop-class du package AlphaSimR. A chaque individu est associé un nom, le nom de son parent femelle, de son parent mâle, la valeur de son phénotype pour le caractère étudié, ses haplotypes pour l'ensemble des marqueurs génomiques, son genre (mâle ou femelle) et sa valeur génétique. L'élément *generation* doit tout d'abord représenter la population en entrée du simulateur, c'est à dire l'ensemble des individus des 13 descendances AKER étudiées. Il est possible de récupérer les haplotypes de ces individus lors de l'imputation des données génomiques manquantes avec le logiciel Beagle, décrite dans le chapitre précédent. Tous les attributs nécessaires à la création de cette population dans le simulateur sont donc connus, excepté le genre et la valeur génétique de chaque individu.

La position du marqueur de stérilité est connue : il s'agit du marqueur situé sur le chromosome 1 en position 42.82cM. Grâce à cette information, il est possible de créer une fonction permettant d'extraire le génotype au marqueur de stérilité pour chaque individu, et ainsi de créer l'élément *sterility marker*. Le genre de l'individu peut alors être déduit de l'élément *sterility marker* : les individus (aa), codés « 0 » dans l'élément *sterility marker*, sont considérés comme des femelles (car mâles stériles) et les individus (Aa), codés « 1 », et (AA), codés « 2 » , sont considérés comme des mâles. En effet, même s'ils sont mâles et femelles fertiles, ils ne sont utilisés dans les croisements impliquant deux parents qu'en tant que mâles.

Le seul attribut manquant pour créer la population de départ est à présent la valeur génétique de chaque individu, elle est calculée selon le modèle choisi pour l'architecture génétique.

Dans les simulations, la valeur génétique d'un individu correspond à sa valeur génétique additive, calculée selon la formule suivante :

$$g_{di} = \sum_{l} (m_{dil} - \mu_l) a_{dl}$$

avec g_{di} la valeur génétique g de l'individu i dans la descendance d, m_{dil} la dose allélique exotique de l'individu i dans la descendance d au QTL l, μ_l la moyenne des doses alléliques exotiques au QTL l et a_{dl} l'effet additif au QTL l dans la descendance d.

Lors de l'étude de l'architecture génétique du rendement racinaire, 24 QTLs ont été détectés. Certains QTLs existants n'ont probablement pas été détectés dû à un manque de puissance. 26 QTLs sont donc échantillonnés aléatoirement afin d'obtenir une architecture génétique du RY composée de 50 QTLs. Ces 26 QTLs supplémentaires sont distants d'au moins 5cM des QTLs détectés, et considérés comme étant présents dans toutes les

descendances. Les effets de ces 26 QTLs sont tirés au sort dans une loi normale :

$$a_s \sim \mathcal{N}(0, \sigma_a^2)$$

avec a_s l'effet additif exotique au QTL simulé s. Les effets des QTLs simulés sont définis pour être moins important que ceux des QTLs détectés. L'héritabilité du rendement racinaire la plus faible calculée sur l'ensemble des 13 descendances AKER étant de 0.14, une variabilité plus faible est choisie pour échantillonner les effets des QTLs aléatoires. La figure 4.4 présente les effets des QTLs détectés et échantillonnés. Pour vérifier la validité de notre architecture génétique, la corrélation entre les valeurs de RY observées et les valeurs génétiques estimées est calculée. Elle équivaut à 0,61, ce qui est cohérent avec la valeur de l'héritabilité du caractère RY obtenue en espérance dans l'ensemble des 13 descendances AKER donnée dans le chapitre 3 (0,60).

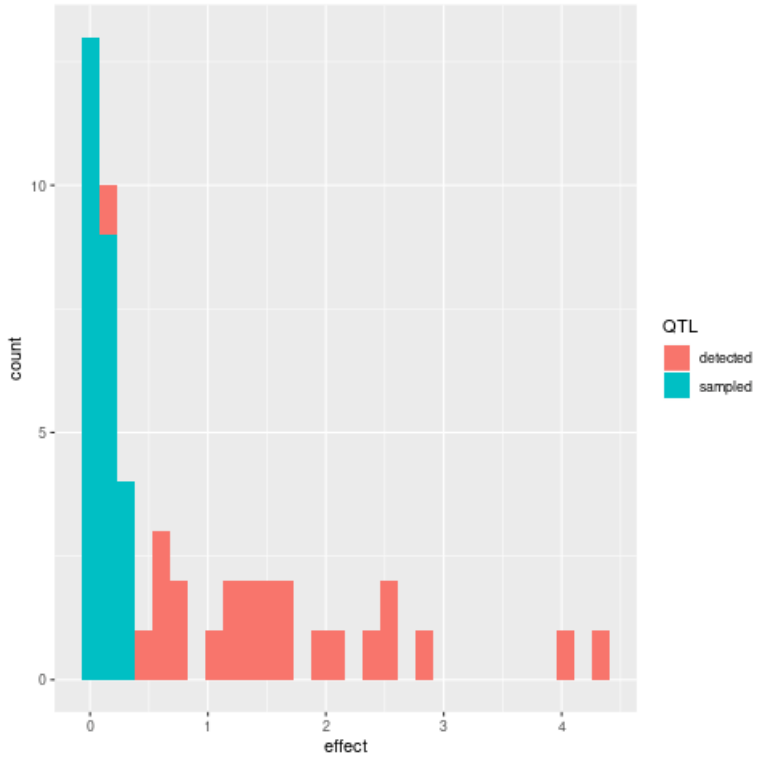


FIGURE 4.4 – Valeur absolue des effets des QTLs détectés en rose et des QTLs échantillonnés en bleu.

Les QTLs n'ont pas les mêmes effets au sein des différentes descendances. En effet, tous les QTLs n'ont pas été détectés par l'étude d'association dans toutes les descendances, et un même QTL ayant été détecté dans deux descendances peut avoir un effet favorable de l'allèle exotique dans l'une et défavorable dans l'autre. Il est donc nécessaire de pouvoir suivre un QTL transmis par une descendance donnée au sein des générations suivantes, afin de déterminer si ce QTL a un effet, et si oui si l'effet de l'allèle exotique à ce QTL est favorable ou défavorable. Pour chaque QTL à suivre, deux pseudo-marqueurs sont créés et

colocalisés à la même position génétique que le QTL. L'un représente l'effet favorable de l'allèle exotique au QTL, l'autre l'effet défavorable. Ainsi, si le QTL a un effet favorable de l'allèle exotique dans la descendance étudiée, le pseudo-marqueur favorable qui y est associé est identique au QTL tandis que le pseudo-marqueur défavorable correspondant n'est composé que du génotype élite homozygote (0). A l'inverse, si l'effet de l'allèle exotique au QTL dans la descendance est défavorable au caractère RY, le pseudo-marqueur défavorable associé au QTL a les mêmes informations génotypiques que le QTL tandis que le pseudo-marqueur favorable n'est constitué que du génotype élite homozygote (0). Dans le cas où le QTL n'est pas détecté dans une descendance, les deux pseudo-marqueurs comportent uniquement le génotype exotique (0). A partir des informations issues des deux pseudo-marqueurs, il est possible de déterminer si l'effet de l'allèle exotique au QTL est favorable, défavorable ou nul (Figure 4.5).

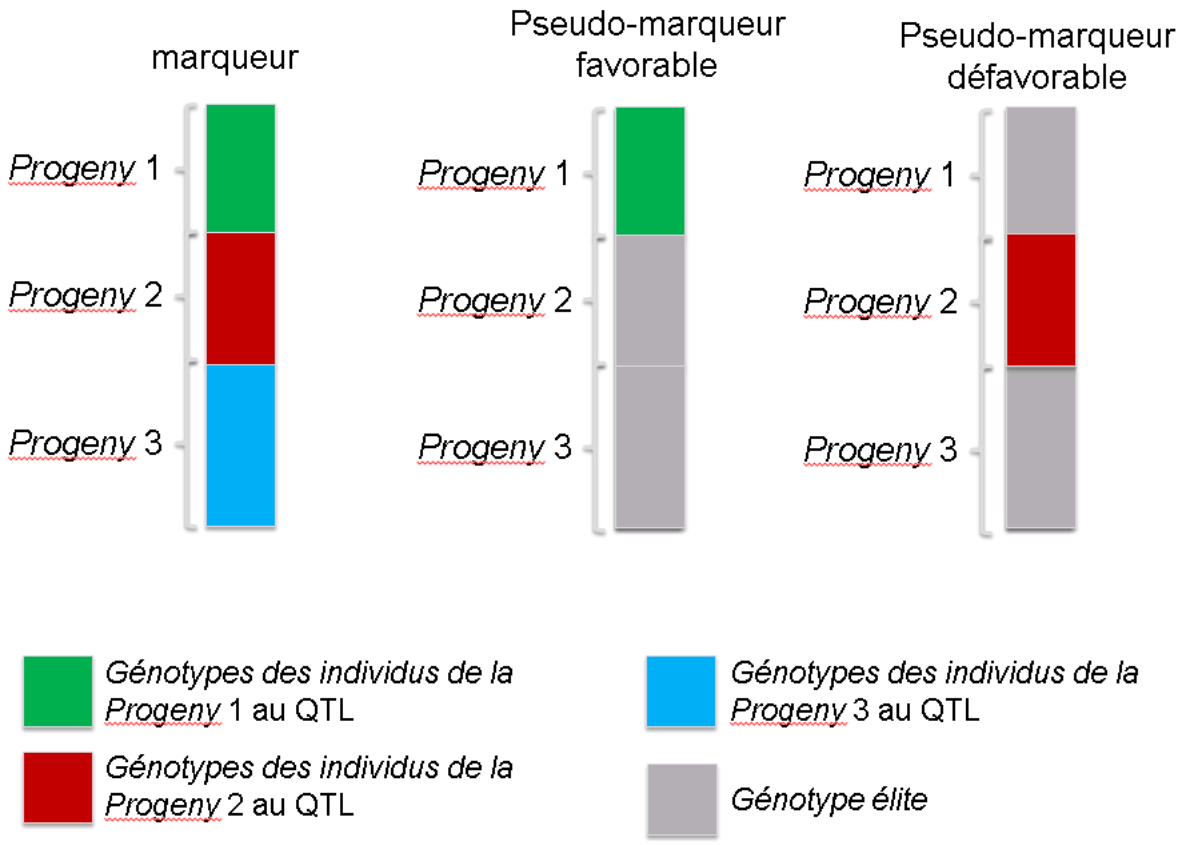


FIGURE 4.5 – Pseudo-marqueurs créés à partir de l'information de l'effet de l'allèle exotique à un marqueur pour trois descendances. L'effet de l'allèle exotique au marqueur est favorable dans la descendance 1, défavorable dans la descendance 2, et nul dans la descendance 3.

Pour chacun des QTLs échantillonés, l'effet favorable ou défavorable de l'allèle exotique est tiré au sort avec une probabilité de 1/5 de présenter un effet favorable.
Tous les attributs permettant de créer une classe *Pop-class* sont maintenant connus, le premier objet *generation* peut être généré.

L'élément *list generation* est une liste regroupant tous les objets *generation* créés pendant la simulation du schéma de pre-breeding. Il s'agit de l'élément qui sera enregistré au format RData au terme de la simulation. La population de départ est nommée « BC2S1 » et la population de fondateurs est nommée « founders » . Chacune des autres générations est nommée selon la nomenclature suivante : identifiant de la génération (« G1 » pour Génération 1 par exemple), suivi par un underscore, puis par l'identifiant de l'action ayant généré la population. Les identifiants des actions sont les suivants :

— S pour *Selfing*

— D pour *Random Pollination*, car on fait ici référence au schéma de Doggett présenté dans la partie 1.2.1.

— P pour *Hybrid Production*, car le phénotype des hybrides générés est simulé

— GS pour sélection génomique

— PS pour sélection phénotypique

— GV pour sélection génomique

Le code R de chacune de ces actions est donné en annexes (Figure A.8).
Le dernier élément de la catégorie objet à conceptualiser est *lineage pedigree*. Il s'agit d'un tableau comportant une ligne par individu et quatre colonnes : *population*, *individual*, *founder* et *sterility_marker*. La colonne *population* indique le nom de la génération d'appartenance de l'individu nommé dans la colonne *individual*. La colonne *founder* indique le nom du fondateur dont descend l'individu selon son lignage maternel, et la colonne *sterility_marker* renseigne sur le genre de l'individu.
On s'intéresse à présent à la conception des différentes actions à réaliser au sein du simulateur.

Actions

Les actions peuvent être distinguées en deux catégories : les actions de croisement et les actions de sélection. Le package AlphaSimR propose plusieurs fonctions adaptées pour réaliser ces actions. La fonction *self* permet ainsi de réaliser des autofécondations, *selectOP* modélise une sélection dans une population de plantes à pollinisation ouverte et peut ainsi simuler une pollinisation aléatoire, *hybridCross* génère des hybrides, et la fonction *selectInd* permet de sélectionner des individus en fonction de leur valeur phénotypique, de leur valeur génomique estimée, ou de leur vrai valeur génétique. Cependant toutes ces actions doivent pouvoir être influencées par les différents scénarios. S'il est possible de changer l'effectif des populations générées par les fonctions du package AlphaSimR et donc de jouer sur le nombre de fondateurs et sur le nombre d'individus dans la population

finale, il n'est pas possible de suivre le lignage maternel des individus. De plus, le suivi du marqueur de stérilité mâle n'est pas prévu dans le package AlphaSimR : l'attribution du genre de chaque individu se fait de façon aléatoire. A chaque génération produite, il faut donc vérifier le génotype au marqueur de stérilité de chaque individu afin de lui attribuer le genre adéquat. Enfin, tous les objets *generation* créés par ces fonctions sont par la suite ajoutée à l'objet *list generations*, il doivent donc être formatés selon les règles édictées précédemment. Afin d'ajouter toutes ces possibilités, les fonctions issues du package AlphaSimR sont « enrobées » avec d'autres fonctions pour former une seule fonction par action, offrant toutes les fonctionnalités requises à la simulation des différents scénarios. Ces fonctions d'enrobage sont au nombre de quatre et sont décrites ci-dessous.

Fonctions d'enrobage

select_on_lineage

Dans le cas où le lignage maternel des individus doit être suivi, la fonction *select_on_lineage* est appelée. Cette fonction permet de générer le vecteur des noms des individus d'une génération qui ont le génotype au marqueur de stérilité spécifié par l'utilisateur, et parmi lesquels chaque fondateur a au moins un descendant par lignage maternel. Cette fonction permet ainsi de vérifier qu'il y a bien au minimum un individu descendant de chacun des fondateurs initiaux parmi les individus à récolter (dans le cadre d'une action de croisement) ou à sélectionner (dans le cadre d'une action de sélection). Dans le cas d'une autofécondation, la récolte se fait sur les individus mâles fertiles : la fonction renvoie une erreur si aucun mâle fertile n'est trouvé pour au moins un des fondateurs. Dans le cas d'une pollinisation aléatoire ou d'une production d'hybride, les récoltes se font sur les femelles : la fonction *select_on_lineage* renvoie une erreur si aucune femelle n'est trouvée pour au moins un des fondateur initiaux. Dans le cas de sélection sur un génotype au marqueur de stérilité mâle particulier, la fonction renvoie une erreur s'il n'existe pas au moins un individu descendant de chacun des fondateurs avec le génotype au marqueur de stérilité mâle désiré.

Population_format

La fonction *Population_format* permet de renommer les individus d'une génération nouvellement produite selon la nomenclature : identifiant de la génération (« G1 » pour Génération 1 par exemple), suivi par un underscore, puis par l'identifiant de l'action ayant généré la population. Elle permet également de définir le genre mâle ou femelle de chaque individu de cette génération à partir de leurs informations haplotypiques au marqueur de stérilité mâle.

recode_sterility_allele

La fonction *recode_sterility_allele* permet de créer l'élément *sterility marker*, le vecteur indiquant le génotype au marqueur de stérilité mâle de chaque individu : 0 si l'individu est mâle stérile, c'est-à-dire homozygote pour l'allèle de stérilité mâle (aa), 1 si l'individu est mâle fertile hétérozygote au marqueur de stérilité (Aa), et 2 si l'individu est mâle fertile homozygote pour l'allèle de fertilité mâle (AA).

lineage_recovery_from_mother

La fonction *lineage_recovery_from_mother* est appelée lorsque le lignage maternel des individus doit être suivi. Cette fonction produit un tableau rassemblant les informations sur le lignage maternel et la stérilité mâle des individus de la génération produite par l'action. Ce tableau comporte autant de lignes que d'individus dans la génération, et quatre colonnes : le nom de la génération, le nom de chaque individu, le nom du fondateur duquel descend chaque individu en fonction de son lignage maternel, et son génotype au marqueur de stérilité mâle. Ce tableau est par la suite ajouté à l'élément *lineage pedigree*, qui rassemble ces mêmes informations pour toutes les générations produites au cours de la simulation du schéma de pre-breeding.

Les paramètres d'entrée de chacune de ces fonctions sont présentées en annexes (Figure A.9).

Fonctions de croisement

Selfing

La figure 4.6 présente l'organisation des différentes fonctions associées pour composer la fonction *Selfing*. La modélisation de l'action d'autofécondation repose sur la fonction *self* du package AlphaSimR. Cette fonction nécessite de renseigner le paramètre *nProgeny* qui indique le nombre de descendants issus de chaque individu de chaque plante autofécondée. Cette fonction ainsi que les différentes fonctions d'enrobage décrites précédemment composent la fonction *Selfing*.

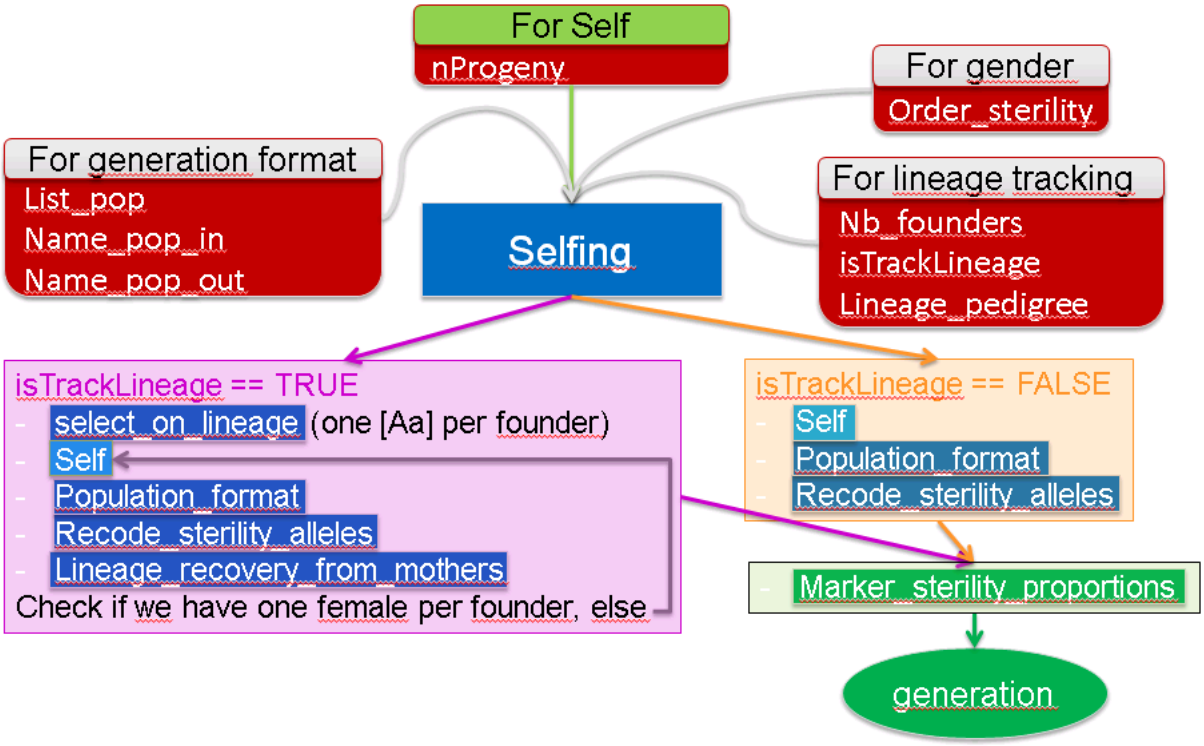


FIGURE 4.6 – Schéma de conception de la fonction *Selfing*. Les paramètres d'entrée sont représentés par un fond rouge. La fonction *Self* issue du package AlphaSimR est représentée par un fond bleu clair. Toutes les fonctions d'enrobage sont représentées par un fond bleu foncé. La fonction de validation est représentée par un fond vert. L'objet *generation* en sortie de la fonction *Selfing* est représenté par un ovale vert.

Dans le cas où le lignage maternel est suivi, l'action *Selfing* fait tout d'abord appel à la fonction *select_on_lineage* pour vérifier qu'il existe bien un individu (Aa) descendant de chaque fondateur. La fonction *self* du package AlphaSimR est alors appelée, les individus mâles fertiles de la population en entrée sont autofécondés et génèrent autant de descendants que précisé par le paramètre *nProgeny*. L'objet *generation* obtenu est ensuite formaté grâce à la fonction *Population_format*, et la fonction *Recode_sterility_alleles* permet de déterminer le genre de chacun des individus générés. Enfin, la fonction *Lineage_recovery_from_mothers* est appelée pour produire le Tableau d'information sur la nouvelle génération produite : nom de la population, nom de chaque individu, nom du fondateur dont provient chaque individu selon son lignage maternel, et génotype au marqueur de stérilité. Dans tous les scénarios des schémas de pre-breeding à simuler, une action d'autofécondation est suivie d'une action de pollinisation aléatoire ou de production d'hybrides. Dans ces deux cas, la génération suivante sera produite par la récolte sur des femelles. Dans le cas où le lignage maternel ne doit pas être suivi, l'autofécondation est réalisée par l'appel à la fonction *self* du package AlphaSimR sans vérification préalable de l'origine des femelles. Cependant, lorsque le lignage maternel est suivi, il est nécessaire de s'assurer que l'autofécondation a permis de générer au moins une femelle descendant

de chacun des fondateurs initiaux. Si ce n'est pas le cas, les étapes à partir de l'appel à la fonction *self* sont réitérées jusqu'à obtenir un objet *generation* contenant au moins un individu femelle provenant de chacun des fondateurs. C'est cet objet *generation* qui est ensuite enregistré dans la liste *list generations*.
L'objet *generation* produit est ensuite formaté et le genre de chaque descendant attribué respectivement par les fonctions *Population_format* et *Recode_sterility_alleles*. L'objet *generation* est alors enregistré dans la liste *list generations*.

Random pollination

La modélisation de l'action de pollinisation aléatoire repose sur la fonction *selectOP* du package AlphaSimR. Cette fonction nécessite de renseigner les paramètres *nInd* et *nSeeds* qui correspondent respectivement au nombre d'individus sur lesquels récolter les graines produites, et combien de graines récolter par individu. Les parents des croisements sont choisis de la façon suivante : un vecteur de femelles est créé par échantillonnage aléatoire de *nInd* individus dans la population. Chaque femelle est répétée *nSeeds* fois dans le vecteur. Un vecteur de mâles de même longueur que celui des femelles est créé par échantillonnage aléatoire avec remise des individus dans la population. La pollinisation de chaque individu du vecteur des femelles par l'individu du vecteur des mâles correspondant crée ainsi une nouvelle génération composée de $nInd \times nSeeds$ individus issus d'une pollinisation aléatoire croisée (les autofécondations sont évitées en veillant à ne pas mettre un même individu à la même position dans le vecteur des femelles et des mâles). Cependant ce choix parfaitement aléatoire des parents ne permet pas de considérer le fait que certains individus sont des mâles stériles, ne pouvant pas polliniser d'autres individus. Le code de cette fonction a donc été repris et un nouveau paramètre d'entrée a été ajouté : un vecteur contenant le nom des individus femelles. Dans cette nouvelle fonction, appelée *random_pollination*, le vecteur des femelles est créé par échantillonnage aléatoire parmi les mâles stériles, et le vecteur des mâles est généré par échantillonnage aléatoire parmi les mâles fertiles. Cette fonction ainsi que les quatre fonctions d'enrobage décrites précédemment constituent la fonction *Random_pollination*, dont l'organisation est représentée dans la Figure 4.7. L'ordre d'appel aux différentes fonctions est le même que celui détaillé pour l'action *Selfing*, en remplaçant la fonction *self* par la fonction *random_pollination*.

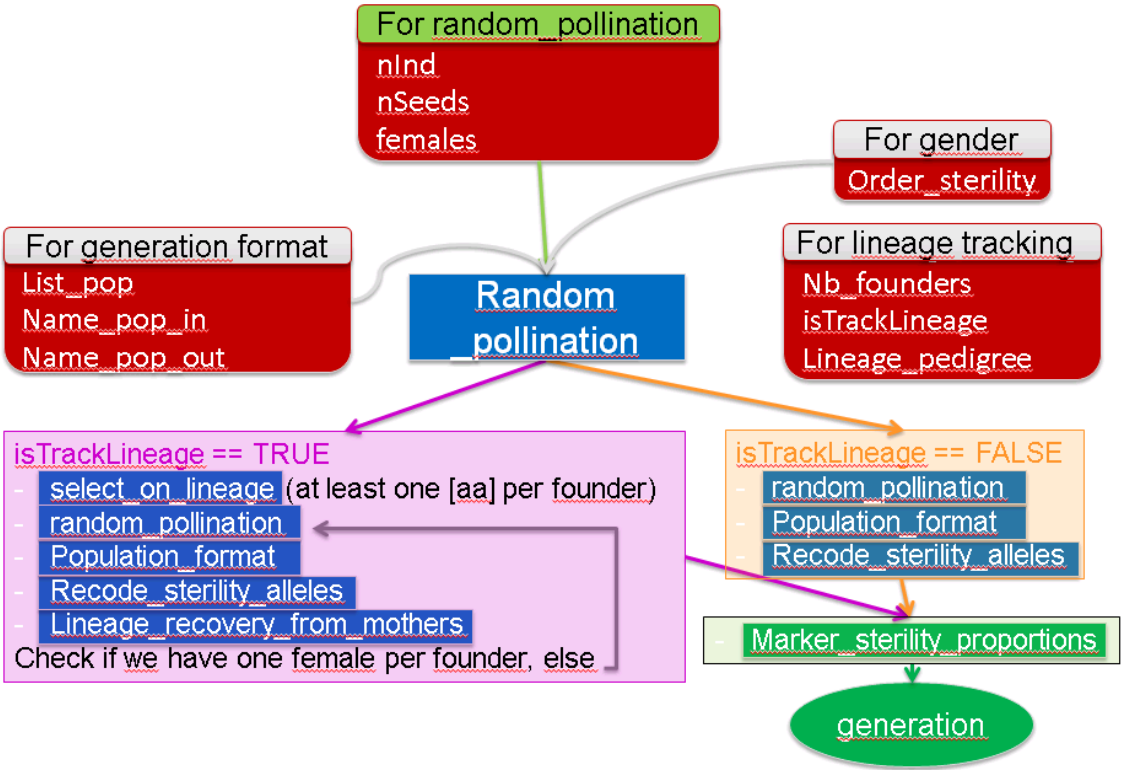


FIGURE 4.7 – Schéma de conception de la fonction de *Random Pollination*. Les paramètres d'entrée sont représentés par un fond rouge. Toutes les fonctions d'enrobage ainsi que la fonction *random_pollination* créée à partir de la fonction *selectOP* du package AlphaSimR sont représentées par un fond bleu foncé. La fonction de validation est représentée par un fond vert. L'objet *generation* en sortie de la fonction *Random Pollination* est représenté par un ovale vert.

Hybrid_production

La production d'Hybrides repose sur la fonction *hybridCross* du package AlphaSimR. Cette fonction modélise le croisement de deux éléments *generation*, l'un correspondant à la population de femelles, l'autre à une population de mâles. Dans ce schéma de pre-breeding, les hybrides sont créés par le croisement d'une population issues du schéma avec un pollinisateur élite. Un objet *generation* est donc créé pour représenter la population élite, composée d'un unique individu mâle fertile (AA). Cet objet est ajouté dans l'élément *list generations* sous le nom « elite » . Les éléments *generation* correspondant à la génération contenant les femelles à polliniser et celle correspondant au pollinisateur élite sont ainsi passés en entrée de la fonction *hybridCross*. Cette fonction est enrobée par les quatre fonctions d'enrobage, dont l'ordre d'appel est décrit dans la Figure 4.8.

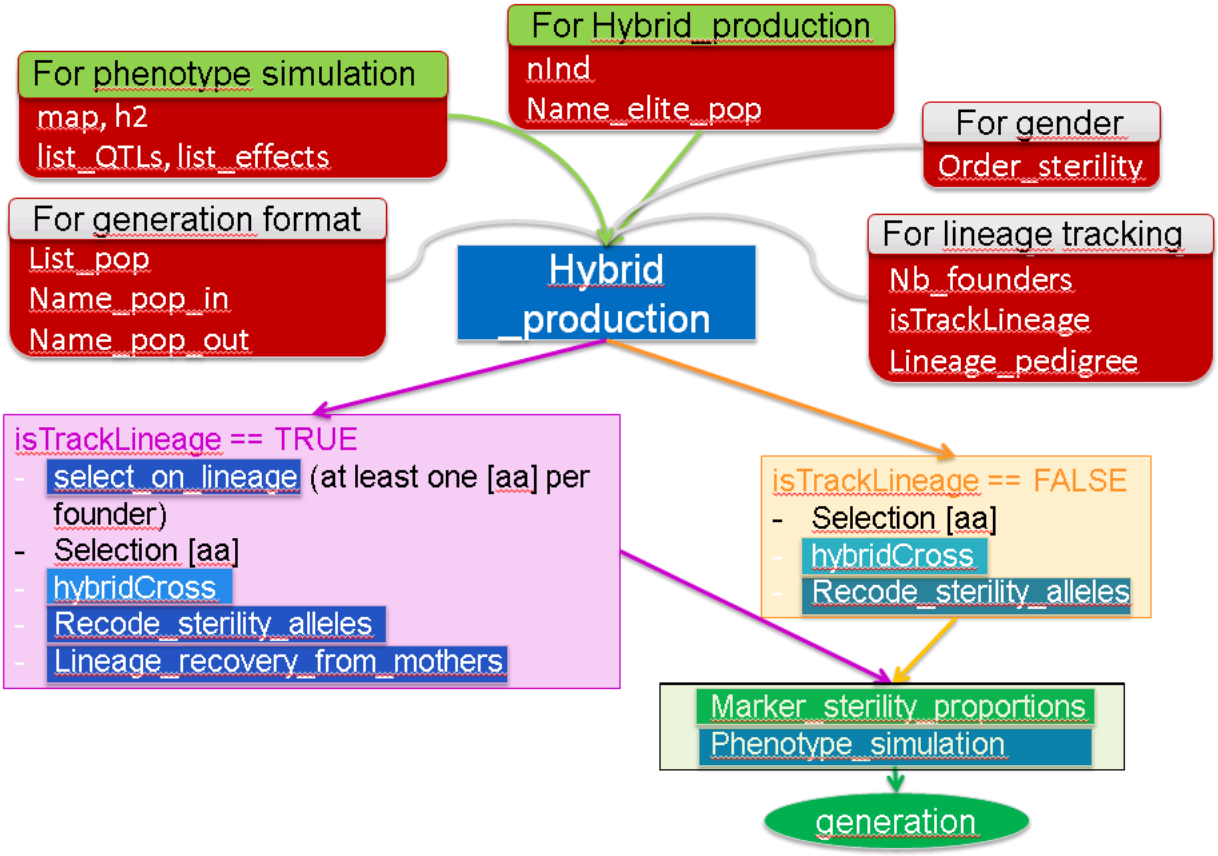


FIGURE 4.8 – Schéma de conception de la fonction de *Hybrid production*. Les paramètres d'entrée sont représentés par un fond rouge. Toutes les fonctions d'enrobage sont représentées par un fond bleu foncé, tandis que la fonction *hybridCross* provenant du package AlphaSimR est représentée par un fond bleu clair. La fonction de validation est représentée par un fond vert. L'objet *generation* en sortie de la fonction *Random Pollination* est représenté par un ovale vert.

L'intérêt lié à la production d'hybrides est la possibilité d'évaluer le phénotype de ces derniers. Une fonction de simulation du phénotype doit donc être intégrée à la fonction *Hybrid_production*. Cette action crée deux générations, celle de génération des hybrides et celle qui permet de les phénotyper. Par souci d'économie de mémoire, seul l'objet *generation* comprenant les hybrides avec un phénotype simulé est enregistré dans la liste *list generations*. Ce phénotype simulé est déterminé à partir de la valeur génétique de l'individu, à laquelle est ajoutée la valeur moyenne du phénotype RY évalué dans l'ensemble des 13 descendances AKER en entrée du schéma, ainsi qu'une erreur calculée selon la formule suivante :

$$\epsilon_i \sim \mathcal{N}(0, (1 - H^2))$$

avec ϵ_i l'erreur attribué à l'individu i, et H^2 l'héritabilité du caractère RY estimée sur l'ensemble des 13 descendances AKER.

Fonctions de sélection

L'un des intérêts du simulateur est de pouvoir comparer la performance des générations sélectionnées par les différentes méthodes, il est donc nécessaire de connaître la moyenne phénotypique des générations sélectionnées. Pour ce faire, le phénotype des individus issus d'une action de sélection est simulé.

GS

L'action de sélection génomique consiste à réaliser deux étapes successives : prédire la valeur génétique des individus d'une génération à partir de leurs informations génomiques, puis sélectionner les meilleurs individus vis-à-vis de cette valeur. La fonction de prédiction RRBLUP du package AlphaSimR a été abandonnée car elle utilise l'ensemble des informations génomiques des marqueurs présents dans l'élément *generation*, dont les pseudo-marqueurs ajoutés pour suivre les effets favorables ou défavorables des QTLs en fonction de leur provenance. C'est la fonction kin.blup du package rrBLUP qui a été utilisée (`https://CRAN.R-project.org/package=rrBLUP`), fonction basée sur l'approche EMMA détaillée dans le chapitre 1 partie 2.2. L'équation de prédiction établie est appliquée pour prédire la valeur génotypique des individus de la génération étudiée à partir de leur génotype. Les meilleurs individus sont alors sélectionnés sur cette valeur génotypique prédite, ainsi que sur le lignage maternel dans le cas ou ce lignage doit être suivi. Le phénotype de ces individus est ensuite simulé.
Un prérequis indispensable à l'utilisation de l'action de sélection génomique est d'avoir utilisé au préalable l'action de production d'hybrides, qui permet d'obtenir la population d'entraînement utilisée par la fonction kin.blup pour apprendre le modèle de sélection génomique.

PS

Les scénarios basés sur la sélection phénotypique ont été envisagés à un stade très tardif de la programmation du simulateur. Toutefois, tous les éléments nécessaires à la constitution d'une telle fonction étaient déjà intégrés au simulateur, la fonction de sélection phénotypique *PS* a donc pu être ajoutée aisément. La méthode de sélection phénotypique consiste à sélectionner les descendants des meilleurs individus (Aa) dont le phénotype est estimé sur descendance.
Trois étapes successives sont réalisées au cours de cette action : l'estimation de la valeur phénotypique des individus (Aa), la sélection des meilleurs de ces individus, et enfin la constitution d'un objet *generation* composé d'un nombre choisi des descendants de chacun de ces meilleurs individus. Pour ce faire, chaque individu (Aa) de la génération à

sélectionner est autofécondé. Les femelles qui en résultent sont croisées avec le pollinisateur élite pour créer des hybrides, dont le phénotype est simulé. La valeur phénotypique de l'individu (Aa) est alors en espérance la moyenne des hybrides. Cette première étape est représentée dans la Figure 4.9. Les meilleurs individus sont ensuite sélectionnés en fonction d'un taux d'intensité de sélection défini, et un nombre choisi des descendants de chacun de ces individus retenu de façon à constituer une génération composée en espérance de 1/2 de mâles et 1/2 de femelles (Figure 4.10). Le phénotype de ces individus est ensuite simulé.

GV

La vraie valeur génétique est connue au sein du simulateur. Il est donc possible de sélectionner les meilleurs individus sur cette vraie valeur génétique, ainsi que sur le lignage maternel dans le cas ou ce lignage doit être suivi. Le phénotype de ces individus est ensuite simulé.

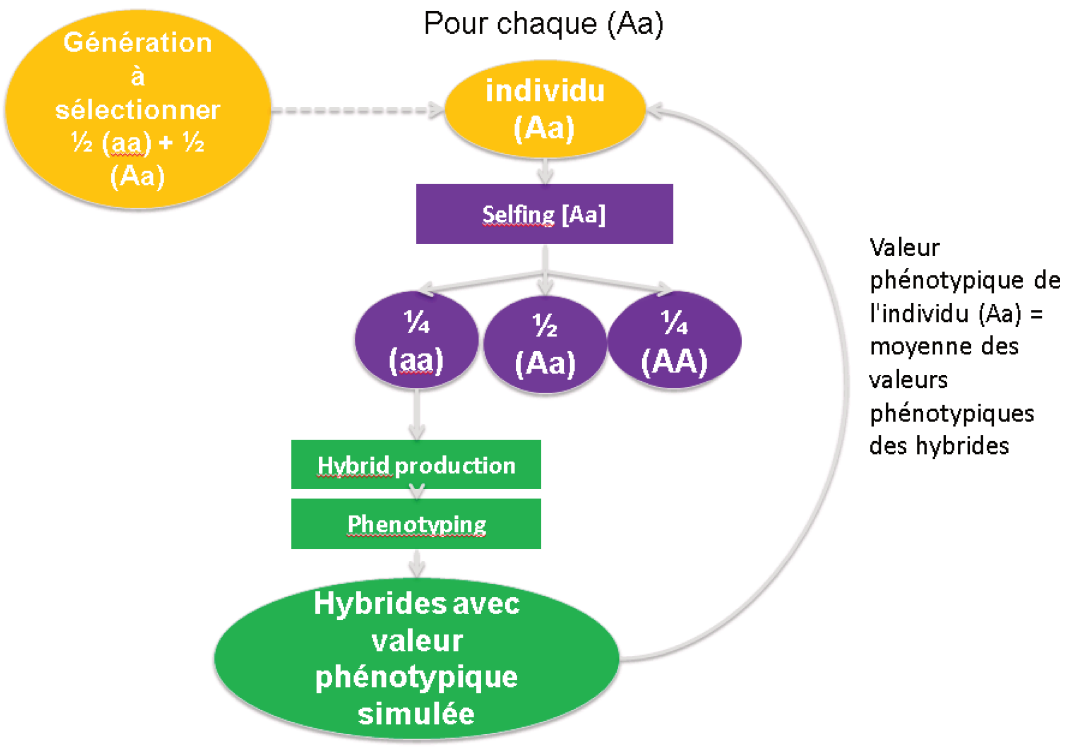


FIGURE 4.9 – Schématisation des étapes requises pour estimer la valeur phénotypique d'un individu dans le cadre d'un schéma de pre-breeding utilisant la sélection phénotypique.

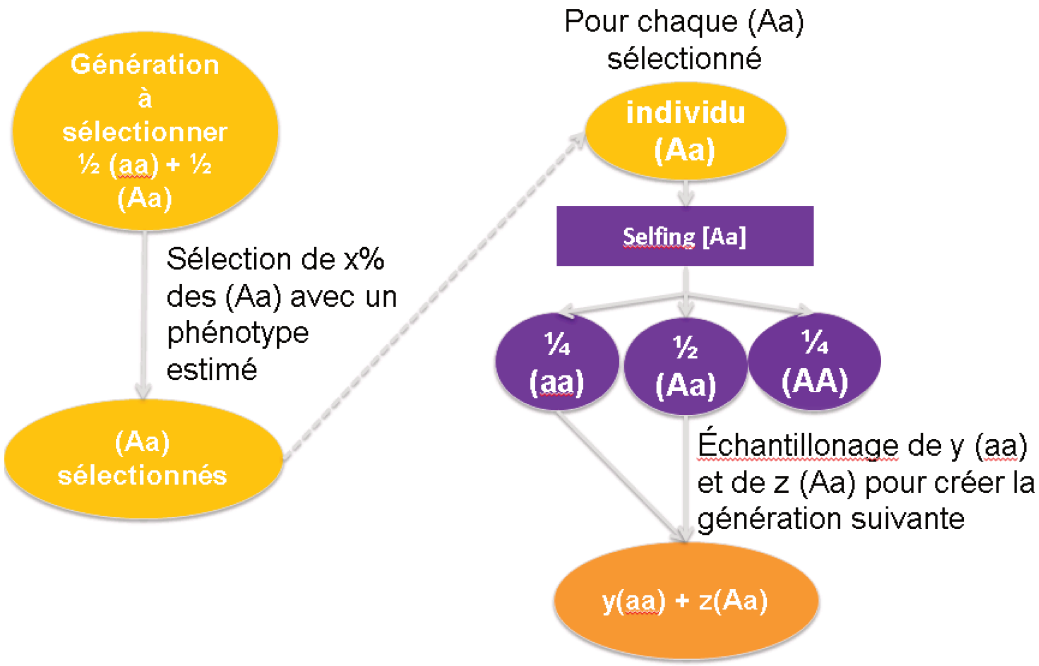


FIGURE 4.10 – Sélection des individus dans le cadre d'un schéma de pre-breeding utilisant la sélection phénotypique.

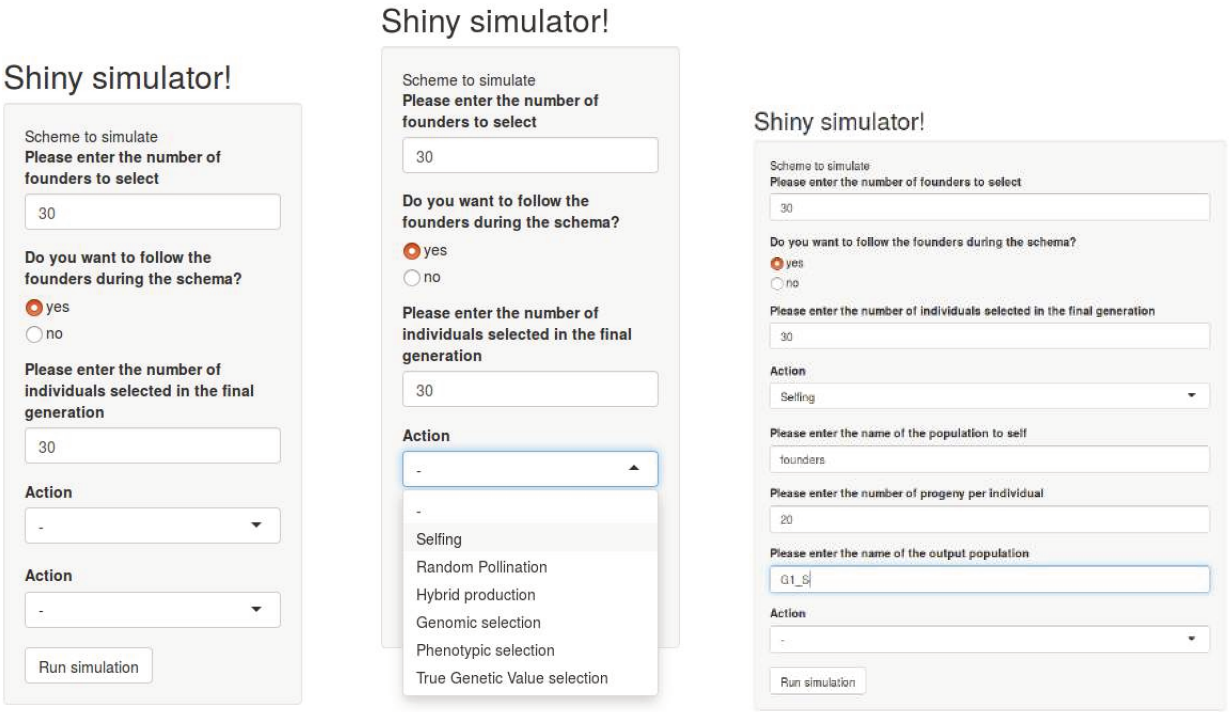
4.2.2 Phase de programmation

La phase de programmation est divisée en plusieurs scripts R. Le premier script permet de créer l'objet *generation* initial. Puisque chacun des 12 scénarios doit être simulé 30 fois, ce script permet la création de 30 éléments *generation* qui diffèrent vis-à-vis des 26 QTLs échantillonnés pour compléter l'architecture génétique. Un second script rassemble l'ensemble des fonctions nécessaires à la simulation des actions. Enfin, les différentes actions sont agencées pour former les différents schémas de pre-breeding : trois scripts sont écrits, un pour chaque schéma basé sur une méthode de sélection différente. Dans ces trois scripts, il est possible de préciser le nombre de fondateurs à sélectionner, le nombre d'individus désirés dans la population de pre-breeding finale, et si le lignage maternel des individus doit être suivi ou non. Les 12 scénarios identifiés lors de la phase d'analyse des besoins peuvent ainsi être simulés à partir de chacun des 30 éléments *generation* créés. Au total 360 simulations sont ainsi réalisées, chacune donnant en sortie l'élément *list generations* comprenant tous les objets *generation* créés au cours de la simulation.

Afin de permettre l'utilisation du simulateur à des personnes peu familiarisées avec le langage de programmation R, une transformation du code en interface graphique a été envisagée. Cette interface graphique doit permettre à l'utilisateur de choisir le nombre

et l'ordre des actions constituant le schéma qu'il souhaite simuler, ainsi que de choisir le scénario à réaliser.

Le package R Shiny (`https://CRAN.R-project.org/package=shiny`) a tout d'abord été expérimenté. Ce package gratuit permet la création de pages web interactives permettant de réaliser les analyses R. L'application est composée de deux fichiers : un fichier « ui.R » pour User Interface script, qui permet de déterminer la mise en page et l'apparence de l'application, et un fichier « server.R » qui contient les instructions nécessaires pour lancer l'application et effectuer les analyses. Cependant ce package nécessite de coder au préalable le nombre de blocs maximum utilisables par l'utilisateur. Un exemple de l'utilisation de ce package est présenté dans la Figure 4.11.



(a) (b) (c)

FIGURE 4.11 – Simulateur Shiny avec seulement 2 blocs Action possibles (a). L'utilisateur est invité à choisir le scénario à simuler : nombre de fondateurs, nombre d'individus dans la population finale, suivi ou non des fondateurs (b). Chaque bloc Action contient toutes les actions possibles, ce qui implique un grand nombre de répétitions dans le code R. Lorsque l'utilisateur choisit une action, il doit renseigner les paramètres d'entrée (c).

L'interface graphique créée ne permet pas une grande flexibilité dans la création des schémas à simuler, utiliser le package Shiny semble ne pas être une solution très rapide ni optimale vis-à-vis du code. Le logiciel R AnalyticFlow a alors été testé (`https://r.analyticflow.com/en/#feature`). Ce logiciel gratuit intégrant le langage R permet de générer très rapidement une interface graphique très intuitive, où l'utilisateur peut déplacer, dupliquer, supprimer, des blocs représentant chaque action pour créer le schéma à simuler. Le schéma intégrant la sélection génomique a ainsi été représenté dans la Figure

4.12.

Chaque Action est précédée par une icône représentant les paramètres d'entrée de l'Action. Lorsque l'utilisateur clique sur cet icône de paramètres, un tableau apparaît (Figure 4.13). La première colonne contient le nom des variables qui seront utilisées dans l'Action suivante. L'utilisateur doit renseigner la valeur de ces variables dans la 2nde colonne. Enfin, la 3ème colonne décrit succinctement la variable.

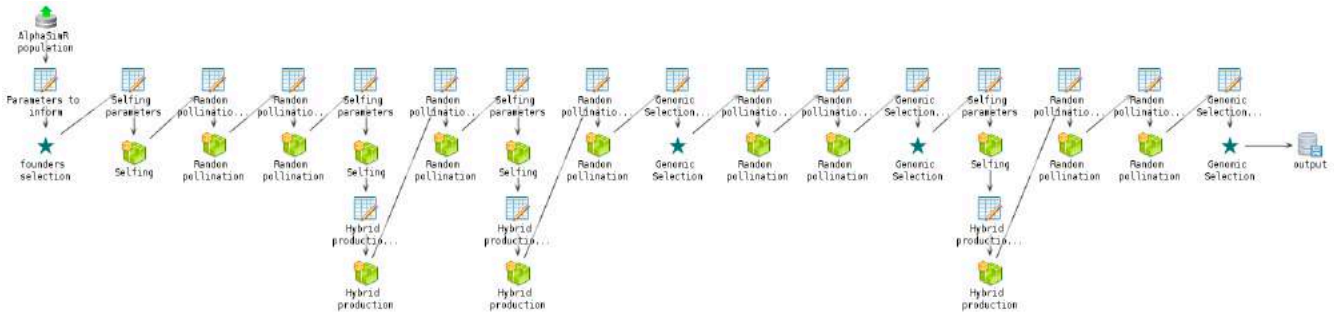


FIGURE 4.12 – Interface graphique du logiciel R AnalyticFlow représentant le schéma de pre-breeding basé sur la sélection génomique. La première et la dernière icône représentent respectivement la population en entrée du simulateur et la liste des générations en sortie du simulateur. Les icônes en forme d'étoiles correspondent à une action de sélection, et les icônes vertes sont des actions de croisement. Les icônes représentant un tableau et un crayon désignent les paramètres à renseigner pour effectuer l'action qui suit.

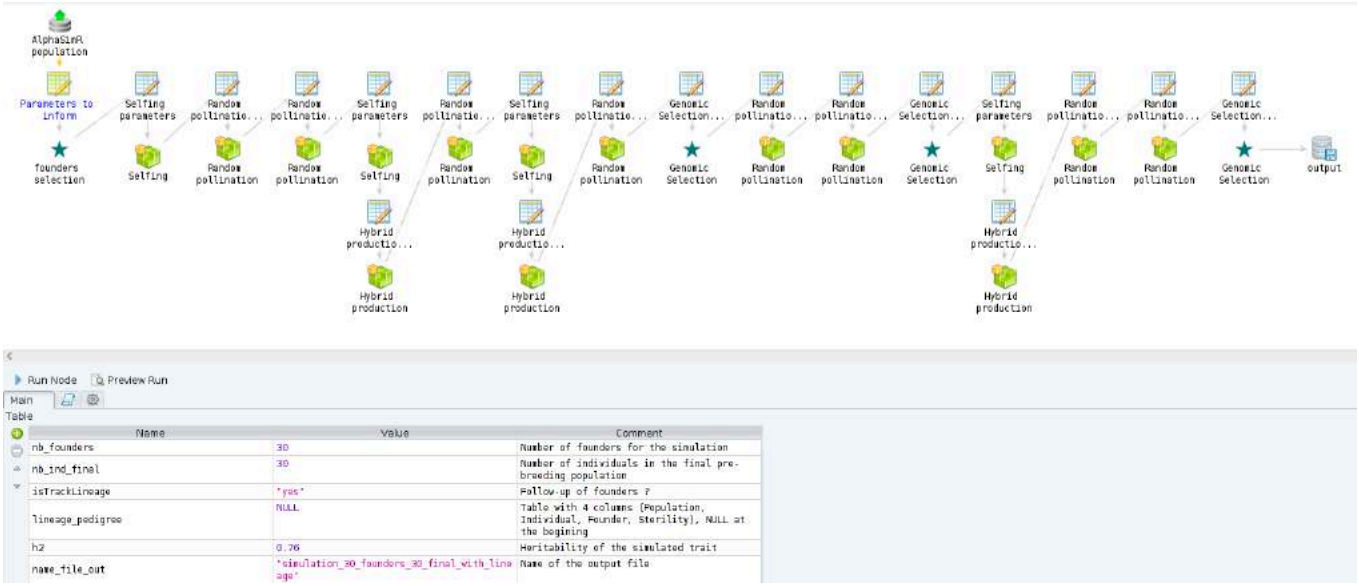


FIGURE 4.13 – Interface graphique du logiciel R AnalyticFlow représentant le schéma de pre-breeding basé sur la sélection génomique. L'icône des paramètres précédant l'action de sélection des fondateurs est sélectionnée, permettant l'affichage du tableau présentant les différents paramètres sous le schéma.

Lorsque l'utilisateur clique sur une Action, le code R composant cette Action est alors affiché. Ceci permet une bonne traçabilité du code, lisible à tout moment (Figure 4.14).

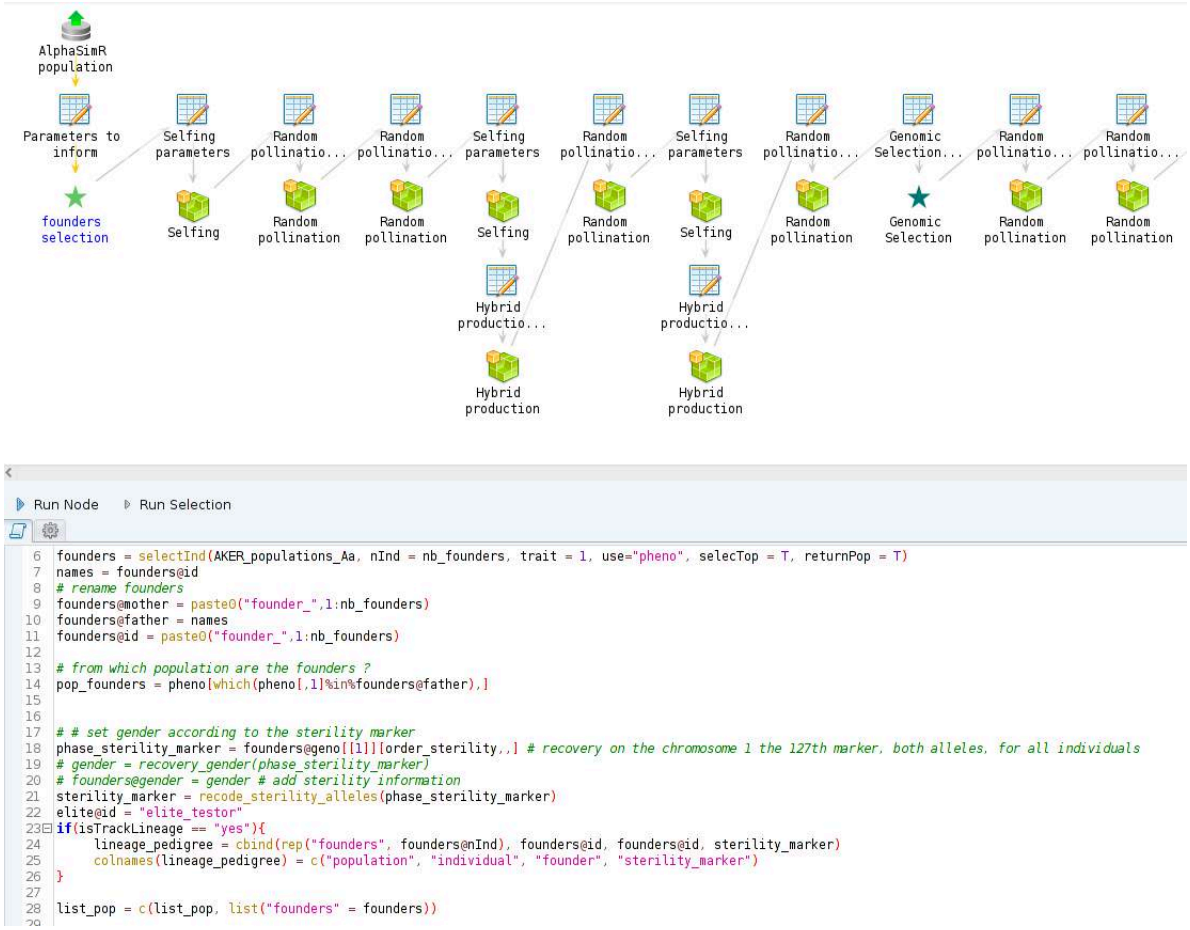


FIGURE 4.14 – Interface graphique du logiciel R AnalyticFlow représentant le schéma de pre-breeding basé sur la sélection génomique.

L'utilisation du logiciel R AnalyticFlow permet donc de produire une interface graphique intuitive, flexible et permettant la traçabilité du code.

4.2.3 Phase de validation

La phase de validation consiste à vérifier que la phase de programmation a bien produit ce qui était attendu pour chaque phase de conception. Selon la Figure présentant le cycle en V (Figure 4.3), à la phase de conception détaillée correspond un test de validation unitaire. Cette validation a pour objectif de vérifier le bon fonctionnement de chacune des actions décrites. Lors du développement des actions de sélection, une vérification manuelle a par exemple été réalisée pour vérifier que les individus sélectionnés par la fonction étaient bien ceux attendus au vu des valeurs. La fonction *Marker_sterility_proportions* a également été développée et intégrée dans chacune des actions, permettant de connaître les proportions d'individus femelles (aa) et mâles (Aa) et (AA) produites par l'action et de vérifier qu'elles correspondent aux valeurs espérées. A la phase de conception architecturale est

associée un test d'intégration. Ce test consiste à vérifier que les différentes actions peuvent être utilisées successivement sans provoquer de problème, ou en provoquant des erreurs attendues. Par exemple il n'est pas possible d'utiliser l'action de sélection génomique si aucune population d'entraînement n'a été générée au préalable. La phase de spécifications externes est vérifiée par un test de validation. Ce test permet de vérifier que tous les scénarios définis fonctionnent. Enfin, la phase d'analyse des besoins est associée à la recette, c'est à dire aux délivrables. Le simulateur est livré sous la forme de script R ainsi que sous la forme de l'interface R AnalyticFlow.

Chapitre 5

Simulations et comparaison des schémas de pre breeding

Dans le chapitre précédent, trois actions de croisements et trois actions de sélection ont été définies. L'enchaînement de ces actions permet de constituer trois schémas de pre-breeding distincts : un schéma utilisant la sélection génomique (GS) (Figure 5.1), un schéma s'appuyant sur la sélection phénotypique (PS) (Figure 5.2), et un dernier schéma basé sur la sélection sur la vraie valeur génétique (GV) (Figure 5.3). Le schéma de GS a été décrit en détail dans le chapitre 4. Le schéma de GV correspond au schéma de GS sans les étapes permettant de produire les populations d'entraînement, inutiles ici puisque les actions de sélection génomique utilisant ces populations sont remplacées par des actions de sélection sur la vraie valeur génétique. Le schéma de PS est constitué de façon à ce que le taux de sélection et l'effectif des générations sélectionnées soient identiques entre le schéma de PS et de GS. Concernant les effectifs, le schéma des sélectionneurs correspond au scénario débutant avec 300 fondateurs pour aboutir à 300 individus dans la population finale (scénario 300 $\rightarrow$ 300). Chaque génération du scénario 30 $\rightarrow$ 30 comporte 10 fois moins d'individus que celle qui lui correspond dans le scénario 300 $\rightarrow$ 300. Dans le scénario 30 $\rightarrow$ 300 les effectifs des générations sont équivalents à ceux des mêmes générations dans le scénario 30 $\rightarrow$ 30 jusqu'en G6, puis égaux à ceux des générations produites par le scénario 300 $\rightarrow$ 300. Le tableau ci-dessous présente les taux de sélection appliqués à chaque étape de sélection et les effectifs des générations qui en résultent en fonction des différents scénarios 5.1. Les effectifs de chacune des générations simulées sont donnés en annexes (Figure A.2 pour la GS, Figure A.3 pour la GV et Figure A.4 pour la PS).

étape de sélection	intensité de sélection	nom de la génération sélectionnée GS	GV	PS	effectif 30–>30	effectif 30–>300	effectif 300 –> 300
1	20%	G7_GS	G7_GV	G7_PS	120	1200	1200
2	20%	G9_GS	G9_GV	G12_PS	60	600	600
3	5%	G12_GS	G12_GV	G17_PS	30	300	300

TABLE 5.1 – Taux de sélection, nom et effectif des générations issues d'actions de sélection

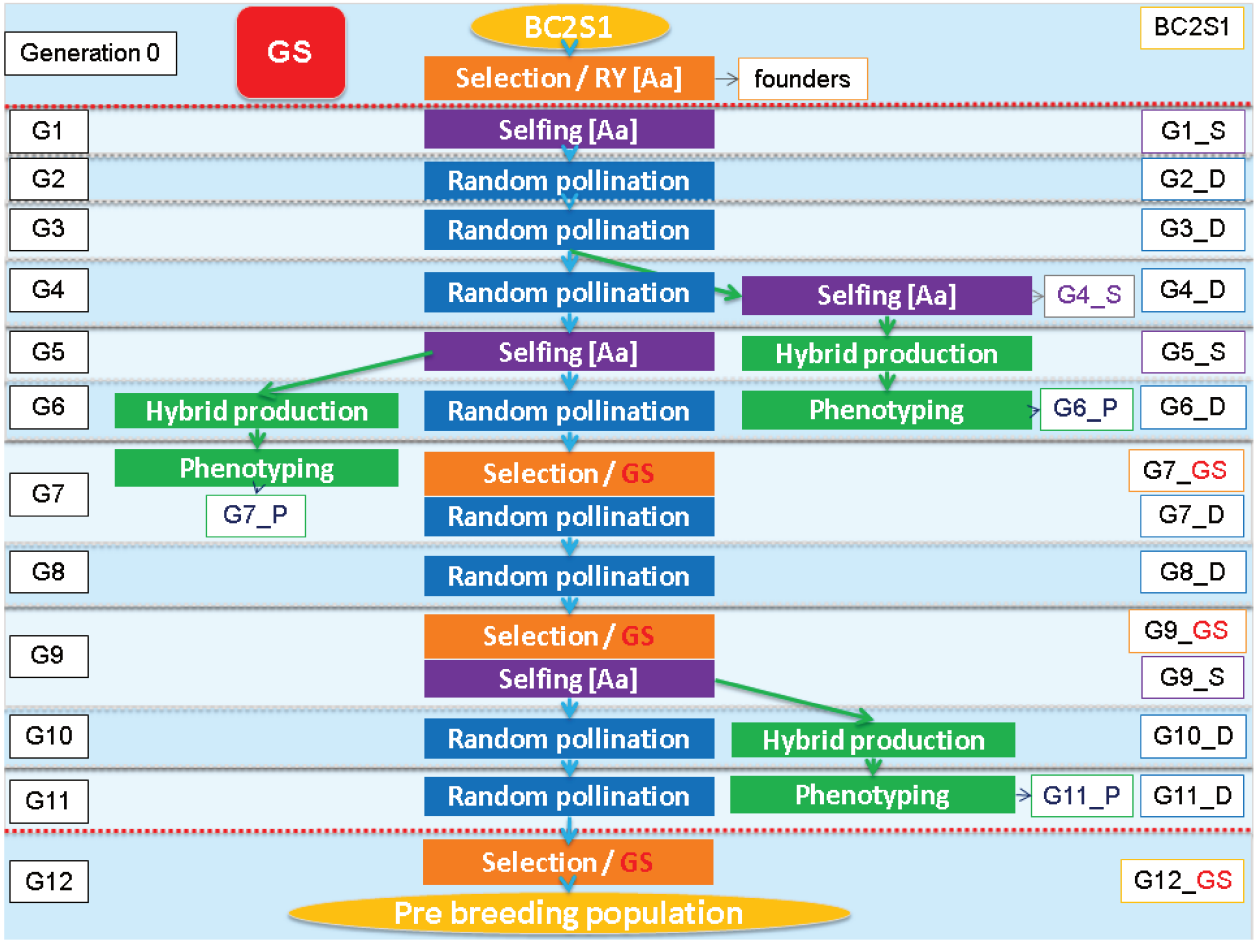


FIGURE 5.1 – Enchaînement d'actions formant le schéma de sélection basé sur la sélection génomique (GS)

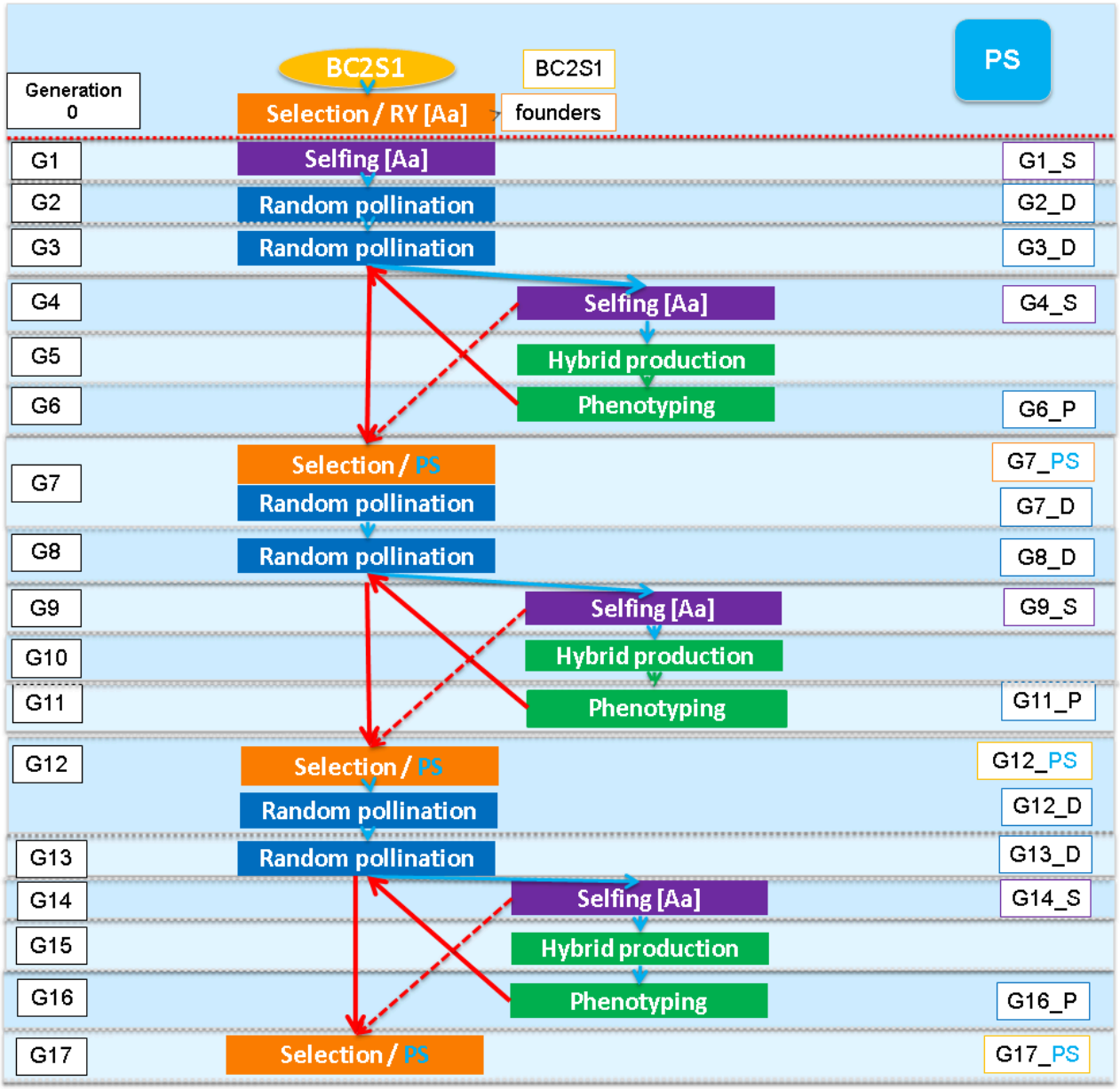


FIGURE 5.2 – Enchaînement d'actions formant le schéma de sélection basé sur la sélection phénotypique (PS)

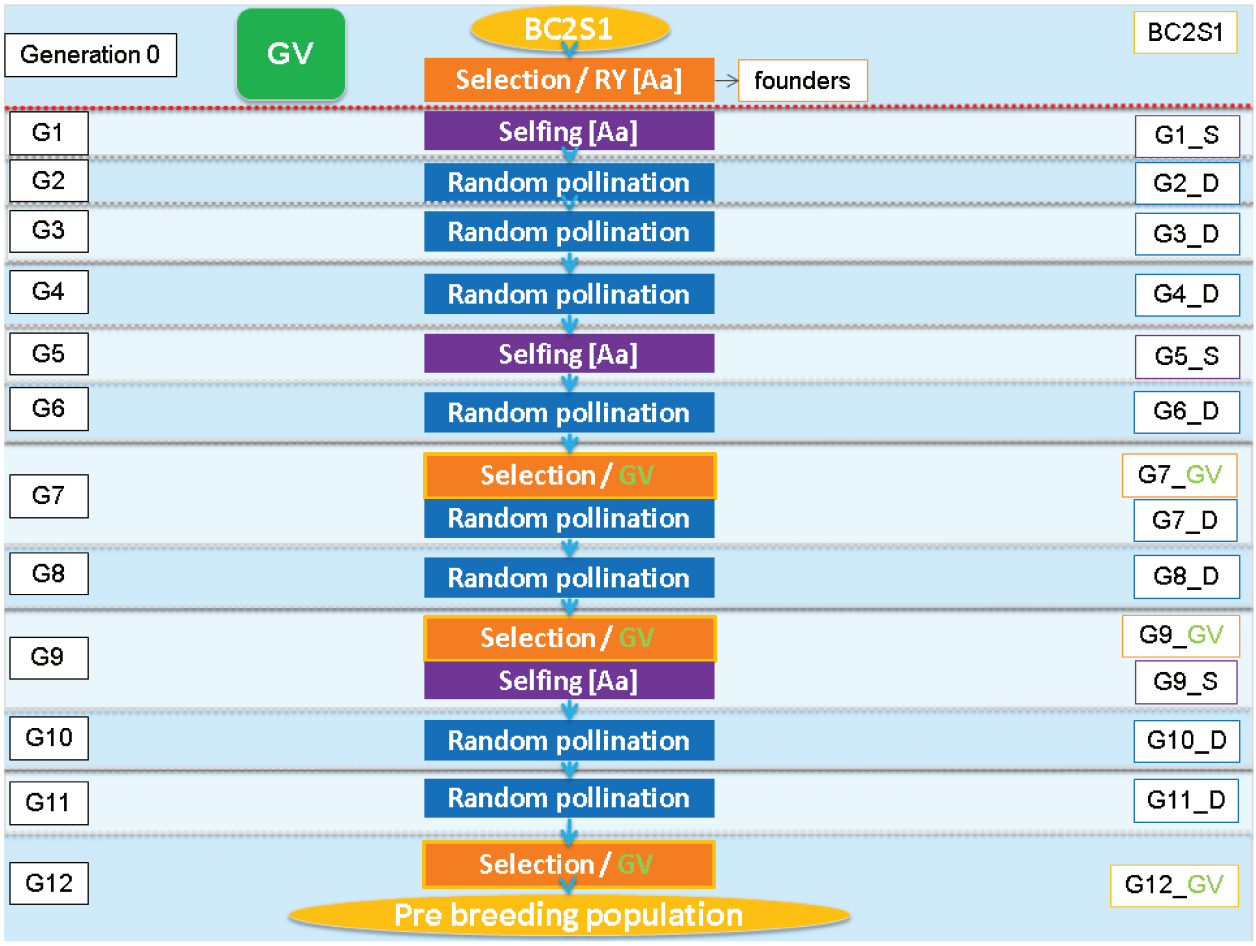


FIGURE 5.3 – Enchaînement d'actions formant le schéma de sélection basé sur la vraie valeur génétique (GV)

Dans le chapitre précédent, 12 scénarios de schéma de pre-breeding ont été définis (Tableau 4.1), chaque scénario étant simulé 30 fois. La population en entrée de ces différents scénarios est composée des 13 populations AKER étudiées vis-à-vis de l'architecture génétique du rendement racinaire (RY). Cette population intègre ainsi une nouvelle variabilité génétique par rapport aux populations de breeding existantes, diversité apportée par les 13 accessions exotiques utilisées dans le programme AKER. Les simulations doivent permettre de comparer ces 12 scénarios afin de choisir le scénario optimal, c'est-à-dire produisant une population de pre-breeding ayant une bonne performance vis-à-vis du caractère étudié, ici le rendement racinaire, tout en évitant au maximum l'érosion de la diversité génétique. Afin de pouvoir comparer les différents scénarios en fonction de ces deux critères, il nous faut définir des indicateurs permettant d'évaluer leur évolution. La première partie de ce chapitre est consacrée à la description des différents indicateurs mis en place pour mesurer le gain génétique et la diversité génétique au cours des simulations. La seconde partie présente les comparaisons des différents scénarios et les conclusions qui en découlent.

5.1 Indicateurs

5.1.1 Gain génétique

Le premier objectif du schéma de pre-breeding est de produire une population pouvant servir de réservoir de matériel génétique nécessaire à la création de nouvelles variétés élites. Il faut donc que les individus de la population de pre-breeding produite aient une bonne performance vis-à-vis du caractère étudié, ici le rendement racinaire. L'estimation du gain génétique dans chaque génération créée au cours des simulations permet de mesurer l'évolution de cette performance. Afin de comparer le gain génétique obtenu avec ces différentes méthodes de sélection, il est nécessaire de définir un estimateur du gain génétique. Deux choix se présentent : calculer la moyenne des valeurs phénotypiques de la génération sélectionnée, ou la moyenne des valeurs génétiques de cette génération. La moyenne des valeurs génétiques constituerait un estimateur plus précis du gain génétique, mais ces valeurs ne sont connues qu'au sein du simulateur, *in silico*. Afin de rester au plus près des valeurs réellement observables par les sélectionneurs, c'est donc la moyenne sur les valeurs phénotypiques qui est retenue comme estimateur.

5.1.2 Diversité de combinaison allélique

Le schéma de pre-breeding optimal doit éviter l'érosion de la diversité génétique. Les schémas implémentés sont inspirés du schéma de type Doggett décrit dans la partie 1.2.1 : les *random pollination* successives n'apportent pas de nouveaux allèles, mais permettent de favoriser l'apparition de combinaisons nouvelles grâce aux crossing overs. Comme décrit dans la partie 4.2.1, la diversité de combinaison allélique peut être influencée par l'effectif des populations, et le fait de suivre le lignage maternel des fondateurs ou non. Afin de comparer les différents scénarios, cinq indicateurs de cette diversité de combinaison allélique sont définis.

Fréquence de l'allèle exotique aux locus « favorables »

Lors de l'étude de l'architecture génétique du rendement racinaire, plusieurs locus ont été identifiés comme ayant un effet favorable de l'allèle exotique. Le maintien de la diversité de combinaison allélique utile apportée par l'exotique peut être évalué grâce au calcul de la fréquence de l'allèle exotique au niveau de ces locus dits « favorables » au sein des différentes générations produites au cours du schéma de pre-breeding simulé. Pour plus de précision, la fréquence du génotype homozygote exotique aux locus « favorables » ainsi que la fréquence du génotype hétérozygote à ces locus sont également calculées.

Calcul de la moyenne de la fréquence du génotype homozygote exotique sur les locus « favorables » dans une génération :

$$fr(\mathrm{XX}_f) = \frac{\sum_{l_f} \frac{\#\{x^{l_f}=\mathrm{XX}^{l_f}\}}{N}}{L_f}$$

où $fr(\mathrm{XX}_f)$ est la fréquence du génotype homozygote exotique aux locus « favorables » dans une génération, x^l est le génotype x au locus « favorable » l_f, L_f est le nombre de locus « favorables » avec $l_f = 1, ..., L_f$, et N est le nombre d'individus composant la génération avec $i = 1, ..., N$.

Calcul de la moyenne de la fréquence du génotype hétérozygote sur les locus « favorables » au sein d'une génération :

$$fr(\mathrm{EX}_f) = \frac{\sum_{l_f} \frac{\#\{x^{l_f}=\mathrm{EX}^{l_f}\}}{N}}{L_f}$$

où $fr(\mathrm{EX}_f)$ est la fréquence du génotype hétérozygote aux locus « favorables » dans une génération, x^{l_f} est le génotype x au locus « favorable » l_f, L_f est le nombre de locus « favorables » avec $l_f = 1, ..., L_f$, et N est le nombre d'individus composant la génération avec $i = 1, ..., N$.

Calcul de la moyenne de la fréquence de l'allèle exotique sur les locus « favorables » dans une génération :

$$fr(\mathrm{X}_f) = \frac{2fr(\mathrm{XX}_f) + fr(\mathrm{EX}_f)}{2N}$$

où $fr(\mathrm{X}_f)$ est la fréquence de l'allèle exotique dans une génération, $fr(\mathrm{XX}_f)$ est la fréquence du génotype homozygote exotique dans une génération, $fr(\mathrm{EX}_f)$ est la fréquence du génotype hétérozygote dans une génération, et N est le nombre d'individus composant la génération avec $i = 1, ..., N$.

Fréquence de l'allèle exotique sur l'ensemble des locus

L'érosion du nombre d'allèle dans une population est un témoin de la dérive génétique, et indique donc une réduction de la diversité de combinaison allélique. L'évolution de la diversité de combinaison allélique peut ainsi être évaluée grâce au calcul de la fréquence d'exotique sur l'ensemble des locus au sein des différentes générations produites au cours

du schéma de pre-breeding simulé. Un schéma de pre-breeding favorable est un schéma permettant d'éviter l'érosion de la diversité génétique, la fréquence de l'allèle exotique sur l'ensemble des locus ne doit pas diminuer au cours des générations. La fréquence du génotype homozygote exotique sur l'ensemble des locus ainsi que la fréquence du génotype hétérozygote sur ces mêmes locus sont également calculés.

Calcul de la moyenne de la fréquence du génotype homozygote exotique sur l'ensemble des locus dans une génération :

$$fr(\mathrm{XX}) = \frac{\sum_{l} \frac{\#\{x^l = \mathrm{XX}^l\}}{N}}{L}$$

où $fr(\mathrm{XX})$ est la fréquence du génotype homozygote exotique dans une génération, x^l est le génotype x au locus l, L est le nombre de locus avec $l = 1, ..., L$, et N est le nombre d'individus composant la génération avec $i = 1, ..., N$.

Calcul de la moyenne de la fréquence du génotype hétérozygote sur l'ensemble des locus au sein d'une génération :

$$fr(\mathrm{EX}) = \frac{\sum_{l} \frac{\#\{x^l = \mathrm{EX}^l\}}{N}}{L}$$

où $fr(\mathrm{EX})$ est la fréquence du génotype hétérozygote aux locus « favorables » dans une génération, x^l est le génotype x au locus l, L est le nombre de locus avec $l = 1, ..., L$, et N est le nombre d'individus composant la génération avec $i = 1, ..., N$.

Calcul de la moyenne de la fréquence de l'allèle exotique sur l'ensemble des locus au sein d'une génération :

$$fr(\mathrm{X}) = \frac{2fr(\mathrm{XX}) + fr(\mathrm{EX})}{2N}$$

où $fr(\mathrm{X})$ est la fréquence de l'allèle exotique dans une génération, $fr(\mathrm{XX})$ est la fréquence du génotype homozygote exotique dans une génération, $fr(\mathrm{EX})$ est la fréquence du génotype hétérozygote dans une génération, et N est le nombre d'individus composant la génération avec $i = 1, ..., N$.

Longueur moyenne des fragments exotiques

Chaque individu BC2S1 composant la population en entrée du schéma de pre-breeding

comporte un fragment de l'accession exotique dont il est issu. Plus les fragments exotiques sont courts, plus il est aisé de repérer les individus comportant des allèles exotiques avec un effet favorable. Un schéma de pre-breeding est donc considéré comme favorable lorsqu'il permet de créer une population de pre-breeding intégrant des fragments exotiques relativement courts. La longueur de ces fragments exotiques peut être évaluée par le calcul de la longueur moyenne des fragments exotiques dans chaque génération simulée.

Pour un individu et sur un chromosome, la longueur d'un segment exotique est calculée selon la formule suivante :

$$LF = d_{end} - d_{start}$$

avec LF la longueur du fragment exotique, d_{end} la position du locus de rupture exotique/élite l (où l'individu étudié présente l'allèle exotique sous forme homozygote ou hétérozygote, alors qu'il ne présente que l'allèle élite au locus $l+1$ ou qu'il constitue le dernier locus du chromosome), et d_{start} la position du locus de rupture élite/exotique l (où l'individu présente l'allèle exotique sous forme homozygote ou hétérozygote, alors qu'il ne présente que l'allèle élite au locus $l-1$ ou qu'il s'agit du premier locus sur le chromosome étudié).
La longueur moyenne des fragments exotiques au sein de la génération est alors calculée selon la formule suivante :

$$\mu_{LF} = \frac{\sum_{e} LF_e}{n_e}$$

où μ_{LF} est la longueur moyenne des fragments exotiques dans la génération étudiée, LF_e la longueur du fragment exotique e avec $e = 1, ..., n_e$, et n_e est le nombre de fragments exotiques dans la population.

Nombre de dimensions significatives de l'ACP

L'Analyse en Composantes Principales, ou ACP est une méthode statistique très utilisée en génétique des populations pour déterminer l'ensemble des axes représentant les différences entre les individus d'une population. Elle consiste à transformer orthogonalement un ensemble de variables potentiellement corrélées en différentes entités indépendantes, représentant chacune une combinaison linéaire de variables corrélées. Ces vecteurs sont appelés des composantes principales. Un indicateur de la structuration de la diversité de combinaison allélique présente au sein de la génération étudiée peut ainsi être défini comme le nombre de composantes principales significatives de l'ACP, c'est-à-dire

le nombre de composantes principales qui pour un risque de 5% sont significativement non réduites à un espace nul. Pour déterminer ce nombre de composantes principales significative, le test de Tracy-Widom est utilisé. En effet, (Patterson et al. 2006) indiquent que pour une population structurée en K populations, l'ACP révèle K-1 axes de variation orthogonaux correspondant à K-1 valeurs propres significatives au sens de Tracy-Widom. Le package (Frichot and Francois 2014) est utilisé pour réaliser ce test.
Le nombre d'individus présents dans une génération peut impacter le nombre de composantes principales significative. Les générations n'étant pas constituées du même nombre d'individus, il est nécessaire de diviser le nombre de composantes principales significatives obtenues par le nombre d'individus afin d'obtenir un indicateur comparable entre les générations.
L'objectif du schéma de pre-breeding étant de constituer un réservoir de diversité génétique non structuré, les scénarios produisant une génération de pre-breeding où le nombre de dimension significatives de l'ACP rapporté au nombre d'individus est élevé seront considérés comme favorables, témoignant d'une faible structuration de la population.

Indicateurs de la matrice de Kinship

La matrice de covariance génétique, aussi appelée matrice de Kinship, peut être estimée à partir des marqueurs moléculaires. Une façon simple de calculer la parenté entre les individus sur la base des marqueurs moléculaires est de considérer la proportion d'allèles qu'ils partagent. Ce coefficient est appelé AIS pour *Alike In State*. La matrice AIS est calculée selon la formule suivante (Maenhout et al. 2009) :

$$AIS(i_1, i_2) = \frac{\boldsymbol{G_1}'\boldsymbol{G_2} + (\mathbf{2} - \boldsymbol{G_1})'(\mathbf{2} - \boldsymbol{G_2})}{4L}$$

avec L le nombre total de marqueurs, $\boldsymbol{G_1}$ et $\boldsymbol{G_2}$ le vecteur de génotypes pour les individus i_1 et i_2 de longueur L, codé 0 pour le génotype homozygote élite, 2 pour le génétoype homozygote exotique, et 1 pour le génotype hétérozygote, et $\mathbf{2}$ un vecteur de « 2 ».

Trois indicateurs de la diversité de combinaison allélique dans la génération étudiée peuvent être déduits de cette matrice : le maximum de la kinship, qui indique le degré de proximité génétique entre les deux individus les plus proches, le minimum de la kinship, qui renseigne sur le degré de divergence des deux individus les plus éloignés, et la moyenne

de la kinship, qui représente la ressemblance moyenne entre les individus d'une génération donnée. L'objectif du schéma de pre-breeding étant de produire un réservoir de diversité génétique, un schéma sera considéré comme favorable si la valeur maximale de la kinship est faible, indiquant que les individus qui se ressemblent le plus sont tout de même assez différents, si la valeur minimale de la kinship est faible, signe que les individus les plus éloignés sont vraiment très différents du point de vue génétique, et si la valeur moyenne de la kinship est faible, illustrant que dans l'ensemble de la population les individus de la population sont éloignés génétiquement.

5.2 Comparaisons des différents schémas de pre-breeding

5.2.1 Etude de l'évolution du gain génétique au cours du schéma de pre-breeding selon les différentes méthodes de sélection et l'effectif des générations

5.2.1.1 Impact des méthodes de sélection

Différentes méthodes de sélection ont été définies lors de la phase de conception du simulateur : la sélection génomique, la sélection phénotypique, et la sélection sur la vraie valeur génétique. La vraie valeur génétique est la valeur génétique intrinsèque des individus sans influence de facteurs extérieurs. Elle est exclusivement connue au sein du simulateur : il n'est pas possible pour les sélectionneurs de mettre en place un programme de pre-breeding basé sur cette sélection. Les populations obtenues dans le simulateur par sélection sur la vraie valeur génétique représentent les générations optimales qu'il est possible d'obtenir par sélection. Comparer la moyenne phénotypique de ces populations avec celle des populations obtenues par sélection génomique permet de se rendre compte de l'efficacité de la sélection génomique. L'évolution du gain génétique des schémas de pre-breeding basés sur la sélection génomique est également comparée à celle issue des schémas basés sur la sélection phénotypique afin de déterminer la méthode de sélection la plus avantageuse vis-à-vis de l'évolution de la performance des populations.
Les schémas issus des trois méthodes de sélection sont comparés dans les Figures 5.4, 5.5 et 5.6 pour les scénarios sans suivi du lignage maternel débutant respectivement avec 30 fondateurs pour aboutir à une population de pre-breeding composée de 30 individus ($30 \rightarrow 30$), 30 fondateurs et 300 individus dans la population de pre-breeding finale ($30 \rightarrow 300$), et 300 fondateurs pour 300 individus dans la population de pre-breeding finale ($300 \rightarrow 300$). Chaque scénario a été simulé 30 fois. La moyenne phénotypique de

chacune des générations issue d'une action de sélection est donnée dans les Tableaux 5.2, 5.3 et 5.4.

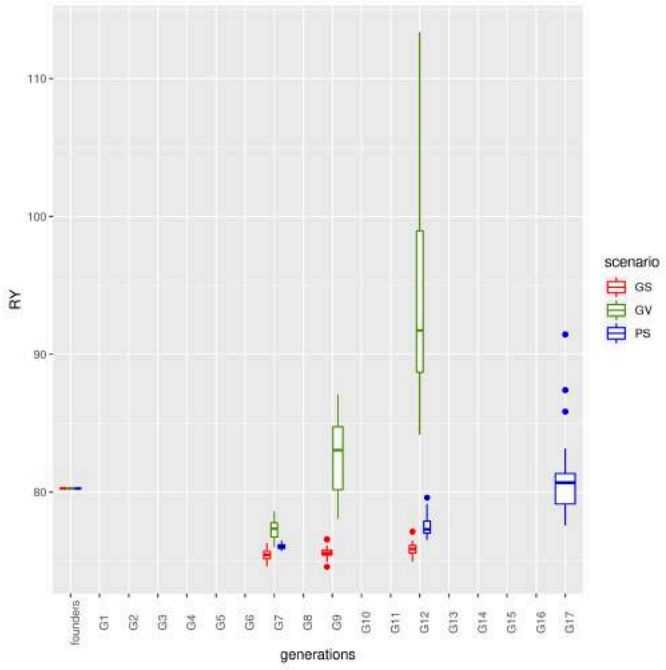


FIGURE 5.4 – Moyenne phénotypique des générations sélectionnées dans un schéma de pre-breeding 30 → 30 sans suivi du lignage maternel par sélection génomique (GS), sélection sur la vraie valeur génétique (GV) ou sélection phénotypique (PS)

generation	GS	GV	PS
founders	80.26	80.26	80.26
G7	75.44	77.31	76.06
G9	75.58	82.66	-
G12	75.86	93.87	77.53
G17	-	-	80.95

TABLE 5.2 – Moyenne phénotypique des générations sélectionnées dans un schéma de pre-breeding 30 → 30 sans suivi du lignage maternel par sélection génomique (GS), sélection sur la vraie valeur génétique (GV) ou sélection phénotypique (PS)

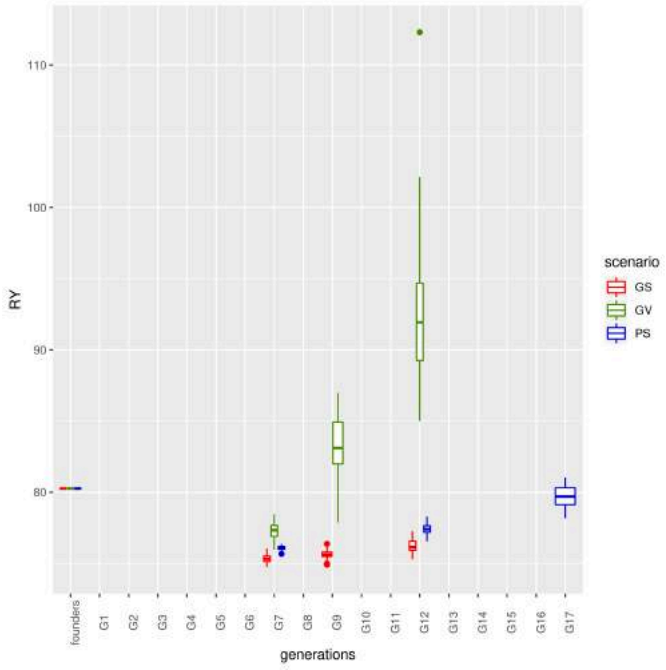


FIGURE 5.5 – Moyenne phénotypique des générations sélectionnées dans un schéma de pre-breeding 30 → 300 sans suivi du lignage maternel par sélection génomique (GS), sélection sur la vraie valeur génétique (GV) ou sélection phénotypique (PS)

generation	GS	GV	PS
founders	80.26	80.26	80.26
G7	75.34	77.26	76.09
G9	75.63	82.80	-
G12	76.24	92.98	77.41
G17	-	-	79.71

TABLE 5.3 – Moyenne phénotypique des générations sélectionnées dans un schéma de pre-breeding 30 → 300 sans suivi du lignage maternel par sélection génomique (GS), sélection sur la vraie valeur génétique (GV) ou sélection phénotypique (PS)

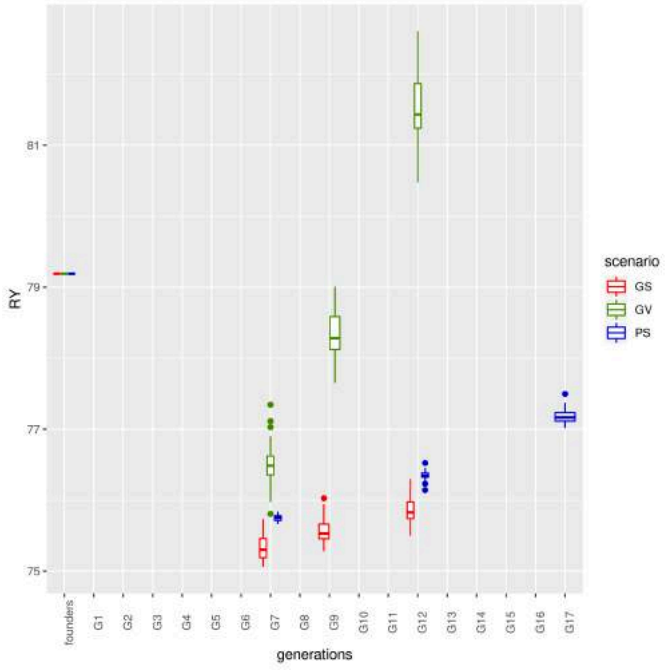


FIGURE 5.6 – Moyenne phénotypique des générations sélectionnées dans un schéma de pre-breeding 300 → 300 sans suivi du lignage maternel par sélection génomique (GS), sélection sur la vraie valeur génétique (GV) ou sélection phénotypique (PS)

generation	GS	GV	PS
founders	79.19	79.19	79.19
G7	75.33	76.51	75.75
G9	75.57	78.32	-
G12	75.85	81.48	76.35
G17	-	-	77.18

TABLE 5.4 – Moyenne phénotypique des générations sélectionnées dans un schéma de pre-breeding 300 → 300 sans suivi du lignage maternel par sélection génomique (GS), sélection sur la vraie valeur génétique (GV) ou sélection phénotypique (PS)

La première génération, appelée « *founders* », représente les fondateurs : le meilleur individu de chacune des 13 descendances AKER étudiées, et les meilleurs individus toutes descendances confondues pour atteindre le nombre de fondateur requis par le scénario. La génération *founders* est composée des mêmes individus quelle que soit la méthode de sélection utilisée par la suite du schéma, la moyenne phénotypique de cette génération est donc la même pour GS, PS et GV.

Dans les deux scénarios débutant avec 30 fondateurs, la moyenne phénotypique des fondateurs est de 80.26 t/ha, alors qu'elle est seulement de 79.19 t/ha dans le scénario débutant à 300 fondateurs. Cette performance plus faible peut être aisément expliquée : les populations composées de 300 fondateurs comprennent 270 individus moins performants que les populations seulement composées de 30 fondateurs, ce qui fait chuter la moyenne phénotypique de cette population.

La moyenne phénotypique des 30 simulations par scénario est ensuite représentée pour les trois générations issues d'une action de sélection, avec une pression de sélection de 20% en G7 et G9 pour les méthodes GS et GV ainsi qu'en G7 et G12 pour la méthode PS, et une pression de sélection de 5% en G12 pour les méthodes GS et GV ainsi qu'en G17 pour la méthode PS.

Une augmentation progressive de la moyenne phénotypique est observée entre les générations G7 et G12 quelle que soit la méthode de sélection utilisée. Cela indique que les actions de GS, PS et GV permettent bien de retenir les individus favorisant une augmentation du gain génétique au cours du schéma de pre-breeding.

Le gain génétique est défini comme étant la performance de la population par unité de temps. Il faut donc comparer les performances des populations obtenues avec les différentes méthodes de sélection à une même génération. La génération G12 représente la fin des schémas de GS et de GV, c'est à cette génération que les performances des différents scénarios sont comparés.
La moyenne phénotypique de la population sélectionnée par la GS en G12 est inférieure à celle issue de la PS à la même génération et ce quelques soient les effectifs des fondateurs et de la population de pre-breeding finale : les Tableaux 5.2, 5.3 et 5.4 indiquent une moyenne de 75.86 t/ha en GS contre 77.53 t/ha en PS pour le scénario 30 $\rightarrow$ 30, de 76.24 t/ha contre 77.41 t/ha pour le scénario 30 $\rightarrow$ 300, et de 75.30 t/ha contre 76.35 t/ha pour le scénario 300 $\rightarrow$ 300. Le gain génétique est donc plus élevé en utilisant la sélection phénotypique plutôt que la sélection génomique. Le rendement racinaire a une héritabilité élevée, établie à 0.60 dans les simulations. D'après la littérature, la sélection génomique présente un avantage vis-à-vis de la sélection phénotypique principalement pour la sélection de caractères peu héritables (Rajsic et al. 2016), ce qui n'est pas le cas dans cette étude.
Pourtant, la comparaison des performances obtenues avec la GS et la GV fait apparaître que la sélection génomique aurait pu être bien meilleure que la sélection phénotypique. En effet, les valeurs de la performance obtenues avec la GV indiquent la performance théorique optimale qu'il aurait été possible d'atteindre avec la GS, soit 93.87 t/ha contre 77.53 t/ha avec la PS dans le scénario 30 $\rightarrow$ 30, 92.98 t/ha contre 77.41 t/ha avec la PS dans le scénario 30 $\rightarrow$ 300, et 81.48 t/ha contre 76.35 t/ha avec la PS dans le scénario 300 $\rightarrow$ 300. Les modèles de prédiction utilisés lors de la GS peuvent donc être améliorés. La première étape de GS s'appuie sur un modèle appris sur la première population d'entraînement, la seconde est basée sur un modèle appris à partir des deux premières populations d'entraînement et la dernière étape est réalisée à partir d'un modèle appris sur l'ensemble des trois populations d'entraînement. L'effectif de ces populations d'entraînement, donné dans le Tableau 5.5, est probablement trop faible pour établir une équation de prédiction optimale, ce qui peut expliquer une partie de la perte de performance observée entre la GS et la GV.

Etape de GS	Population d'entraînement	Effectif de la population d'entrainement 30 → 30	Effectif de la population d'entrainement 30 → 300	Effectif de la population d'entrainement 300 → 300
1	G6_P	30	30	300
2	G6_P + G7_P	60	330	600
3	G6_P + G7_P + G11_P	90	630	900

TABLE 5.5 – Nombre d'individus dans la population d'entrainement utilisée à chacune des étapes de la GS en fonction du scénario simulé

Trois nouveau scénarios de GS sont alors simulés pour vérifier cette hypothèse. Les effectifs de fondateurs et d'individus dans la population de pre-breeding finale sont inchangées, mais l'effectif des populations d'entraînement est augmenté de façon à correspondre au nombre d'individus phénotypés dans les schémas de sélection phénotypiques. Ces effectifs sont donnés dans le Tableau 5.6.

Etape de GS	Population d'entraînement	Effectif de la population d'entrainement augmentée 30 → 30	Effectif de la population d'entrainement augmentée 30 → 300	Effectif de la population d'entrainement augmentée 300 → 300
1	G6_P	120	120	1200
2	G6_P + G7_P	240	1320	2400
3	G6_P + G7_P + G11_P	360	2520	3600

TABLE 5.6 – Nombre d'individus dans la population d'entrainement augmentée utilisée à chacune des étapes de la GS en fonction du scénario simulé

L'évolution de la performance au cours de ces nouvelles simulations est donnée dans les Figures 5.7, 5.8 et 5.9, et la valeur phénotypique moyenne est donnée dans les tables 5.7, 5.8 et 5.9.

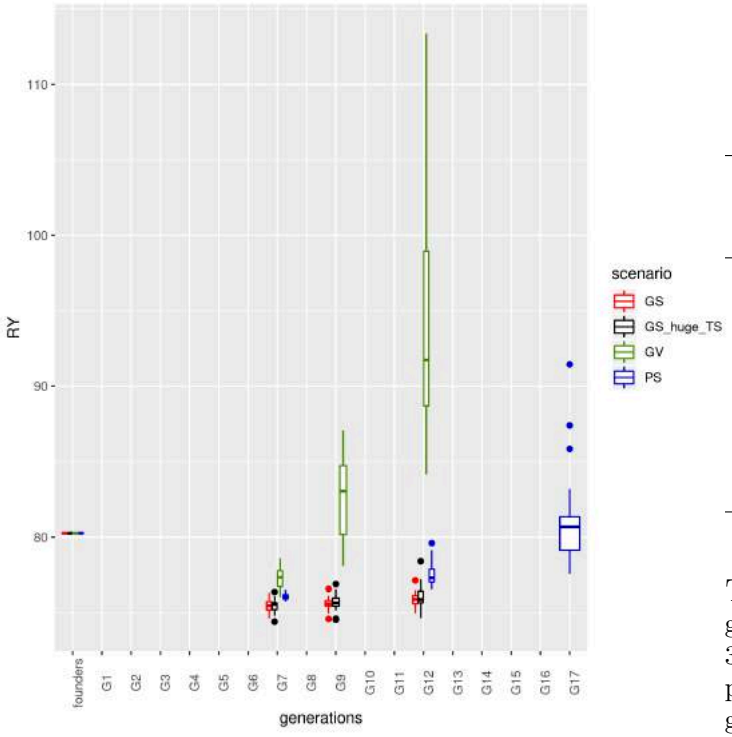


FIGURE 5.7 – Moyenne phénotypique des générations sélectionnées dans un schéma 30 → 30 sans suivi du lignage maternel par sélection génomique (GS), sélection génomique avec de grandes populations d'entraînement (GS_huge_TS), sélection sur la vraie valeur génétique (GV) ou sélection phénotypique (PS)

generation	GS	GV	PS	GS huge_TS
founders	80.26	80.26	80.26	80.26
G7	75.44	77.31	76.06	75.46
G9	75.58	82.66	-	75.69
G12	75.86	93.87	77.53	76.07
G17	-	-	80.95	-

TABLE 5.7 – Moyenne phénotypique des générations sélectionnées dans un schéma 30 → 30 sans suivi du lignage maternel par sélection génomique (GS), sélection génomique avec de grandes populations d'entraînement (GS_huge_TS), sélection sur la vraie valeur génétique (GV) ou sélection phénotypique (PS)

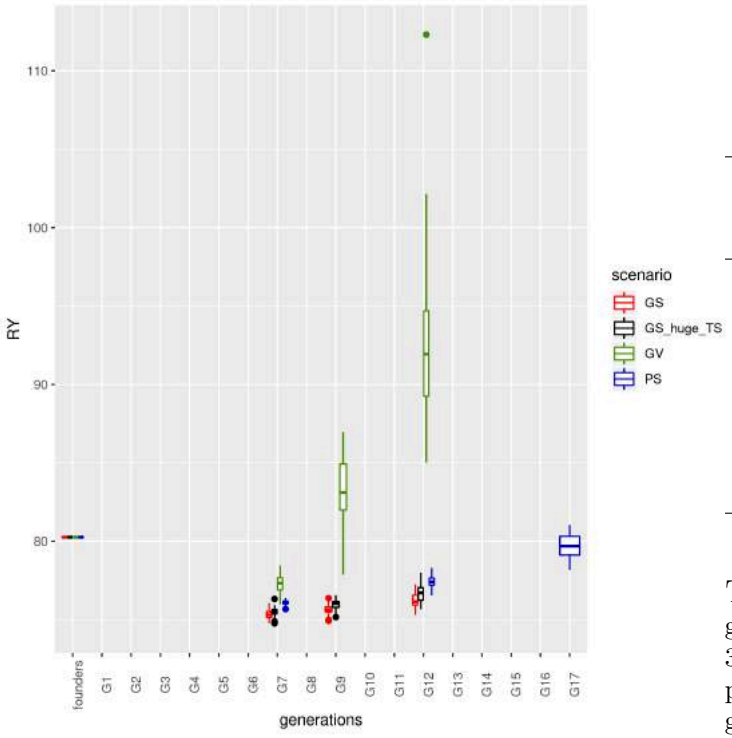


FIGURE 5.8 – Moyenne phénotypique des générations sélectionnées dans un schéma 30 → 300 sans suivi du lignage maternel par sélection génomique (GS), sélection génomique avec de grandes populations d'entraînement (GS_huge_TS), sélection sur la vraie valeur génétique (GV) ou sélection phénotypique (PS)

generation	GS	GV	PS	GS huge_TS
founders	80.26	80.26	80.26	80.26
G7	75.34	77.26	76.09	75.50
G9	75.63	82.80	-	75.94
G12	76.24	92.98	77.41	76.68
G17	-	-	79.71	-

TABLE 5.8 – Moyenne phénotypique des générations sélectionnées dans un schéma 30 → 300 sans suivi du lignage maternel par sélection génomique (GS), sélection génomique avec de grandes populations d'entraînement (GS_huge_TS), sélection sur la vraie valeur génétique (GV) ou sélection phénotypique (PS)

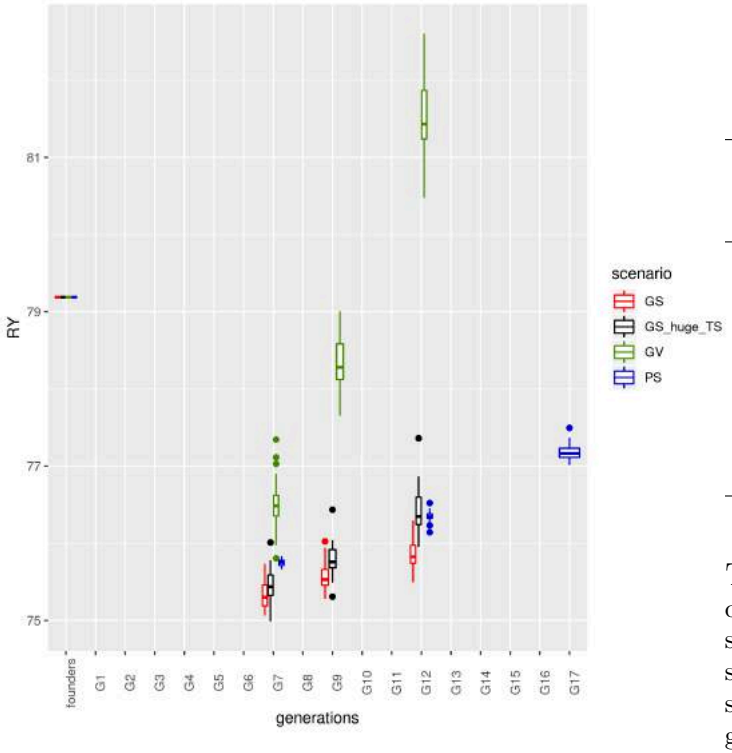


FIGURE 5.9 – Moyenne phénotypique des générations sélectionnées dans un schéma de pre-breeding 300 → 300 sans suivi du lignage maternel par sélection génomique (GS), sélection génomique avec de grandes populations d'entraînement (GS_huge_TS), sélection sur la vraie valeur génétique (GV) ou sélection phénotypique (PS)

generation	GS	GV	PS	GS huge_TS
founders	79.19	79.19	79.19	79.19
G7	75.33	76.51	75.75	75.45
G9	75.57	78.32	-	75.78
G12	75.85	81.48	76.35	76.41
G17	-	-	77.18	-

TABLE 5.9 – Moyenne phénotypique des générations sélectionnées dans un schéma de pre-breeding 300 → 300 sans suivi du lignage maternel par sélection génomique (GS), sélection génomique avec de grandes populations d'entraînement (GS_huge_TS), sélection sur la vraie valeur génétique (GV) ou sélection phénotypique (PS)

L'augmentation de l'effectif des populations d'entraînement a entraîné une amélioration de la performance des populations obtenues par sélection génomique dans tous les scénarios, amélioration d'autant plus importante que le nombre de fondateur et le nombre d'individus composant la population de pre-breeding sont élevés (amélioration en espérance en G12 de 0.21 t/ha dans le scénario 30 → 30, de 0.44 t/ha dans le scénario 30 → 300, et de 0.56 t/ha dans le scénario 300 → 300). Augmenter la taille des populations d'entraînement a donc permis d'améliorer l'estimation de l'effet des marqueurs lors de l'apprentissage du modèle de prédiction.

Malgré cette amélioration, la sélection génomique ne montre toujours pas d'avantage vis-à-vis de la sélection phénotypique. En effet, la sélection phénotypique reste plus performante que la sélection génomique dans les scénarios débutant avec 30 fondateurs, et dans le scénario 300 → 300 les deux méthodes semblent équivalentes en G12 (76.41 t/ha pour la GS contre 76.35 t/ha pour la PS). Le modèle de sélection génomique doit donc encore être amélioré.

Plusieurs pistes, qui restent à explorer, sont ici proposées pour complexifier ce modèle.

Certains QTLs ne sont issus que d'un unique fondateur, les allèles exotiques à ces QTLs sont donc rares. Pondérer plus fortement ces allèles rares pour les favoriser, en utilisant

par exemple la kinship de VanRaden type 3 (centrée et réduite par colonne (VanRaden 2008)) aurait pu améliorer le modèle. Cependant l'utilisation de différentes matrices de kinship ne semble pas avoir un impact significatif sur la précision de la prédiction (Rincent et al. 2012).

Certains QTLs présentent des effets antagonistes importants en fonction de la descendance où ils ont été détectés (chapitre 3). Dans la sélection phénotypique, ce n'est pas l'estimation de ces effets qui a une importance puisque la sélection se fait sur la valeur des parents, cette valeur étant bien estimée par la moyenne phénotypique des hybrides descendant du couple des parents. Au contraire, même avec une taille de population d'entraînement infinie, le modèle de sélection génomique est faux pour l'estimation des effets antagonistes de ces QTLs, puisqu'une une valeur unique est attribuée à un QTL. La valeur prédite des individus au QTL est donc identique, que ces individus portent le bon ou le mauvais allèle. Or ces QTLs antagonistes font partie des QTLs ayant des effets très importants, leur mauvaise estimation a donc un impact majeur pour la prédiction de la valeur génétique des individus. Le modèle de sélection génomique pourrait être complexifié pour capturer les effets antagonistes des QTLs, en passant dans un modèle multiallélique. Ce modèle multiallélique permettrait de distinguer les individus porteurs de l'allèle favorable ou défavorable au QTL transmis par leur mère. En effet, il est possible de suivre cet allèle grâce à la récolte des semences sur les individus mâles stériles. La voie mâle, qui ne peut être connue du fait des pollinisation aléatoires, apporterait toutefois du bruit dans ce modèle. Cependant la fréquence des allèles exotiques à un QTL étant faible, suivre une partie de la transmission, ici la voie femelle, devrait permettre de bien améliorer la prédiction de la valeur génétique des individus à sélectionner.

5.2.1.2 Impact de l'effectif des générations

L'impact de l'effectif des générations sur le gain génétique est étudié dans les schémas intégrant la sélection génomique. L'évolution du gain génétique dans chaque génération issue d'une action de sélection des scénarios $30 \rightarrow 30$, $30 \rightarrow 300$ et $300 \rightarrow 300$ est présentée dans la Figure 5.10. La moyenne de ce gain génétique est donné dans le Tableau 5.10.

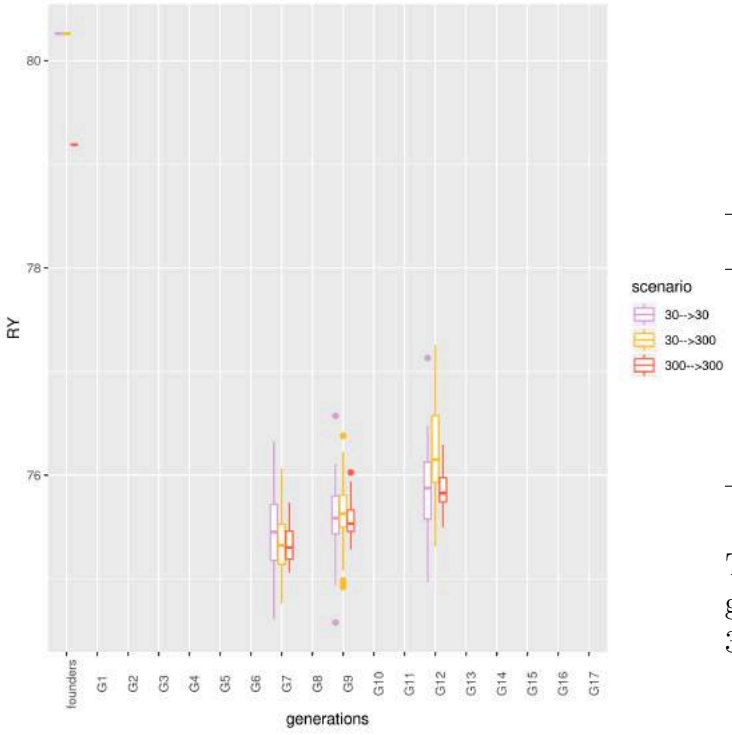


FIGURE 5.10 – Evolution du gain génétique dans trois scénarios, 30 → 30 en rose ; 30 → 300 en jaune, 300 → 300 en rouge

	30 → 30	30 → 300	300 → 300
founders	80.26	80.26	79.19
G7	75.44	75.34	75.33
G9	75.58	75.63	75.56
G12	75.86	76.24	75.85

TABLE 5.10 – Evolution du gain génétique dans trois scénarios, 30 → 30, 30 → 300 et 300 → 300

Dans les deux scénarios débutant avec 30 fondateurs, la moyenne phénotypique des fondateurs est de 80.26 t/ha, alors qu'elle est seulement de 79.19 t/ha dans le scénario débutant à 300 fondateurs. Comme expliqué précédemment, cette performance plus faible est expliquée par le fait que les populations composées de 300 fondateurs comprennent 270 individus moins performants que les populations seulement composées de 30 fondateurs, qui font ainsi chuter la moyenne phénotypique de cette population.

En G12, la performance de la population est plus faible que celle à la génération des fondateurs quel que soit le scénario. Un léger avantage du scénario 30 → 300 est observé. La moyenne phénotypique de la population atteint en espérance 76.24 t/ha, contre 75.86 t/ha dans le scénario 30 → 30 et 75.85 t/ha dans le scénario 300 → 300. Le scénario 30 → 300 semble donc avantager légèrement la performance de la population de pre-breeding générée.

5.2.2 Etude de l'évolution de la diversité de combinaison allélique au cours du schéma de pre-breeding selon le suivi du lignage maternel ou non et les effectifs des populations

Le schéma de pre-breeding optimal doit éviter l'érosion de la diversité de combinaison allélique. L'impact du suivi du lignage maternel ainsi que de l'effectif de la population sur l'évolution de la diversité de combinaison allélique est étudié. Chacun des cinq indicateurs

définis dans la partie 5.1 est calculé pour chaque génération issue d'une action de sélection dans les schémas intégrant la sélection génomique : la génération des fondateurs, la génération G7_GS issue de la première sélection génomique avec une intensité de sélection de 20%, la génération G9_GS issue de la deuxième sélection génomique ayant également une intensité de sélection de 20%, et la génération G12_GS issue de la dernière sélection génomique avec une intensité de sélection de 5% et représentant la population de pre-breeding obtenue à la fin du schéma de pre-breeding. Pour rappel, la génération des fondateurs comporte le meilleur individu issu de chacune des 13 descendances AKER, auxquels sont ajoutés les meilleurs individus toutes descendances confondues pour obtenir l'effectif de fondateurs donné par le scénario. Ce sont ces fondateurs qui sont ensuite suivis ou non par lignage maternel : à la génération des fondateurs aucune différence ne peut être faite entre les scénarios suivant ou non le lignage maternel.

5.2.2.1 Impact du suivi du lignage maternel

L'impact du suivi du lignage maternel est tout d'abord étudié : il faut pour cela comparer l'évolution des indicateurs selon le fait que le lignage maternel soit suivi ou non, et ce pour les trois différents effectifs de génération possibles. Chacun de ces six scénarios est simulé 30 fois. Deux estimateurs de l'espérance de chaque indicateur sont calculés pour chaque scénario : la moyenne et la médiane.

Fréquence de l'allèle exotique aux locus « favorables »

L'évolution de la fréquence de l'allèle exotique aux locus « favorables » avec et sans suivi du lignage maternel est présentée dans les Figures 5.13, 5.12 et 5.11, qui représentent respectivement les simulations de 30 fondateurs et 30 individus dans la population de pre-breeding finale ($30 \rightarrow 30$), 30 fondateurs et 300 individus dans la population de pre-breeding finale ($30 \rightarrow 300$), et 300 fondateurs et 300 individus dans la population de pre-breeding finale ($300 \rightarrow 300$). La moyenne de cet indicateur dans chacune des générations issues d'une action de sélection est donnée dans les Tableaux 5.11, 5.12 et 5.13.

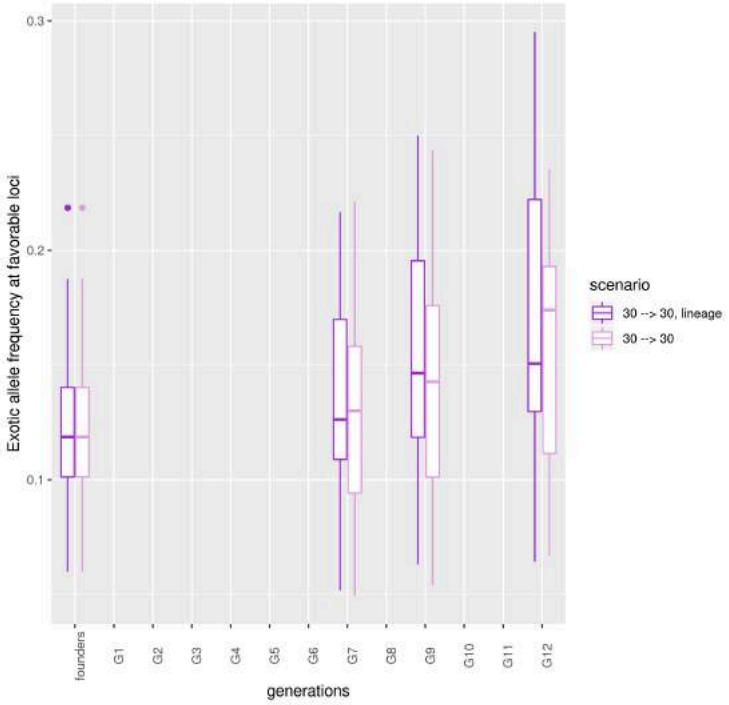


FIGURE 5.11 – Evolution de la fréquence de l'allèle exotique aux locus « favorables » dans deux scénarios, 30 → 30 avec suivi du lignage maternel en violet ; 30 → 30 sans suivi du lignage maternel en rose

	30 → 30 lineage	30 → 30
founders	0.12	0.12
G7	0.14	0.13
G9	0.15	0.14
G12	0.17	0.16

TABLE 5.11 – Evolution de la fréquence de l'allèle exotique aux locus « favorables » dans deux scénarios, 30 → 30 avec ou sans suivi du lignage maternel

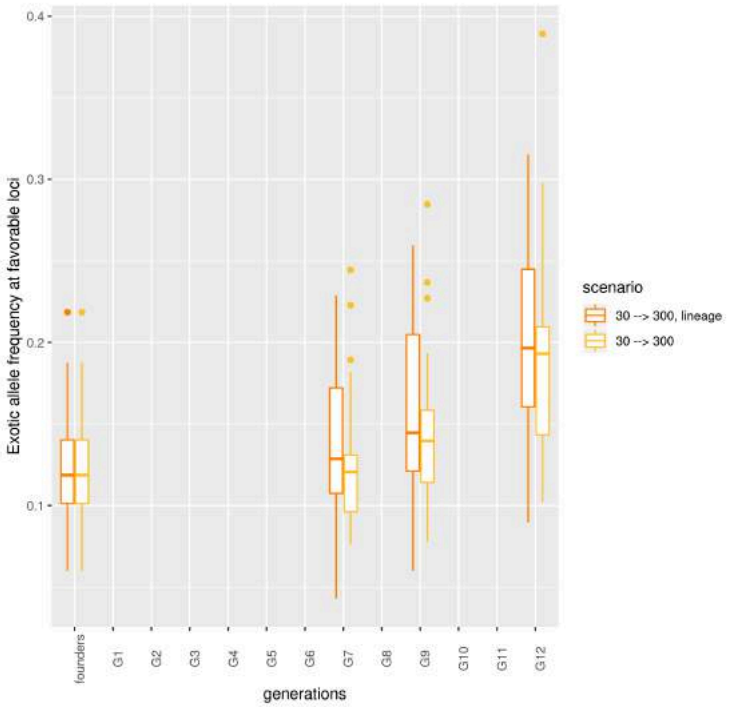


FIGURE 5.12 – Evolution de la fréquence de l'allèle exotique aux locus « favorables » dans deux scénarios, 30 → 300 avec suivi du lignage maternel en orange ; 30 → 300 sans suivi du lignage maternel en jaune

	30 → 300 lineage	30 → 300
founders	0.12	0.12
G7	0.14	0.13
G9	0.16	0.14
G12	0.20	0.19

TABLE 5.12 – Evolution de la fréquence de l'allèle exotique aux locus « favorables » dans deux scénarios, 30 → 300 avec ou sans suivi du lignage maternel

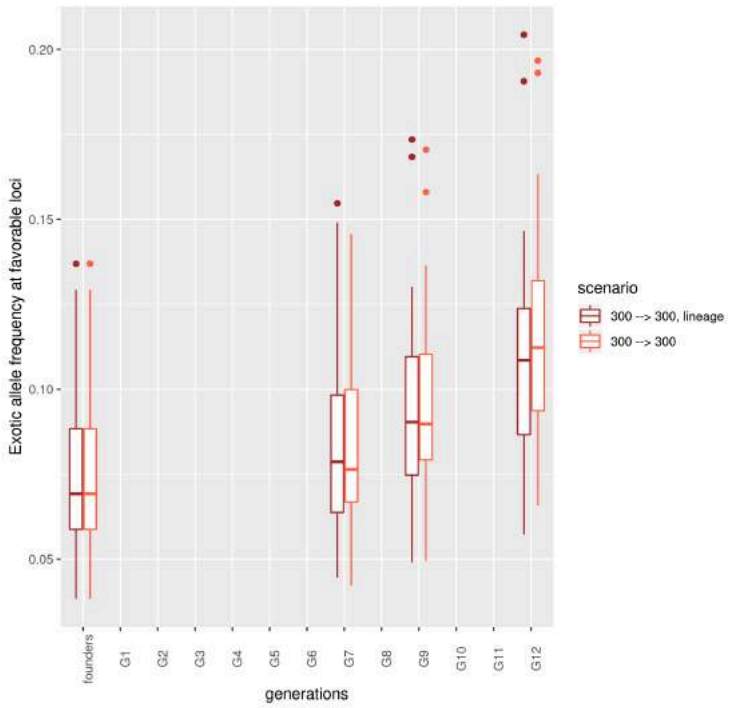


FIGURE 5.13 – Evolution de la fréquence de l'allèle exotique aux locus « favorables » dans deux scénarios, 300 → 300 avec suivi du lignage maternel en marron ; 300 → 300 sans suivi du lignage maternel en rouge

	300 → 300 lineage	300 → 300
founders	0.08	0.08
G7	0.08	0.08
G9	0.10	0.10
G12	0.11	0.12

TABLE 5.13 – Evolution de la fréquence de l'allèle exotique aux locus « favorables » dans deux scénarios, 300 → 300 avec ou sans suivi du lignage maternel

En espérance la fréquence de la présence de l'allèle exotique aux locus « favorables » augmente au cours des simulations et ce dans tous les scénarios, passant en moyenne de 12 à 18% entre la génération des fondateurs et la génération 12 dans les scénarios débutant avec 30 fondateurs, et de 8 à 12% dans les scénarios commençant avec 300 fondateurs. Cette augmentation indique que la sélection génomique permet bien de choisir des individus portant des allèles exotiques aux locus « favorables ».

L'impact du suivi du lignage maternel est ensuite étudié à la génération G12, génération constituant la population de pre-breeding en sortie du schéma de sélection. Aucun impact significatif du suivi du lignage maternel n'est observée dans les 3 scénarios. **Le suivi du lignage maternel n'a pas d'impact significatif sur la fréquence d'allèles exotiques aux locus « favorables ».**

Fréquence de l'allèle exotique pour l'ensemble des locus

L'évolution de la fréquence de l'allèle exotique pour l'ensemble des locus avec et sans suivi du lignage maternel dans chacune des générations issues d'une action de sélection est présentée dans les Figures 5.14, 5.15 et 5.16, qui représentent respectivement les scénarios intégrant la sélection génomique avec les effectifs 30 → 30, 30 → 300 et 300 → 300. Pour rappel, chaque scénario est simulé 30 fois. La moyenne de cet indicateur est donnée dans les Tableaux 5.14, 5.15 et 5.16.

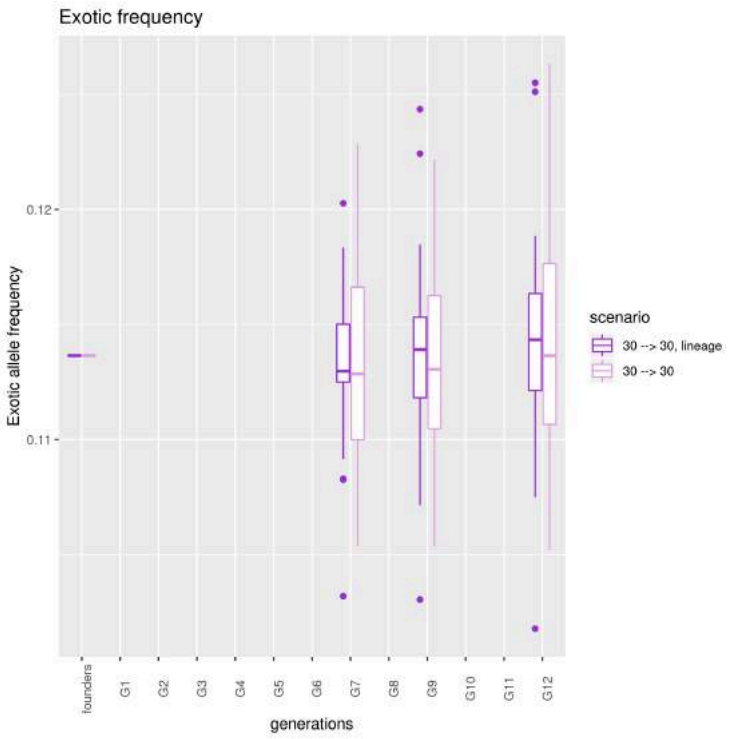


FIGURE 5.14 – Evolution de la fréquence de l'allèle exotique lors de deux scénarios, 30 → 30, avec suivi du lignage maternel en violet ; 30 → 30, sans suivi du lignage maternel en rose

	30 → 30 lineage	30 → 30
founders	0.11	0.11
G7	0.11	0.11
G9	0.11	0.11
G12	0.11	0.11

TABLE 5.14 – Evolution de la fréquence de l'allèle exotique lors de deux scénarios, 30 → 30, avec ou sans suivi du lignage maternel

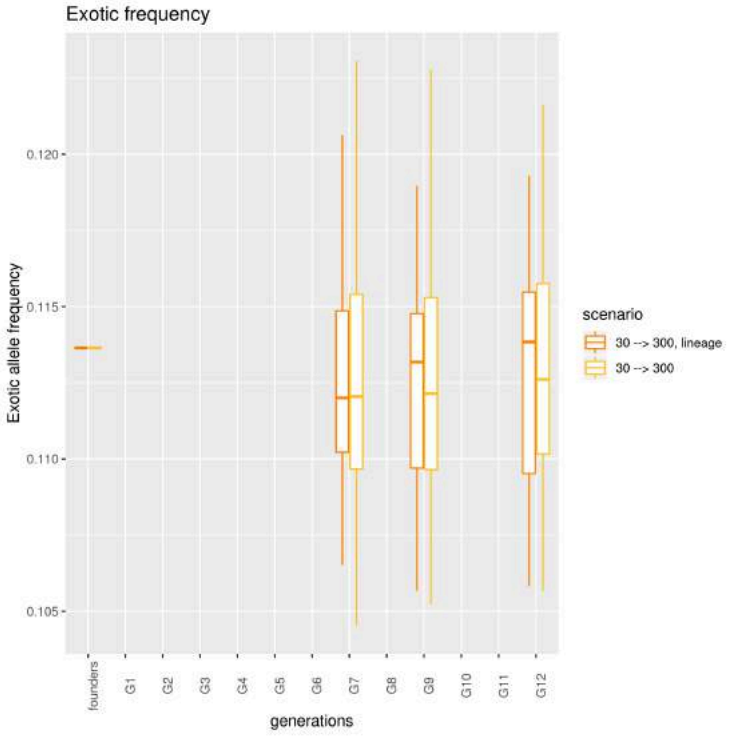


FIGURE 5.15 – Evolution de la fréquence de l'allèle exotique lors de deux scénarios, 30 → 300 individus avec suivi du lignage maternel en orange ; 30 → 300 sans suivi du lignage maternel en jaune

	30 → 300 lineage	30 → 300
founders	0.11	0.11
G7	0.11	0.11
G9	0.11	0.11
G12	0.11	0.11

TABLE 5.15 – Evolution de la fréquence de l'allèle exotique lors de deux scénarios, 30 → 300 individus dans la populations de pre-breeding finale, avec ou sans suivi du lignage maternel

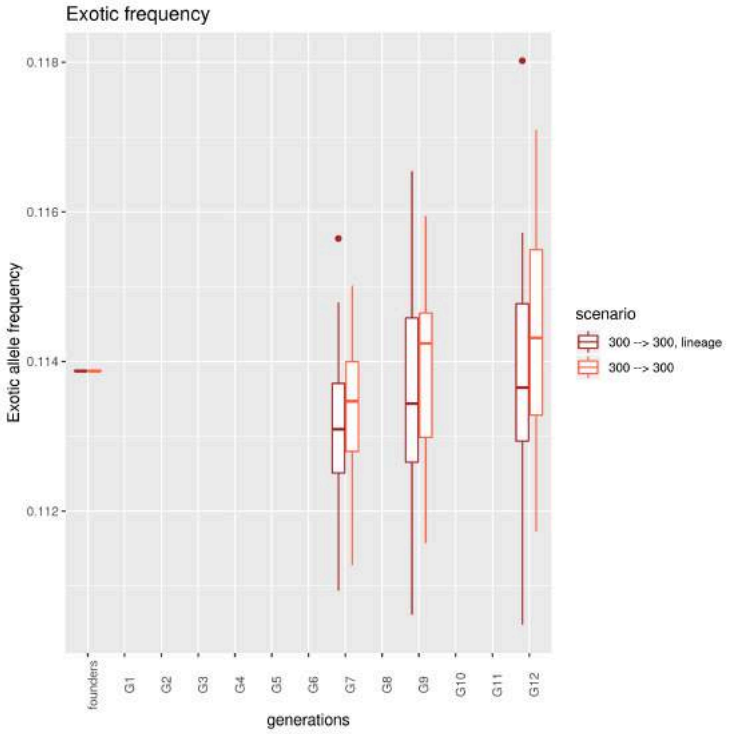


FIGURE 5.16 – Evolution de la fréquence de l'allèle exotique lors de deux scénarios, 300 → 300 avec suivi du lignage maternel en marron ; 300 → 300 sans suivi du lignage maternel en rouge

	300 → 300 lineage	300 → 300
founders	0.11	0.11
G7	0.11	0.11
G9	0.11	0.11
G12	0.11	0.11

TABLE 5.16 – Evolution de la fréquence de l'allèle exotique lors de deux scénarios, 300 fondateurs, 300 individus dans la populations de pre-breeding finale, avec ou sans suivi du lignage maternel

La fréquence de l'allèle exotique au sein de la population est stable au cours des simulations. En effet elle ne semble varier ni en fonction des générations ni en fonction du suivi du lignage maternel, présentant une moyenne de 11% pour chacune des générations produites par les différents scénarios. **Le suivi du lignage maternel n'a pas d'impact significatif sur la fréquence d'allèles exotiques sur l'ensemble des locus.**

Longueur moyenne des fragments exotiques

L'évolution de la longueur moyenne des fragments exotiques avec et sans suivi du lignage maternel dans chacune des générations issues d'une action de sélection est présentée dans les Figures 5.17, 5.18 et 5.19, qui représentent respectivement les scénarios intégrant la sélection génomique avec les effectifs 30 → 30, 30 → 300 et 300 → 300. La moyenne de cet indicateur est donnée dans les Tableaux 5.17, 5.18 et 5.19.

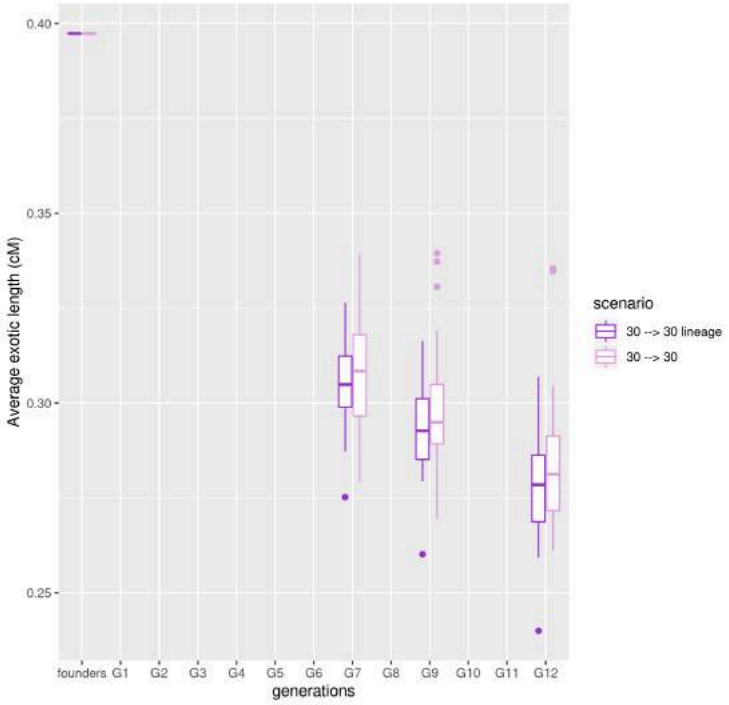


FIGURE 5.17 – Evolution de la longueur moyenne des fragments exotiques lors de deux scénarios, 30 → 30 avec suivi du lignage maternel en violet ; 30 → 30 sans suivi du lignage maternel en rose

	30 → 30 lineage	30 → 30
founders	0.40	0.40
G7	0.30	0.31
G9	0.29	0.30
G12	0.28	0.28

TABLE 5.17 – Evolution de la longueur moyenne des fragments exotiques lors de deux scénarios, 30 → 30 avec ou sans suivi du lignage maternel

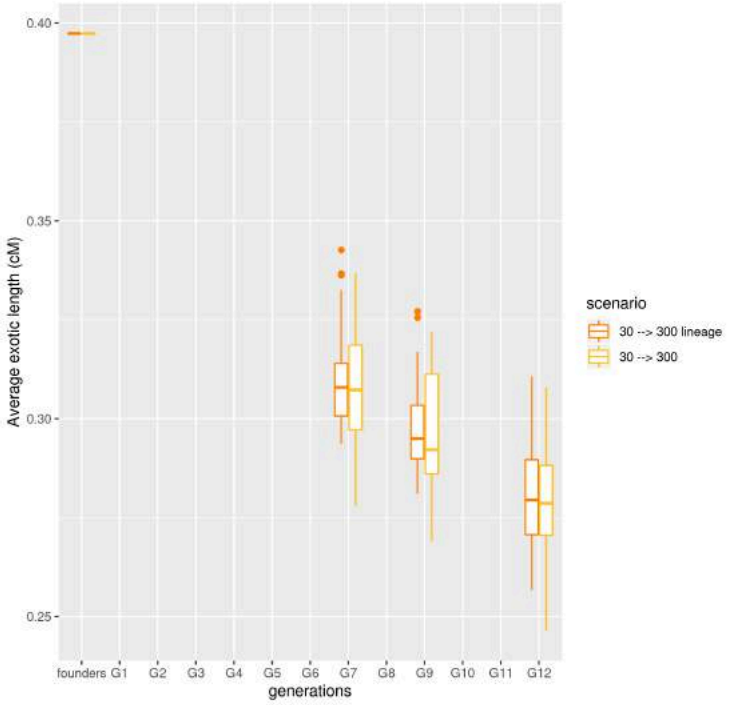


FIGURE 5.18 – Evolution de la longueur moyenne des fragments exotiques lors de deux scénarios, 30 → 300 avec suivi du lignage maternel en orange ; 30 → 300 sans suivi du lignage maternel en jaune

	30 → 300 lineage	30 → 300
founders	0.40	0.40
G7	0.31	0.31
G9	0.30	0.30
G12	0.28	0.28

TABLE 5.18 – Evolution de la longueur moyenne des fragments exotiques lors de deux scénarios, 30 → 300 avec ou sans suivi du lignage maternel

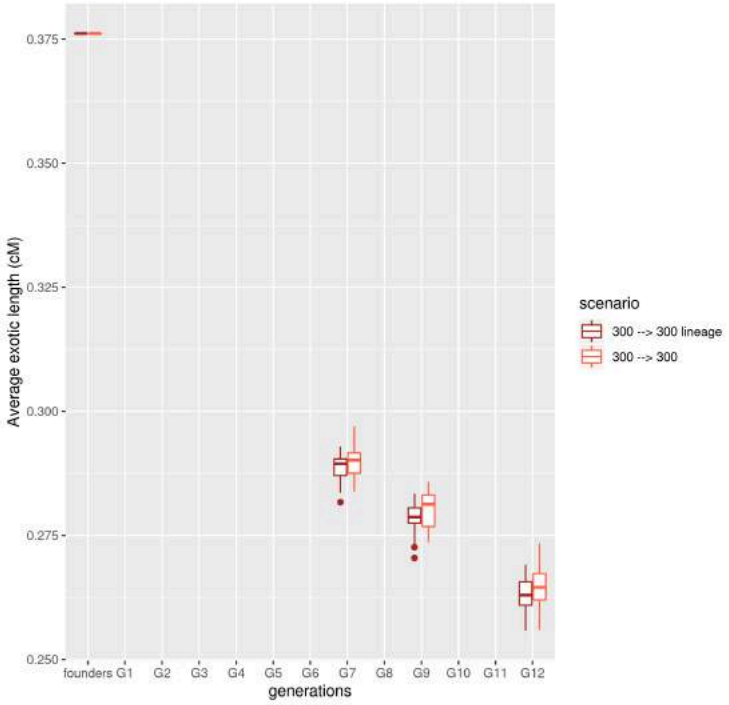


FIGURE 5.19 – Evolution de la longueur moyenne des fragments exotiques lors de deux scénarios, 300 → 300 avec suivi du lignage maternel en marron ; 300 → 300 sans suivi du lignage maternel en rouge

	300 → 300 lineage	300 → 300
founders	0.38	0.38
G7	0.29	0.29
G9	0.28	0.28
G12	0.26	0.26

TABLE 5.19 – Evolution de la longueur moyenne des fragments exotiques lors de deux scénarios, 300 → 300 avec ou sans suivi du lignage maternel

L'évolution de la longueur moyenne des fragments exotiques dans les trois figures indique une bonne recombinaison au cours du schéma de pre-breeding. En effet la valeur de cet indicateur diminue dans tous les scénarios, passant en moyenne de 0.40cM dans la génération des fondateurs à 0.28cM en G12 dans les scénarios débutant avec 30 fondateurs, et de 0.38cM à 0.26cM dans le scénario comportant 300 fondateurs. Il n'y a pas de différence significative entre la longueur moyenne des fragments exotiques selon que l'on suit ou non le lignage maternel à une génération et pour un effectif de population donné. **Aucun impact du suivi du lignage maternel sur le taux de recombinaison n'est mis en évidence**.

Nombre de dimensions significatives de l'ACP

L'évolution du nombre de dimensions significatives de l'analyse en composantes principales rapporté au nombre d'individus composant la génération dans chacune des générations issues d'une action de sélection, avec ou sans suivi du lignage maternel, est représentée dans les Figures 5.20, 5.21 et 5.22, qui représentent respectivement les scénarios intégrant la sélection génomique avec les effectifs 30 → 30, 30 → 300 et 300 → 300. La moyenne de cet indicateur est donnée dans les Tableaux 5.20, 5.21 et 5.22.

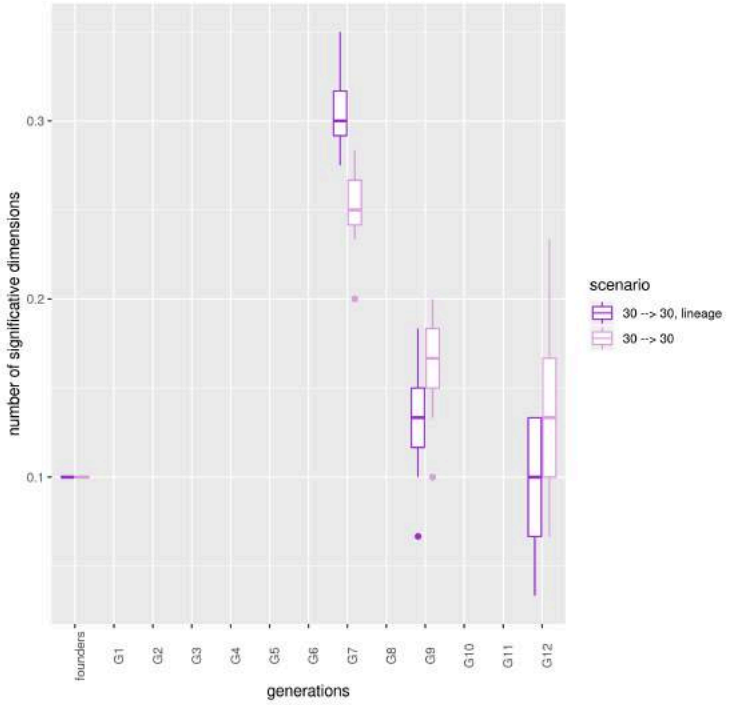


FIGURE 5.20 – Evolution du nombre de dimensions significatives de l'ACP rapporté au nombre d'individus lors de deux scénarios, 30 → 30 avec suivi du lignage maternel en violet ; 30 → 30 sans suivi du lignage maternel en rose

	30 → 30 lineage	30 → 30
founders	0.10	0.10
G7	0.30	0.25
G9	0.14	0.17
G12	0.09	0.15

TABLE 5.20 – Evolution du nombre de dimensions significatives de l'ACP rapporté au nombre d'individus lors de deux scénarios, 30 → 30 individus avec ou sans suivi du lignage maternel

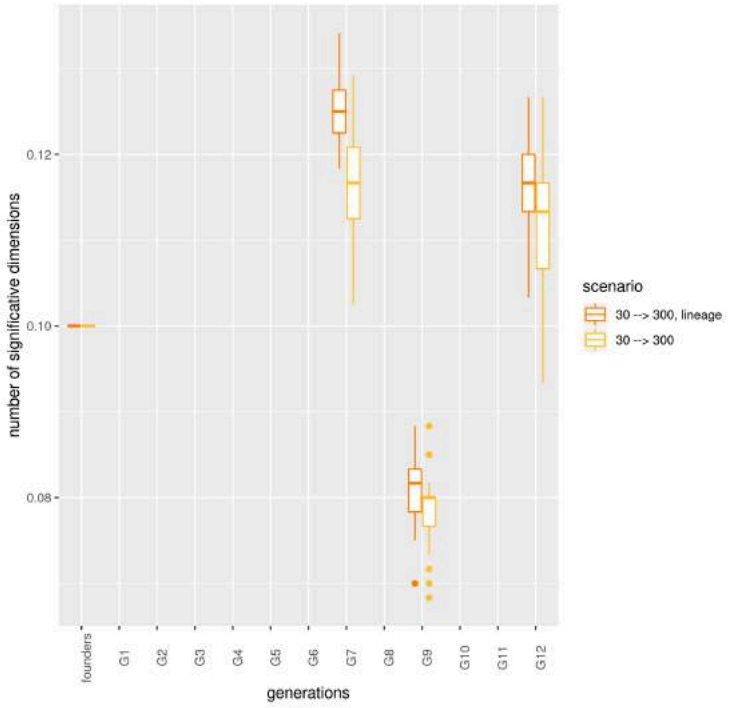


FIGURE 5.21 – Evolution du nombre de dimensions significatives de l'ACP rapporté au nombre d'individus lors de deux scénarios, 30 → 300 avec suivi du lignage maternel en orange ; 30 → 300 sans suivi du lignage maternel en jaune

	30 → 300 lineage	30 → 300
founders	0.10	0.10
G7	0.12	0.12
G9	0.08	0.08
G12	0.12	0.11

TABLE 5.21 – Evolution du nombre de dimensions significatives de l'ACP rapporté au nombre d'individus lors de deux scénarios, 30 → 300 avec ou sans suivi du lignage maternel

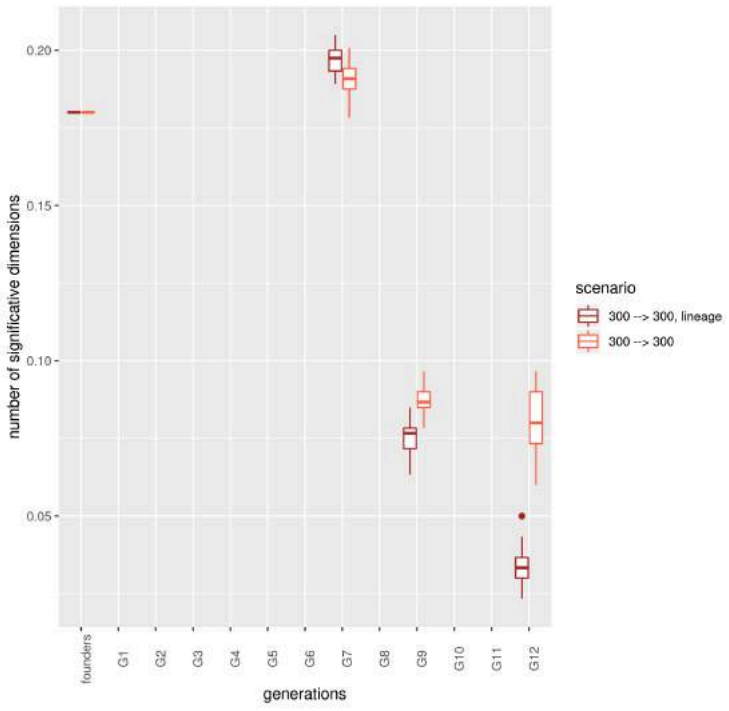


FIGURE 5.22 – Evolution du nombre de dimensions significatives de l'ACP rapporté au nombre d'individus lors de deux scénarios, 300 → 300 avec suivi du lignage maternel en marron ; 300 → 300 sans suivi du lignage maternel en rouge

	300 → 300 lineage	300 → 300
founders	0.18	0.18
G7	0.20	0.19
G9	0.08	0.09
G12	0.03	0.08

TABLE 5.22 – Evolution du nombre de dimensions significatives de l'ACP rapporté au nombre d'individus lors de deux scénarios, 300 fondateurs, 300 individus dans la populations de pre-breeding finale, avec ou sans suivi du lignage maternel

Dans le scénario 300 → 300, l'effectif des générations est suffisamment important pour réfléchir comme en situation de population infinie. Les fondateurs représentent donc la diversité génétique avec le moins de structuration atteignable dans ce scénario et la génération G7, issue de la recombinaison des fondateurs, a approximativement la même diversité (0.18 à la génération des fondateurs, contre 0.19 en G7 sans suivi du lignage maternel et 0.20 avec suivi du lignage maternel). En revanche le scénario 30 → 30 se comporte en population vraiment finie. La recombinaison des fondateurs peut produire des extrêmes qui présentent des profils génétiques très différents. La moyenne de l'indicateur indique ainsi une très forte augmentation entre la génération des fondateurs et la génération G7 dans le scénario 30 → 30 (0.10 à la génération des fondateurs, 0.25 en G7 sans suivi du lignage maternel, 0.30 avec suivi du lignage maternel).

Au cours des générations suivantes la valeur de l'indicateur diminue avec et sans suivi du lignage maternel, jusqu'à atteindre 0.08 sans suivi du lignage maternel et 0.03 avec suivi du lignage maternel en G12 dans le scénario 300 → 300, et 0.09 en G12 avec suivi du lignage maternel et 0.15 sans suivi du lignage maternel dans le scénario 30 → 30. En effet, au fur et à mesure des croisements et des sélections, les individus retenus sont de plus en plus proches au niveau génétique. La population est donc de plus en plus structurée. En

G12 le nombre de dimensions significatives de l'ACP rapporté au nombre d'individus est plus élevé dans le scénario où le lignage n'est pas suivi. Suivre le lignage maternel impose de sélectionner au moins un descendant de chaque fondateur, descendant qui devient alors l'un des parents de la génération suivante et transmet ainsi une partie de son génome à ses descendants. Au fur et à mesure des recombinaisons, les descendants seront issus de l'ensemble des fondateurs : il seront proches du point de vue génétique, ce qui explique la diminution des valeurs de l'indicateur. Ne pas suivre le lignage maternel permet d'avoir plus de liberté dans le choix des individus à chaque action de sélection, ce qui ralentit la diminution des valeurs de l'indicateur.
Le scénario $30 \rightarrow 300$ ne présente pas beaucoup de variation en valeur entre les différentes générations (plus ou moins 0.02 par rapport à la valeur obtenue à la génération des fondateurs). L'indicateur est difficilement interprétable dans ce scénario.
Le nombre de dimensions significatives de l'ACP rapportée au nombre d'individus semble indiquer que la population de pre-breeding obtenue est moins structurée dans les scénarios ne suivant pas le lignage maternel, mais cet indicateur est difficilement interprétable dans le scénario $30 \rightarrow 300$.

Indicateurs de la matrice de Kinship

L'évolution des valeurs minimales, et maximales de la matrice de Kinship des individus de chaque génération issue d'une action de sélection avec et sans suivi du lignage maternel est présentée dans les Figures 5.23, 5.24 et 5.25, qui représentent respectivement les scénarios intégrant la sélection génomique avec les effectifs $30 \rightarrow 30$, $30 \rightarrow 300$ et $300 \rightarrow 300$. La moyenne de ces indicateurs est donnée dans les Tableaux 5.23, 5.24 et 5.25.

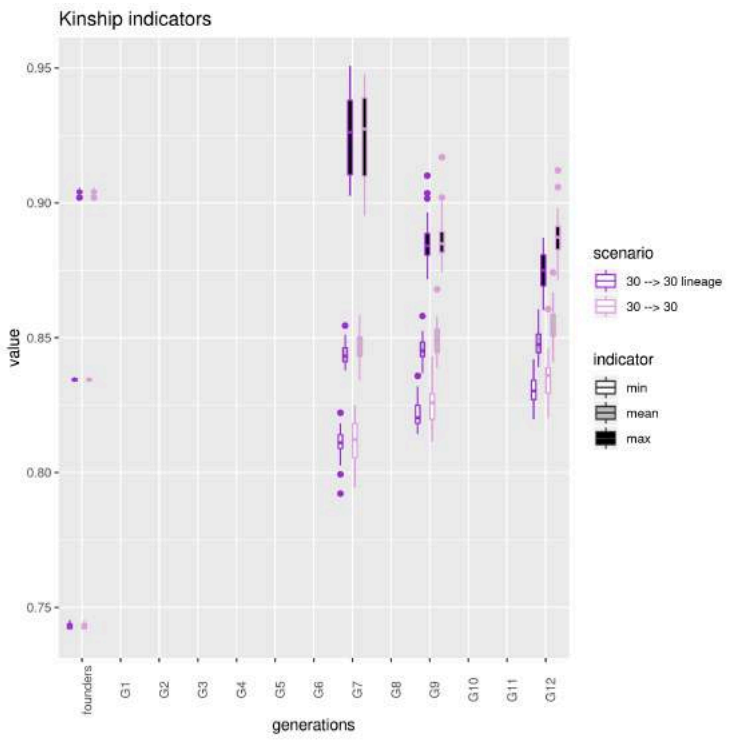


FIGURE 5.23 – Evolution de la valeur minimale (en blanc), moyenne (en gris) et maximale (en noir) de la kinship lors de deux scénarios, 30 → 30 avec suivi du lignage maternel en violet ; 30 → 30 sans suivi du lignage maternel en rose

	min	
	30 → 30 lineage	30 → 30
founders	0.74	0.74
G7	0.81	0.81
G9	0.82	0.83
G12	0.83	0.84
	mean	
	30 → 30 lineage	30 → 30
founders	0.83	0.83
G7	0.84	0.85
G9	0.85	0.85
G12	0.85	0.85
	max	
	30 → 30 lineage	30 → 30
founders	0.90	0.90
G7	0.93	0.92
G9	0.89	0.89
G12	0.87	0.89

TABLE 5.23 – Valeurs minimales, moyennes et maximales de la kinship dans les scénarios : 30 → 30 avec ou sans suivi du lignage maternel

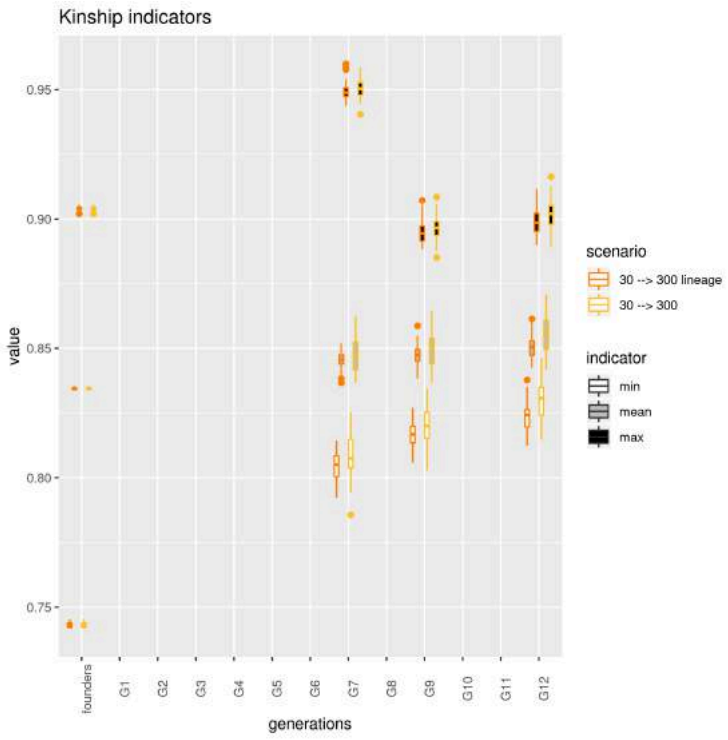


FIGURE 5.24 – Evolution de la valeur minimale (en blanc) et maximale (en noir) de la kinship lors de deux scénarios, 30 → 300 avec suivi du lignage maternel en orange ; 30 → 300 sans suivi du lignage maternel en jaune

	min	
	30 → 300 lineage	30 → 300
founders	0.74	0.74
G7	0.80	0.81
G9	0.82	0.82
G12	0.82	0.83
	mean	
	30 → 300 lineage	30 → 300
founders	0.83	0.83
G7	0.85	0.85
G9	0.85	0.85
G12	0.85	0.86
	max	
	30 → 300 lineage	30 → 300
founders	0.90	0.90
G7	0.95	0.95
G9	0.90	0.90
G12	0.90	0.90

TABLE 5.24 – Valeur minimale et maximale de la kinship dans les scénarios : 30 → 300 avec ou sans suivi du lignage maternel

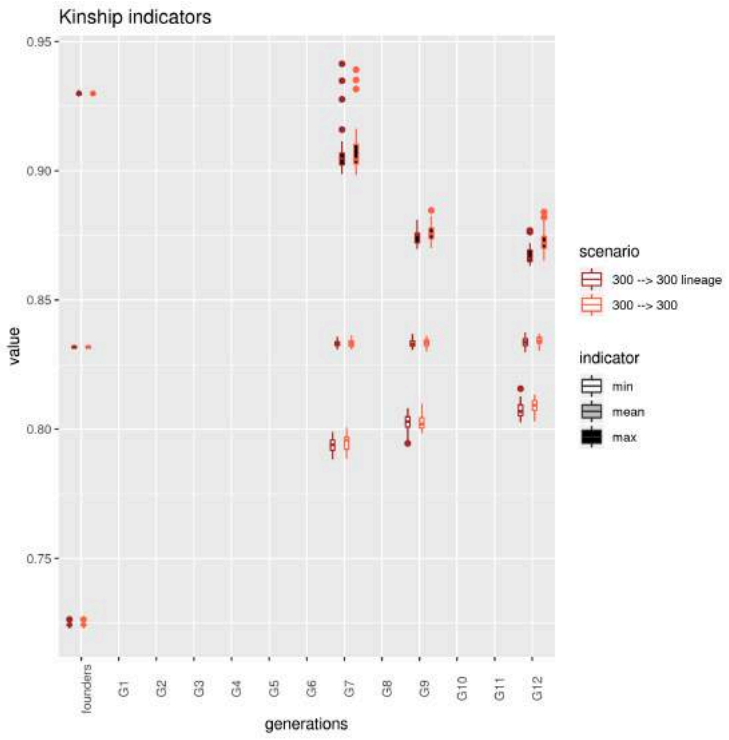


FIGURE 5.25 – Evolution de la valeur minimale (en blanc) et maximale (en noir) de la kinship lors de deux scénarios, 300 → 300 avec suivi du lignage maternel en marron ; 300 → 300 sans suivi du lignage maternel en rouge

	min	
	300 → 300 lineage	300 → 300
founders	0.72	0.72
G7	0.79	0.79
G9	0.80	0.80
G12	0.81	0.81
	mean	
	300 → 300 lineage	300 → 300
founders	0.83	0.83
G7	0.83	0.83
G9	0.83	0.83
G12	0.83	0.83
	max	
	300 → 300 lineage	300 → 300
founders	0.93	0.93
G7	0.91	0.91
G9	0.87	0.88
G12	0.87	0.87

TABLE 5.25 – Valeur minimale et maximale de la kinship dans les scénarios : 300 → 300 avec ou sans suivi du lignage maternel

La matrice de kinship AIS calculée représente la proportion espérée d'allèles identiques entre deux individus. La valeur minimale de cette kinship représente ainsi l'éloignement génétique maximal entre deux individus d'une génération donnée. Cette valeur augmente entre la génération des fondateurs et la génération G12 dans tous les scénarios, ce qui traduit un rapprochement des individus les plus distants au fil des générations. A l'inverse, la valeur maximale de la kinship représente la proximité maximale de deux individus dans une génération donnée. Cette valeur diminue dans tous les scénarios, indiquant que les individus les plus proches s'éloignent. La moyenne quant à elle reste stable au cours des simulations, ce qui signifie qu'aucun des scénarios ne conduit à créer un goulot d'étranglement génétique.

En espérance, le suivi du lignage maternel n'a pas d'impact sur le différentes valeurs de la kinship calculées à la génération G12. En effet ces valeurs sont strictements identiques

ou très proches que le lignage maternel soit suivi ou non, et ce dans tous les scénarios. **Le suivi du lignage maternel ne permet pas de favoriser la diversité génétique de la population de pre-breeding produite.**

5.2.2.2 Impact de l'effectif des générations

L'impact de l'effectif des générations sur la diversité de combinaison allélique est à présent étudié dans les scénarios utilisant la sélection génomique. Il faut pour cela comparer l'évolution des cinq indicateurs selon le fait de démarrer le schéma avec 30 ou 300 fondateurs, et que l'on veuille 30 ou 300 individus dans la population de pre-breeding finale. L'étude de l'impact du lignage a démontré que la diversité génétique n'était pas impactée par le suivi du lignage maternel. L'impact de l'effectif des générations est donc étudié dans les scénarios sans suivi du lignage maternel (les résultats des scénarios avec suivi du lignage maternel sont toutefois donnés en annexes).

Fréquence de l'allèle exotique aux locus « favorables »

L'évolution de la fréquence de l'allèle exotique aux locus « favorables » selon les effectifs $30 \rightarrow 30$, $30 \rightarrow 300$ et $300 \rightarrow 300$ sans suivi du lignage maternel est présentée dans la Figure 5.26. La moyenne de cet indicateur dans chacune des générations issues d'une action de sélection est donnée dans le Tableau 5.26. Les résultats de cette même étude avec suivi du lignage sont donnés en annexes (Figure A.10, Tableau A.5).

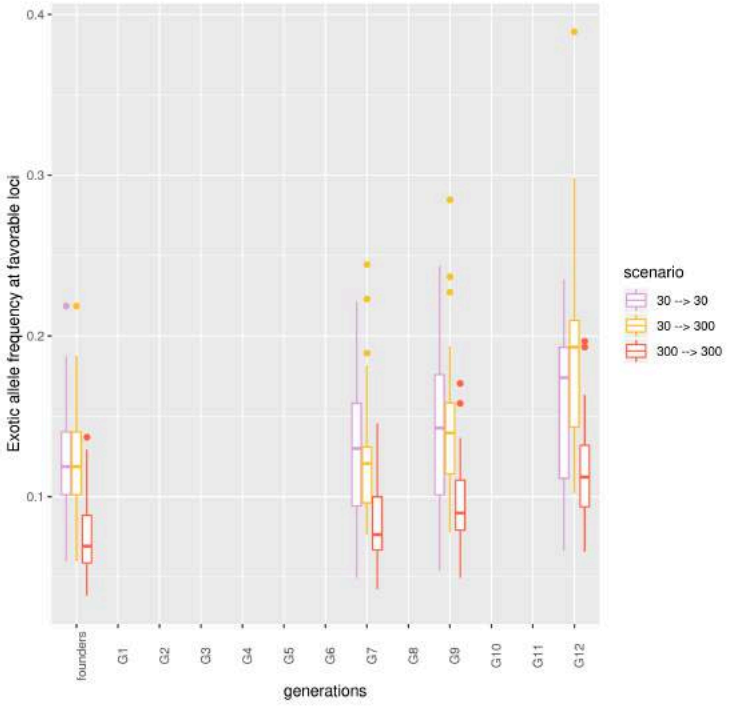


FIGURE 5.26 – Evolution de la fréquence de l'allèle exotique aux locus « favorables » dans trois scénarios sans suivi du lignage maternel, 30 → 30 en rose ; 30 → 300 en jaune ; 300 → 300 en rouge

	30 → 30	30 → 300	300 → 300
founders	0.12	0.12	0.08
G7	0.13	0.13	0.08
G9	0.14	0.14	0.10
G12	0.16	0.19	0.12

TABLE 5.26 – Evolution de la fréquence de l'allèle exotique aux locus « favorables » dans trois scénarios sans suivi du lignage maternel, 30 → 30, 30 → 300 et 300 → 300

En espérance la fréquence de l'allèle exotique aux locus « favorables » augmente au cours des simulations dans les trois scénarios, passant de 12 à 16% entre la génération des fondateurs et la génération 12 dans le scénario 30 → 30, de 12 à 19% dans le scénario 30 → 300 et de 8 à 21% dans le scénario commençant avec 300 fondateurs. Cette augmentation indique que la sélection génomique permet bien de choisir les individus portant l'allèle exotique aux locus « favorables ». La valeur moyenne de cet indicateur est plus faible dans le scénario débutant avec 300 fondateurs que dans les deux autres scénarios. En effet, les populations composées de 300 fondateurs comprennent 270 individus moins performants que les populations seulement composées de 30 fondateurs, ces 270 individus comportent probablement moins d'allèles exotiques aux locus « favorables ».

C'est dans le scénario 30 → 300 que la valeur moyenne de l'indicateur augmente le plus (augmentation de 4% dans les scénarios 30 → 30 et 300 → 300, de 7% dans le scénario 30 → 300). En effet ce scénario présente une valeur moyenne de l'indicateur haute dès la génération des fondateurs (12%), il présente donc à ce stade un avantage vis-à-vis du scénario 300 → 300. Le fait d'augmenter les effectifs dès la génération 7 permet d'augmenter les possibilités de recombinaisons par rapport au scénario 30 → 30 et ainsi de produire puis de sélectionner des individus accumulant les allèles exotiques aux locus

« favorables » plus rapidement qu'avec le scénario 30 → 30. 19% d'allèles exotiques sont ainsi obtenus aux locus « favorables » dans le scénario 30 → 300 contre 16% dans le scénario 30 → 30. **Choisir un faible effectif de fondateurs et un fort effectif en sortie de schéma de pre-breeding est donc le scénario le plus favorable vis-à-vis de la fréquence de l'allèle exotique aux locus « favorables ».**

Fréquence de l'allèle exotique pour l'ensemble des locus

L'évolution de la fréquence de l'allèle exotique sur l'ensemble des locus selon les effectifs 30 → 30, 30 → 300 et 300 → 300 sans suivi du lignage maternel est présentée dans la Figure 5.27. La moyenne de cet indicateur dans chacune des générations issues d'une action de sélection est donnée dans le Tableau 5.27. Les résultats de cette même étude avec suivi du lignage sont donnés en annexes (Figure A.11, Tableau A.6).

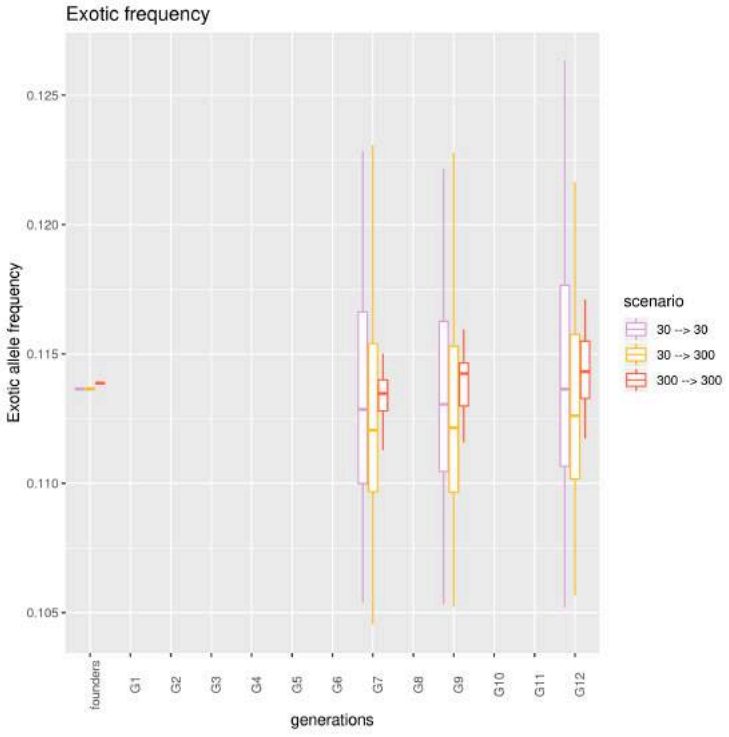


FIGURE 5.27 – Evolution de la fréquence de l'allèle exotique sur l'ensemble des locus dans trois scénarios sans suivi du lignage maternel, 30 → 30 en rose ; 30 → 300 en jaune ; 300 → 300 en rouge

	30 → 30	30 → 300	300 → 300
founders	0.11	0.11	0.11
G7	0.11	0.11	0.11
G9	0.11	0.11	0.11
G12	0.11	0.11	0.11

TABLE 5.27 – Evolution de la fréquence de l'allèle exotique sur l'ensemble des locus dans trois scénarios sans suivi du lignage maternel, 30 → 30, 30 → 300 et 300 → 300

Comme vu précédemment, la moyenne de la fréquence de l'allèle exotique sur l'ensemble des locus ne semble pas varier au cours des simulations, ni en fonction du scénario choisi. Cet indicateur a une valeur de 11% dans l'ensemble des scénarios. **L'effectif de la génération des fondateurs et de la population de pre-breeding finale n'a pas**

d'impact sur la fréquence d'allèles exotiques sur l'ensemble des locus.

Longueur moyenne des fragments exotiques

L'évolution de la longueur moyenne des fragments exotiques selon les effectifs 30 → 30, 30 → 300 et 300 → 300 sans suivi du lignage maternel est présentée dans la Figure 5.27. La moyenne de cet indicateur dans chacune des générations issues d'une action de sélection est donnée dans le Tableau 5.28. Les résultats de cette même étude avec suivi du lignage sont donnés en annexes (Figure A.12, Tableau A.7).

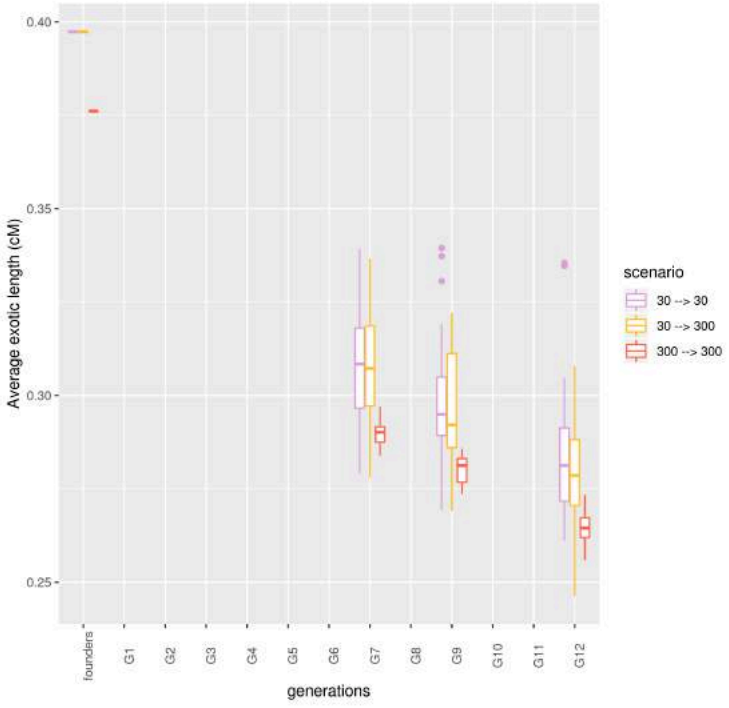


FIGURE 5.28 – Evolution de la longueur moyenne des fragments exotiques lors de trois scénarios sans suivi du lignage maternel, 30 → 30 en rose ; 30 → 300 en jaune ; 300 → 300 en rouge

	30 → 30	30 → 300	300 → 300
founders	0.40	0.40	0.38
G7	0.31	0.31	0.29
G9	0.30	0.30	0.28
G12	0.28	0.28	0.26

TABLE 5.28 – Evolution de la longueur moyenne des fragments exotiques lors de trois scénarios sans suivi du lignage maternel, 30 → 30, 30 → 300 et 300 → 300

L'évolution de la longueur moyenne des fragments exotiques dans la Figure indique une bonne recombinaison au cours du schéma de pre-breeding : la valeur de cet indicateur diminue au fil des générations dans tous les scénarios (de 0.40 à 0.28cM en moyenne dans les scénarios débutant avec 30 fondateurs, de 0.38 à 0.26 en moyenne dans les scénarios commençant avec 300 fondateurs).

La valeur moyenne de l'indicateur est plus faible dans le scénario 300 → 300 que dans les autres scénarios (0.38cM contre 0.40cM), ce qui indique qu'en moyenne les fragments exotiques sont plus courts lorsque le nombre de fondateurs est important. Aucune différence n'est observée entre les scénarios 30 → 30 et 30 → 300, le fait d'augmenter le nombre d'individus à partir de la génération G7 ne présente aucun avantage. Par ailleurs la longueur moyenne des fragments exotiques diminue de 0.12cM entre la génération des

fondateurs et la génération G12, et ce dans les trois scénarios. Ces résultats indiquent que **débuter un schéma de pre-breeding avec un effectif conséquent de fondateurs permet d'obtenir des fragments exotiques plus courts**.

Nombre de dimensions significatives de l'ACP

L'évolution de la longueur moyenne des fragments exotiques selon les effectifs 30 → 30, 30 → 300 et 300 → 300 sans suivi du lignage maternel est présentée dans la Figure 5.27. La moyenne de cet indicateur dans chacune des générations issues d'une action de sélection est donnée dans le Tableau 5.28. Les résultats de cette même étude avec suivi du lignage sont donnés en annexes (Figure A.13, Tableau A.8).

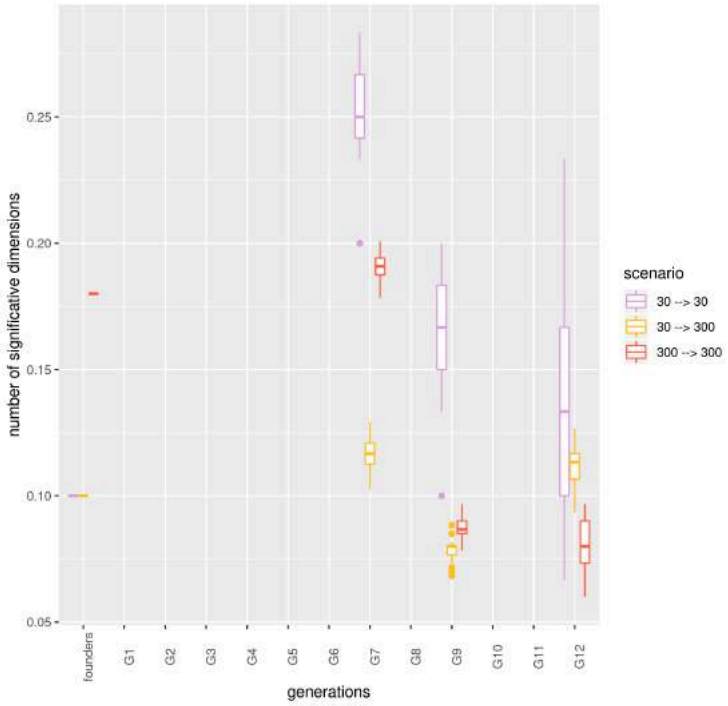


FIGURE 5.29 – Evolution du nombre de dimensions significatives de l'ACP rapporté au nombre d'individus lors de trois scénarios sans suivi du lignage maternel, 30 → 30 en rose ; 30 → 300 en jaune ; 300 → 300 en rouge

	30 → 30	30 → 300	300 → 300
founders	0.10	0.10	0.18
G7	0.25	0.12	0.19
G9	0.17	0.08	0.09
G12	0.14	0.11	0.08

TABLE 5.29 – Evolution du nombre de dimensions significatives de l'ACP rapporté au nombre d'individus lors de trois scénarios sans suivi du lignage maternel, 30 → 30, 30 → 300 et 300 → 300

Les deux scénarios comportant 30 fondateurs présentent une moyenne de cet indicateur de 0.10, contre 0.18 pour le scénario commençant avec 300 fondateurs. Il y a donc une diversité génétique plus importante entre les 300 fondateurs du scénario 300 → 300 qu'entre les 30 fondateurs des scénarios 30 → 30 et 30 → 300.

Le nombre de dimensions significatives de l'ACP rapporté au nombre d'individus présents dans la génération diminue entre la génération G7 et la génération G12 dans les scénarios 30 → 30 et 300 → 300. Les fluctuations dans le scénario 30 → 300, comme discuté dans la partie 5.2.2.1, sont difficilement interprétables. En G12 la valeur de l'indicateur est

plus faible dans le scénario 300 → 300 (0.08) que dans le scénario 30 → 30. Démarrer avec un faible effectif de fondateurs permettrait donc d'obtenir une population de pre-breeding finale moins structurée. Cependant ce résultat peut être expliqué par la taille des effectifs choisis : débuter un schéma de pre-breeding à partir de 30 fondateurs, soit un effectif très faible, permet de générer une population en G7 plus diverse que la population des fondateurs, du fait de la recombinaison des quelques individus de départ (0.25 en G7 contre 0.10 à la génération des fondateurs). Dans le scénario débutant avec 300 individus, l'effectif des fondateurs est suffisamment important pour capturer toute la structuration de la diversité génétique qu'il est possible d'obtenir lors de la recombinaison de ces fondateurs. C'est pourquoi aucune différence n'est observée entre la génération G7 et la génération des fondateurs dans le scénario 300 → 300 (0.19 contre 0.18). Cependant, cet indicateur semble dépendre du nombre d'individus présents dans la génération étudiée, il est donc susceptible d'entraîner de mauvaises conclusions lorsqu'il s'agit de comparer des générations composées d'effectifs différents. En effet, il est plus probable de trouver des individus se ressemblant fortement parmi 300 individus que parmi 30, d'autant plus que la diversité allélique en entrée du schéma de pre-breeding est assez restreinte puisque les fondateurs proviennent des 13 descendances AKER, créées à partir d'un même parent élite croisé avec 13 accessions exotiques différentes. La valeur de l'indicateur est donc probablement fortement pénalisée dans le scénario 300 → 300. Par conséquent **le nombre de dimensions significatives de l'ACP rapporté au nombre d'individus dans la génération étudiée n'est pas un bon indicateur de la structuration de diversité allélique pour comparer des populations de tailles différentes**.

Indicateurs de kinship

L'évolution de la longueur moyenne des valeurs minimales, moyennes et maximales de la matrice de Kinship selon les effectifs 30 → 30, 30 → 300 et 300 → 300 sans suivi du lignage maternel est présentée dans la Figure 5.27. Les moyennes de ces trois indicateurs dans chacune des générations issues d'une action de sélection sont données dans le Tableau 5.30. Les résultats de cette même étude avec suivi du lignage sont donnés en annexes (Figure A.14, Tableau A.9).

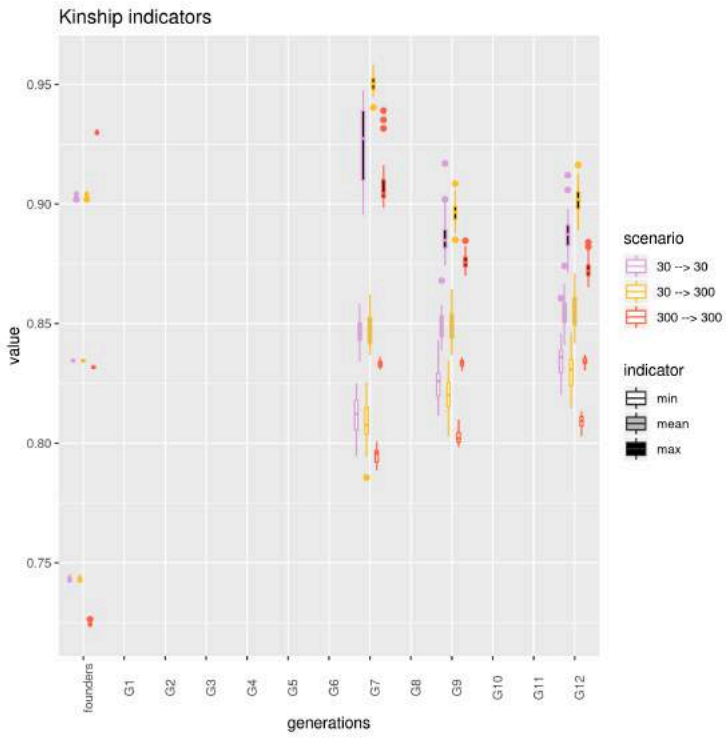


FIGURE 5.30 – Evolution de la valeur minimale (en blanc), moyenne (en gris) et maximale (en noir) de la matrice de kinship lors de trois scénarios, 30 → 30 en rose ; 30 → 300 en jaune ; 300 → 300 en rouge

	min		
	30 → 30	30 → 300	300 → 300
founders	0.74	0.74	0.72
G7	0.81	0.81	0.79
G9	0.83	0.82	0.80
G12	0.84	0.83	0.81
	mean		
	30 → 30	30 → 300	300 → 300
founders	0.83	0.83	0.83
G7	0.85	0.85	0.83
G9	0.85	0.85	0.83
G12	0.85	0.86	0.83
	max		
	30 → 30	30 → 300	300 → 300
founders	0.90	0.90	0.93
G7	0.92	0.95	0.91
G9	0.89	0.90	0.88
G12	0.89	0.90	0.87

TABLE 5.30 – Evolution du nombre de la valeur minimale, moyenne et maximale de la matrice de kinship lors de trois scénarios sans suivi du lignage maternel, 30 → 30, 30 → 300 et 300 → 300

A la génération des fondateurs, le scénario débutant avec 300 individus présente une valeur maximale de la kinship plus élevée que les scénarios débutant avec 30 individus (0.93 contre 0.90). Il est en effet plus probable de trouver des individus très proches parmi 300 fondateurs que parmi 30. De la même façon, le scénario commençant avec 300 fondateurs montre une valeur minimale de la kinship plus faible que celle obtenue avec les scénarios commençant avec 30 fondateurs (0.72 contre 0.74), car il est plus probable de trouver des individus très éloignés parmi 300 fondateur que parmi 30. La valeur moyenne de la kinship, qui représente quant à elle la valeur moyenne de similarité génétique entre deux individus dans une génération donnée, est identique à la génération des fondateurs que le schéma débute avec 30 ou 300 individus.

Une augmentation de la valeur minimale de la kinship est observée au cours des générations, conséquence du rapprochement génétique des individus produits par des croisements successifs sans introduction de nouveau matériel génétique. Cependant, la valeur maximale de la kinship diminue dans tous les scénarios, démontrant que les

nombreuses recombinaisons présentes dans le schéma permettent d'éviter la création d'un goulot d'étranglement génétique. En G12, le scénario présentant en espérance la valeur minimale la plus faible, c'est-à-dire les individus les plus éloignés, est le scénario 300 → 300 (0.81 contre 0.84 pour le scénario 30 → 30 et 0.83 pour le scénario 30 → 300). C'est également le scénario présentant en espérance la valeur maximale de la kinship la plus faible à cette même génération (0.87 contre 0.89 pour le scénario 30 → 30 et 0.90 pour le scénario 30 → 300), les deux individus les plus proches dans la population de pre-breeding produite par ce scénario sont donc plus éloignés que les deux individus les plus proches produits dans les populations de pre-breeding avec les deux autres scénarios.
La valeur moyenne de la kinship reste stable dans le scénario 300 → 300 (0.83), tandis qu'elle augmente légèrement dans les scénarios débutant avec 30 fondateurs, passant de 0.83 à la génération des fondateurs à 0.85 pour le scénario 30 → 30 et 0.86 pour les scénarios 30 → 300. En moyenne les individus composant la population de pre-breeding générée avec le scénario 300 → 300 sont donc plus distants au niveau génétique que ceux constituant les populations de pre-breeding produites avec les deux autres scénarios.
La diversité génétique semble donc être favorisée par un nombre de fondateurs important. En revanche, augmenter l'effectif des générations en cours de schéma ne semble pas avoir d'impact sur la diversité génétique de la population de pre-breeding finale, les indicateurs de kinship calculés à la génération G12 pour les scénarios 30 → 30 et 30 →300 n'étant que peu différents. **Le scénario 300 → 300 permet donc de produire la population de pre-breeding avec la plus grande diversité génétique.**

5.2.2.3 Discussion

L'étude de l'évolution des cinq indicateurs de diversité dans les différents scénarios de sélection génomique a permis d'étudier l'impact du suivi du lignage maternel et l'impact de l'effectif des générations sur la diversité de combinaison allélique. Les principales conclusions sont regroupées ci-dessous.

La fréquence de l'allèle exotique aux locus « favorables » augmente au cours des simulations dans tous les scénarios. La sélection génomique permet donc de sélectionner des individus portant l'allèle exotique aux locus « favorables ». Le suivi du lignage maternel n'a pas d'impact sur cette fréquence. En revanche les différents effectifs semblent avoir un effet sur cette fréquence, la population de pre-breeding finale présentant une fréquence exotique aux locus « favorables » plus élevée lorsque le schéma commence à partir d'un

nombre restreint de fondateurs et que le nombre d'individus par génération est augmenté à partir de la génération G7. Les meilleurs scénarios vis-à-vis de cet indicateur sont donc les scénarios **30 → 300 avec ou sans suivi du lignage maternel**.

La fréquence de l'allèle exotique au sein de la population est stable au cours des simulations. Aucun impact du suivi du lignage maternel ni de l'effectif des fondateurs ou de la population de pre-breeding n'a été mis en évidence. Le schéma de pre-breeding, constitué de nombreuses étapes de recombinaisons, permet donc d'éviter un phénomène de dérive qui aurait pu conduire à la perte de l'allèle exotique. Cet indicateur ne permet pas de distinguer les différents scénarios.

La longueur moyenne des fragments exotiques diminue au cours des simulations dans tous les scénarios. Ce résultat illustre les multiples recombinaisons ayant lieu au cours du schéma de pre-breeding et montre qu'il n'y a pas plus de recombinaison en fonction du suivi ou non du lignage maternel, ni en fonction de l'effectif des générations, la pente étant la même dans tous les scénarios. Des fragments exotiques courts permettent d'identifier plus facilement les individus qui possèdent les allèles exotiques favorables. Débuter ce schéma avec un effectif conséquent d'individus permet d'obtenir en espérance des fragments exotiques plus courts dans la génération des fondateurs, et par conséquent des fragments plus courts dans la population de pre-breeding finale. Les scénarios **300 → 300, avec ou sans suivi du lignage maternel**, sont donc les meilleurs scénarios vis-à-vis de cet indicateur.

Le nombre de dimensions significatives de l'ACP rapportées au nombre d'individus nécessaires pour représenter la diversité des individus est plus importante lorsque le lignage maternel n'est pas suivi. Cet indicateur étant sensible au nombre d'individu dans la génération considérée, il n'est pas adapté pour évaluer l'impact de l'effectif des générations. Les scénarios plébiscités par cet indicateur sont donc les **scénarios ne suivant pas le lignage maternel**.

La valeur maximale de la kinship, qui indique le degré de ressemblance génétique le plus important entre deux individus d'une génération donnée, diminue au cours des générations et ce quel que soit le scénario considéré. Ce résultat est la conséquence de l'éloignement génétique des individus créés au cours des générations, par les recombinaisons successives. Cependant la valeur minimale de la kinship, qui indique à l'inverse le degré de divergence génétique le plus important entre entre deux individus dans une génération donnée, augmente au fil des générations et ce dans tous les scénarios, traduisant une diminution

de la distance génétique entre les individus au cours des simulations sous l'effet de la sélection, que le lignage maternel soit suivi ou non et quels que soient les effectifs des fondateurs et de la population de pre-breeding en sortie du schéma. Le suivi du lignage maternel a un impact défavorable ou nul sur les valeurs de la kinship calculées, et donc sur la diversité génétique de la population de pre-breeding. Concernant l'impact de l'effectif des populations, le scénario permettant de minimiser les trois indicateurs de la kinship et donc de maximiser la diversité génétique de la population de pre-breeding produite est le scénario 300 → 300. Selon les indicateurs basés sur la kinship, le scénario **300 → 300 sans suivi du lignage maternel** est donc le scénario permettant de produire une population de pre-breeding comportant le maximum de diversité génétique.

Les cinq indicateurs de diversité génétique permettent donc d'identifier le scénario aboutissant à la création de la population de pre-breeding contenant le plus de diversité de combinaison allélique. Le suivi du lignage maternel a un effet défavorable ou nul sur la diversité de combinaison allélique finale, le scénario le plus favorable est donc un scénario ne tenant pas compte du lignage maternel. Si le scénario 30 → 300 permet d'obtenir une population de pre-breeding contenant une fréquence d'allèles exotiques aux locus « favorables » plus importante que les autres scénarios, le scénario 300 → 300 permet quant à lui d'obtenir une population de pre-breeding avec des fragments exotiques relativement courts, ce qui permet d'identifier plus aisément les individus détenteurs d'un allèle exotique favorable, et permet également de minimiser les valeurs des indicateurs de kinship et donc de maximiser la diversité génétique.
Le meilleur scénario est donc le scénario 300 → 300 sans suivi du lignage maternel.

Conclusion générale et perspectives

La sélection variétale a pour but de produire des variétés toujours plus performantes vis-à-vis de caractères agronomiques d'intérêt à partir de la diversité génétique existante dans les programmes de sélection. Cependant l'accumulation des caractères prérequis à la commercialisation des variétés de betterave sucrière a eu pour conséquence de réduire considérablement la variabilité génétique disponible dans les programmes de pre-breeding. Le projet AKER a été mis en place pour permettre d'élargir cette diversité génétique grâce à une approche originale d'utilisation de ressources génétiques exotiques, ce qui a conduit à l'identification de 16 accessions exotiques représentant l'ensemble de la diversité allélique qui n'est pas déjà présente au sein des programmes de sélection de betteraves sucrières. L'objectif de cette thèse était de favoriser l'intégration de cette nouvelle diversité génétique découverte dans le cadre du projet AKER dans un programme de pre-breeding en utilisant la sélection génomique.

Ce travail de thèse est constitué de plusieurs étapes. Dans le chapitre 2, les analyses portant sur la comparaison de l'architecture génétique entre une descendance (élite × exotique) et un panel élite ont permis de définir une méthodologie pour identifier l'architecture génétique d'un caractère, et de mettre en évidence l'existence de régions génomiques favorables apportées par l'introduction d'une accession exotique dans un germoplasme élite. Une partie de ces analyses, axée sur l'étude des impuretés diminuant l'extractibilité du sucre blanc des betteraves sucrières, a été valorisée par un article (Pegot-Espagnet et al. 2019). Les études effectuées dans le chapitre 3 ont permis d'identifier l'architecture génétique du rendement racinaire à partir de 13 descendances AKER, chacune de ces descendances provenant de l'introgression d'une accession exotique particulière au sein d'un germoplasme élite. 24 QTLs associés au rendement racinaire ont ainsi été identifiés dans une ou plusieurs descendances. Deux de ces QTLs présentent des effets antagonistes, l'allèle exotique à ces QTLs pouvant avoir un effet favorable ou défavorable important sur le rendement racinaire en fonction de l'accession exotique dont il provient. Le chapitre 4 permet de comprendre le processus qui a conduit à l'élaboration du simulateur en partenariat avec les sélectionneurs de Florimond Desprez Veuve & Fils, et comment l'architecture génétique déterminée au chapitre précédent est utilisée pour

simuler l'évolution du rendement racinaire au cours de 12 scénarios de pre-breeding. La comparaison de l'évolution de plusieurs indicateurs définis dans le chapitre 5 a permis d'évaluer l'intérêt de chacun de ces scénarios vis-à-vis de la performance ainsi que de la diversité de combinaison allélique de la population de pre-breeding générée. Ainsi, la comparaison de l'évolution de la performance en fonction de l'utilisation de méthodes de sélection génomique ou phénotypique vis-à-vis du rendement racinaire a permis de conclure à un avantage de la sélection phénotypique, alors qu'un avantage de la sélection génomique était attendu compte tenu du caractère complexe du rendement racinaire (Heffner et al. 2010). Plusieurs explications peuvent justifier ce résultat. Tout d'abord, la sélection portait sur le caractère de rendement racinaire, dont l'héritabilité calculée dans le chapitre 3 s'avère élevée (0.60). Or, Rajsic et al. (2016) ont montré que la sélection génomique présente un avantage sur la sélection phénotypique principalement quand l'héritabilité du caractère à sélectionner est faible. De plus le modèle de sélection génomique utilisé attribue un effet unique à chaque QTL, alors qu'un QTL pour lequel l'allèle exotique a un effet favorable ou défavorable sur le rendement racinaire en fonction de la descendance dont il provient a été identifié lors de l'étude de l'architecture génétique du caractère. Ce modèle est donc faux. La performance de la population de pre-breeding obtenue par sélection sur la vraie valeur génomique, valeur dont on dispose exclusivement *in silico* et qui représente la population optimale théorique qu'il est possible d'obtenir par sélection génomique, est largement supérieure à la performance obtenue par sélection génomique et par sélection phénotypique. Ce résultat indique que la sélection génomique avec un modèle amélioré peut amener à la création d'une population de pre-breeding plus performante que la sélection phénotypique.

Concernant la diversité génétique de la population de pre-breeding générée, deux paramètres ont été évalués : l'impact du suivi du lignage maternel, et l'impact de l'effectif des fondateurs de départ et de la population de pre-breeding finale. Ces études ont permis d'établir que la diversité allélique est favorisée par un scénario ne suivant pas le lignage maternel et débutant avec un effectif important d'individus, soit 300 fondateurs, pour aboutir à une population de pre-breeding composée de 300 individus. Les différents indicateurs ont également permis de montrer qu'aucun des différents schémas simulés ne conduit à la création d'un goulot d'étranglement génétique.

De nombreux autres aspects du schéma de pre-breeding ont une influence sur le gain génétique et sur la diversité génétique. D'autres pistes de réflexion auraient pu être étudiées grâce au simulateur. Plusieurs d'entre elles sont abordées ci-dessous.

Caractères d'intérêt agronomiques

Les simulations effectuées au cours de cette thèse portaient sur la sélection d'un unique caractère d'intérêt agronomique, le rendement racinaire. Il pourrait néanmoins être intéressant de sélectionner deux caractères en parallèle. Définir des fonctions permettant la sélection conjointe de plusieurs caractères est possible grâce à la structure de l'élément *generation*, qui permet d'avoir accès aux valeurs phénotypiques de plusieurs caractères distincts ainsi qu'au génotype de chacun des individus de la génération. Il serait ainsi envisageable de sélectionner simultanément sur un caractère complexe, tel que le rendement racinaire, et sur un caractère monogénique, comme la résistance à une maladie. En ce cas une nouvelle action devrait être définie au sein du simulateur afin de sélectionner les individus ayant le meilleur phénotype de rendement racinaire tout en possédant l'allèle de résistance pour la maladie au niveau d'un marqueur identifié sur le génome. Il serait également possible de simuler une sélection conjointe de deux caractères complexes, tels que le rendement racinaire et le taux de sucre blanc extractible.

Nouveau modèle de sélection génomique

Les simulations réalisées portent sur 1435 marqueurs distincts. Lors de l'étude de l'architecture génétique du rendement racinaire, des QTLs antagonistes ont été détectés. Une meilleure couverture du génome pourrait éventuellement permettre de diviser ces QTLs antagonistes en QTLs distincts, et ainsi d'améliorer l'estimation de l'effet des marqueurs lors de l'établissement du modèle de prédiction génomique. Dans le cas contraire, un modèle multiallélique permettant de prendre en compte les différents effets apportés par chacune des descendances pourrait être envisagé pour améliorer la prédiction de la valeur génomique des individus à sélectionner.

Prise en compte du plan d'expérience

Dans cette thèse, l'impact des différentes méthodes de sélection n'a été étudié que vis-à-vis de la performance de la population de pre-breeding obtenue. Pourtant, l'aspect économique de la mise en place de ces schémas représente bien évidemment un argument important dans le choix d'un schéma de sélection. Dans les simulations effectuées, seul le phénotype moyen des individus est considéré. Simuler des phénotypes observés sur divers environnements, et donc prendre en compte le plan d'expérience, permettrait d'effectuer une étude économique pour établir s'il est plus utile de faire beaucoup de phénotypage, dans un but de sélection phénotypique, ou au contraire de diminuer le phénotypage au profit du génotypage, en vue de faire de la sélection génomique. Rajsic et al. (2016) proposent ainsi un modèle permettant d'étudier l'effet de plusieurs paramètres sur la performance économique de la sélection génomique et de la sélection phénotypique, comme

par exemple le coût relatif du phénotypage et du génotypage et la taille de la population d'entraînement.

Choix des individus à croiser

Lors des simulations des différents schémas de pre-breeding, les fondateurs ont été sélectionnés en conservant le meilleur individu issu de chacune des 13 descendances AKER, puis en choisissant les meilleurs individus toutes descendances confondues. Cependant, certaines descendances AKER ne présentent pas d'allèle exotiques à des locus « favorables ». Ne pas inclure d'individu issu de telles descendances dans la génération des fondateurs permettrait de réduire le nombre d'allèles exotiques défavorables, et ainsi favoriserait la création d'individus présentant peu d'allèles exotiques défavorables.

D'autres façons de choisir les fondateurs peuvent également être envisagées. Plutôt que de les sélectionner en fonction de leur valeur phénotypique, il pourrait être intéressant de choisir des couples de fondateurs qui produiraient les descendants les plus intéressants. En effet, Allier et al. (2019) ont montré sur le maïs que prédire l'utilité des croisements en termes de gain génétique et de diversité allélique attendus est intéressant pour garantir les performances de la descendance et pour maintenir un gain génétique à long terme dans la sélection végétale.

De la même façon, plutôt que d'effectuer une pollinisation aléatoire il pourrait être intéressant de sélectionner et de croiser le meilleur mâle avec la meilleure femelle vis-à-vis de la valeur du caractère étudié. Cependant, croiser les meilleurs individus entre eux ne garantit pas d'obtenir les meilleurs individus sélectionnés, pour un caractère dont les QTLs ont des effets additifs. En effet, il faut s'intéresser à la valeur des extrêmes d'un couple de parents et donc de choisir les parents en fonction de la distribution de la valeur de leur descendance. Pour ce faire il est nécessaire de pouvoir prédire la distribution de la valeur des descendants d'un couple pour le caractère d'intérêt. Il est possible d'estimer cette valeur à partir d'une équation de prédiction établie à partir d'une population d'entraînement, comme celles produites dans le schéma basé sur la sélection génomique. Une autre façon de procéder, plus simple mais moins précise, serait de se baser uniquement sur la valeur des effets des QTLs, valeur déterminée lors de l'étude de l'architecture génétique du caractère.

Flexibilité du simulateur

L'intérêt de simuler des schémas de pre-breeding est de pouvoir tester *in silico* différentes alternatives qui seraient trop longues et coûteuses à expérimenter en champ. De nouvelles idées peuvent émerger à la vue des résultats, c'est pourquoi le simulateur élaboré dans le cadre de cette thèse a été conçu de façon à être flexible. Les pistes de réflexion abordées

ci-dessus peuvent aisément y être implémentées grâce à la structure de l'objet *generation* utilisé, qui permet d'accéder facilement à l'ensemble des informations des individus au sein d'une génération simulée. Cet outil a ainsi pour vocation d'être amélioré progressivement afin d'offrir la possibilité de simuler et comparer tous les schémas de pre-breeding qui pourront être imaginés par les sélectionneurs.

Bibliographie

Ibraheem Adetunji, Glenda Willems, Hendrik Tschoep, Alexandra Bürkholz, Steve Barnes, Martin Boer, Marcos Malosetti, Stefaan Horemans, and Fred van Eeuwijk. Genetic diversity and linkage disequilibrium analysis in elite sugar beet breeding lines and wild beet accessions. Theoretical and applied genetics, 127(3) :559–571, 2014.

Antoine Allier, Laurence Moreau, Alain Charcosset, Simon Teyssèdre, and Christina Lehermeier. Usefulness criterion and post-selection parental contributions in multi-parental crosses : application to polygenic trait introgression. G3 : Genes, Genomes, Genetics, 9(5) :1469–1479, 2019.

John C Avise. Molecular markers, natural history and evolution. Springer Science & Business Media, 2012.

Ellen Barzen, Rainer Stahl, Elke Fuchs, Dietrich C. Borchardt, and Francesco Salamini. Development of coupling-repulsion-phase scar markers diagnostic for the sugar beet rr1 allele conferring resistance to rhizomania. Molecular Breeding, 3(3) :231–238, Jun 1997. ISSN 1572-9788. doi : 10.1023/A:1009626214058. URL `https://doi.org/10.1023/A:1009626214058`.

Filippo M Bassi, Alison R Bentley, Gilles Charmet, Rodomiro Ortiz, and Jose Crossa. Breeding schemes for the implementation of genomic selection in wheat (triticum spp.). Plant Science, 242 :23–36, 2016.

William D Beavis. Qtl analyses : power, precision, and accuracy. Molecular dissection of complex traits, 1998 :145–162, 1998.

Rex Bernardo. Genomewide selection for rapid introgression of exotic germplasm in maize. Crop Science, 49(2) :419–425, 2009.

Rex Bernardo and Jianming Yu. Prospects for genomewide selection for quantitative traits in maize. Crop Science, 47(3) :1082–1090, 2007.

Nils Olof Bosemark. l0. la betterave à sucre. 1979.

Alain Bouquet, Paul Truel, and Robert Wagner. Application des méthodes de sélection récurrente à l'amélioration génétique de la vigne. 1981.

Brian L Browning and Sharon R Browning. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. The American Journal of Human Genetics, 84(2) :210–223, 2009.

Emily Combs and Rex Bernardo. Genomewide selection to introgress semidwarf maize germplasm into us corn belt inbreds. Crop Science, 53(4) :1427–1436, 2013.

Elizabeth H Corder, Ann M Saunders, Waren J Strittmatter, Donald E Schmechel, P Craig Gaskell, GWet Small, Allen D Roses, JL Haines, and Margaret A Pericak-Vance. Gene dose of apolipoprotein e type 4 allele and the risk of alzheimer's disease in late onset families. Science, 261(5123) :921–923, 1993.

Charles Darwin. The effects of cross and self fertilisation in the vegetable kingdom. D. Appleton, 1877.

Noel Deerr et al. The history of sugar. volume ii. The history of sugar. Volume II., 1950.

Desplanque. Betteraves mauvaises herbes et rudérales : diversité génétique, traits d'histoire de vie et flux de gènes au sein du complexe d'espèces cultivées-sauvages Beta vulgaris ssp. PhD thesis, Université de LILLE I, 1999.

M Desprez and B Desprez. Évolution de methods de sélection de la betterave sucrière des origines à nos jours. Compte Rendu de l'Academie d'Agriculture Francais, 79(6) : 71–84, 1993.

Zeratsion Abera Desta and Rodomiro Ortiz. Genomic selection : genome-wide prediction in plant improvement. Trends in plant science, 19(9) :592–601, 2014.

H Doggett and SA Eberhart. Recurrent selection in sorghum. Crop Science, 8(1) :119–121, 1968.

Juliane C Dohm, André E Minoche, Daniela Holtgräwe, Salvador Capella-Gutiérrez, Falk Zakrzewski, Hakim Tafer, Oliver Rupp, Thomas Rosleff Sörensen, Ralf Stracke, Richard Reinhardt, et al. The genome of the recently domesticated crop plant sugar beet (beta vulgaris). Nature, 505(7484) :546, 2014.

C Doré and F Varoquaux. Histoire et amélioration de cinquante plantes cultivées. Editions Quae, July 2006. ISBN 9782738012159. Google-Books-ID : 6pbtFya2zy8C.

Eric Frichot and Olivier Francois. LEA : an R package for Landscape and Ecological Association studies. under review, 2014. URL http://membres-timc.imag.fr/Olivier.Francois/lea.html.

Ganiere. La bataille du sucre. Revue du Souvenir Napoléonien, 1971. URL https://www.napoleon.org/histoire-des-2-empires/articles/la-bataille-du-sucre/.

Mike Goddard. Genomic selection : prediction of accuracy and maximisation of long term response. Genetica, 136(2) :245–257, 2009.

IL Goldman. Inbreeding and outbreeding in the development of a modern heterosis concept. Genetics and Explortation of Heterosis in Crops. ASA, CSSA, and SSSA, Madison, WI, pages 7–18, 1999.

Gregor Gorjanc, Janez Jenko, Sarah J Hearne, and John M Hickey. Initiating maize pre-breeding programs using genomic selection to harness polygenic variation from landrace populations. BMC genomics, 17(1) :30, 2016.

Ben John Hayes, Peter M Visscher, and Michael E Goddard. Increased accuracy of artificial selection by using the realized relationship matrix. Genetics research, 91(1) : 47–60, 2009.

Elliot L Heffner, Aaron J Lorenz, Jean-Luc Jannink, and Mark E Sorrells. Plant breeding with genomic selection : gain per unit time and cost. Crop science, 50(5) :1681–1690, 2010.

Christa M. Hoffmann. Root quality of sugarbeet. Sugar Tech, 12(3) :276–287, Dec 2010. doi : 10.1007/s12355-010-0040-6.

F Hospital, I Goldringer, and S Openshaw. Efficient marker-based recurrent selection for multiple quantitative trait loci. Genetical research, 75(3) :357, 2000.

Frédéric Hospital. Selection in backcross programmes. Philosophical Transactions of the Royal Society B : Biological Sciences, 360(1459) :1503–1511, 2005.

Muhammad Iqbal and Abdul Saleem. Sugar beet potential to beat sugarcane as a sugar crop in pakistan. American-Eurasian Journal of Agricultural & Environmental Sciences, 15 :36–44, 01 2015. doi : 10.5829/idosi.aejaes.2015.15.1.12480.

SP Jefferies, BJ King, AR Barr, P Warner, SJ Logue, and P Langridge. Marker-assisted backcross introgression of the yd2 gene conferring resistance to barley yellow dwarf virus in barley. Plant Breeding, 122(1) :52–56, 2003.

Hyun Min Kang, Noah A Zaitlen, Claire M Wade, Andrew Kirby, David Heckerman, Mark J Daly, and Eleazar Eskin. Efficient control of population structure in model organism association mapping. Genetics, 178(3) :1709–1723, 2008.

Russell Lande and Robin Thompson. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics, 124(3) :743–756, 1990.

KNUD Larsen. Self-incompatibility in beta vulgaris l. i. four gametophytic, complementary s-loci in sugar beet. Hereditas, 85(2) :227–248, 1977.

François Le Cochec and Pierre Soreau. Aptitudes générale et spécifique à la combinaison de lignées de betterave fourragère et sucrière (beta vulgaris l.). 1982.

Steven Maenhout, Bernard De Baets, and Geert Haesaert. Marker-based estimation of the coefficient of coancestry in hybrid breeding programmes. Theoretical and applied genetics, 118(6) :1181–1192, 2009.

Brigitte Mangin, Renaud Rincent, Charles-Elie Rabier, Laurence Moreau, and Ellen Goudemand-Dugue. Training set optimization of genomic prediction by means of ethacc. PLOS ONE, 14(2) :1–21, 02 2019. doi : 10.1371/journal.pone.0205629. URL `https://doi.org/10.1371/journal.pone.0205629`.

Meuwissen, Hayes, and Goddard. Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157(4) :1819–1829, 2001.

GFJ Milford. The growth and development of the storage root of sugar beet. Annals of Applied Biology, 75(3) :427–438, 1973.

Bernward Märländer, T. Lange, and A. Wulkow. Dispersal principles of sugar beet from seed to sugar with particular relation to genetically modified varieties. Journal fur Kulturpflanzen, 63 :349–373, 01 2011.

J Ødegård, MH Yazdi, AK Sonesson, et al. Incorporating desirable genetic characteristics from an inferior into a superior population using genomic selection. Genetics, 181(2) : 737–745, 2009a.

Jørgen Ødegård, Anna K Sonesson, M Hossein Yazdi, and Theo HE Meuwissen. Introgression of a major qtl from an inferior into a superior population using genomic selection. Genetics Selection Evolution, 41(1) :38, 2009b.

FV Owen. Cytoplasmically inherited male-sterility in sugar beets. Jour. Agr. Res., 71 : 423–440, 1945.

FV Owen. Mendelian male sterility in sugar beets. In Proc Am Soc Sugar Beet Technol, volume 7, pages 371–376, 1952.

FV Owen and George K Ryser. Some mendelian characters in beta vulgaris l. and linkages observed in the yrb group. J Agric Res, 65(3) :155–171, 1942.

Nick Patterson, Alkes L Price, and David Reich. Population structure and eigenanalysis. PLoS genetics, 2(12) :e190, 2006.

Prune Pegot-Espagnet, Olivier Guillaume, Bruno Desprez, Brigitte Devaux, Pierre Devaux, Karine Henry, Nicolas Henry, Glenda Willems, Ellen Goudemand, and Brigitte Mangin. Discovery of interesting new polymorphisms in a sugar beet (elite × exotic) progeny by comparison with an elite panel. Theoretical and Applied Genetics, 132(11) : 3063–3078, 2019.

N Piyasatian, RL Fernando, and JCM Dekkers. Introgressing multiple qtl in breeding programmes of limited size. Journal of Animal Breeding and Genetics, 125(1) :50–56, 2008.

Dean W Podlich, Christopher R Winkler, and Mark Cooper. Mapping as you go. Crop Science, 44(5) :1560–1571, 2004.

Jonathan K Pritchard, Matthew Stephens, Noah A Rosenberg, and Peter Donnelly. Association mapping in structured populations. The American Journal of Human Genetics, 67(1) :170–181, 2000.

Antoni Rafalski. Applications of single nucleotide polymorphisms in crop genetics. Current opinion in plant biology, 5(2) :94–100, 2002.

Predrag Rajsic, Alfons Weersink, Alireza Navabi, and K Peter Pauls. Economics of genomic selection : the role of prediction accuracy and relative genotyping costs. Euphytica, 210(2) :259–276, 2016.

Renaud Rincent, Denis Laloë, Stéphane Nicolas, Thomas Altmann, Dominique Brunel, Pedro Revilla, Victor M Rodriguez, J Moreno-Gonzalez, A Melchinger, Eva Bauer, et al. Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals : comparison of methods in two diverse groups of maize inbreds (zea mays l.). Genetics, 192(2) :715–728, 2012.

VF Savitsky et al. Monogerm sugar beets in the united states. In Proc. Am. Soc. Sugar Beet Technol, volume 6, pages 156–159, 1950.

T Schulz-Streeck, JO Ogutu, Z Karaman, C Knaak, and HP Piepho. Genomic selection using multiple populations. Crop Science, 52(6) :2453–2461, 2012.

Vincent Segura, Bjarni J Vilhjálmsson, Alexander Platt, Arthur Korte, Ümit Seren, Quan Long, and Magnus Nordborg. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nature genetics, 44(7) : 825–830, 2012.

Bertrand Servin, Olivier C Martin, Marc Mézard, et al. Toward a theory of marker-assisted gene pyramiding. Genetics, 168(1) :513–523, 2004.

George Harrison Shull. Duplicate genes for capsule-form inbursa bursa-pastoris. Molecular and General Genetics MGG, 12(1) :97–149, 1914.

PM VanRaden. Efficient methods to compute genomic predictions. Journal of dairy science, 91(11) :4414–4423, 2008.

Claude Viel and Denis Brançon. Le sucre de betterave et l'essor de son industrie : Des premiers travaux jusqu'à la fin de la guerre de 1914-1918. 1999. ISSN 0035-2349. doi : 10.3406/pharm.1999.4743.

Louis de Vilmorin. Note sur la création d'une nouvelle race de betterave-considération sur l'hérédité des végétaux. Comptes Rendus des séances hebdomadaires de l'Académie des Sciences, 43(1856) :871–874, 1856.

Manish K Vishwakarma, VK Mishra, PK Gupta, PS Yadav, H Kumar, and Arun K Joshi. Introgression of the high grain protein gene gpc-b1 in an elite wheat variety of indo-gangetic plains through marker assisted backcross breeding. Current Plant Biology, 1 : 60–67, 2014.

Shi-Bo Wang, Jian-Ying Feng, Wen-Long Ren, Bo Huang, Ling Zhou, Yang-Jun Wen, Jin Zhang, Jim M Dunwell, Shizhong Xu, and Yuan-Ming Zhang. Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. Scientific reports, 6 :19444, 2016.

John C Whittaker, Robin Thompson, and Mike C Denham. Marker-assisted selection using ridge regression. Genetics Research, 75(2) :249–252, 2000.

Tobias Würschum, Jochen C Reif, Thomas Kraft, Geert Janssen, and Yusheng Zhao. Genomic selection in sugar beet breeding populations. BMC genetics, 14(1) :85, 2013.

Yunbi Xu and Jonathan H Crouch. Marker-assisted selection in plant breeding : from publications to practice. Crop science, 48(2) :391–407, 2008.

Yunbi Xu, Yanli Lu, Chuanxiao Xie, Shibin Gao, Jianmin Wan, and Boddupalli M Prasanna. Whole-genome strategies for marker-assisted plant breeding. Molecular breeding, 29(4) :833–854, 2012.

Jianming Yu, Gael Pressoir, William H Briggs, Irie Vroh Bi, Masanori Yamasaki, John F Doebley, Michael D McMullen, Brandon S Gaut, Dahlia M Nielsen, James B Holland, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature genetics, 38(2) :203–208, 2006.

Gang Zhou, Ying Chen, Wen Yao, Chengjun Zhang, Weibo Xie, Jinping Hua, Yongzhong Xing, Jinghua Xiao, and Qifa Zhang. Genetic composition of yield heterosis in an elite rice hybrid. Proceedings of the National Academy of Sciences, 109(39) :15847–15852, 2012.

Annexe A

Annexes

Dispositif expérimental de chacun des 7 environnements AKER :

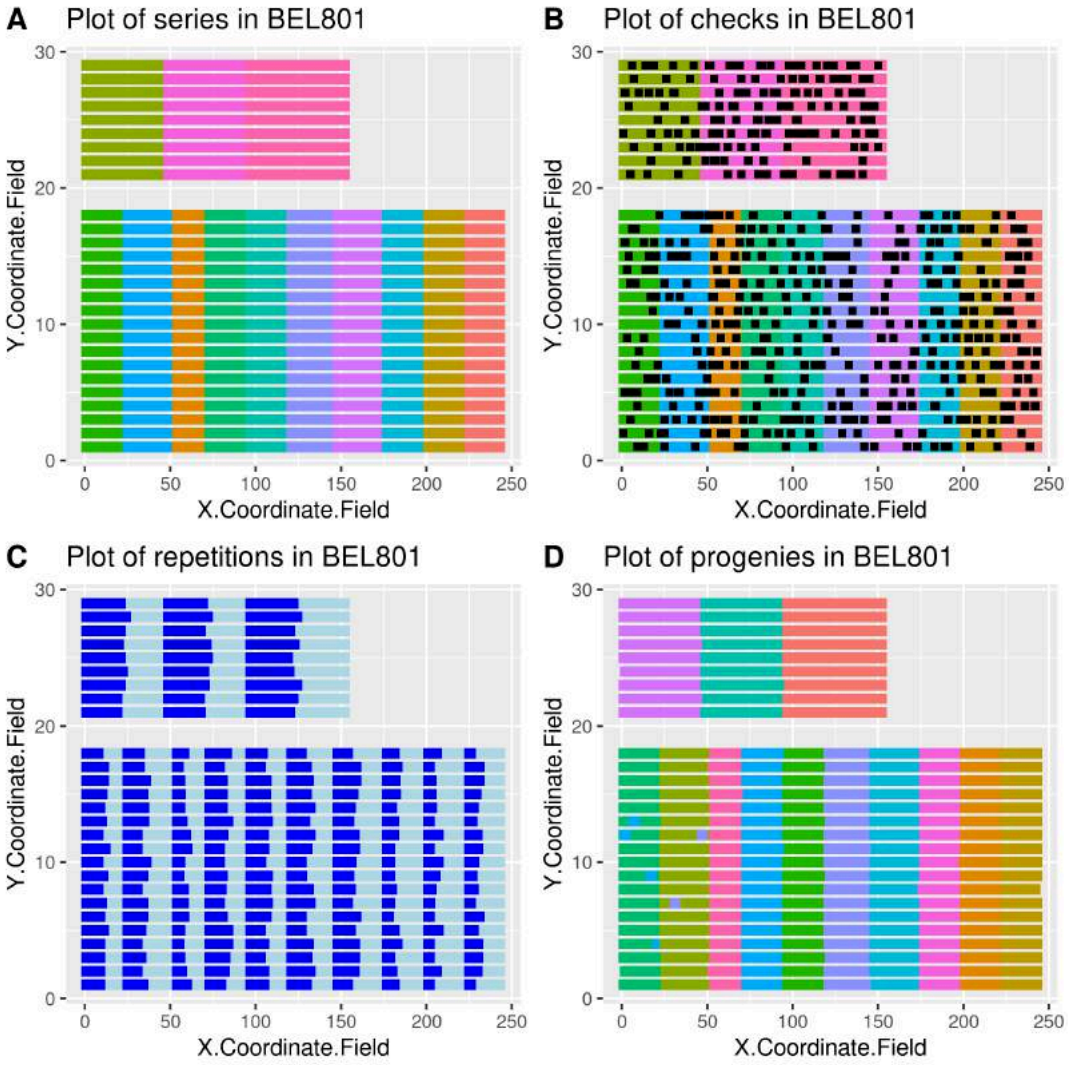


FIGURE A.1 – Dispositif expérimental de l'environnement BEL801 : A. Répartition des individus dans le champ, chaque couleur correspondant à une série ; B. Position des témoins (en noir) dans les séries ; C. Répartition des répétitions par série dans le champ (bleu foncé pour la répétition 1, bleu clair pour la répétition 2) ; D. répartition des individus des 13 descendances dans le champ, chaque couleur correspondant à une descendance

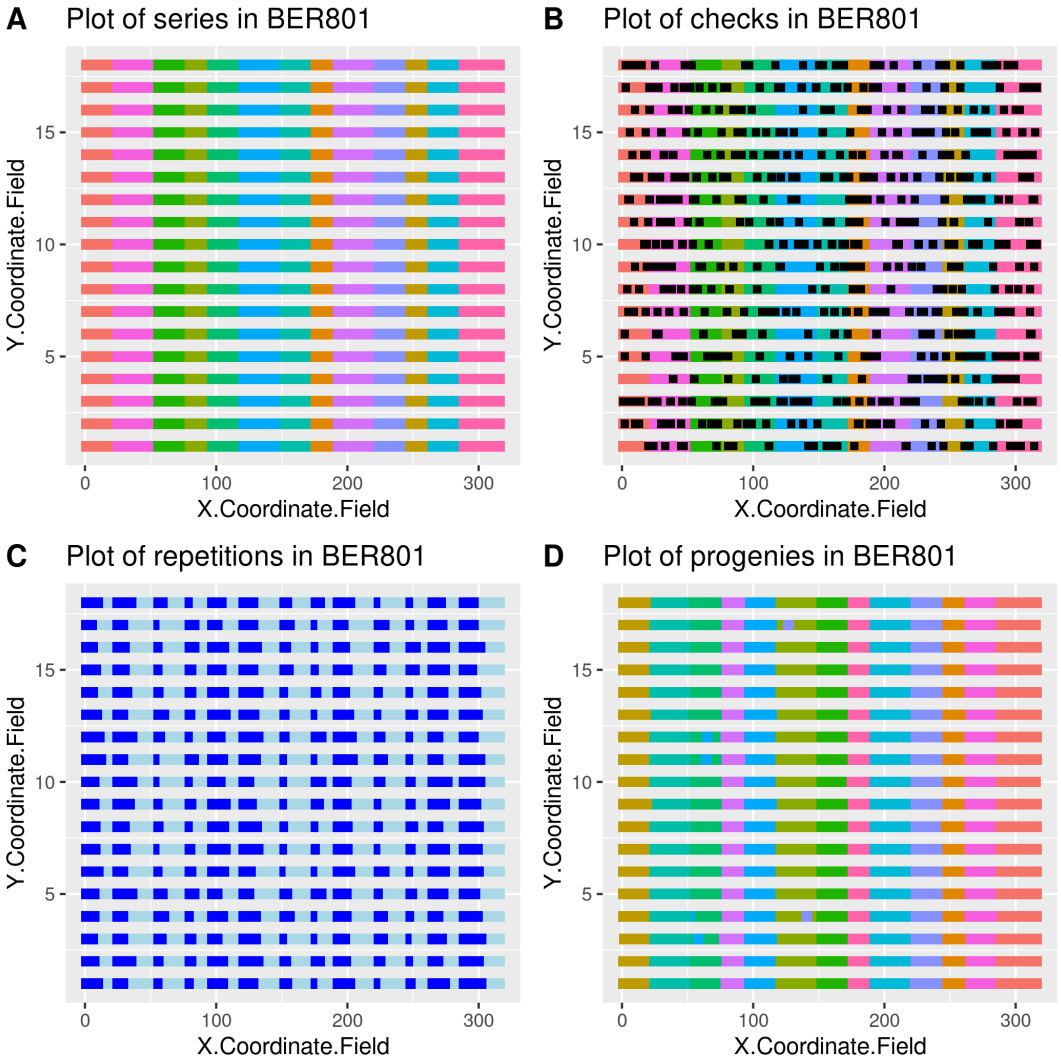


FIGURE A.2 – Dispositif expérimental de l'environnement BER801 : A. Répartition des individus dans le champ, chaque couleur correspondant à une série ; B. Position des témoins (en noir) dans les séries ; C. Répartition des répétitions par série dans le champ (bleu foncé pour la répétition 1, bleu clair pour la répétition 2) ; D. répartition des individus des 13 descendances dans le champ, chaque couleur correspondant à une descendance

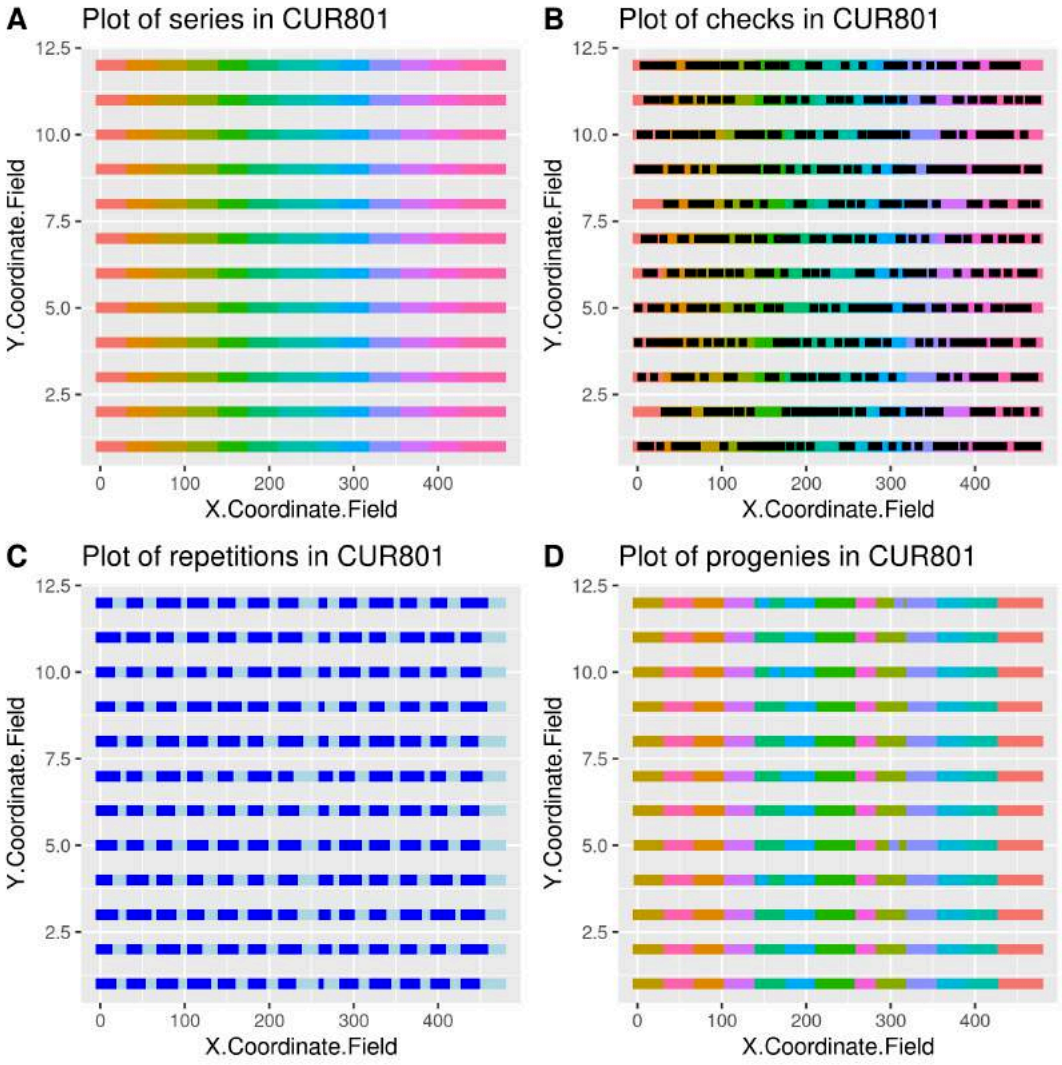


FIGURE A.3 – Dispositif expérimental de l'environnement CUR801 : A. Répartition des individus dans le champ, chaque couleur correspondant à une série ; B. Position des témoins (en noir) dans les séries ; C. Répartition des répétitions par série dans le champ (bleu foncé pour la répétition 1, bleu clair pour la répétition 2) ; D. répartition des individus des 13 descendances dans le champ, chaque couleur correspondant à une descendance

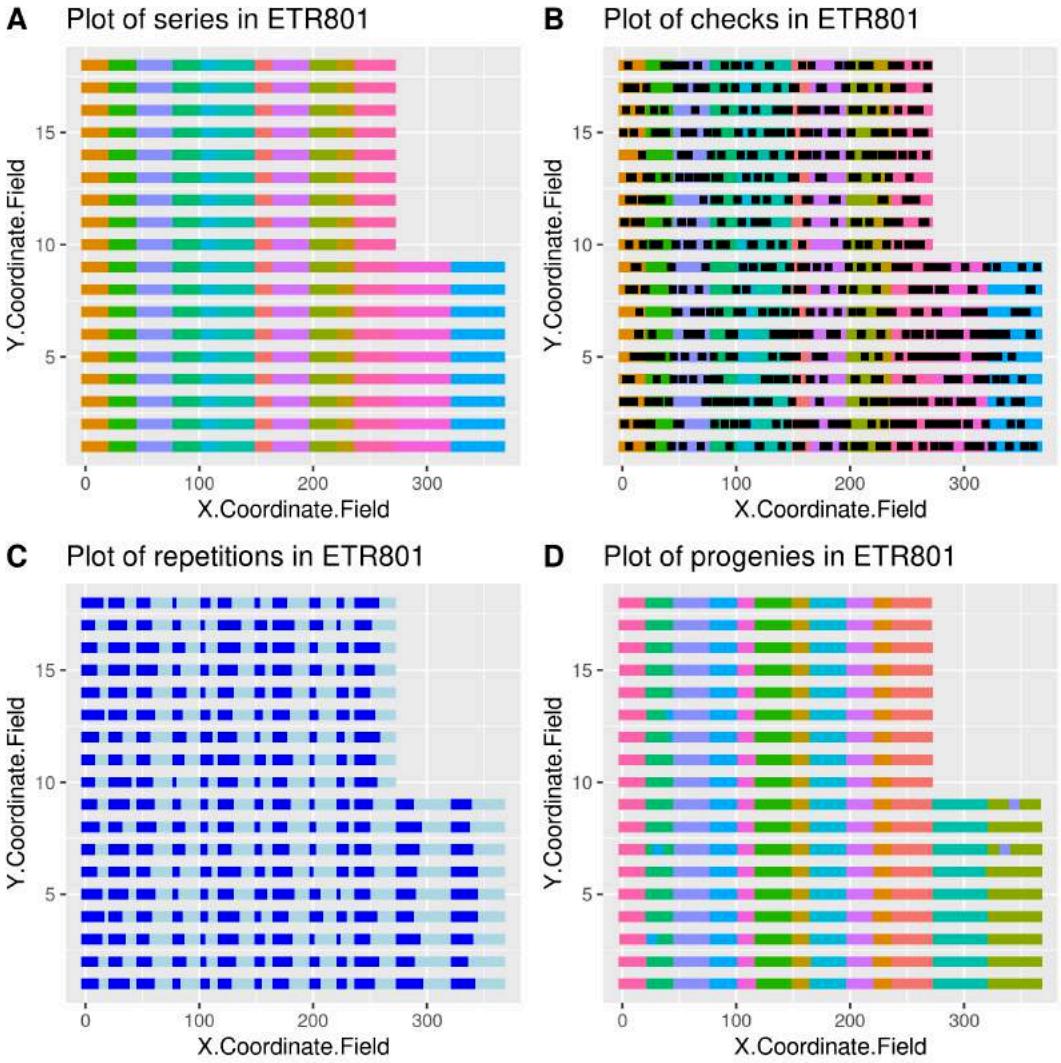


FIGURE A.4 – Dispositif expérimental de l'environnement ETR801 : A. Répartition des individus dans le champ, chaque couleur correspondant à une série ; B. Position des témoins (en noir) dans les séries ; C. Répartition des répétitions par série dans le champ (bleu foncé pour la répétition 1, bleu clair pour la répétition 2) ; D. répartition des individus des 13 descendances dans le champ, chaque couleur correspondant à une descendance

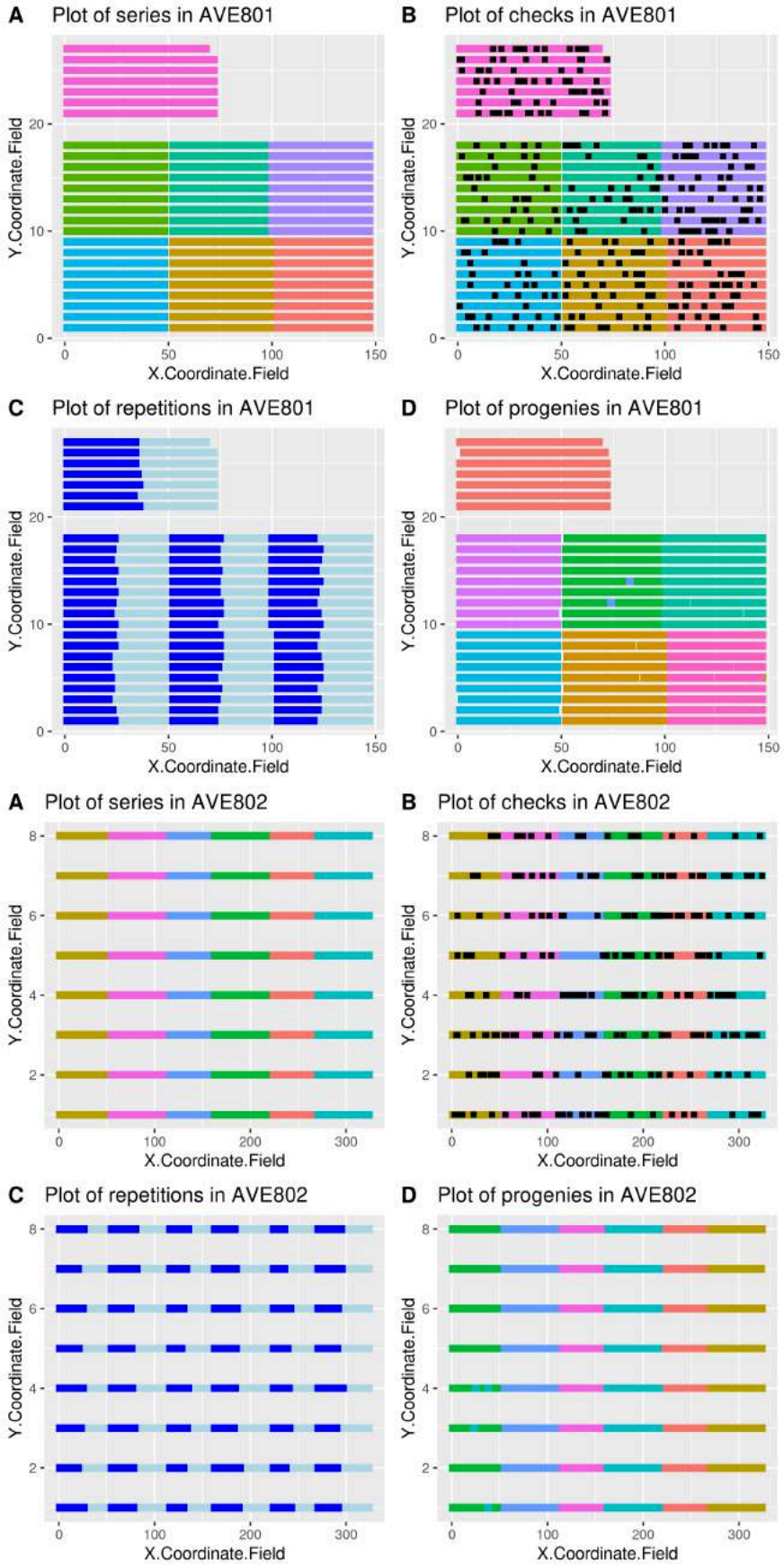


FIGURE A.5 – Dispositif expérimental de l'environnement AVE801, divisé en deux champs par une route. Les champs sont nommés AVE801 et AVE802. Pour chacun des champs : A. Répartition des individus dans le champ, chaque couleur correspondant à une série ; B. Position des témoins (en noir) dans les séries ; C. Répartition des répétitions par série dans le champ (bleu foncé pour la répétition 1, bleu clair pour la répétition 2) ; D. répartition des individus des 13 descendances dans le champ, chaque couleur correspondant à une descendance

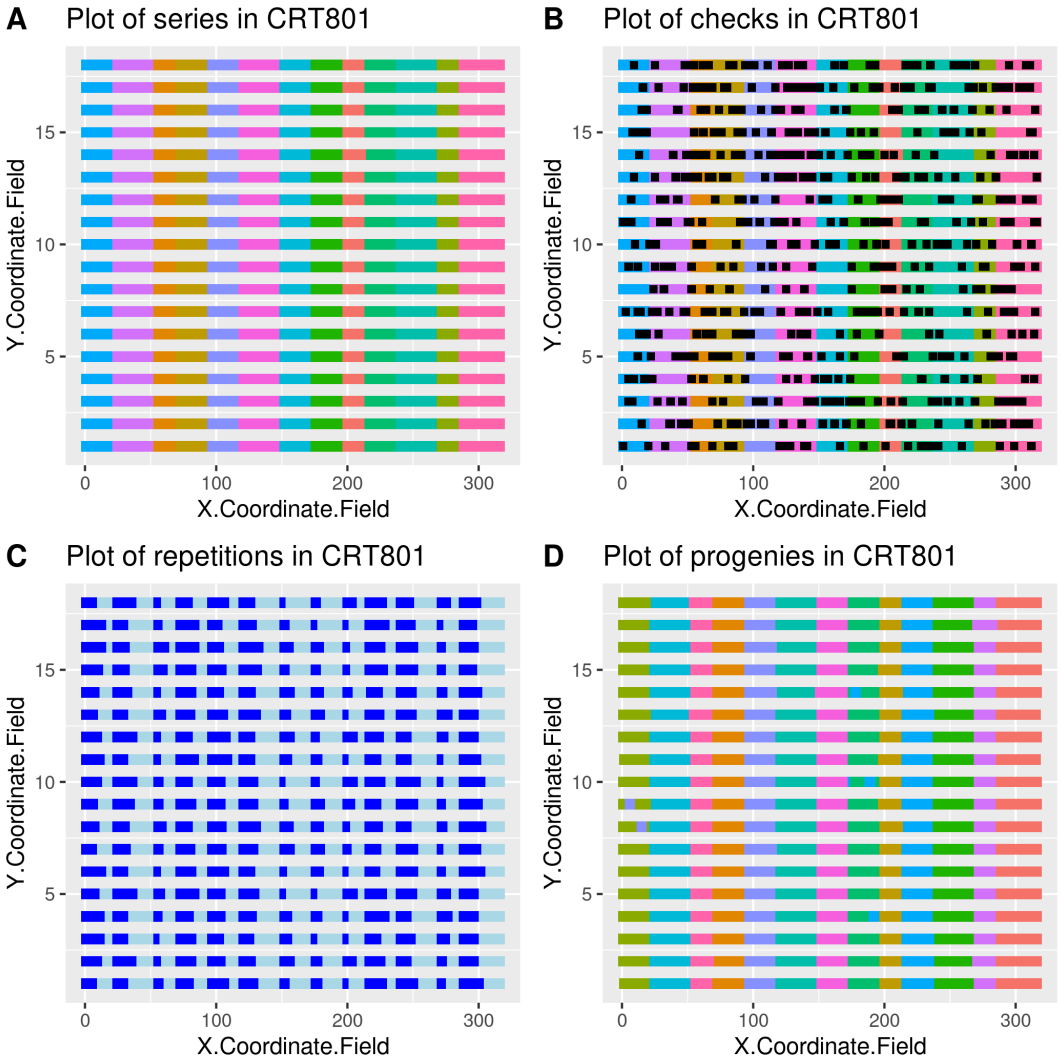
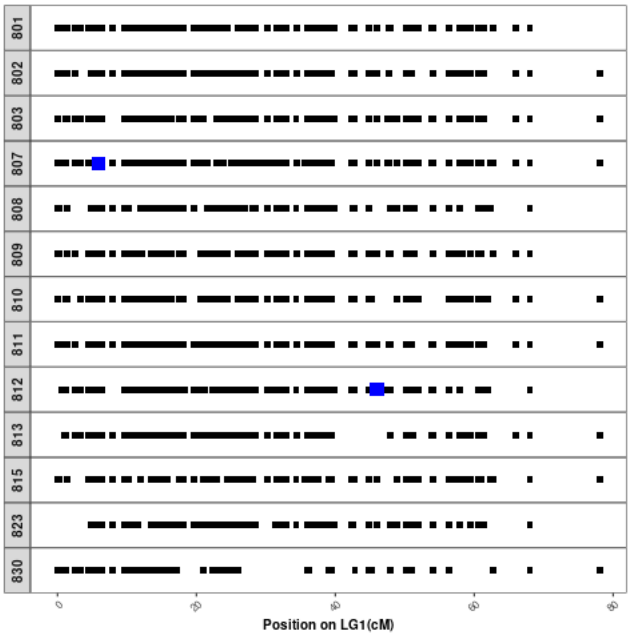
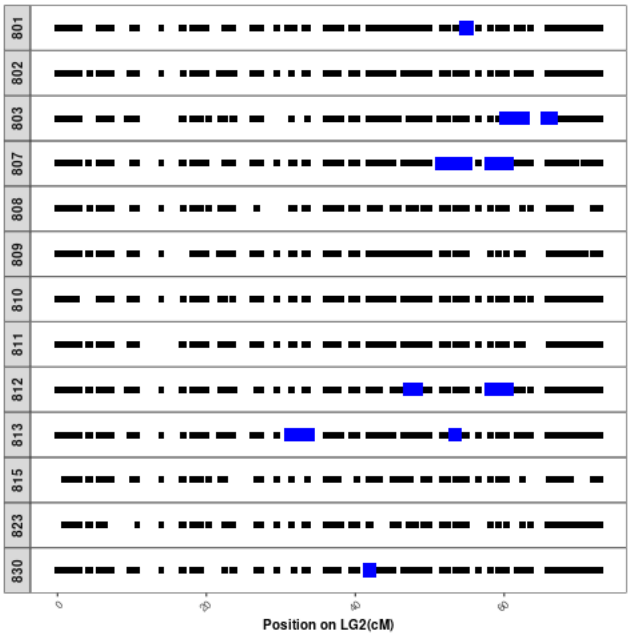


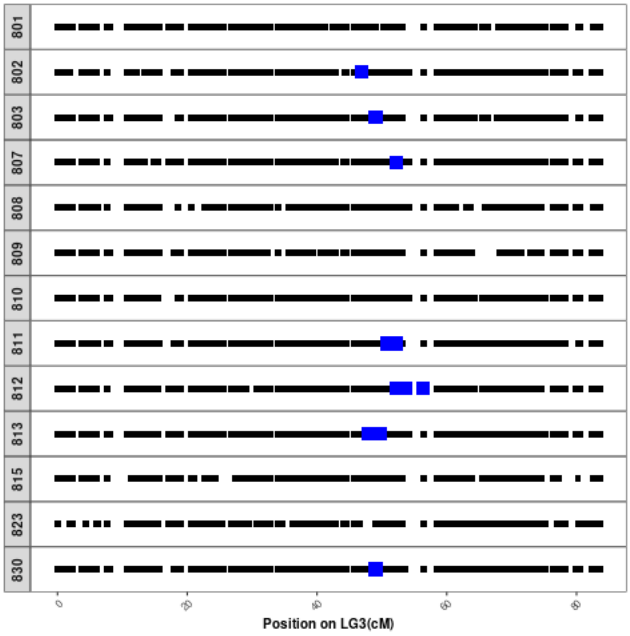
FIGURE A.6 – Dispositif expérimental de l'environnement CRT801 : A. Répartition des individus dans le champ, chaque couleur correspondant à une série ; B. Position des témoins (en noir) dans les séries ; C. Répartition des répétitions par série dans le champ (bleu foncé pour la répétition 1, bleu clair pour la répétition 2) ; D. répartition des individus des 13 descendances dans le champ, chaque couleur correspondant à une descendance

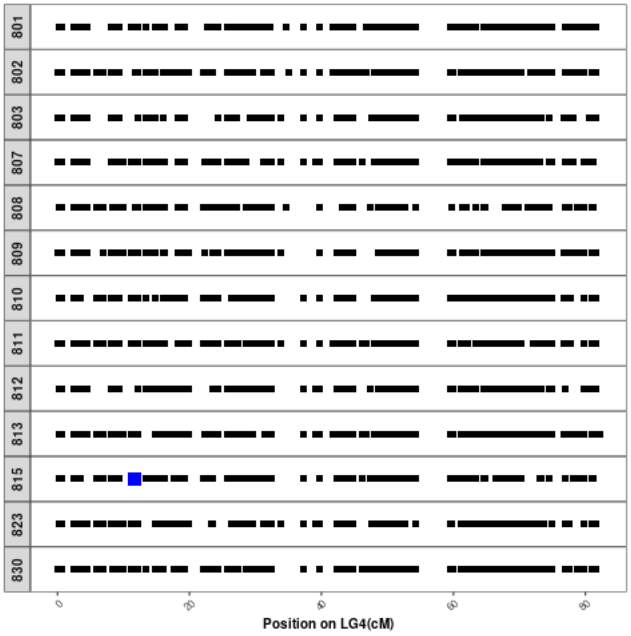
	progeny	SNP	lg	start	end
1	801	AX_124324125	2	54.93	54.93
2	802	AX_124331965	7	45.99	45.99
3	802	AX_124326121	3	46.91	46.91
4	803	AX_124324480	2	60.32	66.32
5	803	AX_124329646	6	32.39	33.78
6	803	AX_124332240	8	12.65	13.71
7	803	AX_124341522	3	49.02	49.02
8	803	AX_124328171	5	19.47	21.51
9	803	AX_124329055	5	68.11	68.11
10	807	AX_124336063	2	58.21	60.32
11	807	AX_124329106	5	36.25	38
12	807	AX_124324024	2	51.69	53.45
13	807	AX_124328249	5	25.43	25.78
14	807	AX_124341093	2	53.45	54.72
15	807	AX_124325629	3	52.19	52.19
16	807	AX_124323553	1	5.92	5.92
17	808	AX_124333010	8	53.17	53.23
18	810	AX_124333118	8	64.03	64.03
19	811	AX_124326245	3	50.78	52.19
20	812	AX_124324167	2	58.21	60.32
21	812	AX_124323948	2	47.26	48.17
22	812	AX_124332257	8	16.53	17.3
23	812	AX_124328447	5	32.78	32.78
24	812	AX_124328455	5	32.78	34.96
25	812	AX_124313705	1	45.94	45.97
26	812	AX_124315511	5	11.31	13.35
27	812	AX_124328086	5	10	10
28	812	AX_124341386	3	52.17	56.38
29	813	AX_124329747	6	42.31	53.59
30	813	AX_124324046	2	53.45	53.45
31	813	AX_124326112	3	48.04	49.72
32	813	AX_124323777	2	31.4	33.6
33	813	AX_124332257	8	16.53	17.65
34	815	AX_124313575	4	11.72	11.72
35	830	AX_124326218	3	49.02	49.02
36	830	AX_124316288	2	41.9	41.95

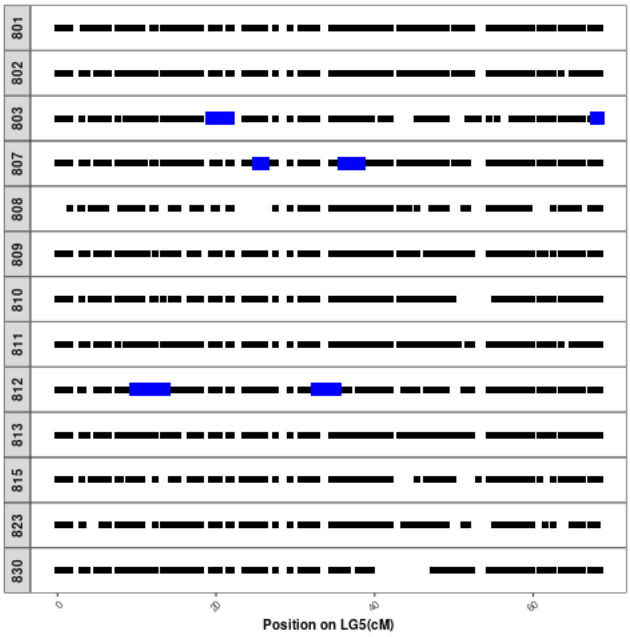
TABLE A.1 – SNPs détectés comme associés au rendement racinaire lors de l'analyse d'association et intervalle dans lequel sont situés leurs redondants.

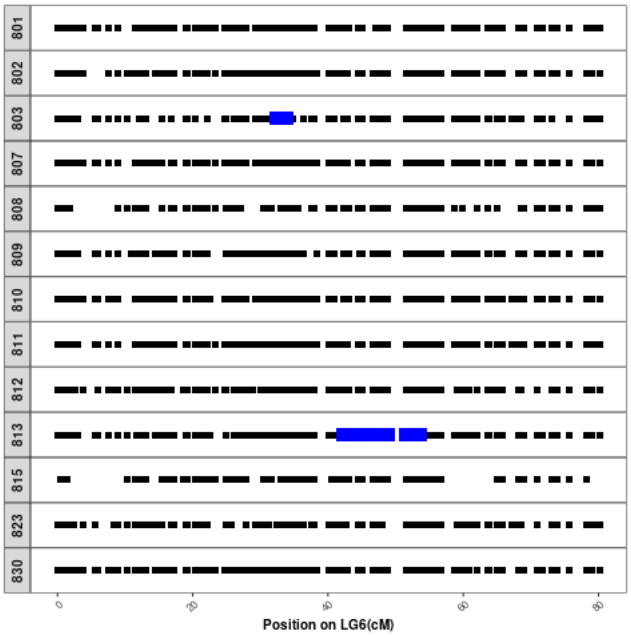

801
802
803
807
808
809
810
811
812
813
815
823
830
0
20
40
60
80
Position on LG1(cM)


801
802
803
807
808
809
810
811
812
813
815
823
830
0
20
40
60
80
Position on LG2(cM)


801
802
803
807
808
809
810
811
812
813
815
823
830
0
20
40
60
80
Position on LG3(cM)


801
802
803
807
808
809
810
811
812
813
815
823
830
0
20
40
60
80
Position on LG4(cM)


801
802
803
807
808
809
810
811
812
813
815
823
830
0
20
40
60
80
Position on LG5(cM)


801
802
803
807
808
809
810
811
812
813
815
823
830
0
20
40
60
80
Position on LG6(cM)

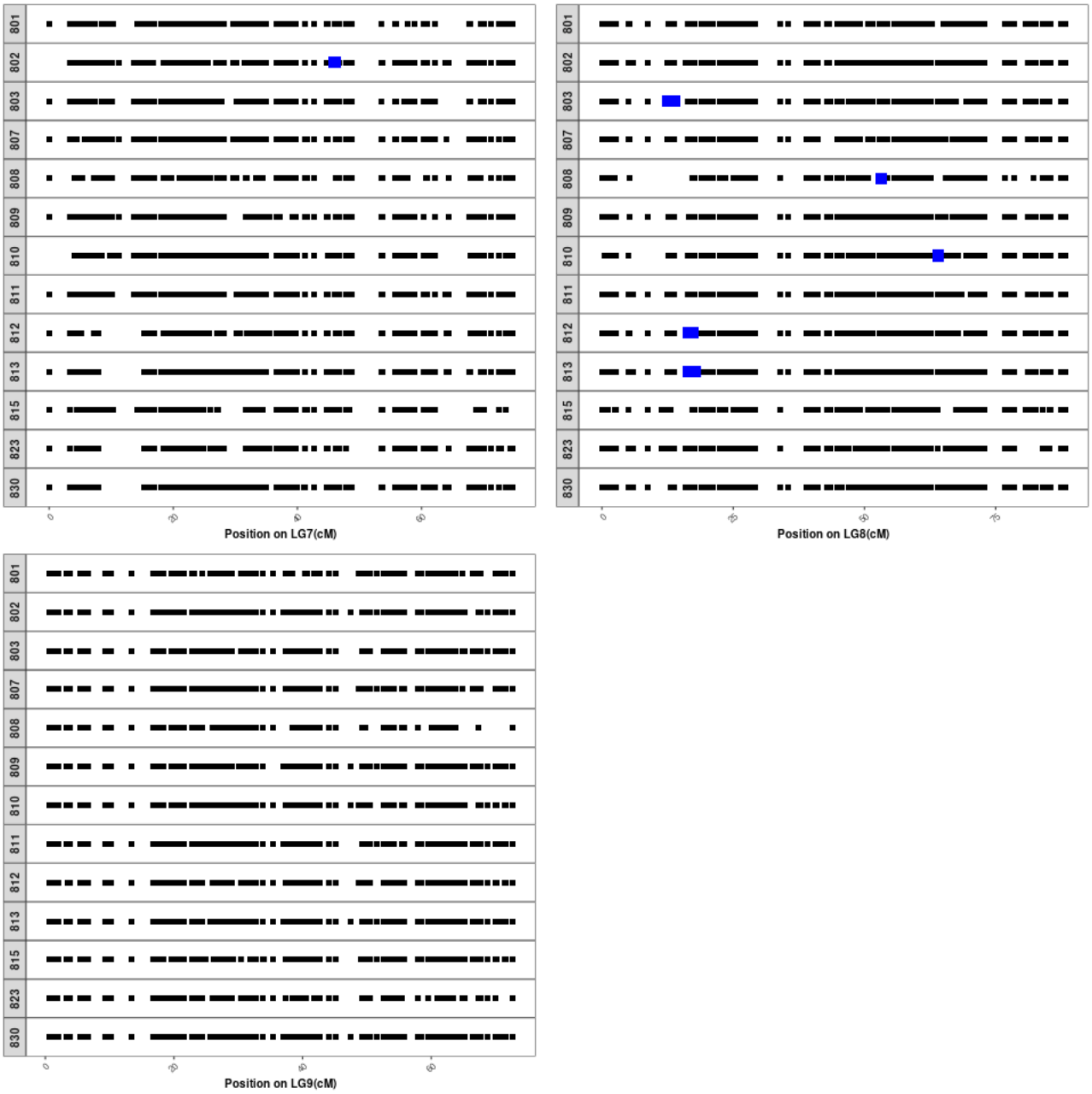


FIGURE A.7 – Ensemble des SNPs génotypés pour chacune des 13 descendances AKER étudiées pour chacun des 9 LG (un LG par figure). Les SNPs détectés comme associés au rendement racinaire lors de l'analyse d'association et leurs redondants sont représentés en bleu.

```
Selfing = function(nProgeny, name_pop_in, name_pop_out, list_pop,
   isTrackLineage, lineage_pedigree, order_sterility, nb_founders, nPlant){
# nProgeny : number of seeds created by the selfing on one individual
# nProgeny : number of seeds created by the selfing on one individual
# name_pop_in : name of the population to self
# name_pop_out : name of the population in output
# list_pop : list with all created population indexed by name
# isTrackLineage : "yes" or "no", should the founders be followed
# lineage_pedigree : lineage_pedigre eelement
# order_sterility : nb of the marker linked with the genic male sterility on chromosome 1
# nb_founders : nb of founders
# nPlant : nb of plant to self

pop = list_pop[[name_pop_in]] # load the generation to self

### with lineage ###
# select plants to self
if(isTrackLineage == "yes"){
   sterility_marker = recode_sterility_alleles(pop@geno[[1]][order_sterility,])
   pop = pop[which(sterility_marker == 1)]
   ind_founders = select_on_lineage(name_pop = name_pop_in,
      name_founders = paste0("founder_",1 :nb_founders), pedigree = lineage_pedigree,
      gender_selec = 1) # (error if not one individual per founder)
   if(nPlant == nb_founders){
      pop = pop[ind_founders]
   }else{
      completion = pop@id[which( !pop@id%in%ind_founders)]
      completion = sample(completion, (nPlant-nb_founders), replace = F)
      pop = pop[c(ind_founders, completion)]}
   all_founders = F
   nbTest = 1
   while( !all_founders){
      # cross
      pop_out = self(pop = pop, nProgeny = nProgeny)
      # format population output
      pop_out = population_format(pop=pop_out, name_pop=name_pop_out,
         rename = T, order_sterility)
```

```
      # rename parents
       S = as.data.frame(paste(pop_out@mother, pop_out@father, sep="_"),
          stringsAsFactors = F)
      colnames(S) = "mother_father"
      D = data.frame(cbind(paste(pop@mother, pop@father, sep="_"), pop@id),
          stringsAsFactors =F)
      colnames(D) = c("mother_father", "id")
      parent_names = NULL
      for(couple in unique(S[,1])){
          parent_names = c(parent_names, rep(D[which(D[,1] == couple),2],
              each=nProgeny))
      }
      pop_out@mother = parent_names
      pop_out@father = parent_names
      # recover lineage of each individual
      sterility_marker = recode_sterility_alleles(pop_out@geno[[1]][order_sterility,])
      lineage_population = lineage_recovery_from_mothers(population_name =
          name_pop_out, name_ind = pop_out@id, mothers = pop_out@mother,
          pedigree = lineage_pedigree, sterility_marker = sterility_marker)
      # check if we have one female per founder
      females = lineage_population[lineage_population[,4] == 0,]
      if(length(unique(females[,3])) == nb_founders){
          all_founders = T
      }else{
          nbTest = nbTest + 1 # redo crosses
      }}
}else{

    ### without lineage ###
    # select plants to self
    sterility_marker = recode_sterility_alleles(pop@geno[[1]][order_sterility,])
    pop = pop[which(sterility_marker == 1)]
    pop = pop[sample(pop@id, nPlant, replace = F)]
    # cross
    pop_out = self(pop = pop, nProgeny = nProgeny)
```

```
    # rename parents
    S = as.data.frame(paste(pop_out@mother, pop_out@father, sep="_"),
        stringsAsFactors =F)
    colnames(S) = "mother_father"
    D = data.frame(cbind(paste(pop@mother, pop@father, sep="_"), pop@id),
        stringsAsFactors =F)
    colnames(D) = c("mother_father", "id")
    parent_names = NULL
    for(couple in unique(S[,1])){
         parent_names = c(parent_names, rep(D[which(D[,1] == couple),2],
           each=nProgeny))
    }
    pop_out@mother = parent_names
    pop_out@father = parent_names
    # format population output
    pop_out = population_format(pop=pop_out, name_pop=name_pop_out,
        rename = T, order_sterility)
    }

### output
sterility_marker = recode_sterility_alleles(pop_out@geno[[1]][order_sterility,])
marker_sterility_proportions(sterility_marker)
return(pop_out)
}
```

```
Random_pollination = function(name_pop_in, name_pop_out, nInd,
    nSeeds, list_pop, isTrackLineage, lineage_pedigree,
    TrackLineageGender = 0, order_sterility, nb_founders, AA=T){
# prepare data to call the function random_pollination and format the output
# name_pop_in : name of the population to random pollinate
# name_pop_out : name of the population in output
# nInd : number of female to harvest
# nSeeds : number of seeds to harvest per individual
# list_pop : list with all created population indexed by name
# isTrackLineage : "yes" or "no", should the founders be followed
# lineage_pedigree : lineage_pedigree element
# TrackLineageGender : gender of progenies for whom at least one representative per
founder is required
# order_sterility : nb of the marker linked with the genic male sterility on chromosome 1
# AA : T or F, should the AA be taken as parent

pop = list_pop[[name_pop_in]] # load the generation to self
sterility_marker = recode_sterility_alleles(pop@geno[[1]][order_sterility,])

if(AA == F){ # keep only Aa as pollinators
    pop = pop[which(sterility_marker !=2)]
}

### with lineage ###
if(isTrackLineage == "yes"){ # select one female per founder for the random pollination
    and create at least one female per founder after random pollination
    candidates = select_on_lineage(name_pop = name_pop_in,
        name_founders = paste0("founder_",1 :nb_founders),pedigree = lineage_pedigree,
        gender_selec = 0)
    if(length(candidates)<nInd){
        ind_aa = pop@id[which(sterility_marker == 0 & !pop@id%in%candidates)]
        completion = sample(ind_aa, (nInd-length(candidates)), replace = F)
        candidates = c(candidates, completion)
    }
    all_founders = F
    nbTest = 1
```

```
    while( !all_founders){
      # cross
      pop_out = random_pollination(pop=pop, nInd = nInd, nSeeds = nSeeds,
        probSelf = 0, trait = 1, use="rand", female = candidates)
      # format population output
      pop_out = population_format(pop=pop_out, name_pop=name_pop_out,
        rename = T, order_sterility)
      # recover lineage of each individual
      sterility_marker = recode_sterility_alleles(pop_out@geno[[1]][order_sterility,])
      lineage_population = lineage_recovery_from_mothers(population_name =
        name_pop_out, name_ind = pop_out@id, mothers = pop_out@mother,
        pedigree = lineage_pedigree, sterility_marker = sterility_marker)
      # check if we have one female per founder
      females = lineage_population[lineage_population[,4] == TrackLineageGender,]
      if(length(unique(females[,3])) == nb_founders){
        all_founders = T
      }else{
        nbTest = nbTest + 1 # redo crosses
      }
    }
}else{

    ### without lineage : select 300 females at random ###
    pop_out = random_pollination(pop=pop, nInd = nInd, nSeeds = nSeeds,
      probSelf = 0, trait = 1, use="rand")
    # format population output
    pop_out = population_format(pop=pop_out, name_pop=name_pop_out,
      rename = T, order_sterility)
    }

# output
sterility_marker = recode_sterility_alleles(pop_out@geno[[1]][order_sterility,])
marker_sterility_proportions(sterility_marker)
return(pop_out)
}
```

```
Hybrid_production = function(name_pop_in, name_pop_out,
    name_elite, mean_pheno, nb_founders, nb_ind, list_pop, isTrackLineage,
    lineage_pedigree, map, order_sterility, list_QTLs, list_effects, h2){
# name_pop_in : name of the population from which we select mothers for hybrids
# name_pop_out : name of the population in output
# name_elite : name of the elite population
# mean_pheno : mean pheno of the population where GV effect are calculated
# nb_founders : number of founders
# nb_ind = number of individuals to cross
# list_pop : list_pop element
# isTrackLineage : isTrackLineage
# lineage_pedigree : lineage_pedigree
# map : genetic map with 3 columns LOCUS, CHROMOSOME, POSITION
# order_sterility : nb of the marker linked with the genic male sterility on chromosome 1
# list_QTLs : list with names of markers = QTLs
# list_effects : list with effect of each QTL

pop = list_pop[[name_pop_in]] # load generation with females to pollinate
sterility_marker = recode_sterility_alleles(pop@geno[[1]][order_sterility,])
elite = list_pop[[name_elite]] # load elite pollinator

### select mothers ###
# with lineage
if(isTrackLineage == "yes"){
    candidates = select_on_lineage(name_pop = name_pop_in,
        name_founders = paste0("founder_",1 :nb_founders), pedigree = lineage_pedigree,
        gender_selec = 0)
    completion = pop[which(sterility_marker == 0)]
    if(nb_ind > completion@nInd){
        pop_aa = completion
    }else{
        completion = completion[which( !completion@id%in%candidates)]
        completion = sample(completion@id, nb_ind-nb_founders, replace = F)
        candidates = c(candidates, completion)
        pop_aa = pop[candidates]
    }
```

```
}else{
   # without lineage
   pop_aa = pop[which(sterility_marker == 0)]
   if(nb_ind < pop_aa@nInd){
      candidates = sample(pop_aa@id, size = nb_ind, replace = F)
      pop_aa = pop_aa[candidates]
   }
}

# hybrid production
pop_out = hybridCross(females = pop_aa, males = elite, crossPlan = "testcross",
   returnHybridPop = F)
# format population output
pop_out = population_format(pop = pop_out, name_pop=name_pop_out,
   rename = T, order_sterility)
if(isTrackLineage == "yes"){
   # recover lineage of each individual
   sterility_marker = recode_sterility_alleles(pop_out@geno[[1]][order_sterility,])
   lineage_population = lineage_recovery_from_mothers(population_name =
      name_pop_out, name_ind = pop_out@id, mothers = pop_out@mother,
      pedigree = lineage_pedigree, sterility_marker = sterility_marker)
}

# output
sterility_marker = recode_sterility_alleles(pop_out@geno[[1]][order_sterility,])
marker_sterility_proportions(sterility_marker)
# Phenotype simulation
pop_out = Phenotype_simulation(mean_pheno, pop_out, list_pop, map, list_QTLs,
   list_effects, h2)
return(pop_out)
}
```

```
GS = function(gender_to_selec, nInd, name_pop_in, name_pop_out,
   name_pop_TS, list_pop, isTrackLineage, lineage_pedigree,
   order_sterility){
# gender_to_selec : "", "AA", "Aa", or "aa"
# nInd : number of individuals to select
# name_pop_in : name of the population where the GS is applied
# name_pop_out : name of the population in output
# name_pop_TS : vector with names of populations in the training set
# list_pop : list with all created population indexed by name
# isTrackLineage : "yes" or "no", should the founders be followed
# lineage_pedigree : lineage_pedigree element
# order_sterility : nb of the marker linked with the genic male sterility on chromosome 1

# load population to select (pop) and training populations (TS)
pop = list_pop[[name_pop_in]]

TS = list_pop[[name_pop_TS[1]]]
if(length(name_pop_TS)>1){
   for(p in 2 :length(name_pop_TS)){
      TS = c(TS, list_pop[[name_pop_TS[p]]])
   }}

# format for kin.blup
### training ###
Xa_TS = genotype_matrix(TS)
colnames(Xa_TS) = map$LOCUS
# delete pseudo markers
Xa_TS = Xa_TS[,-grep("_fav", colnames(Xa_TS))]
Xa_TS = Xa_TS[,-grep("_unf", colnames(Xa_TS))]
# recode 0 -> -1 ; 1 -> 0 ; 2 -> 1
Xa_TS[which(Xa_TS == 0)] = -1
Xa_TS[which(Xa_TS == 1)] = 0
Xa_TS[which(Xa_TS == 2)] = 1
# recover phenotype
data_TS = cbind(TS@pheno[,1], TS@id)
```

```
### test ###
Xa_test = genotype_matrix(pop)
colnames(Xa_test) = map$LOCUS
# delete pseudo markers
Xa_test = Xa_test[,-grep("_fav", colnames(Xa_test))]
Xa_test = Xa_test[,-grep("_unf", colnames(Xa_test))]
# recode 0 -> -1 ; 1 -> 0 ; 2 -> 1
Xa_test[which(Xa_test == 0)] = -1
Xa_test[which(Xa_test == 1)] = 0
Xa_test[which(Xa_test == 2)] = 1
# set phenotype to NA
data_test = cbind(rep(NA, pop@nInd), pop@id)

# merge TS and test genotype matrices
Xa = rbind(Xa_TS, Xa_test)
# merge TS and test phenotype data
data = rbind(data_TS, data_test)
data = as.data.frame(data)
colnames(data) = c("pheno", "id")
data$pheno = as.numeric(as.character(data$pheno))

# function to compute VanRanden kinship
geno = Xa
geno[which(geno == 1)] = 2
geno[which(geno == 0)] = 1
geno[which(geno == -1)] = 0
x.center<-scale(geno,center=TRUE,scale=FALSE)
KK<-x.center%*%t(x.center)
cst.VR<-sum(apply(x.center,2,var))
KK<-KK/cst.VR
K = KK

# kin.blup
eq1 = kin.blup(data=data, geno = "id", pheno = "pheno", K = K)

# prediction recovery
predicted.test = eq1$pred[which(names(eq1$pred)%in%pop@id)]
predicted.test = as.matrix(predicted.test)
```

```
# save predictions in AlphaSimR ebv slot
pop@ebv = predicted.test

# Selection on predicted sugar yield
if(isTrackLineage == "yes"){ # with lineage
   if(gender_to_selec == "Aa"){ # Aa selection
      founder_Aa = select_on_lineage_and_ebv(pop = pop,
         name_pop = name_pop_in,
         name_founders = paste0("founder_",1 :nb_founders),
         pedigree = lineage_pedigree, gender_selec = 1)
      sterility_marker = recode_sterility_alleles(pop@geno[[1]][order_sterility„])
      Aa = pop[which(sterility_marker == 1)]
      if(nInd>nb_founders){
         pop_completion = selectInd(Aa[which( !Aa@id %in% founder_Aa)],
            nInd = nInd-nb_founders, trait = 1, use = "ebv", selecTop = T,
            returnPop = T)
         pop_ebv = c(pop_completion, pop[which(pop@id %in% founder_Aa)])
      }else{
         pop_ebv = pop[which(pop@id %in% founder_Aa)]
      }

   }else if(gender_to_selec == "AA"){ # AA selection
      founder_AA = select_on_lineage_and_ebv(pop = pop,
         name_pop = name_pop_in,
         name_founders = paste0("founder_",1 :nb_founders),
         pedigree = lineage_pedigree, gender_selec = 2)
      sterility_marker = recode_sterility_alleles(pop@geno[[1]][order_sterility„])
      AA = pop[which(sterility_marker == 2)]
      if(nInd>nb_founders){
         pop_completion = selectInd(AA[which( !AA@id %in% founder_AA)],
            nInd = nInd-nb_founders, trait = 1, use = "ebv", selecTop = T,
            returnPop = T)
         pop_ebv = c(pop_completion, pop[which(pop@id %in% founder_AA)])
      }else{
         pop_ebv = pop[which(pop@id %in% founder_AA)]
      }
```

```
    }else if(gender_to_selec == "aa"){ # aa selection
        founder_aa = select_on_lineage_and_ebv(pop = pop,
            name_pop = name_pop_in,
            name_founders = paste0("founder_",1 :nb_founders),
            pedigree = lineage_pedigree, gender_selec = 0)
        sterility_marker = recode_sterility_alleles(pop@geno[[1]][order_sterility,])
        aa = pop[which(sterility_marker == 0)]
        if(nInd>nb_founders){
            pop_completion = selectInd(aa[which( !aa@id %in% founder_aa)],
                nInd = nInd-nb_founders, trait = 1, use = "ebv", selecTop = T,
                returnPop = T)
            pop_ebv = c(pop_completion, pop[which(pop@id %in% founder_aa)])
        }else{
            pop_ebv = pop[which(pop@id %in% founder_aa)]
        }

    }else{ # AA, Aa or aa selection
        founder_aa = select_on_lineage_and_ebv(pop = pop,
            name_pop = name_pop_in,
            name_founders = paste0("founder_",1 :nb_founders),
            pedigree = lineage_pedigree, gender_selec = 0)
        if(nInd>nb_founders){
            pop_completion = selectInd(pop[which( !pop@id %in% founder_aa)],
                nInd = nInd-nb_founders, trait = 1, use="ebv", selecTop = T,
                returnPop = T)
            pop_ebv = c(pop_completion, pop[which(pop@id %in% founder_aa)])
        }else{
            pop_ebv = pop[which(pop@id %in% founder_aa)]
        }
    }

}else{ # without lineage
    if(gender_to_selec == "aa"){ # aa selection
        females = pop[which(pop@gender == "F")]
        pop_ebv = selectInd(females, nInd = nInd, trait = 1, use = "ebv", selecTop = T,
            returnPop = T)
```

```
    }else if(gender_to_selec == "Aa"){ # Aa selection
        sterility_marker = recode_sterility_alleles(pop@geno[[1]][order_sterility,])
        Aa = pop[which(sterility_marker == 1)]
        pop_ebv = selectInd(Aa, nInd = nInd, trait = 1, use = "ebv", selecTop = T,
            returnPop = T)

    }else if(gender_to_selec == "AA"){ # AA selection
        sterility_marker = recode_sterility_alleles(pop@geno[[1]][order_sterility,])
        AA = pop[which(sterility_marker == 2)]
        pop_ebv = selectInd(AA, nInd = nb_founders*4, trait = 1, use = "ebv",
            selecTop = T, returnPop = T)

    }else{ # AA, Aa or aa selection
        pop_ebv = selectInd(pop, nInd = nInd, trait = 1, use="ebv", selecTop = T,
            returnPop = T)
    }
}

# output
sterility_marker = recode_sterility_alleles(pop_ebv@geno[[1]][order_sterility,])
marker_sterility_proportions(sterility_marker)
pop_ebv = population_format(pop=pop_ebv,
    rename = F, name_pop=name_pop_out, order_sterility)
# Phenotype simulation
pop_ebv = Phenotype_simulation(mean_pheno, pop_ebv, list_pop, map, list_QTLs,
list_effects, h2)
return(pop_ebv)
}
```

```
PS = function(selection_rate = 20, name_pop_in,
   name_pop_hybrid_mothers, name_pop_hybrid_mother_parents,
   name_pop_out, list_pop, isTrackLineage, lineage_pedigree,
   order_sterility, map, list_QTLs, list_effects, h2){
# selection_rate : percentage of selfing parents to select (default = 20%)
# name_pop_in : name of the hybrid population with simulated phenotype
# name_pop_hybrid_mothers : name of the population with mothers of hybrids
   (= after selfing)
# name_pop_hybrid_mother_parents : name of the population with parents of
   hybrid mothers (= before selfing)
# name_pop_out : name of the population in output
# list_pop : list with all created population indexed by name
# isTrackLineage : "yes" or "no", should the founders be followed
# lineage_pedigree : lineage_pedigree element
# order_sterility : nb of the marker linked with the genic male sterility on chromosome 1
# map : genetic map with 3 columns LOCUS, CHROMOSOME, POSITION
# list_QTLs : list with names of markers = QTLs
# list_effects : list with effect of each QTL
# h2 : heritability

pop = list_pop[[name_pop_in]] # load the generation to select

# recover phenotype value for hybrid mothers
pheno_mothers = data.frame(pop@mother, pop@pheno, stringsAsFactors = F)
colnames(pheno_mothers) = c("mother", "pheno")
# recover phenotype value for parents of hybrid mothers
mothers = list_pop[[name_pop_hybrid_mothers]]
mothers_parents= data.frame(mothers@id, mothers@mother, stringsAsFactors = F)
colnames(mothers_parents) = c("mother", "parent")
pheno_parents = left_join(pheno_mothers, mothers_parents, by="mother")
parents = list_pop[[name_pop_hybrid_mother_parents]]
test = sapply(parents@id, function(x){
if(x %in% pheno_parents[,3]){
   return(mean(as.numeric(pheno_parents[which(pheno_parents[,3] == x),2])))
}else{
   return(0)}})
parents@pheno = as.matrix(test)
```

```
# phenotypic selection
pop = parents[which(test !=0)] # only parent with a phenotypic value
nInd = ceiling(pop@nInd*selection_rate/100)

# with lineage
if(isTrackLineage == "yes"){
   founder_Aa = select_on_lineage_and_pheno(pop = pop,
      name_pop = name_pop_hybrid_mother_parents,
      name_founders = paste0("founder_",1 :nb_founders),
      pedigree = lineage_pedigree, gender_selec = 1)
   if(pop@nInd > nInd){
      pop_completion = selectInd(pop[which( !pop@id %in% founder_Aa)],
         nInd = (nInd-length(founder_Aa)), trait = 1, use = "pheno", selecTop = T,
         returnPop = T)
      selec = c(pop[founder_Aa], pop_completion)
   }else{
      selec = selectInd(pop[which( !pop@id %in% founder_Aa)],
         nInd = pop@nInd-length(founder_Aa), trait = 1, use = "pheno", selecTop = T,
         returnPop = T)
   }
}else{

   # without lineage
   selec = selectInd(pop, nInd = nInd, trait = 1, use = "pheno", selecTop = T,
      returnPop = T)
}

pop_out = mothers
pop_out = pop_out[which(pop_out@mother%in%selec@id)]

# format population output
pop_out = population_format(pop = pop_out, name_pop=name_pop_out,
   rename = T, order_sterility)
if(isTrackLineage == "yes"){ # with lineage
   # recover lineage of each individual
   sterility_marker = recode_sterility_alleles(pop_out@geno[[1]][order_sterility,])
```

```
    lineage_population = lineage_recovery_from_mothers(population_name =
        name_pop_out, name_ind = pop_out@id, mothers = pop_out@mother,
        pedigree = lineage_pedigree, sterility_marker = sterility_marker)
}
sterility_marker = recode_sterility_alleles(pop_out@geno[[1]][order_sterility,])
marker_sterility_proportions(sterility_marker)
pop_out = Phenotype_simulation(mean_pheno, pop_out, list_pop, map, list_QTLs,
    list_effects, h2)
return(pop_out)
}
```

```
GV = function(gender_to_selec, nInd, name_pop_in, name_pop_out,
   list_pop, isTrackLineage, lineage_pedigree, order_sterility){
 # gender_to_selec : "", "AA", "Aa", or "aa"
# nInd : number of individuals to select
# name_pop_in : name of the population where the GV is applied
# name_pop_out : name of the population in output
# list_pop : list with all created population indexed by name
# isTrackLineage : "yes" or "no", should the founders be followed
# lineage_pedigree : lineage_pedigree element
# order_sterility : nb of the marker linked with the genic male sterility on chromosome 1

pop = list_pop[[name_pop_in]] # load population to select

# genetic value computation
Xa = genotype_matrix(pop)
colnames(Xa) = map$LOCUS
GV = compute_GV(Xa, qtls = list_QTLs, effects = list_effects)
GV = GV*(as.numeric(h2/var(GV)))**0.5
names(GV) = row.names(Xa)
GV = as.matrix(GV)
pop@gv = GV

# Selection on genetic value
if(isTrackLineage == "yes"){ # with lineage
   if(gender_to_selec == "Aa"){ # Aa selection
      founder_Aa = select_on_lineage_and_gv(pop = pop,name_pop = name_pop_in,
         name_founders = paste0("founder_",1 :nb_founders),
         pedigree = lineage_pedigree, gender_selec = 1)
      sterility_marker = recode_sterility_alleles(pop@geno[[1]][order_sterility,])
      Aa = pop[which(sterility_marker == 1)]
      if(nInd>nb_founders){
         pop_completion = selectInd(Aa[which( !Aa@id %in% founder_Aa)],
            nInd = nInd-nb_founders, trait = 1, use = "gv", selecTop = T,
            returnPop = T)
         pop_gv = c(pop_completion, pop[which(pop@id %in% founder_Aa)])
```

```
    }else{
      pop_gv = pop[founder_Aa]
    }

  }else if(gender_to_selec == "AA"){ # AA selection
    founder_AA = select_on_lineage_and_gv(pop = pop,
      name_pop = name_pop_in,
      name_founders = paste0("founder_",1 :nb_founders),
      pedigree = lineage_pedigree, gender_selec = 2)
    sterility_marker = recode_sterility_alleles(pop@geno[[1]][order_sterility,])
    AA = pop[which(sterility_marker == 2)]
    if(nInd>nb_founders){
      pop_completion = selectInd(AA[which( !AA@id %in% founder_AA)],
        nInd = nInd-nb_founders, trait = 1, use = "gv", selecTop = T,
        returnPop = T)
      pop_gv = c(pop_completion, pop[which(pop@id %in% founder_AA)])
    }else{
      pop_gv = pop[founder_AA]
    }

  }else if(gender_to_selec == "aa"){ # aa selection
    founder_aa = select_on_lineage_and_gv(pop = pop,name_pop = name_pop_in,
      name_founders = paste0("founder_",1 :nb_founders),
      pedigree = lineage_pedigree, gender_selec = 0)
    sterility_marker = recode_sterility_alleles(pop@geno[[1]][order_sterility,])
    aa = pop[which(sterility_marker == 0)]
    if(nInd>nb_founders){
      pop_completion = selectInd(aa[which( !aa@id %in% founder_aa)],
        nInd = nInd-nb_founders, trait = 1, use = "gv", selecTop = T,
        returnPop = T)
      pop_gv = c(pop_completion, pop[which(pop@id %in% founder_aa)])
    }else{
      pop_gv = pop[founder_aa]
    }
```

```
  }else{ # AA, Aa or aa selection
    founder_aa = select_on_lineage_and_gv(pop = pop,name_pop = name_pop_in,
      name_founders = paste0("founder_",1 :nb_founders),
      pedigree = lineage_pedigree, gender_selec = 0)
    if(nInd>nb_founders){
      pop_completion = selectInd(pop[which( !pop@id %in% founder_aa)],
        nInd = nInd-nb_founders, trait = 1, use="gv", selecTop = T,
          returnPop = T)
      pop_gv = c(pop_completion, pop[which(pop@id %in% founder_aa)])
    }else{
      pop_gv = pop[founder_aa]
    }}

  }else{ # without lineage
    if(gender_to_selec == "aa"){ # aa selection
      females = pop[which(pop@gender == "F")]
      pop_gv = selectInd(females, nInd = nInd, trait = 1, use = "gv", selecTop = T,
        returnPop = T)

     }else if(gender_to_selec == "Aa"){ # Aa selection
      sterility_marker = recode_sterility_alleles(pop@geno[[1]][order_sterility,])
      Aa = pop[which(sterility_marker == 1)]
      pop_gv = selectInd(Aa, nInd = nInd, trait = 1, use = "gv", selecTop = T,
        returnPop = T)

    }else if(gender_to_selec == "AA"){ # AA selection
      sterility_marker = recode_sterility_alleles(pop@geno[[1]][order_sterility,])
      AA = pop[which(sterility_marker == 2)]
      pop_gv = selectInd(AA, nInd = nb_founders*4, trait = 1, use = "gv",
        selecTop = T, returnPop = T)

    }else{ # AA, Aa or aa selection
      pop_gv = selectInd(pop, nInd = nInd, trait = 1, use="gv", selecTop = T,
        returnPop = T)
    }}
```

```
# output
sterility_marker = recode_sterility_alleles(pop_gv@geno[[1]][order_sterility,])
marker_sterility_proportions(sterility_marker)
pop_gv = population_format(pop=pop_gv, rename = F, name_pop=name_pop_out,
   order_sterility)
pop_gv = Phenotype_simulation(mean_pheno, pop_gv, list_pop, map, list_QTLs,
      list_effects, h2)
return(pop_gv)
}
```

FIGURE A.8 – Codes des actions *Selfing*, *Random_pollination*, *Hybrid_production*, *GS*, *PS*, *GV*.

```
select_on_lineage = function(name_pop, pedigree,
  gender_selec, name_founders)
# name_pop = population from which we want select several individuals
# pedigree = matrix with 4 columns : population, individual, founder, sterility marker
  and with n rows = number of individuals in the scheme
# gender_selec # 0 for aa, 1 for Aa, 2 for AA
# name_founders = vector with names of founders

population_format = function(pop, rename = T, name_pop, order_sterility)
# pop = population to format
# rename = T or F, default = T, F if we want to rename individuals in the population
# name_pop = name of the population to format
# order_sterility = n of the sterility marker on the chromosome 1

recode_sterility_alleles = function(sterility_marker)
# sterility_marker = data frame with 2 rows and one column per marker on
  chromosome 1
# 1st row = first haplotype and 2nd row = second haplotype

lineage_recovery_from_mothers = function(population_name, name_ind,
mothers, pedigree, sterility_marker)
# population_name = name of the population
# name_ind = vector (character with id of individuals)
# mothers = vector (characters) with id of population mothers
# pedigree = matrix with 4 columns : population, individual, founder, sterility marker
  and with n rows = number of individuals in the scheme
# sterility_marker for the population
# output : lineage pedigree of the population
```

FIGURE A.9 – Paramètres d'entrée des fonctions d'enrobage select_on_lineage, population_format, recode_sterility_allele, lineage_recovery_from_mother.

génération	30 → 30	30 → 300	300 → 300
founders	30	30	300
G1_S	300	300	3000
G2_D	300	300	3000
G3_D	300	300	3000
G4_D	300	300	3000
G5_S	300	300	3000
G6_P	30	30	300
G6_D	600	6000	6000
sélection de 20%			
G7_GS	120	1200	1200
G7_D	600	6000	6000
G7_P	30	300	300
G8_D	600 → 300(Aa)	6000 → 3000 (Aa)	6000 → 3000 (Aa)
sélection de 20%			
G9_GS	60	600	600
G9_S	600	6000	6000
G10_D	300	3000	3000
G11_P	30	300	300
G11_D	600	6000	6000
sélection de 5%			
G12_GS	30	300	300

TABLE A.2 – Effectif de chaque génération produite dans les scénarios de GS : 30 fondateurs 30 individus dans la population de pre-breeding finale (30 → 30), 30 fondateurs 300 individus dans la population de pre-breeding finale (30 → 300), 300 fondateurs 300 individus dans la population de pre-breeding finale (300 → 300). « S » désigne une action d'autofécondation, « D » une action de pollinisation aléatoire, « P » une génération d'hybrides phénotypés, « GS » une action de sélection génomique. (Aa) désigne les plantes mâles fertiles hétérozygotes au marqueur de stérilité.

génération	30 → 30	30 → 300	300 → 300
founders	30	30	300
G1_S	300	300	3000
G2_D	300	300	3000
G3_D	300	300	3000
G4_D	300	300	3000
G5_S	300	300	3000
G6_D	600	6000	6000
sélection de 20%			
G7_GV	120	1200	1200
G7_D	600	6000	6000
G8_D	600 → 300(Aa)	6000 → 3000 (Aa)	6000 → 3000 (Aa)
sélection de 20%			
G9_GV	60	600	600
G9_S	600	6000	6000
G10_D	300	3000	3000
G11_D	600	6000	6000
sélection de 5%			
G12_GV	30	300	300

TABLE A.3 – Effectif de chaque génération produite dans les scénarios de GV : 30 fondateurs 30 individus dans la population de pre-breeding finale (30 → 30), 30 fondateurs 300 individus dans la population de pre-breeding finale (30 → 300), 300 fondateurs 300 individus dans la population de pre-breeding finale (300 → 300). « S » désigne une action d'autofécondation, « D » une action de pollinisation aléatoire, « GV » une action de sélection sur la vraie valeur génétique. (Aa) désigne les plantes mâles fertiles hétérozygotes au marqueur de stérilité.

génération	30 → 30	30 → 300	300 → 300
founders	30	30	300
G1_S	300	300	3000
G2_D	300	300	3000
G3_D	300 → 150 (Aa) → 120 max	300 → 150 (Aa) → 120 max	3000 → 1500 (Aa) → 1200 max
G4_S	2400 → 600 (aa)	2400 → 600 (aa)	24 000 → 6000 (aa)
G6_P	600	6000	6000
sélection de 20% des G3_D			
récupération de 5 descendants issus de l'autofécodantion de chaque G3_D sélectionné			
G7_PS	120	1200	1200
G7_D	600	6000	6000
G8_D	300 → 150 (Aa) → 120 max	3000 → 1500 (Aa) → 1200 max	3000 → 1500 (Aa) → 1200 max
G9_S	2400 → 600 (aa)	24 000 → 6000 (aa)	24 000 → 6000 (aa)
G11_P	600	6000	6000
sélection de 20% des G8_D			
récupération de 2.5 descendants issus de l'autofécodantion de chaque G8_D sélectionné			
G12_PS	60	600	600
G12_D	600	6000	6000
G13_D	300 → 150 (Aa) → 120 max	3000 → 1500 (Aa) → 1200 max	3000 → 1500 (Aa) → 1200 max
G14_S	2400 → 600 (aa)	24 000 → 6000 (aa)	24 000 → 6000 (aa)
G16_P	600	6000	6000
sélection de 5% des G13_D			
récupération de 5 descendants issus de l'autofécodantion de chaque G13_D sélectionné			
G17_PS	30	300	300

TABLE A.4 – Effectif de chaque génération produite dans les scénarios de PS : 30 fondateurs 30 individus dans la population de pre-breeding finale (30 → 30), 30 fondateurs 300 individus dans la population de pre-breeding finale (30 → 300), 300 fondateurs 300 individus dans la population de pre-breeding finale (300 → 300). « S » désigne une action d'autofécondation, « D » une action de pollinisation aléatoire, « P » une génération d'hybrides phénotypés, « PS » une action de sélection phénotypique. (Aa) désigne les plantes mâles fertiles hétérozygotes au marqueur de stérilité, (aa) les plantes mâles stériles. Une contrainte est données par les sélectionneurs : 1200 plantes maximum peuvent être évaluées en sélection phénotypique. Le scénario 30 → 30 contient 10 fois moins d'invidus que le scénario 300 → 300 pour une même génération, 120 plantes (Aa) au maximum sont donc évaluées en sélection phénotypique dans ce scénario.

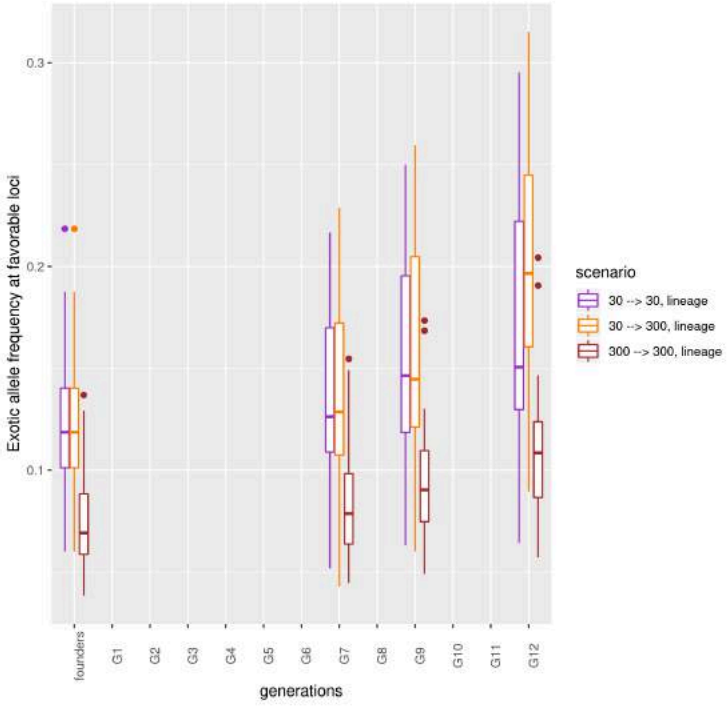


FIGURE A.10 – Evolution de la fréquence de l'allèle exotique aux locus « favorables » dans trois scénarios, 30 → 30 avec suivi du lignage maternel en violet ; 30 → 300 avec suivi du lignage maternel en orange ; 300 → 300 avec suivi du lignage maternel en marron.

	30 → 30	30 → 300	300 → 300
founders	0.12	0.12	0.08
G7	0.14	0.14	0.08
G9	0.15	0.16	0.10
G12	0.17	0.20	0.11

TABLE A.5 – Evolution de la fréquence de l'allèle exotique aux locus « favorables » dans trois scénarios, 30 → 30, 30 → 300 et 300 → 300 avec suivi du lignage maternel.

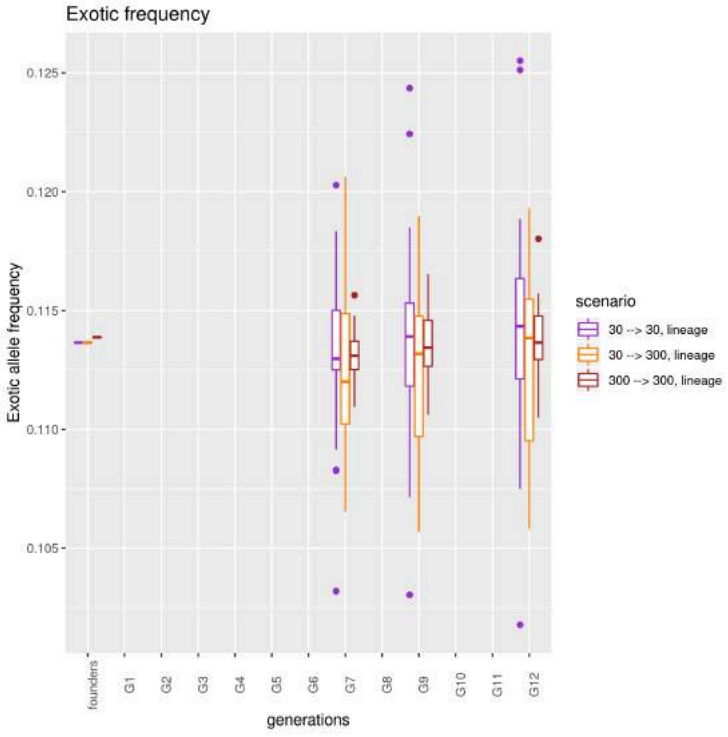


FIGURE A.11 – Evolution de la fréquence de l'allèle exotique sur l'ensemble des locus dans trois scénarios avec suivi du lignage maternel, 30 → 30 en violet ; 30 → 300 en orange ; 300 → 300 en marron.

	30 → 30	30 → 300	300 → 300
founders	0.11	0.11	0.11
G7	0.11	0.11	0.11
G9	0.11	0.11	0.11
G12	0.11	0.11	0.11

TABLE A.6 – Evolution de la fréquence de l'allèle exotique sur l'ensemble des locus dans trois scénarios avec suivi du lignage maternel, 30 → 30, 30 → 300 et 300 → 300.

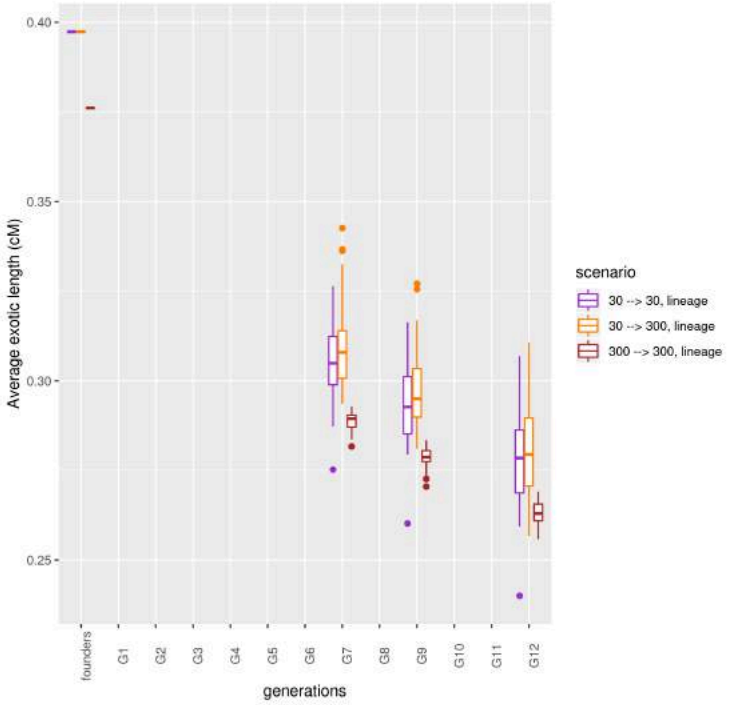


FIGURE A.12 – Evolution de la longueur moyenne des fragments exotiques lors de trois scénarios avec suivi du lignage maternel, 30 → 30 en violet ; 30 → 300 en orange ; 300 → 300 en marron.

	30 → 30	30 → 300	300 → 300
founders	0.40	0.40	0.38
G7	0.30	0.31	0.29
G9	0.29	0.30	0.28
G12	0.28	0.28	0.26

TABLE A.7 – Evolution de la longueur moyenne des fragments exotiques lors de trois scénarios avec suivi du lignage maternel, 30 → 30, 30 → 300 et 300 → 300.

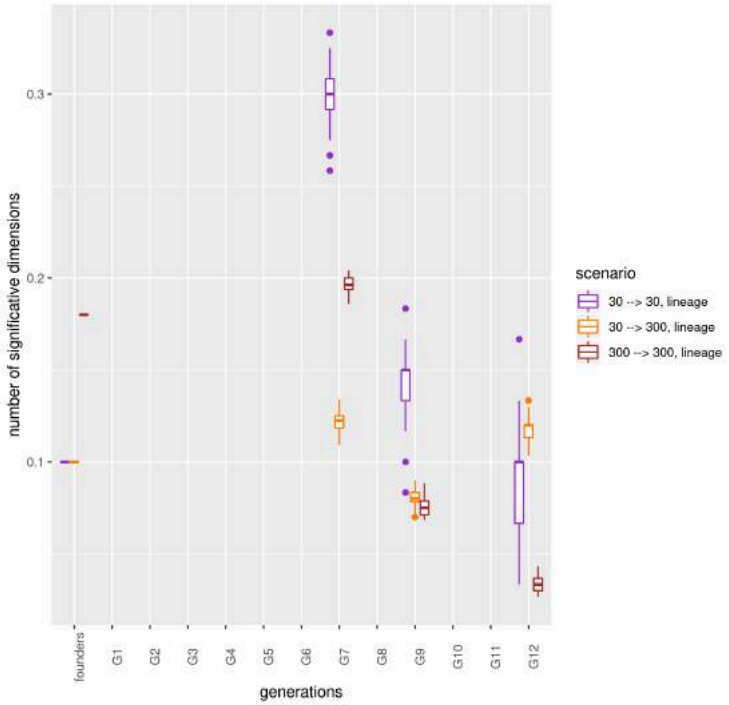


FIGURE A.13 – Evolution du nombre de dimensions significatives de l'ACP rapporté au nombre d'individus lors de trois scénarios avec suivi du lignage maternel, 30 → 30 en violet ; 30 → 300 en orange ; 300 → 300 en marron.

	30 → 30	30 → 300	300 → 300
founders	0.10	0.10	0.18
G7	0.30	0.12	0.20
G9	0.14	0.08	0.08
G12	0.09	0.12	0.03

TABLE A.8 – Evolution du nombre de dimensions significatives de l'ACP rapporté au nombre d'individus lors de trois scénarios avec suivi du lignage maternel, 30 → 30, 30 → 300 et 300 → 300.

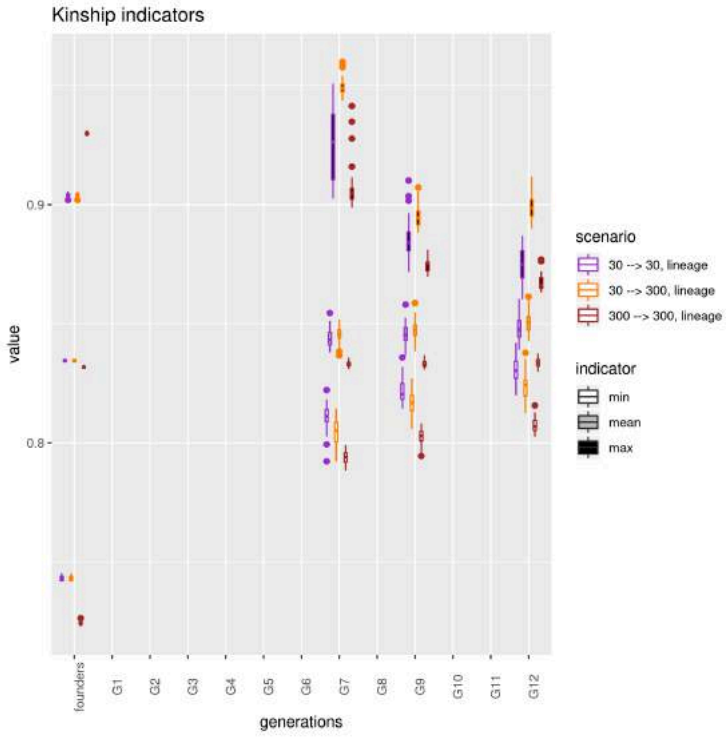


FIGURE A.14 – Evolution de la valeur minimale (en blanc), moyenne (en gris) et maximale (en noir) de la matrice de kinship lors de trois scénarios avec suivi du lignage maternel, 30 → 30 en violet ; 30 → 300 en orange ; 300 → 300 en marron.

	min		
	30 → 30	30 → 300	300 → 300
founders	0.74	0.74	0.72
G7	0.81	0.81	0.79
G9	0.83	0.82	0.80
G12	0.84	0.83	0.81
	mean		
	30 → 30	30 → 300	300 → 300
founders	0.83	0.83	0.83
G7	0.85	0.85	0.83
G9	0.85	0.85	0.83
G12	0.85	0.86	0.83
	max		
	30 → 30	30 → 300	300 → 300
founders	0.90	0.90	0.93
G7	0.92	0.95	0.91
G9	0.89	0.90	0.88
G12	0.89	0.90	0.87

TABLE A.9 – Evolution du nombre de la valeur minimale, moyenne et maximale de la matrice de kinship lors de trois scénarios avec suivi du lignage maternel, 30 → 30, 30 → 300 et 300 → 300.