Détection de gènes coadaptés par analyse pangénomique de signatures de sélection épistatique : application chez la légumineuse modèle Medicago truncatula

Léa Boyrie

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'Université Toulouse 3 - Paul Sabatier

Présentée et soutenue par

Léa BOYRIE

Le 19 octobre 2020

Détection de gènes coadaptés par analyse pangénomique de signatures de sélection épistatique: application chez la légumineuse modèle Medicago truncatula

---

Ecole doctorale : **SEVAB - Sciences Ecologiques, Vétérinaires, Agronomiques et Bioingenieries**

Spécialité : **Interactions plantes-microorganismes**

Unité de recherche :
LRSV - Laboratoire de Recherche en Sciences Végétales

Thèse dirigée par
Christophe JACQUET et Maxime BONHOMME

Jury

M. Mathieu GAUTIER, Rapporteur
Mme Maud TENAILLON, Rapporteure
Mme Joëlle RONFORT, Examinatrice
M. Maxime BONHOMME, Directeur de thèse
M. Christophe JACQUET, Co-directeur de thèse

Abstract

Adaptation by natural selection is central to the evolution of species. By targeting differences in survival and / or reproduction of individuals according to changes in the environment, selection filters genetic variants in populations. Extremely conserved genes are subjected to purifying selection which eliminates deleterious mutations, while other more polymorphic genes will carry positively selected mutations in a certain environmental context. For more than 20 years, modeling in population genetics and the emergence of sequencing technologies allowing the identification of genetic variants at the genome level (e.g. Single Nucleotide Polymorphisms - SNP) have allowed the development of numerous statistical methods analyzing polymorphism to identify genes or regions of the genome presenting selection signatures, while taking into account the other evolutionary forces (genetic drift, gene flow) influencing this polymorphism. However, they test the selection hypothesis independently at each locus and do not allow the interaction between the alleles of genes to be explored as a potential target for epistatic selection. Still, quantitative genetics and modern biology unquestionably show that genes are not functionally independent entities, but that they are interacting elements in larger networks allowing the expression of biological characteristics. The purpose of this thesis is to propose a new statistical test which makes it possible to identify epistatic selection signatures, therefore coadapted genes, on the basis of linkage disequilibrium (DL) using SNP markers. In the first part, we describe the proposed statistics, T_{r_v} and T_{corPC1_v}, which compare pairs of SNPs or genomic regions. They are based on recent work showing that the correlation coefficient (r), strongly influenced by the genetic structure of populations and the degree of genetic similarity between individuals, must be corrected (r_v) by the relationship matrix between individuals (V). Coupled with intensive calculations, simulations of genome-wide SNP data in structured populations have made it possible to demonstrate that T_{r_v} and T_{corPC1_v} follow a Student distribution $\tau_{(n-2)}$, greatly reduce noisy DL background generated by non-selective evolutionary forces, and show good detection power. In a second part, we use the "Genome-Wide Epistatic Selection Scan" (GWESS) approach in the model plant *Medicago truncatula*, where a candidate gene is used as bait to calculate its correlation with all the other genes in the genome. Following the identification of an epistatic selection signature between *MtSUNN* and *MtCLE02*, coding respectively for a receptor and a signaling peptide, a proof of concept is provided by the experimental demonstration (collaboration) that *MtCLE02* has a *MtSUNN*-dependent negative role on

nodulation. The GWESS approach applied to SNP data in humans shows an epistatic selection signature between the *SLC24A5* and *EDAR* genes, involved in skin pigmentation and the development of ectodermal organs (hair, teeth). In *M. truncatula*, 30% of genes potentially under epistatic selection show classical selection signatures, indicating that a majority of apparently "neutral" genes may in fact show adaptive interaction with other genes. Finally, analyzes of enrichment and of genomic subnetworks of interactions anchored on 98 symbiotic genes (nodulation, mycorrhization) demonstrate a significant role of epistatic selection in the evolution of some biological pathways. This work initiates the identification and the exploration of coadapted gene networks using genome-wide SNP data, in model and non-model organisms.

Résumé

L'adaptation par sélection naturelle est centrale dans l'évolution des espèces. En ciblant les différences de survie et/ou de reproduction des individus en fonction des changements de l'environnement, la sélection filtre les variants génétiques dans les populations. Les gènes extrêmement conservés sont soumis à la sélection purifiante qui élimine les mutations délétères, alors que d'autres gènes plus polymorphes porteront des mutations positivement sélectionnées dans un certain contexte environnemental. Depuis plus de 20 ans, la modélisation en génétique des populations et l'émergence de technologies de séquençage accélérant l'identification de variants génétiques à l'échelle du génome (e.g. les Single Nucleotide Polymorphisms – SNP) ont permis le développement de nombreuses méthodes statistiques analysant le polymorphisme pour identifier des gènes ou régions du génome présentant des signatures de sélection, tout en tenant compte des autres forces évolutives (dérive génétique, flux géniques) influençant ce polymorphisme. Cependant, elles testent l'hypothèse de sélection indépendamment sur chaque locus et ne permettent pas d'explorer l'interaction entre les allèles des gènes comme cible potentielle de la sélection épistatique. Or, la génétique quantitative et la biologie moderne montrent indiscutablement que les gènes ne sont pas des entités fonctionnellement indépendantes, mais qu'ils sont des éléments interagissant dans des réseaux plus vastes permettant l'expression des caractéristiques biologiques. Cette thèse a pour objectif de proposer un nouveau test statistique qui permet d'identifier des signatures de sélection épistatique, donc des gènes coadaptés, sur la base du déséquilibre de liaison (DL) à l'aide de marqueurs SNP. Dans une première partie, nous décrivons les statistiques proposées, T_{r_v} et T_{corPC1_v}, qui comparent des paires de SNP ou de régions génomiques. Elles sont basées sur des travaux récents montrant que le coefficient de corrélation (r), fortement influencé par la structuration génétique des populations et l'apparentement des individus, doit être corrigé (r_v) par la matrice d'apparentement entre les individus (V). Couplées à des calculs intensifs, des simulations de données SNP pangénomiques en populations structurées ont permis de démontrer que T_{r_v} et T_{corPC1_v} suivent une distribution de Student $\tau_{(n-2)}$, réduisent fortement le bruit de fond de DL généré par les forces évolutives non sélectives, et ont une bonne puissance de détection. Dans une deuxième partie, nous utilisons l'approche de « Genome-Wide Epistatic Selection Scan » (GWESS) chez la plante modèle *Medicago truncatula*, où un gène candidat est utilisé comme appât pour calculer sa corrélation avec tous les autres gènes du génome. Suite à

l'identification d'une signature de sélection épistatique entre *MtSUNN* et *MtCLE02*, codant respectivement pour un récepteur et un peptide de signalisation, une preuve de concept est apportée par la démonstration expérimentale (collaboration) que *MtCLE02* a un rôle négatif sur la nodulation et dépendant de *MtSUNN*. L'approche GWESS appliquée sur des données SNP chez l'homme montre une signature de sélection épistatique entre les gènes *SLC24A5* et *EDAR*, impliqués dans la pigmentation de la peau et le développement des organes ectodermiques (cheveux, dents). Chez *M. truncatula*, 30% des gènes potentiellement sous sélection épistatique présentent des signatures de sélection classiques, indiquant qu'une majorité de gènes en apparence « neutres » peuvent en fait être en interaction adaptative avec d'autres gènes. Enfin, des analyses d'enrichissement et de sous-réseaux génomiques d'interactions ancrés sur 98 gènes symbiotiques (nodulation, mycorhization) démontrent un rôle non négligeable de la sélection épistatique dans l'évolution de certaines voies biologiques. Ces travaux initient l'identification et l'exploration de réseaux de gènes coadaptés à l'aide de données SNP pangénomiques, chez des organismes modèles ou non modèles.

Remerciements

Je voudrais tout d'abord remercier les membres du jury, Mathieu Gautier, Maud Tenaillon et Joelle Ronfort. Merci d'avoir pris le temps de lire mon manuscrit et accepté d'évaluer mon travail.

Merci à Maxime, un incroyable directeur de thèse ! Merci de m'avoir accueillie dans ton équipe et de m'avoir fait confiance tout au long de ce travail. Merci d'avoir été aussi présent tout au long de ces trois ans et demi et notamment pendant ces derniers mois un peu particuliers avec le covid-19. Merci pour toutes les connaissances que tu m'as apportées et toutes les compétences que j'ai pu développer pendant ces trois ans et demi, notamment grâce à ta bienveillance et ton écoute. Les trois ans passés sous ta direction m'ont permis de passer de l'état d'étudiante à celui d'une adulte apprentie chercheuse. Les prochains étudiants que tu auras en thèse seront très chanceux comme j'ai pu l'être. Merci également à Christophe pour m'avoir suivie durant cette thèse. Merci pour ta confiance et le regard extérieur que tu as pu apporter sur ce travail.

Je tiens également à remercier Pierre-Marc pour avoir accepté de faire partie de mon comité de thèse et pour tous les conseils avisés que tu as pu me donner au cours de ce travail et notamment ton expertise sur les gènes de symbiose. Je voudrais aussi remercier Brigitte Mangin qui m'a accueillie en stage de M1 dans son équipe. Merci d'avoir accepté de faire partie de mon comité de thèse et merci pour tes conseils en statistique. Ce travail de thèse n'aurait pas été possible sans tes mesures du déséquilibre de liaison. Je remercie également Stéphane De Mita qui a accepté de faire partie de mon comité de thèse.

Je remercie aussi chaleureusement Marie-Laure Martin-Magniette qui m'a accueillie dans son laboratoire pendant quelques jours, qui a passé du temps à comprendre mon travail (dans les détails) et m'a donné de précieux conseils qui m'ont permis d'avancer. J'ai pu également, grâce à Marie-Laure, participer aux journées Netbio, qui ont été très enrichissantes.

Je remercie aussi Fabrice Roux pour le travail que nous avons pu faire *sur A. thaliana* et pour nos échanges avec Nathalie Aoun.

Je remercie Elodie Gaulin et Catherine Mathé pour leurs conseils et leurs soutiens pour les enseignements qui ont été une part très enrichissante de la thèse.

Je tiens aussi à remercier Florian Frugier et Corentin Moreau pour leurs manips sur *SUNN* et *CLE02* qui ont vraiment apporté une plus-value à mon travail. Merci pour cette collaboration.

Je souhaite remercier la Genotoul pour les ressources informatiques qui sont mises à notre disposition et je remercie Didier Laborie et Marie-Stéphane Trotard pour leur disponibilité, pour toutes les réponses à mes questions et pour leur support technique.

Je remercie également l'école Doctorale SEVAB et surtout Dominique Pantalacci pour sa disponibilité et toutes ses réponses à mes (nombreuses) questions.

Je remercie tous les membres de l'ancienne équipe IPM ainsi que tous les membres de la nouvelle équipe EVO et toutes les personnes du laboratoire avec qui j'ai pu partager de bons moments. Je remercie tous les perms et non-perms qui ont fait partie de ma vie au LRSV et notamment Emilie, Laurent C, Quentin, Nico, Nathanael, Laurent K, Salimata, Simon, Duchesse, Camille, Loïc, Julie, Mélanie, Tatiana, Maxime, Arthur et j'en oublie. Je remercie aussi les bioinfo Jean et Cyril, qui sont une source de bons conseils ! Je remercie aussi chaleureusement le labcom ; Rémi, Aurélien, Damien, Elodie, Alexandra, Olivier et Thomas. Rémi pour nos échanges sur les films, les livres ou les vacances autour d'un café. Damien, Aurélien et les filles pour les pauses café et les discussions qu'on a pu avoir. Je fais également une mention spéciale pour les (ex) occupants du bureau 50, Charlène, Nathanael et Marion, on a passé de très bons moments autour de la senseo. Marion, pour nos sorties piscine, pour les discussions intenses sur nos sélections vinted, pour les soirées et pour tous les moments qu'on a partagés avec Quentin et les autres. Charlène pour tout ce qu'on a partagé aussi depuis notre premier jour de stage M2, on s'est suivies tout du long. Merci pour nos discussions pour tous tes conseils et tes encouragements surtout pour la partie administrative, tu m'as appris à ne plus procrastiner. Et bien sûr pour tous les moments partagés au labo et hors du labo.

Enfin je souhaite remercier mes amis et ma famille. Toutes mes copines et mes copains d'enfance ; Emilie (super colloc'), Alice, Elie, Agnès, Elisa, Baptiste, Marion, Léa, Camille, Maud, Cindy, Lucie, Sandy et Flore. On se suit depuis toujours (nos 3 ans pour certaines) et j'espère que ça continuera comme ça.

Je remercie toutes les personnes qui m'ont aidé pour relire ce manuscrit et corriger mes fautes d'orthographes : Pascale, Maurice, Martine, Papa, Maman, Emilie et Elisa !

Je remercie mes parents pour leur soutien infaillible et leurs encouragements. Merci de m'avoir permis d'en être là aujourd'hui. Je remercie aussi Marie, ma super belle-mère !

Enfin, je remercie Jonathan, mon binôme de vie. Tu m'as accompagné tout au long de ces trois ans et a été d'un soutien vital. Maintenant une nouvelle aventure commence et la famille s'agrandit.

Table des matières

Introduction

1. Diversité génétique et forces évolutives

1.1 La diversité génétique

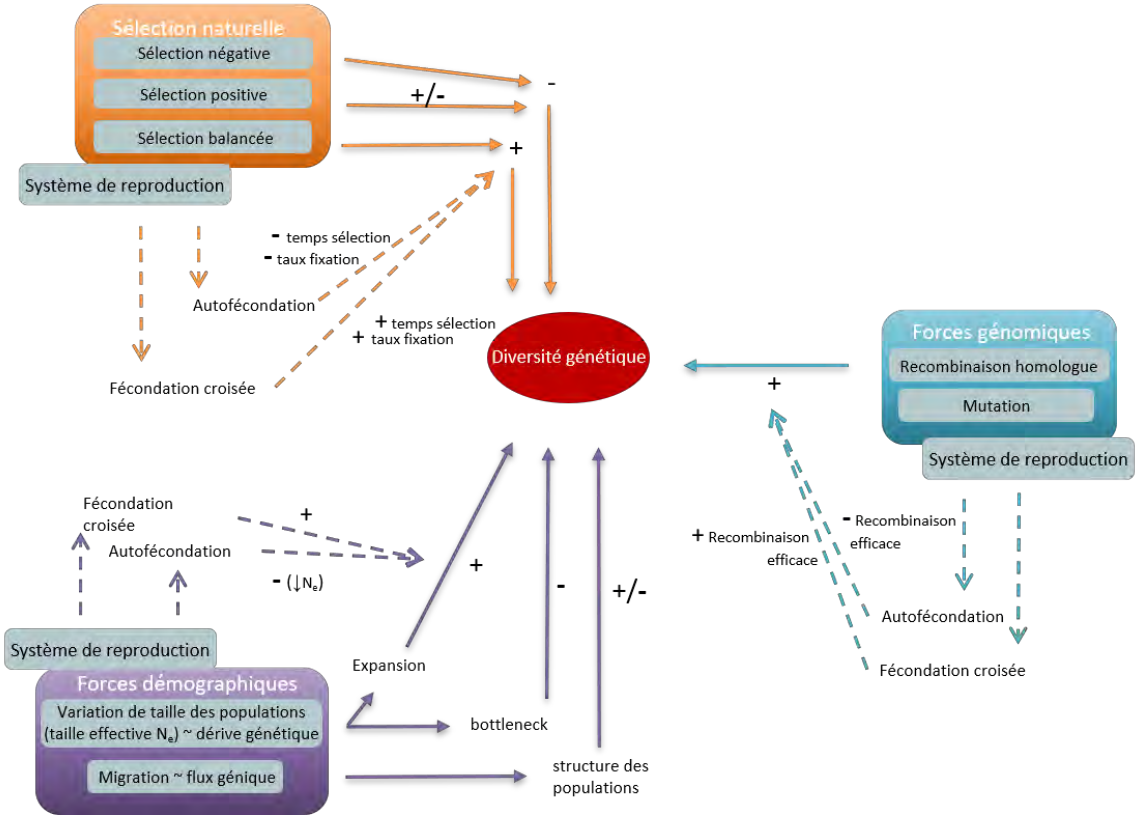
Depuis l'apparition de la vie sur terre et l'hypothétique premier ancêtre commun LUCA (Last Universal Commun Ancestor), les organismes vivants ont évolué et se sont diversifiés. La diversité génétique est à la base de l'évolution, elle représente toutes les variations naturelles possibles entre les génomes ou les chromosomes des individus d'une même espèce ou de différentes espèces. Cette diversité génétique est à la source de la diversité phénotypique, c'est-à-dire d'une multitude de variations biologiques interindividuelles qui évoluent dans des environnements spécifiques. Les individus d'une même espèce ne sont pas génétiquement identiques, leurs séquences d'ADN diffèrent en un grand nombre de positions et ces différences constituent la diversité génétique, appelée également le polymorphisme génétique. La diversité génétique est essentielle pour permettre l'adaptation des espèces. Elle contribue par exemple au développement des différences en termes de résistance aux agents pathogènes, de stratégies de survie ou de reproduction, ou plus généralement de caractéristiques moléculaires, cellulaires, physiologiques ou morphologiques qui évoluent en fonction des modifications de l'environnement d'un organisme. De manière générale, la diversité génétique varie entre les espèces ; par exemple, le génome de *Drosophila simulans* présente une variabilité génétique moyenne de 3% (Begun et al., 2007; Ellegren & Galtier, 2016; Lack et al., 2015), tandis que le génome humain est variable à 0.1% (McVean et al., 2005, The 1000 Genomes Project Consortium 2015). Le polymorphisme varie également au sein des génomes entre différents locus ou entre les chromosomes, cela a pu être montré par exemple chez la plante modèle *Arabidopsis thaliana* (Magnus Nordborg et al., 2005) ainsi que chez le maïs (Tenaillon et al., 2001), chez l'homme (Sachidanandam et al., 2001) ou le poulet (Wong et al., 2004). Le polymorphisme varie aussi selon la fonction des séquences d'ADN. Ainsi, les séquences géniques codant pour des protéines dont les fonctions sont souvent préservées dans l'évolution sont généralement plus conservées que les séquences intergéniques.

1.2 Les forces évolutives

Plusieurs forces évolutives influencent la diversité génétique en modifiant la fréquence des allèles à l'échelle temporelle mais aussi spatiale (**Figure 1**). Ces forces évolutives se classent en trois catégories : (i) les forces génomiques, dont la mutation et la recombinaison,

qui sont les seules forces évolutives créatrices de diversité génétique, (ii) les forces liées aux processus démographiques de variation de taille des populations et de migration que sont la dérive génétique et les flux de gènes, ainsi que (iii) la sélection naturelle.

Figure 1 : Représentation des déterminants de la diversité génétique. Les forces génomiques, démographiques et la sélection naturelle sont les principaux facteurs influençant la diversité génétique et ces facteurs sont eux même déterminés par différents paramètres. Les symboles + et – indiquent si les paramètres désignés augmentent ou diminuent la diversité génétique (Inspirée de H Ellegren and N Galtier, 2016).



1.2.1 Les forces génomiques

Les mutations sont à l'origine de toutes nouvelles variations génétiques. À chaque génération, de nouveaux allèles apparaissent par des mutations à la suite d'erreurs de réplication de l'ADN ou de dommages à l'ADN induits par des éléments mutagènes. Les variations génétiques les plus répandues dans les génomes et qui sont le plus couramment utilisées comme marqueurs génétiques sont les Single Nucleotide Polymorphism (SNP). Ce sont des substitutions d'un nucléotide par un autre. Le taux de substitution nucléotidique μ n'est pas constant entre les espèces ni le long des génomes, mais il se situe néanmoins entre 10^{-10} et 10^{-8} mutation par nucléotide par réplication (Sung et al., 2012). La recombinaison

homologue, aussi appelée « crossing-over » est également source de diversité génétique, en effet elle consiste en un échange de certaines portions entre deux chromosomes homologues, provoqué par des cassures de brins d'ADN (Creighton & McClintock, 1931). La recombinaison homologue permet de rompre les associations d'allèles portés par chacun des chromosomes homologues à différents locus (ou haplotypes) et d'en créer de nouvelles. C'est par ces nouvelles associations d'allèles que la recombinaison crée de la diversité génétique à l'échelle des chromosomes.

1.2.2 Les forces démographiques

La dérive génétique et les flux de gènes sont des forces évolutives qui influencent également la diversité génétique en réponse à des facteurs démographiques tels que la migration, l'expansion de population ou les goulots d'étranglements (« botteneck »). Selon la théorie neutraliste de l'évolution, la plupart des mutations sont neutres et n'affectent pas la valeur adaptative, ou *fitness*, des individus (M. Kimura, 1968; M. Kimura, 1983). Ces mutations neutres qui apparaissent peuvent augmenter en fréquence jusqu'à la fixation ou être éliminées par dérive génétique, un phénomène aléatoire qui modifie les fréquences alléliques car certaines copies d'un gène sont plus (ou moins) transmises que d'autres par hasard, au cours des générations (Wright, 1931). La probabilité de fixation d'un allèle neutre dans une population finie est égale à sa fréquence dans la population de départ, et donc si une mutation apparaît dans une population diploïde de taille N, la probabilité qu'elle se fixe est de $1/2N$. Le temps de fixation d'une telle mutation est de 4N générations et correspond aussi au temps moyen de coalescence des 2N copies du gène dans la population (Kimura, 1969). Durant ce temps, le taux de polymorphisme du locus considéré est égale à $\theta = 4N\mu$ avec μ le taux mutation par génération. Dans les populations naturelles, tous les individus ne participent pas à la reproduction, si bien que la taille de la population qui détermine la dérive génétique n'est pas égale à l'effectif total. La taille efficace d'une population (N_e) correspond à la taille d'une population appelée population de Wright-Fisher qui est panmictique et de taille constante et où seule la dérive génétique influence les fréquences des allèles. Ainsi la taille efficace (N_e) de la population correspond à la taille d'une population de Wright-Fisher qui a le même taux de dérive génétique que la population étudiée. Dans une population de taille constante N, l'intensité de la dérive génétique est inversement proportionnelle à N (M. Kimura, 1983), donc

si la taille de la population évolue au cours des générations, cela va impacter le taux de dérive génétique. Par exemple, si une population subit un goulot d'étranglement (*bottleneck*), cela va entraîner une diminution de la diversité allélique par la fixation ou la perte d'allèles favorisés par la dérive génétique (Ellegren & Galtier, 2016). Inversement, si une population est en expansion, la dérive génétique sera moins importante et la diversité allélique sera favorisée par l'apparition et le maintien de nouvelles mutations (Nei, 1988; Wright, 1931). Divers degrés d'isolement ou de connexion entre populations de tailles variables influencent aussi la diversité génétique des populations en générant de la structure génétique à différentes échelles. En effet, un isolement peut entraîner une diminution de la diversité génétique par dérive génétique et absence de flux de gènes localement dans les sous-populations. Paradoxalement, cela peut aussi permettre le maintien d'une certaine diversité si l'on considère l'ensemble des sous-populations. Ainsi, la migration qui consiste en un transfert d'allèles entre différentes populations d'une même espèce, influence aussi la diversité génétique. Les flux de gènes dus à des événements de dispersion suivis de reproduction permettront d'augmenter - ou limiter la perte de - la diversité génétique dans la population qui reçoit ces flux. En particulier, si la migration est suffisamment importante, elle peut permettre de diminuer la consanguinité, et donc l'homozygotie dans la population receveuse. Par exemple, dans un modèle infini en « îles » dont les populations panmictiques de taille N sont connectées de manière homogène par un taux de migration m, l'indice de différenciation génétique entre population (ou l'indice de fixation) est $F_{ST} = 1/(1+4Nm)$ à l'équilibre (Wright, 1931). Ce modèle permet de montrer qu'un migrant par génération ($Nm = 1$) suffira à limiter le coefficient de consanguinité de 20%, ce qui démontre la forte capacité des flux de gènes à s'opposer aux effets de la dérive génétique.

Les statistiques F de Wright : Les sous-populations du modèle infini en « île » peuvent aussi ne pas être panmictique et dans ce cas, le déficit en hétérozygotie résulte de deux effets ; la structure des populations et l'effet des croisements non aléatoires des individus au sein de chaque sous population. Nous pouvons ainsi définir les statistiques F de Wright ou indice de fixation de Wright. Le F_{ST} correspond à la différenciation génétique entre sous populations, il mesure l'effet de structuration des populations c'est-à-dire la part d'homozygotie des individus de la population totale (indice T) provenant de la subdivision de ces mêmes individus en sous-populations de taille limitées (indice S). Le F_{IS} mesure la part d'homozygotie qui

provient d'une déviation par rapport au régime de reproduction en panmixie dans les sous-populations, il mesure le déficit en hétérozygotie à l'échelle intra-populationnelle. Enfin, le F_{IT} mesure l'homozygotie dans la population globale provenant des deux effets précédents.

$$F_{ST} = \frac{H_T - H_S}{H_T}$$

$$F_{IS} = \frac{H_S - H_O}{H_S}$$

$$F_{IT} = \frac{H_T - H_O}{H_T}$$

Où H_T correspond à l'hétérozygotie attendue (théorique) dans la population globale sous panmixie, H_S correspond à l'hétérozygotie attendue dans chaque sous population sous panmixie et H_0 correspond à l'hétérozygotie observée au sein des sous-populations. Ces statistiques mesurent la part de consanguinité entre les individus qui peut être dû à la taille finie des populations ou au nombre restreint des individus des sous-populations (F_{ST}) ainsi que la part de consanguinité dû aux déviations par rapport au régime de reproduction panmictique (F_{IS}).

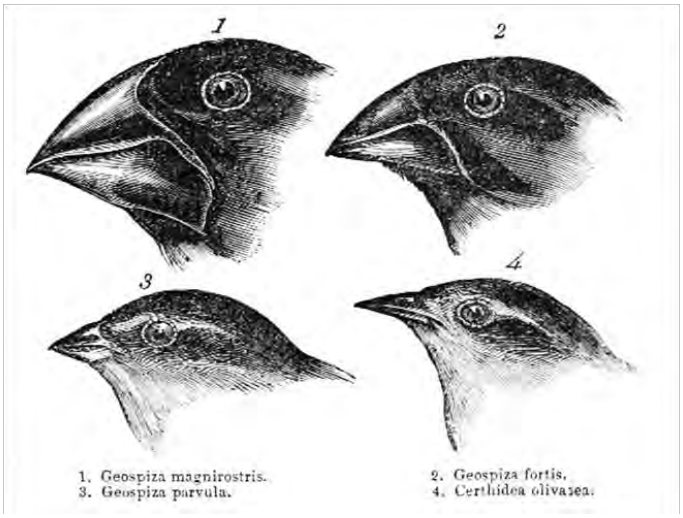
Les systèmes de reproduction, en modifiant les fréquences génotypiques, vont aussi modifier la diversité génétique via les forces évolutives. Les espèces qui s'autoféconden, ou qui présentent un fort degré de consanguinité, montrent moins de variabilité génotypique que les espèces à fécondation croisée car l'autofécondation et la consanguinité augmentent l'homozygotie. À chaque génération d'autofécondation, la proportion d'hétérozygotie est réduite de moitié, et si des allèles récessifs délétères sont présents dans la population, cela peut entraîner une diminution de la *fitness* appelée dépression de consanguinité ('inbreeding depression'). Les populations présentant un fort taux de consanguinité sont elles aussi sujettes à la dépression de consanguinité. Si l'on considère σ comme le taux d'autofécondation, la relation $F_{IS} = \sigma/(2 - \sigma)$ quantifie l'excès relatif d'homozygotie intra-populationnelle par rapport à ce qui attendu sous un régime en reproduction panmictique. Pour une population à 100% autoféconde ($\sigma = 1$), $F_{IS} = 1$ et la population n'est composée que d'individus homozygotes pour les différents allèles. Ce régime d'autofécondation, en réduisant le nombre de gamètes indépendants échantillonnables pour la reproduction, conduit à une diminution de la taille efficace de la population (N_e) d'un facteur $N_e = N/(1 + F_{IS})$, entraînant une augmentation de la dérive génétique (M Nordborg, 2000; Pollak, 1987). Ainsi, pour une

population qui est à 100% autoféconde, N_e est réduite de 50% par rapport à une population à fécondation croisée. Comme le niveau d'équilibre de la variabilité génétique neutre est proportionnel à N_e, et donc à la dérive génétique (Hudson, 1990; M. Kimura, 1971), une population autoféconde présentera en moyenne 50% de la diversité génétique d'une population équivalente mais à fécondation croisée (Burgarella & Glémin, 2017).

1.2.3 La sélection naturelle

La diversité génétique est un élément essentiel pour l'adaptation des organismes à leur environnement. La sélection naturelle agit sur les traits phénotypiques pour permettre l'adaptation et le maintien de fonctions moléculaires essentielles.

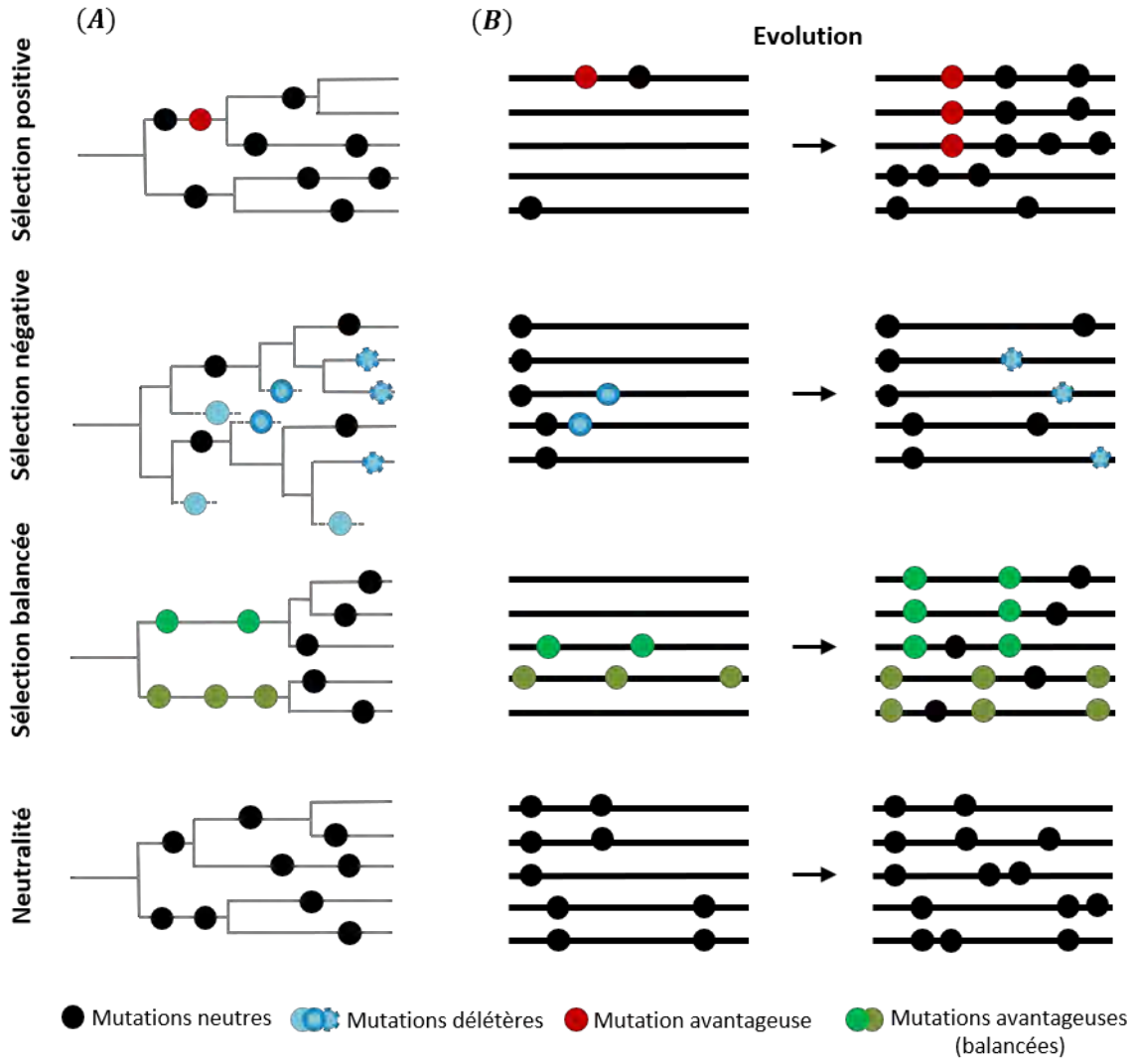
Figure 2 : Les caractéristiques phénotypiques des pinsons des îles Galapados observées par Charles Darwin. Darwin observa des formes de bec variées chez différentes espèces de pinsons et il fit l'hypothèse qu'une forme de bec chez une espèce ancestrale s'est adaptée au fil du temps pour permettre aux pinsons d'acquérir différentes sources de nourriture. Cette illustration montre la forme de bec de quatre espèces de pinsons: 1 Geospiza magnirostris (le grand pinson brodé), 2 G. fortis (le pinson brodé moyen), 3 G. parvula (le petit pinson des arbres), et 4 Certhidea olivacea (le pinson vert).



Charles Darwin a introduit le concept de sélection naturelle au début du 19ème siècle pour décrire l'évolution des espèces (« *On the Origin of species* » Darwin et al., 1859). Il montre comment de nouvelles espèces apparaissent à partir d'espèces préexistantes en réponse à des changements de l'environnement. Un des premiers exemples démontrant ce phénomène d'adaptation par sélection naturelle est la description de l'évolution de la taille du bec chez les pinsons des Galápagos, en fonction de la taille des graines disponibles dans un

environnement donné. Le principe de sélection naturelle décrit par Darwin repose sur la variabilité interindividuelle. Les caractéristiques ou les phénotypes des individus qu'il observe sont transmis(es) des parents aux descendants, et ces caractéristiques sont variables entre tous les descendants. De plus, les ressources pour la survie et la reproduction dans un environnement donné sont limitées ; il existe donc une compétition pour ces ressources, à chaque génération. Ainsi, les individus qui présentent les caractéristiques héritées leur permettant de mieux rivaliser pour ces ressources survivront mieux et auront plus de descendants. Comme ces caractéristiques sont héritables, elles seront plus représentées à la génération suivante et cela produira des changements dans les populations au cours des générations. Darwin observa plusieurs espèces de pinsons qui ont évolué à partir d'une première espèce originelle en colonisant différentes îles du Galápagos. La taille et la forme des becs des pinsons sont corrélées à leurs nouveaux environnements, ce qui correspond ici, à la taille et la nature des graines dont ils se nourrissent (**Figure 2**). L'adaptation par sélection naturelle se base sur un principe simple : les individus qui possèdent des traits phénotypiques leur conférant de meilleures capacités à survivre et à se reproduire par rapport à un environnement spécifique, ont une meilleure valeur sélective, ou *fitness.* Les variants génétiques à l'origine de ces traits phénotypiques plus adaptés seront la cible de la sélection naturelle, et ils vont augmenter en fréquence au cours des générations. La sélection naturelle permet donc l'évolution par adaptation. Parallèlement, Mendel a introduit la théorie de l'hérédité et de la génétique à partir de ses travaux réalisés sur le pois. Il comprend l'existence d'unités héritables, aujourd'hui appelées les gènes, et il en tire un certain nombre de principes connus sous le nom de lois de Mendel. Les travaux de Mendel permettent de compléter la théorie de la sélection naturelle de Darwin pour qui le mécanisme de transmission des caractères reste inexpliqué. Puis au 20ème siècle, R. A. Fisher, S. Wright, et J. B. S. Haldane ont développé les principes de la génétique des populations basés sur la théorie de l'évolution et de la sélection naturelle. Ils ont montré comment différentes forces influencent l'évolution des caractères génétiques et héritables des populations au cours des générations. Ainsi, les variants génétiques qui affectent la *fitness* sont la cible de la sélection et la distribution des fréquences de ces mutations évolue en fonction des valeurs sélectives (**Figure 3**).

Figure 3 : Distribution des variants génétiques sous différents modèles de sélection. (A) Les arbres représentent les généalogies de quatre locus sous sélection positive, négative, balancée et sous neutralité. Chaque cercle représente une mutation (variant génétique) qui apparaît sur la séquence du locus et les couleurs correspondent aux différents modes de sélection. (B) Représentation des haplotypes contenant les mutations au cours du processus d'évolution par sélection ou neutralité sur plusieurs générations schématisées par les flèches. Chaque cercle correspond à une mutation et chaque trait schématise la séquence du locus chez un individu.



Si une mutation avantageuse apparaît, elle est soumise à un processus de sélection positive. À l'inverse, les mutations défavorables sont contre-sélectionnées et leurs fréquences diminuent rapidement ; on parle alors de sélection négative. Les mutations, qui apparaissent aléatoirement, sont pour la plupart délétères et elles sont éliminées plus ou moins efficacement par sélection négative suivant leur degré de récessivité. Cette sélection négative, ou purifiante, est à l'origine de la forte conservation de certains gènes ou de régions génomiques à l'échelle interspécifique. Enfin, la sélection balancée favorise quant à elle le maintien de plusieurs allèles à un locus dans une population et tend à augmenter la diversité

par un mécanisme d'avantage aux hétérozygotes ou par sélection fréquence-dépendante (Bodmer, 1972; Takahata & Nei, 1990). Les différents modes de sélection auront des effets variés sur la distribution des fréquences des mutations aux locus sélectionnés. La sélection négative provoque un excès d'allèles rares, en faibles fréquences, du fait de la contre-sélection. Sous sélection positive (ou balayage sélectif) on observe à la fois un excès d'allèles en forte fréquence au niveau des mutations sélectionnées, mais aussi un excès d'allèles en faibles fréquences qui sont les mutations qui apparaissent pendant la sélection. Enfin, sous sélection balancée, il y a un excès d'allèles en fréquences intermédiaires (**Figure 3**, et voir chapitre 2.3 **Figure 12**).

Un grand nombre de méthodes permettant de détecter des locus influencés par la sélection naturelle ont été développées car elles permettent d'identifier des gènes fonctionnellement importants à l'échelle interspécifique, tout comme des gènes impliqués dans l'adaptation des populations à des modifications de leur environnement. Le problème central que chaque méthode tente de résoudre est de quantifier le rôle des forces évolutives non sélectives sur le polymorphisme des populations, qui peuvent « mimer » l'effet de la sélection. En effet, les mutations neutres qui apparaissent peuvent augmenter en fréquence dans la population en réponse à des facteurs aléatoires tels que la dérive génétique même si elles ne sont pas avantageuses (Nielsen, 2005) et les régions du génome qui sont la cible de la sélection sont aussi influencées par la dérive génétique. Ainsi, la plupart des méthodes développées pour détecter des signatures de sélection cherchent à distinguer les variations génétiques qui évoluent selon un modèle neutre, des variations soumises à la sélection (Haasl & Payseur, 2016; Harris & Meyer, 2006; Nielsen, 2005; Pavlidis & Alachiotis, 2017; Vitti et al., 2013; Weigand et al., 2018).

2. Détection des bases génétiques de l'adaptation

Avec le développement des technologies récentes de séquençage haut-débit, un grand nombre de marqueurs génétiques sont aujourd'hui disponibles pour étudier la sélection en utilisant des approches globales à l'échelle génomique. Ainsi, diverses méthodes ont été développées pour détecter les bases génétiques de l'adaptation et identifier les gènes ou les locus sous-jacents à l'échelle du génome. Deux grands types d'approches ont été développées : (i) les approches qui recherchent des associations entre le génotype et des variables

quantitatives associées à la *fitness*, telles que les analyses de cartographie de Quantitative Trait Loci (QTL), dont les analyses d'association génotype-phénotype « Genome-Wide Association Study » (GWAS), ainsi que les analyses de « Gene Environment Association » (GEA) et (ii) les approches basées exclusivement sur les données génétiques. Parmi ces approches, il y a notamment les analyses phylogénétiques qui comparent les taux de substitutions synonymes et non-synonymes dans les séquences codantes, ainsi que les méthodes de génétique des populations qui recherchent des signatures de sélection à l'échelle intraspécifique, à l'aide de « Genome-Wide Selection Scan » (GWSS). Les méthodes de GEA et de GWAS permettent d'identifier les marqueurs génétiques associés respectivement à des variables environnementales ou à des traits phénotypiques, tandis que les méthodes de génétique des populations permettent d'identifier les signatures génétiques de sélection naturelle par l'analyse des fréquences alléliques aux positions polymorphes.

2.1 Méthodes d'associations génotype-phénotype et génotype-environnement

Les analyses de GEA exploitent les données de polymorphisme entre les individus de différentes populations qui évoluent dans des environnements spécifiques et qui présentent des phénotypes localement adaptés. Ces analyses se focalisent sur l'identification de locus dont les variants alléliques sont associés à des variables de l'environnement. Ces dernières peuvent être, par exemple, des variables climatiques qui sont considérées comme des agents sélectifs auxquels les populations vont répondre par un processus d'adaptation locale (Coop et al., 2010). Deux grands types d'approches sont proposées pour réaliser des analyses de GEA : les approches 'individus-centrées' dont les données de génotypes sont constituées d'un individu par sous-population et les 'approches populationnelles', plus récentes, qui prennent en compte plusieurs individus par population et qui intègrent donc la variation génétique intra-populationnelle (Coop et al., 2010; Frachon et al., 2018, 2019; Gautier, 2015; Günther & Coop, 2013). Dans la littérature, il existe plusieurs exemples d'analyses GEA notamment chez les plantes modèles *Arabidopsis thaliana* (Hancock et al., 2011) et *Medicago truncatula* (Burgarella et al., 2016; Yoder et al., 2014). Une étude récente menée chez *A. thaliana* a montré l'association entre des variants génétiques et la diversité α, qui est la composition et l'abondance des espèces végétales au contact des populations naturelles de cette espèce

(Frachon et al., 2019). Cette analyse de GEA a permis d'identifier des gènes candidats impliqués dans les interactions plante-plante dont, notamment, des gènes de réponse à l'altération de la lumière. La plupart des analyses de GEA sont aussi réalisées à partir de variables environnementales abiotiques telles que la température, le taux d'humidité ou la luminosité (Ferrero-Serrano & Assmann, 2019; Hancock et al., 2011). Les corrélations génotype-environnement identifiées au cours d'analyses GEA sont le reflet de la variation des fréquences alléliques entre les populations analysées en réponse aux changements de l'environnement et cela peut-être la signature génétique d'adaptation locale par sélection naturelle. Toutefois, les variations génétiques qui sont associées à une variable environnementale ne sont pas nécessairement adaptatives mais il est possible de tester cette hypothèse en recherchant des signatures de sélection sur ces mêmes locus. Une étude menée chez la plante *Medicago truncatula* a identifié un certain nombre de SNP en association avec des variables climatiques, et les locus ainsi identifiés présentaient également des signatures génétiques de balayage sélectif (Yoder et al., 2014). Par une approche ciblée sur les gènes de floraison, il a aussi été montré que chez *Medicago truncatula,* le temps de floraison intervient dans l'adaptation au régime annuel de précipitations. Ceci se fait principalement par la variation allélique sur des gènes dont la position dans les voies de signalisations de la floraison est proche des stimuli environnementaux (Burgarella et al., 2016).

Les approches de cartographie de QTL et de GWAS étudient les relations entre les variations génétiques et les variations pour des traits phénotypiques quantitatifs. Les approches de cartographie de QTL sont réalisées à partir de populations « artificielles » issues de croisements contrôlés, telles que les RIL (Recombinant Inbred Lines), ou simplement à partir de populations dont on connait le pedigree des individus génotypés. .Les approches de GWAS sont réalisées avec des populations naturelles. Les populations artificielles étant difficiles à mettre en place chez certaines espèces et notamment les espèces animales, la GWAS présente un avantage majeur par rapport aux méthodes de cartographie de QTL traditionnelles. De plus, les GWAS reflètent mieux la variation génétique à l'échelle des populations, contrairement aux approches de cartographie de QTL qui sont limitées à la diversité génétique parentale des populations issues de croisements. Les analyses de GWAS identifient donc des associations génotype-phénotype en recherchant, dans une population, des corrélations significatives entre les génotypes aux marqueurs SNP et les valeurs

(quantitatives ou qualitatives) phénotypiques mesurées sur les individus (Bergelson & Roux, 2010; Bonhomme & Jacquet, 2020). La probabilité de détecter des QTL associés à une variable phénotypique dépend de l'héritabilité de ce trait, autrement dit, de la proportion de la variation du trait phénotypique qui est due à des différences génétiques entre les individus. Cette probabilité va dépendre aussi de la fréquence des allèles aux QTL. Chez les plantes, les premières analyses de GWAS ont été réalisées chez *Arabidopsis thaliana* sur une variété de phénotypes liés à la résistance aux pathogènes, au développement et à la floraison (Atwell et al., 2010; Hancock et al., 2011). Les premières études de GWAS chez *M. truncatula* portent sur des phénotypes de nodulation, de croissance, de floraison (Stanton-Geddes et al., 2013) et de résistance à l'oomycète *Aphanomyces euteiches* (Bonhomme et al., 2014, 2019). Les variations génétiques identifiées en association avec un trait phénotypique ne sont pas nécessairement des variations génétiques adaptatives, tout comme il n'est pas toujours démontré que le trait phénotypique étudié est adaptatif. Afin de tester si les mutations identifiées par GWAS sont des mutations adaptatives, il est possible de rechercher des signatures de sélection aux locus identifiés par GWAS. Ainsi, les approches de génétique d'association visent à identifier les déterminants génétiques de la variation phénotypique, qui peut elle-même être la cible de la sélection naturelle (Flood & Hancock, 2017; Josephs et al., 2017; Magnus Nordborg & Weigel, 2008; Schork et al., 2009).

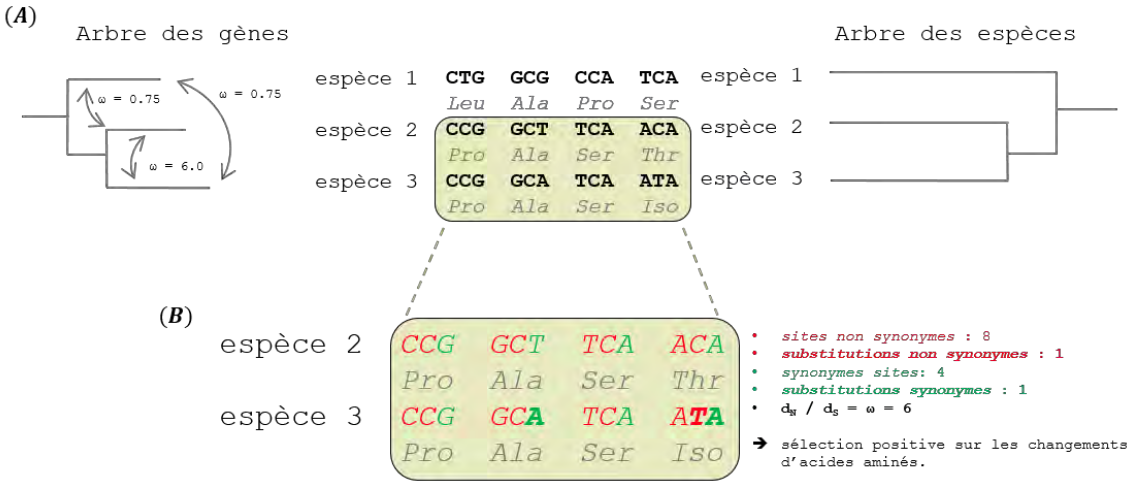
2.2 Méthodes basées exclusivement sur les données génétiques

Les approches strictement basées sur les données génétiques exploitent les signatures des différents types de sélection sur le polymorphisme des séquences pour détecter les bases génétiques de l'adaptation. On discerne deux grands types d'approches : (i) les méthodes basées sur les comparaisons interspécifiques des séquences codantes et (ii) les méthodes de génétique des populations qui permettent de détecter la sélection sur des séquences codantes ou non codantes au sein d'une même espèce. Les approches comparatives interspécifiques sont utilisées pour détecter des événements de sélection passés tandis que les approches intraspécifiques permettent de détecter des événements de sélection récents ou en cours au sein des populations d'une même espèce.

2.2.1 Les approches interspécifiques

Les méthodes de détection inter-espèces reposent généralement sur des comparaisons de séquences codantes homologues entre des taxons apparentés. Par exemple, les séquences homologues qui sont très conservées parmi un ensemble de taxons sont potentiellement sous sélection négative, tandis qu'une forte divergence au niveau de la séquence d'acides aminés entre taxons assez proches peut traduire l'effet de la sélection positive (dite aussi diversifiante, à l'échelle interspécifique).

Figure 4 : Méthode pour détecter la sélection naturelle à l'échelle interspécifique. (A) Représentation d'un gène conservé entre deux espèces et qui présente une forte divergence chez l'espèce 3. L'arbre phylogénétique de gauche est basé sur la séquence du gène. Le paramètre **ω** correspond au rapport d_N/d_S qui compare les taux de substitutions non synonymes et synonymes du gène sur la base de comparaisons par paires d'espèces (détails des calculs non représentés). L'arbre de droite correspond à la phylogénie entre les espèces. (B) Présentation du calcul d_N/d_S (**ω**) entre les espèces 2 et 3 en appliquant le modèle évolutif de Juke et Cantor. Les positions nucléotidiques représentées en rouge correspondent à des positions non synonymes qui modifient l'acide aminé en cas de substitution et les positions nucléotidiques représentées en vert correspondent à des positions synonymes qui ne modifient pas l'acide aminé. Entre les espèces 2 et 3, le rapport d_N/d_S est supérieur à 1, indiquant un évènement de sélection positive qui a favorisé la fixation d'acides aminés substitués.



Parmi les méthodes développées pour détecter la sélection inter-espèce, les tests statistiques les plus couramment utilisés sont basés sur le rapport K_a/K_s autrement appelé d_N/d_S. Ce rapport compare le taux de mutations non synonymes (i.e. modifiant l'acide aminé) au taux de substitutions synonymes (i.e. ne modifiant pas l'acide aminé) qui est supposé être sélectivement neutre (**Figure 4**) (Nielsen, 2005; Vitti et al., 2013). S'il n'y a pas de sélection, le rapport des taux de substitutions non synonymes et synonymes, ω, est égal à 1 (i.e. $\omega = d_N/d_S = 1$) mais s'il y a un excès de substitutions non synonymes (i.e. $\omega > 1$), cela indique un événement de sélection positive favorisant les changements d'acides aminés. Ces évènements de substitution se font, par exemple, en raison de l'apparition ou de l'optimisation de

fonctions ou structures moléculaires adaptées au niveau d'une protéine. À l'inverse, ω < 1 indique une conservation de la séquence d'acides aminés par sélection purifiante, qui vise à maintenir des structures ou fonctions moléculaires essentielles à l'échelle interspécifique. La **Figure 4** illustre comment le rapport d_N/d_S est calculé. Cet exemple montre qu'il y a une conservation des séquences du gène entre les espèces 1 et 2 car elles présentent des valeurs ω < 1 indiquant un maintien des séquences à l'échelle interspécifique par sélection purifiante. En revanche, l'espèce 3 semble montrer des signaux de sélection positive ou divergente par rapport à l'espèce 2 avec une valeur ω > 1 indiquant que la fixation de certaines substitutions nucléotidiques non synonymes a été favorisée chez l'espèce 3.

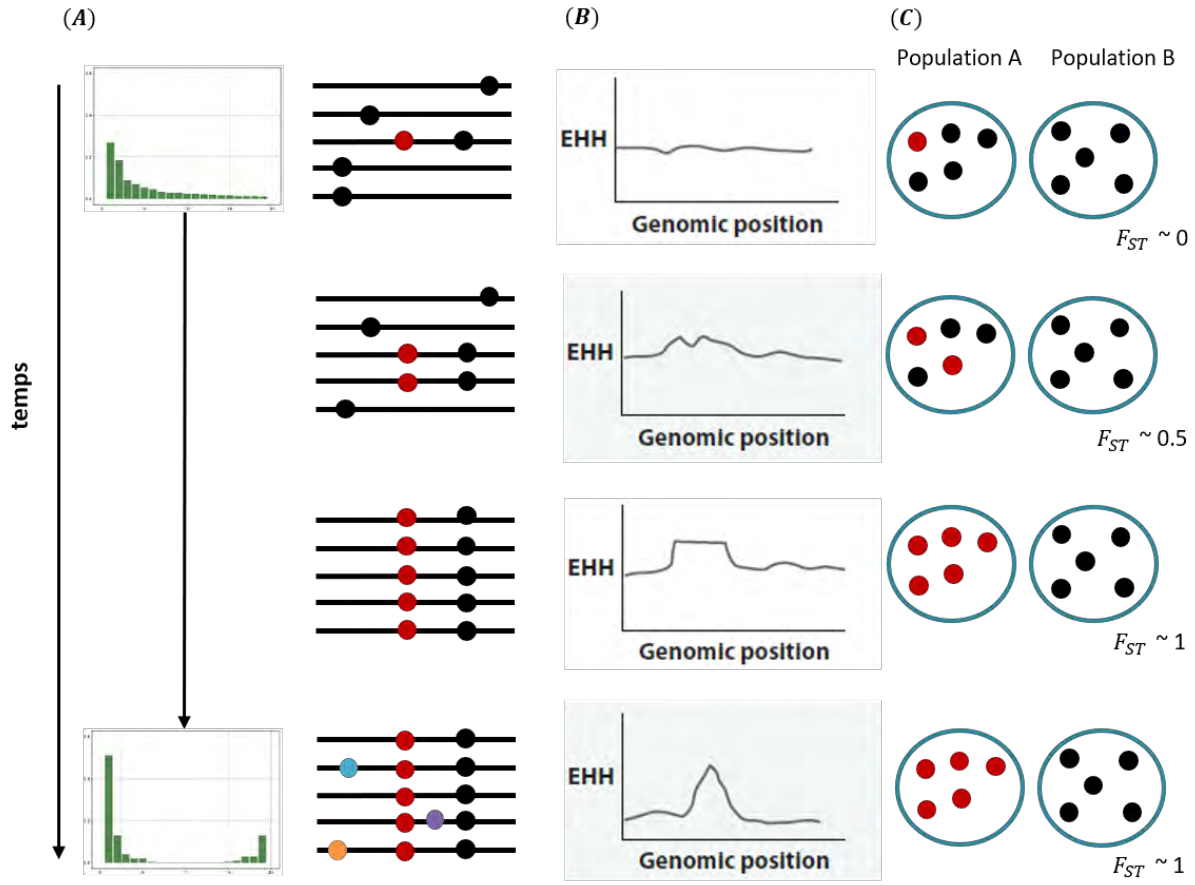
Dans ce contexte, le test statistique de McDonald-Kreitman (MKT) a été développé pour détecter la sélection naturelle en utilisant les données de polymorphismes intra et inter-espèces. Il se base sur l'hypothèse que le taux de divergence inter-espèce et le taux de polymorphisme intra-espèce sont équivalents selon un taux de mutations constant sauf s'il y a de la sélection naturelle ou d'autre forces évolutives telles que les variations de taille de population. En d'autres termes, la méthode compare les proportions de mutations synonymes et non synonymes inter et intra-espèces et permet d'identifier des séquences qui présentent un excès de substitutions non synonymes inter-espèce par rapport au taux intra-espèce (McDonald & Kreitman, 1991).

2.2.2 Les approches intraspécifiques de génétique des populations.

A l'échelle des populations, l'effet de la sélection naturelle s'observe par des changements de fréquence des allèles. Les méthodes développées visent à distinguer les locus neutres des locus sous sélection naturelle et à déterminer quels sont les types de sélection qui influencent les fréquences des allèles à ces locus (Vitti et al., 2013). Les technologies de séquençage haut-débit ont engendré une abondance d'études visant à détecter la sélection à l'échelle globale avec des approches de génomique des populations et permettant l'analyse de plus en plus de gènes. Les méthodes les plus répandues visent à détecter la sélection positive. En effet, lorsqu'une mutation avantageuse sous sélection positive se fixe dans une population, cela va entraîner une réduction de la variabilité génétique car les sites polymorphes neutres qui ségrégent autour de la mutation vont également se fixer par liaison physique (effet d'auto-stop, ou « hitch-hiking »). Ce phénomène local de réduction de la variabilité appelé balayage sélectif (« selective sweep ») va persister jusqu'à ce que la

recombinaison ou de nouvelles mutations apportent à nouveau du polymorphisme (**Figure 5A**). Ainsi, par l'analyse comparative de grands jeux de données de SNP, il est possible d'identifier les signatures moléculaires de la sélection naturelle telles que les balayages sélectifs et de déterminer comment et où la sélection affecte la variabilité génétique.

Figure 5 : Méthodes pour détecter un balayage sélectif à l'échelle intraspécifique. (A) Représentation des méthodes basées sur le spectre de fréquence. Les deux figures de gauche représentent deux spectres de fréquence; la première figure (en haut) représente la distribution des allèles dérivés sous neutralité (avant la sélection) et la seconde figure (en bas) représente la distribution des allèles dérivés après un balayage sélectif. Le spectre de fréquence montre qu'il y a à la fois un excès d'allèles dérivés rare et un excès d'allèles dérivés en forte fréquence. Les haplotypes représentent l'évolution des fréquences des allèles lorsqu'une mutation avantageuse apparait. Les cercles correspondent aux mutations ; les cercles noirs sont des mutations neutres et les cercles rouges les mutations sous sélection positive. La mutation avantageuse augmente en fréquence au cours du temps jusqu'à atteindre la fixation. La mutation sélectionnée est accompagnée de mutations neutres qui augmentent également en fréquence par un effet d'hitch-hiking et on observe une diminution de la diversité autour de la mutation sélectionnée jusqu'à ce que de nouvelles mutations apparaissent (cercles bleus, violets et orange) en faibles fréquences. (B) Un balayage sélectif provoque une augmentation de l'« *Extended Haplotype Homozygosity* » (EHH) qui est une mesure de déséquilibre de liaison qui prend en compte les haplotypes contenant l'allèle sélectionné. L'EHH augmente au cours de la sélection et atteint un plateau puis ce plateau diminue lorsque de nouvelles mutations neutres apparaissent (mutations bleus, violettes et orange). (C) Représentation des méthodes basées sur la différenciation entre sous-populations avec la mesure du F_{ST}. Les différences de fréquences alléliques entre sous-populations reflètent l'action de la sélection naturelle à l'échelle des sous-populations (adapté de Vitti et al., 2013).



Divers tests statistiques ont été développés et les méthodes les plus couramment employées se classent en trois groupes principaux : (i) les approches basées sur les propriétés du spectre de fréquence des allèles (Site Frequency Spectrum – SFS) qui se réfère à la distribution des fréquences des positions polymorphismes de séquences d'ADN via les SNP (Boitard et al., 2009; Nielsen, 2005; Pavlidis & Alachiotis, 2017) (**Figure 5A**), (ii) les approches basées sur la taille et la structure des haplotypes (Sabeti et al., 2002; Voight et al., 2006) (**Figure 5B**), (iii) ainsi que les méthodes qui reposent sur la différenciation génétique des populations, avec notamment la mesure du F_{ST} (Beaumont & Nichols, 1996; Bonhomme et al., 2010; Excoffier et al., 2009; Fariello et al., 2013; Lewontin & Krakauer, 1973) (**Figure 5C**). Les méthodes basées sur le SFS permettent de détecter les différents types de signatures de sélection (balayage sélectif, sélection purifiante et sélection balancée), cependant elles requièrent la connaissance du statut ancestral/dérivé des allèles à chaque SNP et peuvent être sensibles à des scénarios démographiques « mimant » les effets de la sélection (goulot d'étranglement, expansion de population, structuration génétique des populations). Les méthodes basées sur la différenciation génétique des populations et sur la taille des haplotypes permettent essentiellement de détecter divers régimes de sélection positive (« soft sweep », « hard sweep ») dans un contexte d'adaptation locale des populations mais sont moins sensibles aux effets démographiques.

La sélection naturelle laisse un certain nombre d'empreintes sur le génome et chacun de ces tests statistiques est conçu pour capter ces différents signaux. Les approches de génétique des populations ainsi décrites exploitent exclusivement des données génétiques et elles sont notamment complémentaires aux approches de génétique d'association qui visent à identifier les déterminants génétiques de la variation phénotypique. **Cependant, ces approches testent l'hypothèse de neutralité <u>par locus</u> mais ne prennent pas en compte <u>l'interaction entre les allèles à différents locus</u> et le « background génétique » dans lequel évoluent les mutations ciblées par la sélection naturelle**. Or les gènes d'un même génome peuvent former différents réseaux d'interactions, et de ce fait, une mutation sur un gène peut avoir des conséquences fonctionnelles, et donc évolutives, sur les autres gènes du réseau (Hansen, 2013). Par exemple, en réponse à un signal externe, l'activation ou la (co)régulation de l'expression de beaucoup de gènes est précédée de cascades de signalisation et de mécanismes transcriptionnels impliquant des interactions entre protéines ainsi qu'entre

protéines et acides nucléiques. Ainsi, les gènes dont les protéines sont en interaction peuvent être sensibles aux mutations qui pourraient apparaître sur l'une ou l'autre des séquences de ces gènes. Les effets de l'interaction entre les allèles à ces différents gènes sont une caractéristique importante pour l'étude des réseaux et cela a notamment permis d'ordonner les gènes au sein des réseaux (Avery & Wasserman, 1992; Phillips, 2008; Phillips et al., 1997). Le métabolisme cellulaire étant plutôt bien caractérisé, cela a également permis d'étudier l'organisation hiérarchique des réseaux métaboliques et de montrer l'importance des interactions entre gènes. Il a été montré que les réseaux métaboliques de la levure et d'*Escherichia Coli* sont caractérisés par ces interactions (He et al., 2010; Jagdishchandra Joshi & Prasad, 2014; Segrè et al., 2005). De nombreuses études montrent que les interactions entre gènes sont ubiquitaires et elles peuvent être utilisées pour comprendre l'évolution de réseaux géniques complexes. **Dans un cadre évolutionniste, nous nous intéresserons aux allèles présentant des interactions « adaptatives » entre locus et qui seront donc sous l'influence de la sélection, ce qui nous amène à considérer le phénomène d'épistasie**.

3. L'épistasie

3.1 Définition de l'épistasie

Le terme « épistasie » a été introduit par William Bateson au début du 20ème siècle pour décrire l'effet de masquage d'un allèle par un autre (Phillips, 1998). L'épistasie décrit le fait qu'un allèle à un locus puisse bloquer l'expression d'un allèle à un autre locus. Par exemple, si l'on considère deux locus bi-alléliques B et G responsables de la couleur des poils chez la souris, qui possèdent respectivement les allèles *B*/*b* et *G*/*g* (Cordell, 2002) :

Tableau 1: Exemple d'interaction épistatique selon la définition de Bateson (1909). Le phénotype de couleur des poils chez la souris est obtenu à partir de différents génotypes sur deux locus en interaction.

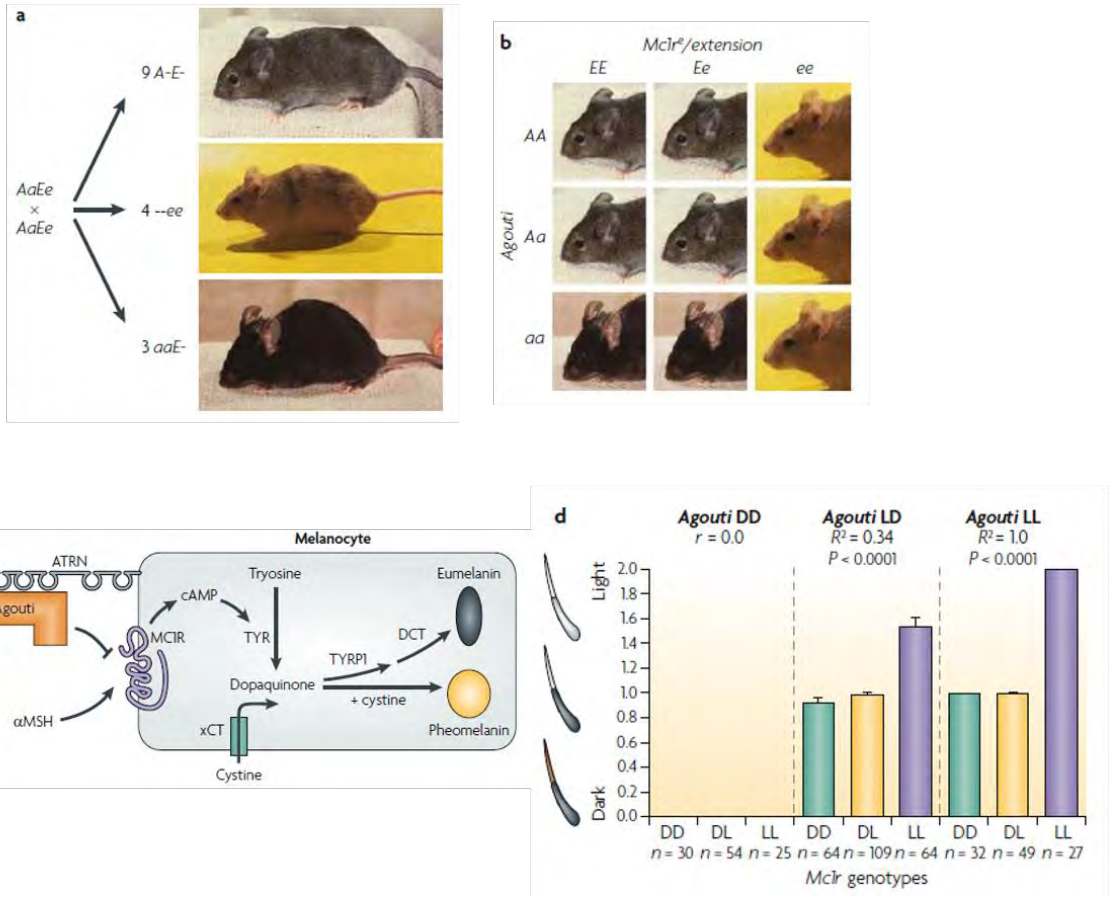
	Genotype at locus G		
Genotype at locus B	*g/g*	*g/G*	*G/G*
b/b	White	Grey	Grey
b/B	Black	Grey	Grey
B/B	Black	Grey	Grey

Les trois phénotypes possibles (poils blanc, noir ou gris) sont représentés dans le **Tableau 1** : les individus qui portent l'allèle *G* au locus G ont un pelage gris quel que soit le génotype au locus B. L'allèle *G* est dominant car il masque tout effet de l'allèle *g*. De plus, si le génotype au locus G est *g/g*, les individus qui possèdent l'allèle *B* au locus B ont un pelage noir. L'allèle *B* est dominant sur l'allèle *b* mais si toutefois le génotype au locus G n'est pas *g/g*, le phénotype du locus B n'est pas observable car tous les individus possédant un allèle *G* ont un pelage gris quel que soit le génotype au locus B. L'effet du locus B est donc masqué par le locus G et celui-ci est donc épistatique sur le locus B (Cordell, 2002). Le terme d'épistasie a ensuite pris un sens plus général au cours du 20ème siècle. Il décrit la notion d'interaction entre gènes ou la dépendance des effets de différents locus combinés qui vont aboutir à un phénotype. Dans la littérature, il existe plusieurs définitions de l'épistasie et on utilise la même terminologie pour décrire en fait trois concepts interconnectés : (a) les relations fonctionnelles entre gènes ou l'épistasie fonctionnelle, (b) les relations génétiques entre locus ou épistasie compositionnelle ou biologique (Cordell, 2002; Phillips, 2008), et enfin (c) l'épistasie statistique qui est définie à la fois en génétique quantitative et en génétique des populations. En génétique quantitative, l'épistasie statistique désigne les relations non-additives entre locus où l'interaction est définie comme un écart à l'additivité et en génétique des populations, l'épistasie statistique est définie comme la non-indépendance des valeurs sélectives à différents locus. L'ensemble des définitions de l'épistasie impliquent des relations entre gènes selon différents points de vue. La **Figure 6** illustre les différentes définitions de l'épistasie à l'aide d'un exemple bien caractérisé entre deux locus responsables de la couleur du pelage chez la souris.

L'**épistasie fonctionnelle** désigne les interactions moléculaires entre protéines. C'est une description strictement fonctionnelle qui ne fait pas de lien direct avec les interactions génétiques même si de manière logique nous devrions observer une perturbation de l'interaction fonctionnelle entre deux protéines si des modifications au niveau génétique (mutations) se manifestaient (**Figure 6C**). L'**épistasie biologique ou compositionnelle** se réfère à la définition de l'épistasie proposée par Bateson (Bateson, 1909) qui décrit l'effet de masquage d'un allèle à un locus par un autre allèle situé sur un second locus. Cela peut être vu comme un élargissement du concept de dominance entre locus (**Figure 6A,B**). Comme les interactions entre gènes peuvent être différentes, plusieurs termes ont été développés pour décrire des formes particulières d'épistasie. Par exemple, l'épistasie est dite « synergique » si

l'effet d'une mutation est renforcé par la présence d'une autre mutation à un autre locus et à l'inverse, l'épistasie est dite « antagoniste » si l'effet d'une mutation est plus faible en présence d'une autre mutation à un autre locus (Weinreich et al., 2005). De manière générale, l'épistasie biologique existe lorsque l'expression d'un allèle à un locus est dépendante de la présence ou de l'absence d'un allèle à un autre locus (Niel et al., 2015) (**Figure 6B**).

Figure 6 : Les différentes définitions de l'épistasie à travers l'exemple du pelage. La couleur du pelage chez les mammifères a souvent été utilisée pour illustrer les différents concepts liés à l'épistasie. (**a**) Exemple de croisement entre les locus A et E bi-alléliques (*A/a* et *E/e*) responsables de la couleur du pelage chez la souris. Le croisement entre deux individus doubles hétérozygotes produit des ratios, à la génération suivante, qui ne sont pas Mendéliens (9:4:3 au lieu de 9:3:3:1, au niveau des associations de gamètes). Il y a un excès d'individus au pelage marbré (jaune) montrant qu'en absence de l'allèle *E*, l'allèle *e* est épistatique sur le locus A (*Agouti* et *Mc1r*^{e}/extension sont les noms des locus A et E). (**b**) Les résultats de ce croisement sont présentés dans une table d'interaction des génotypes (3*3), illustrant l'***épistasie compositionnelle***. (**c**) La biochimie de cette voie dans les mélanocytes a été caractérisée: l'activation de *Mc1r* initie la production de l'eumélanine (couleur sombre) par opposition à la pheomélanine (couleur jaune/beige). *Agouti* agit comme antagoniste à *Mc1r* de sorte que l'activation périodique de *Agouti* entraine la formation de bandes de couleurs sur le pelage (pelage marbré-jaune des génotypes *ee*). Les interactions entre les protéines de cette voie sont représentatives de ***l'épistasie fonctionnelle***. (**d**) Steiner et al. (2007) ont analysé le polymorphisme génétique à partir des données de cette voie biologique (i.e. *Agouti* et *Mc1r*^{e}*/extension*) entre les souris des forêts au pelage sombre et les souris des plages au pelage clair. Ils ont montré que la transition entre le pelage sombre et le pelage clair qui a accompagné le mouvement des souris de la forêt à la plage, s'est faite par une interaction, associée à des changements structuraux et de régulation, entre les locus *Agouti* et *Mc1r*^{e}*/extension*. Les effets de ces interactions génétiques ont été quantifiés pour les génotypes au marqueur *Mc1r*^{e} en faisant la moyenne sur chaque fond génétique (génotypes) *Agouti* échantillonné, afin de donner une estimation de l'***épistasie statistique*** entre ces locus. Les figures sont extraites de Phillips et al (2008).



En génétique moléculaire ou en biochimie, l'épistasie telle qu'elle est décrite ci-dessus est généralement observée pour des gènes qui produisent séquentiellement des substrats ou des catalyseurs d'une même voie biochimique. Ces interactions épistatiques ont également pu être montrées au cours du développement pendant l'activation séquentielle de gènes (Phillips, 2008). L'**épistasie statistique en génétique quantitative** est un concept populationnel qui décrit les relations de non-additivité entre les variants génétiques de plusieurs locus par rapport à leurs contributions à un phénotype (Fisher, 1918) (**Figure 6D**). Ces relations génétiques peuvent être définies selon un modèle de régression linéaire qui pour un organisme haploïde peut s'écrire $P = b_1A + b_2B + b_{12}AB + e$, où b_1 et b_2 sont les effets moyens des allèles aux locus A et B sur le phénotype P et b_{12} décrit l'effet de l'interaction entre les allèles aux deux locus. Le terme e reflète simplement les variations stochastiques provenant de l'environnement ou de locus non pris en compte. Les relations entre locus décrites par ces modèles de régression linéaire sont plus complexes dans les populations diploïdes car ils prennent en compte les effets de dominance entre allèles. Ainsi, les interactions épistatiques sont un écart à l'additivité car l'effet combiné des allèles à deux locus est différent de la somme des effets de ces allèles. Dans ce modèle, les combinaisons d'allèles aux locus A et B influencent le phénotype (P) et en absence d'épistasie, b_{12} est égal à zéro c'est-à-dire que les allèles aux deux locus ont des effets additifs (Phillips, 2008; Wade et al., 2001). D'autre part, **en génétique des populations**, l'épistasie statistique est définie comme la non-indépendance des valeurs sélectives (ou *fitness*) des allèles à différents locus. Un modèle fréquemment utilisé est le modèle multiplicatif qui donne une définition probabiliste de l'indépendance (Phillips, 2008). En effet, deux locus sont sélectivement indépendants (i.e. pas d'épistasie) si la *fitness* des combinaisons génotypiques à ces deux locus est égale au produit des *fitness* des génotypes de chaque locus. Par exemple, si l'on considère deux locus haploïdes A et B où les allèles *A* et *B* possèdent chacun, respectivement, une valeur sélective $W_A = (1 - s)$ et $W_B = (1 - t)$ (s et t sont les coefficients de contre-sélection, avec $0 < s, t < 1$), alors ces locus sont sélectivement indépendants si la valeur sélective de l'haplotype AB est $W_{AB} = (1 - s)(1 - t)$. En présence d'épistasie, la valeur sélective du génotype sera égale à $W_{AB} = (1 - s)(1 - t) + \varepsilon$, avec ε la déviation épistatique (Phillips et al., 1997; Wade et al., 2001). L'épistasie est modélisée en ajoutant un terme de *fitness* sur l'interaction entre les génotypes aux deux locus (ε, positif ou négatif) et la *fitness* combinée des allèles à deux locus est différente du produit des *fitness* de ces allèles.

Ces travaux de thèse abordant l'épistasie sous l'angle de la génétique des populations, nous avons modélisé l'épistasie par la définition de valeurs de *fitness* pour différentes associations d'allèles à différents locus, telles que proposées par les travaux les plus récents dans le domaine (Takahasi, 2009; Takahasi & Tajima, 2005 ; voir la section suivante et le chapitre 1.1). Comme cela a pu être montré, la sélection sur ces génotypes multi-locus va augmenter ou créer de la corrélation entre les allèles des locus sélectionnés (Felsenstein, 1965; Phillips et al., 1997). On parle alors de **sélection épistatique**. La détection de génotypes ainsi cosélectionnés se base sur les fréquences alléliques et/ou génotypiques observées au sein des populations étudiées à ces différents locus. Les mesures de la sélection épistatique seront donc aussi influencées par les autres forces évolutives ainsi que par les systèmes de reproduction (Phillips et al., 1997).

3.2 La sélection épistatique

En génétique des populations, la sélection épistatique désigne les interactions adaptatives entre des variants génétiques qui ségrégent à différents locus. L'étude de ces interactions repose sur des modèles génétiques de sélection épistatique entre des locus non liés (Takahasi, 2009; Takahasi & Innan, 2008; Takahasi & Tajima, 2005). Dans une population panmictique, deux modèles de sélection épistatique sont considérés pour deux locus bi-alléliques A et B, avec *A* et *B* les allèles ancestraux initialement présents dans la population et *a* et *b* les allèles mutants (voir chapitre 1.1 pour plus de détails) : le modèle de **coadaptation**, où *a* et *b* sont individuellement neutres mais forment un haplotype coadapté avec une meilleure *fitness* lorsqu'ils sont combinés chez un même individu ; et le modèle **compensatoire**, où les deux mutations *a* et *b* sont individuellement délétères mais induisent une *fitness* normale (équivalent à l'haplotype ancestral *AB*) quand elles sont combinées chez un même individu (*ab*). Il a été montré que dans ces deux modèles, l'interaction adaptative entre deux locus non liés génère du **déséquilibre de liaison (DL)** par l'association non aléatoire des allèles à différents locus. Dans ce contexte, la détection de la sélection épistatique entre locus non liés se fait à l'aide du DL et des simulations ont été réalisées pour étudier les mécanismes évolutifs de ces interactions adaptatives ainsi que pour évaluer les outils statistiques permettant leur détection. Dans les modèles de simulation, la détection de la sélection épistatique se fait à partir de locus simulés indépendants c'est-à-dire sur des

chromosomes différents ou ayant des taux de recombinaison importants afin de limiter les biais dus à la liaison physique. Takahasi et Tajima (Takahasi & Tajima, 2005) ont étudié la dynamique de fixation de l'haplotype *ab* double mutant en fonction des fréquences initiales de ces allèles mutants cosélectionnés et ils ont pu montrer que la sélection épistatique ne peut promouvoir à elle seule l'évolution d'un système épistatique mais que l'accumulation préalable de mutations neutres qui ségrégent dans la population (« standing variation ») crée un contexte génétique favorable pour ce type d'évolution dans les générations futures. Si les mutations cosélectionnées sont des mutations *de novo* en faibles fréquences, la sélection épistatique sera faiblement efficace. Aussi, il a été montré que l'intervalle de temps entre l'apparition des deux mutations *a* et *b* influence l'intensité de la sélection épistatique. En effet, dans un modèle à deux locus avec deux mutations coadaptées, si l'intervalle de temps est long entre l'apparition des deux mutations, la première va évoluer sous neutralité et risque d'être perdue ou fixée par dérive génétique avant que la seconde mutation apparaisse et que la sélection épistatique entre en jeu (Takahasi, 2009).

Pour détecter la sélection épistatique entre paires de locus bi-alléliques, les statistiques classiques de DL comme le coefficient de corrélation r, r^2 et la statistique D peuvent être utilisées (voir chapitre 1.2). Il a été montré que la statistique *r* a une meilleure puissance de détection si l'état ancestral ou dérivé (i.e. mutant) des allèles est connu. En effet, pour deux locus A et B avec *a* et *b* comme allèles dérivés, l'interaction adaptative positive entre ces deux allèles va créer un excès d'haplotypes *ab* doubles mutants (et donc du double sauvage *AB*) par rapport aux haplotypes recombinants *Ab* et *aB*. Takahasi et Innan (Takahasi & Innan, 2008) ont montré par simulations que la mesure directionnelle du DL permet d'évaluer si les associations d'allèles dérivés *ab* (ou ancestraux *AB*) sont en excès par rapport aux associations recombinantes. Les auteurs ont pu détecter la sélection épistatique avec la statistique *r* aussi bien dans le modèle coadapté que dans le modèle compensatoire. Plus récemment, Id Lahoucine et al. (Id-Lahoucine et al., 2019) ont évalué la détection de signatures de sélection épistatique à partir de données génomiques simulées et structurées en sous-populations dont le régime de sélection épistatique diverge entre les sous-populations. Pour mesurer le déséquilibre de liaison, ils ont utilisé la statistique D'^2_{IS} (voir chapitre 1.2.2) précédemment introduite par Otha (Ohta, Mar 1982a, May 1982b). Cette analyse présente toutefois des limites pour distinguer la sélection épistatique de la sélection

indépendante (ou additive) entre deux locus, ainsi que pour réduire la détection de faux positifs.

L'ensemble de ces travaux sont des contributions importantes à la compréhension des mécanismes d'adaptation impliquant la sélection épistatique dans les populations. Cependant, ils sont basés sur un modèle de reproduction panmictique et il a été montré que le mode de reproduction peut influencer la dynamique d'adaptation (Glémin & Ronfort, 2013; Ronfort & Glemin, 2012) et accroître le DL longue distance. En effet, l'autogamie crée du DL longue distance dans le génome car la recombinaison homologue s'effectue entre des segments chromosomiques identiques (homozygotes) ou presque (Burgarella & Glémin, 2017). De plus, le DL est également influencé par la structure des populations et la dérive génétique du fait de l'isolement et de la variation de taille des populations (Grillo et al., 2016; Ronfort et al., 2006; Slatkin, 2008). Afin de considérer les biais dus à la structure de la population et à l'apparentement entre les individus, la statistique r_v^2 a été proposée pour mieux estimer le DL physique entre des paires de marqueurs situées sur un même chromosome en incluant une matrice d'apparentement *V* qui tient compte de l'apparentement entre les individus (Mangin et al., 2012). Le r_s^2 est également une mesure corrigée du r^2 qui tient compte de la structure des populations, elle est équivalente au r/a, la mesure du « ancestry disequilibrium (AD) » (voir chapitre 1.2.2). Ces deux statistiques utilisent l'information de structure entre les sous-populations comme covariable pour l'inclure dans le calcul du déséquilibre de liaison (corrélation partielle). La valeur de DL est ainsi pondérée en tenant compte du fait que les individus d'une même population ont tendance à partager plus d'allèles que les individus provenant de populations différentes (Mangin et al., 2012; Schumer & Brandvain, 2016). Bien que la statistique r/a ait permis de réduire le nombre de faux positifs dans les scans génomiques et détecter des locus pour lesquels la sélection épistatique et non la structure des populations maintien des combinaisons alléliques ancestrales, un excès de faux positifs était toujours observé.

3.3 Exemples de sélection épistatique dans la littérature

Il existe quelques exemples de recherche de sélection épistatique dans la littérature. Chez les plantes, les études utilisent généralement des approches avec des gènes candidats identifiés, par exemple en GWAS. Une étude chez *Arabidopsis thaliana* a mis en évidence des

interactions adaptatives entre des gènes de réponse à l'herbivorie (Brachi et al., 2015) et une autre a montré des interactions adaptatives entre les gènes *FRI* et *FLC* impliqués dans la floraison (Caicedo et al., 2004). Une autre étude, sur la tomate sauvage, a montré que trois gènes (Pto, Fen, Prf) du complexe de résistance *Prf* sont eux aussi soumis à la sélection épistatique (Grzeskowiak et al., 2014). Pour mesurer les interactions adaptatives entre les gènes, ce sont des statistiques classiques de DL qui ont été utilisées entre des paires de SNP. Pool (Pool, 2015) a ensuite testé une nouvelle approche chez la drosophile. Dans son étude, il a mesuré des associations préférentielles entre des paires d'haplotypes en calculant l'« Ancestry Disequilibrium » entre des locus non liés à l'échelle génomique et il a pu montrer l'importance de la sélection épistatique sur les variations génétiques entre différentes populations de drosophile. Chez l'homme, il existe également des études qui ont été menées pour rechercher de la sélection épistatique à l'échelle du génome en utilisant le déséquilibre de liaison. Hu & Hu (2015) ont comparé des paires de SNP dans différentes populations et ils ont pu montrer que certains SNP identifiés sous sélection épistatique à l'échelle des sous-populations (Afrique, Europe ou Asie) ne sont globalement pas retrouvés à l'échelle de la population entière, montrant ainsi des signaux de sélection et d'adaptation locale (Hu & Hu, 2015). Koch et al. (2013) ont mesuré le DL longue distance (LRLD – « long range LD ») sur l'ensemble des chromosomes humains de la population YRI (« the Yoruba in Ibadan », Nigeria) et ils ont pu identifier du LRLD significatif sur l'ensemble des 22 chromosomes (Koch et al., 2013) pouvant être associé à des signatures de sélection épistatique. Enfin, Daub et al. (Daub et al., 2013) ont mis en évidence des interactions adaptatives entre des gènes impliqués dans les mêmes voies biologiques reliées notamment au système immunitaire. Plus récemment, Zan et al. (2018) ont mesuré le DL à l'échelle globale chez *Arabidopsis thaliana* en construisant des haplotypes sur des fenêtres génomiques. Ils ont notamment recherché des signatures de sélection épistatique à partir de marqueurs préalablement identifiés par des analyses de GWAS réalisées à partir de la concentration de molybdène dans les feuilles. L'analyse GWAS a permis d'identifier 4 marqueurs sur le chromosome 2 et ils ont pu trouver un signal de sélection épistatique entre l'un de ces 4 marqueurs du chromosome 2 (*2 : 10 928 720 bp*) et un marqueur sur le chromosome 1 (*1 : 5 315 502 bp*). À partir de ce résultat, ils ont montré que les accessions qui portent le génotype AA au marqueur du chromosome 1 ont des concentrations intermédiaires en molybdène tandis que les accessions qui ont un génotype GG à ce même marqueur (chromosome 1) ont des concentrations différentes en molybdène

dans les feuilles qui dépendent du génotype (CC ou TT) au marqueur du chromosome 2 (*2 : 10 928 720 bp*). Avec ce résultat, ils ont pu observer des relations épistatiques entre des marqueurs génétiques non liés. Dans le contexte d'analyse GWAS, une étude menée chez *Arabidopsis thaliana* à laquelle j'ai pu contribuer durant ma thèse, a révélé des interactions épistatiques entre des QTL identifiés par GWAS pour la réponse des plantes à la bactérie pathogène *Ralstonia solanacearum* dans différentes conditions de température (Aoun et al. 2020, sous-presse).

Bien que les exemples de sélection épistatique énoncés ci-dessus sont interprétables d'un point de vue statistique, les mécanismes biologiques et fonctionnels sous-jacents ne sont pas toujours explicites. Il existe quelques exemples d'interactions épistatiques validées biologiquement parmi lesquels on retrouve celui décrit chez la levure *Saccharomyces cerevisiae.* Il a été montré que la sporulation de *S. cerevisae* est régulée par 3 SNP situés respectivement dans les régions promotrices des gènes RME1 et IME1 et dans le gène IME1. Ces 3 SNP sont en épistasie et la liaison physique entre la protéine RME1, facteur de transcription, et la région promotrice du gène IME1 dépend des combinaisons alléliques à ces SNP. Ces combinaisons influencent l'efficacité de sporulation (Gertz et al., 2010; Niel et al., 2015). Un autre exemple d'interaction épistatique a été décrit chez le virus du VIH pour le maintien des structures secondaires des ARN viraux dont les interactions entre nucléotides sont sous sélection épistatique. Si une mutation apparait au niveau d'un nucléotide où la structure secondaire se forme, l'appariement Watson-Crick pour former cette structure ne pourra plus se faire et la sélection épistatique agit alors pour favoriser le maintien de ces interactions entre nucléotides (Assis, 2014).

Un défi majeur dans la recherche de signatures de sélection épistatique est de pouvoir interpréter les signatures génomiques, telles que le DL longue distance, d'un point de vue biologique car les interactions statistiques ne sous-entendent pas automatiquement des interactions physiques. Donc, si une interaction statistique est identifiée entre deux gènes, elle peut être causée par une interaction biologique qui elle-même peut s'effectuer de diverses façons. En effet, l'interaction peut être fonctionnelle (directe ou indirecte) entre deux protéines d'une même voie biologique, comme cela a été montré chez *A. thaliana* entre les protéines de la voie du glucosinolate impliquées dans la réponse à l'herbivorie (Brachi et al., 2015). L'interaction peut également être physique, entre deux protéines, entre une protéine

et une séquence d'ADN (Gertz et al., 2010; Niel et al., 2015) ou entre deux séquences d'ADN ou d'ARN (Assis, 2014). Par ailleurs, la cosélection de gènes est possible au sein d'une même voie biologique, ou de sous-réseaux génétiques, sans pour autant qu'une interaction soit démontrée (Daub et al., 2013, 2015; Gouy et al., 2017). À l'échelle de réseaux moléculaires d'interactions, il a été montré, chez l'homme, que des gènes sous sélection positive cartographiés dans des réseaux d'interactions protéine-protéine sont plus souvent localisés à proximité, indiquant un « clustering » des gènes sélectionnés définis à partir des réseaux de gènes(Qian et al., 2015). À contrario, l'ensemble des interactions protéine-protéine, protéine-ADN ou ADN-ADN, qu'elles soient physiques ou fonctionnelles, ne sont pas nécessairement cosélectionnées, notamment en raison de plasticités conformationnelles et de redondances fonctionnelles fréquentes. Ceci suggère de ne pas réduire le concept d'interaction entre gènes à une stricte épistasie évolutive. De plus, les réseaux de gènes ou de molécules en interactions présentent des organisations hiérarchiques, topologiques et modulaires impliquant des degrés d'interaction très variables (Baryshnikova et al., 2013; Boucher & Jenna, 2013; Pritykin & Singh, 2013).

Un autre défi important se pose en amont de l'interprétation biologique ou fonctionnelle des signaux de sélection épistatique ; c'est celui d'améliorer la confiance en ces signaux en adaptant les outils statistiques de détection de ce type de sélection. En effet, les statistiques de DL qui sont utilisées sont influencées par des facteurs démographiques tels que la structure des populations et les différents systèmes de reproduction. Les méthodes de détection doivent tenir compte de ces facteurs, autre que la sélection épistatique, qui influencent le DL afin de permettre une identification plus fiable des gènes sous sélection épistatique.

4. Projet de thèse

4.1 Contexte scientifique et modèle biologique

Le Laboratoire de Recherche en Sciences Végétales (LRSV) où j'ai effectué ma thèse est un laboratoire de biologie végétale où les espèces les plus étudiées sont *Arabidopsis thaliana* et *Medicago truncatula*. *M. truncatula* est l'espèce modèle des légumineuses pour l'étude des interactions entre les plantes et les micro-organismes, notamment pour l'étude de la

symbiose racinaire avec les bactéries fixatrices d'azote du genre *Rhizobium* mais aussi de la symbiose endomycorhizienne avec des champignons à arbuscules. Plusieurs études portent également sur la résistance d*e M. truncatula* face aux pathogènes et notamment la résistance face à l'oomycete *Aphanomyces euteiches* (Bonhomme et al., 2014, 2019). *M. truncatula* est un bon modèle pour l'étude des bases génétiques de l'adaptation dans le contexte de l'interaction entre les plantes et leur environnement biotique. De nombreuses ressources génétiques sont disponibles dont les collections de mutants, ainsi que plusieurs millions de marqueurs SNP, répartis sur les 8 chromosomes de *M. truncatula* et identifiés dans une collection de plus de 262 accessions (Ronfort et al., 2006). Ces accessions ont été collectées sur le pourtour du bassin méditerranéen et ont été séquencées par le Medicago HapMap Project (http://www.medicagohapmap.org/). Ces données sont disponibles en téléchargement libre (http://www.medicagohapmap.org/downloads/mt40). Parmi les 262 accessions séquencées, deux populations clairement distinctes génétiquement ont été identifiées (Bonhomme et al., 2014; Burgarella et al., 2016; De Mita et al., 2011; Ronfort et al., 2006) ; la population « Far-West » (FW) qui se situe à l'ouest du bassin Méditerranéen et la population « Circum » (C) qui est répartie tout autour du bassin Méditerranéen avec seulement quelques accessions à l'ouest. Dans le même temps, le génome de référence A17 de *M. truncatula* a également été publié en haute qualité (Pecrix et al., 2018; Tang et al., 2014; Young et al., 2011). Ces données sont à l'origine d'un certain nombre d'études d'associations (GWAS) et de génomique des populations (GWSS). La première recherche de signatures de sélection à l'échelle du génome de *Medicago truncatula* a été réalisée sur environs 20 000 gènes annotés avec la statistique d_N/d_S (Paape et al., 2013). Cette étude a montré qu'environ 1% des gènes présentent des signatures de sélection positive tandis que 50 à 75% des gènes présentent des signatures de sélection purifiante. Une autre étude visant à détecter des signatures de balayages sélectifs à l'échelle du génome chez *Medicago truncatula* a aussi montré qu'une faible proportion de gènes présente des signatures de « hard sweep ». Cette analyse de GWSS, combinée à l'annotation fonctionnelle et aux données d'expression des gènes, a permis d'identifier un set de gènes soumis à de fortes pressions de sélection positive dans un contexte d'interactions biotiques avec des micro-organismes mutualistes et pathogènes (Bonhomme et al., 2015). D'autre part, les premières GWAS réalisées chez *Medicago truncatula* à partir du jeu de données de SNP issu du projet Hapmap l'ont été sur des phénotypes de nodulation (Stanton-Geddes et al., 2013) et de résistance à l'oomycète

pathogène *Aphanomyces euteiches* (Bonhomme et al., 2014, 2019). Depuis, diverses analyses d'associations ont été réalisées sur différents traits phénotypiques parmi lesquels on peut citer la résistance à l'oomycète *Phytophthora palmivora* (Rey et al., 2017), la teneur en protéine des semences (Signor et al., 2017) ou la résistance à un stress osmotique (Kang et al., 2015).

4.2 Objectifs de la thèse

Dans la continuité de ces travaux, mon projet de thèse s'inscrit dans le projet de recherche ANR DeCoD ("Detecting Networks of coadapted genes by genome-scale analysis of DNA polymorphisms: application to the model legume *Medicago truncatula*"), dont le premier objectif a été de développer un test statistique permettant d'identifier par des signatures de sélection épistatique des gènes ou régions génomiques coadaptés. L'objectif suivant était de contribuer à identifier chez *Medicago truncatula* de nouveaux réseaux géniques impliqués dans l'adaptation aux micro-organismes ainsi que de caractériser de nouveaux gènes candidats en association avec des gènes connus.

Mon projet de thèse s'articule en deux grands chapitres. Dans le premier chapitre méthodologique, l'objectif a été de développer un test statistique pour détecter la sélection épistatique à l'aide du DL en tenant compte de la structure des populations et de l'apparentement entre les individus. Nous avons décrit les différentes statistiques utilisées, puis, avons réalisé des simulations à l'échelle de génomes diploïdes. Il s'agissait de comprendre les dynamiques évolutives de plusieurs régimes de sélection, l'effet de la structuration des populations, de deux régimes de reproduction contrastés, ainsi que de plusieurs modes d'interaction entre les allèles aux locus sélectionnés. Ainsi, quatre modèles de sélection ont été simulés ; deux modèles de sélection épistatique coadapté et compensatoire, un modèle de sélection additive et un modèle neutre. Deux régimes de reproduction ont été simulés ; un modèle panmictique et un modèle à 95% d'autofécondation. Et enfin, trois modes d'interaction entre les allèles de chaque locus sélectionné ont été simulés ; dominance, codominance et récessivité. L'ensemble de ces paramètres a été simulé afin de comprendre l'influence de chacun sur la dynamique évolutive des mutations sous sélection épistatique. Dans cet objectif, nous avons, dans un premier temps, suivi l'évolution des fréquences des allèles cosélectionnés et nous avons calculé les taux de cofixation. Ensuite,

l'objectif des simulations a été d'évaluer si les statistiques de DL que nous avons développées permettent effectivement de détecter la sélection épistatique en tenant compte de l'effet des facteurs démographiques (structure des populations) et des régimes de reproduction (degré d'apparentement). Deux types de mesures ont été évalués ; une mesure de DL basée sur les comparaisons entre paires de SNP (r/r_v) et une mesure basée sur les comparaisons entre paires de fenêtres génomiques (cor_{PC1}/cor_{PC1v}). Enfin, dans le but d'évaluer si la sélection épistatique influence le polymorphisme des locus et si ces locus présentent également des signatures de sélection classiques, nous avons calculé des statistiques de tests de neutralité (*D*, *H*, *E*) sur les données simulées.

Dans le second chapitre, l'objectif a été d'identifier, chez *Medicago truncatula,* de nouvelles interactions adaptatives entre gènes par des approches ciblées, avec des gènes connues et par des approches systémiques, où tous les gènes du génome ont été analysés. Une analyse plus délimitée portant sur des données génétiques humaines est aussi présentée afin d'illustrer les possibilités d'application de l'approche à divers organismes. Dans une première partie nous présentons les données SNP utilisées ainsi que les populations analysées chez *Medicago truncatula* et chez l'homme. Dans une deuxième partie, nous décrivons les résultats de l'approche pangénomique ciblée que nous avons développé : l'analyse « Genome-Wide Epistatic Selection Scan » (GWESS). Les analyses de GWESS visent à calculer le DL entre un gène candidat appât et tous les autres gènes du génome à l'aide des statistiques de DL, r/r_v et cor_{PC1}/cor_{PC1v}, basées respectivement sur la comparaison de paires de SNP et de fenêtres génomiques (fixées à 10kb). Chez *Medicago truncatula*, nous avons particulièrement focalisé les analyses GWESS (avec les statistiques cor_{PC1}/cor_{PC1v}) sur un set de 98 gènes candidats caractérisés dans la littérature et impliqués, pour la plupart, dans les interactions entre les plantes et les micro-organismes. Ce sont, notamment, des gènes impliqués dans les réponses des plantes en présence de micro-organismes symbiotiques tels que les bactéries *Rhizobium* fixatrices d'azote lors de la nodulation, ainsi que les champignons endomycorhiziens à arbuscules (« Arbuscular Mycorrhizal Fungi ») lors de la mycorhization. Sur l'ensemble des résultats que nous avons obtenus, nous avons analysé plus en détails les résultats de GWESS pour trois gènes candidats : *MtSUNN, MtCRA2* et *MtNIN. MtSUNN et MtCRA2* sont deux récepteurs LRR-RLK (« LRR receptor-like kinase ») impliqués dans la régulation systémique de la formation des nodules pendant la symbiose rhizobienne et *MtNIN* est un facteur de transcription essentiel à l'activation du programme de développement du

nodule. L'analyse GWESS de *MtSUNN* comme appât a révélé une interaction adaptative avec le gène *MtCLE02.* Dans le cadre d'une collaboration avec l'IPS2 de Paris-Saclay (Equipe de F. Frugier), nous avons pu obtenir une preuve expérimentale sur le rôle de régulation négative de la nodulation par *MtCLE02* dépendant du gène *MtSUNN*. Chez l'homme, des analyses de GWESS ont également été mises en place à l'aide des statistiques r/r_v. Plusieurs gènes candidats ont été testés, et nous nous sommes particulièrement intéressés à deux gènes connus pour être sous sélection positive dans les populations humaines : les gènes *SLC24A5* et *EDAR*, respectivement impliqués dans la pigmentation de la peau et dans le développement des organes ectodermiques (cheveux, dents, ...). L'approche GWESS révèle une signature de sélection épistatique ou de cosélection entre ces deux gènes dans certaines populations du sud de l'Asie Centrale. Dans une troisième partie, après avoir calculé le DL entre toutes les paires de gènes du génome de *Medicago truncatula*, nous avons recherché des corrélations potentielles entre les signatures de sélection épistatique et les signatures de sélection « classiques ». Dans cet objectif, nous avons calculé les statistiques de tests de neutralité (*D*, *H*, *E*) sur l'ensemble des gènes de *M. truncatula* et nous avons comparé les résultats sur des ensembles de gènes neutres et sous sélection épistatique. Les résultats montrent que 30% des gènes potentiellement sous sélection épistatique chez *M. truncatula* présentent des signatures de sélection classiques, indiquant qu'une majorité de gènes en apparence « neutres » peuvent en fait être en interaction adaptative avec d'autres gènes. Enfin, dans une quatrième partie, nous avons exploré les signatures génomiques de sélection épistatique chez *M. truncatula* à travers deux approches systémiques. La première approche consiste à comparer les valeurs de DL entre des gènes candidats de même voie biologique (symbiose rhizobienne, symbiose mycorhizienne) ou de même fonction moléculaire (facteurs de transcription, récepteurs kinases, îlots symbiotiques - (Pecrix et al., 2018)), à celles d'un set de gènes sélectionnés aléatoirement. Les résultats montrent des enrichissements en signaux épistatiques dans les sets de gènes candidats. La seconde approche consiste à construire des réseaux génomiques d'interactions génétiques significatives entre gènes de *Medicago truncatula*. Nous décrivons quelques caractéristiques de ces réseaux, mais du fait de la masse de données générée, de la complexité et de la taille de ces réseaux, nous avons aussi analysé des sous-réseaux génomiques d'interactions significatives incluant des gènes connus et faisant partie du set de 98 gènes symbiotiques. Ces analyses permettent de visualiser des connexions directes entre certains gènes symbiotiques et les connexions indirectes via des interactions

intermédiaires avec d'autres gènes. L'analyse de ces sous-réseaux ouvre la voie vers la caractérisation biologique de nouveaux gènes de ces voies biologiques.

Dans l'objectif de faciliter la lecture et l'interprétation des résultats de ce manuscrit, les différentes parties qui composent ces deux chapitres - l'un méthodologique, l'autre plutôt analytique – comportent, chacune, des éléments de discussion nécessaires à la compréhension et à la critique des résultats. À la fin de ce manuscrit, une synthèse générale reprend les principaux résultats qui sont également discutés et mis en perspective, à l'aune de la littérature.

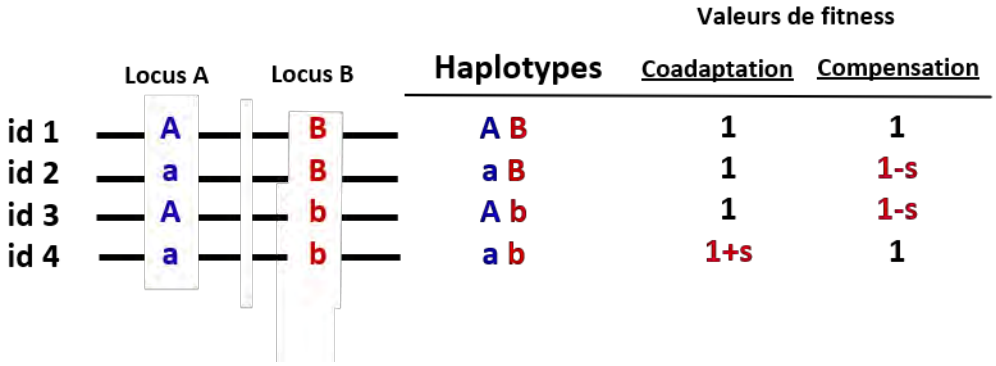
Chapitre 1 : Simulations génétiques et détection statistique de la sélection épistatique entre paires de locus

L'objectif de ce premier chapitre a été de développer une approche par simulation afin d'évaluer les capacités de détection statistique de la sélection épistatique à partir de données SNP simulées sur des locus indépendants. Parallèlement, et sur ces mêmes données simulées, nous avons évalué si les locus sous sélection épistatique présentent également des signatures de sélection de manière indépendante sur chacun des locus. Les deux premières parties de ce chapitre présentent les modèles génétiques de sélection épistatique ainsi que les outils statistiques qui ont été développés pour mesurer la sélection épistatique sur les gènes (déséquilibre de liaison et statistiques de tests de neutralité). Une brève revue bibliographique présente les statistiques qui permettent de mesurer le DL de façon classique mais aussi en tenant compte de l'apparentement entre les individus et de la structure génétique des populations. Cette partie nous permet d'introduire les statistiques de test qui ont été utilisées au cours de ce projet afin de détecter la sélection épistatique (i) entre des paires de locus bi-alléliques avec les mesures r et r_v et (ii) entre des paires de fenêtres génomiques avec les mesures cor_{PC1} et cor_{PC1v}. Puis, la troisième partie présente les simulations qui ont été réalisées. Deux modèles de sélection épistatique ont été simulés dans une population structurée en deux sous-populations et selon deux régimes de reproductions possibles ; un régime panmictique et un régime à 95% d'autofécondation. Enfin, la quatrième partie présente les résultats détaillés des simulations avec l'analyse de contrôle qualité des données générées (structuration génétique, apparentement, évolution des fréquences alléliques et taux de cofixation des allèles sélectionnés). Nous présentons ensuite les résultats de DL obtenus dans les différents modèles de sélection et dans le modèle neutre ainsi que les proportions de faux positifs des statistiques de DL et les puissances de détection. Enfin, nous abordons les résultats des statistiques de neutralité.

1.1 Présentation des modèles théoriques de sélection épistatique

La détection d'interactions adaptatives repose sur un modèle génétique de sélection épistatique entre deux locus non liés physiquement. Deux modèles de sélection épistatique ont été proposés : le modèle de coadaptation (Takahasi, 2009; Takahasi & Innan, 2008; Takahasi & Tajima, 2005) et le modèle compensatoire (Takahasi & Innan, 2008) (**Figure 7**).

Figure 7 : Présentation des modèles de sélection épistatique coadapté et compensatoire. Représentation des *fitness* attribuées aux individus en fonction des haplotypes aux SNP sous sélection épistatique. Les deux SNP A et B sont bi-alléliques et indépendants. Les allèles *A* et *B* sont les allèles sauvages ancestraux et *a* et *b* sont les allèles dérivés ou mutant. (Ce code ne signifie pas que A domine a (respectivement B/b)



1.1.1 Le modèle de coadaptation

Le modèle de coadaptation a été proposé par Takahasi et Tajima (2005) et Takahasi (2009). Dans ce modèle haploïde, deux mutations *a* et *b*, individuellement neutres sur 2 locus A et B indépendants, forment ensemble un haplotype coadapté. Pour deux locus bi-alléliques avec les allèles *A/a* et *B/b*, il y a quatre haplotypes possibles. Si *a* et *b* sont les mutations, les individus qui portent ces deux allèles ont une meilleure *fitness*. La *fitness*, ou valeur sélective (w) des individus portant cet haplotype est plus élevée, telle que

$$w_{AB} = 1$$

$$w_{Ab} = 1$$

$$w_{aB} = 1$$

$$w_{ab} = 1 + s$$

où s est le coefficient de sélection ($s \geq 0$). Au cours des générations, les individus portant l'haplotype *ab* seront de plus en plus fréquents car ils ont de meilleures capacités à survivre et se reproduire. Si seul l'allèle *a* est présent dans la population, cela n'aura pas d'effet sur la *fitness* mais quand l'allèle *b* apparaît, il y a une coévolution des fréquences de ces allèles. La combinaison de ces deux allèles est favorable et l'haplotype *ab* tend à augmenter sous l'effet de la sélection épistatique (Takahasi, 2009; Takahasi & Tajima, 2005). Dans une population diploïde, les *fitness* relatives aux génotypes des deux locus A et B cosélectionnés dans un

modèle coadapté dépendent de la récessivité, de la dominance ou de la codominance des allèles (**Tableau 2**).

1.1.2 Le modèle compensatoire

Dans le modèle compensatoire, deux mutations *a* et *b* individuellement délétères, se compensent lorsqu'elles sont combinées. Si une mutation apparaît à un locus, il ne pourra plus interagir avec le second locus, entraînant une baisse de *fitness*, sauf si une autre mutation se produit également au second locus pour rétablir l'interaction (i.e. compensation) :

$$w_{AB} = 1$$
$$w_{Ab} = 1 - s$$
$$w_{aB} = 1 - s$$
$$w_{ab} = 1$$

où s est le coefficient de contre-sélection ($s \geq 0$). Les individus portant les haplotypes *ab* (haplotype dérivé, et « compensatoire ») et *AB* (haplotype ancestral) ont une meilleure *fitness* et sont donc plus fréquents dans la population. Dans les populations diploïdes, les *fitness* relatives à ces deux locus sont présentées dans le **Tableau 2** et comme dans le modèle coadapté, les *fitness* relatives aux combinaisons génotypiques des locus A et B dépendent de la récessivité, dominance ou codominance des allèles dérivés.

1.1.3 Le modèle additif

Le modèle additif est un modèle de sélection positive indépendant entre les paires de locus. C'est un modèle de sélection polygénique indépendante qui permet de comparer et différencier les modèles de sélection épistatique et le modèle neurte. Dans ce modèle, les haplotypes qui possèdent au moins un allèle mutant (a et/ou b) sont sélectionnés et possèdent une meilleure *fitness*.

$$w_{AB} = 1$$
$$w_{Ab} = 1 + s$$
$$w_{aB} = 1 + s$$
$$w_{ab} = 1+2s$$

Ainsi, trois haplotypes différents sont sélectionnés dans le modèle additif et notamment les haplotypes recombinants.

Tableau 2: Fitness des génotypes aux deux locus (A et B) dans les modèles de sélection épistatique (coadapté et compensatoire) et un modèle de sélection additive, pour les mutations récessives, dominantes et codominantes (allèles dérivés *a* et *b*).

		COADAPTE			COMPENSATOIRE			ADDITIF		
		BB	*Bb*	*bb*	*BB*	*Bb*	*bb*	*BB*	*Bb*	*bb*
récessivité	*AA*	1	1	1	1	1	1-2s	1	1	1+s
	Aa	1	1	1	1	1	1-2s	1	1	1+s
	aa	1	1	1+2s	1-2s	1-2s	1	1+s	1+s	1+2s
codominance	*AA*	1	1	1	1	1-s	1-2s	1	1+s/2	1+s
	Aa	1	1+s	1+s	1-s	1-s	1-s	1+s/2	1+s	1+1.5s
	aa	1	1+s	1+2s	1-2s	1-s	1	1+s	1+1.5s	1+2s
dominance	*AA*	1	1	1	1	1-2s	1-2s	1	1+s	1+s
	Aa	1	1+2s	1+2s	1-2s	1	1	1+s	1+2s	1+2s
	aa	1	1+2s	1+2s	1-2s	1	1	1+s	1+2s	1+2s

1.1.4 Le modèle neutre

Dans le modèle neutre, toutes les combinaisons génotypiques ont la même fitness ($w = 1$). Seule la dérive génétique pourra influencer les fréquences alléliques et génotypiques dans les populations simulées. Ce modèle correspond à l'hypothèse nulle de neutralité.

1.1.5 Influence d'autres facteurs : structure génétique, système de reproduction, interaction entre les allèles d'un même locus

Les deux modèles génétiques de sélection épistatique sont comparés au modèle neutre afin d'étudier les dynamiques de fixation des allèles cosélectionnés et le DL qui s'établit entre les deux locus. Cependant, ces processus sont aussi influencés par l'interaction entre les allèles d'un même locus, par la structure génétique des populations et par les systèmes de reproduction. Ainsi, dans ce travail, nous avons évalué l'effet de deux systèmes de reproduction contrastés; un régime panmictique (« *random-mating* ») et un régime en autofécondation (« *self-mating* ») dans une population structurée en deux sous-populations isolées. Nous avons également évalué l'effet de l'interaction entre les allèles d'un même locus: dominance, codominance, récessivité (**Tableau 2**).

.

1.2 Les outils statistiques de détection de la sélection épistatique

Les interactions adaptatives entre locus peuvent être le résultat de la sélection épistatique suivant les modèles présentés ci-dessus. Il a été montré que dans ces deux modèles de sélection, un déséquilibre de liaison (DL) significatif s'établit entre les locus cosélectionnés (Takahasi & Innan, 2008). Plusieurs statistiques de DL existent, parmi lesquelles on peut distinguer ; (i) les mesures classiques telles que D, D', r et r^2 et (ii) les mesures de DL tenant compte de la structure des populations et/ou de l'apparentement entre les individus (r_s, r_{vs}, $r_{v,}$ $r|a$, D'^2_{IS}). La plupart des statistiques de DL présentées dans cette partie n'ont pas été initialement développées pour détecter des interactions adaptatives associées à de la sélection épistatique entre deux locus. Historiquement, la plupart de ces statistiques ont été développées pour mesurer le DL lié à la proximité physique entre marqueurs génétiques (Hill & Robertson, 1968; Lewontin & Kojima, 1960; Mangin et al., 2012). Cependant, les statistiques $r|a$ ou D'^2_{IS} ont été développées pour mesurer le DL associé à la sélection épistatique dans un contexte de populations subdivisées où la dérive génétique influence aussi le DL dans les sous-populations (Id-Lahoucine et al., 2019; Ohta, Mar 1982a, May 1982b). De façon générale, on distinguera deux types de déséquilibre de liaison ; le DL 'physique' entre des marqueurs liés physiquement et qui présentent un faible taux de recombinaison et le DL 'évolutif' qui se forme entre des marqueurs indépendants (non liés physiquement) qui se trouvent sur des chromosomes différents ou qui sont très éloignés sur un même chromosome. Bien entendu, on ne peut exclure que la sélection épistatique puisse agir sur des locus liés, mais dans ce contexte, il sera plus difficile d'évaluer son importance. L'ensemble de ces statistiques sont présentées ci-après.

1.2.1 Les statistiques classiques de déséquilibre de liaison

Le déséquilibre de liaison mesure l'association non aléatoire entre les allèles à deux locus différents. Par le calcul du DL, il est possible de détecter des combinaisons favorables d'allèles à deux locus indépendants formant des haplotypes ayant une meilleure *fitness*. Dans le cadre de la sélection épistatique, les individus portant ces haplotypes ont de meilleures capacités à se reproduire et à survivre dans leur environnement. Si l'on considère deux locus A et B à deux allèles *A/a* et *B/b* (i.e. 1 SNP par locus), il y a quatre combinaisons possibles, soit

quatre haplotypes. Le DL mesure si un haplotype est en excès ou en déficit par rapport à l'hypothèse d'indépendance sous équilibre de liaison (Lewontin & Kojima, 1960) :

$$D_{ab} = P_{ab} - P_a P_b$$

où P_{ab} est la fréquence de l'haplotype *ab*, et P_a et P_b sont les fréquences des allèles *a* et *b*. Il est à noter que $D_{ab} = D_{AB} = -D_{aB} = -D_{Ab} = D$. La statistique *D* dépend des fréquences des allèles comparés et il n'est donc pas judicieux de comparer des paires de locus qui ont des fréquences alléliques différentes. En effet, *D* dépend directement des fréquences alléliques, et si les allèles aux deux locus sont en fréquences intermédiaires (i.e. 0.5), la valeur maximale de D atteinte est 0.25 mais si un allèle est en fréquence très faible (ou forte) par rapport à l'autre, le D sera proche de zéro. La statistique D' permet de normaliser la valeur du D en la divisant par sa valeur maximale théorique suivant les fréquences alléliques observées.

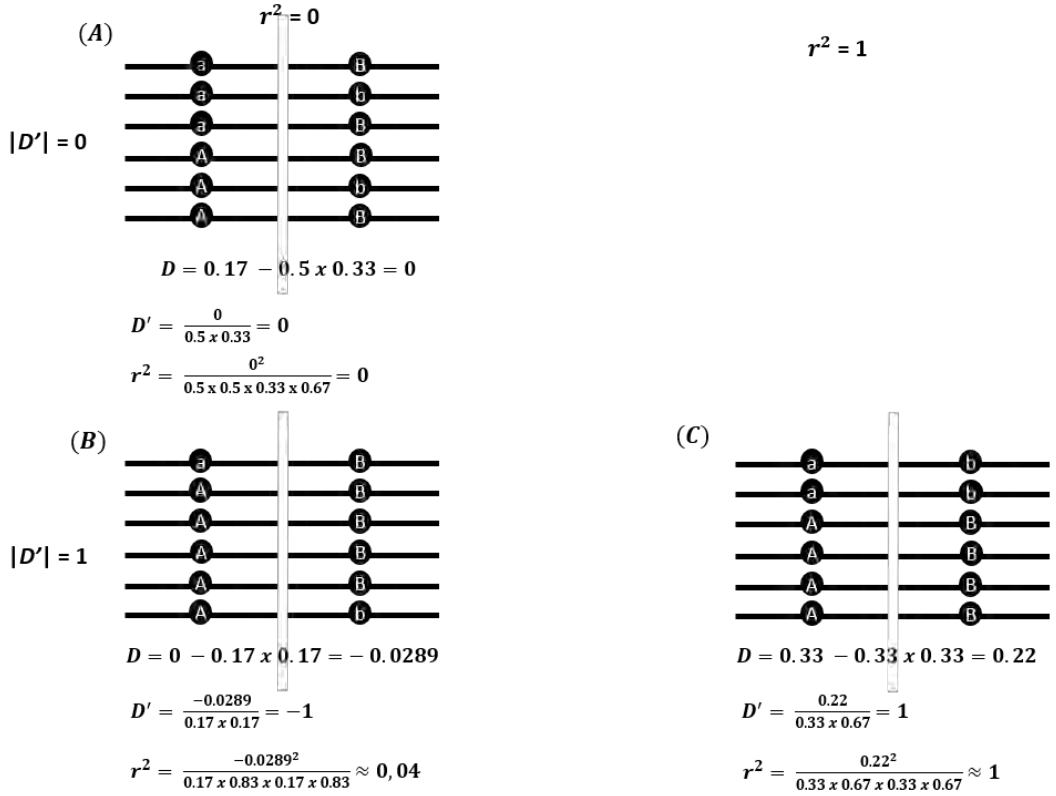
$$D' = D/D_{max} \text{ avec :}$$

$$\text{Si } D_{ab} > 0 \qquad D'_{ab} = D_{ab}/\min(p_a p_B, p_A p_b)$$

$$\text{Si } D_{ab} < 0 \qquad D'_{ab} = D_{ab}/\min(p_a p_b, p_A p_B)$$

D' varie entre 0 et 1. Cette statistique est toutefois sensible aux fréquences alléliques extrêmes qui peuvent être liées à de faibles tailles d'échantillons (**Figure 8**).

Figure 8 : Mesure de la différence des statistiques de DL entre *r²*, *D* et *D'*. Représentation schématique du DL entre deux SNP pour illustrer les différentes sensibilités des statistiques aux variations de fréquences alléliques pour trois cas extrêmes: (A) les deux allèles *a* et *b* sont en fréquences intermédiaires mais les 4 haplotypes sont représentés; D' et r^2 sont à 0, (B) *a* et *b* sont en fréquences faibles et parmi les 4 haplotypes (*AB*, *Ab*, *aB*, *ab*) seulement 3 sont représentés, dans ce cas D' est à 1 tandis que r^2 est proche de 0. (C) *a* et *b* sont en fréquences intermédiaires et les deux statistiques D' et r^2 ont atteint leur valeurs maximales à 1.



Le r^2 est une mesure complémentaire pour estimer le DL.

$$r^2 = D^2 / (p_a p_b p_A p_B)$$

Le r^2 correspond au carré du coefficient de corrélation, il est compris entre 0 et 1. Ainsi, les statistiques D' et r^2 ne sont pas sensibles de la même manière aux variations de fréquences. Si parmi les 4 haplotypes possibles (*AB*, *Ab*, *aB*, *ab*), seulement 3 sont représentés, le D' est égal à 1, sa valeur maximale, tandis que le r^2 ne sera que faiblement modifié (**Figure 8B**). Le D' est plus sensible à de faibles tailles d'échantillons, ce qui influence la présence ou l'absence d'haplotypes rares. De plus, le r^2 peut se calculer sur des données génotypiques de phase inconnue (le cas des doubles hétérozygotes) et dont on ne connaît pas l'état ancestral/dérivé des allèles en utilisant la corrélation au carré entre deux marqueurs SNP (Hill & Robertson, 1968; Rogers & Huff, 2009). En revanche, si les données génotypiques sont diploïdes, et que l'état ancestral et dérivé des allèles est connu, le r^2 ne donne pas d'information sur le sens de l'interaction contrairement à la statistique r.

$$r = D_{ab} / \sqrt{p_a p_b p_A p_B}$$

Le r est une mesure directionnelle du DL qui permet de quantifier s'il y a un excès d'haplotypes ayant des allèles mutants ou ancestraux, il correspond également à la corrélation des fréquences des allèles aux deux locus ($1 \leq r \leq 1$) et sa valeur au carré correspond aux r^2. Ainsi, si l'on considère deux locus A et B ayant deux allèles mutants *a* et *b*, l'interaction entre ces deux allèles va créer un excès d'haplotypes *ab* double mutant (et donc aussi du double sauvage *AB*) par rapport aux haplotypes recombinants *aB* et *Ab* et ainsi donner une valeur positive de *r*. Il a été montré, dans le cadre d'interactions adaptatives liées à de la sélection épistatique entre deux locus indépendants, que la mesure directionnelle du DL améliore l'efficacité de la détection de la sélection épistatique par rapport à l'attendu neutre (Takahasi & Innan, 2008).

Les statistiques présentées ci-dessus sont des mesures classiques du déséquilibre de liaison et parmi toutes celles qui ont été décrites, la statistique de r^2 *e*st la plus couramment utilisée. Mais comme cela a été évoqué, le DL est largement influencé par la dérive génétique dans les petites populations ainsi que par la structure génétique des populations et les systèmes de reproduction fermés qui augmentent la corrélation des fréquences génotypiques et alléliques aux différents locus. L'ensemble de ces forces évolutives auront tendance à accroitre le DL et notamment le DL 'longue distance' au sein des chromosomes conduisant à une estimation biaisée du DL c'est-à-dire différente de zéro sous l'hypothèse nulle d'indépendance. Plusieurs statistiques ont été développées afin de considérer les biais dus à l'apparentement entre les individus et à la structure des populations et obtenir, ainsi une mesure corrigée du DL.

1.2.2 Les statistiques de déséquilibre de liaison qui prennent en compte la structure des populations et l'apparentement entre les individus

1.2.2.1 Mesures du DL dans les populations structurées.

La sélection épistatique et la dérive génétique sont deux des principaux mécanismes responsables de la formation du DL entre deux locus. La dérive génétique entraîne des fluctuations aléatoires des fréquences des allèles au sein des populations causant une

diminution de la diversité génétique. Au sein de populations structurées, la dérive génétique sera plus forte dans les sous-populations du fait de l'isolement locales et de la taille de population plus faible conduisant à une augmentation du DL à l'échelle globale. En effet, des corrélations se créent entre les fréquences des allèles aux différents locus dans les sous-populations. Ohta proposa une série de mesures qui permettent de calculer le DL dans diverses sous-populations et de comparer ces sous-populations. C'est une approche analogue aux statistiques F de Wright qui sont fréquemment utilisées pour mesurer la structure génétique des populations et identifier des signatures de sélection à partir de la variance des fréquences alléliques entre sous-populations. Les statistiques F (F_{IS}, F_{ST} et F_{IT}) comparent les fréquences alléliques et génotypiques, sur un locus, entre individus, sous-populations et population globale, tandis que les statistiques D d'Ohta comparent les fréquences alléliques et haplotypiques dans un système à deux locus. Ainsi, une série de mesures ont été proposées :

$$D_{IS}^2 = E\left\{\sum_{ij}(g_{ij,k} - x_{i,k}y_{j,k})^2\right\}$$

$$D_{ST}^2 = E\left\{\sum_{ij}(x_{i,k}y_{j,k} - \overline{x}_i\overline{y}_j)^2\right\}$$

$$D_{IT}^2 = E\left\{\sum_{ij}(g_{ij,k} - \overline{x}_i\overline{y}_j)^2\right\}$$

$$D'^2_{IS} = E\left\{\sum_{ij}(g_{ij,k} - \overline{g_{ij}})^2\right\}$$

$$D'^2_{ST} = E\left\{\sum_{ij}(\overline{g_{ij}} - \overline{x}_i\overline{y}_j)^2\right\}$$

Dans une population divisée en sous-populations, on considère deux locus A et B, avec $x_{i,k}$, $y_{j,k}$ les fréquences des allèles *i* et *j* aux deux locus respectifs dans la population *k*; $g_{ij,k}$ est la fréquence de l'haplotype *ij* (gamète A_iB_j) dans la population *k* aux locus A et B; $\overline{g_{ij}}$, $\overline{x}_i$ et $\overline{y}_j$ les valeurs moyennes sur l'ensemble des populations. Sur la base de ces définitions, D_{IS}^2 est la variance du DL dans la sous-population *k*, D_{ST}^2 est la variance de la corrélation entre les locus

A et B dans la sous-population *k* relative à la corrélation dans la population totale, et D_{IT}^2 est la variance totale du DL, c'est-à-dire la variance de la corrélation entre les locus A et B d'un même gamète dans une sous-population *k* relative à la corrélation dans la population totale. D'^2_{IS} est la variance de la corrélation entre les locus A et B d'un même gamète dans une sous-population *k* par rapport à celle du gamète moyen de la population totale, et D'^2_{ST} est la variance du DL dans la population totale. Ces statistiques sont reliées de manière hiérarchique et la statistique D'^2_{IS} est utilisée pour identifier des paires de locus cosélectionnés dans une sous-population *k*. En effet, le D'^2_{IS} compare les fréquences haplotypiques d'une paire de locus dans une sous-population *k* aux fréquences haplotypiques moyennes dans la population entière. Ainsi, la comparaison des D'^2_{IS} calculés dans chaque sous-population peut permettre d'identifier des paires de locus qui présentent des signatures de sélection épistatique dans certaines sous-populations. Cependant, en moyennant la fréquence d'un haplotype sur l'ensemble des sous-populations ($\overline{g_{ij}}$), le D'^2_{IS} suppose que les sous-populations subissent le même degré de dérive génétique. Ainsi, dans une sous-population qui présente un fort degré de dérive génétique par rapport aux autres, la valeur de D'^2_{IS} est forte (i.e. une forte variation des fréquences haplotypiques par rapport à la population totale « moyenne ») et cela suggère un effet de la sélection épistatique alors que cette valeur est explicable par un effet plus important de la dérive génétique dans cette sous-population. L'utilisation du D'^2_{IS} se restreint donc à des modèles de structuration génétique simples - du type « infinite island model » - en dehors desquels le taux de faux positifs sera significativement accru. De plus, il a été montré que le D'^2_{IS} distingue difficilement les cas de signatures de sélection épistatique des cas de signatures de balayages sélectifs liés à la sélection positive sur deux locus indépendants, dans une sous-population donnée (Beissinger et al., 2016; Id-Lahoucine et al., 2019; Ohta, Mar 1982a, May 1982b).

1.2.2.2 Mesures du DL corrigées par la structure des populations

Une autre mesure de DL a été développée pour analyser les populations hybrides, appelée « Ancestry Disequilibrium » (AD) (Pool, 2015; Schumer & Brandvain, 2016). La mesure de AD a été développée pour détecter les incompatibilités génétiques qui pourraient être contre-sélectionnées dans les populations hybrides. Dans les populations hybrides, les allèles peuvent être caractérisés en fonction de leurs origines (deux populations « ancestrales ») et

le signe (+/-) de l'AD a un sens. Ainsi, une valeur positive de AD signifie qu'il y a une surreprésentation des combinaisons alléliques qui proviennent du même ancêtre, ce qui dans notre modèle de sélection épistatique pourrait correspondre aux haplotypes ancestraux et mutants (*AB*, *ab*) et si l'AD est négatif, cela signifie qu'il y a une surreprésentation des haplotypes recombinants qui proviennent des deux ancêtres différents (*Ab*, *aB*). Pour calculer les valeurs de AD, les auteurs ont recherché les associations non aléatoires entre les allèles à deux locus non liés physiquement dans ces populations hybrides. Les valeurs positives de AD sont ainsi le reflet de la sélection naturelle qui élimine les combinaisons interspécifiques d'allèles qui entrainent une diminution de la *fitness* (Pool, 2015; Schumer & Brandvain, 2016). À partir de ce AD, trois statistiques ont été proposées pour mesurer le DL au sein de ces populations. La première est la statistique *r* (voir chapitre 1.2.1), une mesure classique de DL qui correspond au coefficient de corrélation de Pearson. La deuxième est la statistique $r|a$ qui est une corrélation partielle de *r*. Cette mesure pondérée de la corrélation tient compte de la proportion d'allèles (*a*) provenant de chacune des populations ancestrales de la population hybride. Le score *a* est la somme pour chaque individu des allèles de chaque locus provenant de l'une ou l'autre des populations ancestrales. La troisième statistique de AD est $r|a_f$, c'est une mesure équivalente au $r|a$ à l'exception que pour le calcul de *a* (« genome-wide ancestry proportion ») les deux chromosomes sur lesquels sont situés les locus d'intérêt ne sont pas pris en compte. La statistique $r|a$ permet de pondérer les valeurs d'AD et réduire le taux de faux positifs liés à la structure des populations en prenant comme information le fait que les individus qui proviennent d'une même population ancestrale ont tendance à partager plus d'allèles.

La statistique r_s^2 précédemment introduite par Mangin et al. (2012) est une mesure équivalent au $r|a$ qui tient compte de la structure des populations mais r_s^2 permet de calculer le DL sur des données non phasées. La connaissance des génotypes aux deux locus est suffisante. Le r_s^2 peut se calculer en utilisant la corrélation au carré pondérée par une covariable de structure. En effet, le r_s^2 prend en compte un vecteur de structure S qui donne pour chaque individu sa population d'origine ou une probabilité d'appartenance à chaque population. Par exemple, si l'on considère une population (autogame pour simplifier) échantillonnées pour N = 10 individus diploïdes avec deux marqueurs bi-alléliques homozygotes (génotypes codés 0 et 2), X^l et X^m. La population est structurée en deux sous-

populations et le vecteur S de structure est égal à 0 pour les individus qui proviennent de la première population et égal à 1 pour les individus qui proviennent de l'autre population.

$$X^l = \begin{pmatrix} 2 \\ 2 \\ 2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 2 \end{pmatrix} \quad X^m = \begin{pmatrix} 2 \\ 2 \\ 2 \\ 0 \\ 2 \\ 0 \\ 0 \\ 0 \\ 2 \\ 0 \end{pmatrix} \quad \text{et} \quad S = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

Alors $r^2 = cor^2(X^l, X^m) = 0.17$ ($r = 0.41$), et $r_s^2 = cor^2(X^l, X^m; S) = 0.05$ ($r_s = 0.22$). La corrélation $\text{cor}(X, Y; S)$ correspond à la corrélation partielle entre X et Y (Mangin et al., 2012, R package LDcorSV). En pratique, r_s^2 peut se calculer en prenant le coefficient de corrélation au carré des résidus des vecteurs X^l et X^m dans un modèle linéaire où le génotype au SNP est régressé sur la variable S. La valeur de r^2 est biaisée lorsque l'échantillon est structuré car le fait que des individus d'un même échantillon ont des origines génétiques différentes, cela génère du DL même si les locus ne sont pas liés mais simplement à cause des différences de fréquences alléliques entre les deux populations. En effet, la subdivision des populations due à des flux génétiques limités entraîne une augmentation de la dérive génétique (diminution de la taille efficace) au sein des sous-populations du fait de l'isolement de ces populations et à l'échelle de la population globale, la structure génétique augmente le DL.(Grillo et al., 2016; Ronfort et al., 2006). Par conséquent, la structuration des populations peut conduire à la surestimation du DL et la statistique r_s^2 limite cette inflation.

1.2.2.3 Mesures du DL corrigées pour la structure et l'apparentement

Lorsque les populations structurées étudiées sont constituées d'individus dont l'apparentement est hétérogène, les mesures r^2 et r_s^2 de la corrélation ne sont pas de bons très estimateurs du DL car celui-ci est également influencé par l'apparentement. En effet, chez les espèces autogames, ou lorsqu'il y a de la consanguinité ou un fort taux d'apparentement, l'effet de la recombinaison homologue, bien que présente, est plus difficilement observable du fait de l'homozygotie chromosomique. En effet, cela provoque une extension du DL dans le génome sur de grandes régions chromosomiques, voire entre les chromosomes (Burgarella & Glémin, 2017). La statistique r_v^2 a été proposée par Mangin et al., (Mangin et al., 2012) afin de capturer le DL physique entre paires de marqueurs en incluant une matrice

d'apparentement *V* (ou matrice « kinship ») afin de réduire le DL lié à l'apparentement et à la structure génétique. Les statistiques r_v^2 et r_s^2 ont été développées pour mesurer le DL entre des SNP adjacents, dans un contexte de cartographie génétique. Ainsi, la statistique r_v^2 inclut dans le calcul du DL la matrice d'apparentement *V* qui est calculée à partir des données génotypiques à l'échelle du génome. La matrice *V* est construite en comparant chaque paire d'individus *ij* et correspond à la proportion d'allèles identiques partagés sur un grand nombre de SNP. Par exemple, si l'on considère une population constituée d'individus diploïdes avec des SNP bi-alléliques dont les génotypes sont codés 0, 1 ou 2 (ou dose allélique) en fonction du génotype diploïde (00, 01, 11), la valeur entre deux individus *ij* pour un SNP est : 2 si les deux individus sont homozygotes et partagent les mêmes allèles (génotypes bi-locus 0/0 ou 2/2); 1 si les deux individus ne partagent qu'un allèle (génotypes 0/1, 1/2 ou 1/1) et 0 si les individus ne partagent aucun allèle (génotypes 0/2). Puis, la valeur de l'apparentement entre deux individus correspond à la somme des valeurs obtenues à chaque SNP divisée par le nombre total d'allèles soit le nombre de SNP multiplié par 2 (population diploïde). Cette mesure de l'identité par état (Identity In State – IIS) calculée à partir des SNP est une estimation de l'identité par descendance (Identity By Descent – IBS) car ils suivent l'hypothèse d'un modèle de mutation de type « infinite sites model ». La matrice ainsi obtenue est une matrice carrée symétrique (**Figure 9B**). Pour calculer r_v^2, les vecteurs de génotypes aux SNP (X^l, X^m) sont multipliés par l'inverse de la racine carrée de la matrice d'apparentement, $(V^{-1/2} * X^l)$ et $(V^{-1/2} * X^l)$ et r_v^2 correspond au coefficient de corrélation au carré obtenu à partir de ces deux vecteurs: $cor^2\left(V^{-1/2}X^l, V^{-1/2}X^m\right)$ (**Figure 9A,C**). La statistique r_v^2 ainsi décrite permet de réduire le DL lié à l'apparentement et à la structure des populations (Mangin et al., 2012). La **Figure 10** montre les heatmap de deux matrices d'apparentement réordonnées par clustering hiérarchique et obtenues à partir de populations simulées de 500 individus diploïdes (voir chapitre 1.4). Les heatmap montrent l'apparentement entre tous les individus simulés dans la population structurée en deux sous-groupes génétiques et pour deux régimes de reproduction: un modèle en autogamie (a) et un modèle en panmixie (b).

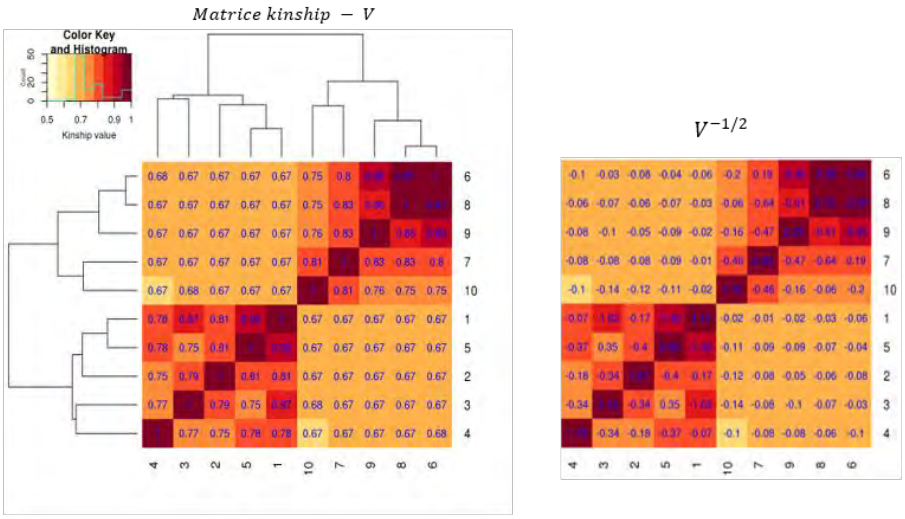
Figure 9 : Calcul des statistiques r (r^2) et r_v (r_v^2) dans une population virtuelle autogame de 10 individus structurée en deux sous-populations. (A) X^l et X^m sont deux SNP bi-alléliques dont les génotypes sont codés 0 ou 2 en fonction de la quantité d'allèle dérivé. Les vecteurs $V^{-1/2} * X^l$ et $V^{-1/2} * X^m$ correspondent au produit des vecteurs X^l (et X^m) et de $V^{-1/2}$. (**B**) La heatmap (« Matrice kinship – *V* ») représente la matrice kinship calculée à partir des données génétiques simulées à l'échelle du génome et réduite aux 10 individus représentés (la similarité varie entre 0 et 1). La heatmap montre la distribution de l'apparentement réordonné par un clustering hiérarchique entre les 10 individus et structuré en deux sous-groupes génétiques. La heatmap $V^{-1/2}$ montre les valeurs de l'inverse de la racine carré de la matrice kinship. **(C)** Calcul des statistiques r^2 (r) et $r_v^2(r_v)$, r correspond au coefficient de corrélation de Pearson entre les deux vecteurs de SNP X^let X^m, et r_v correspond coefficient de corrélation de Pearson entre les vecteurs $(V^{-1/2} * X^l)$ et $(V^{-1/2} * X^m)$ issus du produit matriciel entre les vecteurs de SNP et de la matrice $V^{-1/2}$.

(A) Vecteurs des génotypes aux locus *l* et *m* (X^l et X^m) corrigés par la matrice d'apparentement ($V^{-1/2} * X^l$ et $V^{-1/2} * X^m$).

$$X^l = \begin{pmatrix} \text{Id } 4 & 0 \\ \text{Id } 3 & 2 \\ \text{Id } 2 & 2 \\ \text{Id } 5 & 0 \\ \text{Id } 1 & 2 \\ \text{Id } 10 & 2 \\ \text{Id } 7 & 0 \\ \text{Id } 9 & 0 \\ \text{Id } 8 & 0 \\ \text{Id } 6 & 0 \end{pmatrix} \quad V^{-1/2} * X^l = \begin{pmatrix} \text{Id } 4 & -1.39 \\ \text{Id } 3 & 1.31 \\ \text{Id } 2 & 2.47 \\ \text{Id } 5 & -3.16 \\ \text{Id } 1 & 3.86 \\ \text{Id } 10 & 2.96 \\ \text{Id } 7 & -1.27 \\ \text{Id } 9 & -0.67 \\ \text{Id } 8 & -0.44 \\ \text{Id } 6 & -0.73 \end{pmatrix}$$

$$X^m = \begin{pmatrix} \text{Id } 4 & 0 \\ \text{Id } 3 & 2 \\ \text{Id } 2 & 2 \\ \text{Id } 5 & 2 \\ \text{Id } 1 & 2 \\ \text{Id } 10 & 0 \\ \text{Id } 7 & 0 \\ \text{Id } 9 & 2 \\ \text{Id } 8 & 0 \\ \text{Id } 6 & 0 \end{pmatrix} \quad V^{-1/2} * X^m = \begin{pmatrix} \text{Id } 4 & -2.10 \\ \text{Id } 3 & 2.09 \\ \text{Id } 2 & 1.80 \\ \text{Id } 5 & 2.11 \\ \text{Id } 1 & 1.01 \\ \text{Id } 10 & -1.09 \\ \text{Id } 7 & -1.47 \\ \text{Id } 9 & 4.05 \\ \text{Id } 8 & -0.48 \\ \text{Id } 6 & -2.32 \end{pmatrix}$$

(B) Heatmap de la matrice kinship *V* clusterisée et de l'inverse de la racine carré de *V*.

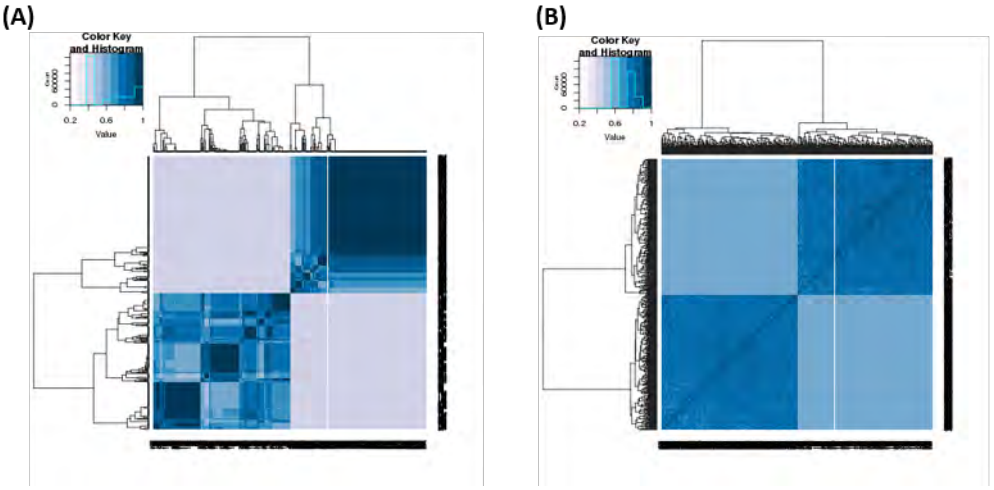


(C) Calcul des corrélations r et r_v entre les SNP X^l et X^m.

$$\boldsymbol{r} = \boldsymbol{cor}(\boldsymbol{X^l}, \boldsymbol{X^m}) = \boldsymbol{cor}\left(\overbrace{\begin{pmatrix} X_1^l \\ \vdots \\ X_n^l \end{pmatrix}}^{X^l}, \overbrace{\begin{pmatrix} X_1^m \\ \vdots \\ X_n^m \end{pmatrix}}^{X^m}\right) = 0.41 \ (r^2 = 0.17\ )$$

$$\boldsymbol{r_v} = \boldsymbol{cor}(\boldsymbol{V^{-1/2} * X^l}, \boldsymbol{V^{-1/2} * X^m}) = \boldsymbol{cor}\left(\overbrace{\begin{pmatrix} V_{11} & \cdots & V_{1n} \\ \vdots & \ddots & \vdots \\ V_{n1} & \cdots & V_{nn} \end{pmatrix}}^{V^{-1/2}} * \overbrace{\begin{pmatrix} X_1^l \\ \vdots \\ X_n^l \end{pmatrix}}^{X^l}, \overbrace{\begin{pmatrix} V_{11} & \cdots & V_{1n} \\ \vdots & \ddots & \vdots \\ V_{n1} & \cdots & V_{nn} \end{pmatrix}}^{V^{-1/2}} * \overbrace{\begin{pmatrix} X_1^m \\ \vdots \\ X_n^m \end{pmatrix}}^{X^m}\right) = 0.05 \ (r_v^2 = 0.0025)$$

Figure 10 : Représentation de la matrice kinship – V calculée sur l'ensemble des individus simulés à la génération 300 pour deux simulations réalisées sous neutralité en panmixie et en autogamie.
Les heatmap (« Matrice kinship – *V* ») représentent les matrices kinship calculées à partir des données génétiques simulées à l'échelle du génome à partir du modèle neutre (la similarité varie entre 0 et 1) à la génération 300 en autogamie (A) et en panmixie (B). Les heatmap montrent les distributions de l'apparentement réordonné par un clustering hiérarchique entre tous les individus simulés et structurés en deux sous-groupes génétiques.



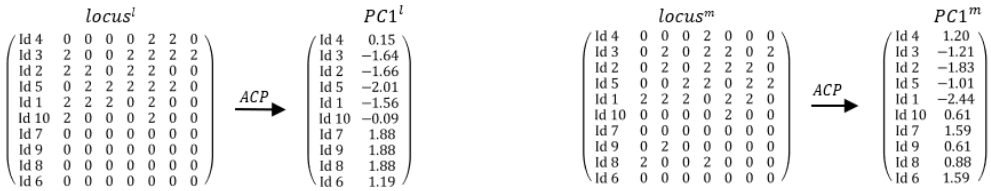
Ainsi, la mesure r_v^2 tient compte de l'apparentement et permet de corriger les biais dus à la structure des populations. De plus, l'avantage de la statistique r_v^2 est qu'elle peut se calculer uniquement à partir des données génotypiques et il n'est pas nécessaire de connaître la phase des haplotypes. Comme cela a été évoqué ci-dessus, ces statistiques r_s^2 et r_v^2 ont été développées pour obtenir une mesure sans biais du DL entre des SNP adjacents au sein de populations structurées et composées d'individus plus ou moins apparentés. Durant ce travail, nous avons utilisé la statistique r_v pour mesurer le DL entre des SNP non liés physiquement afin de détecter des interactions adaptatives qui pourraient être liées à de la sélection épistatique. En effet, la statistique r_v capture aussi bien l'apparentement que la structure des populations avec la matrice *V* (Figure 10), sachant que r_s requiert par ailleurs une modélisation préalable de la structure génétique des populations (par exemple avec le logiciel STRUCTURE – (Pritchard et al., 2000)). Nous avons également utilisé cette méthode de correction par la matrice d'apparentement pour mesurer le DL entre des paires de fenêtres génomiques qui possèdent des SNP potentiellement sous sélection épistatique.

1.2.2.4 Mesures du DL sur des fenêtres génomiques

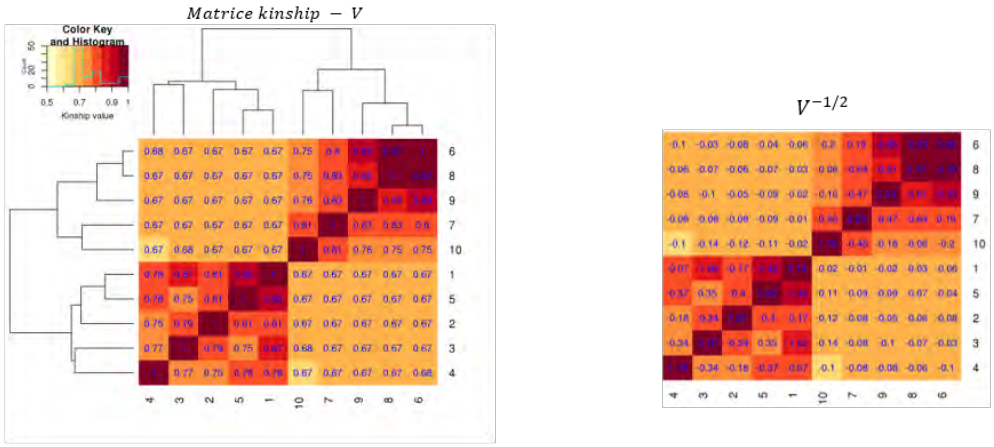
Nous avons développé une méthode afin de mesurer le DL entre des fenêtres génomiques à l'échelle des haplotypes. Contrairement aux méthodes de comparaisons par paires de SNP, cette méthode se base sur les associations haplotypiques entre locus (ou gènes) et utilise l'information de corrélation entre les SNP de chaque locus pour « résumer » le génotype d'un individu à chaque locus. Pour une fenêtre génomique données, les haplotypes sont modélisés à l'aide d'une analyse multivariée ; l'analyse en composante principale (ACP). L'ACP est utilisée pour l'exploration de données de génotypes afin d'extraire une variable continue (une composante principale) et résumer l'information de variation génétique contenue dans le locus en diminuant le nombre de dimensions. Dans la littérature, il existe plusieurs exemples où les ACP ont été réalisées sur des données génomiques. Les ACP ont été utilisées pour représenter l'apparentement entre les individus à l'échelle du génome (Patterson et al., 2006) et les résultats permettent de produire des « cartes » qui reflètent la structure des populations et l'histoire démographique (par exemple les flux de gènes). Dans ce contexte, Price et al (2006) ont réalisé des ACP sur un ensemble de marqueurs génomiques afin d'en extraire les composantes principales (PC), puis ces vecteurs de PC ont été utilisés dans le cadre d'analyses d'associations pour ajuster les génotypes de chaque marqueurs pris individuellement en fonction des coordonnés des individus sur les différents axes de variation. Les ACP ont également été utilisées pour décrire les variations de la diversité génétique en fonction des régions géographiques ou de la structure des populations. (François et al., 2010; Jombart et al., 2009; H. Li & Ralph, 2019). Pour cette étude, nous proposons d'utiliser l'ACP afin d'en extraire la première composante principale (PC1) que nous considérons comme un vecteur qui résume pour chaque individu l'information haplotypique et donc la variation génétique sur une fenêtre génomique (**Figure 11A**).

Figure 11 : Méthode pour modéliser quantitativement les haplotypes et calculer le DL non corrigé cor_{PC1} et corrigé cor_{PC1v} sur les fenêtres génomiques. (A) $locus^l$ et $locus^m$ sont deux locus constitués de 7 SNP bi-alléliques codés 0 ou 2 en fonction de la quantité d'allèles dérivés chez 10 individus (n=10). $PC1^l$ et $PC1^m$ sont les deux vecteurs de PC1 issus des ACP réalisées sur les données génétiques des $locus^l$ et $locus^m$ respectivement. Les deux vecteurs de PC1 sont les haplotypes quantitatifs. **(B)** La heatmap (Matrice kinship - V) représente la matrice kinship calculée à partir des données génétiques simulées à l'échelle du génome et réduite aux 10 individus représentés (la similarité varie entre 0 et 1), la heatmap présente la distribution de l'apparentement réordonné par un clustering hiérarchique entre les 10 individus et structuré en deux groupes génétiques. La heatmap $V^{-1/2}$ montre les valeurs de l'inverse de la racine carré de la matrice kinship. **(C)** Calcul des statistiques cor_{PC1} et cor_{PC1v}. cor_{PC1} correspond au coefficient de corrélation de Pearson entre les deux vecteurs de $PC1$ ($PC1^l$ et $PC1^m$), et cor_{PC1v} correspond coefficient de corrélation de Pearson entre les vecteurs ($V^{-1/2} * PC1^l$) et ($V^{-1/2} * PC1^m$) issus du produit matriciel entre les vecteurs de $PC1$ et de la matrice $V^{-1/2}$.

(**A**) Résumé des haplotypes aux locus *l* et *m* par la première composante principale de l'ACP.



(**B**) Heatmap de la matrice kinship *V* clusterisée et de l'inverse de la racine carré de *V*.



(**C**) Calcul des corrélations cor_{PC1} et cor_{PC1v} entre les locus l et m.

$$\boldsymbol{cor_{PC1} = cor(\overbrace{PC1^l}, \overbrace{PC1^m}) = cor\left(\begin{pmatrix} PC1_1^l \\ \vdots \\ PC1_n^l \end{pmatrix}, \begin{pmatrix} PC1_1^m \\ \vdots \\ PC1_n^m \end{pmatrix}\right)} = 0.86$$

$$\boldsymbol{cor_{PC1v} = cor(\overbrace{V^{-1/2}} * \overbrace{PC1^l}, \overbrace{V^{-1/2}} * \overbrace{PC1^m}) = cor\left(\begin{pmatrix} V_{11} & \cdots & V_{1n} \\ \vdots & \ddots & \vdots \\ V_{n1} & \cdots & V_{nn} \end{pmatrix} * \begin{pmatrix} PC1_1^l \\ \vdots \\ PC1_n^l \end{pmatrix}, \begin{pmatrix} V_{11} & \cdots & V_{1n} \\ \vdots & \ddots & \vdots \\ V_{n1} & \cdots & V_{nn} \end{pmatrix} * \begin{pmatrix} PC1_1^m \\ \vdots \\ PC1_n^m \end{pmatrix}\right)} = 0.15$$

Les vecteurs de PC1 sont des variables continues ; à chaque individu est attribué une valeur en fonction de son génotype sur la fenêtre génomique et les individus qui portent les mêmes haplotypes (ou génotypes multi-SNP), auront la même valeur de PC1. Puis, le DL entre deux fenêtres génomiques est la corrélation entre les deux vecteurs de PC1 ; $cor_{PC1} = cor(PC1^l, PC1^m)$ (**Figure 11C**). Pour calculer cor_{PC1v}, sur le même principe que r_v, les vecteurs de PC1 sont multipliés par l'inverse de la racine carrée de la matrice d'apparentement (**Figure 11B,C**). Le DL corrigé correspond à la corrélation calculée entre les vecteurs $V^{-1/2} * PC1^l$ et $V^{-1/2} * PC1^m$ (voir chapitre 1.2.2).

Ainsi, pour détecter la sélection épistatique, nous utilisons une mesure classique de DL basée sur les comparaisons par paires de SNP avec le coefficient de corrélation $\boldsymbol{r = cor(X^l, X^m)}$ et la mesure $\boldsymbol{cor_{PC1} = cor(PC1^l, PC1^m)}$ du DL entre des fenêtres génomiques de plusieurs SNP. Pour prendre en compte le DL généré par l'apparentement entre les individus et la structure des populations, nous utilisons la correction par la matrice *V* (Mangin et al., 2012) et nous calculons le DL avec $\boldsymbol{r_v = cor(V^{-1/2}X^l, V^{-1/2}X^m)}$ et $\boldsymbol{cor_{PC1v} = cor(V^{-1/2} * PC1^l, V^{-1/2} * PC1^m)}$ qui se basent respectivement sur les SNP et sur les fenêtres génomiques. Enfin, nous faisons un test de nullité du coefficient de corrélation pour évaluer la significativité de la corrélation entre deux variables X^l et X^m, ou $PC1^l$ et $PC1^m$ (respectivement $V^{-1/2}X^l$ et $V^{-1/2}X^m$, ou $V^{-1/2} * PC1^l$ et $V^{-1/2} * PC1^m$) en obtenant une p-valeur. Pour cela, nous calculons la statistique $T = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}$ qui suit une distribution de Student $\tau_{(n-2)}$ si les deux variables X^l et X^m (respectivement $PC1^l$et $PC1^m$) sont indépendantes. Dans un modèle neutre d'évolution du DL, les observations au sein de X^l et X^m (respectivement $PC1^l$et $PC1^m$) ne sont pas indépendantes si les individus sont apparentés et si la population est structurée. Nous faisons donc l'hypothèse que T va suivre approximativement la distribution de Student $\tau_{(n-2)}$ pour les statistiques r_v et cor_{PC1v} uniquement. De plus, si l'état ancestral/dérivé des allèles est connu, le signe du coefficient de corrélation r (ou r_v) reflète le sens de l'interaction. Un r (ou r_v) positif s'établit lorsque les SNP sont cosélectionnés selon les modèles épistatiques de coadaptation et compensatoire, c'est-à-dire lorsque l'association d'allèles dérivés est sélectionnée (coadaptation) ou lorsque les associations d'allèles dérivés et ancestraux sont sélectionnées (compensatoire). Dans ce cas, le test de corrélation peut être réalisé de manière unilatérale avec l'hypothèse alternative « r (ou r_v) est significativement supérieur à zéro ». Si l'état ancestral/dérivé n'est pas connu, le

signe de r (ou r_v) n'est pas interprétable et la p-valeur du test de corrélation est calculée du côté négatif ou positif de la distribution nulle qui est symétrique. De même, le signe de cor_{PC1} (ou cor_{PC1v}) n'étant pas interprétable puisque les PC1 correspondent à un classement des individus en fonction de leurs génotypes multi-SNP, la p-valeur du test de corrélation est calculée comme dans le cas de r (ou r_v). Ainsi, la statistique r_v (respectivement cor_{PC1v}) nous permet de se placer dans la cadre d'un test statistique utilisant la distribution de Student pour tester la nullité du coefficient de corrélation. Des simulations ont été réalisées afin d'évaluer la puissance de détection des statistiques ainsi que l'adéquation à la loi de Student $\tau_{(n-2)}$ (voir chapitre 1.5.3). Un modèle d'évolution neutre entre deux locus indépendants a été simulé pour évaluer la distribution du DL sous l'hypothèse nulle et trois modèles de sélection ont été simulés; deux modèles de sélection épistatique (coadapté et compensatoire) et un modèle de sélection additive (indépendante sur chaque locus). Pour chaque simulation, nous estimons le DL avec les statistiques présentées ci-dessus. Les simulations nous permettent également d'évaluer l'effet des différents modes de reproduction et de la structure des populations.

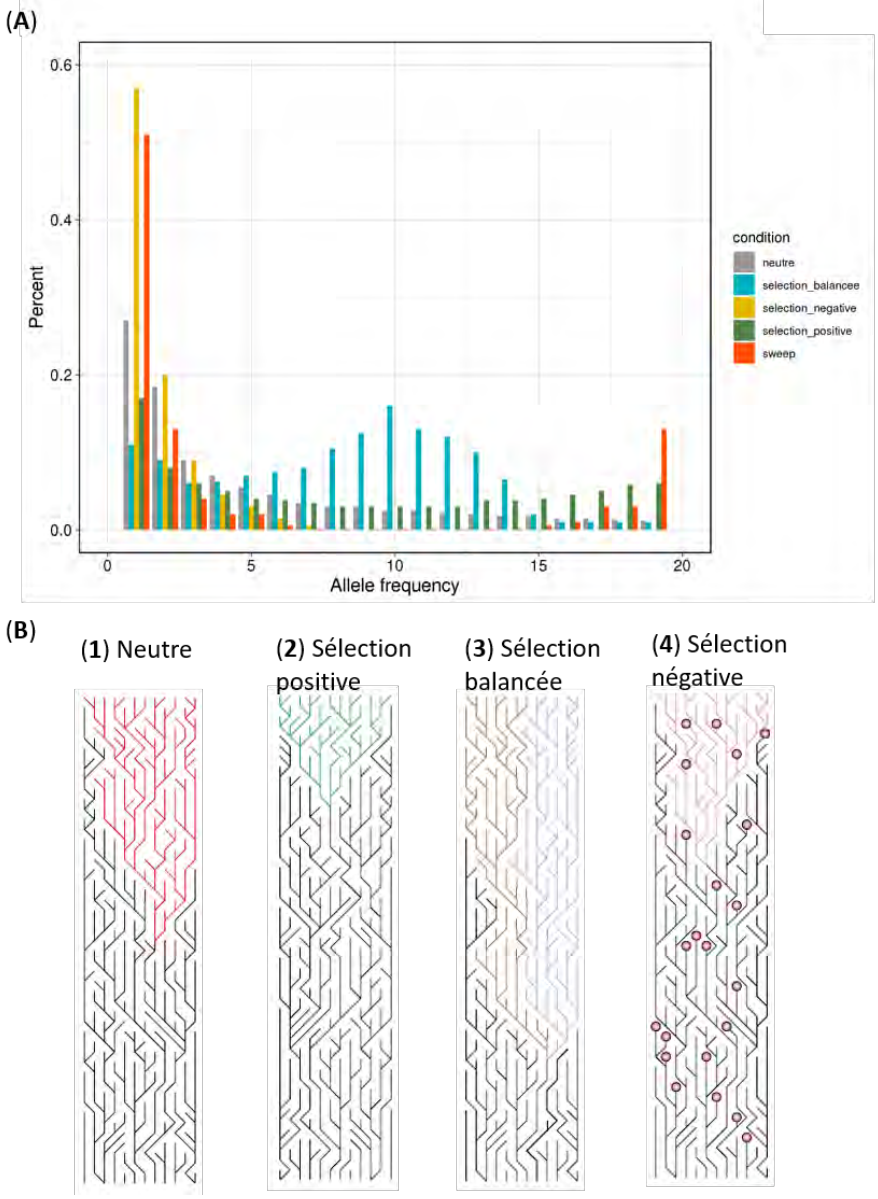
1.3 Les outils statistiques de détection de la sélection naturelle

Nous nous sommes également intéressés à l'évolution du polymorphisme des séquences lorsque celles-ci présentent des variants génétiques sous sélection épistatique. En effet, la sélection épistatique contribue à l'adaptation des organismes à leur environnement en induisant des changements de fréquences alléliques aux locus ciblés. L'objectif est d'évaluer si les locus sous sélection épistatique présentent également des signatures de sélection « classique » de manière indépendante. Dans une étude de 2009, Takahasi a analysé l'effet de la sélection épistatique sur le polymorphisme des séquences. Il considère un modèle de sélection épistatique par coadaptation (voir chapitre 1.1.1). L'objectif a été d'évaluer l'effet de la sélection épistatique sur le polymorphisme de ces locus et notamment l'effet d'auto-stop ou « hitch-hiking » induit par la sélection épistatique sur le polymorphisme des variants neutres qui sont physiquement liés aux locus cosélectionnés (Takahasi 2009). D'après les résultats de cette étude, les signatures de sélection épistatique sur le polymorphisme des gènes, (modèle coadapté) varient entre les locus selon le temps qui sépare l'apparition de ces deux mutations. En d'autres termes, si la mutation *b* est introduite peu de temps après l'apparition de la première mutation *a*, l'auteur observe une diminution locale de variabilité

dans la région où se trouve la mutation *a* et l'augmentation de la fréquence de l'allèle *a* est liée à la sélection épistatique, elle-même possible par la présence de l'allèle *b*. À partir de ces résultats préliminaires, nous avons recherché des signatures de sélection indépendamment sur les locus identifiés sous sélection épistatique. Pour identifier de telles signatures de sélection et comme cela a été décrit en introduction, un grand nombre de tests statistiques ont été développés et parmi les approches qui se basent strictement sur les données génétiques, nous avons utilisé des méthodes qui permettent de détecter la sélection au sein d'une même espèce et qui se basent sur les propriétés du spectre de fréquence des allèles – SFS. Les méthodes basées sur le SFS permettent de détecter une large gamme de signatures de sélection (sélection positive, balayages sélectifs, sélection purifiante et sélection balancée) en fonction des caractéristiques de la sélection et du contexte démographique. Ces méthodes basées sur SFS sont sensibles aux processus démographiques ainsi qu'à la structure des populations et elles doivent être employées avec précaution, dans des populations stables et faiblement structurées, ou tout du moins, sur des populations dont l'histoire démographique est bien caractérisée. De plus, Takahasi (2009) a utilisé ces mêmes statistiques de tests de neutralité pour étudier les différents patron de polymorphisme des locus sous sélection épistatique à partir du modèle coadapté (calcul du ***D*** de Tajima et du ***E*** de Zeng) et ces tests sont facilement réalisables à l'échelle du génome pour identifier des signatures de sélection à partir des séquences des gènes.

Ainsi, ces tests de neutralité se basent sur les propriétés du spectre de fréquence des allèles (Site Frequency Spectrum) qui mesure la distribution du nombre de sites polymorphes (SNP) selon les classes de fréquence des allèles dérivés dans un échantillon de séquences d'un locus donné (**Figure 12**). Les différentes formes de sélection modifient le spectre de fréquence et le patron de polymorphisme d'un locus est comparé à ce qui est attendu sous l'hypothèse nulle de neutralité afin d'identifier des signatures de sélection (Nielsen, 2005; Sabeti et al., 2007). La **Figure 12A** montre les distributions du nombre d'allèles dérivés allèles mutants en fonction de trois grands modes de sélection. Sous neutralité, le SFS possède une forme en « L », avec un nombre plus important de SNP ayant des allèles dérivés en faibles fréquences que de SNP avec des allèles dérivés en fortes fréquences.

Figure 12 : **Spectre des fréquences alléliques et généalogie des gènes sous neutralité, sous sélection négative, balancée, positive et sous balayage sélectif.** (A) L'axe des x représente les classes de fréquence *i* de l'allèle dérivé sur un échantillon de n=20 séquences ($1 \leq i \leq 19$) et l'axe des Y indique la proportion de SNP appartenant à chaque classe de fréquence de l'allèle dérivé. (B) Les quatre panels représentent les généalogies d'une population de 12 individus haploïdes. Chaque ligne retrace l'ascendance d'un lignée et les lignes colorées représentent tous les descendants qui ont hérités soit d'une mutation neutre (1) soit de mutations sous sélection naturelle (2-4). (1) représente la généalogie d'un allèle neutre (rouge) qui se fixe par un effet de dérive génétique. (2) représente la généalogie d'un allèle sous sélection positive (vert) et dont la fixation est plus rapide par rapport à l'allèle neutre. (3) représente la généalogie de deux allèles (bleu et marron) qui sont sous sélection balancées. Ces deux allèles ségrègent sur toute la généalogie et ne sont ni perdus ni fixés. La généalogie de ces deux allèles a ainsi une coalescence plus ancienne. (4) représente la généalogie d'un locus sous sélection négative où un allèle (violet) tend à se fixer, tandis que les mutations délétères (cercles) sont éliminées par sélection négative. Les Figures sont extraites et adaptées de Bamshad et Wooding (2003).



Ceci est directement relié à la généalogie (coalescent) sous neutralité (**Figure 12B**). La sélection négative provoque un excès de SNP avec des allèles mutants rares, par rapport au modèle neutre, en raison de la présence de mutations délétères qui sont contre-sélectionnées. Sous sélection positive et aussi lors d'un balayage sélectif, on observe un excès

de SNP avec des allèles dérivés en fortes fréquences. Ce sont les allèles directement sélectionnées ou les allèles sélectionnées accompagnées des mutations neutres en hitch-hiking au début du processus de sélection positive. Sous balayage sélectif on observe, en plus, un excès de SNP avec des allèles dérivés rares, qui sont les mutations qui apparaissent au cours du balayage sélectif sur les haplotypes portant l'allèle sélectionnée. Enfin, sous sélection balancée, on observe un excès de SNP avec des allèles dérivés en fréquences intermédiaires du fait de la sélection sur les hétérozygotes (**Figure 12**). Tous ces patrons de polymorphisme associés aux différentes formes de sélection sont directement reliés à la structure du coalescent au niveau du locus sous sélection. Les tests de neutralité permettent de comparer des estimateurs du polymorphisme de séquence d'un locus dont on connait la distribution sous l'hypothèse d'évolution neutre dans une population de taille constante (équilibre mutation-dérive), pour un locus de même taille et pour une même taille d'échantillon. Les statistiques utilisées sont le ***D*** de Tajima (Tajima, 1989), le ***H*** de Fay & Wu (Fay & Wu, 2000), et le ***E*** de Zeng (Zeng et al., 2006).

Le degré de polymorphisme d'un locus au sein d'une population est donné par la relation $\theta = 4N\mu$. N représente la taille efficace de la population et μ est le taux de mutation au locus par génération. $4N\mu$ $(= 2N * 2\mu)$ est une estimation du nombre de mutations accumulées entre deux séquences d'un même gène, depuis leur plus récent ancêtre commun et selon l'hypothèse d'un modèle de mutation de type infinite-site (M. Kimura, 1968). Si l'on considère une population diploïde de taille N, c'est aussi le nombre moyen de mutations qui ségrégent à un locus dans cette population à un temps donné. Les statistiques de test utilisées comparent trois estimateurs de θ. Ces estimateurs sont sensibles à différentes fréquences alléliques; les fréquences faibles, intermédiaires et les fortes fréquences.

$$\theta_s = S \Big/ \sum_{i=1}^{n-1} 1/i$$

$$\theta_\pi = \binom{n}{2}^{-1} \sum_{i=1}^{n-1} i(n-i) S_i$$

$$\theta_\pi = \binom{n}{2}^{-1} \sum_{i<j}^{n-1} d_{ij}$$

$$\theta_L = \frac{1}{n-1} \sum_{i=1}^{n-1} iS_i$$

S_i est le nombre de sites avec *i* allèles dérivés, c'est le spectre de fréquence. θ_s ($S = \sum_{i=1}^{n-1} S_i$) est le nombre de sites qui ségrégent au locus (i.e. le nombre de SNP), quelle que soit la fréquence des variants. À l'échelle du génome, les sites polymorphes avec des allèles dérivés rares sont toujours plus fréquents que les sites polymorphes portant des allèles dérivés en plus forte fréquence ; θ_S comptabilisera donc tous les types de fréquence des variants, dont les variants en faible fréquence. θ_π mesure le nombre moyen de différences entre deux paires de séquences prises au hasard dans la population, d_{ij} correspondant au nombre de différences entre les séquences *i* et *j* (Tajima, 1989). θ_π mesure donc le degré moyen d'hétérozygotie sur l'ensemble du locus. Cette statistique est donc très sensible aux variants de fréquences intermédiaires. Enfin, θ_L est le nombre moyen de mutations (allèles dérivés) accumulées depuis le plus récent ancêtre commun de ces séquences (Zeng et al., 2006). Il est sensible à la proportion d'allèles dérivés ayant atteint des fréquences élevées. θ_L ne peut se calculer que si l'on connaît l'état ancestral/dérivé des allèles aux SNP présents sur le locus. Les statistiques des tests de neutralité comparent ces trois estimateurs :

$$D = \frac{\theta_\pi - \theta_S}{\sqrt{Var(\theta_\pi - \theta_S)}}$$

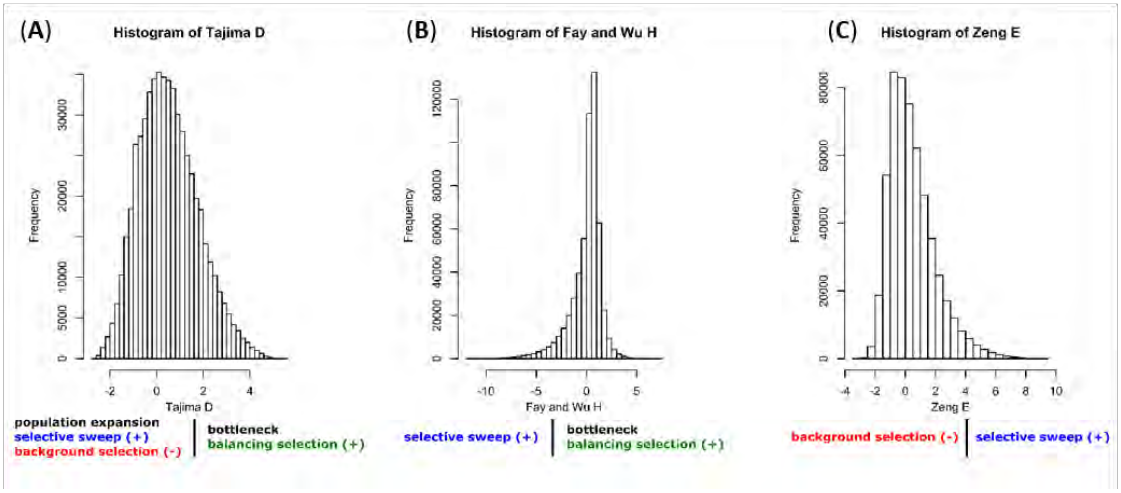
$$H = \frac{\theta_\pi - \theta_L}{\sqrt{Var(\theta_\pi - \theta_L)}}$$

$$E = \frac{\theta_L - \theta_S}{\sqrt{Var(\theta_L - \theta_S)}}$$

Le ***D*** de Tajima contraste deux estimateurs de θ qui sont sensibles aux variants présents en fréquences faibles et intermédiaires, le ***H*** de Fay & Wu contraste deux estimateurs de θ qui

sont sensibles aux variants présents en fréquences intermédiaires et fortes et le ***E*** de Zeng contraste deux estimateurs de θ qui sont sensibles aux variants présents en fréquences fortes et faibles. Sous neutralité, ces trois statistiques auront des espérances similaires, c'est-à-dire proches de 0, mais seulement si la population est stable (à l'équilibre mutation-dérive). En effet, les distributions de ces statistiques sous neutralité ne sont pas toujours centrées sur zéro car elles sont sensibles à différents scénarios démographiques pouvant mimer les effets de la sélection (goulot d'étranglement, expansion démographique ou structure génétique des populations). De manière générale, les trois statistiques ont des comportements différents en fonction des modèles de sélection (**Figure 13**).

Figure 13 : **Distributions théoriques des tests de Neutralité.** Distribution du *D* de Tajima (A), du *H* de Fay & Wu (B) et du *E* de Zeng (C). Sous chacune des distributions, il est indiqué comment les différents modes de sélection influencent ces distributions vers des valeurs positives ou négatives. Les forces démographiques de « bottleneck », d'expansion et de structuration génétique des populations influencent également ces statistiques.



Le ***D*** de Tajima est négatif en cas de balayage sélectif ou de sélection négative car il y a une réduction de l'hétérozygotie mais il n'est pas possible de discriminer ces deux évènements si l'on ne sait pas lequel des 2 allèles à un SNP donné est l'allèle dérivé. Si *H* est négatif et *E* positif sur le même locus, cela confirme l'hypothèse d'un balayage sélectif. Après un évènement de balayage sélectif, le retour à la neutralité est plus ou moins long selon les statistiques. *E* détecte plus facilement des balayages sélectifs après fixation, tandis que *H* détecte plus facilement des balayages sélectifs en cours. Inversement, lorsque le *D* et *H* sont positifs, c'est une signature de sélection balancée car il y a un maintien de l'hétérozygotie. Enfin, la sélection négative sur les mutations délétères est détectée par un *D* et un *E* négatifs (**Figure 13**).

Ainsi, ces statistiques de neutralité permettent de détecter différents types de signatures de sélection, mais deux de ces statistiques (*H* et *E*) requièrent la connaissance du statut ancestral/dérivé des allèles et, les trois statistiques peuvent être sensibles à différents scénarios démographiques pouvant « mimer » la sélection et induire des faux positifs. Pour ce travail, nous avons calculé les statistiques de neutralité *D*, *H* et *E* à l'échelle des gènes chez *Medicago truncatula* grâce aux données SNP du Medicago HapMap Project. L'objectif a été de comparer les patrons de polymorphisme de différents gènes impliqués ou non dans des associations épistatiques significatives. Parallèlement, les statistiques de neutralité ont aussi été calculées sur les données simulées sous sélection épistatique. Le modèle simulé sous neutralité nous permet d'obtenir la distribution de ces statistiques sous l'hypothèse nulle à laquelle nous comparons les distributions obtenues sous sélection épistatique (coadapté et compensatoire) ainsi que sous sélection additive. Les statistiques ***D***, ***H*** et ***E*** sont calculées sur les mêmes fenêtres génomiques simulées pour calculer les statistiques de DL (cor_{PC1} et cor_{PC1v}). Enfin, comme les statistiques de neutralité supposent qu'il n'y a pas de structure de population elles ont été calculées à l'échelle des sous-populations simulées, à l'intérieur desquelles il n'y a pas ou peu de sous-structuration (voir chapitre 1.4 – détails des simulations). Chez *M. truncatula* nous avons également réalisé ces calculs dans les sous-populations décrites dans la littérature.

1.4 Description des Simulations

La recherche de sélection épistatique chez *Medicago truncatula* vise à identifier des gènes dont l'interaction est soumise à la sélection. Afin de comprendre les mécanismes évolutifs de la sélection épistatique et de développer les outils statistiques permettant sa détection, des simulations ont été réalisées pour modéliser ces interactions adaptatives. La première étape des simulations consiste à simuler une population ancestrale en « backward » par coalescence afin de générer une série d'haplotypes. Puis la population ancestrale diploïde, évolue en « forward ». Pendant 100 générations la population évolue sous neutralité puis nous faisons 200 générations avec ou sans sélection en fonction des modèles évolutifs. Deux programmes sont utilisés pour réaliser les simulations « backward » et « forward ».

1.4.1 Les simulations « backward » par coalescence

Les simulations « backward » sont générées avec le logiciel SCRM (Staab et al., 2015). À chaque simulation, le logiciel génère des données de polymorphisme avec une approche par coalescence ; deux copies d'un gène qui coalescent ont un ancêtre commun. SCRM utilise une méthode d'approximation, la coalescence séquentielle de Markov (SMC). Cette méthode a été développée pour simuler la coalescence avec de la recombinaison à l'échelle d'un chromosome entier car le modèle de coalescence standard est trop coûteux en temps de calcul pour de longues régions génomiques (Kelleher et al., 2016; Marjoram & Wall, 2006). SMC a été introduit par McVean et Cardin (G. A. T. McVean & Cardin, 2005), le programme génère un arbre généalogique puis des événements de recombinaison sont rajoutés uniformément pour modifier cet arbre et créer la généalogie finale. Cette méthode produit des patrons de DL similaires, au moins sur de courtes distances, à ceux générés par les méthodes exactes de reconstruction du coalescent avec recombinaison. (Yang et al., 2014). SCRM autorise la mutation ($\theta = 4N\mu$), la recombinaison ($\rho = 4Nc$) et il est également possible de choisir la taille de l'échantillon ainsi que la taille des locus simulées (**Tableau 3**). En sortie, SCRM génère un fichier texte contenant les génotypes haploïdes aux SNP bi-alléliques, codés 0 pour les allèles ancestraux et 1 pour les allèles dérivés. Les positions des SNP sur les locus sont connues et les locus simulés font une taille de 5Mb afin d'être à l'échelle d'une large région chromosomique et de pouvoir calculer une matrice d'apparentement entre les individus simulés (voir chapitre 1.2.2). Pour chaque simulation d'une population ancestrale, quatre chromosomes sont générés et les chromosomes contiennent environ 15 000 SNP, soit 1 SNP pour 333 paires de bases en moyenne. Les paramètres de simulations sont présentés dans le **Tableau 3**.

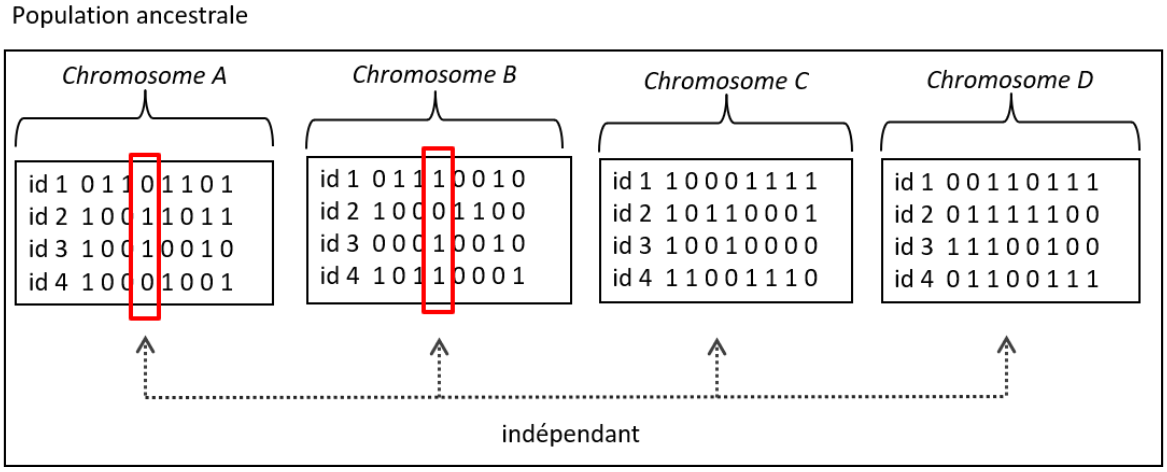
Commande SCRM : *scrm $N_{(nhap)}$ l -t theta -r rho l -p 8 > fichier_output.txt*

Tableau 3: Paramètres des simulations « backward » (logiciel SCRM) et « forward » (SIMUPOP).

Paramètres	Valeurs	Description
l	5Mb	Taille d'un chromosome
N_0	100000	Taille de la population ancestrale
$N_{(nhap)}$	1000	Taille de la population échantillonnée dans SCRM
Nrep	4	Nombre de chromosomes simulés
c	10^{-8}	Probabilité de recombinaison par paire de base par génération
$\rho = 4 * N_0 * c * l$	20000	Taux de recombinaison à l'intérieur du locus dans la population
μ	10^{-9}	Probabilité de mutation par paire de base par génération
$\theta = 4 * N_0 * \mu * l$	2000	Paramètre de mutation pour le locus dans la population

Parmi les quatre chromosomes, deux SNP au milieu des deux chromosomes A et B sont désignés comme les SNP qui seront soumis à la sélection épistatique (**Figure 14**).

Figure 14 : **Présentation schématique des chromosomes simulés.** Les quatre chromosomes simulés sont indépendants. Ils portent environs 15000 SNP représentés en colonne sur ce schéma et sont constitués de 500 individus diploïdes représentés en ligne. Les deux SNP situés au milieu des deux chromosomes A et B sont les SNP qui sont soumis à la sélection dans les différents modèles (SNP encadrés en rouge).



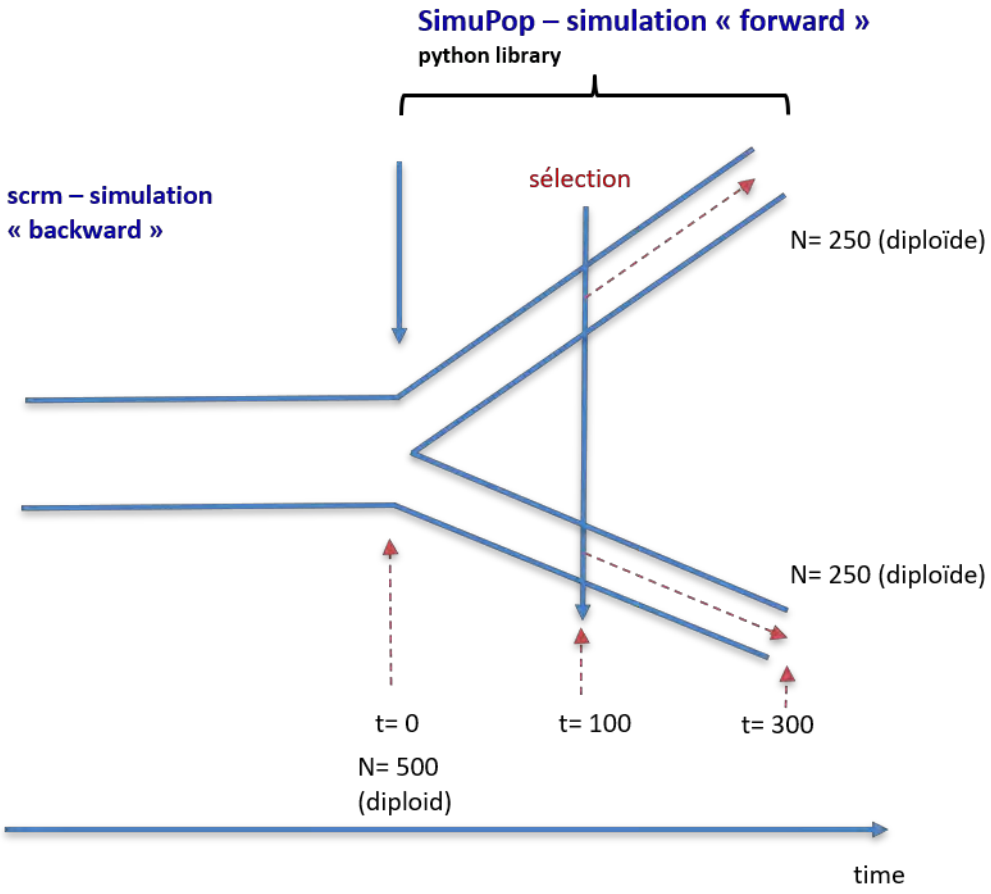
De plus, nous avons échantillonné les simulations SCRM pour lesquelles les mutations à ces deux SNP sont en fréquences intermédiaires (0.25 < fréquence < 0.75) afin d'améliorer l'efficacité de la sélection épistatique (standing variation) lors des simulations « forward ». En effet, si les allèles soumis à la sélection épistatique sont préalablement en « standing variation » cela améliore l'efficacité de la sélection épistatique par rapport à la dérive génétique (Takahasi, 2007, 2009; Takahasi & Innan, 2008). Ainsi, pour chaque simulation, une population ancestrale de quatre chromosomes est générée par coalescence puis nous faisons

évoluer la population pendant 300 générations avec un programme de simulation « forward in time ».

1.4.2 Simulations « forward » avec SimuPop – python

Les simulations « forward » sont réalisées avec la bibliothèque python simuPop (Peng & Amos, 2008; Peng & Kimmel, 2005). Les fichiers de sortie SCRM sont le point de départ de la simulation « forward » à partir desquels la population va évoluer pendant 300 générations (**Figure 15**). Le programme modélise des individus diploïdes avec leurs génotypes et simule la transmission de ces génotypes individuels aux cours des générations. Plusieurs opérateurs sont appliqués à la population durant l'évolution et ils s'appliquent avant, après ou pendant la reproduction des individus diploïdes. Ces opérateurs sont les probabilités de recombinaison et de mutation, les différents régimes de reproduction et la sélection. Les probabilités de mutation et de recombinaison sont identiques aux probabilités tableau 3. Le nombre de mutations appliquées à la population à chaque génération est modélisé par une fonction « *infinite site model* » préalablement crée dans l'équipe (Stage Master 2 M. Negretto, 2015). Ainsi, chaque nouvelle mutation apparait à une nouvelle position et le nombre de mutations générées par génération est modélisé par une loi de Poisson de paramètre $\lambda = 4 * N * \mu * l$, où N est la taille de la population échantillonnée avec SCRM soit 500 individus diploïdes (ou 1000 haplotypes), l correspond à la taille d'un chromosome, soit 5Mb et μ est la probabilité de mutation par paire de base par génération (**Tableau 3**).

Figure 15 : Représentation graphique de l'évolution des populations simulées. La population ancestrale est générée avec une méthode par coalescence puis la population évolue pendant 300 générations avec une méthode de simulation «forward in time».



Au début de la simulation, les deux SNP au milieu de deux chromosomes A et B (**Figure 14**) sont désignés comme les SNP qui seront sous sélection et à chaque génération, une *fitness* est attribuée à tous les individus en fonction des allèles qu'ils portent à ces deux SNP. C'est ainsi que les génotypes des individus qui ont de meilleures *fitness* seront davantage transmis aux descendants aux cours de la reproduction.

A la génération 0, la population est divisée en deux sous-populations afin de générer de la structuration génétique et la reproduction ne peut se faire qu'entre les individus d'une même sous-population. Ainsi, nous avons simulé un modèle d'évolution neutre entre deux locus indépendants afin d'avoir la distribution du DL sous l'hypothèse nulle. Parallèlement, les modèles de sélection ont été simulés afin d'évaluer les puissances de détection (1-β) des statistiques de DL en fonction de l'erreur de type I (α) calculée sous l'hypothèse nulle. Trois modèles de sélection ont été simulés ; les modèles de sélection épistatique coadapté et

compensatoire et un modèle de sélection additive, c'est-à-dire indépendante entre les locus. Les valeurs de *fitness* attribuées aux individus en fonction des allèles aux deux locus sélectionnés sont présentées dans le **Tableau 2**, sachant que nous avons attribué une valeur sélective *s* de 0.1. Cette valeur de *s* a été choisie (analyses pilotes) afin d'avoir une efficacité de sélection acceptable sachant que la population simulée est de taille modeste (250 individus diploïdes par sous-population) et compte tenu des temps de calcul nécessaires pour simuler sur 300 générations de petits génomes avec une telle densité de SNP. De plus, nous avons simulé deux modes de reproduction ; un régime en panmixie (« random-mating ») où les individus se reproduisent de manière aléatoire et un régime à 95% d'autofécondation (« self-mating ») qui nous permet de correspondre au mieux à une espèce comme *Medicago truncatula* (Siol et al., 2008) ou *Arabidopsis thaliana*. Ces deux modes de reproduction situés aux deux extrêmes nous permettent d'évaluer l'influence de l'apparentement sur les signatures de sélection épistatique et le DL. Enfin, les simulations de chaque modèle de sélection (i.e. coadapté, compensatoire et additif) ont été réalisées avec des mutations dominantes, codominantes ou récessives (voir **Tableau 2**).

Pour résumer, une simulation est constituée de 4 chromosomes d'environ 15 000 SNP et de 500 individus diploïdes répartis en deux sous-populations, et chacune des simulations est répétée 1 000 fois pour toutes les combinaisons de paramètres suivants : (i) les deux modes de reproduction, (ii) les trois modèles de sélection plus le modèle neutre (iii) et les trois modes d'interaction entre allèles dérivés et ancestraux au sein de chaque locus (dominance, codominance et récessivité). Au cours des simulations, nous suivons les fréquences alléliques des SNP cosélectionnés pendant les 300 générations et les statistiques de DL (i.e. r, r_v, cor_{PC1} et cor_{PC1v}) sont calculées toutes les 20 générations sur la population entière (les 500 individus diploïdes répartis en deux sous-populations). Les statistiques cor_{PC1} et cor_{PC1v} sont calculées sur des fenêtres génomiques de 10kb (i.e. 5kb de part et d'autre du SNP focal). La matrice d'apparentement est calculée (voir chapitre 1.2.2.3) toutes les 20 générations, à chaque fois que le DL est calculé, et à partir des 4 chromosomes simulés. Pour le calcul de la matrice d'apparentement, nous appliquons une Minor Allele Frequency (MAF) supérieure à 5% sur les SNP et pour le calcul du DL, nous ne faisons pas de filtre sur la MAF sur les régions génomiques autour des SNP cosélectionnés. De plus, les statistiques de tests de neutralité (*D*, *H* et *E*) sont également calculées sur les mêmes fenêtres génomiques à partir desquelles nous avons

calculé le DL. Enfin, les 1000 simulations réalisées pour chacune des combinaisons de paramètres nous permettent d'obtenir les distributions empiriques de toutes les statistiques de DL.

1.5 Résultats des simulations

Les simulations ont été réalisées sur le cluster de calcul de la Genotoul et nous remercions les administrateurs sans quoi ce travail n'aurait pas été possible. Ainsi, plusieurs scénarios démographiques et plusieurs modèles de sélection ont été simulés et sur l'ensemble de ces paramètres, les populations simulées ont généré une masse de données importante d'environs 865Go. La population ancestrale simulée avec une méthode de coalescence est constituée de 1000 simulations de 4 chromosomes pour 1000 individus haploïdes chacune, et elle représente environ 115Go. Le logiciel SCRM permettant de générer la population ancestrale par coalescence a été utilisé sur le cluster de calcul de la Genotoul. Les simulations sont réalisées à l'aide d'un script python nous permettant de filtrer les données générées pour ne garder que les simulations où les chromosomes A et B portent des mutations en fréquences intermédiaires (i.e. 0.25 < freq(a) et freq(b) < 0.75) aux deux SNP qui seront cosélectionnés par la suite. Le logiciel SCRM est directement utilisé depuis les scripts python (**Annexe 1**). Une fois la population ancestrale créée, elle sert de point de départ pour les différentes populations simulées avec la méthode « forward in time ». Les simulations « forward in time » ont été générées également sur le cluster de calcul à l'aide de scripts python avec la bibliothèque simuPop (Peng & Amos, 2008 ; Peng & Kimmel, 2005) installée sur le cluster. Au total 1000 simulations ont été effectuées pour chacune des 20 combinaisons de paramètres qui ont été décrites (reproduction, modèles de sélection, et interaction de dominance entre les allèles dans les modèles de sélection), soit 20000 simulations ce qui représente une masse de données d'environ 750Go. Pour chaque condition, les 1000 simulations ont été réalisées à partir d'un script python qui fait évoluer la population génération après génération (**Annexe 2**).

Afin de vérifier que les simulations ont été correctement réalisées, la première partie des résultats présente les données produites. Nous avons (i) estimé le spectre de fréquences alléliques (SFS) neutre dans la population ancestrale, (ii) suivi l'évolution de la structure et de l'apparentement des populations simulées à l'aide des statistiques F_{IS} et F_{ST} et (iii) suivi l'évolution des fréquences alléliques des SNP destinés à être les cibles de la sélection. Dans

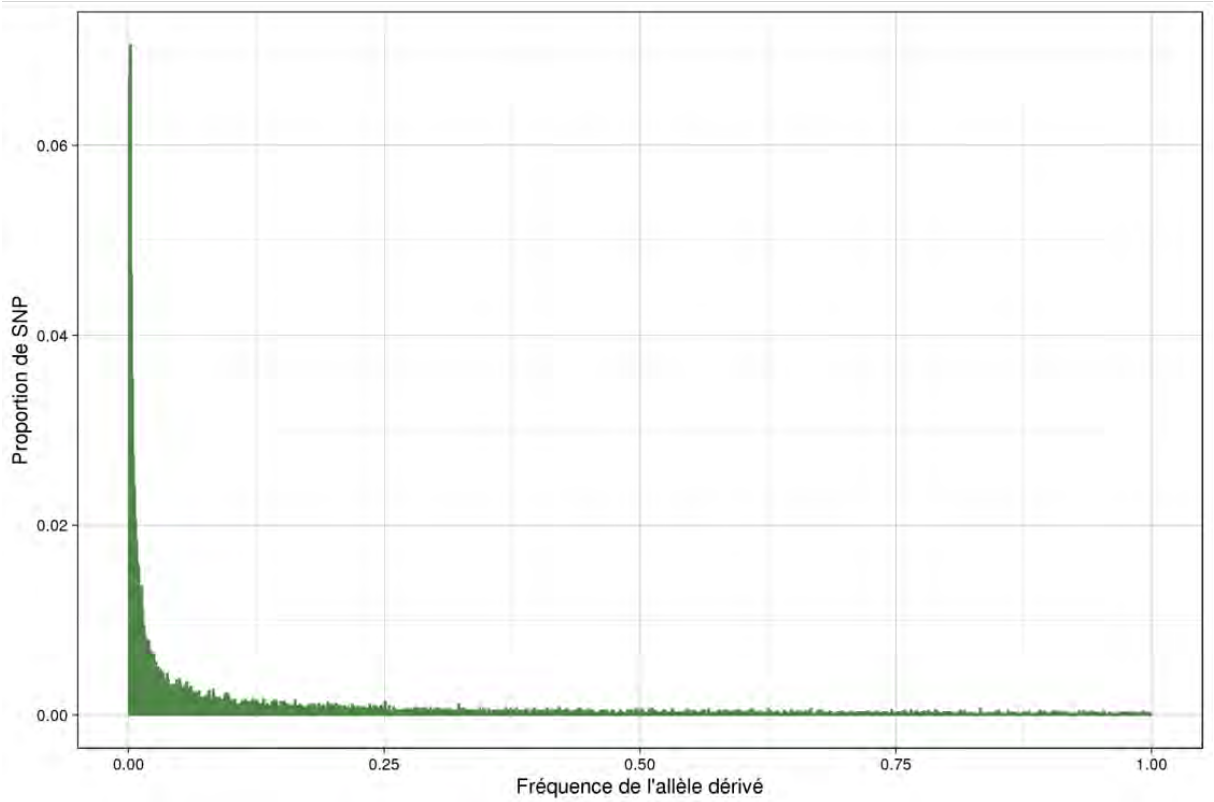
une deuxième partie, nous présentons les résultats de DL entre les paires de locus simulés, ainsi que les puissances de détection de la sélection épistatique et les proportions de faux positifs des statistiques de DL. Enfin, la dernière partie des résultats est consacrée aux statistiques de tests de neutralité qui ont été calculées sur les données simulées.

1.5.1 Contrôle qualité des simulations

Les simulations ont été réalisées afin de comprendre les mécanismes évolutifs liés à la sélection épistatique et tester les outils statistiques permettant sa détection.

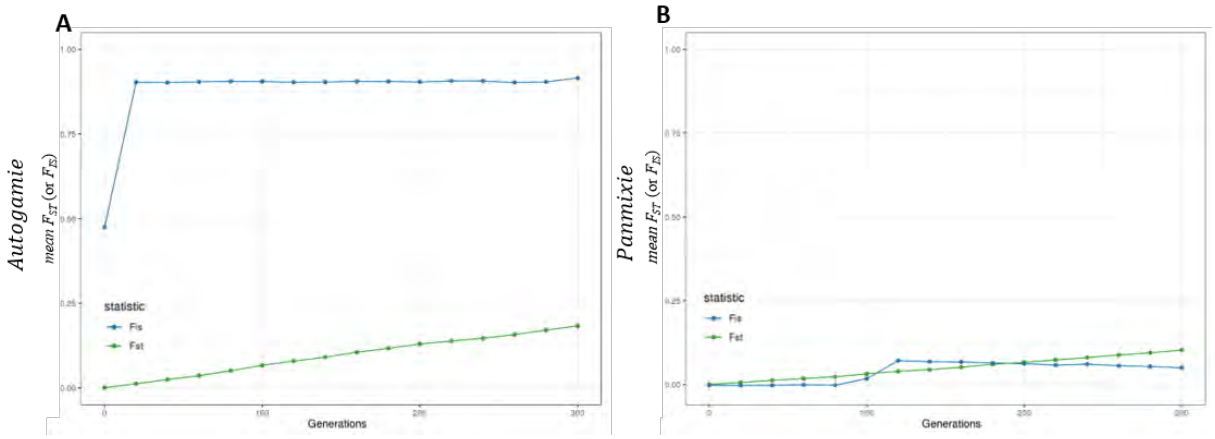
Afin de vérifier que le polymorphisme génétique de la population ancestrale de départ (avant la séparation en deux populations) reflète bien celui d'une population de taille constante évoluant sous un modèle neutre, nous avons calculé son spectre de fréquences alléliques (SFS). La **Figure 16** présente un spectre de fréquences (SFS) calculé à la génération 0 sur un chromosome simulé par coalescence. Le SFS mesure la distribution du nombre de sites polymorphes (SNP) selon les classes de fréquences des mutations (allèle dérivé) chez les 500 individus diploïdes simulés. Ce spectre a une forme en « L » indiquant une forte proportion de SNP ayant des mutations en faibles fréquences et un nombre décroissant de SNP avec des mutations en fréquences croissantes dans la population, typique de ce qui est attendu dans un modèle neutre d'évolution (**Figure 12**). Les chromosomes générés par coalescence ont été correctement simulés et peuvent servir de point de départ pour les simulations « forward in time ».

Figure 16 : Spectre des fréquences alléliques obtenu sur un chromosome simulé, à la génération 0, dans la population ancestrale. L'axe des X représente les classes de fréquence *i* de l'allèle dérivé sur les 1000 haplotypes simulés - ou 500 individus diploïdes - ($1 \leq i \leq 1000$). L'axe des Y indique la proportion de SNP appartenant à chaque classe de fréquence de l'allèle dérivé.



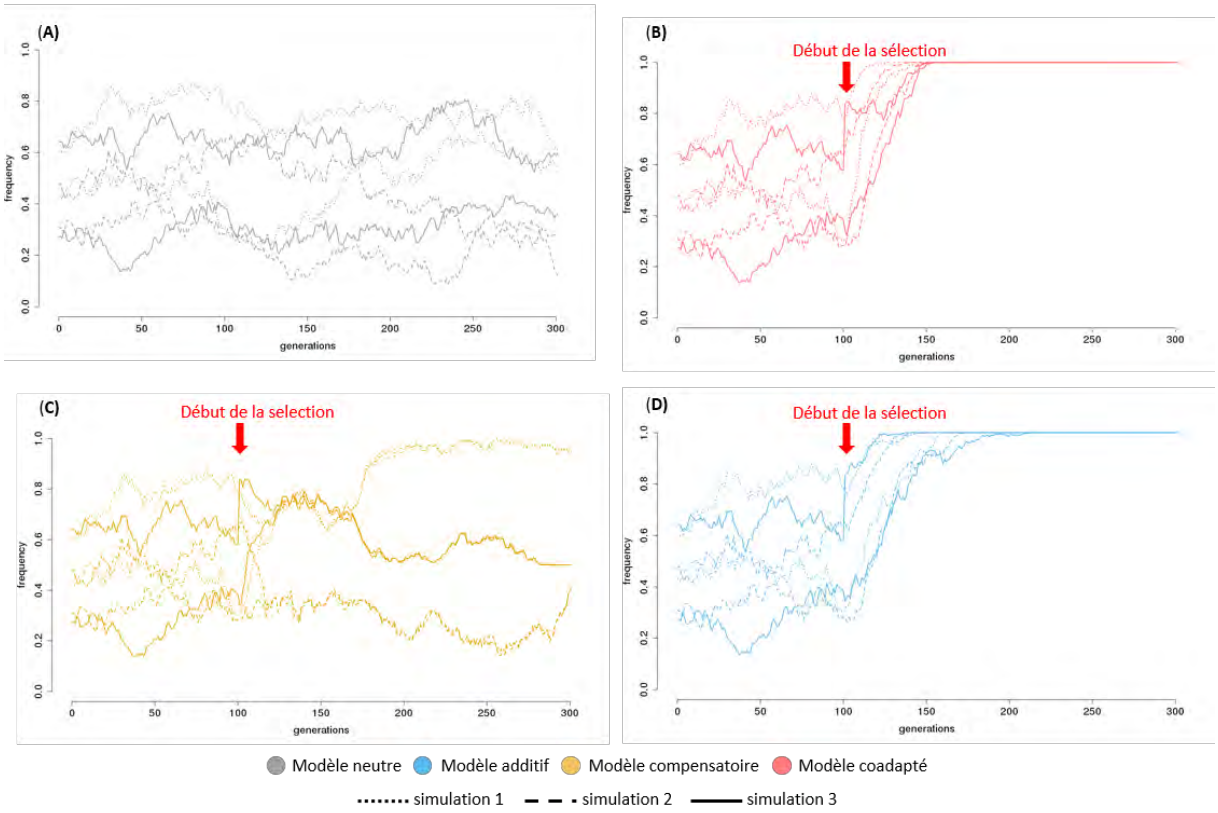
Afin de vérifier que les simulations « forward in time » ont été correctement réalisées, nous avons calculé les statistiques F_{ST} $\left(F_{ST} = \frac{var(p_t)}{p_0(1-p_0)}\right)$ et F_{IS} $\left(F_{IS} = \frac{H_e - H_o}{H_e} = 1 - \frac{H_o}{H_e}\right)$ pour estimer le degré de structuration génétique des populations et l'apparentement au sein des populations, au cours des générations. À l'issue des simulations (génération 300), les F_{IS} moyens obtenus dans les populations simulées en autogamie et panmixie étaient respectivement de 0.92 et 0.07, tandis que les F_{ST} moyens étaient de 0.19 et 0.10 (**Figure 17**). Ces résultats sont cohérents avec les paramètres des simulations que nous avons fixés, les populations autogames à 95% présentant un taux de consanguinité élevé montrant un déficit global en hétérozygotie et les valeurs moyennes de F_{ST} montrent une différenciation des populations modérée à forte selon les régimes de reproduction (Wright, 1978).

Figure 17 : Evolution de l'indice de fixation F_{ST} et du coefficient de consanguinité F_{IS} des populations simulées sur 300 générations. Evolution des coefficients F_{ST} et F_{IS} en autogamie (**A**) et en panmixie (**B**), calculés dans le modèle d'evolution neutre.



Enfin, nous avons suivi l'évolution des fréquences alléliques des SNP cosélectionnés. Des exemples d'évolution des fréquences alléliques sont montrés en **Figure 18** ; nous suivons l'évolution des fréquences des allèles dérivés *a* et *b* dans les populations globales des SNP situés au milieu des deux chromosomes A et B et qui sont la cible de la sélection dans les différents modèles évolutifs. Les fréquences alléliques sont représentées pour trois simulations prises au hasard dans chacun des quatre modèles de sélection et pendant les 300 générations de simulation. Dans le modèle neutre, on observe une évolution aléatoire des fréquences entre les paires d'allèles *a* et *b* aux trois simulations. Dans le modèle de sélection épistatique coadapté, il y a une cofixation rapide des allèles *a* et *b* aux trois simulations et dans le modèle de sélection additif, on observe une cofixation des allèles mais l'évolution des fréquences semble moins corrélée entre *a* et *b*. Enfin, dans le modèle compensatoire, on observe une coévolution des fréquences entre les paires d'allèles *a* et *b* dans les trois simulations avec un maintien du polymorphisme. Cette figure montrant l'évolution des fréquences alléliques ne représente pas les différents modèles de reproduction ni les différentes formes d'interaction entre les allèles dérivés et ancestraux (les simulations montrées ici sont réalisées en autogamie avec codominance des allèles), mais elle permet de visualiser le type d'évolution des fréquences des allèles au cours des générations.

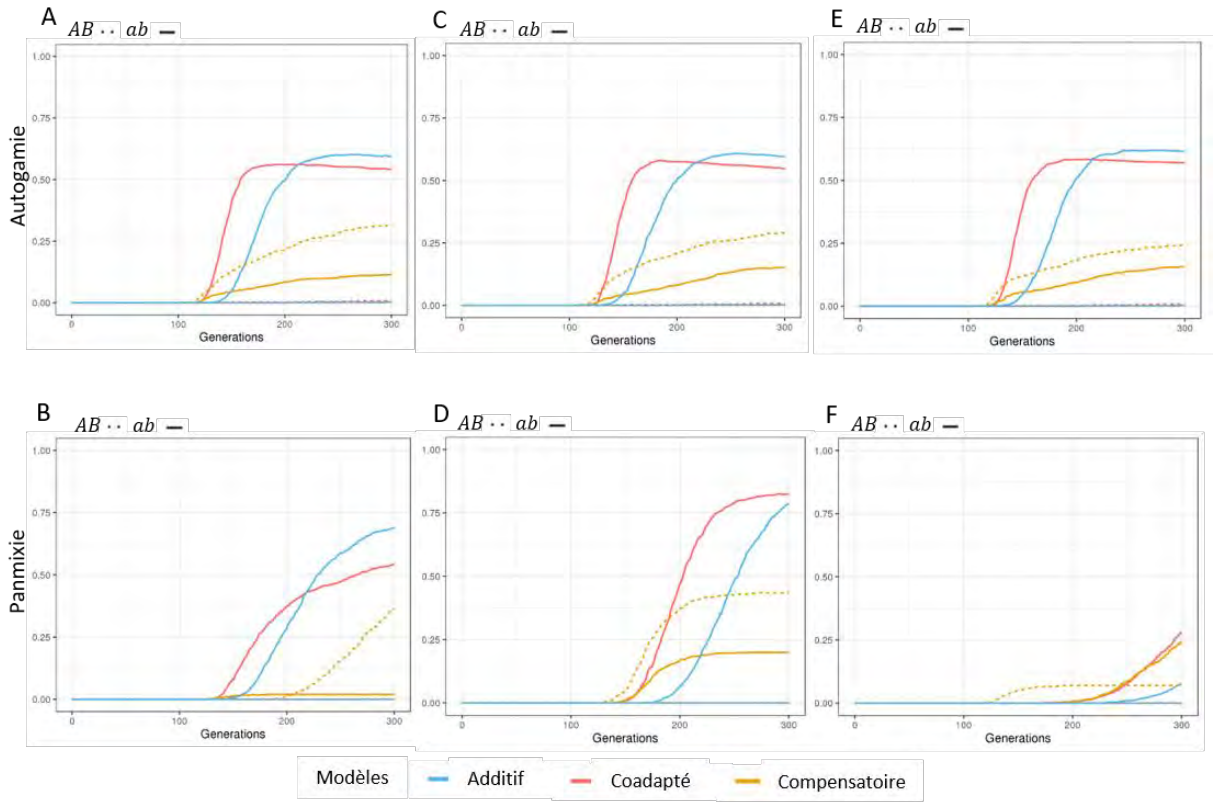
Figure 18 : Evolution des fréquences alléliques *a* et *b* des SNP cosélectionnés au cours des simulations. Les graphiques (A), (B), (C) et (D) représentent l'évolution des fréquences alléliques des allèles dérivés *a* et *b* au niveau des SNP co-sélectionnés sur trois simulations prises au hasard et dans le modèle neutre, dans les modèles de sélection épistatique coadapté et compensatoire et dans le modèle additif. Pour chacune des trois simulations représentées, nous suivons l'évolution des fréquences des deux allèles dérivés (*a* et *b*) aux deux locus A et B. Les fréquences des allèles dérivés qui sont représentées sont issues des simulations réalisées en autogamie avec le modèle de mutation codominant.



Afin de rendre compte de l'efficacité de la sélection épistatique simulée, la **Figure 19** présente les taux de cofixation des haplotypes *ab* cosélectionnés (respectivement *AB*) dans les populations globales. Les taux de cofixation sont calculés sur la base des 1000 simulations réalisées pour les quatre modèles évolutifs (neutre, coadapté, additif et compensatoire) ainsi que pour les deux régimes de reproduction (autogamie et panmixie) et les trois modes d'interactions entre les allèles (récessif, codominant et dominant). La **Figure 19** montre que les modèles coadapté et additif ont des taux de cofixation plus élevés que le modèle compensatoire car ce sont deux modèles de sélection positive. Dans le modèle compensatoire, il y a un maintien du polymorphisme car la sélection se fait sur les haplotypes *ab* et *AB* et les deux allèles aux deux SNP ségrégent la plupart du temps pendant les 300

générations, bien que dans certaines simulations l'un des deux haplotypes *ab* ou *AB* peut se fixer dans l'une ou l'autre (ou les deux) sous-populations.

Figure 19 : Taux de cofixation des haplotypes mutants (*ab*) et ancestraux (*AB*) des locus simulés.
Les taux de cofixations des haplotypes mutants (*ab*) et ancestraux (*AB*) au cours des générations montrent l'efficacité de la sélection dans le modèle en autogamie (**A** - **C** - **E**) et dans le modèle en panmixie (**B** - **D** - **F**). Les dynamiques de cofixation sont représentées pour les mutations récessives (**A**- **B**), codominantes (**C** - **D**) et dominantes (**E** – **F**). Les allèles *A*/*a* and *B*/*b* sont les allèles des SNP A et B sous sélection épistatique et qui sont situés au milieu des chromosomes A et B simulés.



D'autre part, la fixation est plus rapide dans le modèle coadapté par rapport au modèle additif notamment en autogamie où la récessivité, la codominance et la dominance des mutations n'a pas d'effet car il y a très peu d'individus hétérozygotes ($F_{IS} = 0.92$). En autogamie, le taux de cofixation des allèles dérivés atteint plus rapidement un plateau par rapport à la panmixie. En panmixie, le mode d'interaction entre les allèles d'un même locus (récessif, codominant et dominant) influence fortement la dynamique de cofixation car les allèles sélectionnés sont présents dans plusieurs génotypes sélectionnés notamment chez les individus hétérozygotes ($F_{IS} = 0.07$) (**Tableau 2**). Si les mutations sont codominantes, le taux de cofixation est plus élevé en panmixie qu'en autogamie mais si les mutations sont dominantes le taux de cofixation est plus faible en panmixie. En effet, en cas de dominance, les individus

hétérozygotes qui portent l'allèle ancestral ont le même coefficient de sélection que les individus homozygotes pour l'allèle dérivé et cela favorise le maintien des deux allèles dans la population (**Tableau 2**). Enfin, les taux de cofixation des allèles ancestraux *A* et *B* sont proches de zéro dans les modèles coadapté et additif (puisque l'on sélectionne *a* et *b*) tandis que dans le modèle compensatoire, les allèles ancestraux se maintiennent et le taux de cofixation de l'haplotype *AB* est même supérieur à celui de *ab*. Ce résultat peut être expliqué par le fait que les allèles aux deux SNP soumis à la sélection sont en fréquences intermédiaires dès de début des simulations (0.25 < freq(*a*) ou freq(*b*) < 0.75) mais cette distribution n'est pas uniforme, et les allèles ancestraux sont toutefois plus fréquents (**Annexe 3**). En effet, si l'on se place à la génération 0, la médiane de la fréquence de l'allèle *a* dans la population totale est de 0.43 et à la génération 100 (modèle neutre) elle est de 0.446 en autogamie et 0.435 en panmixie (**Annexe 3**). Pour résumer, le modèle de sélection épistatique compensatoire présente un faible taux de fixation des allèles dérivés cosélectionnés car il y a maintien du polymorphisme tandis que le modèle de sélection épistatique coadapté et le modèle de sélection additive amènent plus souvent à la fixation des allèles dérivés. De plus, le modèle codominant est le modèle pour lequel nous observons les taux de cofixation les plus élevés notamment en panmixie ; nous détaillerons principalement les résultats obtenus en codominance pour la suite des résultats. Enfin, les valeurs de cofixation que nous avons obtenues doivent être interprétées au regard des populations simulées qui sont de taille relativement faible (N=250 individus par sous-population) et qui ont des régimes de reproduction différents. En effet, l'efficacité de la sélection dépend du produit $N_e s$ où s est le coefficient de sélection et N_e la taille efficace de la population. Ainsi l'efficacité de la sélection augmente avec la taille de la population d'un facteur Ns qui dépend de N_e lui-même étant directement relié au mode de reproduction (Glémin, 2007).

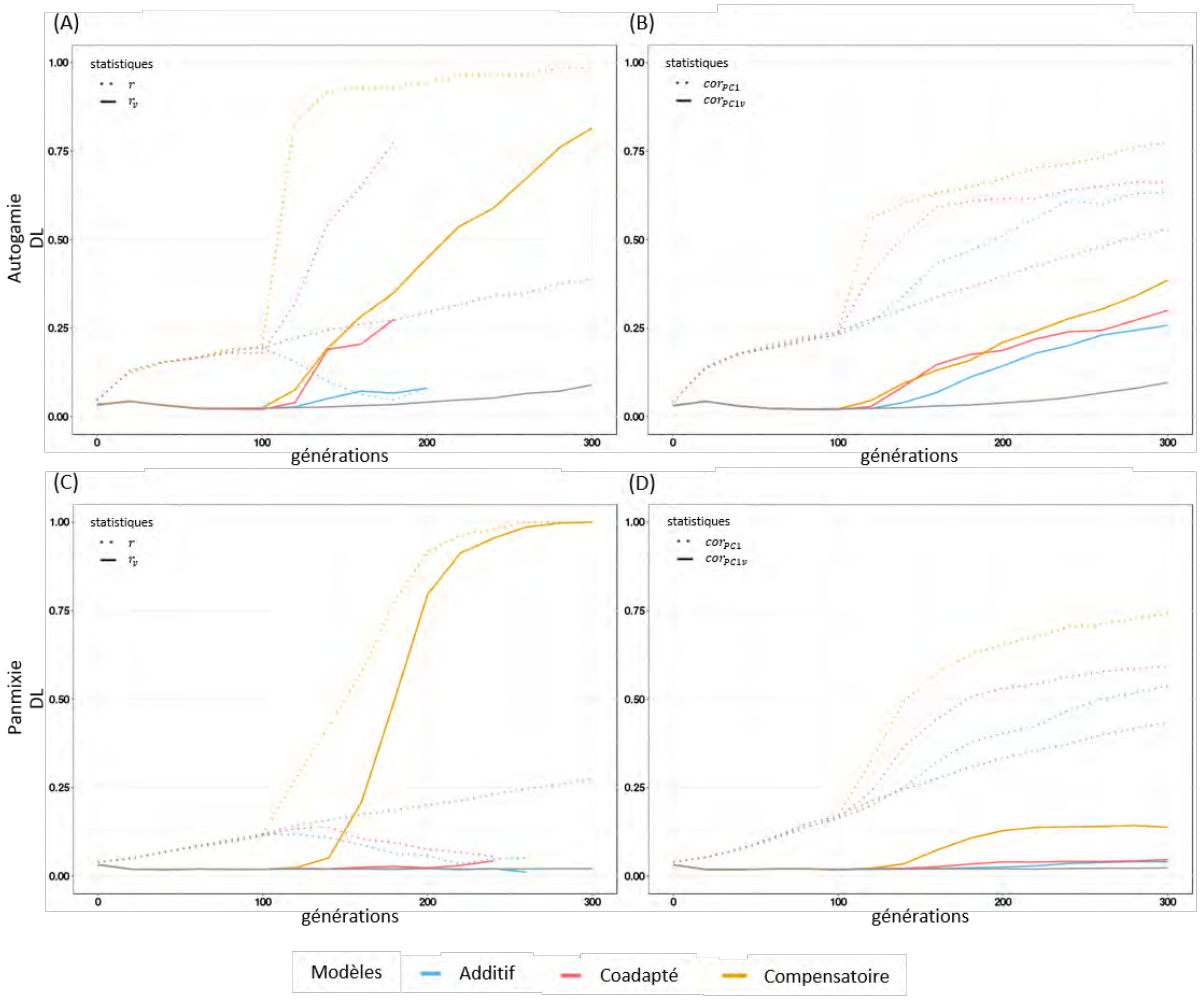
1.5.2 Déséquilibre de liaison entre paires de locus sous sélection épistatique

L'évolution du DL moyen calculé entre paires de locus simulés est présentée sous sélection épistatique, additive et sous neutralité pour les deux sous-populations évoluant soit en autogamie, soit en panmixie et sous des modèles de sélection avec des allèles codominant (à chaque locus) (**Figure 20**). Nous avons sélectionné ce modèle en codominance car il présente les meilleurs taux de cofixation (**Figure 19**) dans les populations panmictiques

sachant que les taux de cofixation en autogamie sont similaires quel que soit le mode d'interaction entre les allèles (récessifs, codominants ou dominants). Afin de comparer les modèles de sélection et éviter les biais d'échantillonnages dans le calcul du DL moyen sur l'ensemble des simulations, 500 simulations qui cofixent les allèles dérivés dans les 2 sous-populations, avant la génération 300, sont échantillonnées parmi les 1000 simulations des modèles coadapté et additif. Dans le modèle compensatoire, 500 simulations sont échantillonnées aléatoirement. Pour plus de lisibilité, le DL moyen entre deux locus est calculé sur la base des valeurs absolues du coefficient de corrélation à chaque simulation. Enfin, les résultats de DL obtenus avec les simulations réalisées sous les modèles de sélection avec les allèles récessifs ou dominants sont représentés en **Annexe 4**. Dans le modèle de sélection avec les mutations codominantes, (**Figure 20**) on observe tout d'abord que sous neutralité (courbe grise) la structure des populations et l'apparentement génèrent du DL entre les locus indépendants et les statistiques r et cor_{PC1} peuvent atteindre des valeurs autour de 0.25 à 0.5 à la dernière génération sans que nous ayons appliqué de sélection. Ce DL de fond observé dans le modèle neutre est réduit et devient en moyenne proche de zéro lorsqu'il est calculé avec les statistiques r_v et cor_{PC1v} qui utilisent la correction par la matrice d'apparentement V. Ce résultat montre bien l'importance d'utiliser les statistiques r_v et cor_{PC1v} pour calculer le DL et réaliser ensuite le test statistique afin d'évaluer la significativité du coefficient de corrélation en tenant compte du DL de fond lié aux facteurs démographiques et à l'apparentement. Les modèles avec les mutations dominantes et récessives (**Annexe 4**) montrent également que le DL est globalement réduit dans le modèle neutre lorsqu'il est calculé avec les statistiques r_v et cor_{PC1v} corrigées pour la structure et l'apparentement. D'autre part, les modèles de sélection génèrent plus de DL que le modèle neutre bien que cela dépende des modèles et du régime de reproduction. En autogamie, les trois modèles de sélection génèrent plus de DL que le modèle neutre mais le DL observé en additif reste inférieur au DL observé dans les modèles de sélection épistatique (**Figure 20**). Cependant, dans le modèle additif-autogamie, r diminue dès le début de la sélection tandis que r_v augmente et le r observé dans le modèle additif reste inférieur au r observé sous neutralité. Ce résultat surprenant peut s'expliquer par un effet de structure marqué sous neutralité qui serait atténué dans le modèle additif par l'effet de la sélection et la fixation indépendante sur les allèles dérivés. En panmixie, en revanche, seul le modèle compensatoire génère du DL détectable avec r et r_v. Les modèles coadapté et additif génèrent peu de DL. De plus, les

statistiques r et r_v ne sont plus calculables après quelques générations de sélection notamment dans les modèles coadapté et additif car les allèles aux deux SNP cosélectionnés sont fixés (**Figure 20, Annexe 4**). En revanche, avec cor_{PC1} et cor_{PC1v}, le DL peut se calculer pendant toute la durée des simulations car nous utilisons une approche par fenêtre qui prend plusieurs SNP sur 10kb autour de chacun des SNP soumis à la sélection.

Figure 20 : Evolution du DL entre les locus simulés sous sélection épistatique, additive et sous neutralité pendant 300 générations. Données observées pour 500 simulations avec N=500 individus diploïdes, s=0.1 et des mutations codominantes. (A, B) système de reproduction autogame à 95% et (C, D) 100% panmixie. (A, C) le DL moyen est calculé au niveau des SNP sous sélection avec les statistiques r et r_v. (B, D) le DL moyen est calculé sur des fenêtres génomiques avec les statistiques cor_{PC1} et cor_{PC1v}. Les courbes grises correspondent au modèle neutre, les courbes rouges au modèle coadapté, les courbes jaunes au modèle compensatoire et les courbes bleus au modèle additif. Les traits pleins représentent les statistiques r_v et cor_{PC1v} corrigées par la matrice kinship et les traits pointillés correspondent aux statistiques r et cor_{PC1} non corrigées par la matrice.



Pendant toute la durée des simulations et même après que les SNP cosélectionnés ont été fixés, nous observons toujours une différence entre le modèle neutre et les modèles de sélection épistatique montrant qu'il y a un effet d'auto-stop (« hitch-hiking ») sur les SNP qui

environnent les SNP sélectionnés dans la fenêtre de 10kb. Cependant, nous ne pouvons pas bien distinguer les modèles de sélection épistatique du modèle additif en autogamie lorsque le DL est calculé à l'échelle des haplotypes (statistiques cor_{PC1} et cor_{PC1v}) car il y a une forte structure haplotypique. En effet, les SNP sous sélection portent bien les allèles *a* et *b* mais il y a une hétérogénéité des haplotypes sous sélection dans les sous-populations, et ce pour tous les modèles d'interaction entre les allèles. La différence entre les modèles est plus grande si le DL est calculé à l'échelle des SNP (statistiques r et r_v). En panmixie, lorsque le DL est calculé avec la méthode des haplotypes, les modèles additif et coadapté ne montrent pas de DL par rapport au modèle neutre tandis que le modèle compensatoire présente une évolution du DL légèrement supérieure au modèle neutre avec les mutations codominantes et dominantes (**Figure 20, Annexe 4**).

Pour résumer, les statistiques de DL qui se situent à l'échelle des SNP (r et r_v) semblent plus efficaces pour détecter la sélection épistatique que les statistiques basées sur les haplotypes (cor_{PC1} et cor_{PC1v}) mais malgré cela, les statistiques r et r_v ne peuvent détecter la sélection si un des allèles à un SNP a été fixé (ou un allèle à chaque SNP). Les statistiques cor_{PC1} et cor_{PC1v} prennent en compte le polymorphisme des SNP environnant les SNP sous sélection dans une fenêtre génomique définie afin d'exploiter l'effet d'auto-stop même après la fixation des allèles des SNP sous sélection.

1.5.3 Contrôle du taux de faux positifs et puissance de détection des statistiques de DL

Sous l'hypothèse nulle d'indépendance entre deux variables X^l et X^m, ou $PC1^l$ et $PC1^m$ (respectivement $V^{-1/2}X^l$ et $V^{-1/2}X^m$, ou $V^{-1/2} * PC1^l$ et $V^{-1/2} * PC1^m$), les valeurs du coefficient de corrélation, à travers la statistique $T = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}$, sont supposées suivre la distribution de Student $\tau_{(n-2)}$. Nous avons évalué l'ajustement des statistiques T_r, T_{r_v}, $T_{cor_{PC1}}$ et $T_{cor_{PC1v}}$ à cette distribution théorique de Student dans le modèle neutre afin de déterminer le taux de faux positifs. À partir des corrélations (r, r_v, cor_{PC1} et cor_{PC1v}) calculées dans le modèle simulé sous neutralité, les proportions de faux positifs (FP) de T_r, T_{r_v}, $T_{cor_{PC1}}$ et $T_{cor_{PC1v}}$ sont données par rapport à différents quantiles de rejet calculés sur la distribution théorique de Student $\tau_{(n-2)}$. Les résultats sont présentés dans le **Tableau 4**. Pour mesurer les proportions de faux positifs des statistiques T_r, T_{corPC1}, T_{r_v} et T_{corPC1_v} ainsi que pour l'analyse

de puissance, nous avons sélectionné deux temps ; la génération 140 qui est un point transitoire dans les dynamiques de sélection, et la génération 300, finale. À la génération 140, les statistiques T_r et T_{corPC1} montrent une forte proportion de faux positifs dans les modèles en panmixie et autogamie. Pour une erreur de type I de 1%, T_r et T_{corPC1} présentent entre 55 et 81% de faux positifs tandis qu'à la même génération T_{r_v} et T_{corPC1_v} sont en adéquation avec la distribution de Student et les proportions de faux positifs varient entre 0.1 et 3% (**Tableau 4**). La génération 300 présente des proportions similaires, avec des taux de faux positifs qui varient entre 74 et 91% pour une erreur de type I à 1% pour T_r et T_{corPC1} tandis que T_{r_v} et T_{corPC1_v} ont des proportions de faux positifs allant de 1.1 à 22% pour la même erreur de type I.

Tableau 4 : Proportion de faux positifs (FP) des statistiques T_r, T_{corPC1}, T_{r_v} and T_{corPC1_v} en comparaison avec la distribution théorique de Student utilisée pour tester la significativité du coefficient de corrélation.

Génération	**Mode de reproduction**	**Statistique**	**proportions de FP**		
			10% (1.283)	**5% (1.648)**	**1% (2.334)**
140	Autogamie	T_r	85 %	82 %	74 %
		T_{corPC1}	89 %	86 %	81 %
		T_{r_v}	**13 %**	**8 %**	**3 %**
		T_{corPC1_v}	**13 %**	**7 %**	**3 %**
	Panmixie	T_r	72 %	66 %	55 %
		T_{corPC1}	83 %	78 %	70 %
		T_{r_v}	**2.8 %**	**0.6 %**	**0.2 %**
		T_{corPC1_v}	**2.5 %**	**0.6 %**	**0.1 %**
300	Autogamie	T_r	92 %	91 %	87 %
		T_{corPC1}	95 %	93 %	91 %
		T_{r_v}	**31 %**	**26 %**	**20 %**
		T_{corPC1_v}	**37 %**	**31 %**	**22 %**
	Panmixie	T_r	85 %	81 %	74 %
		T_{corPC1}	93 %	91 %	87 %

T_{r_v}	**4.6 %**	**2.3 %**	**1.2 %**
T_{corPC1_v}	**5.2 %**	**3.1 %**	**1.1 %**

Les proportions de faux positifs correspondent aux proportions de simulations dont les statistiques ont une valeur supérieure ou égale aux quantiles de rejet défini par la distribution $\tau_{(n-2)}$ pour différentes erreurs de type I : 10%, 5%, et 1% (les quantiles sont indiqués entre parenthèses). Dans nos simulations, la taille de l'échantillon n est égale à 500. Comme le signe du coefficient de corrélation n'est pas interprétable, en particulier pour T_{corPC1} et T_{corPC1_v}, ce sont les valeurs absolues de T_r, T_{corPC1}, T_{r_v} et T_{corPC1_v} et de la distribution de Student $\tau_{(n-2)}$ qui ont été utilisées pour calculer les proportions de faux positifs.

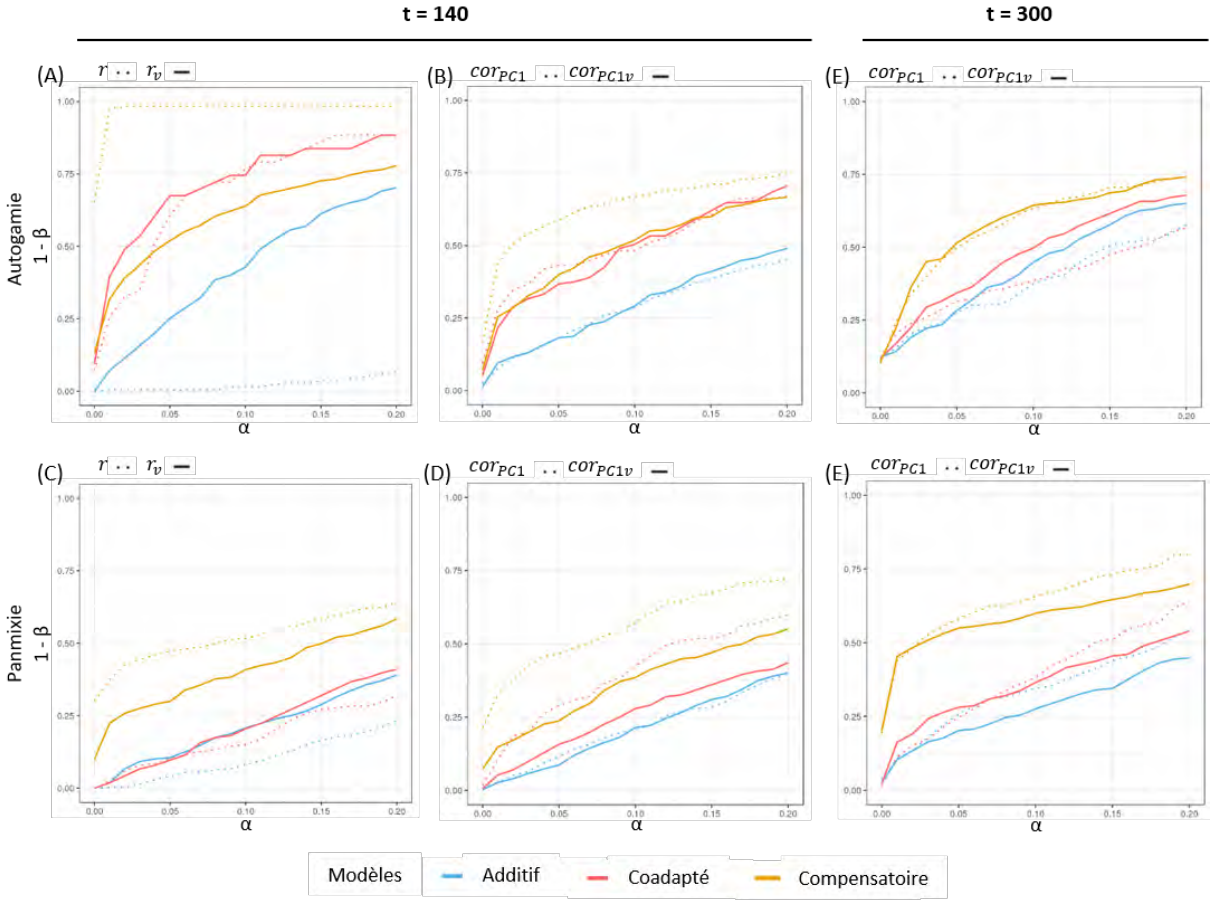
Les figures en **Annexe 5** permettent également de visualiser que les distributions des statistiques T_{r_v} et T_{corPC1_v} (courbes noires pointillées) sont en adéquation avec la distribution théorique de Student $\tau_{(n-2)}$ (courbe rouge) tandis que T_r et T_{corPC1} (courbes noires pleines) ne suivent pas la distribution théorique et montrent une forte variance. Les statistiques T_{r_v} et T_{corPC1_v} dans le modèle en autogamie sont malgré tout moins conservatives à la génération 300 puisqu'elles présentent des taux de faux positifs qui varient autour de 20% mais cela reste bien inférieur aux proportions de FP observées avec les statistiques T_r et T_{corPC1} qui se situent entre 87 et 91%.

Ces résultats montrent qu'il est essentiel d'utiliser la correction pour la structure des populations et l'apparentement entre les individus avec la matrice *V* afin de réaliser correctement les tests statistiques de corrélation pour mesurer le DL entre des paires de SNP ou de fenêtres génomiques indépendantes. Il semble toutefois que lorsque l'on combine les effets de l'apparentement et de la structure, les distributions de T_{r_v} et de T_{corPC1_v} ont tendance à montrer une plus forte variance sur le long terme et il y a une augmentation du taux de FP que l'on observe à la génération 300.

Enfin, nous avons réalisé une analyse de puissance à partir des données simulées sous l'hypothèse nulle de neutralité et d'indépendance et des données simulées sous chacun des modèles de sélection (et d'indépendance). Pour ce faire, nous avons calculé la puissance (1-β) de détection de l'hypothèse H1 (sélection épistatique ou additive) des statistiques de DL en fonction de l'erreur de type I (α) pour H0 sous neutralité. Les simulations réalisées sous neutralité ont été utilisées pour estimer les quantiles de l'erreur de type I α de 0.1% à 20% en utilisant les valeurs absolues de chacune des statistiques dans les deux régimes de reproduction. Enfin, pour chacun des modèles de sélection sous panmixie et autogamie avec les mutations codominantes, nous avons calculé la proportion des simulations dont les valeurs

absolues de r (respectivement r_v, cor_{PC1} et cor_{PC1v}) étaient supérieures aux quantiles. Les puissances de détection ont été calculées pour les statistiques T_r, T_{corPC1}, T_{r_v} et T_{corPC1_v} à la génération 140 où l'ensemble des allèles sélectionnés n'étaient pas encore fixés, ainsi qu'à la génération 300 pour les statistiques T_{corPC1} et T_{corPC1_v}. La **Figure 21** montre tout d'abord que les puissances de détection des statistiques r/r_v et cor_{PC1}/cor_{PC1v} sont globalement plus fortes dans le modèle compensatoire par rapport aux modèles coadapté et additif (respectivement 25-50%, 10-65%, et 10-30%, pour un α=5% avec les statistiques r_v et cor_{PC1v}) notamment en panmixie. En revanche, la correction par la matrice d'apparentement (statistiques r_v et cor_{PC1v}) n'augmente pas les puissances de détection de la sélection épistatique, elle tend plutôt à les réduire surtout pour le modèle compensatoire. En effet, le DL observé dans le modèle compensatoire est probablement lié à la structure car les haplotypes sélectionnés peuvent être différents (*AB* ou *ab*) entre les deux sous-populations, et leur fixation (*AB* ou *ab*) est plus fréquente dans les sous-populations que dans l'ensemble de la population.

Figure 21 : Puissance de détection de la sélection épistatique avec les statistiques de DL basées sur les SNP ou les fenêtres génomiques, par rapport au modèle neutre. Les puissances de détection de la sélection épistatique en autogamie et en panmixie sont calculées à l'échelle des SNP (r et r_v) (**A, C**) et à l'échelle des fenêtres génomiques (cor_{PC1} et cor_{PC1v}) (**B, E, et D, F**). Les figures (**A, B, C, D**) représentent les puissances de détection calculées à la génération 140 et les figures (**E, D**) à la génération 300 (r et r_v ne sont plus calculables à cette génération dans les modèles de sélection coadapté et additif). L'axe des x correspond à l'erreur de type I (α) et l'axe des y correspond à la puissance de détection (1-β). Les mutations sont codominantes.



Par exemple, dans certains cas, l'haplotype *AB* (ou *ab*) est proche de la fixation dans une sous-population, et l'haplotype *ab* dans l'autre sous-population, ou bien le polymorphisme est maintenu dans la seconde sous-population avec *ab* et *AB* qui ségrégent en fréquences intermédiaires. Ces différences entre les sous-populations conduisent à de fortes valeurs de DL lorsque la structure n'est pas prise en compte.

D'autre part, si l'on compare les statistiques r et r_v dans le modèle additif en autogamie, on observe une augmentation de la puissance de détection de r_v par rapport à r. Nous observons la même tendance dans le modèle panmictique mais la différence entre r et r_v est moins forte. Cet effet peut être mis en relation avec les résultats obtenus dans la **Figure 19** (taux de cofixation) qui montre 60% de fixation de l'haplotype *ab* dans le modèle additif tandis que l'haplotype *AB* est très peu fréquent (taux de cofixation nul). Le reste de la

population porte les haplotypes recombinants *aB* ou *Ab*, qui sont aussi sélectionnés, bien que plus faiblement (**Tableau 2**). Ainsi, si l'on se place à l'échelle de la population globale, le DL sera faible si une sous-population porte majoritairement l'haplotype *ab* fortement sélectionné et l'autre sous-population un haplotype recombinant *aB* ou *Ab* aussi sélectionné. En revanche, si l'on se place à l'échelle des sous-populations, le DL sera plus fort car l'autogamie a tendance à limiter la recombinaison étant donné qu'il y a très peu d'individus hétérozygotes et le DL s'étend au-delà des chromosomes car les haplotypes sont transmis en bloc. On observera ainsi un maintien de *ab vs AB* dans une sous-population, et de *aB vs Ab* dans l'autre sous-population. Les différences en termes de puissance entre r et r_v sont moins forte en panmixie qu'en autogamie car la fixation des haplotypes se fait plus facilement en autogamie puisqu'il y a moins de brassage entre les allèles à chaque génération.

Enfin les puissances de cor_{PC1v} (et cor_{PC1}) sont un peu plus faibles que celles obtenues avec r_v (et r) car les statistiques cor_{PC1v} et cor_{PC1} prennent en compte plusieurs SNP situés dans la fenêtre génomique et sont dépendantes de l'hétérogénéité des haplotypes portant les SNP sous sélection. De plus, les puissances obtenues avec cor_{PC1} sont globalement équivalentes aux puissances de cor_{PC1v}.

Pour résumer, le calcul des proportions de faux positifs par rapport à la distribution théorique de Student montre l'intérêt de calculer le DL avec les statistiques r_v et cor_{PC1v} afin de réduire le DL de fond c'est-à-dire le DL inter-locus qui n'est pas lié à la sélection épistatique mais à des facteurs démographiques (structure des populations) et aux modes de reproduction (degré d'apparentement entre les individus). En revanche, la correction par la matrice d'apparentement ne permet pas d'augmenter les puissances de détection.

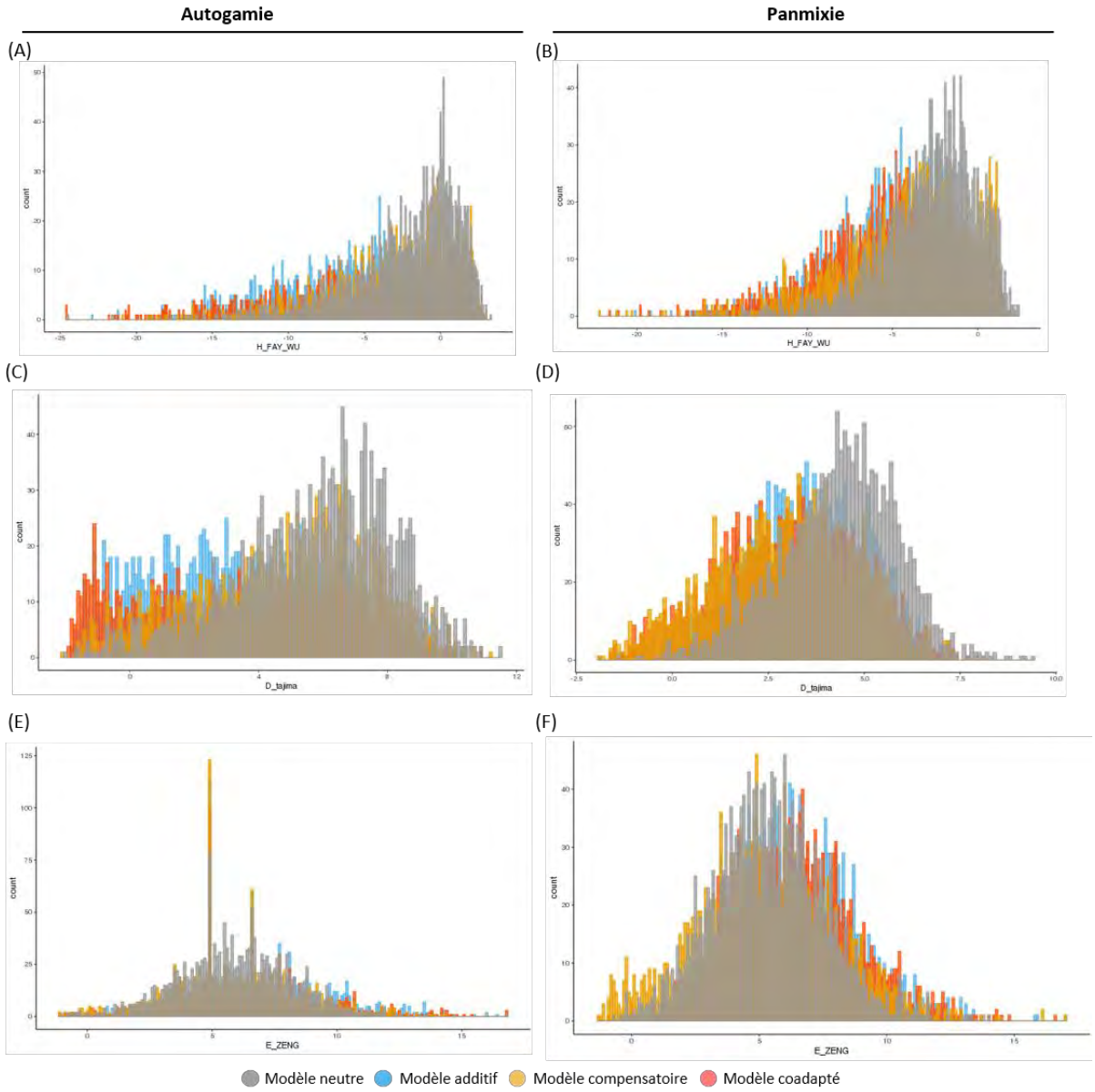
1.5.4 Signatures de sélection sur les locus en épistasie dans les simulations

Afin d'étudier l'évolution du polymorphisme des locus sous sélection épistatique et évaluer si ces locus, pris séparément, présentent des signatures de sélection, nous avons calculé les statistiques de tests de neutralité sur les locus simulés dans les différents modèles de sélection. Les statistiques ont été calculées sur les fenêtres de 10kb des chromosomes simulés sur lesquelles sont également calculées les statistiques de déséquilibre de liaison et

qui portent les SNP soumis à la sélection épistatique. La **Figure 22** présente les distributions des statistiques de neutralité ; ***D*** de Tajima, ***H*** de Fay & Wu et ***E*** de Zeng, pour les quatre modèles de sélection, en régime de reproduction autogame ou panmictique. Les calculs ont été faits à la génération 140, qui est un point transitoire dans la dynamique de fixation où l'ensemble des allèles sélectionnés ne sont pas encore fixés, et où le polymorphisme neutre reste substantiel. Enfin, les statistiques ***D***, ***H*** et ***E*** ont été calculées au sein d'une sous-population simulée de 250 individus afin d'éviter les biais dus à la structure génétique de la population simulée.

La **Figure 22** montre tout d'abord que sous neutralité les distributions des statistiques ***D*** et ***E*** ne sont pas centrées sur 0 et sont fortement décalées vers des valeurs positives (moyenne ***D*** = 5.74 et 4.31, et moyenne ***E*** = 5.94 et 5.36, en autogamie et panmixie). Les distributions de ***H*** sont néanmoins plus proches de 0 en moyenne (moyenne ***H*** = -2.17 et -2.77, respectivement en autogamie et panmixie). Ceci indique une sous-représentation des allèles rares dans chaque sous-population, et les populations simulées n'ont donc pas atteint l'équilibre mutation-dérive attendu pour une population stable évoluant sous neutralité. Ce phénomène est observable au niveau des locus de 10kb analysés sur les chromosomes A et B (**Figure 22**), mais aussi au niveau de locus de même taille sur les chromosomes n'étant pas la cible de la sélection (chromosomes C et D, **Annexe 6**). Ceci permet d'affirmer que le filtre de fréquence allélique que nous avons effectué à la génération 0, uniquement sur les SNP des chromosomes A et B destinés à être sélectionnés (i.e. $0.25 < \text{freq}(a)$ et $\text{freq}(b) < 0.75$) - et pas ceux des chromosomes C et D - n'est pas responsable de ce biais. Les spectres de fréquence calculés à la génération 140, sur la base d'un chromosome neutre en autogamie et panmixie, à l'échelle des deux sous-populations ou dans une sous-population, confirment cette sous-représentation des allèles rares à l'échelle du génome lors des simulations « forward in time » (**Annexe 7**).

Figure 22 : Distribution des statistiques de tests de neutralité calculées sur les locus simulés dans les différents modèles de sélection à la génération 140, dans les modèles de reproduction en autogamie et en panmixie. Les statistiques ***H*** de Fay & Wu, ***D*** de Tajima et ***E*** de Zeng sont calculées sur les fenêtres génomiques de 10kb entourant les SNP simulés sous sélection dans les quatre modèles de sélection; les modèles de sélection épistatiques compensatoire et coadapté, le modèle de sélection additif et le modèle neutre. Les distributions correspondent aux statistiques calculées sur les fenêtres génomiques extraites des chromosomes A et B contenant les deux SNP sous sélection. Les figures (A, B) représentent les distributions de ***H*** de Fay & Wu, les figures (C, D) représentent les distributions du ***D*** de Tajima et les figures (E, F) représentent les distributions du ***E*** de Zeng. (A, C, E) sont extraites du modèle de reproduction en autogamie et (B, D, F) du modèle panmictique. L'ensemble des ces statistiques sont calculées à l'échelle d'une sous-population simulée de 250 individus. L'axe des abscisses correspond aux valeurs de statistiques calculées sur les données simulées et l'axe des ordonnées correspond à la courbe de densité. Les mutations sous sélection sont codominantes.



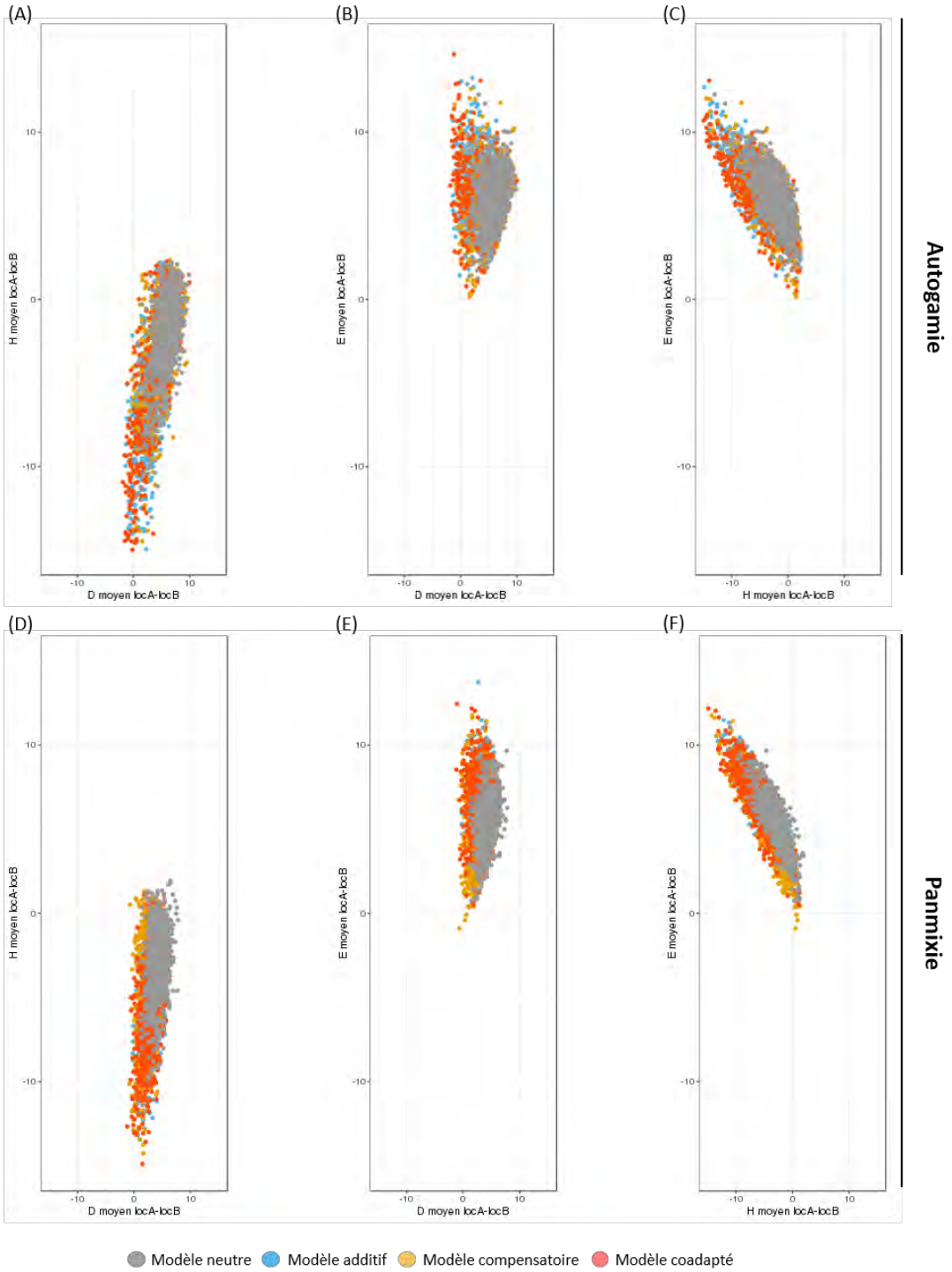
Cette sous-représentation des allèles rares à l'échelle du génome peut s'expliquer par la structuration de la population ancestrale (N=500) en deux sous-populations isolées de taille réduite de moitié (N=250) et donc par un goulot d'étranglement qui a favorisé la perte des allèles rares. Le fait que les sous-populations n'ont pas atteint l'équilibre mutation-dérive au

moment où la sélection opère ne pose pas de problème pour l'analyse, puisque l'objectif est de comparer la distribution du polymorphisme des locus sous sélection à celle de locus sous neutralité, dans le même scénario d'évolution. Cette sous-représentation globale des allèles rares s'observe aussi dans les modèles de sélection. Cela a pour conséquence de décaler les distributions de ***D*** et ***E*** vers des valeurs positives (**Figure 22**) bien que les distributions dans ces modèles présentent une surreprésentation des fréquences alléliques extrêmes en comparaison avec le modèle neutre.

Si l'on compare maintenant les modèles de sélection pour chacune des statistiques, la statistique ***H*** indique des signaux de balayage sélectif avec un décalage des distributions vers des valeurs faibles dans les modèles de sélection épistatique coadapté et additif pour les deux régimes de reproduction. Les signaux de sélection semblent toutefois plus extrêmes en autogamie. Les modèles additif et coadapté montrent des signatures de sélection similaires qu'il n'est pas possible de distinguer. Le modèle compensatoire, en revanche, semble montrer un signal de balayage sélectif surtout en panmixie. Les distributions de la statistique ***D*** calculées dans les différents modèles confirment les signaux de balayage sélectifs avec des distributions décalées vers des valeurs faibles de ***D*** dans les modèles coadapté et additif par rapport au modèle neutre. Là encore, il est difficile de distinguer ces deux modèles et les signaux de balayage sélectifs sont également plus forts en autogamie qu'en panmixie. Le modèle compensatoire montre également un signal de balayage sélectif en panmixie. Enfin, les distributions de la statistique ***E*** calculées dans les différents modèles de sélection montrent aussi des signatures de balayage sélectif dans les modèles additif et coadapté mais le décalage des distributions entre ces deux modèles de sélection et le modèle neutre est plus faible qu'avec les statistiques précédentes. En effet, ***E*** distingue mieux les signaux de sélection après la fixation des allèles, or nous nous situons ici à la génération 140 et l'ensemble des allèles sélectionnés sont en cours de fixation donc ces résultats semblent cohérents. Enfin, de façon plus inattendue, le modèle compensatoire semble indiquer un signal de sélection négative en panmixie avec une distribution décalée vers des valeurs faibles de ***E***. En effet, ***E*** ($E = \frac{\theta_L - \theta_S}{\sqrt{Var(\theta_L - \theta_S)}}$) contraste θ_s qui correspond au nombre de SNP qui ségrègent au locus et θ_L qui correspond au nombre moyen d'allèles dérivés, or θ_s est faible car peu de mutations sont apparues sur 10kb à la génération 140 dans nos simulations. C'est donc θ_L qui est

comparativement plus faible car il y a peu d'allèles dérivés en cours de fixation par un effet de la sélection.

Figure 23 : Distributions conjointes des statistiques de neutralité: $DH_{\overline{AB}}$, $DE_{\overline{AB}}$ et $HE_{\overline{AB}}$ calculées dans les modèles neutre, de sélection épistatique coadapté et compensatoire, et dans le modèle et additif. Les distributions conjointes des statistiques sont obtenues à partir des valeurs moyennes aux deux locus A et B pour chaque simulation et pour chacune des statistiques. Les Figures (A, B et C) représentent les distributions en autogamie et les figures (D,E,F) les distributions en panmixie.



En effet, dans le modèle compensatoire, il y a un maintien du polymorphisme mais lorsque les allèles ou les haplotypes (*AB* ou *ab*) se fixent, on observe plus souvent une fixation de l'haplotype *AB* ancestral que de l'haplotype *ab* mutant (voir les taux de cofixation, **Figure 19**).

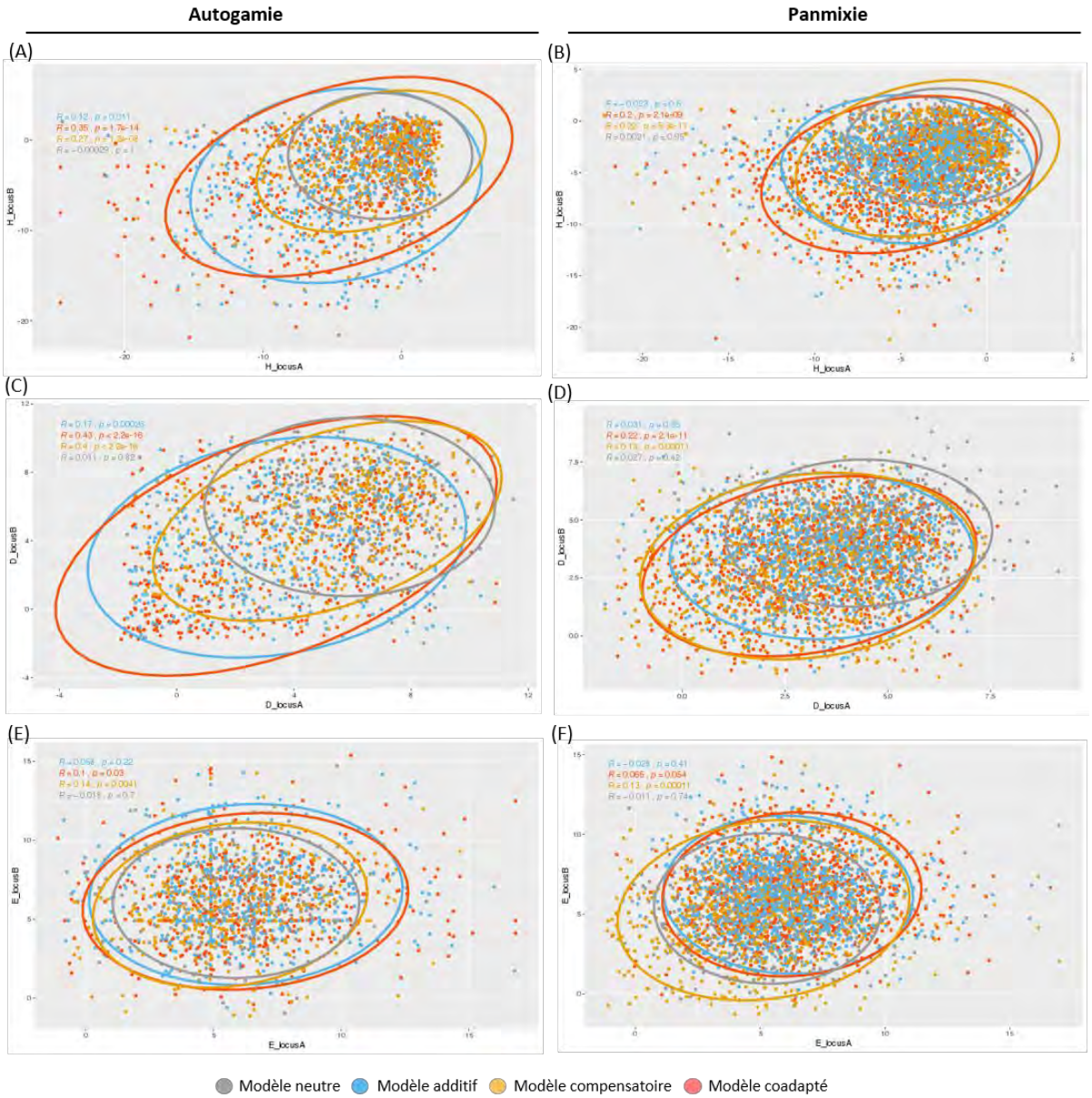
La **Figure 23** présente les distributions conjointes $DH_{\overline{AB}}$, $DE_{\overline{AB}}$ et $HE_{\overline{AB}}$ des statistiques de neutralité à la génération 140. Cette représentation des résultats permet de conjuguer les différentes sensibilités de ces statistiques. La valeur moyenne entre les deux locus A et B est calculée pour chaque statistique et pour chaque simulation, puis les distributions conjointes sont représentées. Les distributions dans les modèles de sélection se distinguent assez bien des distributions dans le modèle neutre mais les différents modèles de sélection sont difficilement distinguables, notamment les modèles de sélection épistatique coadapté et additif. En panmixie, le modèle compensatoire semble se démarquer des autres modèles de sélection avec notamment les statistiques ***H*** et ***E*** qui montrent des signaux similaires à la **Figure 22**. Dans les modèles coadapté et additif, nous faisons les mêmes observations : les statistiques semblent montrer des signaux de balayage sélectif avec des valeurs de ***H*** négatives, et des valeurs de ***E*** positives. La distinction entre les modèles semble, en revanche, moins forte avec le ***D*** de Tajima sous cette représentation même si les modèles de sélection semblent montrer un léger signal de balayage sélectif. Ces résultats montrent l'intérêt de connaître l'état ancestral/dérivé des allèles aux locus, dont tiennent compte les statistiques ***H*** et ***E***, pour ce type d'analyse. La figure en **Annexe 8** représente les distributions conjointes de ces mêmes statistiques mais calculées entre des locus extraits des chromosomes C et D non ciblés par la sélection. Les modèles de sélection ne se distinguent pas du modèle neutre car les statistiques sont réalisées sur les chromosomes neutres quels que soient les modèles de sélection. Dans le modèle en autogamie, on observe malgré tout, une légère différence entre le modèle neutre (gris) et les modèles de sélection par rapport au mode de reproduction panmictique. En effet, le régime d'autogamie favorisant la transmission en bloc de plusieurs chromosomes, les chromosomes non ciblés par la sélection pourront être plus souvent transmis avec les chromosomes ciblés par la sélection, que sous un régime de panmixie.

Ainsi, les statistiques présentées permettent de détecter des signatures de sélection et essentiellement des signatures de balayage sélectif avec ***D*** et ***H*** et également avec ***E*** de façon plus contrastée. En effet, nous nous situons à la génération 140 qui est une phase transitoire de la sélection épistatique alors que ***E*** détecte plus efficacement des signatures de

balayage sélectif après fixation. D'après ces résultats, les modèles coadapté et additif ne sont pas différenciables avec ces statistiques, ils présentent tous deux des signatures de sélection positive liées à la sélection des allèles dérivés *a* et *b* dans les deux modèles. Les signatures de sélection sont aussi plus fortes en autogamie car ce régime de reproduction maintient plus facilement les génotypes *aa/bb* par rapport au modèle panmictique qui présente plus d'individus hétérozygotes. Dans le modèle compensatoire, les résultats sont plus contrastés. En effet, beaucoup de simulations présentent un maintien du polymorphisme aux deux SNP sous sélection, mais la cofixation des haplotypes (*AB* ou *ab*) est observée dans 40% des cas en autogamie et dans 65% des cas en panmixie, et parmi celles-ci, c'est plus souvent *AB* qui est fixé (voir les taux de cofixation, **Figure 19**). Ainsi, la statistique ***H*** qui est plus faible dans le modèle compensatoire par rapport au modèle neutre, surtout en panmixie, montre qu'il y a des cas où les allèles *a* et *b* tendent à se fixer (θ_L correspond au nombre moyen d'allèles dérivés accumulés). Les valeurs de ***D*** faibles dans ce modèle par rapport à la neutralité montrent les cas où il y a une tendance à la fixation de *A* et *B*, ou *a* et *b.* Enfin, les valeurs de ***E*** plus faibles dans le modèle compensatoire, montrent qu'il y a fixation des allèles ancestraux *A* et *B* avec un θ_L faible.

Afin de voir s'il existe une corrélation des patrons de polymorphisme aux locus A et B sous sélection épistatique dans les simulations, nous avons représenté les distributions conjointes pour les paires de locus A et B pour D, H, et E: D_{AB}, H_{AB} et E_{AB}. La **Figure 24** montre globalement que les corrélations des signatures de sélection entre les paires de locus A et B issues des simulations sont faibles. Dans le modèle neutre, les corrélations entre les paires de locus varient entre -0.018 et 0.027 selon les statistiques, et ne sont pas significatives. Les statistiques ***H*** et ***D*** indiquent des corrélations faibles mais significatives entre les paires de locus dans le modèle coadapté ($0.20 < r < 0.43$) et le modèle compensatoire ($0.13 < r < 0.40$), avec un signal plus important que dans le modèle additif ($-0.023 < r < 0.17$, non significatif). Ces valeurs de corrélation sont plus élevées en autogamie qu'en panmixie, indiquant une cofixation des allèles *a* et *b* plus efficace. Avec la statistique ***E***, le modèle compensatoire se démarque ($r = 0.14$ et p-valeur = 0.0041 en autogamie, $r = 0.13$ et p-valeur = 0.00011 en panmixie), par rapport aux autres modèles de sélection qui ne présentent pas de corrélation significative, pouvant être le reflet de la tendance à la cofixation des allèles ancestraux *A* et *B*, comme indiqué plus haut.

Figure 24 : Distributions conjointes des statistiques de neutralité calculées entre les paires de locus simulés en autogamie et panmixie. Les distributions conjointes des statistiques ***H*** de Fay & Wu, ***D*** de Tajima et ***E*** de Zeng entre les paires de locus A et B simulés dans les différents modèles de sélection sont représentées. Les figures (A, B) représentent les distributions conjointes de $H_{\overline{AB}}$ en autogamie et panmixie. Les figures (C, D) représentent les distributions conjointes de $D_{\overline{AB}}$ en autogamie et panmixie. Les figures (E, F) représentent les distributions conjointes de $E_{\overline{AB}}$ en autogamie et panmixie. L'axe des abscisses correspond aux valeurs de statistiques calculées sur les locus A et l'axe des ordonnées correspond aux valeurs de statistiques calculées sur les locus B. Les ellipses sont calculées avec la fonction « *stat_ellipse*()» disponible dans la bibliothèque « *ggplot* » de R qui permet de calculer une ellipse des données et suppose la distribution *t* multivariée.



Globalement, bien que des corrélations existent entre les signatures de sélection identifiées indépendamment, les simulations indiquent qu'elles ne permettent pas d'identifier des « cosignatures de sélection » claires. En effet, les signatures de sélection seront fortement dépendantes des fréquences initiales des allèles aux locus soumis à la sélection épistatique.

De plus, les graphiques de corrélations ainsi que les distributions univariées et conjointes des statistiques de neutralité suggèrent que des locus en apparence neutres peuvent être soumis à la sélection épistatique. Seule la corrélation génétique, analysée par le déséquilibre de liaison, peut permettre d'identifier des signatures de sélection épistatique de manière claire. Les signatures de sélection identifiées indépendamment sur chaque locus peuvent toutefois permettre d'évaluer comment les différents régimes de sélection épistatique peuvent influencer le polymorphisme des gènes.

Chapitre 2 : Détection de gènes sous sélection épistatique

Ce chapitre porte sur l'identification de signatures de sélection épistatique à l'échelle du génome chez la légumineuse modèle *Medicago truncatula* à l'aide de données SNP et via l'approche statistique présentée dans le premier chapitre. Une analyse portant sur des données génétiques humaines est aussi présentée afin d'illustrer les possibilités d'application de l'approche à divers organismes. L'objectif est l'identification de nouvelles interactions adaptatives entre gènes, ou coadaptation, pour des gènes aux fonctions biologiques et moléculaires connues, non étudiées précédemment, ou inconnues. Deux types d'approches ont été mises en place, la première englobant la seconde : une approche « genome-wide » pour laquelle tous les gènes ont été testés et une approche par gènes candidats plus ciblée sur des fonctions biologiques et moléculaires d'intérêt. Les gènes qui ont été particulièrement étudiés chez *Medicago truncatula* sont principalement impliqués dans les interactions plantes-micro-organismes (IPM). Leurs fonctions sont associées à la capacité de réponse de la plante en présence de micro-organismes symbiotiques tels que les bactéries Rhizobia, fixatrices de l'azote atmosphérique lors de la nodulation, ainsi que les champignons endomycorhiziens à arbuscule (« Arbuscular Mycorrhizal Fungi ») lors de la mycorhization. Ces deux symbioses participent à l'amélioration de la nutrition minérale des plantes. La symbiose fixatrice d'azote est restreinte aux clades des Fabales (dont fait partie *Medicago truncatula*), Fagales, Cucurbitales et Rosales (Griesmann et al., 2018) et la symbiose mycorhizienne est très largement répandue au sein des clades de plantes terrestres (Parniske, 2008; Radhakrishnan et al., 2020). Les statistiques de DL qui ont été développées et testées sur les données simulées présentées dans le chapitre précédent, ont été utilisées pour rechercher des signatures de sélection épistatique de manière approfondie chez *M. truncatula* et de manière plus ciblée chez l'homme du fait de nos connaissances restreintes. Les mesures de DL, r/r_v et cor_{PC1}/cor_{PC1v}, basées respectivement sur les SNP et sur des fenêtres génomiques contenant plusieurs SNP, ont été utilisées pour réaliser des Genome-Wide Epistatic Selection Scans (« GWESS ») à l'aide de gènes candidats considérés comme appâts. La première partie de ce chapitre présente les espèces et les données analysées. Dans une seconde partie, nous décrivons l'approche utilisée pour rechercher des gènes ou régions génomiques potentiellement sous sélection épistatique avec un gène appât, ainsi que les résultats obtenus avec quelques gènes candidats. Au cours de ma thèse, nous avons pu tester quelques dizaines de gènes appâts candidats chez l'humain ainsi que 98 gènes chez *M. truncatula*, mais tous les résultats ne sont pas présentés, seulement ceux de 3 gènes chez *Medicago truncatula* et ceux

de 2 gènes chez l'humain. Parmi l'ensemble des gènes appâts qui ont été testés chez *Medicago truncatula*, le gène *MtSUNN* a montré un signal de coadaptation jusqu'alors inconnu avec le gène *MtCLE02* (codant pour un peptide du type CLAVATA-like) et des analyses fonctionnelles effectuées en collaboration avec l'IPS2 (Université Paris-Saclay) ont pu mettre en évidence un rôle négatif de *MtCLE02*, dépendant de *MtSUNN*, sur la nodulation. Sur les données humaines, l'analyse GWESS a permis d'identifier un signal de coadaptation entre les gènes *SLC24A5* et *EDAR*. Ce résultat suggère une interaction génétique adaptative, ou un phénomène de cosélection très marqué chez certaines populations humaines entre la pigmentation de la peau et la voie de l'ectodysplasine, impliquée dans le développement des organes ectodermiques (poils, dents, glandes sudoripares). Dans une troisième partie, des statistiques de tests de neutralité ont été calculées sur les données génétiques de *M. truncatula* afin de décrire le polymorphisme à l'échelle du génome et des populations. Le but est également d'évaluer dans quelle mesure les gènes identifiés potentiellement sous sélection épistatique présentent également des signatures de sélection indépendantes. Enfin, la dernière partie est consacrée à l'analyse « genome-wide » des signatures de sélection épistatique chez *Medicago truncatula.* Les statistiques de DL, cor_{PC1}/cor_{PC1v} ont été calculées à l'échelle du génome entre toutes les paires de gènes, et cette ressource représente une masse de données très importante de $(48\,333x48\,332)/2 \; = 1.1x10^{9}$ (pour 48 333 gène) valeurs de DL dans chacune des populations. Dans une démarche exploratoire du traitement analytique de la masse de résultats ainsi générée deux approches ont été mises en place. La première approche se base sur l'analyse de sets de gènes candidats, nous comparons les patrons de DL entre des gènes de même voie biologique ou de même fonction moléculaire au DL de sets de gènes échantillonnés aléatoirement. La seconde approche est plus systémique, nous faisons une analyse descriptive des réseaux d'interactions génétiques. Les réseaux intègrent toutes les corrélations significatives entre paires de gènes qu'elles soient intrachromosomique et interchromosomique ou uniquement interchromosomique. Nous avons également analysé des sous-réseaux incluant des gènes caractérisés de la symbiose racinaire rhizobienne et dans la symbiose racinaire mycorhizienne. L'analyse de ces résultats est une partie toujours exploratoire, l'objectif serait d'identifier de nouvelles interactions adaptatives entre de nouveaux gènes dans le but de les caractériser.

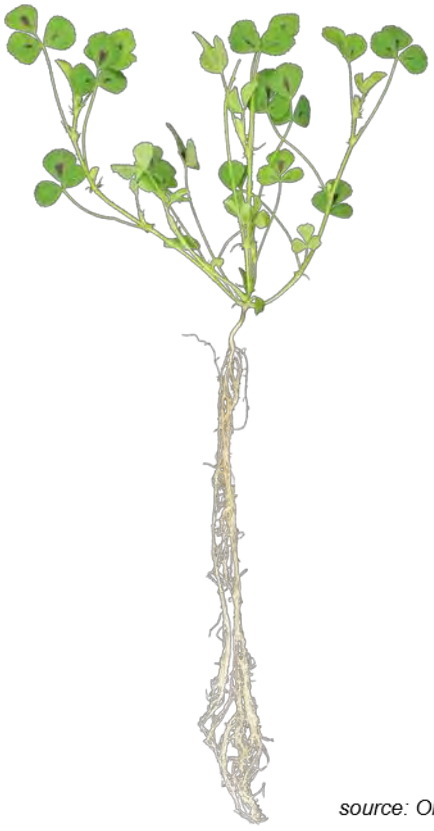
2.1 Présentation des données

Les analyses de GWESS ont été réalisées chez *Medicago truncatula* et chez l'humain à partir de données de SNP. Dans cette première partie, nous présentons les données SNP qui ont été utilisées pour chacun des organismes ; le nombre de SNP, les populations analysées ainsi que leur structuration génétique. Nous présentons également succinctement les espèces, leurs histoires démographiques et les travaux qui ont été publiés à partir des données utilisées.

2.1.1 Description des données de *Medicago truncatula*

Medicago truncatula est une plante diploïde principalement autoféconde (**Figure 25**). Elle sert de modèle principal pour l'étude de la génétique et de l'évolution de la symbiose rhizobienne avec les bactéries fixatrices d'azote et de la symbiose mycorhizienne que l'on retrouve chez la plupart des plantes terrestres mais pas chez *Arabidopsis thaliana*, la plante modèle la plus étudiée.

Figure 25 : La légumineuse modèle *Medicago truncatula*



source: Olivier André

M. truncatula possède un génome relativement petit d'environ 430 millions de paires de bases (Pecrix et al., 2018) et le temps de génération de graine à graine est assez court (environs 4

mois) (Kang et al., 2016). Il existe également une grande diversité d'écotypes issus d'environnements variables et qui présentent des phénotypes divers (Ronfort et al., 2006). Enfin, une large collection de mutants est disponible pour l'analyse génomique (Tadege et al., 2008) ainsi que plusieurs millions de marqueurs génétiques répartis sur les 8 chromosomes et identifiés sur une large collection de 262 accessions.

Les données SNP utilisées pour ce travail ont été générées par le projet international de séquençage haut débit d'accessions de *Medicago truncatula* : le Medicago HapMap Project (http://www.medicagohapmap2.org/). Ces données sont disponibles en téléchargement libre (http://www.medicagohapmap.org/downloads/mt40). La séquence du génome de Medicago (génome de la lignée A17 Jemalong), publiée en 2011 (Young et al., 2011) et mise à jour en 2014 et 2018 (Pecrix et al., 2018; Tang et al., 2014) a été utilisée comme génome de référence pour le mapping et la cartographie des SNP à l'échelle du génome de *M. truncatula*. C'est précisément la version Mt4.0 du génome qui a été utilisée pour le mapping des SNP par le projet HapMap et c'est sur ce jeu de données que nous basons nos analyses. Les SNP qui ont passé le filtre de contrôle qualité classique de validation possèdent un haut degré de confiance pour leur localisation physique (régions intergéniques ou intragéniques, exon, intron, 5' ou 3' UTR) ainsi que pour leurs effets, c'est-à-dire les mutations synonymes ou non synonymes, codon stop (http://www.medicagohapmap.org/downloads/Mt40/Mt4.0_HapMap_README.pdf). 262 accessions ont été séquencées produisant ainsi plusieurs millions de variants génétiques répartis sur l'ensemble du génome. Sur la base de la version Mt4.0 du génome de *M. truncatula*, 22 079 496 SNP ont été identifiés sur les 8 chromosomes (2 775 109, 2 452 016, 3 278 883, 3 091 225, 2 688 969, 2 385 157, 2 831 334, 2 576 803 sur les chromosomes 1 à 8 respectivement). De plus, 49 accessions du genre *Medicago* ont également été séquencées afin de constituer un jeu de données externe pour les analyses évolutives. Ce jeu de données a notamment permis l'identification de l'état ancestral/dérivé des allèles pour 1 283 721 SNPs chez *M. truncatula* (voir (Bonhomme et al., 2015)). Les données SNP réparties sur les 8 chromosomes de *M. truncatula* ont été préalablement imputées avec le logiciel TASSEL (Bradbury et al., 2007). Les bases manquantes ont été remplacées par le nucléotide de l'accession qui partage l'haplotype le plus long entourant la position contenant le nucléotide manquant (Bonhomme et al., 2014).

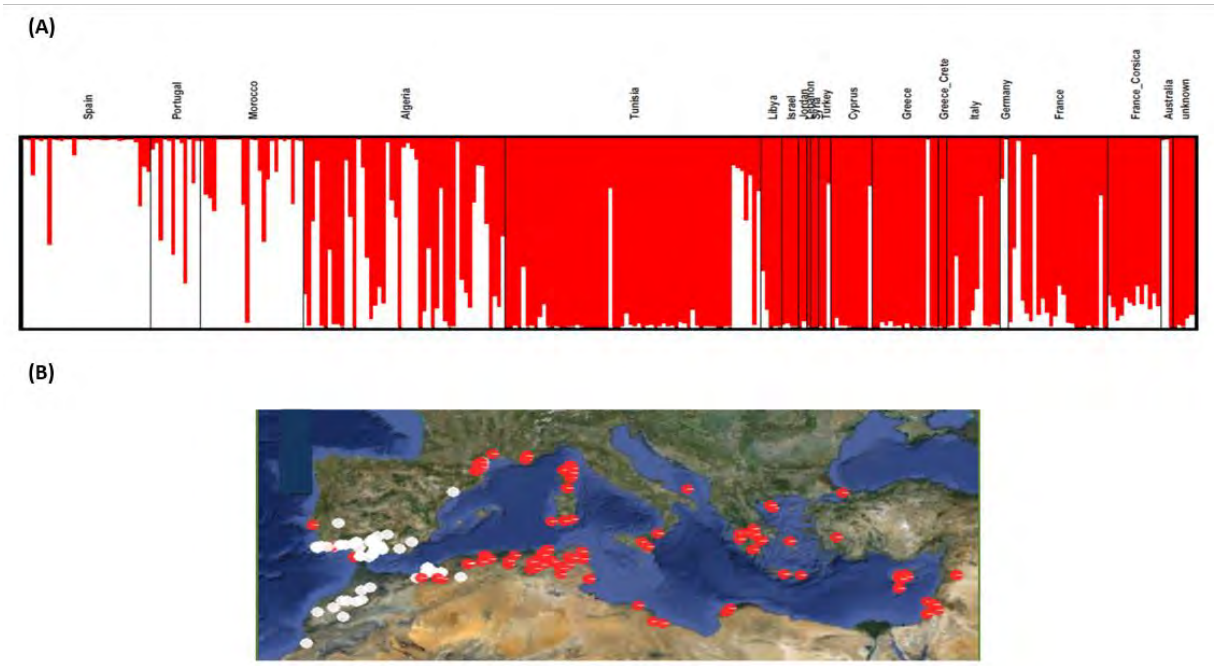
2.1.1.1 Histoire démographique et structure des populations chez M. truncatula

L'étude de l'histoire démographique et de la structure des populations sont des étapes nécessaires si l'on souhaite rechercher des signatures de sélection à l'échelle du génome ou étudier les bases génétiques de l'adaptation car ces forces démographiques peuvent conduire à des estimations biaisées et augmenter le taux de faux positifs. Les premières études de génétique des populations réalisées à l'échelle du génome sur les données de séquençage haut débit ont validé les conclusions apportées par des études plus anciennes montrant que les populations de *Medicago truncatula* ont subi une forte expansion démographique (Branca et al., 2011; De Mita et al., 2007a, 2011). Sur les données génétiques, une telle expansion démographique se traduit par un excès de variants rares dans le génome autrement dit de SNP dont les allèles dérivés sont en faibles fréquences (Bonhomme & Jacquet, 2020). La connaissance de cet aspect de l'histoire démographique globale de l'espèce est importante car cet excès de variants rares peut être interprété, à tort, comme un signal de sélection purifiante ou comme de multiples signaux de balayage sélectif à l'échelle du génome. Une première étude sur la recherche de signatures de sélection à l'échelle du génome de *Medicago truncatula* a été réalisée avec notamment la statistique d_N/d_S (Paape et al., 2013) ; elle a montré que 50 à 70% des mutations non-synonymes sont sous sélection purifiante (i.e. la proportion de mutations non-synonymes est plus importante que la proportion de mutations synonymes pour les allèles dérivés rares dans les séquences codantes), indiquant qu'il est possible de détecter de tels signaux de sélection malgré les effets de l'expansion démographique. Dans cette même étude, les auteurs ont également montré que 1% des gènes analysés sont sous sélection positive. Plus récemment, Bonhomme et al., ((Bonhomme et al., 2015) ont recherché des signatures de balayage sélectif à l'échelle du génome montrant également une faible proportion de gènes sous sélection positive. En revanche, la plupart des régions identifiées ne sont pas communes aux deux études et cela montre l'importance de l'échantillonnage ainsi que des tests statistiques choisis car ils sont sensibles à différentes dynamiques et intensités de la sélection.

Dans le but de rechercher des signatures de sélection épistatique à l'échelle du génome, il est nécessaire de connaître également la structure génétique des populations. En effet, comme cela a été montré dans le chapitre précédent, la structure des populations peut conduire à une estimation biaisée du DL et une augmentation du taux de faux positifs si elle

n'est pas prise en compte. Chez *Medicago truncatula*, plusieurs études ont clairement mis en évidence une structure en deux groupes génétiques de la core-collection (Bonhomme et al., 2015; Burgarella et al., 2016; De Mita et al., 2011; Kang et al., 2016; Paape et al., 2013; Ronfort et al., 2006; Stanton-Geddes et al., 2013) (**Figure 26**) (La core-collection correspond à un ensemble d'accessions génotypées qui visent à représenter la diversité génétique de l'espèce)**.**

Figure 26 : Structure génétique de *Medicago truncatula* répartie en deux principaux groupes (Bonhomme et al. 2014). (A) barplot représentant les probabilités d'appartenance aux groupes génétiques Far West (FW) et Circum (C) pour 288 accessions de *M.truncatula*. (B) Distribution géographique des deux groupes génétiques. Les population FW et C sont représentées en blanc et rouge respectivement.



La sous-population Far-West (FW, représentée en blanc) comprend 80 accessions échantillonnées principalement à l'ouest du bassin méditerranéen (Espagne, Portugal, Maroc et Ouest de l'Algérie) et la sous-population Circum (C, représentée en rouge) comprend 186 accessions échantillonnées sur tout le pourtour méditerranéen avec seulement quelques accessions situées à l'ouest. Quelques accessions sont mixtes et sont considérées dans les deux groupes génétiques.

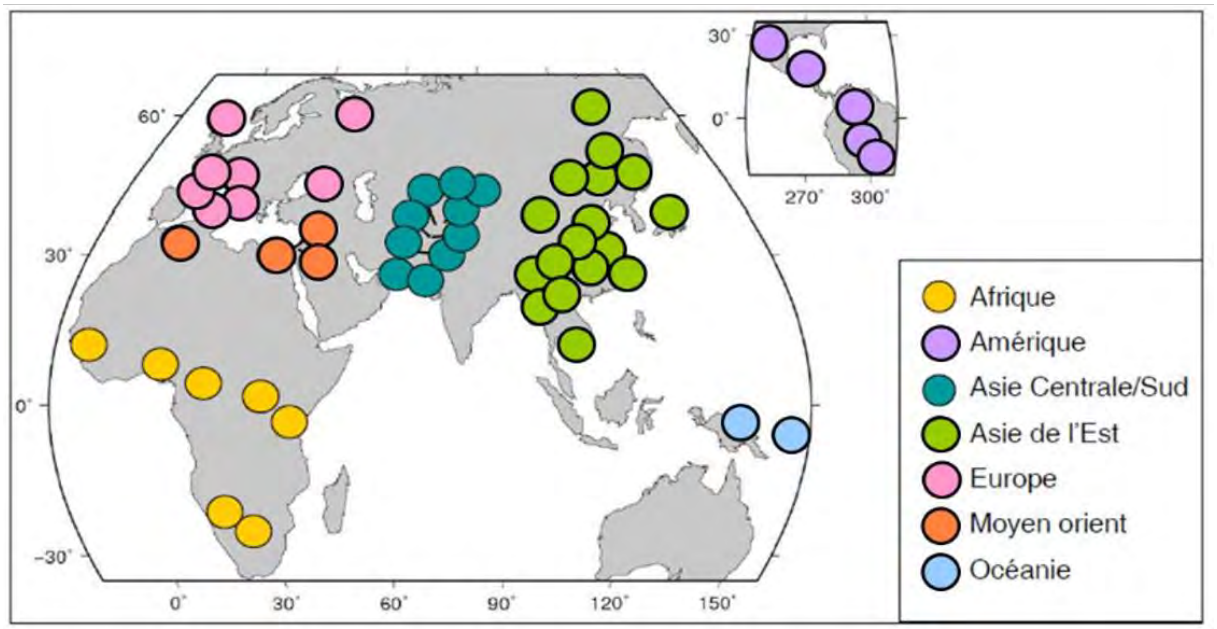
De plus, Branca et al., (Branca et al., 2011) ont étudié les profils de DL à l'échelle du génome chez *M. truncatula* sur la base de 26 accessions avec une couverture 15X à l'échelle du génome. Ils ont pu montrer que le DL moyen entre deux SNP adjacents diminue de moitié lorsque la distance entre ces SNP est de 3kb (i.e. « LD-decay ») et si la distance est de 5kb, le

DL diminue en moyenne de deux-tiers. Ces résultats sont malgré tout très variables selon les régions du génome (Branca et al., 2011), et le LD-decay varie globalement entre 1kb à 10kb. Plus récemment, Bonhomme et al., ((Bonhomme et al., 2015) ont obtenus des résultats similaires en se basant sur les données de SNP à l'échelle du génome et sur l'ensemble de la collection de *M. truncatula*. L'information de LD-decay est importante pour l'analyse de GWAS car cela influence la résolution des pics d'association et donc des possibles gènes candidats sous-jacents à ces pics. Pour ce travail, le LD-decay peut également être une information importante lorsque nous mesurons le DL entre un gène appât et les autres gènes du génome et que nous obtenons des pics de DL aux niveaux inter- ou intrachromosomiques.

2.1.2 Description des données humaines

Pour rechercher des signatures de sélection épistatique chez l'humain, nous avons utilisé les données du HGDP-CEPH (Human Genome Diversity Panel - Centre d'Etude du Polymorphisme Humain) constituées de 644 257 SNP pour 940 individus appartenant à 52 populations issues de 23 pays répartis sur 7 régions: l'Amérique, l'Asie de l'Est, L'Asie centrale Sud, l'Europe, le Moyen-Orient, l'Afrique subsaharienne et l'Océanie (Cann et al., 2002; J. Z. Li et al., 2008) (**Figure 27**). Le génotypage des 940 individus a été réalisé avec une puce Illumina (HumanHap 650K) (J. Z. Li et al., 2008). La version du génome B36 (c'est-à-dire les positions des gènes) a été utilisée pour ce travail afin de correspondre aux positions de SNP du jeu de données HGDP-CEPH. Ces données de polymorphisme représentent une large collection d'ADN humain et elles ont été largement utilisées en génétique des populations pour étudier la diversité génétique, la structure génétique et l'histoire évolutive des populations humaines (Cann et al., 2002; J. Z. Li et al., 2008; N. A. Rosenberg et al., 2002, 2005). Le détail des populations est donné dans le tableau en **Annexe 9**.

Figure 27 : Répartition géographique des populations du « Human Genome Diversity panel ».



2.2 Approche GWESS avec une méthode appât

Pour rechercher des signatures génomiques de sélection épistatique, nous avons mis en place une approche par GWESS (« genome-wide epistatic selection scan ») à une dimension, qui consiste à mesurer le DL entre un gène candidat et tous les autres gènes du génome, chez *Medicago truncatula* ou chez l'humain. Les statistiques de DL corrigées, cor_{PC1v} et r_v ont été utilisées respectivement sur les données de *M. truncatula* et humaine. Pour chaque comparaison de paires de gènes nous calculons le DL et nous faisons un test statistique de nullité du coefficient de corrélation afin d'obtenir une p-valeur. Plusieurs gènes candidats ont été analysés chez *M. truncatula* et chez l'humain, les résultats sont présentés dans cette partie.

2.2.1 Principe – méthode de l'approche appât

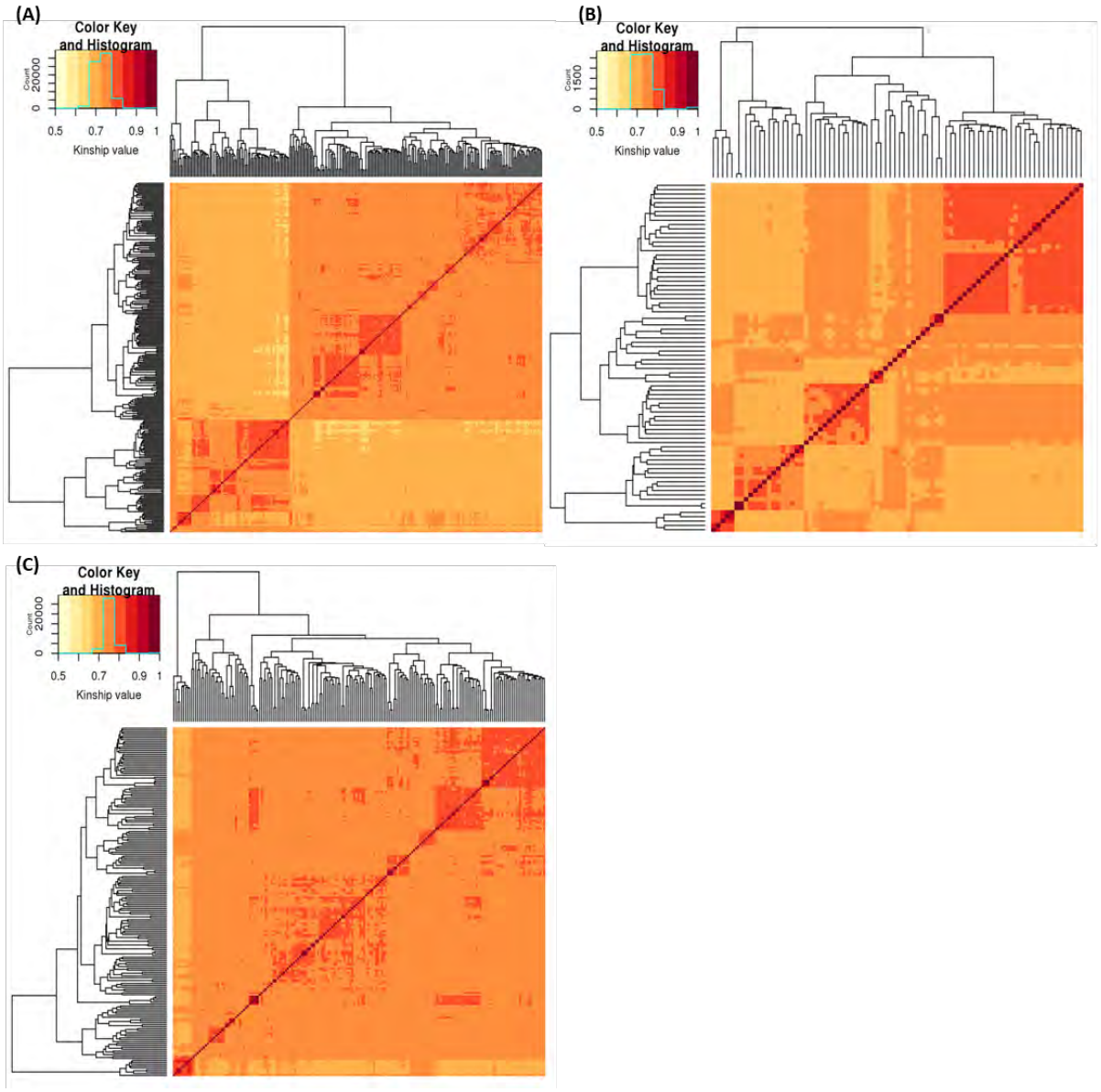
2.2.1.1 Méthode de calcul chez Medicago truncatula

Une approche appât a été mise en place pour rechercher des signatures de sélection épistatique à l'aide du DL entre un gène candidat et tous les autres gènes du génome de *M. truncatula* et l'humain. Chez *Medicago truncatula*, le DL est calculé avec les statistiques cor_{PC1} et cor_{PC1v}. Nous avons réalisé des ACP sur les données génétiques de chacun des 48 333 gènes. L'ACP est faite sur une fenêtre de 10kb centrée sur chaque séquence génique (du codon d'initiation au codon stop, comprenant exons et introns) afin d'augmenter le nombre de SNP notamment pour les gènes de petite taille et ainsi réduire le biais d'échantillonnage. Les

vecteurs de PC1 issus des ACP résument les génotypes multi-SNP à chaque gène. De plus, comme cela a été dit dans le chapitre précédent, la statistique cor_{PC1v} correspond au coefficient de corrélation pondéré par l'inverse de la racine carré de la matrice d'apparentement. La matrice est calculée à partir des 262 accessions en population totale (respectivement 80 et 182 en population FW et C) sur l'ensemble des données SNP du génome et avec une « minor allele frequency » (MAF) supérieure à 5% (**Figure 28**). La matrice V est construite en mesurant la similarité entre chaque paire d'individus *ij* comme sur les données simulées (voir chapitre 1.2.2.3) et correspond à la proportion d'allèles identiques partagés entre les individus. *M. truncatula* est une espèce diploïde, les SNP sont bi-alléliques et les génotypes sont codés 0, 1 ou 2 (dose d'allèle) en fonction du génotype diploïde. La valeur entre deux individus correspond à la somme des valeurs obtenues en comparant chaque SNP, divisé par le nombre total de copies. La **Figure 28** montre les heatmap des trois matrices d'apparentement calculées chez *M. truncatula* dans la population entière (**Figure 28A**) ainsi que dans les deux sous-populations FW et C (**Figure 28B,C**). Les heatmap montrent les matrices d'apparentement réordonnées par un clustering hiérarchique à partir des individus de chaque population. Elles montrent l'apparentement entre les individus et dans la population entière on observe bien les deux groupes génétiques qui correspondent aux deux sous-populations FW et C. Les heatmap qui représentent les populations FW et C montrent aussi des sous-structurations génétiques locales.

Ensuite, le DL est calculé avec les statistiques cor_{PC1} et cor_{PC1v} qui correspondent respectivement aux coefficients de corrélations calculés entre deux vecteurs PC1 ($cor_{PC1} = cor(\ PC1^{l},\ PC1^{m})$) et entre deux vecteurs PC1 multipliés par l'inverse de la racine carré de la matrice d'apparentement; $PC1_V$ ($cor_{PC1v} = cor\left(V^{-1/2} * PC1^{l}, V^{-1/2} * PC1^{m}\right)$ (Voir **Figure 11** – méthode de calcul de cor_{PC1} et cor_{PC1v}). De cette façon, le DL est calculé entre toutes les paires de gènes chez *M. truncatula* en mesurant le coefficient de corrélation ce qui représente un total de $1.1x10^{9}$ comparaisons $(48\,333x48\,332)/2$. Nous effectuons également un test de corrélation (statistiques $T_{cor_{PC1}}$ et $T_{cor_{PC1v}}$) pour mesurer la significativité de chaque valeur de corrélation (cor_{PC1} et cor_{PC1v}) et obtenir une p-valeur. Enfin, nous appliquons une correction de Bonferroni sur les p-valeur au seuil de 10^{-6} lorsque nous comparons les gènes avec l'approche appât (i.e. correction de Bonferroni, en « une dimension », pour α = 5% et 48 333 tests).

Figure 28 : Représentation de la matrice kinship – V calculée chez *Medicago truncatula* sur les 262 individus en population totale (A), dans la population FW composée de 80 individus échantillonnés (B) et dans la population C composée de 182 individus échantillonnés (C). Les heatmap représentent le regroupement des individus par clustering hiérarchique sur la base des matrices kinship calculées à partir des données génétiques de *M. truncatula* (la similarité varie entre 0 et 1). Les distributions de l'apparentement entre tous les individus sont indiquées en haut à gauche de chaque panel. Le panel (A) montre très clairement la structuration en deux sous-groupes génétiques principaux chez *M. truncatula*.



2.2.1.2 *Méthode de calcul chez l'humain*

Pour rechercher des signatures de sélection épistatique chez l'humain, nous avons utilisé les données HGDP-CEPH constituées de 644 257 SNP (431 951 SNP après une MAF à 5%) répartis sur 22 chromosomes et 940 individus répartis en 57 populations de 7 régions du monde. L'approche appât sur les données humaines a été réalisée avec les statistiques r et r_v entre paires de SNP. En effet, les données SNP sont moins denses dans ce jeu de donnée, par rapport aux données SNP de *M. truncatula* et on compte en moyenne un SNP pour 5kb

(3.2x10^9/644 257) contre un SNP pour 30 (6.5x10^8/22 079 496) paires de base chez *Medicago truncatula*. De ce fait, on retrouve en moyenne 3 SNP par gène humain, distants d'environs 5kb. Ainsi, nous ne pouvons pas résumer les génotypes multi-SNP aux différents gènes avec la méthode des ACP entre des SNP trop distants et qui présentent donc un taux de recombinaison plus important. L'approche appât a été réalisée avec les statistiques r et r_v et le SNP appât correspond au SNP le(s) plus proches du gène.

Contrairement à l'approche par fenêtres génomiques ou gènes, nous n'avons pas calculé la corrélation entre toutes les paires de SNP car cela représente un temps de calcul et une masse de données générées trop importants. Le DL est calculé uniquement entre les SNP de gènes candidats qui ont été choisis, versus tous les SNP du génome. Les statistiques r et r_v sont donc calculées à l'échelle des SNP, r correspondant au coefficient de corrélation entre deux vecteurs de SNP et r_v au coefficient de corrélation pondéré par la matrice d'apparentement ; les vecteurs de données génotypiques aux SNP étant chacun multiplié par l'inverse de la racine carrée de la matrice *V* (**Figure 9**). Les calculs de DL ainsi que les matrices d'apparentement ont été réalisés à l'échelle de la population mondiale ainsi qu'à l'échelle des régions géographiques, c'est-à-dire l'Amérique, l'Asie Centrale et l'Asie de l'est, l'Europe, le Moyen-orient, l'Afrique subsaharienne et l'Océanie. Nous effectuons également un test de corrélation (statistiques T_r et T_{r_v}) pour mesurer la significativité de chaque valeur de corrélation (r et r_v) et obtenir une p-valeur. Enfin, nous appliquons une correction de Bonferroni sur les p-valeur au seuil de 10^{-7} lorsque nous comparons les SNP avec l'approche appât (i.e. correction de Bonferroni, en « une dimension », pour α = 5% et 0.05/431 951= 1.15*10^{-7} tests).

2.2.2 Approche appât chez *Medicago truncatula*

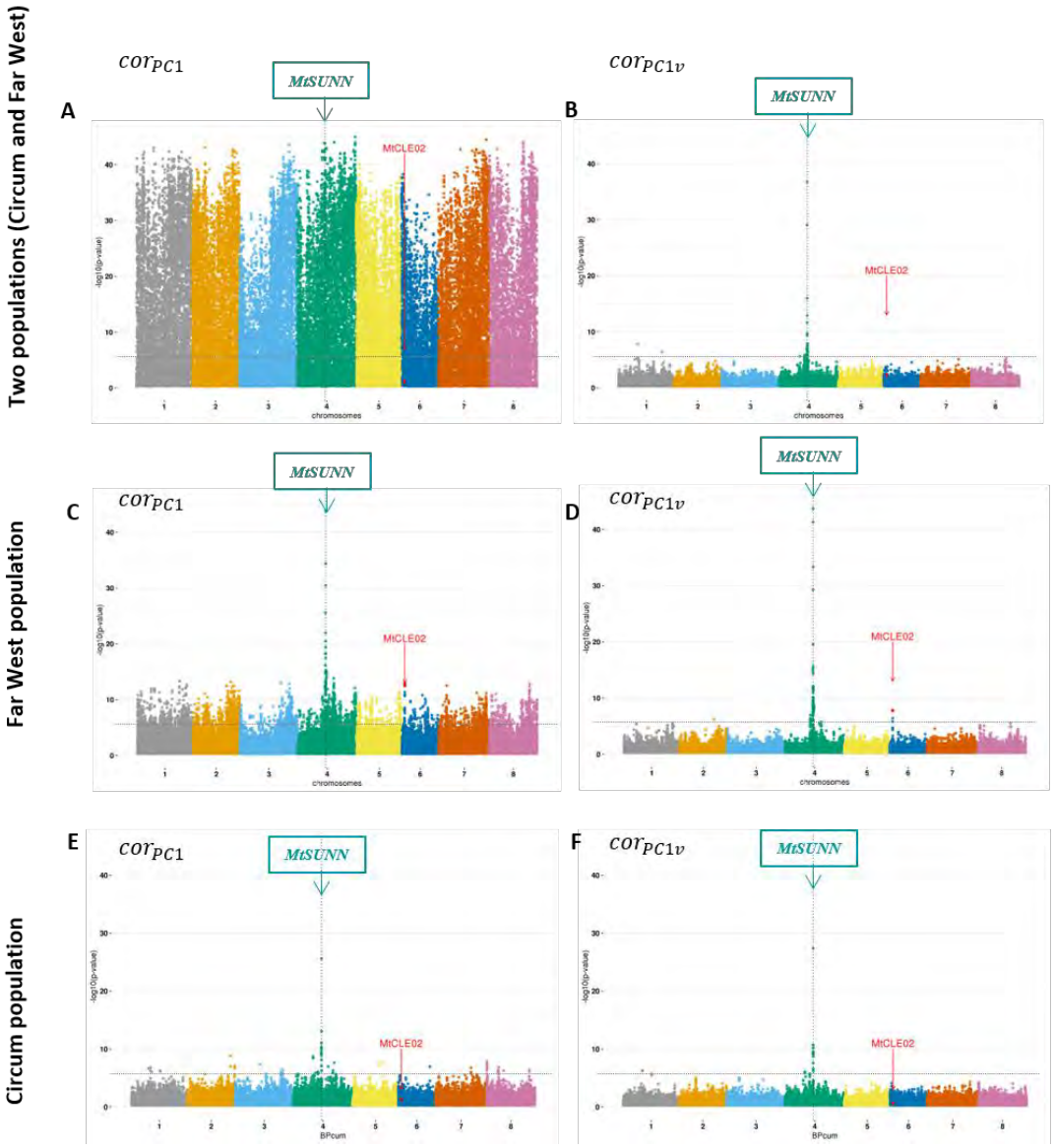
L'approche appât a été testée sur 98 gènes candidats de *M. truncatula*, ces gènes sont classés dans deux grandes catégories biologiques : la symbiose rhizobienne et la symbiose mycorhizienne. Des analyses détaillées sur 3 gènes candidats sont présentées dans cette partie.

2.2.2.1 Association entre le gène candidat MtSUNN et MtCLE02

Le gène *MtSUNN* (Medtr4g070970) a été testé en approche appât ; la protéine qu'il code est un récepteur de type Leucin-Rich Repeat Receptor-Like Kinase (LRR-RLK) impliqué dans la régulation systémique négative de la formation des nodules pendant la symbiose

rhizobienne. Il a été montré que les peptides codés par les gènes *MtCLE12* et *MtCLE13* interagissent avec SUNN pour réguler négativement la formation des nodules (Laffont et al., 2019 ; Mortier et al., 2010, 2012). L'analyse de GWESS que nous avons mise en place avec le gène *MtSUNN* comme appât ne nous a pas permis de retrouver une interaction épistatique entre *MtSUNN* et *MtCLE12*/*MtCLE13*, mais nous avons identifié un signal d'interaction avec un gène membre de la même famille : *MtCLE02* (Medtr6g009390).

Figure 29 : Distribution du DL entre le gène appât *MtSUNN* et tous les autres gènes du génome de *M. truncatula*. Le DL entre le gène *MtSUNN* et les autres gènes de *M. truncatula* est calculé dans la population entière (**A** - **B**), dans la population Far-West (**C** - **D**) et dans la population Circum (**E** - **F**). Les p-valeurs des tests de corrélation sont calculées à partir des statistiques $Tcor_{PC1}$ (**A** – **C** – **E**) et $Tcor_{PC1v}$ (**B** - **D** - **F**) qui prennent en compte la matrice kinship. L'axe des x correspond aux positions des gènes répartis sur les 8 chromosomes de *Medicago truncatula*. L'axe des y est le -$\log_{10}$(p-valeur) des tests de corrélation.

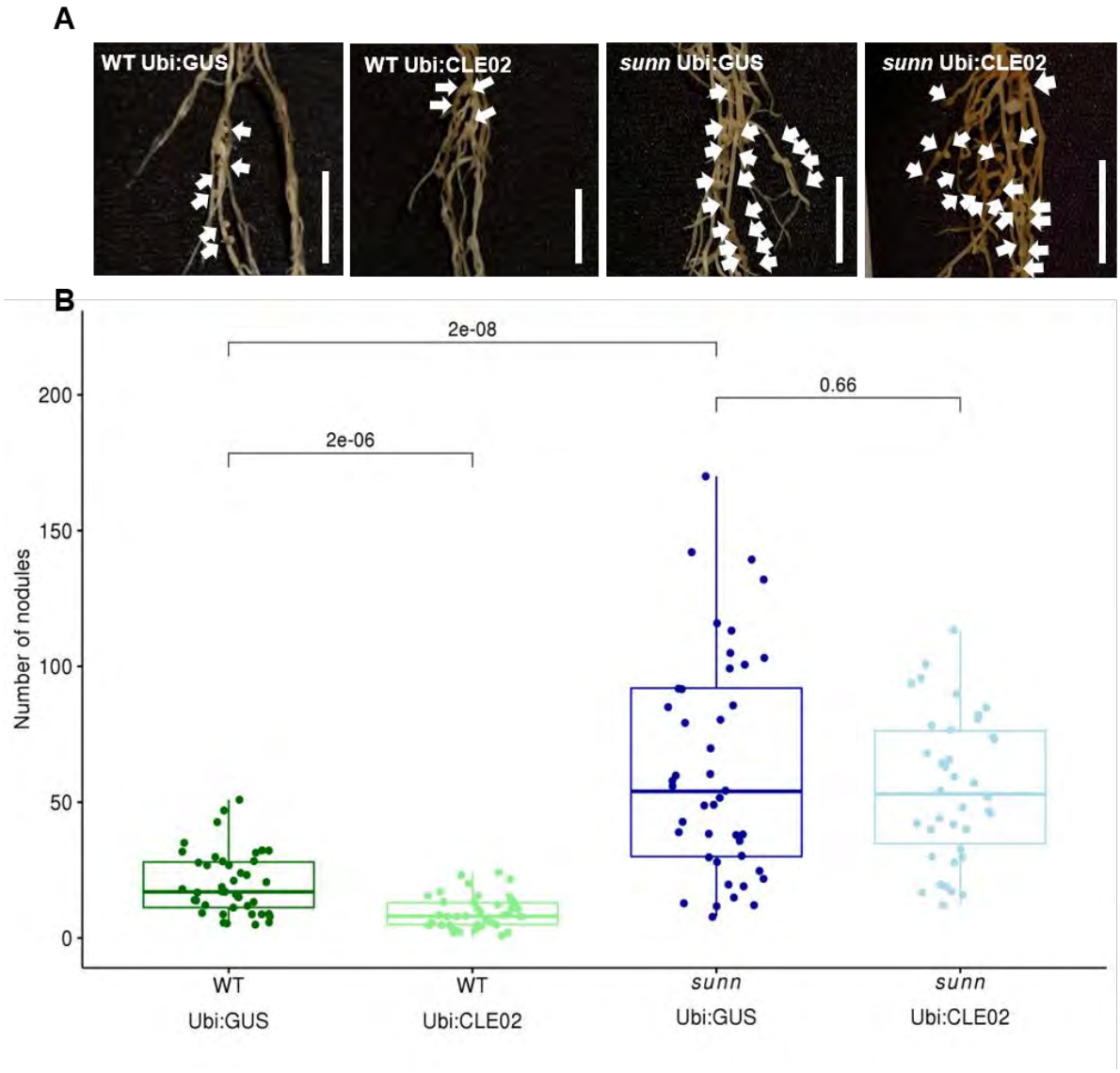


La **Figure 29** montre la distribution du DL obtenu entre le gène appât *MtSUNN* et tous les autres gènes du génome dans les deux sous-populations de *Medicago truncatula* ainsi que

dans la population totale et pour les mesures de DL cor_{PC1} et cor_{PC1v}. Les statistiques T_{corPC1} et T_{corPC1_v} ont été calculées à partir des PC1 issus des ACP réalisées sur les données SNP des gènes de *M. truncatula* et les p-valeurs ont été obtenues à partir de la distribution nulle de Student $\tau_{(n-2)}$. Les figures basées sur les statistiques T_{corPC1} (**Figure 29A,C,E**) montrent, tout d'abord, de fortes valeurs de DL à l'échelle du génome, avec de très faibles p-valeurs de tests par rapport aux figures basées sur T_{corPC1_v} (**Figure 29B,D,F**). Ce résultat montre que la correction par la matrice d'apparentement permet de réduire globalement le DL à l'échelle du génome notamment dans la population entière qui présente un degré de structuration plus élevé. Dans la sous-population FW, un pic sur le chromosome 6 obtenu avec la statistique T_{corPC1_v} montre une association significative entre les gènes *MtSUNN* et *MtCLE02* (**Figure 29D**, p-valeur = 1.7 x 10^{-8}). *MtCLE02* est le principal gène candidat qui montre un signal de sélection épistatique avec *MtSUNN* en dehors du chromosome 4 où est situé *MtSUNN*. Le gène *MtCLE02* présente également des signaux de sélection épistatique avec *MtSUNN* lorsque le DL est calculé avec la statistique T_{corPC1} (**Figure 29C**) mais dans ce cas, de nombreux autres gènes présentent aussi des signaux similaires voire plus significatifs indiquant qu'à l'échelle du génome, si l'on n'utilise pas la correction par la matrice d'apparentement il y a un taux de faux positifs élevé. Enfin, *MtCLE02* ne montre aucun signal de coadaptation avec *MtSUNN* lorsque le DL est calculé à l'échelle de la population entière ou de la population C (**Figure 29A,B**, p-valeurs de 0.077 et 0.006 pour T_{corPC1} et T_{corPC1_v} en population totale; et **Figure 29E,F**, p-valeurs de 0.05 et 0.24 pour T_{corPC1} et T_{corPC1_v} en population C). Ce résultat montre que le signal de sélection épistatique est local et que la sélection opère donc à l'échelle de la sous-population FW. Dans ce cas, la sélection épistatique ne peut être détectée à l'aide des données SNP de la population totale avec la statistique T_{corPC1_v} car le signal est confondu avec la structure génétique. Les figures en **Annexe 9** montrent les distributions du DL entre le gène appât *MtCLE02* et tous les autres gènes du génome, calculées dans les trois populations. Lorsque le DL est calculé à l'échelle de la population entière ou de la population C avec la statistique T_{corPC1_v} (**Annexe 9B,F**), le gène *MtCLE02* ne montre pas de signaux de coadaptation avec le gène *MtSUNN*. En population FW, un pic sur le chromosome 4 contenant le gène *MtSUNN* montre une association significative avec *MtCLE02* (**Annexe 9D**), bien que le gène *MtSUNN* ne se positionne pas en haut du pic mais à 21kb.

Ainsi, le gène *MtSUNN*, codant pour un récepteur LRR-RLK et le gène *MtCLE02*, codant pour un peptide sécrété de type CLAVATA (Mortier et al., 2010, 2012), semblent avoir co-évolué. Le récepteur SUNN intervient dans la régulation systémique de la nodulation ; il régule négativement la symbiose et il a précédemment été identifié en association avec les peptides CLE12 et CLE13. L'expression de SUNN ainsi que des peptides CLE12 et CLE13 est induite par l'inoculation de la bactérie *Rhizobium* pendant l'initiation de la nodulation (Mortier et al., 2010, 2012) mais ça n'est pas le cas pour le peptide CLE02 (**Annexe 10A,B**).

Figure 30 : Validation expérimentale de la relation fonctionnelle entre les gènes *MtCLE02* et *MtSUNN* dans la nodulation symbiotique chez *M. truncatula*. (A) Image représentant les racines nodulées, 14 jours après l'inoculation de rhizobium et surexprimant le gène *MtCLE02* (Ubi:CLE02) ou un gène de contrôle *GUS* (Ubi:GUS) soit dans les plantes sauvages (WT), soit dans les plantes mutantes *sunn*. Les barres d'échelle = 1 cm. **(B)** Boxplots représentant la variation du nombre de nodules dans les mêmes conditions décrites en **A**. Un test de Wilcoxon-Mann-Whitney sur la somme des rangs a été utilisé pour évaluer les différences statistiques par paires de conditions expérimentales.



Les relations entre les peptides CLE et le récepteur SUNN précédemment étudiées nous ont amenés à tester une interaction fonctionnelle supposée entre le peptide CLE02 et le récepteur SUNN. Grâce à une collaboration avec l'équipe de Florian Frugier (IPS2, Université Paris-Saclay) des analyses fonctionnelles ont été effectuées afin de tester l'interaction entre ces deux gènes avec une approche de génétique consistant à surexprimer le gène *MtCLE02* dans des racines de *M. truncatula* sauvages (WT) ou mutantes pour le gène *MtSUNN* (mutant *sunn*) qui n'exprime pas la protéine correspondante. Le matériel et méthode de ces expérimentations est décrit dans l'article publié dans la revue *Heredity*. Les résultats de ces expérimentations montrent tout d'abord que le nombre de nodules sur les racines de plantes WT par rapport aux plantes mutantes *sunn* est significativement inférieur, mettant en évidence le phénotype de supernodulation connu des mutants *sunn* (**Figure 30A,B**, test de Wilcoxon-Mann-Whitney, p-valeur = 2×10^{-8}). Ensuite, le nombre de nodule diminue de façon significative lorsque *MtCLE02* est surexprimé (comme le confirme la qRT-PCR, **Annexe 10C**) dans les racines de WT, ce qui montre un rôle négatif de *MtCLE02* sur la nodulation (**Figure 30A,B**, test de Wilcoxon-Mann-Whitney, valeur $p = 2 \times 10^{-6}$). Enfin, la surexpression de *MtCLE02* dans les racines du mutant *sunn* ne modifie pas le nombre de nodules (test de Wilcoxon-Mann-Whitney, p-value = 0.66) contrairement à ce qui est observé dans les racines WT. Ces résultats montrent que le rôle négatif que peut avoir le peptide CLE02 sur la nodulation est dépendant du récepteur SUNN. Ces résultats expérimentaux établissent donc un lien fonctionnel entre ces deux gènes de *M. truncatula* dans un contexte de nodulation symbiotique des racines. Ce lien fonctionnel a pu être testé suite à l'analyse de la sélection épistatique réalisée à l'aide de notre méthode basée sur le DL. Ces résultats montrent une co-évolution de ces deux gènes. Le polymorphisme observé sur le gène *MtSUNN* au sein de la population FW semble être dû à la sélection balancée car la statistique ***H*** (voir chapitre 1.3) est égale à 1.45 (rang de 8.41% des valeurs les plus élevées du génome). En revanche, dans la population entière et dans la population C, ***H*** est égale à 0.52 et -0.41, ce qui correspond, respectivement, aux rangs de 39.5% et 58.7% des valeurs les plus élevées du génome. Le polymorphisme au niveau du gène *MtCLE02* semble être dû à un balayage sélectif qui est détectable à l'échelle de la population entière et de la population C (***H*** = -2.25 et -4.45; rangs de 8.98% et 4% des valeurs les plus faibles du génome), mais pas dans la population FW où le polymorphisme est de type neutre (***H*** = 0.027; rang de 41.3% des valeurs les plus faibles du

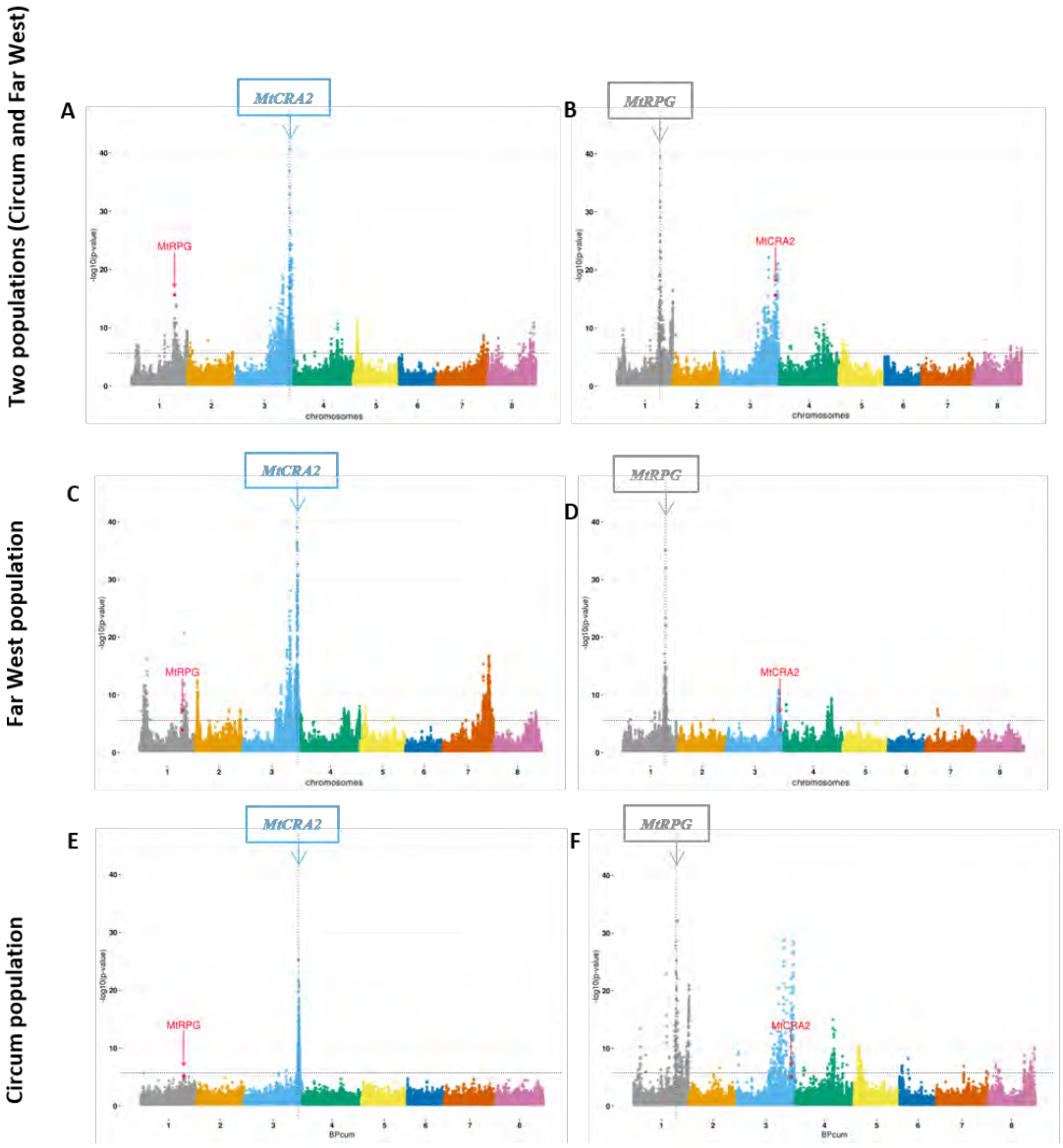
génome). Les deux gènes *MtSUNN* et *MtCLE02* semblent rester polymorphes et sous sélection épistatique dans la population FW, peut être par un mécanisme compensatoire. Comme preuve de concept et pour illustrer le rôle de la sélection épistatique entre ces gènes, une approche de génétique a été utilisée pour démontrer l'interaction fonctionnelle entre le peptide CLE02 et le récepteur SUNN. En effet, le peptide de signalisation CLE02 influence négativement le nombre de nodule de façon dépendante de SUNN comme les peptides CLE12 et CLE13 précédemment étudiés (Mortier et al., 2010, 2012). Enfin, le gène *MtCLE02* n'est pas régulé lors de la nodulation symbiotique contrairement aux gènes *MtCLE12* et *MtCLE13*, et phylogénétiquement *MtCLE02* n'est pas relié à *MtCLE12* ni *MtCLE13* dont le rôle négatif sur le nombre de nodules a été précédemment montré. Ces résultats montrent qu'avec notre méthode de détection de la sélection épistatique, nous avons pu identifier un nouveau peptide CLE agissant dans la même voie biologique que les peptides précédemment étudiés et nous avons pu identifier une interaction fonctionnelle entre deux gènes qui n'ont pas nécessairement les mêmes profils d'expression ou de corégulation.

2.2.2.2 Association entre le gène candidat MtCRA2 et MtRPG

Le gène *MtCRA2* (compact root architecture 2, Medtr3g110840) a été testé en approche appât. Il code pour une protéine de type Leucin-Rich Repeat Receptor-Like Kinase (LRR-RLK) qui, de manière antagoniste et indépendante de *MtSUNN*, régule positivement et de façon systémique le nombre de nodules pendant la symbiose rhizobienne (Laffont et al., 2019). Il a été montré que les peptides CEP interagissent avec CRA2 pour réguler positivement le nombre de nodules au cours de la symbiose. L'analyse GWESS que nous avons réalisée avec le gène *MtCRA2* comme appât ne nous a pas permis d'identifier une interaction épistatique significative entre *MtCRA2* et un ou plusieurs peptides CEP. En revanche, nous avons identifié une interaction significative avec le gène *MtRPG* (rhizobium-directed polar growth, Medtr1g090807). La protéine du gène *MtRPG* contrôle l'infection par rhizobium et est requise pour l'infection et la mise en place de la symbiose rhizobienne chez de nombreuses espèces (Arrighi et al., 2008; Griesmann et al., 2018). La **Figure 31** montre les distributions du DL entre les gènes appâts *MtCRA2* et *MtRPG* et tous les autres gènes du génome de *Medicago truncatula* dans les deux sous-populations FW et C, ainsi que dans la population entière. Seuls les résultats obtenus avec la statistique T_{corPC1_v} sont représentés car les distributions de DL

calculées avec T_{corPC1} présentent des proportions de faux positifs trop élevées comme nous l'avons montré précédemment.

Figure 31 : Distribution du DL entre les gènes appâts *MtCRA2 et MtRPG* et tous les autres gènes du génome de *M. truncatula*. Le DL entre les gènes *MtCRA2* et *MtRPG* et les autres gènes de *M. truncatula* est calculé respectivement dans la population entière (**A - B**), dans la population Far-West (**C - D**) et dans la population Circum (**E - F**). Les p-valeurs des tests de corrélation sont calculées à partir de la statistique $Tcor_{PC1v}$ qui prend en compte la matrice kinship. L'axe des x correspond aux positions des gènes répartis sur les 8 chromosomes de *Medicago truncatula*. L'axe des y est le -$\log_{10}$(p-valeur) des tests de corrélation.



Ainsi, dans la population entière, la distribution du DL obtenue avec *MtCRA2* comme gène appât montre un pic sur le chromosome 1 correspondant au gène *MtRPG* (**Figure 31A**, p-valeur = 2.16 x 10^{-16}). Le pic sur ce chromosome montre une association significative entre les deux gènes *MtCRA2* et *MtRPG* car le point au somment du pic qui correspond à *MtRPG* possède la p-valeur la plus faible sur l'ensemble du génome, en dehors du chromosome 3 où

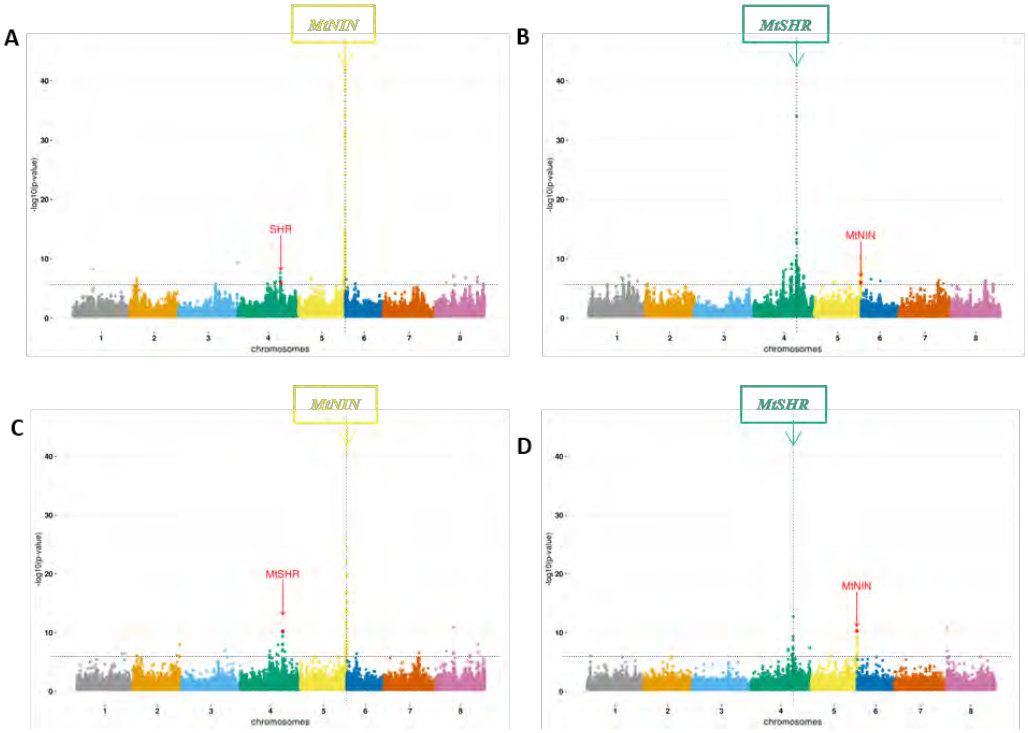
se situe la région génomique de *MtCRA2*. À l'inverse, la distribution du DL obtenue en population entière avec *MtRPG* comme appât montre un pic significatif contenant le gène *MtCRA2* (**Figure 31B**), mais celui-ci ne correspond pas au sommet du pic d'association, bien qu'ayant toujours une p-valeur = 2.16 x 10^{-16}. Ces résultats semblent malgré tout montrer un signal probable de cosélection entre *MtCRA2* et *MtRPG* au sein de la population entière. Les distributions de DL calculées avec *MtCRA2* et *MtRPG* comme appât au sein de la population FW ne montrent pas de signal de cosélection (**Figure 31C,D**) et les distributions au sein de la population C montrent un signal entre *MtRPG* et *MtCRA2*, quand ce dernier est utilisé comme appât (**Figure 31E**, p-value = 8.7 x 10^{-6} avec *MtRPG* qui est le 5^{ième} gène le plus significatif, en dehors du chromosome 3 de *MtCRA2*). Ces résultats mettent en évidence une signature de coadaptation entre deux gènes connus qui sont impliqués dans la nodulation. En effet, l'un régule positivement et de façon systémique la formation des nodules (*MtCRA2*), et l'autre est nécessaire à l'infection par *Rhizobium* au cours de la formation du nodule (*MtRPG*). De plus, dans le cadre de l'ANR PSYCHE à laquelle nous avons participé (coordinateur : Florian Frugier, IPS2, Université Paris-Saclay), ces deux gènes ont été trouvés comme faisant partie d'un même cluster de coexpression lors d'une cinétique d'expression dans les nodules de *Medicago truncatula*. En effet, d'après ces analyses, ces deux gènes ont des profils d'expression très similaires qui résultent de l'application d'un stress hydrique (sécheresse) et d'une carence en azote à distance (système expérimental de split-root). Cette réponse fait partie d'un mécanisme physiologique de compensation systémique de la nodulation (données non publiées). Ainsi, *MtCRA2* et *MtRPG* sont tous deux impliqués dans le processus biologique de nodulation, ils ont des profils d'expression similaires dans certaines conditions de stress et nous les identifions sous sélection épistatique ou cosélection. Il semblerait donc qu'ils soient coadaptés, peut-être dans le cadre d'un mécanisme de régulation de *MtRPG* par *MtCRA2*, mais on ne sait pas si - et comment - ils interagissent au niveau moléculaire.

2.2.2.3 Association entre le gène candidat MtNIN et MtSHR

Le gène *MtNIN* (Nodule Inception, Medtr5g099060) a également été testé en approche appât. *MtNIN* code pour un facteur de transcription impliqué dans la symbiose, il est essentiel pour l'organogène des nodules ainsi que pour l'initiation de l'infection par les bactéries dans les racines (Griesmann et al., 2018; Madsen et al., 2010; Marsh et al., 2007; Schauser et al., 1999). L'analyse de GWESS réalisée avec *MtNIN* comme appât nous a permis d'identifier un

signal significatif avec le gène *MtSHR* (short-root, Medtr4g097080). La protéine du gène *MtSHR* est un facteur de transcription de type GRAS, une famille de facteurs de transcriptions spécifique des plantes et impliqués dans divers processus au cours du développement racinaire. De façon intéressante, *MtSHR* a récemment été associé, d'après une analyse GWAS, à la variation naturelle de la stimulation de la croissance racinaire de *Medicago truncatula* en réponse à des signaux symbiotiques de type lipochitooligosaccharides -LCO- (Bonhomme et al., en préparation). La **Figure 32** montre les distributions de DL entre les gènes appâts *MtNIN* et *MtSHR* et tous les autres gènes du génome de *Medicago truncatula* dans la population FW, calculées avec la statistique T_{corPC1_v}.

Figure 32. Distribution du DL entre les gènes appâts *MtNIN* et *MtSHR* et tous les autres gènes du génome de *M. truncatula* dans la population Far-West. Le DL entre les gènes *MtNIN* et *MtSHR* et tous les autres gènes de *M. truncatula* est calculé respectivement à partir de fenêtres génomiques de 10kb centrées sur chaque gène (**A - B**), ou à partir de fenêtres génomiques comprenant uniquement les SNP dans les gènes (séquence génique) (**C - D**). Les p-valeurs des tests de corrélation sont calculées à partir de la statistique $Tcor_{PC1v}$ qui prend en compte la matrice kinship. L'axe des x correspond aux positions des gènes répartis sur les 8 chromosomes de *Medicago truncatula*. L'axe des y est le -$\log_{10}$(p-valeur) des tests de corrélation.



Les résultats obtenus avec la statistique T_{corPC1_v} en population entière et dans la population C sont représentés en **Annexes 11 et 12**. Les **Figures 32A,B et 32C,D** montrent les distributions du DL obtenues en population FW mais la différence entre ces figures est la taille des fenêtres génomiques sur lesquelles les ACP ont été réalisées pour extraire les valeurs de

PC1. Dans les **Figures 32A et 32B** les ACP sont réalisées sur des fenêtres de 10kb centrées sur chaque gène, et dans les **Figures32C 32D** les ACP sont réalisées sur des fenêtres comprenant uniquement la séquence génique. Ainsi, lorsque l'on modifie la taille des fenêtres génomiques pour extraire les haplotypes multi-SNP, cela influence en partie les résultats de DL. La **Figure 32C** représente la distribution du DL avec le gène *MtNIN* comme appât et sur le chromosome 4, on observe un pic avec le gène *MtSHR* au sommet (**Figure 32C**, p-value = 5.2×10^{-11}). La **Figure 32A** représente également la distribution du DL avec le gène *MtNIN* comme appât mais les ACP ont été réalisées sur des fenêtres de 10kb. On observe aussi un pic sur le chromosome 4 au niveau du gène *MtSHR* mais ce dernier n'est pas au somment du pic (**Figure 32A**, p-value = 1.03×10^{-6}). Inversement, les **Figures 32B et 32D** représentent les distributions du DL avec *MtSHR* comme appât. Sur les deux figures, on observe deux pics correspondant au gène *MtNIN* mais sur la **Figure 32D**, où l'on a utilisé des fenêtres génomiques correspondant aux séquences géniques, le gène *MtNIN* se situe au sommet du pic et il possède la p-valeur la plus faible sur l'ensemble du génome (**Figure 32D**, p-value = 5.2×10^{-11}). Les distributions de DL calculées dans la population entière ainsi que dans la population C ne montrent pas de signaux d'interactions significatifs entre les gènes *MtNIN* et *MtSHR*. Ainsi, nous avons identifié une interaction adaptative significative entre deux autres gènes impliqués dans la symbiose rhizobienne. Dans le cadre l'ANR DeCoD, une analyse du mutant du gène *MtSHR* est prévue afin de comprendre son rôle dans cette symbiose ainsi que son interaction génétique potentielle avec *MtNIN*. Enfin, ces résultats obtenus entre les gènes *MtNIN* et *MtSHR* montrent l'importance du choix des tailles de fenêtres génomiques pour calculer les haplotypes à l'aide de l'ACP. Les fenêtres de 10kb permettent, en principe, de travailler avec un nombre raisonnable de SNP pour le calcul des haplotypes. En revanche, si la fenêtre est trop grande, on est au-delà de la décroissance du DL (« LD decay ») et les SNP utilisés pour les analyses ACP ne sont plus suffisamment en DL car les événements de recombinaison sont plus fréquents. Ainsi selon la densité en SNP des données, le choix de la taille des fenêtres est un compromis entre le nombre de SNP et le DL entre les SNP contenus dans la fenêtre (ou LD-decay).

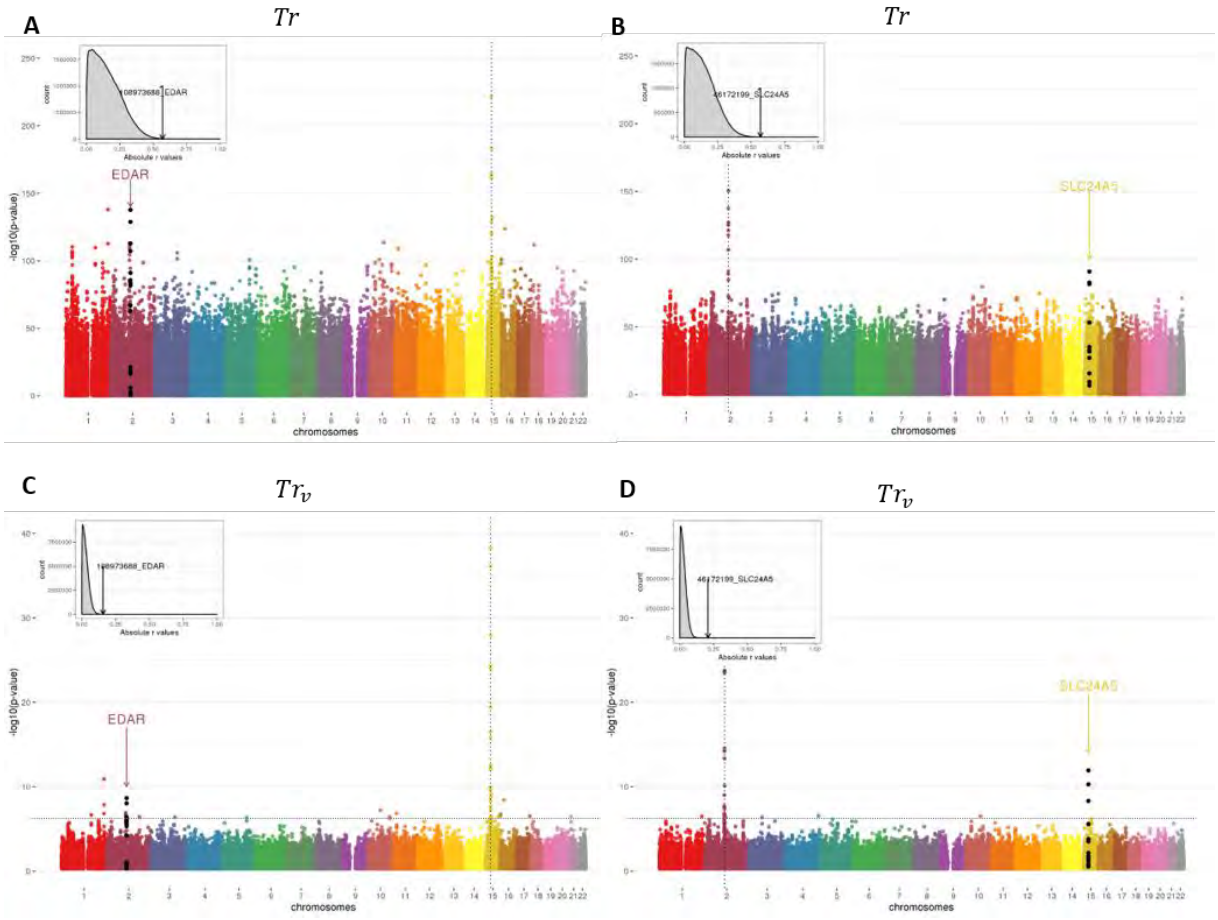
2.2.3 Approche appât chez l'humain

L'approche appât chez l'humain a été testée sur quelques gènes candidats à l'échelle de la population mondiale et de plusieurs régions géographiques. Le DL est calculé à l'échelle des SNP avec les statistiques r et r_v et seulement un exemple est présenté dans cette partie.

2.2.3.1 Association entre les gènes SLC24A5 et EDAR

Les gènes *SLC24A5* et *EDAR* ont tous les deux étés testés en approche appât à partir des données SNP humaines. *SL24A5* code pour une protéine échangeur de cation qui affecte la pigmentation chez l'homme et le poisson zèbre (Lamason et al., 2005) et *EDAR* code pour un récepteur impliqué dans le développement des follicules pileux, des dents et des glandes sudoripares (Botchkarev & Fessing, 2005; Sadier et al., 2014). Les deux gènes *SLC24A5* et *EDAR* ont précédemment été identifiés sous sélection positive dans les populations Européennes et de l'Asie de l'Est, respectivement (Bryk et al., 2008; Sabeti et al., 2007; Speidel et al., 2019). Pour rechercher des signaux de sélection épistatique, les GWESS ont été réalisées avec les mesures r et r_v à partir des données SNP HGDP-CEPH obtenues sur un panel de 940 individus. Deux SNP appâts, 15_46172199 (rs2250072) et 2_108973688 (rs6749207) situés respectivement dans les gènes *SLC24A5* (chromosome 15) et *EDAR* (chromosome 2), ont été testés. Les statistiques T_r et T_{r_v} ont été calculées entre les deux SNP appâts et les 431 951 autres SNP du génome (après un filtre de MAF à 5%). La **Figure 33** montre la distribution du DL obtenue entre les SNP appâts 15_46172199 (rs2250072) et 2_108973688 (rs6749207) et tous les autres SNP de la population humaine à l'échelle mondiale. Les figures basées sur la statistique T_r (**Figures 33A,B**) montrent de fortes valeurs de DL à l'échelle du génome, avec de très faibles p-valeurs de tests, par rapport aux scans réalisés avec T_{r_v} (**Figures 33C,D**), montrant que la correction par la matrice d'apparentement permet de réduire globalement le DL à l'échelle de la population mondiale. Les distributions du DL avec le SNP 15_46172199 de *SLC24A5* comme appât (**Figures 33A,C**) montrent un pic au niveau de chromosome 2 correspondant aux SNP situés dans le gène *EDAR* avec le SNP 2_108946170 en haut du pic (**Figure 33C**; p-valeur = 2.29×10^{-9}). Inversement, les GWESS réalisées avec le SNP 2_108973688 de *EDAR* comme appât (**Figure 33B,D**) montrent un pic sur le chromosome 15 correspondant au gène *SLC24A5* avec le SNP 15_46179457 en haut du pic (**Figure 33D**; p-valeur = 1.2×10^{-12}).

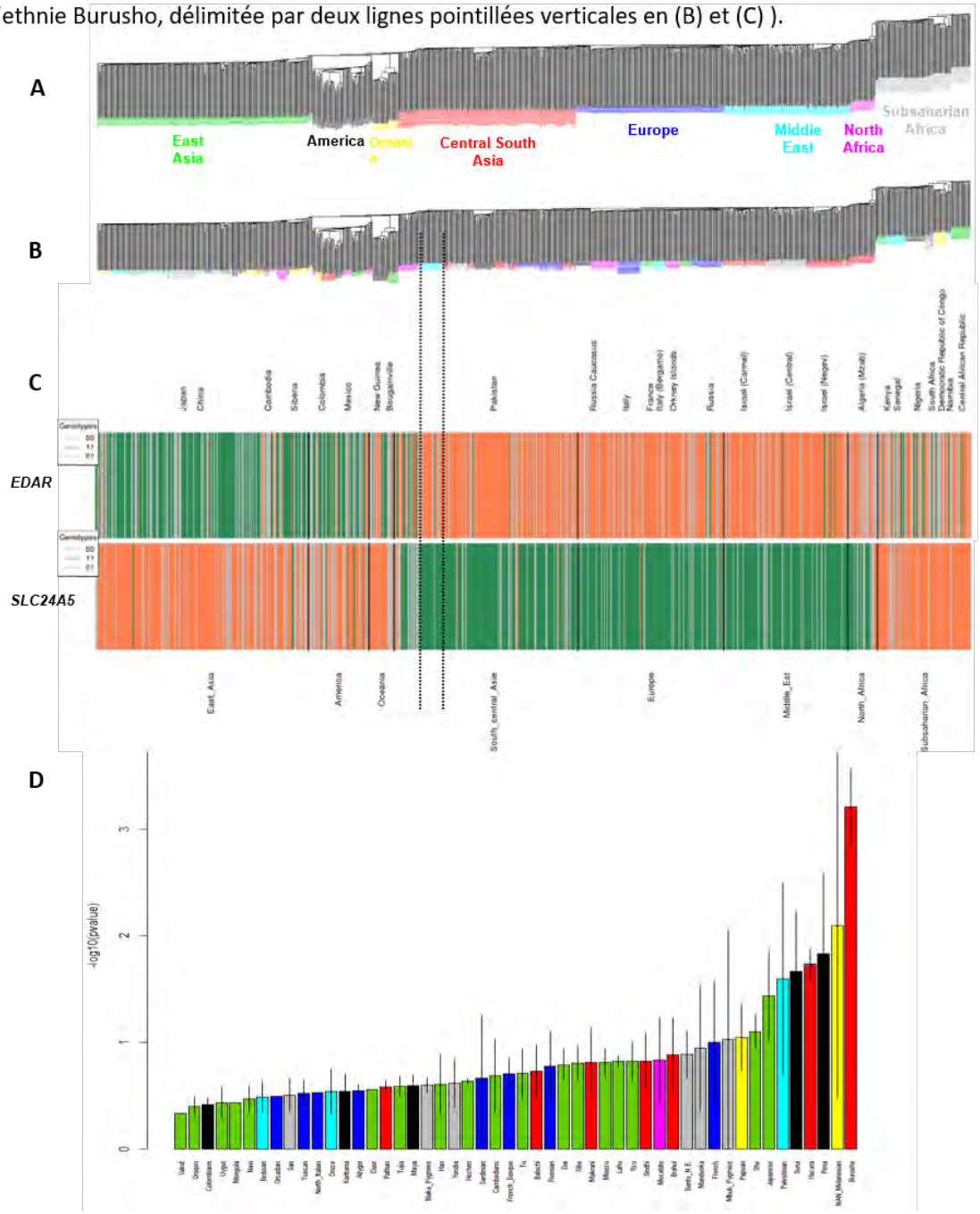
Figure 33 : Distribution du DL entre les SNP appâts des gènes *SLC24A5* et *EDAR* et tous les autres SNP des données HGDP-CEPH de ensemble de la population humaine à l'échelle mondiale (n=940). Le DL entre les SNP appâts 15_46172199 (*SLC24A5*) et 2_108973688 (*EDAR*) et tous les autres SNP du génome est calculé avec les statistiques T_r (**A, B**) ou T_{r_v} (**C, D**), qui prend en compte la matrice d'apparentement. L'axe des x correspond aux positions des SNP répartis sur les 22 autosomes humains, chaque point correspond à un SNP et les points noirs montrent les SNP des gènes candidats qui sont sous sélection épistatique avec le SNP appât de chaque figure (ligne verticale pointillée). L'axe des y est le -$\log_{10}$(p-value) du test de corrélation. Les graphiques en haut à gauche de chaque figure montrent les distributions du DL entre chaque SNP appât et tous les autres SNP du génome (à l'exception des SNP situés dans une fenêtre de 50Kb entourant les SNP appâts). La significativité de la corrélation entre le SNP appât de *SLC24A5* et le SNP cible de *EDAR* (respectivement le SNP appât de *EDAR* et le SNP cible de *SLC24A5*) est représentée par une flèche.



La **Figure 34** montre la répartition géographique des génotypes aux deux SNP appâts, 15_46172199 -*SLC24A5*- et 2_108973688 -*EDAR*-, à l'échelle mondiale. Cette répartition (**Figure 34C**) indique une corrélation avec la structure de la population représentée par l'arbre phylogénétique construit à partir de la matrice d'apparentement (**Figure 34A,B**). En effet, l'allèle dérivé du SNP 15_46172199 de *SLC24A5* qui est associé à l'allèle responsable du phénotype de peau claire du gène *SLC24A5* est présent en Europe, en Afrique du Nord, au Moyen Orient et en Asie centrale du sud. D'autre part, l'allèle dérivé du SNP 2_108973688 qui est associé à l'allèle responsable du phénotype des cheveux épais du gène *EDAR* est présent

en Asie de l'est, en Amérique et en Océanie (**Figure 34C**). Les signatures de DL observées avec la statistique T_r entre les gènes *SLC24A5* et *EDAR* à l'échelle de la population mondiale, sont le reflet de la sélection naturelle sur les allèles dérivés (en vert, **Figure 34C**) dans différentes régions géographiques de façon corrélée à la structure de la population mondiale. Cependant, lorsque le DL est calculé avec la correction pour la structure de la population et l'apparentement entre les individus avec la statistique T_{r_v}, il est toujours significatif entre les gènes *SLC24A5* et *EDAR* (**Figures 33C,D**). Ce résultat indique que la sélection épistatique ou la cosélection a probablement permis la coadaptation entre ces deux gènes au sein de sous-régions géographiques. Afin de localiser l'origine géographique de cette signature de sélection, des analyses de GWESS ont été réalisées au sein de six régions géographiques: l'Asie centrale du sud, l'Asie de l'est, l'Afrique subsaharienne, le Moyen-Orient, l'Europe et l'Amérique. Les résultats sont présentés en **Annexes 13 à 18**. Seule l'analyse de GWESS réalisée dans la population de l'Asie centrale du sud présente des signatures de DL significatives entre les SNP des gènes *SLC24A5* et *EDAR* (**Annexe 13C**, p-valeur = 6.7 x 10^{-6} au SNP 2_108973688 avec le SNP 15_46172199 de *SLC24A5* comme appât ; **Annexe 13D**, p-valeur = 2.8 x 10^{-6} au SNP 15_46174380 avec le SNP 2_108973688 de *EDAR* comme appât ; en utilisant la statistique T_{r_v}). Les populations humaines issues des données HGDP-CEPH de la région de l'Asie centrale du sud sont constituées de 8 ethnies différentes originaires du Pakistan. Les GWESS réalisées avec les statistiques T_r et T_{r_v} au sein de cette région géographique du Pakistan présentent des patrons de DL similaires (**Annexe 13**) montrant une faible structure génétique au sein de cette région. Afin de rechercher les signaux de cosélection ou de sélection épistatique, nous avons calculé le DL et réalisé des tests de corrélation avec T_{r_v} entre deux SNP du panel HGDP-CEPH situés dans le gène *SLC24A5* (15_46179457 -rs1834640- et 15_46172199) et trois SNP situés dans le gène *EDAR* (2_108962124 394 -rs260607-, 2_108982808 -rs17034770- et 2_108973688) pour 50 groupes ethniques répartis dans huit régions géographiques et montrant du polymorphisme pour ces cinq SNP (**Figure 34D**).

Figure 34 : Schéma de la structure de la population humaine déduite de la matrice d'apparentement et distribution géographique des allèles et du DL entre les SNP 15_46172199 (*SLC24A5*) et 2_108973688 (*EDAR*). (A) Les arbres phylogénétiques sont construits avec une méthode de « Neighbor-Joining » à partir de la matrice kinship basée sur les 431 951 SNP des données HGDP-CEPH. Les arbres représentent la structure de la population à l'échelle mondiale. (B) Même arbre que (A) montrant le clustering des différentes sous population ou ethnies. (C) Les barplots représentent les distributions géographiques des allèles aux SNP 2_108973688 (*EDAR*) et 15_46172199 (*SLC24A5*); ils repésentent les patterns de DL entre ces deux SNP qui sont principalement dus à la structure de la population ainsi qu'à la sélection sur les allèles dérivés (1, en vert) au niveau des deux gènes. (D) Moyenne (et écart-type) des valeurs de DL basées sur la statistique T_{r_v} entre les SNP de *SLC24A5* (SNP 15_46172199, SNP 15_46179457) et les SNP de *EDAR* (SNP 2_108962124, SNP 2_108973688, SNP 2_108982808) au sein de chaque sous-population humaine. Le barplot montre que le DL que l'on observe entre ces deux SNP provient principalement de la sous-population de l'Asie centrale du sud car il y a probablement une cosélection des allèles au niveau des ethnies du Pakistan (principalement de l'ethnie Burusho, délimitée par deux lignes pointillées verticales en (B) et (C)).



La moyenne et l'écart-type des -$\log_{10}$(p-valeur) calculés sur la base des 6 comparaisons de paires de SNP montrent que le groupe ethnique où l'on observe les valeurs de DL les plus élevées entre *SLC25A5* et *EDAR* est celui des Burusho au Pakistan (moyenne = 3.2 et écart-type = 0.36), comme le soulignent les associations des génotypes aux SNP 15_46172199 et 2_108973688 dans cette ethnie (**Figure 34C**, valeur de r_v = 0.63 pour les génotypes situés entre les deux droites verticales pointillées). Les patrons de DL observés entre *SLC24A5* et *EDAR* au sein de l'ethnie des Burusho ne semblent pas influencés par une quelconque sous-structure génétique de la population car les valeurs moyennes (et écart-type) de DL calculées avec T_r sont équivalentes aux valeurs de T_{r_v} (**Annexe 19**). Enfin, ces analyses de DL montrent aussi un signal plus faible mais significatif au sein de l'ethnie des Hazara également localisée au Pakistan, avec une moyenne et un écart-type du -$\log_{10}$(p-value) de 1.73 et 0.14, respectivement.

Chez l'humain, le gène *SLC24A5* possède un rôle important dans la variation de la pigmentation de la peau et il a été montré sous sélection positive dans les populations européennes (Deng & Xu, 2017; Izagirre et al., 2006; Sabeti et al., 2007). La mutation causale du phénotype de peau claire n'étant pas présente dans les données HGDP-CEPH (SNP rs1426654, 464 position: 15_46213776), nous avons utilisé les SNP localisés dans le même haplotype du gène *SLC24A5* (i.e. 15_46179457 et 15_46172199) (Basu Mallick et al., 2013; Beleza et al., 2013; Crawford et al., 2017). D'autre part, la mutation (V370A) caractérisée chez *EDAR* qui code pour un récepteur lié au récepteur TNFα impliqué dans la structure des cheveux ainsi que dans le développement des dents et des glandes sudoripares, a été identifiée sous sélection positive en Asie de l'est et dans les populations natives américaines (Bryk et al., 2008; Sadier et al., 2014; Speidel et al., 2019). La mutation causale du gène EDAR n'étant plus présente dans les données HGDP-CEPH, nous avons utilisé les SNP se situant dans la séquence génomique d'*EDAR* (i.e. 2_108962124, 2_108973688 et 2_108982808). Ainsi, la distribution géographique des génotypes aux SNP 15_46172199 et 2_108973688 est fortement corrélée à la structure de la population à l'échelle mondiale ce qui explique le fait que nous ayons obtenus des valeurs de DL élevées sur les scans réalisés avec la statistique T_r non corrigée. Enfin, les scans réalisés avec T_{r_v} montrent également de fortes valeurs de DL entre *SLC24A5* et *EDAR* à l'échelle de la population mondiale, probablement dues à de la sélection épistatique ou à de la cosélection. Le groupe ethnique des Burusho en Asie centrale du sud (Pakistan) semble être l'origine géographique de cette signature de cosélection entre

EDAR et *SLC24A5*. Les GWESS réalisées à l'échelle de la population de l'Asie centrale du sud avec les statistiques T_r et T_{r_v} présentent des valeurs de DL globalement similaires, ce qui montre qu'il y a une faible structure au sein de la population testée, comme observé précédemment dans cette région géographique ainsi qu'en Inde (S. Rosenberg et al., 2006). Dans le groupe ethnique des Burusho, il y a une association prédominante entre les allèles ancestraux des SNP d'*EDAR* et les allèles dérivés des SNP de *SLC24A5* montrant la coexistence des phénotypes de peau claire et de cheveux fins au sein de cette ethnie, avec probablement moins de morphotypes est-asiatiques de peaux plus foncées avec une structure capillaire de cheveux plus épais. Ces patrons phénotypiques pourraient être dus soit à de la sélection épistatique, soit à de la cosélection entre ces deux traits phénotypiques. Etant donné qu'aucun lien fonctionnel n'a pour l'instant été démontré entre la pigmentation de la peau et la voie de l'ectodysplasine, il semble plus probable que ces traits soient cosélectionnés car le lien entre ces deux phénotypes semble spécifique à certaines populations. Ainsi, nous avons pu montrer que le DL entre *EDAR* et *SLC24A5* n'est pas seulement dû à la sélection naturelle sur chacun des gènes de manière indépendante et dans des régions géographiques différentes, mais aussi probablement à de la cosélection entre ces gènes à l'échelle locale.

2.2.4 Conclusion/Discussion approche appât

Les GWESS réalisées chez *Medicago truncatula* et chez l'humain ont mis en évidence de nouvelles interactions évolutives entre plusieurs gènes candidats. Les deux statistiques r/r_v et cor_{PC1}/cor_{PC1v} ont été utilisées, montrant des signaux de sélection épistatique ou de cosélection à l'échelle des SNP ou des gènes. Le choix entre ces statistiques dépendra des données et de la densité en SNP; les signaux de sélection ainsi identifiés pourront varier en fonction des statistiques choisies. Les analyses en approche appât ont montré également l'importance d'utiliser la correction par la matrice d'apparentement afin de réduire le DL lié à la structure des populations et l'apparentement entre les individus. Les scans réalisés aussi bien chez *M. truncatula* que chez l'humain à l'échelle des populations entières montrent une différence significative entre les statistiques T_r/T_{corPC1} et T_{r_v}/T_{corPC1_v}. En revanche, les scans réalisés à l'échelle de certaines régions géographiques ou dans certaines sous-populations, par exemple au sein de l'ethnie des Burusho chez l'humain, ne présentent pas de différence majeure entre T_r et T_{r_v}, montrant ainsi un faible effet de la structure de la population. Chez *Medicago truncatula,* l'approche appât avec le gène *MtSUNN* nous a permis d'identifier une

possible interaction épistatique avec le gène *MtCLE02.* Suite à une collaboration avec l'équipe de Florian Frugier (IPS2, Université Paris-Saclay) nous avons pu démontrer une interaction fonctionnelle entre ces gènes, apportant ainsi une preuve de concept à notre méthode avec une validation fonctionnelle. Dans un autre contexte, il est également possible de tester des gènes candidats identifiés par exemple en GWAS et rechercher des signaux de cosélection ou sélection épistatique avec d'autres gènes du génome (voir le cas de *MtSHR* et *MtNIN*). Les gènes identifiés en coexpression sont aussi de bons candidats pour rechercher des signatures de cosélection (voir le cas de *MtCRA2* et *MtRPG*), mais de façon logique tous les gènes coexprimés ne présentent pas nécessairement des signaux de cosélection ou de coadaptation et inversement les gènes cosélectionnés ne sont pas nécessairement coexprimés ou corégulés. Pour résumer, l'approche appât et les statistiques de DL développées pour détecter des interactions épistatiques sont des outils pour identifier de nouveaux gènes candidats en interaction fonctionnelle ou potentiellement impliqués dans une même voie biologique et dont les données de polymorphismes montrent qu'ils co-évoluent.

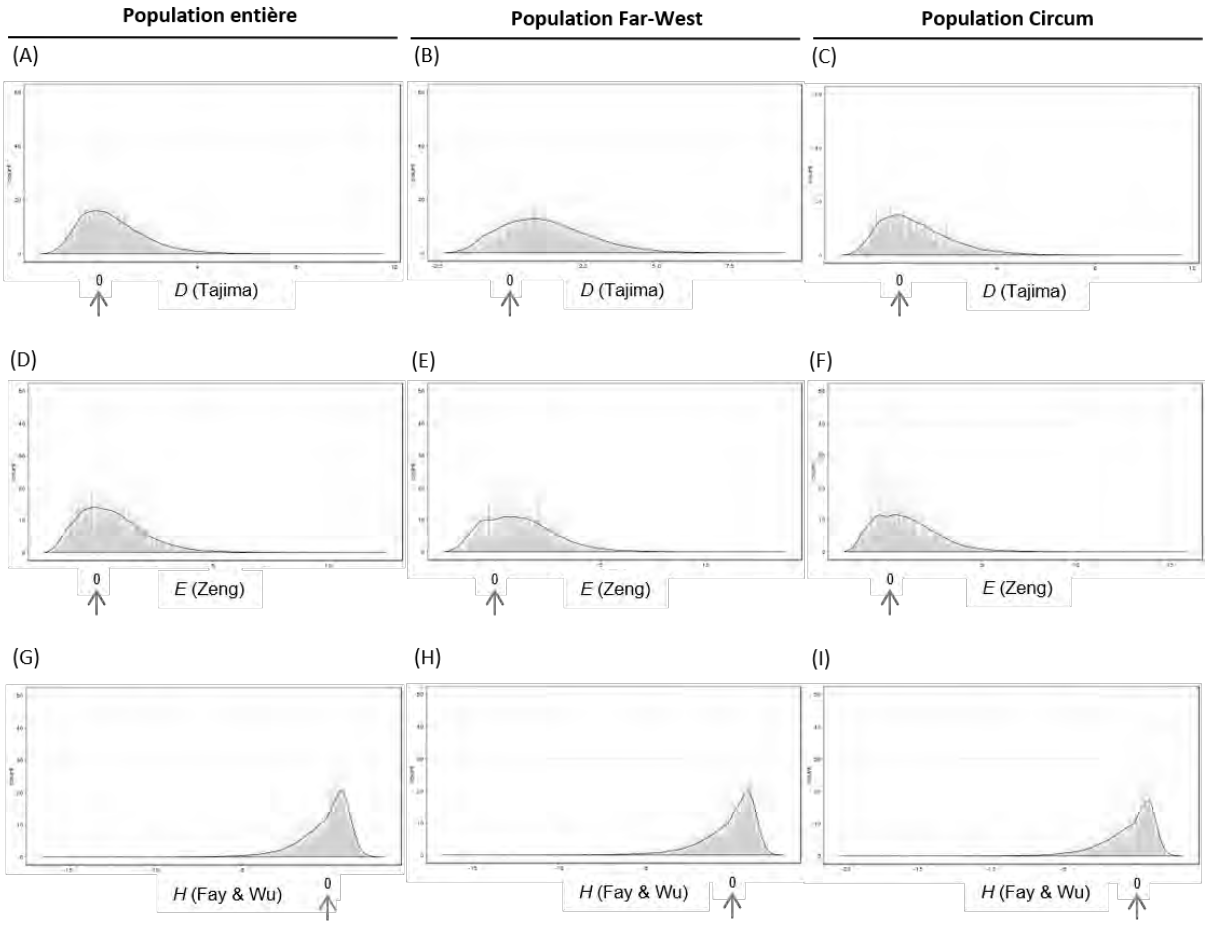
2.3 Polymorphisme moléculaire des gènes de Medicago truncatula et traces de sélection sur les gènes en épistasie

Afin d'étudier les patrons de polymorphisme chez *M. truncatula,* nous avons calculé des statistiques de tests de neutralité sur l'ensemble des gènes disponibles. Les statistiques ont été calculées sur des fenêtres de 10kb centrées sur les gènes, les mêmes fenêtres qui ont été utilisées pour calculer les statistiques de DL. L'objectif a été d'évaluer si les gènes de *M. truncatula* qui montrent des signatures de sélection épistatique, présentent également des signatures de sélection de manière individuelle. Les statistiques de neutralité ont aussi permis d'étudier les signatures de sélection à l'échelle du génome chez *M. truncatula*.

2.3.1 Polymorphisme à l'échelle du génome

Les statistiques de neutralité ont été calculées sur l'ensemble des gènes de *Medicago truncatula* dans les deux populations FW et C ainsi que dans la population entière. La **Figure 35** présente les distributions des statistiques de neutralité ; ***D*** de Tajima, ***E*** de Zeng et ***H*** de Fay & Wu calculées à partir des 48 331 gènes. Le détail des méthodes de calcul des statistiques est présenté en chapitre 1.3.

Figure 35 : Distributions des statistiques de tests de neutralité calculées sur l'ensemble des gènes de *Medicago truncatula* dans la population entière et dans les sous-populations Far-West et Circum. Les statistiques ***D*** de Tajima, ***E*** de Zeng et ***H*** de Fay & Wu sont calculées sur les fenêtres génomiques de 10kb entourant les gènes. Les figures (A, B, C) représentent les distributions du ***D*** de Tajima, les figures (D, E, F) représentent les distributions du ***E*** de Zeng, et les figures (G, H, I) représentent les distributions du ***H*** de Fay & Wu. (A, D, G) correspondent aux valeurs de statistiques calculées dans la population entière, (B, E, H) dans la population Far-West et (C, F, I) dans la population Circum. L'axe des abscisses correspond aux valeurs de statistiques calculées sur les gènes et l'axe des ordonnées correspond à la courbe de densité.



Les **Figures 35A,B,C** présentent les distributions du ***D*** de Tajima au sein des trois populations. Dans la population entière et la population C, le mode (valeur modale) des distributions de *D* (empiriques) est proche de zéro (moyenne de ***D*** en population entière = 0.53 ; moyenne de ***D*** en population C = 0.64). Cependant, dans la population FW, la distribution est décalée vers des valeurs positives (moyenne de ***D*** en population FW = 1.2). Le ***D*** de Tajima ($D = \frac{\theta_\pi - \theta_S}{\sqrt{Var(\theta_\pi - \theta_S)}}$) contraste θ_π qui correspond au degré moyen d'hétérozygotie et θ_S qui correspond au nombre de sites qui ségrégent au locus (i.e. le nombre de SNP). Lorsque la distribution est globalement décalée vers des valeurs positives, cela signifie qu'en moyenne $\theta_\pi > \theta_S$ et donc qu'il y a plus de variants en fréquences intermédiaires que de variants rares à l'échelle du génome. Ce déficit en allèles rares dans la population FW peut

être la conséquence d'un effet de sous-structuration génétique et/ou de l'effet d'un goulot d'étranglement.

Les **Figures 35D,E,F** présentent les distributions du ***E*** de Zeng au sein des trois populations. Là aussi, dans la population entière et la population C, le mode (valeur modale) des distributions de ***E*** est proche de zéro (moyenne de ***E*** en population entière = 0.6 ; moyenne de ***E*** en population C = 0.96), et dans la population FW la distribution est décalée vers des valeurs positives (moyenne de ***E*** en population FW = 1.13). Le ***E*** de Zeng ($E = \frac{\theta_L - \theta_S}{\sqrt{Var(\theta_L - \theta_S)}}$) contraste θ_L et θ_S qui correspondent respectivement au nombre moyen d'allèles dérivés accumulés depuis le plus récent ancêtre commun entre ces séquences et au nombre de sites qui ségrégent au locus. Si la distribution est décalée vers des valeurs positives, comme c'est le cas pour la population FW, cela signifie que $\theta_L > \theta_S$ et qu'il y a en moyenne plus de SNP avec des allèles dérivés en forte fréquence au sein de la population FW, que de SNP qui portent des allèles dérivés en faibles fréquences. Ce déficit en allèles dérivés rares dans la population FW peut être la conséquence d'un effet de sous-structuration génétique et/ou de l'effet d'un goulot d'étranglement, comme aussi indiqué par le ***D*** de Tajima.

Enfin, les **Figures 35G,H,I** montrent que le mode (valeur modale) des distributions de la statistique ***H*** de Fay & Wu est proche de zéro pour les trois populations (moyenne de ***H*** en population entière, FW et C = -0.18, -0.12 et -0.52, respectivement). Ici, l'effet de déficit en allèles rares, visualisé dans la population FW avec les statistiques ***D*** et ***E***, n'est pas visible avec le ***H*** de Fay & Wu ($H = \frac{\theta_\pi - \theta_L}{\sqrt{Var(\theta_\pi - \theta_L)}}$) qui est moins sensible aux allèles rares.

Nous savons que la population entière de *M. truncatula* est structurée en deux grandes populations FW et C. De plus, il a été montré que la population FW possède une taille efficace N_e plus grande que la population Circum (Bonhomme et al., 2015) et elle n'a pas subi de goulot d'étranglement. Le déficit en allèles rares observé dans la population FW est donc très probablement associé à un effet marqué de sous-structuration génétique. La **Figure 28**, qui présente les matrices d'apparentement calculées dans les trois populations montre qu'il y a bien une sous-structuration des populations FW et C qui semble beaucoup plus marquée dans la population FW. Ces résultats appuient les conclusions d'études précédentes sur la structuration génétique chez *M. truncatula* (Gentzbittel et al., 2019; Ronfort et al., 2006). De plus, la relation théorique entre N_e et le degré de structuration génétique est bien documentée et corrobore aussi les résultats obtenus pour les deux grandes populations FW

et C de l'espèce *M. truncatula*. En effet, il a été montré que N_e est positivement corrélée à la taille N de chaque sous-population (i.e. « deme »), au nombre de sous-populations D, ainsi qu'à un faible taux de migration m entre les sous-populations, par la relation $N_e = ND(1 + \frac{1}{2M})$, où $M = \frac{2NDm}{D-1}$ (Wakeley, 1999).

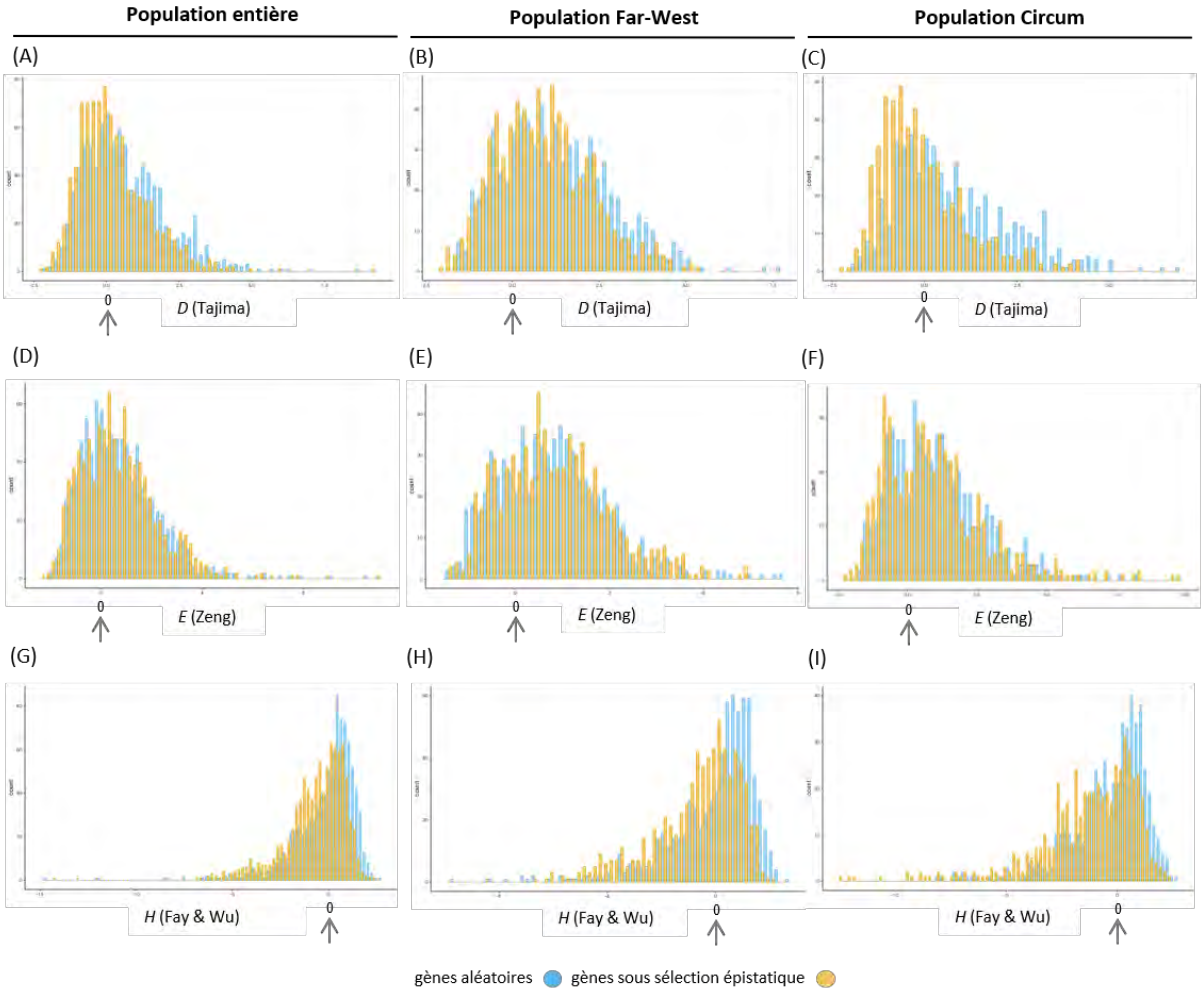
A l'échelle de la population entière, nos analyses de polymorphisme indiquent un masquage de la forte sous-structuration génétique de la population FW, par un nombre important de SNP dont les allèles dérivés sont en faibles fréquences et qui sont apportés par la population Circum, ce qui ramène en quelque sorte les distributions genome-wide de ***D*** et ***E*** à l'équilibre dans la population entière. Ces résultats confirment l'importance de tenir compte de la structuration génétique dans les analyses de signatures de sélection, en effectuant comme nous le faisons chez *M. truncatula* des analyses dans les 2 grandes sous-populations FW et C.

2.3.2 Signatures de sélection sur des gènes en épistasie

L'objectif de cette partie est d'évaluer si les gènes de *M. truncatula* qui montrent des signatures de sélection épistatique, présentent également des signatures de sélection de façon individuelle. Les statistiques ***D***, ***H*** et ***E*** de neutralité ont été calculées sur l'ensemble des gènes de *Medicago truncatula* et nous avons confronté ces résultats avec les résultats d'analyses genome-wide de signatures de sélection épistatique. Pour cela, nous avons comparé deux ensembles de gènes appelés *geneB* et *geneB'*. L'ensemble *geneB* est constitué de gènes qui présentent au moins un signal significatif (avec T_{corPC1_v}) avec un second gène (*geneA*) situé sur un chromosome différent et l'ensemble *geneB'* est constitué de gènes qui ne présentent aucune interaction significative (avec T_{corPC1_v}) avec un autre gène situé également sur un chromosome différent. Pour constituer ces deux ensembles, nous nous sommes basés sur la liste totale de paires de gènes, appelée *geneA-geneB* dont les valeurs de T_{corPC1_v} sont significatives au seuil de 10^{-11} (i.e. correction de Bonferroni pour α = 5% et 48 333 gènes, donc pour $\frac{48\,333\,(48\,333-1)}{2}$ comparaisons ; voir la description de ce jeu de données en chapitre 2.4) et qui se situent sur des chromosomes différents. Cet ensemble de paires de gènes *geneA-geneB* est constitué de toutes les paires de gènes significatives sur des chromosomes différents, au seuil fixé et dont nous avons supprimé tous les doublons. Ensuite, nous nous sommes basés sur la liste *geneA-geneB* et nous avons échantillonné des gènes

geneB' aux conditions ; (i) qu'ils ne soient pas déjà présents dans la liste *geneA-geneB* et (ii) qu'ils ne soient pas sur le même chromosome que le gène de la liste *geneA*. Les distributions des statistiques ***D***, ***E*** et ***H*** sont représentées en **Figure 36** pour les deux ensembles *geneB* et *geneB'*.

Figure 36 : Distribution des statistiques de neutralité calculées sur un ensemble de gènes sous sélection épistatique et sur un ensemble de gènes échantillonnés aléatoirement chez *Medicago truncatula* dans la population entière et dans les sous-populations Far-West et Circum. Les figures (A, B, C) représentent les distributions du ***D*** de Tajima, les figures (D, E, F) les distributions du ***E*** de Zeng et les figures (G, H, I) les distributions du ***H*** de Fay & Wu. (A, D, G) correspondent aux valeurs de statistiques calculées dans la population entière, (B, E, H) dans la population Far-West et (C, F, I) dans la population Circum. Les distributions des statistiques de neutralité sont représentées pour un ensemble de gènes sélectionnés aléatoirement et qui ne présentent pas de signature de sélection épistatique significatives (au seuil de 10^{-11}) (histogrammes bleu) ainsi que pour un ensemble de gènes qui présentent une (ou plusieurs) signature(s) de sélection épistatique significative sur des chromosomes différents (histogrammes jaune). L'axe des abscisses correspond aux valeurs des statistiques calculées sur les gènes et l'axe des ordonnées correspond à la courbe de densité.



Les **Figures 36A,B,C** présentent les distributions du ***D*** de Taijma au sein des trois populations et pour les deux ensembles de gènes (*geneB'* et *geneB*) qui appartiennent à des paires aléatoires ou à des paires sous sélection épistatique. Les distributions des gènes *geneB'* (en bleu, gènes échantillonnés aléatoirement) sont très similaires à celles obtenues sur

l'ensemble des gènes du génome de *Medicago truncatula* dans les trois populations (moyenne de ***D*** sur l'ensemble *geneB'* respectivement en population FW, C et population entière = 1.22, 0.73 et 0.60 ; moyenne sur l'ensemble des gènes respectivement en population FW, C et population entière = 1.2, 0.64 et 0.53). En revanche, les moyennes de ***D*** calculées sur l'ensemble *geneB* de gènes sous sélection épistatique sont plus faibles, notamment en population C (moyenne de ***D*** sur l'ensemble *geneB* respectivement en population FW, C et population entière = 0.95, -0.03 et 0.27). Ces résultats montrent que beaucoup de gènes qui présentent un signal de sélection épistatique semblent également montrer un signal de balayage sélectif.

Les **Figures 36D,E,F** présentent les distributions du ***E*** de Zeng au sein des trois populations et dans les deux ensembles de gènes *geneB* et *geneB'*. Les distributions de ***E*** parmi les ensembles de gènes aléatoires *geneB'* dans les trois populations sont proches des distributions obtenues sur l'ensemble du génome de *M. truncatula* (moyenne de ***E*** sur l'ensemble *geneB'* en population FW, C et population entière = 1.1, 0.84 et 0.67 ; moyenne sur l'ensemble des gènes en population FW, C et population entière = 1.13, 0.96 et 0.6) montrant que ces gènes évoluent probablement sous neutralité. Inversement, les distributions de ***E*** obtenus sur l'ensemble de gène sous sélection épistatique sont sensiblement décalées vers des valeurs positives et semblent montrer également des signaux de balayages sélectifs dans les trois populations (moyenne de ***E*** sur l'ensemble *geneB* en population FW, C et population entière = 1.36, 1.01 et 0.78).

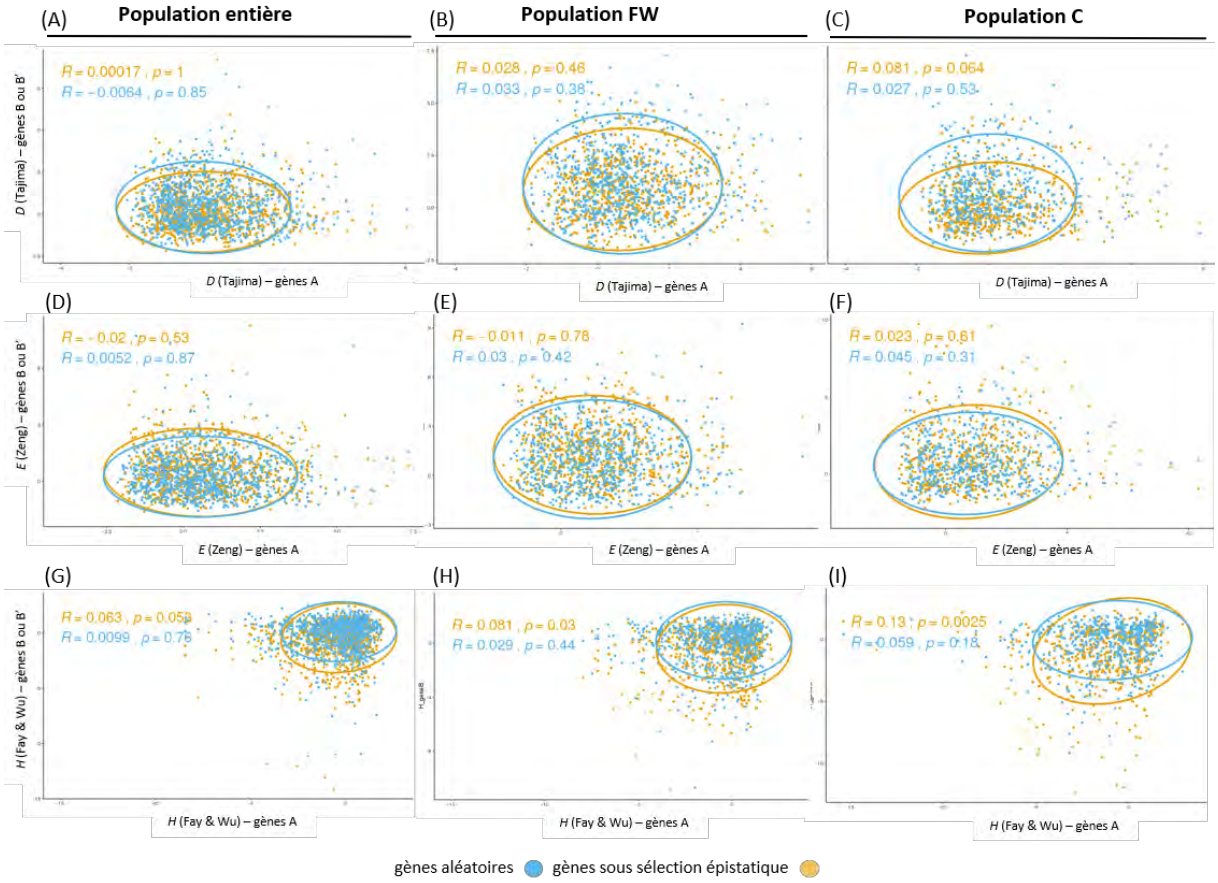
Enfin, les **Figures 36G,H,I** présentent les distributions du ***H*** de Fay & Wu dans les trois populations pour les deux ensembles *geneB* et *geneB'*. Dans l'ensemble de gènes aléatoires *geneB'*, les distributions sont globalement centrées autour de zéro comme nous l'avons observé sur les distributions à l'échelle du génome (moyenne de ***H*** sur l'ensemble *geneB'* en population FW, C et population entière = -0.10, -0.42 et -0.21; moyenne sur l'ensemble des gènes en population FW, C et population entière = -0.12, -0.52 et -0.18). En revanche, dans l'ensemble de gènes sous sélection épistatique, les distributions sont décalées vers des valeurs plus faibles (moyenne de ***H*** sur l'ensemble *geneB* en population FW, C et population entière = -0.63, -1.26 et -0.67) étant le signe ici aussi de signaux de balayages sélectifs.

Pour résumer, les distributions des statistiques de neutralité montrent qu'un grand nombre de gènes de *Medicago truncatula* qui présentent des signatures de sélection

épistatique présentent également des signatures de balayages sélectifs. En effet, dans la population entière, on constate que 30% des gènes sous sélection épistatique présentent également des traces de sélection (seuil empirique top 10% des gènes sous balayage sélectif). Parmi ces 30%, 15% sont sous balayage sélectif. Dans les populations FW et C, il y a respectivement 27% et 35% des gènes sous sélection épistatique qui présentent également des signatures de sélection dont 15% et 19% sont sous balayage sélectif. Il a été montré que les évènements de balayages sélectifs chez *Medicago truncatula* semblent plus anciens dans la population FW par rapport à la population C (Bonhomme et al., 2015) et comme cela a été dit dans le chapitre 1, ***E*** détecte plus facilement des balayages sélectifs anciens et après fixation. Ainsi ***E*** est plus sensible aux signaux de balayages sélectif sur les gènes en épistasie dans la population FW par rapport aux autres populations. Dans la population C, les évènements de balayage sélectif sur les gènes en épistasie semblent plus récents d'après la statistique ***H***, et le ***D*** de Tajima montre une différence plus importante entre les deux ensembles de gènes *geneB* et *gèneB'* car il y a probablement une part plus importante de SNP avec un excès d'allèles rares dans l'ensemble de gènes sous sélection épistatique. Ainsi, les statistiques de neutralité montrent que la sélection épistatique génère globalement, chez *Medicago truncatula,* des signaux de balayage sélectif. Malgré ces signaux, nous ne pouvons savoir lequel du modèle épistatique coadapté ou compensatoire permet la sélection entre des paires de gènes de *M. truncatula*. D'après les résultats des simulations, il semblerait que de tels signaux de sélection positive soient liés à un modèle de sélection épistatique coadapté en autogamie, mais il n'est pas exclu que le modèle compensatoire puisse être également lié à ces signaux. Cependant, d'après les résultats des simulations, les signatures de sélection ne sont pas fortement différenciées et il est difficile de distinguer les modèles de sélection épistatique du modèle de sélection additif en termes de signatures sur le polymorphisme.

La **Figure 37** présente les distributions conjointes des paires de gènes A-B et A-B' pour les statistiques D, H, et E: D_{AB}, H_{AB} et E_{AB} afin de rechercher s'il y a une corrélation des patrons de polymorphisme entre des gènes sous sélection épistatique et entre des gènes sélectionnés aléatoirement.

Figure 37 : Distributions conjointes des statistiques de neutralité calculées sur un ensemble de paires de gènes sous sélection épistatique et sur un ensemble de paires de gènes échantillonnés aléatoirement chez *Medicago truncatula* dans la population entière et dans les sous-populations Far-West et Circum. Les distributions conjointes des statistiques ***H*** de Fay & Wu, ***D*** de Tajima et ***E*** de Zeng entre les paires de gènes A - B et A – B' sont présentées dans chaque graphique. Les Figures (A, B, C) représentent les distributions conjointes de D_{AB} et $D_{AB'}$ dans les trois populations. Les figures (D, E, F) représentent les distributions conjointes de E_{AB} et $E_{AB'}$ dans les trois populations. Les figures (G, H, I) représentent les distributions conjointes de H_{AB} et $H_{AB'}$ dans les trois populations. L'axe des abscisses correspond aux valeurs de statistiques calculées sur les gènes A, l'axe des ordonnées aux valeurs calculées sur les gènes B sous sélection épistatique (jaune) et sur les gènes B' (bleu). Les ellipses sont calculées avec la fonction « *stat_ellipse*()» disponible dans la bibliothèque « *ggplot* » de R et permet de calculer une ellipse des données et suppose la distribution *t* multivariée.



Cette figure montre globalement que les corrélations entre les signatures de sélection sont faibles, voire inexistantes, que ce soit entre les gènes sous sélection épistatique (i.e. *geneA – geneB*, en jaune) ou les gènes aléatoires (i.e. *geneA - geneB'*, en bleu). Les corrélations entre les statistiques ***D***, ***E*** et ***H*** obtenues sur les paires *geneA – geneB'* aléatoires varient en 0.0052 et 0.059 selon les statistiques et ne sont pas significatives. Seule la statistique ***H*** montre des corrélations faibles mais significatives entre les paires de gènes épistatiques (*geneA - geneB*) dans les populations FW et C (r = 0.081 en population FW, r = 0.13 en population C). Les corrélations de ***H*** sont légèrement plus élevées en population C montrant que ***H*** détecte mieux la sélection au sein de cette sous-population ce qui va dans le sens des conclusions précédentes ; la sélection est plus récente dans la population C que dans la population FW et

H détecte mieux les évènements de sélection en cours. En revanche, les statistiques ***E*** et ***D*** ne montrent pas de corrélations significatives notamment dans la population FW où la sélection est plus ancienne et où nous aurions pu attendre un signal de sélection avec ***E***. Enfin, les figures en **Annexe 20** présentent les distributions conjointes $DH_{\overline{AB}}$, $DE_{\overline{AB}}$ et $HE_{\overline{AB}}$ des statistiques de neutralité au sein des trois populations. Les valeurs moyennes entre les paires *geneA-geneB* et *geneA-geneB'* sont calculées pour chaque statistique et les distributions conjointes entre ces statistiques ne montrent pas de différences très marquées entre ces groupes de gènes.

Nos résultats montrent qu'une proportion significative de gènes sous sélection épistatique (30%, 35% et 27% en population entière, C et FW) présente également des traces de sélection individuelles, dont la plupart sont des balayages sélectifs (15%, 19% et 15% en population entière, C et FW) mais les signatures de sélection chez *M. truncatula* ne semblent pas être corrélées entre les paires de gènes analysés. Ainsi les statistiques de neutralité ne permettent pas d'identifier des signatures de cosélection sur les gènes en épistasie. Ces signatures sont en effet fortement dépendantes des fréquences initiales des mutations (Tajima, 1989; Takahasi, 2009). Les résultats chez *Medicago truncatula* semblent montrer que l'on ne se situe pas dans un modèle de coadaptation avec des mutations « *de novo* » car cela créerait des corrélations plus fortes entre les gènes par une dynamique de cosélection plus longue. Les résultats semblent plutôt montrer que la coadaptation entre les gènes se fait à partir de mutations en « standing variation » c'est-à-dire en fréquences intermédiaires sur au moins un des deux gènes sous sélection épistatique. Ainsi nous pourrions imaginer un système avec un gène « recruteur » qui porte une mutation adaptée en fréquence intermédiaire dès le début de la sélection et un gène « recruté » dont la mutation adaptative est en faible fréquence au début de la sélection. Ce modèle de cosélection présenté dans la littérature par Takahasi (Takahasi, 2009; Takahasi & Tajima, 2005) induirait en effet de faibles corrélations entre les valeurs de statistiques de neutralité des paires de gènes cosélectionnés. La sélection positive induirait ainsi une signature de sélection plus forte sur le gène « recruté » si la mutation adaptative n'est pas en forte fréquence dès le début de la sélection tandis que le second gène présente déjà un taux de polymorphisme important au début de la sélection. Par exemple, les gènes *MtSUNN* et *MtCLE02* que nous avons identifiés comme potentiellement sous sélection épistatique (chapitre 2.2) pourraient suivre ce modèle. En effet, le polymorphisme de *MtSUNN* en population FW montre qu'il est sous sélection balancée avec ***H*** = 1.45 ce qui le classe parmi

les 8,41% des valeurs les plus élevées du génome sur les 48331 gènes testés. À l'inverse, la valeur de ***H*** de *MtSUNN* en population entière est de 0.52 le classant parmi les 39,5% les plus élevé du génome. De plus, le polymorphisme de *MtCLE02* semble montrer que ce gène est sous balayage sélectif à l'échelle de la population globale avec une valeur de ***H*** et -2.25 le classant parmi les 8.98% les plus faibles du génome. Il y a donc un maintien du polymorphisme sur l'un de ces deux gènes. La sélection épistatique pourrait agir suivant un modèle « gène recruteur - gène recruté » par un mécanisme compensatoire ou, plus probablement, suivant un modèle de coadaptation à partir de mutations en « standing variation » dans la population FW seulement.

Ainsi, les signatures de sélection seront fortement dépendantes des fréquences initiales des allèles aux locus soumis à la sélection épistatique. Les statistiques de neutralité peuvent apporter une information supplémentaire sur les gènes identifiés sous sélection épistatique et éventuellement sur les modèles de sélection. Parmi l'ensemble de gènes qui ont été analysé au cours de ce travail, le **Tableau 5** présente des gènes connus et caractérisés dans la littérature chez *M. truncatula* et pour lesquels nous avons identifié des signatures de sélection épistatique avec une ou plusieurs régions du génome mais également des signatures de sélection individuelle. Les gènes présentés dans le **Tableau 5** présentent au moins une signature de sélection épistatique (seuil de p-valeur à 10^{-11}) avec une autre région du génome sur un chromosome différent, ils font partie du top 10% des gènes de *Medicago truncatula* sous pression de sélection (balayage sélectif, sélection balancée, ou sélection purifiante) et ils font l'objet d'au moins une publication biologique chez *Medicago truncatula***.** Parmi cette liste, il y a une surreprésentation des gènes dont la fonction est liée à la nodulation car beaucoup de publications chez *Medicago truncatula* portent sur la nodulation. Ainsi, sur l'ensemble des gènes du **Tableau 5**, on peut citer notamment les récepteurs à domaine LysM, *MtLYR2*, *MtLYR3* et *NFP*, qui jouent un rôle important , comme d'autres récepteurs de ce type, dans la perception des signaux symbiotiques de type lipo-chitooligosaccharides (LCO) à l'origine du déclenchement du processus de nodulation (Arrighi et al., 2006). *MtLYR2* et *NFP* présentent des signatures de balayages sélectifs associées à des signatures de sélection épistatique avec d'autres régions du génome. Contrairement à *MtLYR2*, le balayage sélectif observé pour *NFP* est moins intense, comme observé lors d'une étude précédente (De Mita et al., 2007).

Tableau 5 : Exemple de gènes chez *Medicago truncatula* présentant une ou des signatures(s) de sélection épistatique avec une ou plusieurs régions du génome, et faisant partie du top 10% des gènes de *Medicago truncatula* sous pression de sélection (balayage sélectif, sélection balancée, ou sélection purifiante).

Population	Gene_id	Gene_Name	Annotation_Mt4	top_sweep_H	top_balancing_H	top_background_E	Biological_process	Reference
C	Medtr1g020020	MtCEP12	transmembraneproteinputative	66,8443	33,1578	9,9298	nodulation	PMID: 29030416
FW	Medtr1g021845	MtLYR2	LysM-domainreceptor-likekinase	0,6893	99,3128	99,0287	nodulation	PMID: 16844829
C	Medtr1g074370	MtERF1	ethylene-responsivetranscriptionfactor1B	5,2584	94,7436	90,9632	aphid resistance	PMID: 17249425
C	Medtr1g090807	MtRPG	myosinheavychain-likeproteinputative	7,2049	92,7972	83,9311	nodulation	PMID: 18621693
FW	Medtr1g094780	MtPAL2	phenylalanineammonia-lyase-likeprotein	5,9948	94,0073	87,6303	nodulation	PMID: 25527707
FW	**Medtr1g094960**	**MtARF16a**	**auxinresponsefactorputative**	**1,9530**	**98,0491**	**93,4245**	**nodulation**	**PMID: 25527707**
whole	**Medtr1g094960**	**MtARF16a**	**auxinresponsefactorputative**	**5,9442**	**94,0579**	**88,1345**	**nodulation**	**PMID: 25527707**
FW	Medtr1g101680	MtLYE2	peptidoglycan-bindingLysMdomainprotein	5,2867	94,7154	80,4136	nodulation	unpublished
C	Medtr1g106420	MtLBD1	LOBdomainprotein	3,3516	96,6505	92,3725	root response to salt stress	PMID: 21150260
whole	Medtr1g492760	MtDME	HhH-GPDbaseexcisionDNArepairfamilyprotein	6,6296	93,3725	77,7332	nodulation	PMID: 27797357
whole	Medtr2g097580	MtZIP2	ZIPmetaliontransporterfamilyprotein	72,7278	27,2742	7,2672	nodulation	PMID: 28732146
FW	Medtr2g100900	MtHPT5	histidinephosphotransferprotein	9,3452	90,6569	95,9457	nodulation	PMID: 31088345
C	**Medtr2g450070**	**MtRRB5**	**two-componentresponseregulatorARR12-likeprotein**	**2,5605**	**97,4416**	**93,4800**	**nodulation**	**PMID: 32376762**
whole	**Medtr2g450070**	**MtRRB5**	**two-componentresponseregulatorARR12-likeprotein**	**5,5724**	**94,4297**	**90,1699**	**nodulation**	**PMID: 32376762**
C	Medtr2g461240	MtKNOX10	classIIknotted-likehomeoboxprotein	8,4081	91,5940	94,1170	nodulation	PMID: 26351356
whole	**Medtr3g092780**	**MtPRR3**	**PRRresponseregulator**	**5,2339**	**94,7682**	**80,0469**	**nodulation**	**PMID: 32376762**
C	**Medtr3g092780**	**MtPRR3**	**PRRresponseregulator**	**6,4284**	**93,5737**	**75,9394**	**nodulation**	**PMID: 32376762**
whole	**Medtr3g099200**	**EPP1**	**hypotheticalprotein**	**64,9954**	**35,0066**	**0,8349**	**nodulation**	**PMID: 30476329**
C	**Medtr3g099200**	**EPP1**	**hypotheticalprotein**	**65,0415**	**34,9606**	**0,4705**	**nodulation**	**PMID: 30476329**
FW	**Medtr3g099200**	**EPP1**	**hypotheticalprotein**	**86,3854**	**13,6167**	**9,3556**	**nodulation**	**PMID: 30476329**
whole	Medtr3g106485	MtFLOT1	SPFH/band7/PHBdomainmembrane	4,7105	95,2916	88,9134	nodulation	PMID: 20018678
whole	**Medtr3g109610**	**MtCCD8**	**carotenoidcleavagedioxygenase**	**64,4201**	**35,5820**	**8,3991**	**mycorhization_nodulation**	**PMID: 26503135**
FW	**Medtr3g109610**	**MtCCD8**	**carotenoidcleavagedioxygenase**	**67,3399**	**32,6621**	**1,9363**	**mycorhization_nodulation**	**PMID: 26503135**
whole	Medtr3g110840	MtCRA2	LRRreceptor-likekinasefamilyprotein	83,1021	16,9000	7,8882	root, nodule development	PMID: 25521478
FW	Medtr3g117420	2OG-Fe(II)	2OG-Fe(II)oxygenasefamilyoxidoreductase	1,2700	98,7321	93,3201	root, nodule development	PMID: 25901015
FW	Medtr4g069850	MtRab7A2	RABGTPase-likeproteinA1D	9,0131	90,9890	87,6512	nodulation	PMID: 19734435
whole	Medtr4g094730	MtLYM2	LysMdomainGPI-anchoredprotein	0,2015	99,8006	99,8006	nodulation	PMID: 16844829
C	Medtr4g098870	MtRRB24	two-componentresponseregulator-APRR2-likeprotein	6,0641	93,9380	84,4245	nodulation	PMID: 32376762
FW	Medtr4g106990	MtSWEET4	bidirectionalsugartransporter	4,7770	95,2251	91,2230	nodulation	PMID: 27021190
C	**Medtr4g116545**	**MtKNOX9**	**classIIknotted-likehomeoboxprotein**	**4,5902**	**95,4119**	**90,7217**	**nodulation**	**PMID: 26351356**
whole	**Medtr4g116545**	**MtKNOX9**	**classIIknotted-likehomeoboxprotein**	**9,9672**	**90,0349**	**84,6951**	**nodulation**	**PMID: 26351356**
C	Medtr4g130800	MtENOD20	plastocyanin-likedomainprotein	73,5475	26,4546	7,6275	nodulation	PMID: 9526510
FW	Medtr5g011070	MtKNOX4	classIIknotted-likehomeoboxprotein	4,8543	95,1478	86,3624	nodulation	PMID: 26351356
FW	Medtr5g019040	MtNFP	Nod-factorreceptor5putative	7,8037	92,1984	90,1953	nodulation	PMID: 16844829
FW	Medtr5g019050	MtLYR3	LysM-domainreceptor-likekinase	90,5504	9,4517	50,9911	nodulation	PMID: 16844829
FW	Medtr5g096860	armadillo	armadillo/beta-catenin-likerepeatprotein	3,2292	96,7728	75,5467	root, nodule development	PMID: 25901015
FW	Medtr6g007160	MtN24	transmembraneproteinputative	4,2110	95,7911	89,0820	nodulation	PMID: 8634476
FW	Medtr6g007460	MtRRA7	responseregulatorreceiverdomainprotein	6,6005	93,4016	94,2705	nodulation	PMID: 32376762
whole	Medtr6g007637	MtSWEET5b	bidirectionalsugartransporter	8,1104	91,8917	84,3067	nodulation	PMID: 27021190
FW	**Medtr7g088830**	**Cam**	**Caminteractingprotein**	**5,7379**	**94,2642**	**90,1890**	**symbiosis**	**PMID: 25901015**
whole	**Medtr7g088830**	**Cam**	**Caminteractingprotein**	**8,0543**	**91,9477**	**79,3678**	**symbiosis**	**PMID: 25901015**
FW	Medtr7g118260	MtPRR6	PRRresponseregulator	7,4172	92,5849	86,5483	nodulation	PMID: 32376762
FW	Medtr8g024240	MtMTP1	heavymetaltransporterMTP1	8,3258	91,6762	90,3687	nodulation	PMID: 30042781
FW	Medtr8g042490	MtSWEET2a	bidirectionalsugartransporter	97,4329	2,5692	18,5253	nodulation	PMID: 27021190
FW	Medtr8g056900	MtSPIKE1	guaninenucleotideexchangefactorputative	4,8376	95,1645	80,4909	nodulation	PMID: 22683509
C	**Medtr8g097320**	**MtSYMREM1**	**carboxy-terminalregionremorin**	**5,4978**	**94,5042**	**91,3775**	**nodulation**	**PMID: 20133878**
whole	**Medtr8g097320**	**MtSYMREM1**	**carboxy-terminalregionremorin**	**6,8082**	**93,1939**	**93,7754**	**nodulation**	**PMID: 20133878**
whole	Medtr8g107360	MtPIN5	auxineffluxcarrierfamilytransporter	77,6522	22,3498	7,0366	nodulation	PMID: 15375694

Les colonnes “top_sweep_H”, “top_balancing_H” et “top_background_E” présentent le rang (en pourcentage) de chaque gène par rapport à l’ensemble des gènes 48331 analysés. Les valeurs indiquées en rouge représentent les gènes situés dans le top 10%

Citons également le gène *MtEPP1* (Early Phosphorylated Protein 1) qui est important pour l’initiation de la symbiose rhizobienne (Valdés-López et al., 2019) et qui présente des signatures de sélection purifiante importantes au sein des trois populations de *M. truncatula*. C’est l’unique gène de la liste à présenter une telle signature de sélection purifiante dans les trois populations, montrant qu’il est très conservé chez *Medicago truncatula*. Le gène *MtCCD8* (carotenoid cleavage dioxygenase 8) présente également une signature de sélection purifiante indiquant une conservation importante, particulièrement dans la population FW. *MtCCD8* est impliqué dans la synthèse des strigolactones et joue un rôle majeur dans le développement (ramification) ainsi que dans l’établissement de la symbiose mycorhizienne (Gomez-Roldan et

al., 2008; Wang et al., 2011). Enfin, nous pouvons mentionner un certain nombre de gènes de type « PRR response-regulator » ou « two-component response-regulator », *MtRRB5, MtPRR3*, *MtRRB24*, *MtRRA7* et *MtPRR6*, qui présentent des signatures de balayages sélectifs. Ces gènes, impliqués dans la signalisation cytokinine, jouent un rôle critique lors des étapes précoces de la nodulation en coordonnant l'organogénèse des nodules et la progression de l'infection par les bactéries (Tan et al., 2020). Les gènes présentés constituent une liste bien évidemment non exhaustive des gènes qui présentent des signatures de sélection épistatique avec d'autres régions du génome, et dont les valeurs de statistiques de neutralité se classent parmi les 10% les plus extrêmes du génome. Les statistiques de neutralité montrent qu'entre 27% et 30% des gènes de *M. truncatula* sous sélection épistatique présentent également des signatures de sélection individuelles. Ainsi, une grande majorité des gènes ciblés par la sélection épistatique (≈ 70%) présentent des profils de polymorphisme similaires aux gènes sous neutralité, ce qui montre que des gènes en apparence neutres peuvent en fait être soumis à la sélection épistatique. Ainsi il n'est pas envisageable de détecter la sélection épistatique en analysant les gènes seulement avec des statistiques de neutralité classique. Seule la corrélation génétique peut permettre d'identifier des signatures de sélection épistatique. Les signatures de sélection identifiées individuellement à chaque gène peuvent en revanche permettre de quantifier l'influence de la sélection épistatique sur le polymorphisme des gènes, et d'émettre des hypothèses quant aux différents modèles de sélection épistatique qui permettent la cofixation ou le maintien des allèles sélectionnés dans les populations.

2.4 Signatures génomiques de sélection épistatique chez M. truncatula

L'objectif de cette partie est de rechercher des signatures de sélection épistatique chez *Medicago truncatula* à l'échelle du génome. Le DL est calculé entre toutes les paires des 48333 gènes (fenêtres génomiques de 10kb) de *M. truncatula,* ce qui représente un total de $\frac{n*(n-1)}{2} = \frac{48\,333*(48\,333-1)}{2} = 1.17*10^9$ calculs. La statistique cor_{PC1v} a été utilisée pour mesurer le DL (voir **Figure 11** – méthode de calcul de cor_{PC1v}). Nous effectuons un test de corrélation (statistique $T_{cor_{PC1v}}$) pour mesurer la significativité de chaque valeur de corrélation et obtenir une p-valeur. Nous appliquons également une correction de Bonferroni sur les p-valeurs au seuil de 10^{-11} (i.e. correction de Bonferroni, pour α = 5% et $1.17*10^9$ tests). Ce seuil peut être relevé en fonction des objectifs d'analyses de ces données. Ces calculs ont généré une masse de données importante correspondant aux p-valeurs de tests ainsi qu'aux coefficients de corrélation cor_{PC1v} obtenus entre chaque paire de gènes. Cela correspond à $1.17*10^9$ valeurs de corrélation et p-valeurs de tests entre toutes les paires de gènes avant que nous ayons appliqué un filtre sur les p-valeurs significatives. Si l'on ne considère que les gènes et les interactions significatives au seuil de 10^{-11}, cela représente 801 400 interactions significatives en population FW, 961 460 interactions en population entière et 636 804 interactions en population C. Les temps de calculs permettant d'obtenir l'ensemble de ces résultats sont relativement longs. En effet, pour calculer d'abord les 48333 PC1 sur tous les gènes, cela a représenté 80 heures de calcul effectuées en parallèle. Ensuite, pour calculer l'ensemble des valeurs de corrélation et les p-valeurs de test, cela a représenté 215 heures de calculs également effectuées en parallèle et pour chacune des populations soit environ 8 heures pour chaque condition.

Il est important de noter que comme le DL a été calculé entre toutes les paires de gènes, nous identifions des paires significativement corrélées car les gènes sont proches physiquement et des paires significativement corrélées car les gènes sont distants, mais sous sélection épistatique ou sont cosélectionnés. Deux approches ont été mises en place pour initier l'exploration de la masse de données produite. Dans ces deux approches, le problème du DL « physique » s'est posé, et nous avons tenté d'en tenir compte. L'analyse de ces

résultats constitue une partie toujours exploratoire dont l'objectif serait d'identifier de nouvelles interactions adaptatives entre de nouveaux gènes.

La première approche est une approche par set de gènes candidats où nous comparons les valeurs de DL entre des gènes d'une même voie biologique, ou de même fonction moléculaire, à un set de gènes aléatoirement sélectionnés mais présentant les mêmes propriétés de distances physiques inter-locus que pour le set de gènes candidats. La seconde est une approche plus systémique, par l'analyse descriptive de réseaux d'interactions génétiques intégrant toutes les corrélations par paires de gènes significatives à un certain seuil, qu'elles soient intrachromosomiques et interchromosomiques, ou seulement interchromosomiques. Dans cette approche aussi, du fait de la masse de données générée, des sous-réseaux d'interactions sont analysés, incluant pour chaque paire de gènes au moins un gène caractérisé comme étant impliqué dans une voie biologique particulière. Ici, nous nous intéressons aussi à des gènes impliqués dans la symbiose racinaire rhizobienne et dans la symbiose racinaire mycorhizienne.

2.4.1 Approche exploratoire par l'analyse de sets de gènes candidats

Pour cette analyse, nous avons comparé les patrons de DL entre des sets de gènes candidats et des sets de gènes aléatoirement sélectionnés. Dans un premier temps, les gènes candidats que nous avons étudiés sont des gènes de mêmes voies biologiques, impliqués dans les symbioses rhizobienne et mycorhizienne et qui sont classés en deux catégories ; «symbiose RN » et « symbiose AM ». La symbiose RN correspond à la symbiose racinaire rhizobienne (« Rhizobial Nodule -RN- symbiosis ») avec les bactéries du genre *Rhizobium* fixatrices d'azote atmosphérique et impliquées dans la nodulation. La symbiose AM correspond à la symbiose mycorhizienne à arbuscule (« Arbuscular Mycorrhizae -AM- symbiosis ») avec les champignons endomycorhiziens à arbuscule qui améliorent la nutrition minérale des plantes. Une troisième catégorie, appelée « symbiose RN+AM » correspond à certains gènes des deux listes RN et AM impliqués dans les deux symbioses via la voie de signalisation commune, Common Symbiosis Signalling Pathway - CSSP – (Gough & Cullimore, 2011; Oldroyd, 2013). Les gènes appartenant à ces catégories biologiques ont des fonctions moléculaires variées, allant de la perception des micro-organismes symbiotiques, de la signalisation jusqu'à la mise en place au niveau des racines des structures symbiotiques permettant les échanges. Le **Tableau 6** présente ces gènes.

Tableau 6 : Liste des gènes candidats de *Medicago truncatula* impliqués dans les voies symbiotiques rhizobienne et mycorhizienne et analysés dans les approches exploratoires.

Gene_Name	ID_Mt4	Annotation_Mt4	Category	Candidate vs random analysis	Subnetworks analysis
PT4	Medtr1g028600	high affinity inorganic phosphate transporter	AM	oui	oui
RAM2	Medtr1g040500	Glycerol-3-Phosphate Acyl Transferase (GPAT)	AM	oui	oui
keto	Medtr3g085740	3-ketoacyl-(acyl-carrier) reductase	AM	oui	oui
EPP1	Medtr3g099200	hypothetical protein	AM	oui	oui
CCD8	Medtr3g109610	carotenoid cleavage dioxygenase	AM	oui	oui
fatX	Medtr3g111900	fatty acyl-CoA synthetase family protein	AM	oui	oui
MAX2	Medtr4g080020	F-box protein MAX2	AM	oui	oui
PT9	Medtr4g083960	phosphate transporter	AM	oui	oui
DH	Medtr4g097510	enoyl-(acyl carrier) reductase	AM	oui	oui
SCL3/RAD1	Medtr4g104020	GRAS family transcription factor	AM	oui	oui
STR2	Medtr5g030910	white-brown-complex ABC transporter family protein	AM	oui	oui
LFS	Medtr5g081780	polyketide cyclase/dehydrase and lipid transporter	AM	oui	oui
RAM1	Medtr7g027190	GRAS family transcription factor	AM	oui	oui
CCD7	Medtr7g045370	carotenoid cleavage dioxygenase	AM	oui	oui
TGL	Medtr7g081050	triacylglycerol lipase-like protein	AM	oui	oui
SC	Medtr7g105460	short-chain dehydrogenase/reductase	AM	oui	oui
DXS2	Medtr8g068265	1-deoxy-D-xylulose-5-phosphate synthase	AM	oui	oui
GDSL	Medtr8g074560	GDSL-like lipase/acylhydrolase	AM	oui	oui
STR	Medtr8g107450	white-brown-complex ABC transporter family protein	AM	oui	oui
MtCEP12	Medtr1g020020	transmembraneproteinputative	RN	oui	oui
NOOT2	Medtr1g051025	BTB/POZankyrinrepeatprotein	RN	oui	oui
NF-YA1	Medtr1g056530	CCAAT-binding transcription factor	RN	oui	oui
MtCNGC15a	Medtr1g064240	cyclicnucleotide-gatedionchannel-likeprotein	RN	oui	oui
RPG	Medtr1g090807	myosin heavy chain-like protein, putative	RN	oui	oui
MtDME	Medtr1g492760	Putative_DNA-(apurinic_or_apyrimidinic_site)_lyase	RN	oui	oui
MtDML1	Medtr2g008920	Putative_DNA_glycosylase,_helix-turn-helix,_base-excision_DNA_repair	RN	oui	oui
MtCEP17	Medtr2g016800	transmembraneproteinputative	RN	oui	oui
NPR2	Medtr2g028950	regulatoryproteinNPR1	RN	oui	oui
NPR1	Medtr2g039880	regulatoryproteinNPR1	RN	oui	oui
nfya3	Medtr2g041090	CCAAT-binding transcription factor	RN	oui	oui
MtCNGC15c	Medtr2g094860	cyclicnucleotide-gatedionchannel-likeprotein	RN	oui	oui
MtDAS18	Medtr2g103570	putative peptidase S26B, peptidase S24/S26A/S26B/S26C	RN	oui	oui
MtDNF1	Medtr3g027890	putative signal peptidase complex subunit 3	RN	oui	oui
MtDAS12	Medtr3g085510	putative microsomal signal peptidase 12kDa subunit	RN	oui	oui
MtCRA2	Medtr3g110840	LRRreceptor-likekinasefamilyprotein	RN	oui	oui
MtCNGC15b	Medtr4g058730	cyclicnucleotide-gatedionchannelproteinputative	RN	oui	oui
MtSUNN	Medtr4g070970	LRRreceptor-likekinasefamilyprotein	RN	oui	oui
MCM1	Medtr4g071090	putative mini-chromosome maintenance protein	RN	oui	oui
MtCLE13	Medtr4g079610	Clavata3/ESR(CLE)genefamilymemberMtCLE13	RN	oui	oui
MtCLE12	Medtr4g079630	Clavata3/ESR(CLE)genefamilymemberMtCLE12	RN	oui	oui
MCM2	Medtr4g096700	putative mini-chromosome maintenance protein	RN	oui	oui
MCM3	Medtr4g116870	putative mini-chromosome maintenance protein	RN	oui	oui
MCM4	Medtr4g122270	putative mini-chromosome maintenance protein	RN	oui	oui
MtSYMREM2	Medtr5g010590	carboxy-terminal region remorin	RN	oui	oui
MtCEP5	Medtr5g017710	hypotheticalprotein	RN	oui	oui
LYR3	Medtr5g019050	LysM-domain receptor-like kinase	RN	oui	oui
MtCEP4	Medtr5g025730	hypotheticalprotein	RN	oui	oui
MtCEP10	Medtr5g030490	hypotheticalprotein	RN	oui	oui
MtDAS25	Medtr5g081900	putative signal peptidase complex subunit 2	RN	oui	oui
PUB1	Medtr5g083030	ubiquitin-protein ligase, PUB17	RN	oui	oui
NPR3	Medtr5g090770	NPR1/NIM1-likeregulatoryproteinputative	RN	oui	oui
MtDML2	Medtr5g095880	Putative_DNA_glycosylase,_helix-turn-helix,_base-excision_DNA_repair	RN	oui	oui
NIN	Medtr5g099060	nodule inception protein	RN	oui	oui
ERN2	Medtr6g029180	AP2 domain class transcription factor	RN	oui	oui
ERN1	Medtr7g085810	AP2 domain class transcription factor	RN	oui	oui
NOOT1	Medtr7g090020	BTB/POZankyrinrepeatprotein	RN	oui	oui
NF-YA2	Medtr7g106450	CCAAT-binding transcription factor	RN	oui	oui
MtCEP11	Medtr8g072170	transmembraneproteinputative	RN	oui	oui
MtCEP14	Medtr8g087390	transmembraneproteinputative	RN	oui	oui
MCM5	Medtr8g090000	putative mini-chromosome maintenance protein	RN	oui	oui
MtSYMREM1	Medtr8g097320	carboxy-terminal region remorin	RN	oui	oui
MtCHK1_MtCRE1	Medtr8g106150	cytokininreceptorhistidinekinase	RN	oui	oui
NUP85	Medtr1g006690	Nup85 nucleoporin protein	AM + RN	oui	oui
DMI1/POLLUX	Medtr2g005870	ion channel pollux-like protein	AM + RN	oui	oui
DELLA-1	Medtr3g065980	DELLA domain GRAS family transcription factor GAI	AM + RN	oui	oui
NSP2	Medtr3g072710	GRAS family transcription factor	AM + RN	oui	oui
enod11	Medtr3g415670	transmembrane protein, putative	AM + RN	oui	oui
NFP	Medtr5g019040	Nod-factor receptor 5, putative	AM + RN	oui	oui
ipd3/cyclops	Medtr5g026850	cyclops protein, putative	AM + RN	oui	oui
DMI2	Medtr5g030920	nodulation receptor kinase-like protein	AM + RN	oui	oui
LYK3	Medtr5g086130	LysM receptor kinase K1B	AM + RN	oui	oui
NUP133	Medtr5g097260	Nup133/Nup155-like nucleoporin	AM + RN	oui	oui
VAPYRIN	Medtr6g027840	ankyrin repeat RF-like protein, putative	AM + RN	oui	oui
NENA	Medtr6g072020	nucleoporin seh1-like protein	AM + RN	oui	oui
CASTOR	Medtr7g117580	ion channel castor	AM + RN	oui	oui
NSP1	Medtr8g020840	GRAS family transcription factor	AM + RN	oui	oui
MtCCaMK_DMI3	Medtr8g043970	calmodulin-domain kinase CDPK protein	AM + RN	oui	oui
GA2OX6_GA2OX2	Medtr1g086550	plant gibberellin 2-oxidase	symbiosis_(misc)	non	oui
TF80	Medtr3g022830	GRAS family transcription factor	symbiosis_(misc)	non	oui

SOBIR1_like	Medtr3g075440	LRR receptor-like kinase family protein	symbiosis_(misc)	non	oui
LYK9	Medtr3g080050	LysM receptor kinase K1B	symbiosis_(misc)	non	oui
LYR9	Medtr4g058570	LysM-domain receptor-like kinase	symbiosis_(misc)	non	oui
SHR	Medtr4g097080	GRASfamilytranscriptionfactor	symbiosis_(misc)	non	oui
SNF1	Medtr5g018050	sucrose nonfermenting 1(SNF1)-related kinase	symbiosis_(misc)	non	oui
LYR8	Medtr5g042440	LysM-domain receptor-like kinase	symbiosis_(misc)	non	oui
MtCLE45_sweep	Medtr5g056935	Clavata3/ESR(CLE)genefamilymember	symbiosis_(misc)	non	oui
LYR4	Medtr5g085790	LysM-domain receptor-like kinase	symbiosis_(misc)	non	oui
LYK4	Medtr5g086120	LysM receptor kinase K1B	symbiosis_(misc)	non	oui
LYK2	Medtr5g086310	LysM receptor kinase K1B	symbiosis_(misc)	non	oui
MtCLE02	Medtr6g009390	Clavata3/ESR(CLE)genefamilymemberMtCLE02	symbiosis_(misc)	non	oui
feronia	Medtr6g015805	feronia receptor-like kinase	symbiosis_(misc)	non	oui
LYR6	Medtr7g07932	LysM type receptor kinase	symbiosis_(misc)	non	oui
LYR5	Medtr7g079350	LysM type receptor kinase	symbiosis_(misc)	non	oui
LYK11	Medtr8g014500	LysM type receptor kinase	symbiosis_(misc)	non	oui
LEA_P	Medtr8g046000	embryonic abundant-like protein	symbiosis_(misc)	non	oui
LYR1	Medtr8g078300	Nod-factor receptor 5, putative	symbiosis_(misc)	non	oui
TF124	Medtr8g442410	GRAS family transcription factor	symbiosis_(misc)	non	oui
lipoprot_attach	Medtr8g464760	membrane lipoprotein lipid attachment site-like protein putative	symbiosis_(misc)	non	oui

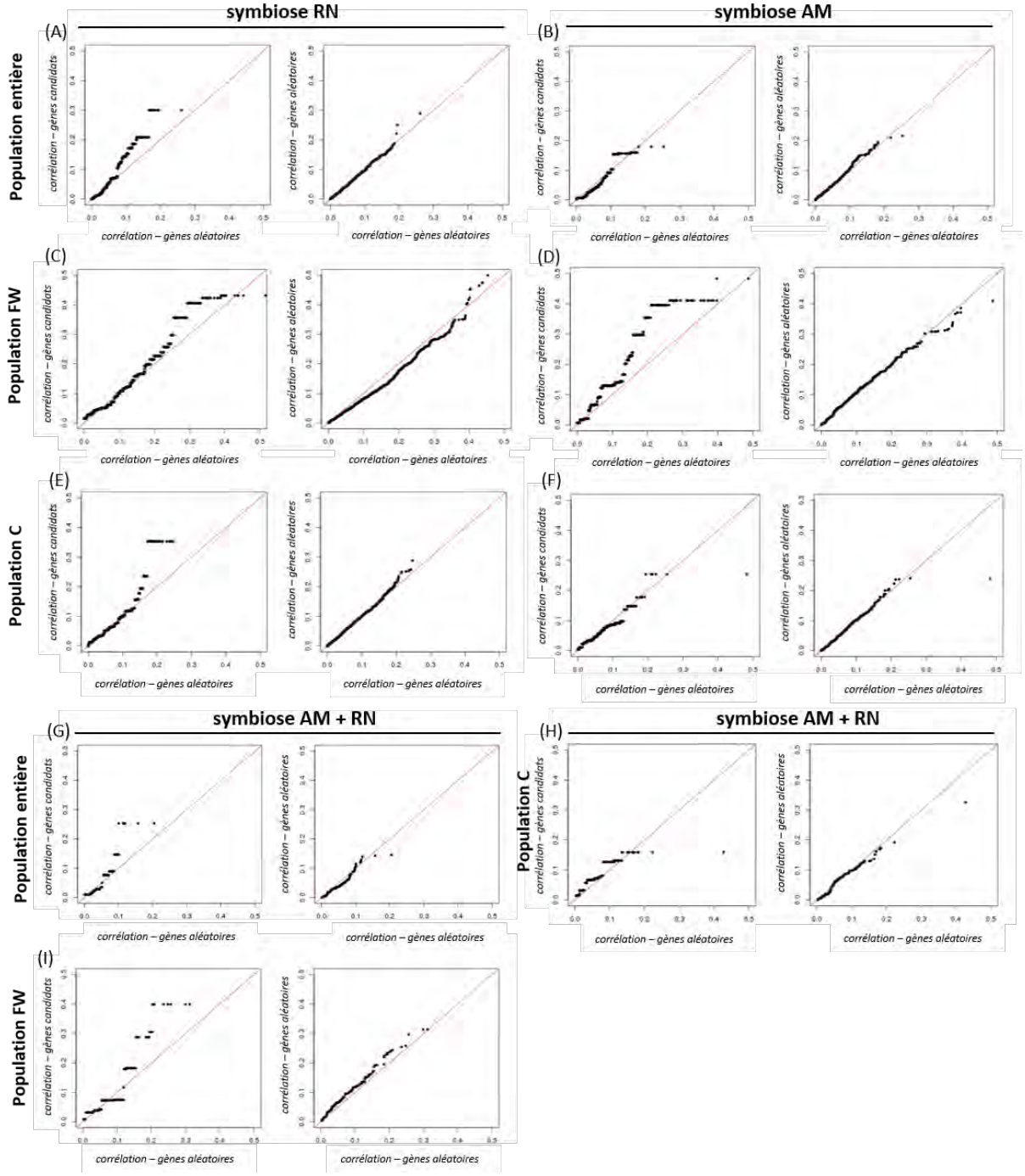
Ensuite, nous avons comparé les patrons de DL sur des sets de gènes de même fonctions moléculaires, tels que l'ensemble des facteurs de transcription (différentes familles) et les gènes de la famille des RLK (Receptor-Like kinase) impliqués dans la détection de signaux (biotiques ou abiotiques) de l'environnement pour réguler l'expression des gènes qui sont impliqués dans le développement et la croissance, dans la réponse de défense des plantes face aux pathogènes et dans les interactions symbiotiques chez les légumineuses (Afzal et al., 2008; Chiu & Paszkowski, 2020; Dievart et al., 2020; Ye et al., 2017). Nous avons également testé les gènes des ilots symbiotiques. Ces ilots correspondent à des régions génomiques comprenant chacune plusieurs gènes impliqués dans le développement du nodule lors de la symbiose rhizobienne et dont les patrons d'expression et les marques épigénétiques (méthylation et modification des histones) indiquent que les gènes d'un même ilot sont corégulés et qu'un grand nombre d'ilots fonctionnent de manière synchronisée (Pecrix et al., 2018).

Les statistiques de DL ont été calculées entre toutes les paires de gènes candidats d'un ensemble testé et entre un nombre équivalent de paires de gènes aléatoires chez *Medicago truncatula*. Les paires de gènes aléatoires sont échantillonnées dans tout le génome de telle manière que la distribution des distances inter-locus et des paires interchromosomiques soit équivalente à celle des paires de gènes candidats. L'objectif de cet échantillonnage est de réduire les biais du DL physique qu'il pourrait y avoir parmi l'ensemble de gènes candidats, en ajustant la même proportion de DL physique entre les deux ensembles de gènes candidats et aléatoires.

2.4.1.1 Analyse de gènes candidats de même voies biologiques

La **Figure 38** présente les résultats de la comparaison de la distribution du DL entre les gènes candidats impliqués dans les symbioses RN, AM, ou RN+AM, avec celle des gènes aléatoires, dans les trois populations de *M. truncatula*. Les gènes de symbiose RN présentent un excès de fortes valeurs de DL par rapport aux gènes aléatoires dans les trois populations (**Figures 38A,C,E**). Cet ensemble de gènes de symbiose est constitué de 58 gènes dont 43 sont impliqués exclusivement dans la symbiose RN et 15 font partie de la voie CSSP (catégorie « symbiose RN+AM »). De plus, les **Figures 38A,C,E** qui représentent le DL entre deux ensembles de gènes aléatoires indiquent qu'il n'y a globalement pas de différence, montrant ainsi que les gènes impliqués dans la symbiose rhizobienne présentent des valeurs de DL globalement plus fortes qui ne sont pas liées à des associations par hasard mais pouvant être liées à de la sélection épistatique ou à de la cosélection au sein des trois populations de *M. truncatula*.

Figure 38 : Q-Q plots des distributions du DL entre les sets de gènes candidats impliqués dans les symbioses rhizobienne (RN) et mycorhizienne (AM) et des sets de gènes aléatoires, dans les trois populations de *Medicago truncatula*. Les distributions de la statistique cor_{PC1v} des paires de gènes aléatoires (axe des abscisses) sont comparées aux distributions de cor_{PC1v} entre les paires de gènes candidats (axe des ordonnées) ou en fonction d'un second ensemble de paires de gènes aléatoires (axe des ordonnées). Les Figures A, C et E impliquent les gènes candidats de la symbiose RN (i.e. les gènes RN exclusivement et les gènes RN du groupe RN+AM). Les Figures B, D et F impliquent les gènes candidats de la symbiose AM (i.e. les gènes AM exclusivement et les gènes AM du groupe RN+AM). Les Figures G, H et I impliquent les gènes RN+AM stricts (i.e. les gènes de la voie de signalisation commune « CSSP »).

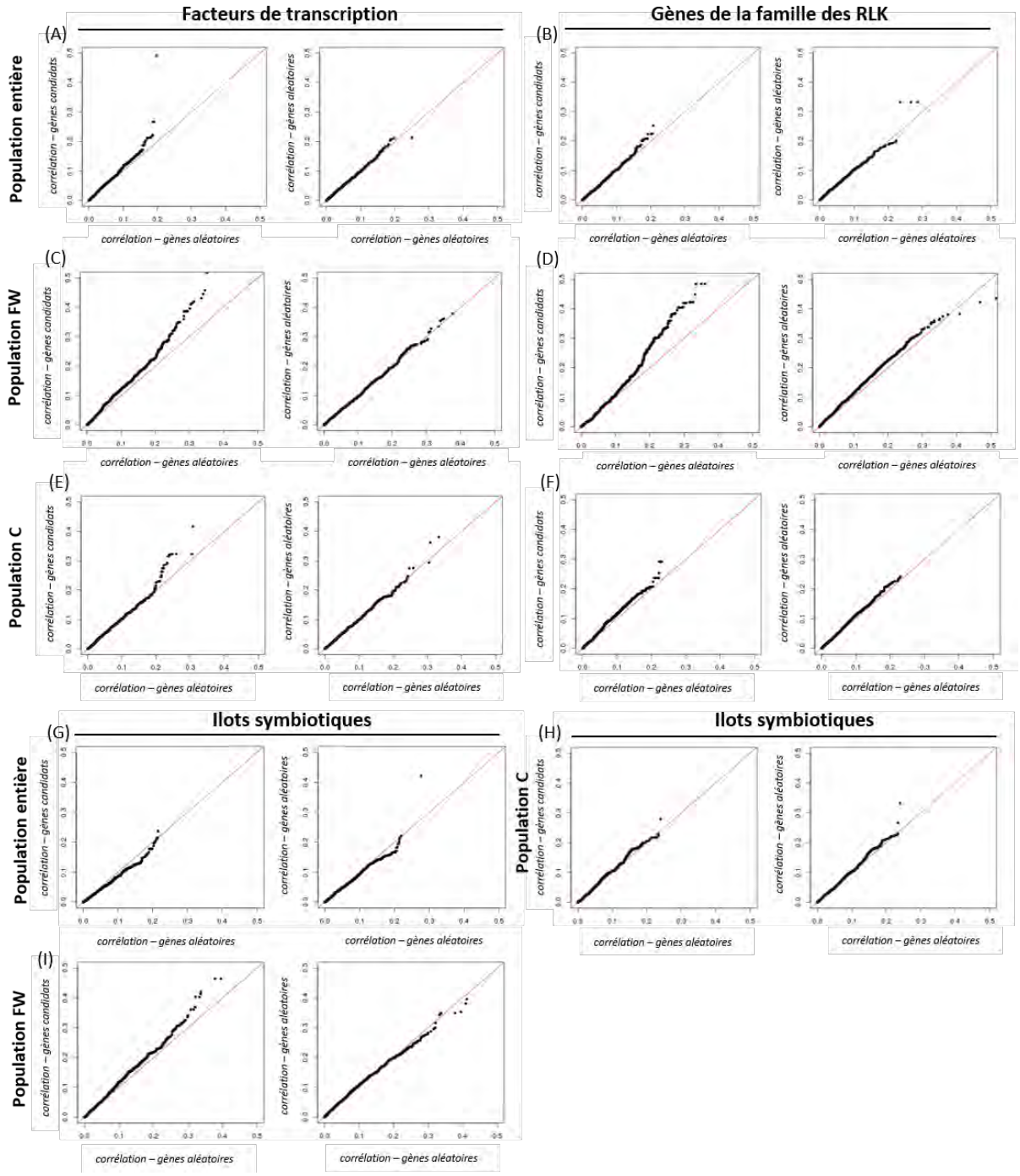


Les gènes impliqués dans la symbiose AM montrent un excès de fortes valeurs de DL par rapport aux gènes aléatoires uniquement au sein de la population FW (**Figure 38D**). Pour la population C et la population entière, les distributions entre les gènes candidats et les gènes

aléatoires sont globalement semblables aux distributions obtenues avec les deux ensembles de gènes aléatoires (**Figure 38B,F**). Cet ensemble de gènes de symbiose est constitué de 34 gènes dont 19 sont impliqués exclusivement dans la symbiose AM et 15 font partie de la voie CSSP (catégorie « symbiose RN+AM »). Enfin, le groupe de 15 gènes impliqué dans les deux symbioses (« symbiose RN+AM ») présente un excès de valeurs fortes de DL par rapport aux gènes aléatoires, au niveau de la population entière et de la population FW et dans une moindre mesure au niveau de la population C.

Les gènes impliqués dans la symbiose RN semblent donc montrer des signaux de sélection épistatique ou de cosélection sur l'ensemble de l'espèce *M. truncatula*, alors que les gènes impliqués dans la symbiose AM ne semblent montrer des signaux de sélection épistatique que dans la population FW. De plus, des signaux de sélection épistatique sont détectés dans le set de gènes de la voie CSSP (« symbiose RN+AM ») à l'échelle de l'espèce, mais de manière plus prononcée dans la population FW. Du point de vue de la détection de ces signaux, il est important de remarquer que si un signal est détecté dans la population entière, il est aussi assez bien détecté dans les deux-sous-populations (exemple de la symbiose RN). Cependant, si aucun signal n'est détecté en population entière, il reste néanmoins possible de détecter un signal dans l'une des deux populations (exemple de la symbiose AM). Ceci indique que cor_{PC1v}, en corrigeant pour la structure des populations, a tendance à détecter des signaux de sélection épistatique à l'échelle globale de l'espèce, tandis que les signaux corrélés à la structure des populations (i.e. adaptation locale) ne peuvent être détectés qu'en analysant les sous-populations séparément. Enfin, en comparant les patrons de DL entre ces ensembles de gènes candidats par rapport à des ensembles de gènes échantillonnés aléatoirement, nous montrons que des gènes impliqués dans une même voie biologique ou des voies biologiques communes peuvent présenter un enrichissement en signaux de sélection épistatique, traduisant ainsi le rôle important que peuvent jouer les interactions génétiques dans une même voie biologique.

Figure 39 : Q-Q plots des distributions du DL entre les sets de gènes candidats de même fonction moléculaire et des sets de gènes aléatoires, dans les trois populations de *Medicago truncatula*. Les distributions de la statistique cor_{PC1v} des paires de gènes aléatoires (axe des abscisses) sont comparées aux distributions de cor_{PC1v} entre les paires de gènes candidats (axe des ordonnées) ou en fonction d'un second ensemble de paires de gènes aléatoires (axe des ordonnées). Les Figures A, C et E impliquent des gènes candidats codant pour des facteurs de transcription. Les Figures B, D et F impliquent les gènes candidats codant pour des protéines de la famille des RLK (« *Receptor-Like kinase* »). Les Figures G, H et I impliquent des gènes candidats des ilots symbiotiques.



2.4.1.2 *Analyse de gènes candidats de même fonctions moléculaires*

La **Figure 39** présente les résultats de la comparaison de la distribution du DL entre les gènes codant pour des facteurs de transcription, les gènes codant pour des RLK, ou appartenant aux ilots symbiotiques, avec celle des gènes aléatoires, dans les trois populations

de *M. truncatula*. Comme dans la sous-partie précédente, les gènes aléatoires ont été échantillonnés dans tout le génome de telle manière que la distribution des distances inter-locus et des paires interchromosomiques soient équivalentes à celle des paires de gènes candidats. Chez *M. truncatula*, environ 2 800 gènes ont été identifiés comme codant pour des facteurs de transcription. Les distributions du DL représentées sur les **Figures 39A,C,E** ne se basent pas sur l'ensemble des comparaisons de toutes les paires de ces gènes (i.e. ~ 3.9 millions de comparaisons) mais nous avons échantillonné 1 500 paires de gènes afin de réduire les temps de calcul pour l'échantillonnage des gènes aléatoires et pour la représentation des Q-Q plots. Ainsi, les distributions du DL parmi les gènes codant pour des facteurs de transcription semblent montrer un excès de valeurs fortes par comparaison avec les ensembles de gènes aléatoires, dans les trois populations de *M. truncatula*. Cette signature, bien que présente dans l'ensemble des populations, est toutefois plus marquée dans la population FW. Les **Figures 39B,D,F** présentent les comparaisons des distributions du DL entre un ensemble de gènes candidats appartenant à la famille des RLK et un ensemble de gènes aléatoires. Nous avons là aussi échantillonné 1 500 paires de gènes de la famille des RLK parmi les 790 gènes (i.e. ~ 312 000 comparaisons). Les distributions du DL montrent un excès de valeurs fortes parmi les gènes de la famille des RLK au sein de la population FW uniquement. Enfin, les **Figures 39H,I,J** présentent les résultats de la comparaison du DL entre les gènes des ilots symbiotiques et un ensemble de gènes aléatoires. 1 500 paires de gènes ont également été échantillonnées parmi les 7 200 gènes des ilots symbiotiques (i.e. ~ 25.9 millions de comparaisons). Les résultats montrent aussi un excès de valeurs fortes de DL uniquement au sein de la population FW sur l'ensemble des gènes candidats.

L'ensemble de ces résultats montre qu'il est possible d'identifier des signaux de sélection épistatique, ou de cosélection, entre des gènes de mêmes voies biologiques ou qui ont des fonctions moléculaires communes. Certains signaux peuvent cependant être spécifiques d'une population. Par exemple, la spécificité d'enrichissement en signaux d'épistasie pour les gènes de la symbiose AM dans la population FW pourrait traduire des pressions de sélection liées à l'environnement biotique et/ou abiotique de cette population, ou bien un relâchement de la pression de sélection dans la population C. Cependant, cette approche visant à tester l'enrichissement en signaux de sélection épistatique via les valeurs de corrélation (i.e. de DL) entre gènes candidats est restrictive et ne nous permet pas

d'explorer et d'identifier, de manière statistiquement robuste et à l'échelle du génome, de nouvelles interactions significatives entre des paires de gènes candidats spécifiques ou des paires de gènes impliquant un gène candidat et un nouveau gène non caractérisé, comme cela a été fait pour les gènes *MtSUNN* et *MtCLE02*. Afin d'explorer plus largement les interactions génétiques façonnées par la sélection épistatique, nous proposons (i) d'initier une description des réseaux d'interactions génétiques à l'échelle du génome, en intégrant toutes les corrélations par paires de gènes significatives à un certain seuil, mais aussi (ii) de nous concentrer sur la description de sous-réseaux d'interactions ancrés sur des voies biologiques particulières, en choisissant là aussi les gènes des symbioses rhizobienne et mycorhizienne ainsi que les gènes avec lesquels ils sont en interactions significatives.

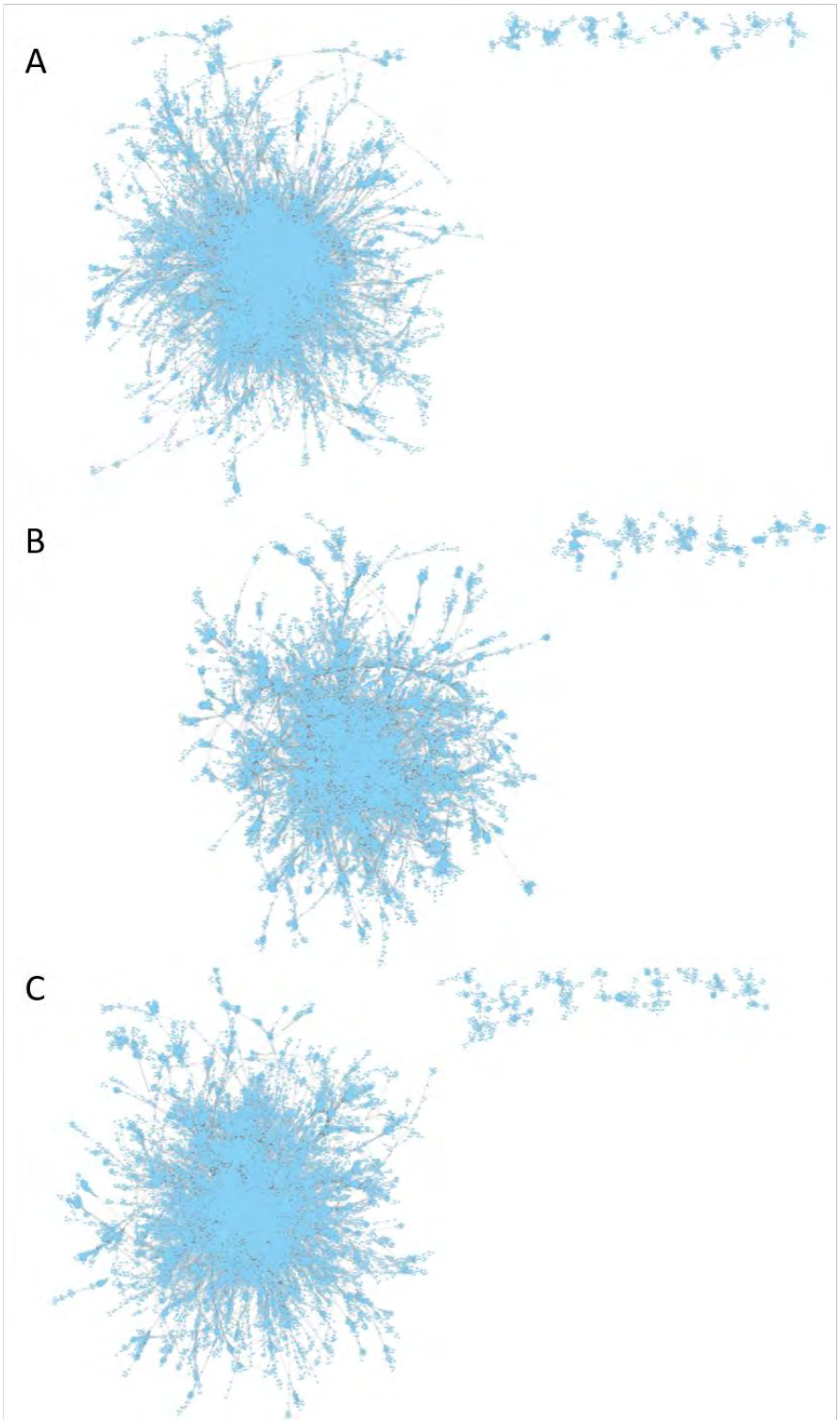
2.4.2 Approche systémique par l'analyse de réseaux génomiques d'interactions entre gènes

2.4.2.1 Description générale des réseaux génomiques d'interactions génétiques et de leurs propriétés

Pour analyser l'ensemble des résultats de DL produits à l'échelle du génome de *Medicago truncatula*, nous avons utilisé une approche réseau. La **Figure 40** présente les réseaux d'interactions construits avec les résultats de DL calculés entre chaque paire de gènes du génome de *M. truncatula* dans la population entière et les populations FW et C. Les réseaux ont été construits avec le logiciel Cytoscape (Shannon et al., 2003), les nœuds correspondent à des gènes et deux gènes sont reliés par une arête si la p-valeur du test de corrélation (statistique $T_{cor_{PC1v}}$) est inférieure au seuil de 10^{-11}. Un filtre supplémentaire est ajouté afin de réduire le bruit de fond au sein des réseaux ; seuls les nœuds (i.e. gènes) possédant au moins trois arêtes significatives au seuil fixé sont conservés. De cette façon, tous les gènes qui ne possèdent qu'une ou deux interactions significatives au seuil de 10^{-11} ne sont pas représentés et cela permet d'éliminer certains gènes. De plus, le DL est calculé entre toutes les paires de gènes de *M. truncatula* impliquant à la fois des interactions longue distance entre paires de gènes pouvant être liées à de la sélection épistatique, mais aussi de la liaison physique entre des gènes proches produisant également du DL. Pour construire ces réseaux, les interactions intrachromosomiques et interchromosomiques ont été représentées sans qu'il n'y ait de filtre sur le DL physique (intrachromosomique). Ainsi lorsque nous représentons le DL à l'échelle globale dans les trois populations, cela produit des réseaux denses avec un nombre de nœuds très important, et constitués d'une composante connexe principale

comportant la plupart des gènes et des petites composantes connexes annexes. Une composante connexe est définie comme un sous-graph connexe maximal c'est-à-dire un ensemble de points qui sont reliés deux à deux par un chemin.

Figure 40 : Réseaux génomiques d'interactions génétiques (au moins trois arêtes par gènes, seuil de significativité à 10^{-11} pour les interactions deux-à-deux) chez *M. truncatula* (A, B, C: population entière, FW et C).



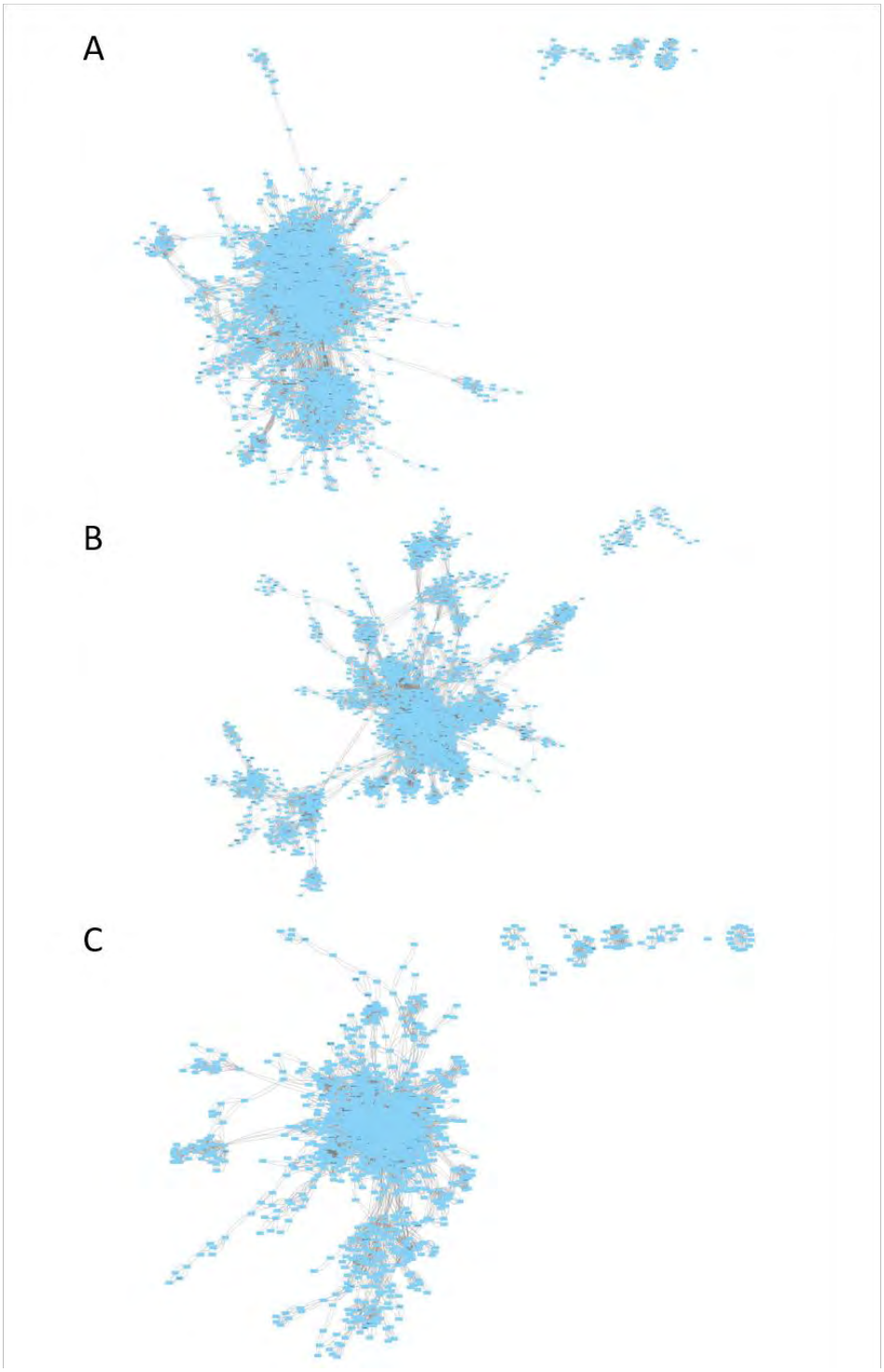
Le réseau construit en population entière est constitué de 44 207 gènes reliés par 961 460 arêtes, le réseau en population FW est constitué de 46 015 gènes reliés par 801 398 arêtes et le réseau en population C est constitué de 45 282 gènes pour 936 804 arêtes. Ainsi, les réseaux complets tels qu'ils sont présentés sont difficilement analysables et interprétables même lorsque l'on ne considère que les interactions significatives au seuil de 10^{-11}, qui est un seuil très conservatif.

Un moyen de réduire la complexité de ces réseaux en grande partie due au DL physique est de ne conserver que les interactions interchromosomiques. La **Figure 41** présente ces mêmes réseaux mais ne représentant que les interactions interchromosomiques. Nous avons également conservé le critère d'au moins trois interactions significatives pour un gène afin de réduire le bruit de fond. Les réseaux ainsi produits sont constitués de 5 207 gènes en population entière reliés par 43 772 arêtes, 4 568 gènes en population FW reliés par 42 114 arêtes et 3058 gènes en population C pour 30 222 arêtes. Le réseau interchromosomique permet de réduire les biais dus au DL physique afin de mieux mettre en évidence les interactions longues distances générées très probablement par de la sélection épistatique. On peut ainsi plus facilement observer des clusters d'interactions. Mais malgré ce filtre appliqué sur les interactions, les réseaux restent très denses et il est difficile de pouvoir les analyser afin d'identifier des interactions spécifiques ou des groupes de gènes particulièrement reliés. De plus, ce filtre que nous avons utilisé en ne conservant que les interactions interchromosomiques est imparfait car d'une part, nous avons potentiellement éliminé des interactions évolutives entre des gènes d'un même chromosome, éloignés ou non physiquement et d'autre part, nous n'avons pas complètement éliminé les interactions liées au DL physique.

Pour illustrer ce dernier point, la **Figure 42** représente un sous-réseau construit à partir du réseau initial de la population FW mais qui est constitué d'un sous-ensemble de gènes (gènes de symbiose) et leurs voisins dans le réseau (voir le détail de l'analyse dans la partie suivante). Aucun filtre sur le nombre d'arêtes par gène n'est appliqué dans ce sous-réseau (**Figure 42A**). Par comparaison avec la **Figure 42B** où le nombre d'arêtes par gène est de deux au minimum, cette représentation permet d'illustrer comment le DL physique influence la structure des réseaux d'interaction génétique, même à longue distance. Ainsi, lorsque l'on

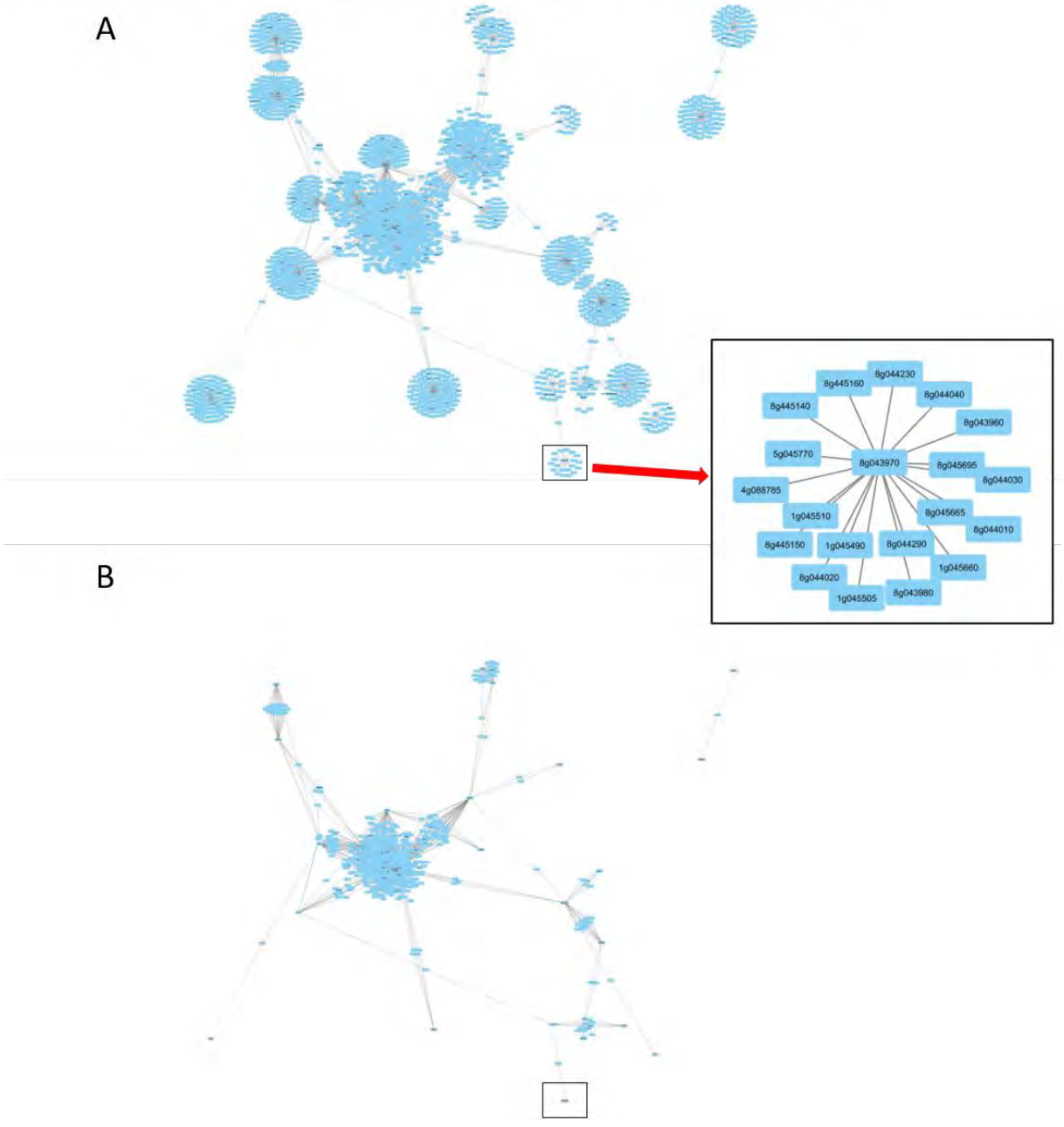
conserve l'ensemble des interactions (**Figure 42A**), le réseau a une topologie avec des formes en étoile.

Figure 41 : Réseaux génomiques d'interactions génétiques interchromosomiques (au moins trois arêtes par gènes, seuil de significativité à 10^{-11} pour les interactions deux-à-deux) chez *M. truncatula* (A, B, C: population entière, FW et C).



Ces formes particulières décrivent en fait le DL physique entre un gène et ses voisins mais également entre deux gènes intrachromosomiques et leurs voisins. Le zoom dans la **Figure 42A** décrit une de ces structures en étoile montrant les interactions (au seuil de 10^{-11}) entre un gène du chromosome 8 (8g043970) et 19 autres gènes situés sur le chromosome 8 mais également sur les chromosomes 1, 4 et 5.

Figure 42 : Illustration de la visualisation du déséquilibre de liaison d'origine physique dans un sous-réseau génomique d'interactions génétiques (i.e. gènes symbiotiques) dans la population FW de *M. truncatula* (A: au moins une arête par gène et zoom sur une structure en étoile, B au moins deux arêtes par gène).



Les interactions entre le gène 8g043970 et les gènes situés sur le chromosome 8 dont les identifiants sont 8g043960, 8g043980, 8g044010, 8g044020, 8g044030, 8g044040, 8g044230 et 8g044290, sont liées au DL physique (distances entre 5kb et 186kb du gène 8g043970). En revanche, les gènes 8g445140, 8g445150 et 8g445160, situés à plus de 300 kb du gène central 8g043970, sont probablement des gènes dont le DL avec 8g043970 a moins de chance d'être généré par une liaison physique, mais plutôt par des interactions longue distance intrachromosomiques dues à la sélection épistatique. En filtrant le sous-réseau sur un nombre d'arêtes par gène ≥ 2 (**Figure 42B**, voir la zone du gène 8g043970 encadré), nous ne visualisons plus le DL physique avec les gènes proches, ni le DL généré par des interactions significatives avec des gènes éloignés sur le même chromosome, ni même l'effet du DL physique sur les interactions à l'échelle interchromosomique, qui lui est intéressant en termes de détection. En effet, si l'on considère les gènes du chromosome 1 dans le zoom de la **Figure 42A** ; 1g045490, 1g045505 et 1g045510 et 1g045660, on remarque qu'ils sont significativement reliés au gène 8g043970 mais ces quatre gènes sont eux-mêmes physiquement proches les uns des autres (distance maximale de 73 kb). Ainsi, il pourra être fréquent d'identifier du DL entre un gène et plusieurs autres gènes proches entre eux, sur un autre chromosome. Dans ce cas, il ne sera pas possible de savoir lequel de ces gènes est significativement en interaction adaptative avec le premier gène. La résolution de l'interaction se fera donc à l'échelle d'une région génomique de plusieurs kb. Cependant, ce cas de figure a l'avantage d'accroître la capacité de détection de la sélection épistatique à l'échelle interchromosomique, car le DL ne se limite pas uniquement aux deux gènes sous sélection mais « s'étend » aussi aux gènes voisins.

La prise en compte du DL généré par la liaison physique pour mieux détecter le DL généré par la sélection épistatique entre gènes distants sera un enjeu important pour l'analyse des réseaux que nous avons construits. Les réseaux d'interactions réalisés avec toutes les interactions significatives, ou seulement avec les interactions significatives interchromosomiques, ou bien en filtrant sur le nombre minimum d'arêtes par gène, constituent une première base pour l'exploration des données produites. Ils pourraient permettre d'identifier des gènes coadaptés ou des groupes de gènes significativement reliés, même si ces réseaux restent imparfaits. En perspective, différentes approches pourraient être mises en place pour tenir compte de la liaison physique qui influence les valeurs de DL, afin

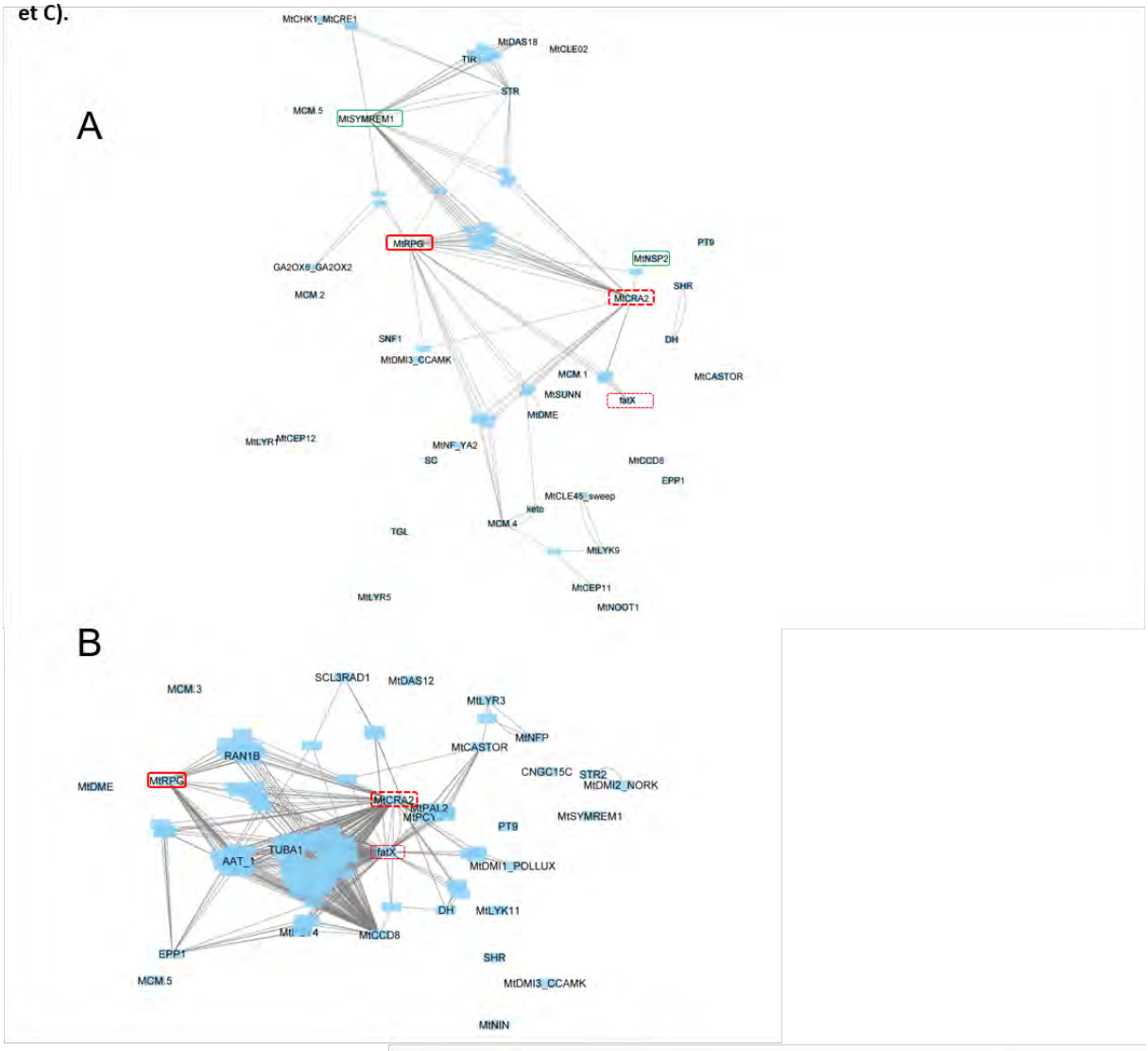
de mieux mettre en évidence les interactions dues à la sélection épistatique. Une approche par « blocs » pourrait être envisagée avec, par exemple, la méthode de « adjacency-constrained clustering » - *adjclust* - (Neuvial et al., 2017) qui a été développée dans la cadre des analyses de GWAS et de Hi-C pour construire des « blocs » le long des chromosomes. Avec cette approche, nous pourrions délimiter des blocs de DL au sein des chromosomes de *M. truncatula*, recalculer les ACP sur ces blocs et extraire de nouveaux vecteurs PC1 pour calculer les corrélations entre blocs. De cette façon, chaque nouvelle fenêtre serait considérée comme indépendante physiquement et nous pourrions considérer de fortes valeurs de DL inter-bloc comme résultant de l'effet de la sélection épistatique. Par ailleurs, une seconde approche de modélisation du DL pourrait être explorée, où, quelle que soit la distance entre deux locus le DL pourrait être décomposé en un DL physique « moyen » estimé par rééchantillonnage (bootstrap) sur une paire de locus, et une composante résiduelle qui correspondrait à du DL maintenu par sélection épistatique.

2.4.2.2 Sous-réseaux génomiques d'interactions ancrés sur des gènes symbiotiques

Afin d'analyser une partie des données produites à l'échelle du génome et identifier des gènes ou groupes de gènes coadaptés, nous avons extrait des sous-réseaux d'interaction à partir du réseau global, comportant un sous-ensemble de 98 gènes candidats impliqués dans les symbioses racinaires chez *M. truncatula* (symbiose AM, RN et RN+AM). Ces sous-réseaux ont été extraits dans les trois populations de *M. truncatula* et ils présentent les interactions significatives au seuil de 10^{-8} entre des gènes candidats et leurs voisins. Nous avons choisi un seuil de significativité moins conservatif par rapport au seuil de 10^{-11} car cela nous permet de représenter plus d'interactions et notamment des interactions avec des gènes connus dont la p-valeur du test de corrélation se situe entre 10^{-11} et 10^{-8}. La liste des gènes est disponible dans le **Tableau 6**. Ces sous-réseaux extraits du réseau global représentent les interactions intra et interchromosomiques et nous avons également conservé le critère d'au moins deux ou trois interactions significatives pour un gène afin de réduire le bruit de fond. Les sous-réseaux ainsi générés sont présentés pour un critère d'au moins trois interactions par gènes dans la **Figure 43**, et pour un critère d'au moins deux interactions par gènes en **Figure 44**. Les Figures en **Annexe 21 et 22** représentent respectivement ces mêmes sous-réseaux mais les gènes (i.e. les nœuds) ne sont pas désignés par leurs noms, lorsque celui-ci est connu, mais

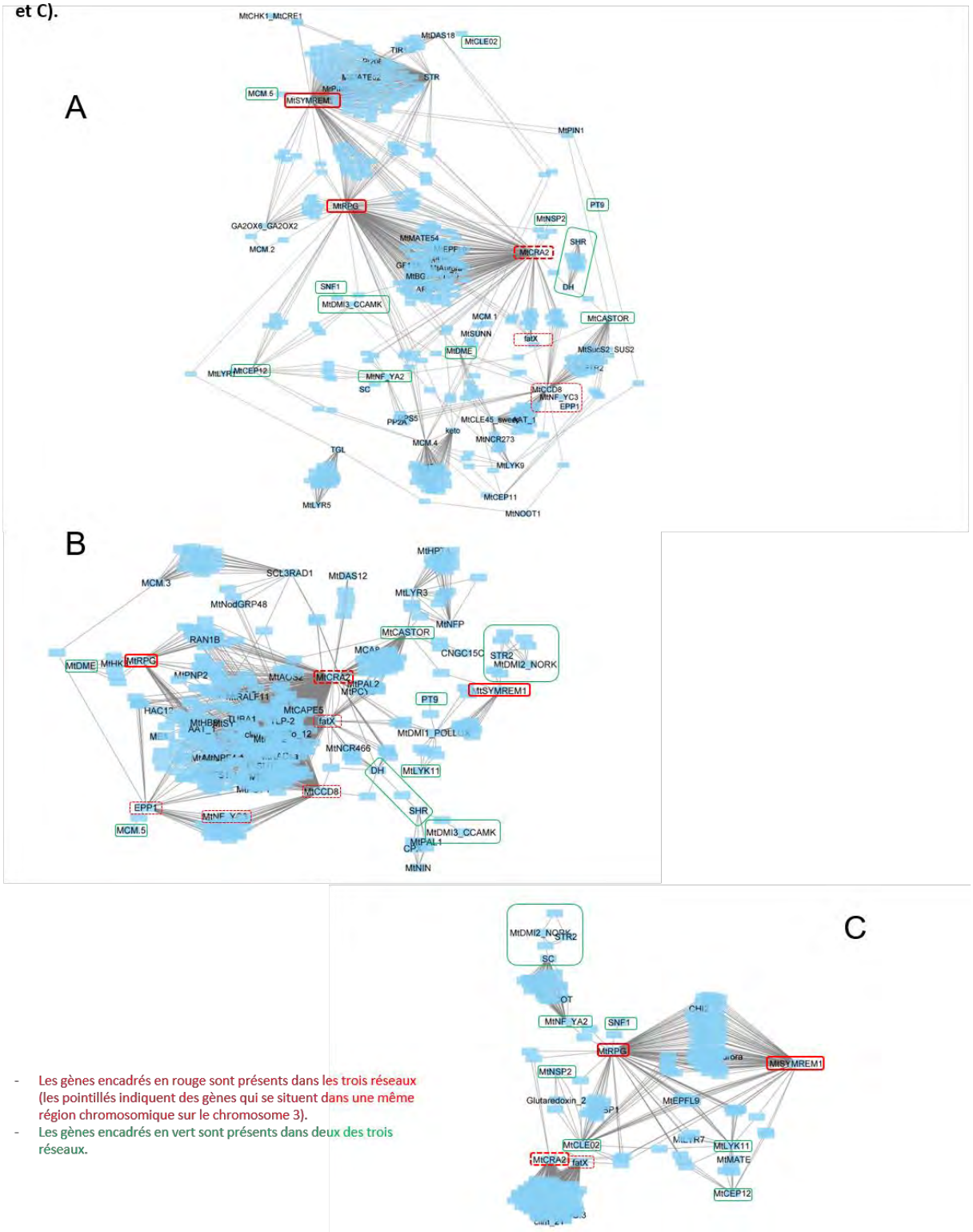
par leurs identifiants chromosomiques afin d'illustrer sur quels chromosomes sont situés les gènes et ainsi différencier les interactions intra et interchromosomiques.

Figure 43 : Sous-réseaux génomiques d'interactions génétiques (au moins trois arêtes par gènes) pour des gènes impliqués dans les symbioses rhizobiennes et mycorhiziennes chez *M. truncatula* (A, B, C: population entière, FW et C).



- Les gènes encadrés en rouge sont présents dans les trois réseaux (les pointillés indiquent des gènes qui se situent dans une même région chromosomique sur le chromosome 3).
- Les gènes encadrés en vert sont présents dans deux des trois réseaux.

Figure 44 : Sous-réseaux génomiques d'interactions génétiques (au moins deux arêtes par gènes) pour des gènes impliqués dans les symbioses rhizobiennes et mycorhiziennes chez *M. truncatula* (A, B, C: population entière, FW et C).



- Les gènes encadrés en rouge sont présents dans les trois réseaux (les pointillés indiquent des gènes qui se situent dans une même région chromosomique sur le chromosome 3).
- Les gènes encadrés en vert sont présents dans deux des trois réseaux.

Les sous-réseaux extraits des réseaux globaux montrent les interactions significatives impliquant au moins un gène candidat connu (et parfois deux gènes candidats connus) ; ainsi avec cette représentation nous pouvons identifier de potentielles nouvelles interactions épistatiques. Lorsque l'on compare les sous-réseaux entre les **Figures 43 et 44**, nous pouvons voir que le filtre de deux ou trois interactions par gène influence la complexité des sous-réseaux. Avec le filtre de trois arêtes par gènes, une partie du bruit de fond est supprimé mais nous perdons également certains gènes connus et certaines interactions qui peuvent être intéressantes. Ces gènes perdus apparaissent « isolés » dans la **Figure 43**, afin de pouvoir comparer avec la **Figure 44**. Avec le filtre de deux interactions par gène, les sous-réseaux en populations entière, FW et C sont respectivement constitués de 3 831, 3 153 et 2 153 gènes reliés par 3 556, 2 882 et 1 246 arêtes. Parallèlement, avec le filtre de trois interactions par gènes, les sous-réseaux en population entière, FW et C sont respectivement constitués de 73, 149 et 38 gènes reliés par 124, 410 et 90 arêtes. Ces sous-réseaux montrent qu'un certain nombre de gènes connus sont significativement reliés de façon directe ou indirecte au sein des différentes populations et ils peuvent avoir des positions « centrales » (ou « hub ») vis-à-vis des autres gènes du sous-réseau. Parmi ces gènes, il y a notamment les gènes *MtRPG*, *MtCRA2* et *MtSYMREM1*, qui sont plus ou moins directement reliés dans les sous-réseaux, qui se situent tous les trois sur des chromosomes différents et qui apparaissent dans les sous-réseaux des trois populations. Ces trois gènes possèdent un degré de connexion important au sein de ces sous-réseaux. Par exemple, pour la population entière (**Figure 44**), 183 arêtes intrachromosomiques (dont la moitié sont longue distance) et 166 arêtes interchromosomiques relient *MtRPG*, dont 112 ciblent la région de *MtCRA2* sur le chromosome 3 (dont *MtCRA2* directement, mais aussi d'autres gènes symbiotiques importants situés dans cette région, comme *MtCCD8* et *MtEPP1*). 283 arêtes intrachromosomiques (dont 279 ciblent des gènes dans une région de ~22Mb incluant *MtCRA2*) et 55 arêtes interchromosomiques relient *MtCRA2*, dont 29 ciblent la région de *MtRPG* (dont *MtRPG* directement, nécessairement) sur le chromosome 1. Enfin, 364 arêtes intrachromosomiques partent de *MtSYMREM1* (chromosome 8) et ciblent les gènes situés dans une région de ~10Mb incluant *MtSYMREM1* - dans laquelle 6 gènes distants de 400 kb

de *MtSYMREM1* ciblent directement *MtCRA2* sur le chromosome 3. 96 arêtes interchromosomiques relient également *MtSYMREM1*, dont 67 ciblent la région incluant *MtRPG* (sur 14 Mb) sur le chromosome 1. Ces trois gènes sont localisés dans des régions de fort DL sur trois chromosomes différents. Ces régions sont en interactions significatives entre elles mais aussi avec d'autres régions génomiques. Le gène *MtCRA2* (compact root architecture 2, Medtr3g110840) code pour une protéine de type Leucin-Rich Repeat Receptor-Like Kinase (LRR-RLK) qui régule positivement le nombre de nodules pendant la symbiose rhizobienne (Laffont et al., 2019), et le gène *MtRPG* (rhizobium-directed polar growth, Medtr1g090807) code pour une protéine qui est également impliquée dans la symbiose rhizobienne en contrôlant l'infection par *Rhizobium.* Avec l'approche réseau, nous avons pu retrouver l'interaction entre *MtRPG* et *MtCRA2* que nous avions identifiée avec l'approche appât (voir chapitre 2.2.2.2). Le gène *MtSYMREM1* (symbiotic remorin1, Medtr8g097320) code pour une protéine de type « receptor-binding protein » qui est requise pour l'infection par *Rhizobium* et pour la nodulation. Il a été montré que *MtSYMREM1* interagit directement avec les récepteurs symbiotiques de type LysM-RLK tels que *MtNFP*, *MtLYK3* et *MtDMI2* à la manière d'une protéine échafaudage (Chiu & Paszkowski, 2020; Lefebvre et al., 2010). De manière intéressante, nous pouvons remarquer que *MtSYMREM1* est très proche de *MtDMI2* et dans une moindre mesure de *MtNFP*, dans le sous-réseau de la population FW. Ceci n'est cependant pas observé dans les sous-réseaux de la population entière et de la population C, indiquant que cette configuration particulière d'interaction est probablement spécifique de la population FW. Il n'en reste pas moins que *MtSYMREM1* semble avoir une position importante dans chacun des sous-réseaux (**Figure 44**).

D'autre part, un certain nombre d'autres gènes symbiotiques apparaissent dans les sous-réseaux d'au moins deux des trois populations montrant qu'il y a un degré important de connectivité parmi ces gènes. Parmi eux, certains ne sont présents que dans une des sous-populations (FW ou C) et en population entière, ce qui peut s'expliquer par un manque de puissance de détection dans une des deux sous-populations. C'est le cas notamment pour les gènes *MtDMI3_CCAMK*, *MtCASTOR*, *MtDME* (DEMETER, Satgé, 2016), *MtSHR*/*MtDH* et *PT9* (population FW et population entière) et pour les gènes *MtNF-YA2* (Laloum et al., 2014) et *MtNSP2* (population C et population entière). Il est intéressant de noter que la plupart des gènes et voies de signalisation symbiotiques caractérisées chez *Medicago truncatula* sont issus d'analyses effectuées sur des mutants du génotype de référence A17 qui appartient à la

population FW. De fait, nous retrouvons uniquement au sein du sous-réseau de la population FW un certain nombre de gènes qui appartiennent à la voie CSSP (*MtNFP*, *MtDMI1*) ainsi que des gènes spécifiques de la nodulation (*MtNIN*) ou à la mycorhization (*MtRAD1*), ce qui pourrait traduire un phénomène de coadaptation locale entre certains gènes symbiotiques. La faible superposition et donc la faible similarité des sous-réseaux issus des sous-populations FW et C montrent que les interactions génétiques qui sous-tendent les mécanismes symbiotiques (et notamment la nodulation) peuvent varier au sein d'une même espèce en fonction des mécanismes d'adaptation locale mais on ne peut exclure un manque de résolution dans l'une des populations. Toutefois, ces sous-réseaux semblent partager des composantes génétiques (voir par exemple le cas des gènes *MtRPG*, *MtCRA2* et *MtSYMREM1*), dont certaines ne sont peut-être pas détectées, et qui constituent probablement une ossature commune.

La représentation des résultats de DL à l'échelle de sous-réseaux incluant des gènes candidats caractérisés permet de visualiser les interactions génétiques dans une ou des voies biologiques particulières, ce qui n'est pas possible avec le réseau global. Nous avons pu visualiser certaines interactions entre gènes symbiotiques, comme *MtCRA2 vs MtRPG*, et *MtSYMREM1* vs *MtDMI2*. Nous avons pu également visualiser des connexions entre les gènes symbiotiques via des interactions intermédiaires avec d'autres gènes, ce qui ouvre la voie vers la caractérisation biologique de nouveaux gènes. En comparaison avec l'approche appât (présentée en chapitre 2.2) qui teste les interactions génétiques à une dimension, l'approche réseau explore les résultats issus d'analyses génomiques en deux dimensions ($\frac{n(n-1)}{2}$, avec n = nombre de gènes) et nous pouvons identifier de multiples interactions potentielles, directes ou indirectes, entre gènes ou groupes de gènes.

Synthèse et perspectives

1. Méthodologie statistique et simulations

L'objectif du premier chapitre a été de développer un test statistique pour détecter la sélection épistatique à l'aide du DL en tenant compte de la structure des populations et de l'apparentement entre les individus. Cet objectif a été atteint en proposant les statistiques T_{r_v} et T_{corPC1_v}, respectivement basées sur la comparaison de paires de SNP et de paires de fenêtres génomiques. Des simulations de données SNP à l'échelle de génomes diploïdes en populations structurées, couplées à des calculs intensifs de DL, ont permis de démontrer que T_{r_v} et T_{corPC1_v} réduisent fortement le bruit de fond de DL généré par les forces évolutives non sélectives. T_{r_v} et T_{corPC1_v} suivent une distribution de Student $\tau_{(n-2)}$ sous l'hypothèse nulle d'absence de corrélation entre les locus testés et présentent une bonne puissance de détection. Les simulations ont également permis de comprendre les dynamiques évolutives de plusieurs régimes de sélection, l'effet de la structure des populations, de deux régimes de reproduction et plusieurs modes d'interactions entre les allèles.

1.1 Evolution des allèles des SNP simulés

L'évolution des fréquences alléliques des SNP simulés dans les différents modèles montrent que la sélection sur deux locus selon les modèles coadapté (épistatique) et additif, tend à fixer les allèles cosélectionnés positivement, comme dans les cas de sélection positive sur un seul locus. Nous avons montré que la cofixation est plus rapide dans le modèle de sélection épistatique coadapté que dans le modèle de sélection additive, indépendante entre les deux locus. Par ailleurs, le modèle compensatoire tend à maintenir le polymorphisme aux 2 locus car la sélection se fait sur les haplotypes *ab* et *AB* et les deux allèles ségrégent dans la plupart des simulations (~60% des cas) pendant les 300 générations. Dans le cas de populations structurées, la sélection épistatique selon un modèle compensatoire peut aussi amener à la cofixation de *AB* ou *ab* dans une des sous-populations, ou les deux sous-populations. Ainsi, la sélection épistatique compensatoire peut conduire à la fixation des allèles cosélectionnés mais elle peut également permettre le maintien du polymorphisme. Les simulations réalisées dans les différents modèles nous ont également permis d'évaluer les effets de l'interaction entre les systèmes de reproduction et le degré de dominance des allèles sélectionnés. Ainsi, dans les populations autofécondes, les taux de fixation des haplotypes sélectionnés (*ab* dans le modèle coadapté, et *ab* ou *AB* dans le modèle compensatoire) est

inchangé quel que soit le degré de dominance des allèles (i.e. récessif – dominant – codominant). Ces résultats sont connus dans les modèles à un locus mais nous l'avons étendu à des modèles à deux locus en épistasie (**Figure 19**) (Glémin & Ronfort, 2013). En revanche, dans les populations panmictiques, le degré de dominance dans les simulations influence les taux de cofixation et ces derniers sont plus élevés lorsque les allèles sont codominants ou récessifs. Là aussi, ce phénomène est connu dans les modèles à un locus pour des mutations codominantes (Glémin & Ronfort, 2013) mais les faibles taux de cofixation que nous avons obtenus pour des mutations dominantes contrastent avec ce qui est connu dans les modèles à un locus. En effet, il a été montré qu'en panmixie de nouvelles mutations dominantes ont plus de chances d'être fixées que des mutations récessives car la sélection agit surtout sur les génotypes hétérozygotes. Ce phénomène appelé « tamis de Haldane » (Haldane, 1927) prédit que l'adaptation dans les grandes populations panmictiques se fait préférentiellement via la fixation de mutations dominantes. Si la mutation adaptative qui apparaît est récessive, elle a peu de chances d'être sélectionnée car il faut, pour cela, qu'elle soit très souvent à l'état homozygote. Dans le cas de nos simulations, nous trouvons un très faible taux de cofixation des mutations dominantes et un taux plus élevé pour les mutations récessives. Ces observations ne sont probablement pas dues aux modèles épistatiques mais au fait que les mutations sélectionnées sont en « standing variation » dès le début des simulations. En panmixie, les individus hétérozygotes sur les locus cosélectionnés sont de ce fait très nombreux et la dominance va donc permettre le maintien de l'allèle non sélectionné ce qui va ainsi freiner la fixation de la mutation sélectionnée. Ce phénomène est aussi accentué par l'effet de la dérive génétique car les populations simulées sont de petite taille et l'allèle non sélectionné peut être en forte fréquence au début du processus de sélection.

Maintenant, si l'on compare les modèles de reproductions, nous constatons que les taux de cofixation sont plus rapides en autogamie qu'en panmixie, même si la fixation peut être moins efficace au final car les taux à l'équilibre sont plus faibles. Ces résultats corroborent là aussi, ceux des modèles de sélection à un locus (Glémin & Ronfort, 2013) mais nous les étendons aux modèles épistatiques. L'efficacité de sélection dépend du coefficient de sélection *s* et de la taille de la population. Des analyses de simulations pilotes nous ont amenés à fixer $s = 0.1$ du fait des tailles des populations simulées ($N = 250$ individus par sous-population). Ainsi, l'efficacité de sélection dépend de $N_e s$ avec N_e la taille efficace de la

population qui est directement reliée au mode de reproduction (Glémin & Ronfort, 2013) car $N_e = N/(1 + F_{IS})$, avec $F_{IS} = \sigma/(2 - \sigma)$ et σ le taux d'autofécondation (M Nordborg, 2000; Pollak, 1987). Ces correspondances montrent pourquoi la fixation est moins efficace en autogamie car la taille efficace est plus faible. Les populations autofécondes à 95% présentent une efficacité de sélection qui sera ainsi réduite d'environ 50% avec une augmentation de la dérive. En perspective, il serait intéressant de simuler différentes valeurs du coefficient de sélection s, à condition de simuler des tailles de population plus grandes telles que $N = 1000$ dans la mesure où ces simulations seraient réalisables en termes de temps de calcul.

1.2 Evolution du déséquilibre de liaison

Nous avons simulé des populations où la structure et l'apparentement génèrent du DL entre des locus indépendants qui évoluent sous neutralité. Ce DL de fond influence fortement les mesures r et cor_{PC1}, mais nous avons montré qu'il est possible de le réduire lorsque l'on utilise les mesures r_v et cor_{PC1v} traduisant l'importance d'utiliser la correction par la matrice *V*. Les simulations que nous avons menées sont originales, dans le sens où nous avons généré délibérément de la structure génétique et différents degrés d'apparentement afin d'estimer leurs effets sur le DL dans un modèle neutre et dans des modèles de sélection épistatique. Les études précédentes, sur lesquelles nous nous sommes basés pour ce travail et qui recherchent des signatures de sélection épistatique à partir du DL, ont simulé des données génétiques sans structure et se sont limitées à un régime panmictique (Takahasi, 2009; Takahasi & Innan, 2008; Takahasi & Tajima, 2005).

Les simulations ont montré que les modèles de sélection épistatique et additive génèrent plus de DL que le modèle neutre, comme l'avaient précédemment montré Takahasi & Innan, 2008. En autogamie, tous les modèles de sélection épistatique génèrent du DL, mais il est difficile de distinguer ces modèles du modèle additif avec l'approche haplotypique (cor_{PC1} et cor_{PC1v}). Les comparaisons directes des SNP sous sélection (r, r_v) permettent de mieux différencier les modèles épistatiques du modèle additif. En panmixie, seul le modèle compensatoire génère significativement plus de DL. Les statistiques r et r_v semblent, ainsi, plus efficaces pour détecter la sélection épistatique que les statistiques cor_{PC1} et cor_{PC1v} car ces dernières sont sensibles à l'hétérogénéité haplotypique. De plus, les évènements de recombinaison dans les fenêtres génomiques étudiées vont tendre à réduire la puissance de

détection mais ce phénomène peut être limité en fixant des tailles de fenêtre petites (i.e. ≤ 10kb). Malgré cela, certains locus peuvent présenter des taux de recombinaison élevés qui vont diminuer le DL entre SNP d'une même fenêtre et augmenter ainsi l'hétérogénéité des haplotypes sélectionnés. La taille des fenêtres génomiques peut toutefois être fixée en fonction des valeurs de LD-decay de l'espèce étudiée. En revanche, l'avantage des statistiques cor_{PC1} et cor_{PC1v} est qu'elles prennent en compte le polymorphisme des SNP environnants et donc l'effet d'auto-stop ce qui permet de détecter un signal même après la fixation des mutations sélectionnées. Enfin, cor_{PC1} et cor_{PC1v} sont rapides à implémenter lorsque l'on teste des centaines ou des milliers de locus. Les statistiques de DL basées sur les SNP (r, r_v) ou sur les fenêtres génomiques (cor_{PC1}, cor_{PC1v}) peuvent s'utiliser de façon complémentaire et permettent de détecter des signaux de sélection à différentes échelles dans le génome.

1.3 Contrôle des faux positifs et puissance de détection de la sélection épistatique

La comparaison des taux de faux positifs entre T_r - T_{corPC1} et T_{r_v} - T_{corPC1_v} montre qu'il est essentiel d'utiliser la correction pour la structure des populations et l'apparentement pour mesurer le DL associé à la sélection épistatique sur deux locus indépendants. Les taux de faux positifs de T_r / T_{corPC1} et T_{r_v} / T_{corPC1_v} varient respectivement autour de 74%-81% et 3%. Ces travaux étendent l'utilisation originelle de la mesure $\boldsymbol{r_v}$ proposée par (Mangin et al., 2012), à la recherche de signatures de sélection épistatique entre locus distants. Il semble toutefois que lorsque l'on combine les effets de l'apparentement et de la structure les statistiques T_{r_v} - T_{corPC1_v} montrent une plus forte variance sur le long terme (génération 300 des simulations) avec une augmentation du taux de faux positifs.

Parallèlement les analyses de puissances ont montré que la correction par la matrice d'apparentement n'augmente pas la puissance de détection de la sélection épistatique. La correction a même tendance à réduire la puissance dans le modèle de sélection épistatique compensatoire. En effet, le DL observé dans le modèle compensatoire semble lié à la structure des populations car les haplotypes sélectionnés entre sous-populations peuvent être différents (*AB* très fréquent dans une sous-population, et *ab* très fréquent dans l'autre sous-population) entrainant de forte valeurs de DL lorsque la structure n'est pas prise en compte. Globalement les puissances varient entre 10% et 65% pour α = 5%. Ces puissances sont en partie limitées par la taille modeste des populations simulées ; elles devraient augmenter dans

les populations de grandes tailles du fait d'une meilleure efficacité de sélection. Dans leur étude, Takahasi et Innan ont mesuré le DL sur des données simulées de paires de locus sous sélection épistatique selon les modèles coadapté et compensatoire (Takahasi and Innan 2008). Ils ont utilisé la statistique *r* pour mesurer le DL entre les paires de locus A et B en fonction de plusieurs classes de fréquences initiales des mutations sous sélection. Contrairement à nos simulations, Takahasi et Innan ont simulé des populations uniquement panmictiques et sans structure. Malgré ces différences, les puissances de détection de la statistique r que nous avons obtenues en panmixie (pour un α=5%) sont respectivement de 0.12 et 0.50 dans les modèles coadapté et compensatoire, contre 0.27 et 0.24 obtenus par Takahasi et Innan pour des classes de fréquences initiales des mutations *a* et *b* entre 0.4 et 0.6. Les différences en termes de puissance entre ces résultats sont probablement liées à la structure des populations, notamment dans le modèle compensatoire, où nous avons obtenu des puissances plus élevées que Takahasi et Innan mais dont les puissances de r_v sont plus faibles (autour de 0.31 pour α=5%) et se rapprochent des valeurs de Takahasi et Innan.

En perspective, il serait intéressant de simuler la sélection épistatique dans une des deux sous-populations (i.e. « adaptation locale ») et de calculer le DL à l'échelle de la sous-population et de la population globale afin d'évaluer les différences de signatures de sélection. Ce type de simulation présenterait l'avantage de ressembler à des cas réels fréquents où la sélection agit localement comme chez *M. truncatula* où l'on trouve fréquemment de la sélection en population FW ou C uniquement (Bonhomme et al., 2015). Cela permettrait d'évaluer l'influence sur le DL corrigé car les signaux de sélection à l'échelle de populations locales peuvent être plus difficile à détecter avec r_v et cor_{PC1v}. En effet, si l'on tient compte de l'ensemble des populations, la sélection sera corrélée à la structure des populations et nous pouvons perdre le signal lorsque nous utilisons la correction par la matrice d'apparentement. Enfin, cela permettrait d'évaluer l'influence des matrices d'apparentement sur la correction du DL à différentes échelles (sous-population, population entière). Les analyses chez *M. truncatula* (**Figure 29**) et chez l'homme (**Figure 33**, **Annexe 13**) montrent en effet que la correction par la matrice est d'autant plus forte si la structuration est forte (i.e. à une échelle de structuration géographique plus large). Enfin, pour explorer l'effet de la liaison physique sur la détection de la sélection épistatique et sur son efficacité, des simulations de sélection épistatique entre deux locus situés sur un même chromosome et à différentes distances (i.e. différents taux de recombinaison) pourraient être effectuées.

1.4 Signatures de sélection sur les locus en épistasie

Nous avons évalué l'effet de la sélection épistatique sur le polymorphisme des locus simulés. Les résultats ont montré que les locus simulés sous les modèles coadapté et additif présentent des signatures de balayages sélectifs liés à la sélection positive des allèles dérivés *a* et *b* dans les deux modèles. Il est cependant très difficile de distinguer ces deux modèles avec les statistiques de neutralité. Dans les différents modèles, les signatures de sélection sont plus fortes en autogamie qu'en panmixie car l'autogamie permet l'accélération de la fixation des génotypes *aa/bb* car le taux d'hétérozygotie est réduit. Le modèle compensatoire montre aussi des signaux de sélection mais les résultats sont plus contrastés car il y a un maintien du polymorphisme tout au long des générations. Globalement, les tests de neutralité ne permettent pas d'identifier des cosignatures de sélection sur des locus soumis à la sélection épistatique car ces signatures dépendent des fréquences initiales des allèles aux deux locus au moment où le processus de sélection entre en jeu. Dans son étude de 2009, Takahasi a également évalué les effets d'un modèle de coadaptation sur le polymorphisme des locus et il a montré qu'ils diffèrent en fonction du temps qui sépare l'apparition des mutations cosélectionnées (Takahasi 2009). Dans un modèle à deux locus A et B, si les mutations *a* et *b* sont toutes deux introduites sur un intervalle de temps réduit, comme c'est le cas dans nos simulations car elles sont présentes dès la génération 0, une diminution locale de la variabilité est observée (Otto & Whitlock, 2009; Takahasi, 2009). Ces résultats sont comparables aux signatures de balayages sélectifs que nous avons obtenues sur les modèles coadapté et additif de sélection positive. Cependant, dans le modèle compensatoire, les résultats sont plus contrastés car le modèle sélectionne ou contre-sélectionne *A*, *B*, *a* et *b* suivant les combinaisons prises : *AB* et *ab* sont sélectionnés positivement, alors que *Ab* et *aB* sont sélectionnés négativement.

Ainsi, seule la corrélation génétique (i.e. le DL) permet réellement d'identifier des signatures de sélection épistatique. Les signatures de sélection identifiées indépendamment sur chaque locus peuvent, toutefois, permettre d'évaluer comment les différents régimes de sélection épistatique peuvent influencer le polymorphisme des gènes. En effet, on peut remarquer que la sélection épistatique peut fréquemment aboutir à la cofixation d'allèles dérivés ou ancestraux, ce qui permet de penser que certains locus identifiés dans un

« selection scan » comme ayant subi un balayage sélectif peuvent aussi bien être en épistasie évolutive (Otto & Whitlock, 2009).

2. Détection de gènes sous sélection épistatique à l'aide de données SNP

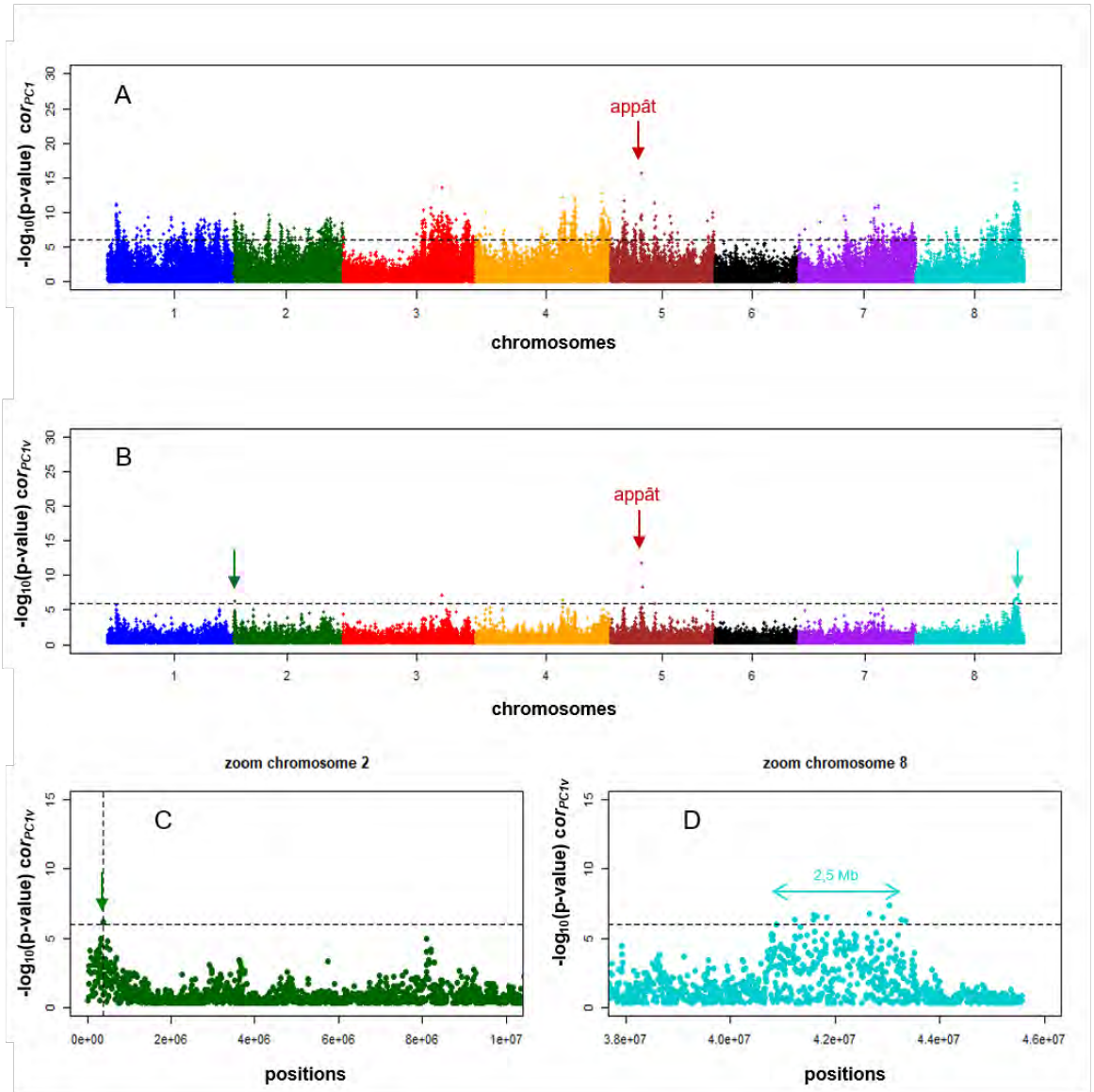
L'objectif du second chapitre a été d'identifier des interactions adaptatives entre gènes dont les fonctions biologiques ou moléculaires sont connues, inconnues ou non étudiées. Nous avons recherché des signatures de sélection épistatique chez *Medicago truncatula* à l'échelle du génome et nous avons également mené une analyse chez l'humain afin d'illustrer les possibilités d'application de la méthode que nous avons développée. Chez *Medicago truncatula*, deux approches ont été mises en place ; une approche « genome-wide » et une approche par groupes de gènes candidats, ciblée sur certaines fonctions biologiques ou moléculaires. Les statistiques T_r / T_{r_v} et T_{corPC1} / T_{corPC1_v} ont été utilisées pour réaliser des Genome-Wide Epistatic Selection Scans (« GWESS ») à l'aide de gènes candidats considérés comme appâts.

2.1 L'approche GWESS avec un gène « appât »

Nous proposons une approche originale, l'approche GWESS, qui consiste à analyser le DL entre un gène (ou une région génomique) « appât » et les autres locus du génome, et de représenter les résultats de l'analyse statistique sous la forme d'un Manhattan plot, de la même façon que pour les analyses GWAS. Par cette approche, il est possible de visualiser l'étendue du DL au niveau de la région du gène appât. Le DL intrachromosomique permet de rechercher des signatures de sélection épistatique longue distance du fait de nombreux événements de recombinaison (**Figure 45A,B**), et le DL interchromosomique permet de rechercher des signatures de sélection épistatique entre différents chromosomes. La détection de régions significativement associées au gène appât sur d'autres chromosomes dépendra de l'intensité du mécanisme de sélection épistatique et du contexte local du DL au niveau de la région cible : la résolution se fera parfois à l'échelle d'un gène candidat (**Figure 45C**) ou à l'échelle d'une région génomique plus large (**Figure 45D**). Ainsi, cette approche GWESS peut aussi permettre d'explorer visuellement la variabilité génomique du DL. Dans le but d'améliorer la détection de gènes coadaptés, notamment lorsque la résolution des GWESS se fait à l'échelle de fenêtres génomiques larges, nous pourrions envisager la possibilité de

traiter le signal de DL avec l'approche du score local comme cela est fait pour des analyses GWAS et des scans génomiques de sélection (Bonhomme et al., 2019; Fariello et al., 2017, Aoun et al., 2020 (sous presse)).

Figure 45 : Manhattan plots d'une analyse GWESS.
A et B illustrent les analyses sans (cor_{PC1}) ou avec (cor_{PC1v}) la correction du DL pour la structure des populations et l'apparentement entre les individus, respectivement. C et D sont des zooms sur des régions des chromosomes 2 et 8, respectivement, significativement associées au gène appât sur le chromosome 5. Sur le chromosome 2 la résolution est à l'échelle d'un gène candidat. Sur le chromosome 8 la résolution est à l'échelle d'une région génomique plus large.



2.2 L'approche GWESS chez Medicago truncatula

Plusieurs dizaines d'analyses GWESS ont été réalisés chez *Medicago truncatula*. Parmi l'ensemble des résultats obtenus, un des résultats majeurs est la découverte d'un signal de coadaptation entre les gènes *MtSUNN* et *MtCLE02* dans la sous-population FW et l'établissement d'un lien fonctionnel entre ces gènes dans un contexte de nodulation symbiotique (collaboration avec l'IPS2). Ces résultats constituent donc une démonstration de faisabilité de notre méthode. Contrairement à *MtSUNN* et à d'autres peptides CLE (*MtCLE12* et *MtCLE13*), *MtCLE02* n'est pas régulé lors de la nodulation. Ces résultats montrent bien l'intérêt de notre méthode du fait de sa complémentarité avec d'autres méthodes, telles que les analyses de transcriptomique. De plus, le gène *MtCLE02* est phylogénétiquement distant des autres peptides CLE qui ont déjà été caractérisés. L'identification de ce nouveau peptide CLE qui joue avec le récepteur SUNN un rôle dans la nodulation, montre que notre méthode peut permettre d'identifier de nouvelles interactions fonctionnelles entre paires de gènes, indépendamment de leurs profils d'expressions ou de corégulation.

Deux autres gènes appâts analysés ont permis d'identifier des interactions significatives. Le gène *MtCRA2* a été identifié en interaction avec le gène *MtRPG* et ils sont tous deux impliqués dans la symbiose rhizobienne. Dans le cadre de l'ANR PSYCHE, ils ont été identifiés comme faisant partie d'un même cluster de coexpression dans les nodules dans un contexte de stress abiotique hydrique et azoté. En perspective, il serait intéressant de tester de quelle façon ils interagissent. Nous avons également identifié des signaux de coadaptation entre les gènes *MtNIN* et *MtSHR*, deux gènes impliqués dans la symbiose rhizobienne. En effet, *MtSHR*, codant pour un facteur de transcription de type GRAS, a récemment été associé d'après une analyse de GWAS à la variation naturelle de la stimulation de la croissance racinaire de *M. truncatula* en réponse à des signaux symbiotiques de type lipochitooligosaccharides -LCO- (Bonhomme et al., en préparation). Par l'analyse de ces gènes, nous avons vu que les résultats des GWESS peuvent légèrement varier selon les tailles de fenêtres génomiques qui sont choisies, montrant l'importance du choix de ces tailles de fenêtres pour réaliser les ACP. Ce choix dépendra d'un compromis entre la densité en marqueurs SNP et la décroissance du DL (LD-decay) selon l'espèce étudiée.

2.3 L'approche GWESS chez l'homme

Les GWESS réalisées chez l'humain ont montré que ces approches peuvent être réalisées avec des données SNP pangénomiques sur n'importe quel organisme étudié. Ainsi, nous avons pu montrer un signal de sélection épistatique, ou de cosélection, entre les gènes *SLC24A5* et *EDAR*. Pour cette analyse, nous avons utilisé les statistiques T_r et T_{r_v} qui mesurent le DL à l'échelle des SNP. Ce choix a été guidé par la plus faible densité en SNP (1 SNP tous les 5kb) du jeu de données utilisé (HGDP-CEPH). Nous avons identifié un signal de coadaptation entre les SNP appâts de *SLC24A5* et *EDAR* à l'échelle de la population mondiale à partir des scans réalisés avec la statistique T_{r_v}. De plus, nous avons pu montrer que la distribution géographique des allèles des deux SNP qui ont été utilisés comme appâts est corrélée à la structure des populations à l'échelle mondiale ce qui explique que les scans avec T_r (non corrigé) présentent des valeurs de DL élevés par rapport aux scans réalisés avec T_{r_v}. En revanche, les scans réalisés à l'échelle de la sous-population de l'Asie centrale du sud présentent peu de différences en termes de distribution du DL entre les statistiques T_r et T_{r_v}, montrant que cette population est faiblement structurée et que la correction par la matrice d'apparentement influence faiblement les valeurs de DL dans le cas où il n'y a pas ou peu de sous-structuration, et si l'apparentement est faible. Enfin, nous avons identifié une des origines possibles de cette signature de cosélection au niveau de l'ethnie des Burusho, où l'association des allèles ancestraux d'*EDAR* et des allèles dérivés de *SLC24A5* est très significative. Ces associations d'allèles reflètent la persistance d'une peau claire de type européenne et d'une structure de cheveux fins dans ce groupe ethnique asiatique. À ce jour, aucun lien fonctionnel n'a encore été démontré entre le gène *SLC24A5* lié à la pigmentation de la peau et la voie de l'ectodysplasine à laquelle appartient *EDAR*. Alors que le DL à longue distance entre *EDAR* et *SLC24A5* reflète largement une forte sélection positive agissant indépendamment sur ces deux gènes dans différentes régions géographiques du monde (Bryk et al., 2008; Deng & Xu, 2017; Izagirre et al., 2006; Sabeti et al., 2007; Sadier et al., 2014; Speidel et al., 2019), nos résultats suggèrent également que la cosélection d'allèles de ces gènes peut avoir localement contribué à la constitution phénotypique des populations humaines.

2.4 Perspectives pour l'approche GWESS

L'approche GWESS développée pour identifier des interactions adaptatives entre des paires de SNP ou des paires de gènes est facilement implémentable et plutôt rapide (https://github.com/leaboyrie/LD_corpc1). Les matrices d'apparentement sont calculables en quelques minutes ou secondes selon les données (i.e. 30 sec à ~ 3 min chez *Medicago truncatula* et l'humain) et pour réaliser une GWESS avec un gène ou un SNP appât, il nous a fallu entre 2 min et 12 min pour analyser respectivement les 48 333 gènes de *M. truncatula* avec T_{corPC1_v} et les 431 951 SNP chez l'humain avec T_{r_v}. Les analyses de GWESS peuvent être facilement implémentées chez des organismes non modèles mais qui possèdent des données SNP disponibles à l'échelle du génome et qui présentent une densité suffisante pour identifier des gènes candidats ou des régions génomiques sous sélection épistatique. Les analyses GWESS avec l'approche appât pourraient être réalisées également chez des espèces non modèles apparentées à des espèces modèles bien caractérisées afin d'identifier des interactions génétiques potentiellement partagées ou perdues entre des espèces proches. Enfin, la validation fonctionnelle des interactions identifiées par GWESS reste limitée aux espèces modèles mais de façon générale, les GWESS sont des analyses complémentaires aux analyses d'expression/co-expression ou de GWAS. Chez les espèces non modèles, si les GWESS ne sont pas possibles car l'on ne dispose pas, par exemple, de données SNP à l'échelle du génome, il est toujours possible d'analyser uniquement des gènes candidats avec des marqueurs génétiques spécifiques mais l'analyse des résultats de DL est limitée en l'absence de référence sur d'autres portions du génome.

2.5 Polymorphisme des gènes de M. truncatula et traces de sélection sur les gènes en épistasie

L'objectif de cette partie a été d'évaluer si les gènes de *Medicago truncatula* qui présentent des signatures de sélection épistatique présentent aussi des signatures de sélection individuellement. Tout d'abord, nous avons calculé des statistiques de tests de neutralité (*D*, *H*, *E*) chez *M. truncatula* sur l'ensemble des gènes et dans les trois populations et nous avons établi les distributions de ces statistiques à l'échelle du génome, ce qui n'avait encore que partiellement été fait chez *M. truncatula*. D'après ces distributions, nous avons mis en évidence le fait que la population FW présente un déficit en allèles rares potentiellement lié à un effet de sous-structuration génétique. Ces résultats appuient les

conclusions d'études précédentes sur la structuration génétique chez *M. truncatula* (Ronfort et al. 2006, Gentzbittel et al., 2019). Dans la population entière, les analyses de polymorphisme montrent qu'il y a un nombre important de SNP avec des allèles rares apportés par la population C et qui masquent les effets de la sous-structure de la population FW. Par le cumul de ces effets, les distributions des statistiques de neutralité *D* et *E* sont à l'équilibre. Ces résultats confirment l'importance de tenir compte de la structuration génétique dans les analyses de signatures de sélection, en effectuant comme nous le faisons chez *M. truncatula* des analyses dans les 2 grandes sous-populations FW et C.

L'analyse du polymorphisme des gènes sous sélection épistatique dans la population entière chez *M. truncatula* nous a permis de montrer que 30% d'entre eux présentent également des signatures de sélection individuelles, dont 15% sont sous balayage sélectif. Dans les populations FW et C, il y a respectivement 27% et 35% des gènes sous sélection épistatique qui présentent également des signatures de sélection dont 15% et 19% sont sous balayage sélectif. Ces résultats montrent qu'entre 65% et 73% des gènes sous sélection épistatique sont en apparence « neutres » lorsqu'ils sont analysés individuellement, alors qu'ils présentent de potentiels signaux de coadaptation avec d'autres gènes. Comme les simulations l'ont montré, les statistiques de neutralité calculées chez *M. truncatula* ne permettent pas d'identifier des signatures de cosélection entre paires de gènes. Cependant, ces statistiques peuvent permettre de quantifier l'influence de la sélection épistatique sur le polymorphisme des gènes et éventuellement de donner des informations supplémentaires sur les gènes identifiés, comme par exemple le mécanisme potentiel de sélection épistatique sous-jacent. Ces résultats montrent l'importance de considérer les gènes en interaction avec d'autres gènes et non pas simplement comme des entités individuelles.

2.6 Signatures génomiques de sélection épistatique chez M. truncatula

Dans une dernière partie, nous avons recherché des signatures de sélection épistatique à l'échelle du génome chez *M. truncatula*, en comparant toutes les paires de gènes, afin d'identifier de nouvelles interactions entre de nouveaux gènes ou des gènes déjà connus. Nous avons ainsi généré une source de données importante. L'analyse de l'ensemble de ces résultats constitue une partie toujours exploratoire dont l'objectif serait d'identifier de nouvelles interactions. Deux approches ont été proposées pour initier l'analyse de ces

données, mais différentes méthodes pourraient être mises en place à l'avenir et notamment, par exemple, des méthodes statistiques d'analyse et d'inférence de réseaux.

La première approche mise en place a été de comparer les patrons de DL entre des ensembles de gènes candidats et des ensembles de gènes échantillonnés aléatoirement. Nous avons ainsi montré que des gènes de mêmes voies biologiques, tels que des gènes impliqués dans les symbioses rhizobienne et mycorhizienne, présentent un enrichissement en signaux de sélection épistatique. Nous avons également identifié des signaux de sélection épistatique entre des gènes de même fonction moléculaire, tels que les facteurs de transcription ou les gènes codant pour des récepteurs de type RLK. Toutefois, l'approche par comparaison de ces ensembles de gènes candidats et aléatoires est une approche globale qui ne permet pas d'identifier précisément de nouvelles interactions significatives et elle ne constitue pas une analyse des réseaux d'interactions.

Dans la partie suivante, nous avons reconstruit des réseaux génomiques d'interactions génétiques significatives afin d'analyser l'ensemble des résultats de DL sur le génome. Les réseaux construits à partir des interactions interchromosomiques et intrachromosomiques significatives au seuil de p-valeur de 10^{-11} sont constitués de 44 207, 46 015 et 45 282 gènes et 961 460, 801 398 et 936 804 arêtes en population entière, FW et C, respectivement. Pour réduire la complexité de ces réseaux, l'objectif a été de réduire l'impact du DL physique et ne conserver que le DL lié à la sélection épistatique. Dans cet objectif, nous n'avons conservé que les interactions interchromosomiques (**Figure 41**) mais cette méthode présente le défaut de supprimer aussi potentiellement du DL lié à la sélection épistatique intrachromosomique à longue distance. Nous avons aussi filtré des sous-réseaux pour un nombre minimum de 2 arêtes, ce qui a permis d'éliminer en grande partie le DL entre proches voisins (**Figure 42**). Ce type d'approche, exploitant certaines propriétés du réseau en lien avec le DL physique mériterait d'être approfondi.

En perspective, d'autres approches de modélisation pourraient permettre de réduire le DL physique dans les réseaux. Par exemple, nous pourrions utiliser la méthode de « adjacency-constrained clustering » - *adjclust* - (Neuvial et al., 2017) qui a été développée pour construire des blocs de DL au sein des chromosomes. Nous pourrions ainsi délimiter des blocs de DL de tailles variables et recalculer les ACP sur des fenêtres génomiques correspondant à ces blocs, pour enfin évaluer le DL entre blocs. Une seconde approche consisterait à modéliser le DL

physique par une approche de rééchantillonnage et décomposer le DL entre deux locus en un DL physique et une composante résiduelle qui correspondrait à du DL relié à des associations épistatiques.

Enfin, pour analyser une partie des données produites à l'échelle du génome, nous avons extrait des sous-réseaux d'interactions significatives ancrés sur des gènes candidats ainsi que tous leurs interactants. Les sous-réseaux ont été extraits à partir de 98 gènes candidats impliqués dans les symbioses racinaires chez *M. truncatula* (symbioses AM, RN et RN+AM). Nous avons fait varier le filtre de deux ou trois interactants (arêtes) par gène afin de réduire la complexité et le bruit de fond des sous-réseaux. Cette représentation a montré les positions centrales (« hub ») de certains gènes tels que *MtRPG*, *MtCRA2* et *MtSYMREM1* qui possèdent un degré de connexion important lié à de nombreuses interactions adaptatives avec d'autres gènes du sous-réseau. D'autres part, nous avons vu qu'un nombre important de gènes candidats sont présents dans au moins deux des trois sous-réseaux populationnels, montrant qu'il y a un remarquable degré de conservation de la connectivité parmi ces gènes à l'échelle de l'espèce chez *M. truncatula*.

L'analyse de sous-réseaux, en utilisant des sous-ensembles de gènes candidats, permet de visualiser les interactions entre des gènes caractérisés et/ou de nouveaux gènes, ce qui n'est pas possible à l'échelle du réseau global. Ces représentations permettent de visualiser des interactions directes ou indirectes entre différents gènes, ouvrant ainsi de nouvelles possibilités pour la caractérisation fonctionnelle. En perspective, il serait intéressant de tester d'autres sous-ensembles de gènes impliqués dans différentes voies biologiques (floraison, voies métaboliques, immunité …).

3. Conclusion

Par ce travail, nous avons pu montrer l'importance de la sélection épistatique pour l'analyse et la compréhension des bases génétiques de l'adaptation des organismes. Pour la prédiction de l'adaptation des populations aux changements de l'environnement, il est important de considérer les interactions entre les gènes car les milliers de gènes du génome d'un organisme sont des composantes de différents réseaux d'interactions fonctionnelles plus ou moins intimes. De ce fait, une mutation qui apparaît sur un gène peut avoir des conséquences fonctionnelles, et donc évolutives, sur les autres gènes du réseau (Boyle et al.,

2017; Hansen, 2013). Dans la recherche de signatures de sélection épistatique, une corrélation génétique significative identifiée entre deux gènes montre qu'ils sont potentiellement coadaptés, mais la nature de leurs interactions ainsi que le processus biologique dans lequel ils interviennent restent hypothétiques. La limite de ces analyses statistiques sur les signatures de sélection entre paires de gènes est que nous ne pouvons souvent qu'émettre des hypothèses sur de potentielles interactions physiques ou fonctionnelles. Ces interactions ne pourront être vérifiées que par des études expérimentales. L'approche que nous proposons (dont l'approche GWESS), est complémentaire, et se situe à mi-chemin des analyses de GWAS et de GWSS basées sur la diversité génétique d'une espèce, ainsi que des analyses génomiques telles que la transcriptomique qui vise à identifier des groupes de gènes co-exprimés. Elle est un outil supplémentaire pour identifier des gènes constituant les bases de l'adaptation et qui peuvent ensuite faire l'objet de validations expérimentales. La seconde limite de ces analyses est de pouvoir considérer les interactions épistatiques de façon intégrées dans le cadre d'analyses génomiques. En effet, la plupart des modèles théoriques qui ont été développés pour étudier les bases génétiques de l'adaptation ignorent souvent les effets épistatiques et font l'hypothèse que les gènes agissent indépendamment sur la *fitness* (analyses de génomique des populations) ou sur l'expression d'un phénotype (analyses de génétique quantitative et de cartographie comme la GWAS). Si ces effets sont rarement pris en compte, c'est en partie dû à la complexité de considérer les gènes en interactions car la problématique d'exploration est décuplée d'un ordre de grandeur $\frac{N^2}{2}$ pour N gènes. L'analyse de génomes entiers demande donc du temps de calcul et de traitement analytique supplémentaire. Dans le but d'analyser ces interactions épistatiques, nous avons développé les analyses de GWESS, avec l'approche « appât », à une dimension, permettant de détecter les interactions possibles entre un gène et tous les autres gènes du génome. Ce sont des outils facilement implémentables chez différentes espèces pour identifier de nouveaux gènes candidats en interaction fonctionnelle avec un gène connu, dont les données de polymorphismes montrent qu'ils co-évoluent. Enfin, les analyses de réseaux d'interactions en sont au stade exploratoire mais elles constituent une vaste ressource de données qu'il sera intéressant d'exploiter à l'avenir. La modélisation de l'évolution des réseaux de gènes en épistasie (> 2 locus) est aussi une étape fondamentale pour comprendre et prédire l'adaptation, à long terme, des populations sur la base de ces entités génétiques (Yukilevich et al., 2008). Les travaux que j'ai

effectués pendant ma thèse contribueront à l'identification, à l'exploration fonctionnelle, et à la modélisation de l'évolution des réseaux de gènes coadaptés à l'aide de données SNP pangénomiques, chez des organismes modèles ou non modèles.

Publications

Article principal de la Thèse

Léa Boyrie, **Corentin Moreau, Florian Frugier, Christophe Jacquet, Maxime Bonhomme**. (2020). A linkage disequilibrium-based statistical test for Genome-Wide Epistatic Selection Scans in structured populations. ***Heredity*****. EN REVISION**

(IF = 3.801)

Article en collaboration

Nathalie Aoun, Henri Desaint, Léa Boyrie, Maxime Bonhomme, Laurent Deslandes, Richard Berthomé, Fabrice Roux. (2020). A complex network of additive and epistatic quantitative trait loci underlies natural variation of *Arabidopsis thaliana* quantitative disease resistance to *Ralstonia solanacearum* under heat stress. ***Molecular Plant Pathology*****. SOUS PRESSE**

(IF = 4.379)

A linkage disequilibrium-based statistical test for Genome-Wide Epistatic Selection Scans in structured populations

Léa Boyrie[1], Corentin Moreau[2], Florian Frugier[2], Christophe Jacquet[1], Maxime Bonhomme[1]

[1] Laboratoire de Recherche en Sciences Végétales (LRSV), Université de Toulouse, Centre National de la Recherche Scientifique (CNRS), Université Paul Sabatier (UPS), Castanet-Tolosan, France

[2] Institute of Plant Sciences-Paris Saclay (IPS2), Centre National de la Recherche Scientifique, Univ Paris-Sud, Univ Paris-Diderot, Univ d'Evry, Institut National de la Recherche Agronomique, Université Paris-Saclay, 91192 Gif-sur-Yvette, France

Corresponding author: Maxime Bonhomme

bonhomme@lrsv.ups-tlse.fr

Word count : 6606

Abstract

The quest for signatures of selection using SNP data has proven efficient to uncover genes involved in conserved and/or adaptive molecular functions, but none of the statistical methods were designed to identify interacting alleles as targets of selective processes. Here, we propose a statistical test aimed at detecting epistatic selection, based on a linkage disequilibrium (LD) measure accounting for population structure and heterogeneous relatedness between individuals. SNP-based (T_{r_v}) and window-based (T_{corPC1_v}) statistics fit a Student distribution, allowing to test the significance of correlation coefficients. As a proof of concept, we use SNP data from the *Medicago truncatula* symbiotic legume plant and uncover a previously unknown gene coadaptation between the *MtSUNN* (*Super Numeric Nodule*) receptor and the *MtCLE02* (*CLAVATA3-Like*) signaling peptide. We also provide experimental evidence supporting a *MtSUNN*-dependent negative role of *MtCLE02* in symbiotic root nodulation. Using human HGDP-CEPH SNP data, our new statistical test uncovers strong LD between *SLC24A5* (skin pigmentation) and *EDAR* (hairs, teeth, sweat glands development) worldwide, which persists after correction for population structure and relatedness in Central South Asian populations. This result suggests that epistatic selection or coselection could have contributed to the phenotypic make-up in some human populations. Applying this approach to genome-wide SNP data will facilitate the identification of coadapted gene networks in model or non-model organisms.

Introduction

In populations, natural selection targets genomic regions with evolutionarily conserved functions or with genetic variants contributing to adaptation to changing environments. Patterns of DNA sequence polymorphisms in these regions are expected to bear the signature of directional or balancing, positive selection on adaptive mutations, or of negative selection against deleterious mutations (Bamshad and Wooding, 2003; Nielsen, 2005; Vitti *et al*, 2013). Identifying genes showing such selection signatures has been a major goal of population genetics over the last decades. Many statistical methods have been developed, accounting for the neutral evolution expected for molecular polymorphisms in populations with varying degrees of genetic structure or with particular demographic histories (Pavlidis and Alachiotis, 2017; Vitti *et al*, 2013; Weigand and Leese, 2018). Thanks to high-throughput sequencing technologies, these methods can now be used to perform Genome-Wide Scans for Selection (GWSS) using Single Nucleotide Polymorphism (SNP) datasets (Ahrens *et al*, 2018; Haasl and Payseur, 2016; Oleksyk *et al*, 2010). Although GWSS have identified cohorts of genes associated with past or ongoing selective processes, they are not designed to identify gene coadaptation, resulting from epistatic selection on interacting genes (Otto and Whitlock, 2009).

Few studies have examined the impact of epistatic selection models in population samples using DNA polymorphisms. Simulations of two-locus epistatic models with different degrees of recombination (i.e. $0 < c < 0.5$) in a panmictic population have shown that the efficiency of epistatic selection and its statistical detection are improved if standing genetic variation already exists, rather than if selection operates on *de novo* mutations (Takahasi, 2009; Takahasi and Tajima, 2005). Simulations of two-locus coadaptation in subdivided populations have also shown that the fixation probability of the coadapted haplotype across all sub-populations increases upon moderate migration and isolation (Takahasi, 2007).

Adaptive epistatic interactions between alleles at two independent loci are expected to generate Linkage Disequilibrium (LD). It has been shown that the correlation coefficient *r*, a LD measure related to r^2, can be used to detect epistatic selection between two bi-allelic loci in a population because it is a directional measure which can indicate an excess of ancestral and derived allelic associations, relative to recombinant allelic associations (Takahasi and Innan, 2008). This will detect epistatic selection either in a coadaptation model where two derived alleles can form a coadapted allelic combination, or in a compensatory model where the two derived alleles are individually deleterious but compensate when combined (Piskol and Stephan, 2008). More recently, a simulation study of epistatic selection in structured populations has suggested the use of the $D'_{IS}{}^2$ measure of LD (Ohta, 1982a; Ohta, 1982b), which quantifies how the frequencies of the different two-locus haplotypes in a sub-population depart from the average frequencies across all sub-populations (Id-Lahoucine *et al*, 2019). However, high $D'_{IS}{}^2$ values were found in simulated models of two-locus epistatic selection but also of single-locus selection at two independent loci. This suggests that $D'_{IS}{}^2$ cannot always distinguish between these two models. In addition, population structure, genetic drift and relatedness among individuals due to non-random mating also act as confounders because they increase genome-wide levels of LD and generate long-distance LD (Glémin *et al*, 2006; Mangin *et al*, 2012; Nordborg, 2000; Slatkin, 2008), which can falsely be interpreted as signatures of epistatic selection (Zhang *et al*, 2004).

A significant improvement towards capturing the LD due to physical linkage in a structured population with various degrees of relatedness among individuals was the introduction of r^2_v, an r^2 measure which includes the kinship matrix into the calculation in order to penalize correlated two-locus genotypic data arising from high levels of relatedness (Mangin *et al*, 2012). Subsequently, the $r|a$ measure of LD was proposed to identify interspecific genetic incompatibilities corresponding to pairs of loci showing an excess of ancestral haplotype

combinations in admixed populations (Schumer and Brandvain, 2016). The $r|a$ measure is a partial correlation coefficient between genotypic data at two loci given the genome-wide ancestry proportion, a, between two species. Although such type of measure represented an improvement, authors acknowledged that genetic relationships in populations with complex demographic histories and genetic structures were not fully taken into account (Schumer and Brandvain, 2016).

In this study, we propose a statistical test to detect epistatic selection in heterogeneously structured populations (i) between two bi-allelic SNPs by using the r_v measure, or (ii) between two genomic regions including each multiple SNPs, by using the cor_{PC1v} measure. This latter measure captures the quantitative correlation between the first principal component (PC1) summarizing the multi-SNP genotypes for each genomic region. Using simulations of genome-wide SNP data in structured diploid populations with random to self-mating processes under two epistatic selection models, we show that, compared with r and cor_{PC1}, r_v and cor_{PC1v} (i) drastically reduced the background LD generated by population structure and relatedness between individuals; (ii) showed an equivalent or a lower power to detect epistatic selection, depending on the mating process, on the dominance of selected mutations and on the selection model; and (iii) T statistics ($T = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}$) derived from r_v or cor_{PC1v} (i.e. T_{r_v} and T_{corPC1_v}) fitted a Student distribution $\tau_{(n-2)}$ under the null hypothesis of independence between the two tested loci. Hence, unlike T statistics derived from r or cor_{PC1} (i.e. T_r and T_{corPC1}), T_{r_v} and T_{corPC1_v} can be used for statistical testing of the correlation coefficient between two loci, while accounting for population structure and heterogeneous relatedness between individuals.

Empirical detections of epistatic selection on SNP data are scarce in the literature (Brachi *et al*, 2015; Caicedo *et al*, 2004; Csilléry *et al*, 2014; Grzeskowiak *et al*, 2014; Hu and Hu, 2015; Pool, 2015). We assayed our statistical test in the frame of genome-wide epistatic selection scans (GWESS), with genomic SNP data from two different model organisms: the

legume plant *Medicago truncatula* and humans. As a proof of concept, we first described in *M. truncatula* the detection of epistatic selection between the Super Numeric Nodule *MtSUNN* gene, encoding a receptor which is central for the negative regulation of symbiotic root nodulation, and the CLAVATA3-like (CLE) signaling peptide *MtCLE02*. Accordingly, an ectopic expression of the *MtCLE02* gene in *M. truncatula* wild-type and *sunn* mutant roots experimentally demonstrated a *MtSUNN*-dependent negative role of the *MtCLE02* gene on nodulation, hence validating functionally the genetic interaction between these two genes. In humans, we illustrated the usefulness of the approach by identifying a significant epistatic or coselection signal in Central South Asian populations between *SLC24A5* and *EDAR* genes, encoding respectively a cation exchanger affecting pigmentation in zebrafish and human (Lamason *et al*, 2005) and a receptor involved in the development of hair follicles, teeth and sweat glands (Botchkarev and Fessing, 2005; Sadier *et al*, 2014). Together with the fact that *SLC24A5* and *EDAR* were previously shown to be under strong positive selection in Europe and East Asian populations, respectively (Bryk *et al*, 2008; Sabeti *et al*, 2007; Speidel *et al*, 2019), our results highlight the role of epistatic selection or coselection in shaping gene coadaptation during the evolution of populations.

Materials and Methods

Genetic models of epistatic selection

We follow fitness genotype formalization under epistatic selection models as in (Takahasi and Innan, 2008; Takahasi and Tajima, 2005). Two independent bi-allelic loci A and B were considered, with ancestral alleles *A* and *B*, and derived alleles *a* and *b*, in a haploid population. The coadaptation model consists in positively selecting the two-locus *ab* combination. The compensatory model consists in selecting against the *Ab* and *aB* two-locus combinations, but not against *AB* and *ab*. The coefficient *s* is used to positively or negatively

149 select two-locus genotypes (**Table 1**), and in the neutral model all fitness values are set up to 1.
150 In a diploid population, the two-locus fitness expression is more complex because it depends
151 on the level of dominance of the derived alleles (**Supplementary Table S1**).

152 **SNP-based and window-based LD measures of epistatic selection**

153 In a diploid organism, at a given bi-allelic SNP with alleles coded 0 and 1, the three
154 possible genotypes are (00, 01, 11), which can be coded as the allelic dose of allele 1 (0, 1, 2).
155 The measure on unphased genotypes between two bi-allelic loci is defined by the correlation
156 coefficient *r* between vectors of genotypes at the SNPs *l* and *m*, X^l and X^m (Hill and Robertson,
157 1968; Rogers and Huff, 2009; Weir, 1979):

158 $$r = \frac{\sum_{i=1}^{n}(X_i^l - \overline{X^l})(X_i^m - \overline{X^m})}{\sqrt{\sum_{i=1}^{n}(X_i^l - \overline{X^l})^2}\sqrt{\sum_{i=1}^{n}(X_i^m - \overline{X^m})^2}}$$

159 In the case where 0 and 1 are the ancestral and derived alleles, respectively, a positive sign of *r*
160 indicates that combinations of ancestral and derived alleles (i.e. 00 and 11) preferentially
161 segregate in individuals at the two SNPs, compared with alternative combinations (i.e. 01 and
162 10). At two physically unlinked loci in a panmictic population, this measure allows to detect
163 fitness interactions between two new mutations under the coadaptation or the compensatory
164 model (Takahasi and Innan, 2008). In the context of GWESS with high-density SNP data, we
165 propose to use the cor_{PC1} measure of LD between two genomic regions containing each multiple
166 SNPs. The first principal component $PC1^l$ is used to summarize quantitatively the multi-SNP
167 genotypes of the genomic region *l* (see (McVean, 2009)). Then, cor_{PC1} is the correlation
168 coefficient between vectors of summarized multi-SNP genotypes of the two genomic regions *l*
169 and *m*, $PC1^l$ and $PC1^m$:

170 $$cor_{PC1} = \frac{\sum_{i=1}^{n}(PC1_i^l - \overline{PC1^l})(PC1_i^m - \overline{PC1^m})}{\sqrt{\sum_{i=1}^{n}(PC1_i^l - \overline{PC1^l})^2}\sqrt{\sum_{i=1}^{n}(PC1_i^m - \overline{PC1^m})^2}}$$

However, as mentioned in (Mangin *et al*, 2012), population structure and relatedness among individuals generate non-independence between individuals and tend to bias upwardly the LD values. This is particularly the case in highly inbred or predominantly selfing species (Glémin *et al*, 2006). At a given locus, Mangin and collaborators proposed to weight the observations by multiplying the vector of genotypes by $V^{-\frac{1}{2}}$, where V is the kinship (or relatedness) matrix among individuals. V is built with the V_{ij} covariance for all pairs (i,j) of individuals. V_{ij} is the average number of identical genotypes between individuals *i* and *j* over all SNPs, in a genome-wide SNP dataset. This measure of Identity In State (IIS) is a good proxy of Identity By Descent (IBS) as SNP markers are likely to be accurately modelled by an infinite site mutation model. Consequently, since r is the Pearson correlation coefficient, r_v can be computed as $cor(V^{-\frac{1}{2}}X^l, V^{-\frac{1}{2}}X^m)$ (Mangin *et al*, 2012) and equivalently cor_{PC1v} as $cor(V^{-\frac{1}{2}}PC1^l, V^{-\frac{1}{2}}PC1^m)$.

Statistical test of epistatic selection based on linkage disequilibrium

Under the hypothesis that observations within X^l and X^m (respectively within $PC1^l$ and $PC1^m$) are independent, then r (respectively cor_{PC1}) can be used to obtain the T statistics:

$$T = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}$$

which follows a Student distribution $\tau_{(n-2)}$. However, in the case were observations are not independent, i.e. when the genotypes at a given locus are correlated within the population due to non-random mating and/or between populations due to structure, we then expect that only the T statistics obtained using r_v or cor_{PC1v} follow the $\tau_{(n-2)}$ distribution. In the case where the ancestral/derived allele status is known at the SNPs, a positive sign of r (or r_v) strictly reflects the coadaptation and compensatory epistatic models (see the previous section), and a unilateral test can be performed with alternative hypothesis being "r (or r_v) > 0". If the ancestral/derived allele status is not known, the sign of r is not interpretable, but the p-value of the test can be

computed on either side of the null distribution. Likewise, whether the ancestral/derived allele status of the SNPs is known or not, the sign of cor_{PC1} (or cor_{PC1v}) is not interpretable since PC1 or PC1 with opposite signs imply an identical ranking of individuals genotypes (or relatedness) in a given genomic region (see for instance (Li and Ralph, 2019)).

Simulation and LD-based detection of epistatic selection in structured population

Simulations of neutral evolution at two independent bi-allelic loci were carried out in order to evaluate the distribution of T_r, T_{corPC1}, T_{r_v} and T_{corPC1_v} statistics under the null hypothesis and their fit to a Student distribution $\tau_{(n-2)}$. In addition, two-locus epistatic selection was simulated in the same framework to evaluate the statistical power (1 - β) of these measures to detect two-locus epistatic selection given a type I error (α) for the null hypothesis. Details of our simulation procedure are provided in the **Supplementary Method File 1**, along with the python code used to run the simulations. Briefly, genome-wide (4 chromosomes) SNP data (~ 15 000 SNPs per chromosome) in a two-population split model with 250 diploid individuals per population during 300 generations (the ancestral population before the split was generated by coalescent simulations) were replicated 1000 times for all combinations of the following parameter settings: (i) selection regimes as neutrality, coadapted (COAD) and compensatory (COMP) two-locus epistatic selection, or additive (ADD) two-locus selection (all selection models starting 100 generations after the split time), (ii) random or self-mating (95% selfing rate) since the initial generation, and (iii) complete recessivity, codominance or dominance of the mutations under selection.

GWESS with SNP data in *Medicago truncatula* and human

GWESS was performed in *M. truncatula* using a raw dataset of 22 079 533 SNP markers identified on the eight chromosomes of the species by the Medicago HapMap Project on a collection of 262 accessions (see http://www.medicagohapmap.org/downloads/mt40). The collection has already been screened for GWAS for different traits (Bonhomme *et al*, 2014;

Bonhomme *et al*, 2019; Burgarella *et al*, 2016; Kang *et al*, 2015; Le Signor *et al*, 2017; Rey *et al*, 2017; Stanton-Geddes *et al*, 2013; Yoder *et al*, 2014) but also for GWSS (Bonhomme *et al*, 2015; Branca *et al*, 2011; Paape *et al*, 2013). This highly self-mating species (95% selfing rate), originating from the Mediterranean basin, is structured in two major sub-populations, the Far West population (FW) concentrated on the West part under Atlantic influence, and the Circum population (C) that spreads over the rest of the Mediterranean basin (Bonhomme *et al*, 2014; Burgarella *et al*, 2016; De Mita *et al*, 2011; Ronfort *et al*, 2006). Samples from the FW sub-population and from the C sub-population consist of 80 and 182 accessions, respectively. We used a bait approach for GWESS, in which T_{corPC1} and T_{corPC1_v} were calculated for a given candidate gene, here *MtSUNN*, with the Medtr4g070970 gene identifier in the genome version 4.0 - http://www.medicagogenome.org/ - (Tang *et al*, 2014) or MtrunA17Chr4g0035451 in the v5 version- https://medicago.toulouse.inra.fr/MtrunA17r5.0-ANR/ - (Pecrix *et al*, 2018), against each of the remaining 48 339 genes of the genome. The T_{corPC1} and T_{corPC1_v} statistics were calculated based on PC1 values from SNP data located in 10 kbp windows spanning each *M. truncatula* gene, and PC1 values from *MtSUNN*. P-values were then obtained from the $\tau_{(n-2)}$ null distribution. For the PC1 calculation on each gene, imputed SNP data were required. Gene-based imputation was performed using the TASSEL software (Bradbury *et al*, 2007), where each missing base was imputed with the accession that shares the longest haplotype surrounding the base, on a window of 30 SNPs maximum (Bonhomme *et al*, 2014). For the calculation of T_{corPC1_v}, the kinship matrix *V* of the 262 individuals was estimated based on 7 252 792 SNPs with a 5% Minor Allele Frequency (MAF).

In human, GWESS was performed by using the dataset of 644 257 SNPs (431 951 SNPs with a 5% MAF) with no missing data from the HGDP-CEPH Human Genome Diversity Panel on a world-wide sample (America, Asia, Europe, Middle East, North Africa, Sub-Saharan Africa and Oceania) of 940 individuals belonging to 57 populations from 23 countries (Cann *et*

al, 2002; Li *et al*, 2008). The genome version (i.e. the gene positions) B36 was used for this analysis in order to fit with SNP positions in the HGDP-CEPH dataset, similarly to (Daub *et al*, 2013). We used a bait approach in which T_r and T_{r_v} were calculated for SNPs located in or near *SLC24A5* and *EDAR* genes (chromosome 15 and 2, respectively), against each of the remaining SNPs of the genome.

Functional genetic validation of the relationship between *MtCLE02* and *MtSUNN* in *Medicago truncatula*

The *M. truncatula sunn* mutant, the *MtCLE02* cloning strategy for overexpression, the root transformation protocol, and the conditions for phenotyping the nodulation are described in the **Supplementary Method File 2**.

Results

Quality control of simulations

To ensure that simulations produced consistent between-population structure and within-population inbreeding levels, the F_{ST} and F_{IS} parameters were calculated (**Supplementary Figure S1**). At the outcome of the simulations (i.e. generation 300), the average F_{IS} in self-mating and panmictic populations was equal to 0.92 and 0.07 respectively, while the average F_{ST} was equal to 0.19 and 0.10 respectively. In order to quantify the fixation of coselected alleles, we tracked down the evolution of the frequency of the derived alleles *a* and *b* at the two SNPs A and B intended to be targeted by selection, located on two different chromosomes. Selection efficiency was measured by the co-fixation rate of *a* and *b* at each generation in each selection model (**Supplementary Figure S2**). The first observation was that the COAD epistatic model generally induced a higher speed of co-fixation than the ADD positive selection model, while, as expected, the COMP model tended to maintain a higher polymorphism due to selection of both *AB* and *ab* combinations at the two selected loci. The

second observation was that co-fixation rates of the derived alleles in self-mating populations reached more rapidly an equilibrium value than in panmictic populations, but more importantly that in self-mating populations the dominance level of the selected mutations had few effect on the co-fixation dynamics because of the very low heterozygosity level (F_{IS} = 0.92 at the onset of selection). A similar result was observed in a simulation study of selective sweeps in self-mating populations (Hartfield and Bataillon, 2020). On the other hand, the dominance level in panmictic populations strongly impacted co-fixation dynamics due to more complex fitness patterns in the presence of heterozygotes (**Supplementary Table S1**). However, despite starting from standing variation at SNPs under selection, values of the co-fixation rates were moderate, and this result must be interpreted in light of the small size of the simulated population (N = 250 in each population), as selection efficiency increases with population size according to a factor *Ns* (Glémin, 2007).

Two-locus linkage disequilibrium under epistatic selection models

We focused hereafter on the evolution of the two-locus average LD across simulations, in self-mating and panmictic populations under selection models with codominance. Codominance of the selected mutations indeed produced the highest co-fixation rates in panmictic populations, while the dominance level had no effect in self-mating populations (**Supplementary Figure S2**). Dominance or recessivity of the selected mutations mainly impacted negatively the two-locus average LD under the COMP epistatic model in panmictic populations (**Supplementary Figure S3)**. In the codominant mutations model, we first observed that under the neutral model, the population structure with or without non-random mating generated LD between two independent loci, as measured using r or cor_{PC1}, that could reach 0.25 to 0.5 at the final generation (**Figure 1**). This background LD was lowered to zero or close to zero on average, when correcting these statistics by the V matrix, as measured using r_v or cor_{PC1v}.

Second, we observed that selection models tended to generate more LD than the neutral model, as measured using r_v or cor_{PC1v}. For instance, in self-mating species, the COAD, COMP and ADD selection models all tended to generate more LD than the neutral model, with COAD and COMP generating more LD than the ADD model (**Figure 1**). In panmictic populations, however, only the COMP model generated a consistent LD, compared to the COAD and ADD models which generated low LD. Nevertheless, despite correcting for population structure, it remained difficult to distinguish epistatic selection from additive selection in self-mating populations at the haplotype level, as different haplotypes were under selection in each sub-population. This artifact was less prominent when focusing on the SNPs targeted by selection.

Finally, SNP-based LD measures (r/r_v) were more efficient than haplotype-based LD measures (cor_{PC1}/cor_{PC1v}) to detect epistatic selection (**Figure 1**). However, these measures could not capture any signal once allele fixation at one SNP or co-fixation at the two SNPs occurred. On the other hand, cor_{PC1}/cor_{PC1v} relied on SNP polymorphisms in the genomic region surrounding SNPs under selection, so that they could benefit from the hitch-hiking effect even after allele fixation at the selected SNPs.

False positive control and power of two-locus correlation statistics

On the assumption that values of the correlation coefficient r follow a Student distribution $\tau_{(n-2)}$ under the null hypothesis of independence between the two variables tested, we examined the fit of the statistics T_r, T_{corPC1}, T_{r_v} and T_{corPC1_v} to such a distribution. False Positive (FP) proportions of these statistics are given for different rejection quantiles of the Student distribution $\tau_{(n-2)}$ in **Table 2**. Two time points were considered in neutral simulations, at generation 140 in the midst of the time course, and at the last generation 300. At generation 140, in both the self-mating and random-mating models, T_r and T_{corPC1} showed excessively large FP proportions. For instance, FP proportions ranging from 55 to 81% were observed for a 1% type I error, while T_{r_v} and T_{corPC1_v} showed adequate, conservative FP proportions ranging

from 0.1 to 3% for the same 1% type I error (**Table 2**). At generation 300, a similar behaviour was observed, with FP proportions ranging from 74 to 91% for a 1% type I error, while T_{r_v} and T_{corPC1_v} showed adequate - though less conservative in the case of the self-mating model - FP proportions ranging from 1.1 to 22% for the same 1% type I error. This indicates that corrections for population structure and heterogeneous relatedness are necessary in order to perform statistical tests of the neutral hypothesis for a null correlation between two independent loci (two SNPs or two genomic regions), accounting for "noisy" neutral processes.

A power analysis of the T_r, T_{corPC1}, T_{r_v} and T_{corPC1_v} statistics was then performed by using simulated data (i) under the null hypothesis of neutrality and independence between loci and (ii) under each of the selection models and independence between loci. At both time-points (generations 140 and 300), a general trend was that the detection power with r/r_v and cor_{PC1}/cor_{PC1v} was higher for the COMP model than for the COAD or the ADD models (i.e. 25-50%, 10-65%, and 10-30%, respectively, for α=5% with r_v or cor_{PC1v} statistics), especially when considering random mating (**Figure 2**). In addition, the correction of LD-based measures by the kinship matrix (r_v/cor_{PC1v}) did not increase the detection power of epistatic selection; rather, it tended to reduce power, especially in the COMP model, but not in the COAD model. This is due to the fact that the fixation of the *AB* allelic combination was more frequent in sub-populations than in the whole population in the COMP model (see **Supplementary Figure S2**) - a consequence of unequal initial frequencies of the ancestral/derived alleles in the simulations -, leading to high LD values when population structure was not taken into account. Finally, cor_{PC1v} tended to show less power than r_v because of haplotype heterogeneity, namely when the same selected allele was associated to different haplotypes within a sub-population (**Figure 2A,C; Figure 2B,D**).

Detection of two-locus coadaptation in the *Medicago truncatula* plant

To illustrate the statistical testing of the correlation coefficient between two loci using T_{corPC1} or T_{corPC1_v}, a one-dimension GWESS was performed using a bait approach with the *MtSUNN* gene, which is a key regulator of nodulation in legumes, against the 48 339 other genes of the *M. truncatula* genome. Two scans were implemented, including SNP data from either the whole *M. truncatula* collection - n=262 individuals - or from the Far West (FW) sub-population - n=80 individuals - (**Figure 3A,C; Figure 3B,D**; respectively). A clear inflation towards small p-values could be observed for scans based on T_{corPC1} (**Figure 3A,B**) compared with scans based on T_{corPC1_v} (**Figure 3C,D**), and this inflation was higher with data from the whole collection that showed a higher degree of population structure. In the FW sub-population scan, a sharp peak was observed using T_{corPC1_v} on the chromosome 6 corresponding to the *MtCLE02* (Medtr6g009390) gene on top of the peak (**Figure 3D,** p-value = 1.7×10^{-8}). *MtCLE02* corresponded to the top candidate gene showing an epistatic selection signal outside of the chromosome 4 where *MtSUNN* was located. Whereas *MtCLE02* was also highly correlated with *MtSUNN* when considering T_{corPC1} (**Figure 3B,** p-value = 2.74×10^{-13}), several other genomic regions displayed similar or even more significant signals, which may indicate spurious genome-wide correlations. Interestingly, scans based on SNP data from the whole *M. truncatula* population did not reveal such strong signal in the genomic region containing *MtCLE02* (**Figure 3A,C;** p-value = 0.077 and 0.006 for T_{corPC1} and T_{corPC1_v}, respectively), indicating that in this specific case, epistatic selection may have occured at the sub-population level.

Experimental evidence for the genetic relationship between *MtSUNN* and *MtCLE02* in *Medicago truncatula*

The *MtSUNN* gene encodes a Leucine-Rich Repeats – Receptor Like Kinase (LRR-RLK) whereas the *MtCLE02* genes encodes a CLAVATA-like secreted signaling peptide. The SUNN receptor function, which is crucial in the systemic negative regulation of nodulation, was previously associated to other CLE secreted signaling peptide encoding genes, *MtCLE12*

and *MtCLE13* (Mortier *et al*, 2012; Mortier *et al*, 2010), but not with *MtCLE02*. Whereas their expression was induced by the inoculation with symbiotic rhizobia bacteria initiating nodulation, this was not the case for *MtCLE02* (**Supplementary Figure S4A,B**). These previously documented CLE/SUNN relationships pointed us to test for a putative functional interaction between CLE02 signaling peptides and the SUNN receptor. As previously performed for *MtCLE12* or *MtCLE13* (Mortier *et al*, 2012; Mortier *et al*, 2010), we used a genetic approach consisting in overexpressing comparatively the *MtCLE02* gene in *M. truncatula* Wild-Type (WT) and *sunn* mutant roots (**Figure 4)**. First, a quantification of the nodule number in WT versus *sunn* mutant roots highlighted the well-known supernodulation phenotype of the *sunn* mutant (Mann & Whitney – Wilcoxon test, p-value = $2x10^{-8}$). Second, the nodule number was significantly decreased when *MtCLE02* was overexpressed in WT roots, as validated by real time RT-PCR (**Supplementary Figure S4C**), indicating a negative role of *MtCLE02* on nodulation (**Figure 4**, Mann & Whitney - Wilcoxon test, p-value = $2x10^{-6}$**)**. Third, *MtCLE02* overexpression in *sunn* mutant roots did not impact the nodule number (Mann & Whitney – Wilcoxon test, p-value = 0.66), in contrast to what was observed in the WT, indicating that the negative role of CLE02 on nodulation relies on the SUNN receptor.

Detection of two-locus coadaptation in human populations

In human, two GWESS were performed on the world-wide sample of 940 individuals with SNP data from *SLC24A5* and *EDAR* genes, two major drivers of the external appearance which have been subjected to strong positive selection in human populations according to different studies (Basu Mallick *et al*, 2013; Beleza *et al*, 2013; Bryk *et al*, 2008; Sabeti *et al*, 2007; Speidel *et al*, 2019). Two bait SNPs, 15_46172199 (rs2250072) and 2_108973688 (rs6749207) located in *SLC24A5* and *EDAR* genes, respectively, were chosen for the GWESS. For each bait SNP, SNP-based statistics T_r and T_{r_v} were calculated for 431 950 genome-wide SNPs. Scans implemented in the world-wide population with SNPs 15_46172199 and

2_108973688, each as bait, were inflated towards small p-values when using the T_r statistic (**Figure 5A,B**), compared with scans implemented with the T_{r_v} statistic (**Figure 5C,D**). Using the SNP 15_46172199 as bait for *SLC24A5* gene, a peak corresponding to the *EDAR* gene was detected, with the SNP 2_108946170 as the top significant SNP (**Figure 5A,C**; T_{r_v}-based p-value = 2.29×10^{-9}). Conversely, when the scan was performed with the SNP 2_108973688 as bait from the *EDAR* gene, a peak corresponding to the *SCL24A5* gene was detected, with the SNP 15_46179457 as the top significant SNP (**Figure 5B,D**; T_{r_v}-based p-value = 1.2×10^{-12}). Genome-wide LD distributions between each bait SNP and all other SNPs (**Figure 5C,D**; top left of each panel) showed very high LD values between SNPs from *EDAR* and *SLC24A5*, thus indicating extremely significant signals (*EDAR* SNPs ranked among the top 0.01 to 0.18% SNPs in LD with *SLC24A5*; and *SLC24A5* SNPs ranked among the top 0.006% to 0.03% SNPs in LD with *EDAR*). The world-wide geographic distribution of genotypes at SNPs 15_46172199 - *SLC24A5* - and 2_108973688 - *EDAR* - (**Figure 6C**) correlated substantially with the global human population structure, as depicted by a phylogenetic tree based on the kinship matrix among individuals (**Figure 6A,B**). Indeed, the derived allele at SNP 15_46172199, associated with the light skin allele at the *SCL24A5* gene, was present in Europe, North Africa, Middle East and Central South Asia; and the derived allele at SNP 2_108973688, associated with the thick hair allele at the *EDAR* gene, was present in East-Asia, America, and Oceania (**Figure 6C**). The strong LD signature observed in the world-wide samples between *SCL24A5* and *EDAR*, as measured with T_r, therefore reflected the selection of derived alleles in different geographic regions, and thus a correlation with the global population structure. However, the LD measured with T_{r_v} was still highly significant between *SCL25A5* and *EDAR*, indicating that epistatic selection may have occurred between both genes at the level of geographic sub-regions.

In order to localize the geographic origin of such selection signature, GWESSs were performed within six geographic regions of the world-wide sample: Central South Asia, East Asia, Sub-Saharan Africa, Middle East, Europe and America (**Supplementary Figures S5-S10**). Only the GWESS performed in Central South Asia indicated a significant LD between SNPs at the *SCL24A5* and *EDAR* genes (**Supplementary Figure S5C**, T_{r_v}-based p-value = 6.7x10^{-6} at SNP 2_108973688; **Supplementary Figure S5D**, T_{r_v}-based p-value = 2.8x10^{-6} at SNP 15_46174380).

Human population samples from the HGDP-CEPH dataset in Central South Asia are composed of eight different ethnic groups from Pakistan. To search for local signals, LD tests were performed with T_{r_v} between two candidate SNPs within *SCL24A5* (15_46179457 (rs1834640) and 15_46172199), and three candidates SNPs within *EDAR* (2_108962124 (rs260607), 2_108982808 (rs17034770) and 2_108973688), for the 50 ethnic groups or populations distributed within eight geographic regions and showing polymorphism at all five SNPs. Average and standard deviation of -log$_{10}$(p-value) across six pairwise SNP comparisons (**Figure 6,D**) strongly supported a high LD between *SCL24A5* and *EDAR* in the Burusho ethnic group from Pakistan (3.2 and 0.36, respectively), as also highlighted by genotypes of Burusho individuals (**Figure 6,C**, r_v = 0.63 for genotypes between the two dotted lines). This pattern of high LD between *SCL24A5* and *EDAR* in Burusho did not seem to be generated by any population sub-structure in this ethnic group, since LD tests performed with T_r resulted in a similar average (3.18) and standard deviation (0.36) of -log$_{10}$(p-value) (**Supplementary Figure S11**).

Discussion

We introduced a statistical method which can detect the signature of epistatic selection using LD between two loci. SNP-based (T_{r_v}) and window-based (T_{corPC1_v}) statistics, which take

443 into account the underlying population structure and relatedness among individuals, are shown
444 to fit a Student distribution $\tau_{(n-2)}$, allowing to easily and quickly test for significance of
445 correlation coefficients in the frame of GWESS using either a candidate SNP, a gene, or a short
446 genomic region as bait. Simulations have shown that T_{r_v} and T_{corPC1_v} showed equivalent or less
447 power than T_r or T_{corPC1} to detect epistatic selection occurring simultaneously in all sub-
448 populations, ranging from 10 to 65% (assuming a 5% type I error) depending on the epistatic
449 selection model and mating process. Thus, selection signals in local populations could be more
450 difficult to detect with T_{r_v} and T_{corPC1_v} because in these cases, selection can be correlated with
451 population structure. In addition, the impact of the kinship matrix on LD correction changes
452 depending on the scale of the sampling, with a stronger impact for large scales of geographic
453 population structure than for smaller, less structured, and less heterogeneous geographic scales.
454 These features suggest that GWESS should be performed not only on a global sample
455 comprising individuals from different populations, but also on samples from different sub-
456 populations in order to search for more population-specific patterns of epistatic selection.
457 Despite the power was not increased when using T_{r_v} and T_{corPC1_v}, simulations as well as
458 analyses performed both in *M. truncatula* and human genomes strongly supported their use, in
459 order to efficiently control for false positives. Interestingly, although SNP-based statistics
460 (T_r, T_{r_v}) may tend to show an increased power than window-based statistics (T_{corPC1}, T_{corPC1_v})
461 because they are not sensitive to haplotype heterogeneity and because mutations under epistatic
462 selection at both loci can be directly tested by SNP-based statistics, window-based statistics
463 also show several advantages. First, window-based statistics are faster to implement at the
464 genome scale, notably in two-dimensional GWESS. Second, the window size can be fixed at a
465 value that fits best the average LD decay in the species studied, even though a standard 10 kbp
466 size can be used by default. Finally, window-based statistics allow detecting coevolving genes

even after the putative fixation of coselected SNPs, because surrounding SNPs within genes or windows also carry a selection signal by hitchhiking.

The applications of our method to SNP data from human populations or from populations of the model plant *M. truncatula* allowed in both cases to identify a couple of genes most probably under epistatic selection, or at least under coselection. In *M. truncatula*, additional experiments revealed a genetic interaction likely shaped by epistatic selection between the *MtSUNN* and *MtCLE02* genes. The polymorphism at the *MtSUNN* gene could be driven by balancing selection at the local level because the *H* statistic (Fay and Wu, 2000) is 1.45 in the Far West population and ranks among the highest 8.41% in the entire genome. On the other hand, $H = 0.52$ based on the whole species, ranking among the highest 39.5%, according to a set of 47 875 genes. The polymorphism at the *MtCLE02* gene seems more affected by an ongoing soft sweep that can be detected at the level of the whole species ($H = -2.25$, ranking among the lowest 8.98% of the genome). Still, both genes maintain polymorphisms, and epistatic selection could drive this pattern. As a proof of concept, a genetic approach was used to demonstrate the functional interaction between the CLE02 signaling peptide and the SUNN receptor in the context of symbiotic root nodulation. Indeed, the CLE02 signaling peptide negatively affected the number of nodule organs on the plant root system depending on the SUNN receptor, as previously shown for other CLE peptide encoding genes, *MtCLE12* and *MtCLE13* (Gautrat *et al*, 2019; Mortier *et al*, 2012; Mortier *et al*, 2010). Interestingly, it should be noted that the *MtCLE02* gene is, in contrast to *MtCLE12* and *MtCLE13*, not regulated by symbiotic nodulation conditions and not phylogenetically closely related to previously characterized CLE peptide encoding genes shown to have a related negative impact on nodule number (Hastwell *et al*, 2017). The identification of a novel CLE peptide acting in this genetic pathway highlights the discovery power of our method to

functionally associate gene pairs independently of their expression pattern or of a coregulation pattern criterion.

In human, the *SLC24A5* gene, a major driver of variation in skin pigmentation, has been shown to be under positive selection in the European population (Deng and Xu, 2018; Izagirre *et al*, 2006; Sabeti *et al*, 2007). The causal mutation for the light skin phenotype was not present in HGDP data (SNP rs1426654, position: 15_46213776), but SNPs used (i.e. 15_46179457 and 15_46172199) were located on the same *SLC24A5* haplotype background that was previously characterized (Basu Mallick *et al*, 2013; Beleza *et al*, 2013; Crawford *et al*, 2017). In addition, the V370A mutation in the *EDAR* gene coding for a receptor related to TNFα receptors and involved in driving hair structure, as well as teeth and sweat glands development, was shown to be under positive selection in East Asia and in native Americans, and to increase hair thickness (Bryk *et al*, 2008; Sadier *et al*, 2014; Speidel *et al*, 2019). However, just as for *SLC24A5*, the causal mutation in the *EDAR* gene was not present in HGDP data, but SNPs used (i.e. 2_108962124, 2_108973688 and 2_108982808) were located within the genomic sequence of *EDAR*. The geographic distribution of genotypes at SNPs 15_46172199 and 2_108973688 strongly correlated with the world-wide human population structure, which explained the high LD observed at this level in the scans implemented with T_r. However, scans implemented with T_{r_v} indicated a persistence of top SNPs in LD between *SLC24A5* and *EDAR* in the world-wide sample, which might be due to epistatic selection or coselection. We identified such selection signature in Central South Asia, with the Burusho ethnic group from Pakistan as being one possible geographic origin. The GWESS performed in Central South Asia, and subsequent LD tests performed between *SLC24A5* and *EDAR* in sub-populations with T_r or T_{r_v} statistics showed similar results, indicating a weak effect of population structure, as previously observed within this geographic region and in India (Rosenberg *et al*, 2006). The Burusho showed a predominant association between the derived alleles at *SLC24A5* and the ancestral alleles at

516 *EDAR*, which is indicative of the persistence of a typical European light skin and thin hair
517 structure in this ethnic group. To date, no functional link is yet demonstrated between the
518 critical skin pigmentation-related *SCL24A5* gene and the ectodysplasin pathway to which
519 *EDAR* belongs. While the long-distance LD between *EDAR* and *SCL24A5* largely reflects
520 strong positive selection acting independently on these two genes in different geographic
521 regions, our results also suggest that coselection of these genes may have locally contributed to
522 the phenotypic make-up of human populations.

523 Our method was fast to implement on a computer equipped with an Intel Xeon E5-2640
524 v4 processor (10 Cores and 2.4-3.4 GHz performance, supplied by 256 Go memory), as only
525 few minutes were needed to estimate the kinship matrix, depending on the sample size (e.g. 30
526 sec and ~ 3 min for *Medicago* and human, respectively), and to implement one GWESS (e.g. 2
527 min and 12 min for 48 339 *Medicago* genes with T_{corPC1_v} and 431 950 human SNPs with T_{r_v},
528 respectively). Although functional analyses tools for genetic interaction are still limited to a
529 few model species, GWESS can easily be performed on non-model organisms, as long as
530 sufficiently dense SNP data are available, to identify candidate genes under epistatic selection
531 that may be relevant to study in related model species for instance. If GWESS is not directly
532 applicable, possible relationships between candidate genes could be directly tested using
533 dedicated SNP markers, but one should be cautious about the lack of benchmark with other
534 genes. A natural extension of the one-dimension use of T_{r_v} or T_{corPC1_v} - based tests is the
535 implementation of two-dimensional GWESS in which the correlation of each polymorphic
536 locus (SNP, gene or genomic region) in the genome would be tested against all remaining
537 polymorphic loci, thanks to parallel computations on bioinformatics platforms. We anticipate
538 that such an approach will open the way towards exploring evolutionary coadapted gene
539 networks.

540

Acknowledgements

This work was supported by the "DeCoD" project funded by the French Agence Nationale de la Recherche (grant number ANR-16-CE20-0017-01). The PhD position of Léa Boyrie was funded by the "DeCoD" project. We thank the bioinformatics platform Toulouse Midi-Pyrenees (Genotoul). This work was performed in the LRSV (Toulouse, France), part of the "Laboratoire d'Excellence" (LABEX) entitled TULIP (grant number ANR-10-LABX-41). We thank Carole Laffont (IPS2, CNRS, Gif-sur-Yvette, France) for providing results about *MtCLE13* expression. Work in the Florian Frugier laboratory has benefited from a French State grant (Saclay Plant Sciences, grant number ANR-17-EUR-0007, EUR SPS-GSR) and an ANR grant ("PSYCHE", grant number ANR-16-CE20-0009-01). We thank Thomas Bataillon, two other anonymous reviewers, and Pierre-Marc Delaux for useful criticisms and comments to improve the manuscript.

Competing interests

The authors declare no financial competing interests.

Data Availability

The *M. truncatula* SNP dataset (hapmap format) used in this study can be retrieved at http://www.medicagohapmap.org/downloads/mt40. The Human SNP dataset from the HGDP-CEPH Human Genome Diversity Panel can be retrieved at ftp://ftp.cephb.fr/hgdp_supp1 (http://www.cephb.fr/hgdp/). R scripts to implement the statistical test based on T_r, T_{r_v}, T_{corPC1} or T_{corPC1_v}, along with an example dataset, are available at https://github.com/leaboyrie/LD_corpc1.

References

Ahrens CW, Rymer PD, Stow A, Bragg J, Dillon S, Umbers KDL *et al* (2018). The search for loci under selection: trends, biases and progress. *Mol Ecol* **27**(6)**:** 1342-1356.

Bamshad M, Wooding SP (2003). Signatures of natural selection in the human genome. *Nat Rev Genet* **4**(2)**:** 99-111.

Basu Mallick C, Iliescu FM, Möls M, Hill S, Tamang R, Chaubey G *et al* (2013). The light skin allele of SLC24A5 in South Asians and Europeans shares identity by descent. *PLoS Genet* **9**(11)**:** e1003912.

Beleza S, Santos AM, McEvoy B, Alves I, Martinho C, Cameron E *et al* (2013). The timing of pigmentation lightening in Europeans. *Mol Biol Evol* **30**(1)**:** 24-35.

Bonhomme M, André O, Badis Y, Ronfort J, Burgarella C, Chantret N *et al* (2014). High-density genome-wide association mapping implicates an F-box encoding gene in Medicago truncatula resistance to Aphanomyces euteiches. *New Phytol* **201**(4)**:** 1328-1342.

Bonhomme M, Boitard S, Clemente HS, Dumas B, Young N, Jacquet C (2015). Genomic Signature of Selective Sweeps Illuminates Adaptation of Medicago truncatula to Root-Associated Microorganisms. *Molecular Biology and Evolution* **32**(8)**:** 2097-2110.

Bonhomme M, Fariello MI, Navier H, Hajri A, Badis Y, Miteul H *et al* (2019). A local score approach improves GWAS resolution and detects minor QTL: application to Medicago truncatula quantitative disease resistance to multiple Aphanomyces euteiches isolates. *Heredity (Edinb)* **123**(4)**:** 517-531.

Botchkarev VA, Fessing MY (2005). Edar signaling in the control of hair follicle development. *J Investig Dermatol Symp Proc* **10**(3)**:** 247-251.

Brachi B, Meyer CG, Villoutreix R, Platt A, Morton TC, Roux F *et al* (2015). Coselected genes determine adaptive variation in herbivore resistance throughout the native range of Arabidopsis thaliana. *Proc Natl Acad Sci U S A* **112**(13)**:** 4032-4037.

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**(19)**:** 2633-2635.

Branca A, Paape TD, Zhou P, Briskine R, Farmer AD, Mudge J *et al* (2011). Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume Medicago truncatula. *Proc Natl Acad Sci U S A* **108**(42)**:** E864-870.

Bryk J, Hardouin E, Pugach I, Hughes D, Strotmann R, Stoneking M *et al* (2008). Positive selection in East Asians for an EDAR allele that enhances NF-kappaB activation. *PLoS One* **3**(5)**:** e2209.

Burgarella C, Chantret N, Gay L, Prosperi JM, Bonhomme M, Tiffin P *et al* (2016). Adaptation to climate through flowering phenology: a case study in Medicago truncatula. *Mol Ecol*.

Caicedo AL, Stinchcombe JR, Olsen KM, Schmitt J, Purugganan MD (2004). Epistatic interaction between Arabidopsis FRI and FLC flowering time genes generates a latitudinal cline in a life history trait. *Proc Natl Acad Sci U S A* **101**(44)**:** 15670-15675.

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L *et al* (2002). A human genome diversity cell line panel. *Science* **296**(5566)**:** 261-262.

Crawford NG, Kelly DE, Hansen MEB, Beltrame MH, Fan S, Bowman SL *et al* (2017). Loci associated with skin pigmentation identified in African populations. *Science* **358**(6365).

Csilléry K, Lalagüe H, Vendramin GG, González-Martínez SC, Fady B, Oddou-Muratorio S (2014). Detecting short spatial scale local adaptation and epistatic selection in climate-related candidate genes in European beech (Fagus sylvatica) populations. *Mol Ecol* **23**(19)**:** 4696-4708.

Daub JT, Hofer T, Cutivet E, Dupanloup I, Quintana-Murci L, Robinson-Rechavi M *et al* (2013). Evidence for polygenic adaptation to pathogens in the human genome. *Mol Biol Evol* **30**(7)**:** 1544-1558.

De Mita S, Chantret N, Loridon K, Ronfort J, Bataillon T (2011). Molecular adaptation in flowering and symbiotic recognition pathways: insights from patterns of polymorphism in the legume Medicago truncatula. *BMC Evol Biol* **11:** 229.

Deng L, Xu S (2018). Adaptation of human skin color in various populations. *Hereditas* **155:** 1.

Fay JC, Wu CI (2000). Hitchhiking under positive Darwinian selection. *Genetics* **155**(3)**:** 1405-1413.

Gautrat P, Mortier V, Laffont C, De Keyser A, Fromentin J, Frugier F *et al* (2019). Unraveling new molecular players involved in the autoregulation of nodulation in Medicago truncatula. *J Exp Bot* **70**(4)**:** 1407-1417.

Glémin S (2007). Mating systems and the efficacy of selection at the molecular level. *Genetics* **177**(2)**:** 905-916.

Glémin S, Bazin E, Charlesworth D (2006). Impact of mating systems on patterns of sequence polymorphism in flowering plants. *Proc Biol Sci* **273**(1604)**:** 3011-3019.

Grzeskowiak L, Stephan W, Rose LE (2014). Epistatic selection and coadaptation in the Prf resistance complex of wild tomato. *Infect Genet Evol* **27:** 456-471.

Haasl RJ, Payseur BA (2016). Fifteen years of genomewide scans for selection: trends, lessons and unaddressed genetic sources of complication. *Mol Ecol* **25**(1)**:** 5-23.

Hartfield M, Bataillon T (2020). Selective Sweeps Under Dominance and Inbreeding. *G3 (Bethesda)* **10**(3)**:** 1063-1075.

Hastwell AH, de Bang TC, Gresshoff PM, Ferguson BJ (2017). CLE peptide-encoding gene families in Medicago truncatula and Lotus japonicus, compared with those of soybean, common bean and Arabidopsis. *Sci Rep* **7**(1)**:** 9384.

Hill WG, Robertson A (1968). Linkage disequilibrium in finite populations. *Theor Appl Genet* **38**(6)**:** 226-231.

Hu XS, Hu Y (2015). Genomic Scans of Zygotic Disequilibrium and Epistatic SNPs in HapMap Phase III Populations. *PLoS One* **10**(6)**:** e0131039.

Id-Lahoucine S, Molina A, Cánovas A, Casellas J (2019). Screening for epistatic selection signatures: A simulation study. *Sci Rep* **9**(1)**:** 1026.

Izagirre N, García I, Junquera C, de la Rúa C, Alonso S (2006). A scan for signatures of positive selection in candidate loci for skin pigmentation in humans. *Mol Biol Evol* **23**(9)**:** 1697-1706.

Kang Y, Sakiroglu M, Krom N, Stanton-Geddes J, Wang M, Lee YC *et al* (2015). Genome-wide association of drought-related and biomass traits with HapMap SNPs in Medicago truncatula. *Plant Cell Environ* **38**(10)**:** 1997-2011.

Lamason RL, Mohideen MA, Mest JR, Wong AC, Norton HL, Aros MC *et al* (2005). SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science* **310**(5755)**:** 1782-1786.

Le Signor C, Aimé D, Bordat A, Belghazi M, Labas V, Gouzy J *et al* (2017). Genome-wide association studies with proteomics data reveal genes important for synthesis, transport and packaging of globulins in legume seeds. *New Phytol* **214**(4)**:** 1597-1613.

Li H, Ralph P (2019). Local PCA Shows How the Effect of Population Structure Differs Along the Genome. *Genetics* **211**(1)**:** 289-304.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S *et al* (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**(5866)**:** 1100-1104.

Mangin B, Siberchicot A, Nicolas S, Doligez A, This P, Cierco-Ayrolles C (2012). Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity (Edinb)* **108**(3)**:** 285-291.

McVean G (2009). A genealogical interpretation of principal components analysis. *PLoS Genet* **5**(10)**:** e1000686.

694 Mortier V, De Wever E, Vuylsteke M, Holsters M, Goormachtig S (2012). Nodule numbers are governed
695 by interaction between CLE peptides and cytokinin signaling. *Plant J* **70**(3)**:** 367-376.

696
697 Mortier V, Den Herder G, Whitford R, Van de Velde W, Rombauts S, D'Haeseleer K *et al* (2010). CLE
698 peptides control Medicago truncatula nodulation locally and systemically. *Plant Physiol* **153**(1)**:** 222-
699 237.

700
701 Nielsen R (2005). Molecular signatures of natural selection. *Annu Rev Genet* **39:** 197-218.

702
703 Nordborg M (2000). Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph
704 with partial self-fertilization. *Genetics* **154**(2)**:** 923-929.

705
706 Ohta T (1982a). Linkage disequilibrium due to random genetic drift in finite subdivided populations.
707 *Proc Natl Acad Sci U S A* **79**(6)**:** 1940-1944.

708
709 Ohta T (1982b). Linkage disequilibrium with the island model. *Genetics* **101**(1)**:** 139-155.

710
711 Oleksyk TK, Smith MW, O'Brien SJ (2010). Genome-wide scans for footprints of natural selection. *Philos*
712 *Trans R Soc Lond B Biol Sci* **365**(1537)**:** 185-205.

713
714 Otto SP, Whitlock MC (2009). The impact of epistatic selection on the genomic traces of selection. *Mol*
715 *Ecol* **18**(24)**:** 4985-4987.

716
717 Paape T, Bataillon T, Zhou P, J Y Kono T, Briskine R, Young ND *et al* (2013). Selection, genome-wide
718 fitness effects and evolutionary rates in the model legume Medicago truncatula. *Mol Ecol* **22**(13)**:** 3525-
719 3538.

720
721 Pavlidis P, Alachiotis N (2017). A survey of methods and tools to detect recent and strong positive
722 selection. *J Biol Res (Thessalon)* **24:** 7.

723
724 Pecrix Y, Staton SE, Sallet E, Lelandais-Brière C, Moreau S, Carrère S *et al* (2018). Whole-genome
725 landscape of Medicago truncatula symbiotic genes. *Nat Plants* **4**(12)**:** 1017-1025.

726
727 Piskol R, Stephan W (2008). Analyzing the evolution of RNA secondary structures in vertebrate introns
728 using Kimura's model of compensatory fitness interactions. *Mol Biol Evol* **25**(11)**:** 2483-2492.

729
730 Pool JE (2015). The Mosaic Ancestry of the Drosophila Genetic Reference Panel and the D.
731 melanogaster Reference Genome Reveals a Network of Epistatic Fitness Interactions. *Mol Biol Evol*
732 **32**(12)**:** 3236-3251.

733
734 Rey T, Bonhomme M, Chatterjee A, Gavrin A, Toulotte J, Yang W *et al* (2017). The Medicago truncatula
735 GRAS protein RAD1 supports arbuscular mycorrhiza symbiosis and Phytophthora palmivora
736 susceptibility. *J Exp Bot* **68**(21-22)**:** 5871-5881.

Rogers AR, Huff C (2009). Linkage disequilibrium between loci with unknown phase. *Genetics* **182**(3)**:** 839-844.

Ronfort J, Bataillon T, Santoni S, Delalande M, David JL, Prosperi JM (2006). Microsatellite diversity and broad scale geographic structure in a model legume: building a set of nested core collection for studying naturally occurring variation in Medicago truncatula. *BMC Plant Biol* **6:** 28.

Rosenberg NA, Mahajan S, Gonzalez-Quevedo C, Blum MG, Nino-Rosales L, Ninis V *et al* (2006). Low levels of genetic divergence across geographically and linguistically diverse populations from India. *PLoS Genet* **2**(12)**:** e215.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C *et al* (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**(7164)**:** 913-918.

Sadier A, Viriot L, Pantalacci S, Laudet V (2014). The ectodysplasin pathway: from diseases to adaptations. *Trends Genet* **30**(1)**:** 24-31.

Schumer M, Brandvain Y (2016). Determining epistatic selection in admixed populations. *Mol Ecol* **25**(11)**:** 2577-2591.

Slatkin M (2008). Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* **9**(6)**:** 477-485.

Speidel L, Forest M, Shi S, Myers SR (2019). A method for genome-wide genealogy estimation for thousands of samples. *Nat Genet* **51**(9)**:** 1321-1329.

Stanton-Geddes J, Paape T, Epstein B, Briskine R, Yoder J, Mudge J *et al* (2013). Candidate genes and genetic architecture of symbiotic and agronomic traits revealed by whole-genome, sequence-based association genetics in *Medicago truncatula*. *PLoS One* **8**(5)**:** e65688.

Takahasi K (2009). Coalescent under the evolution of coadaptation. *Molecular Ecology* **18**(24)**:** 5018-5029.

Takahasi KR (2007). Evolution of coadaptation in a subdivided population. *Genetics* **176**(1)**:** 501-511.

Takahasi KR, Innan H (2008). The direction of linkage disequilibrium: a new measure based on the ancestral-derived status of segregating alleles. *Genetics* **179**(3)**:** 1705-1712.

Takahasi KR, Tajima F (2005). Evolution of coadaptation in a two-locus epistatic system. *Evolution* **59**(11)**:** 2324-2332.

Tang H, Krishnakumar V, Bidwell S, Rosen B, Chan A, Zhou S *et al* (2014). An improved genome release (version Mt4.0) for the model legume Medicago truncatula. *BMC Genomics* **15:** 312.

Vitti JJ, Grossman SR, Sabeti PC (2013). Detecting natural selection in genomic data. *Annu Rev Genet* **47:** 97-120.

Weigand H, Leese F (2018). Detecting signatures of positive selection in non-model species using genomic data. *Zoological Journal of the Linnean Society* **184**(2)**:** 528–583.

Weir BS (1979). Inferences about linkage disequilibrium. *Biometrics* **35**(1)**:** 235-254.

Yoder JB, Stanton-Geddes J, Zhou P, Briskine R, Young ND, Tiffin P (2014). Genomic signature of adaptation to climate in Medicago truncatula. *Genetics* **196**(4)**:** 1263-1275.

Zhang W, Collins A, Gibson J, Tapper WJ, Hunt S, Deloukas P *et al* (2004). Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps. *Proc Natl Acad Sci U S A* **101**(52)**:** 18075-18080.

Figures Legends

Figure 1. Evolution of inter-locus epistatic selection detected with linkage disequilibrium on simulated data.

Evolution of inter-locus LD in a self-mating simulation model calculated on a SNP-based scale with r or r_v (**A**), and on a window-based scale with cor_{PC1} or cor_{PC1v} (**B**). Evolution of inter-locus LD in a random mating simulation model calculated on a SNP-based scale (**C**) and on a window-based scale (**D**). Fixation rates in (A) and (C) depict co-fixation of *a* and *b* coselected mutant alleles over generations, showing the selection efficiency. Note that mutations under selection are codominant. In order to perform proper comparisons between selection models and to avoid sampling bias on the average LD in the COAD and ADD models, we selected 500 simulations among those in which the outcome at the last generation was the co-fixation of the derived alleles *a* and *b* in both sub-populations; and in the COMP selection model, we randomly sampled 500 simulations (*i.e.* simulations showing fixation of the *AB* or *ab* combination, or still showing polymorphism at the last generation). For an increased visibility, the evolution of the

812 average two-locus LD is depicted using absolute correlation values. Note that curves stopped
813 at different time-points for different scenarios, indicating that fixation has been reached for one
814 or both SNPs under epistatic selection, and thus that r and r_v are no longer computable.
815
816 **Figure 2. Detection power of epistatic selection models for SNP-based and window-based**
817 **LD measures.**

818 The detection power of epistatic selection in a self-mating simulation model and in a random
819 mating model was calculated on a SNP-based scale - r_v and r, full and dotted curves,
820 respectively - (**A, B**), and on a window-based scale - cor_{PC1v} and cor_{PC1}, full and dotted curves,
821 respectively - (**C**-**F**). Figures (**A-D**) depict the detection power at generation 140 and figures
822 (**E**, **F**) at generation 300 (r and r_v are no longer computable at this generation in coadapted and
823 additive selection models; see **Figure 1**). The *x*-axis corresponds to the type I error (α) and the
824 y-axis to the detection power (1-β). Mutations under selection are codominant. For each
825 statistic, neutral simulations were used to estimate one-way rejection quantiles by using the
826 absolute values of the statistic, corresponding to type I errors α ranging from 0.001 to 0.20.
827 Then, for each selection model under self-mating or random mating with codominant
828 mutations, we calculated the proportion of simulations where absolute values of each T statistic
829 were higher than each rejection quantile. The power was calculated for T_r, T_{corPC1}, T_{r_v} and
830 T_{corPC1_v} at generation 140, where allele fixation at SNPs under selection was not yet achieved,
831 and also at the last generation 300 for T_{corPC1} and T_{corPC1_v} window-based measures.
832
833 **Figure 3. LD distribution between the bait gene *MtSUNN* and all genes of *M. truncatula***
834 **genome.**

LD between the *MtSUNN* gene (framed) and all *M. truncatula* genes was calculated in the entire population (**A, C**) and in the Far-West population (**B, D**). The p-values of the correlation tests were calculated from T_{corPC1} (**A, B**) and from T_{corPC1_v} statistics **(C, D**). The *x*-axis corresponds to gene positions spanning the eight chromosomes, each point corresponding to a gene and red dots depicting the *MtCLE02* gene in each figure. The y-axis shows the -$\log_{10}$(p-value) of the test of the correlation coefficient.

Figure 4. Experimental validation of the CLE02 signaling peptide / SUNN receptor genetic relationship in *M. truncatula* symbiotic nodulation.

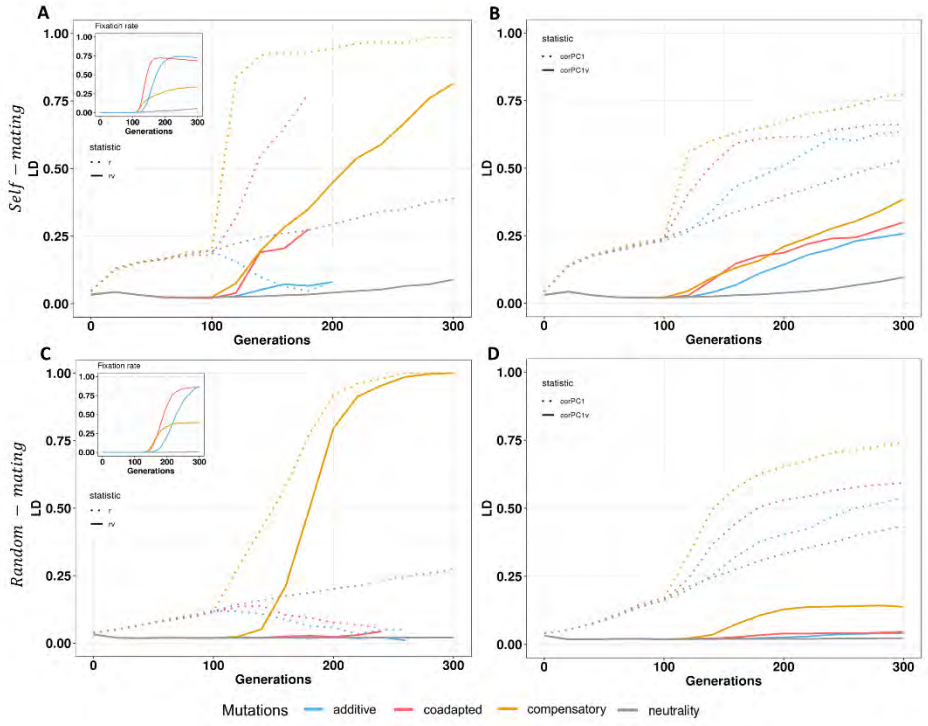
(A) Representative images of nodulated roots, 14 days post rhizobium inoculation, overexpressing the *MtCLE02* gene (Ubi:CLE02) or a *GUS* control gene (Ubi:GUS) either in Wild-Type (WT) plants or in the *sunn* mutant. Scale bar = 1 cm. **(B)** Boxplots of the number of nodules in the same conditions as described in **A**. A Mann & Whitney Wilcoxon rank sum test was used to assess pairwise statistical differences, as indicated within the graph.

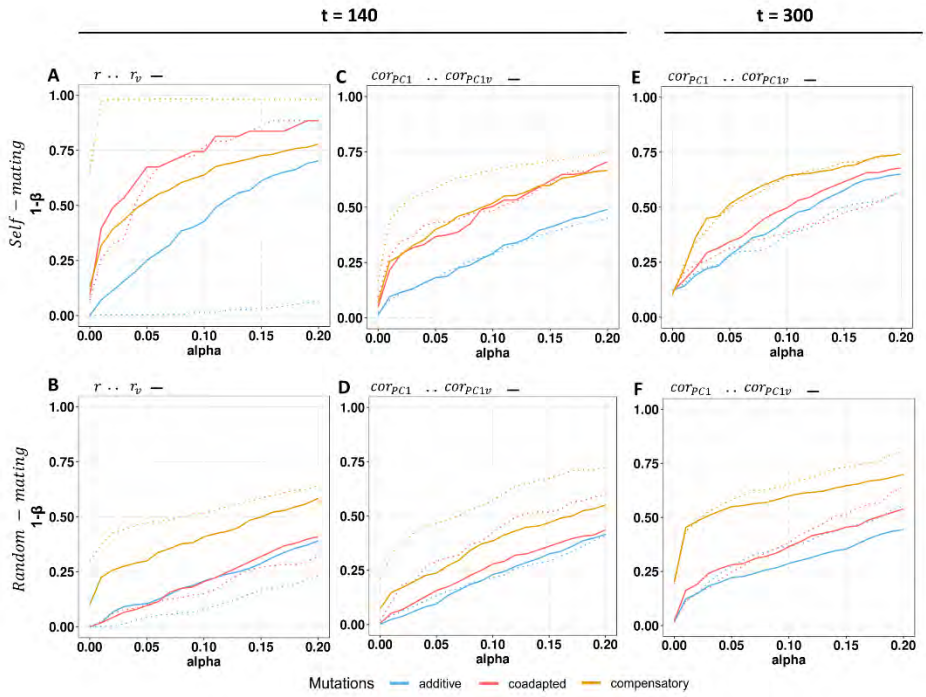
Figure 5. LD distribution between the bait SNPs of *SLC24A5* and *EDAR* genes and all other HGDP-CEPH SNPs in the whole human population samples (n=952).

LD between SNP 15_46172199 (*SLC24A5)* or SNP 2_108973688 (*EDAR*), respectively, and all other SNPs of the genome is tested using T_r (**A, B**) or T_{r_v} (**C, D**). The *x*-axis corresponds to SNP positions spanning the 22 human autosomes, each point corresponds to a SNP and the black points depict SNPs at candidate genes in epistatic selection with one SNP at the bait gene (vertical dotted line) in each figure. The y-axis is the -$\log_{10}$(p-value) of the test of the correlation coefficient. Plots at the top left of each figure show the distribution of LD between each bait SNP and all other SNPs of the genome. LD values between the bait SNP of *SLC24A5* and the target top SNPs of *EDAR* (respectively the bait SNP of *EDAR* and the target top SNPs of *SLC24A5*) is represented by an arrow.

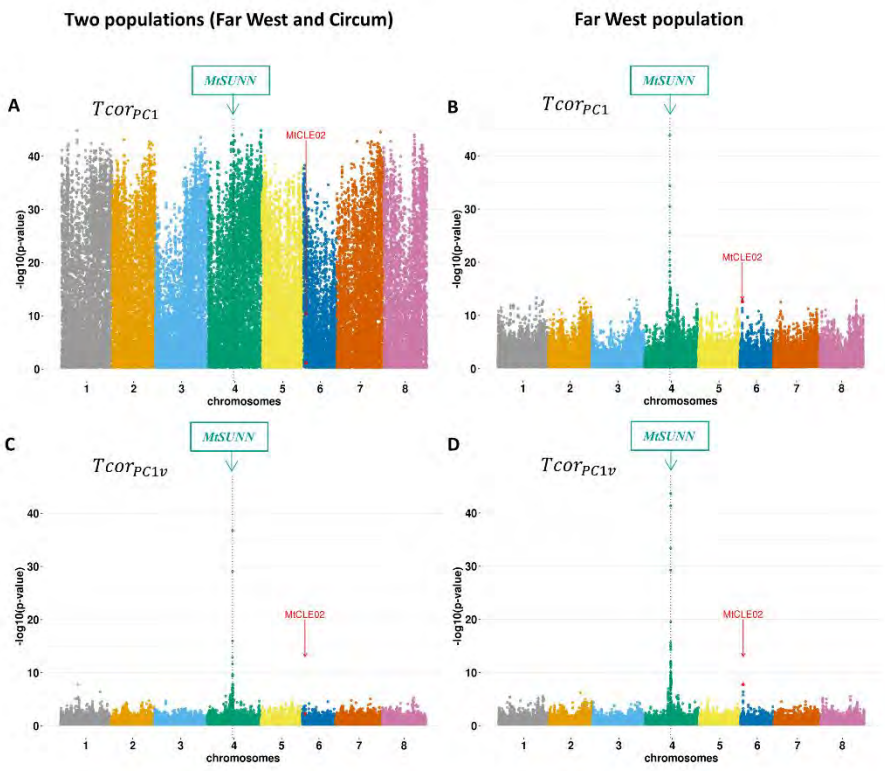
Figure 6. Schematic human population structure inferred from the kinship matrix, the geographic distribution of alleles, and the LD between *SLC24A5* and *EDAR*.

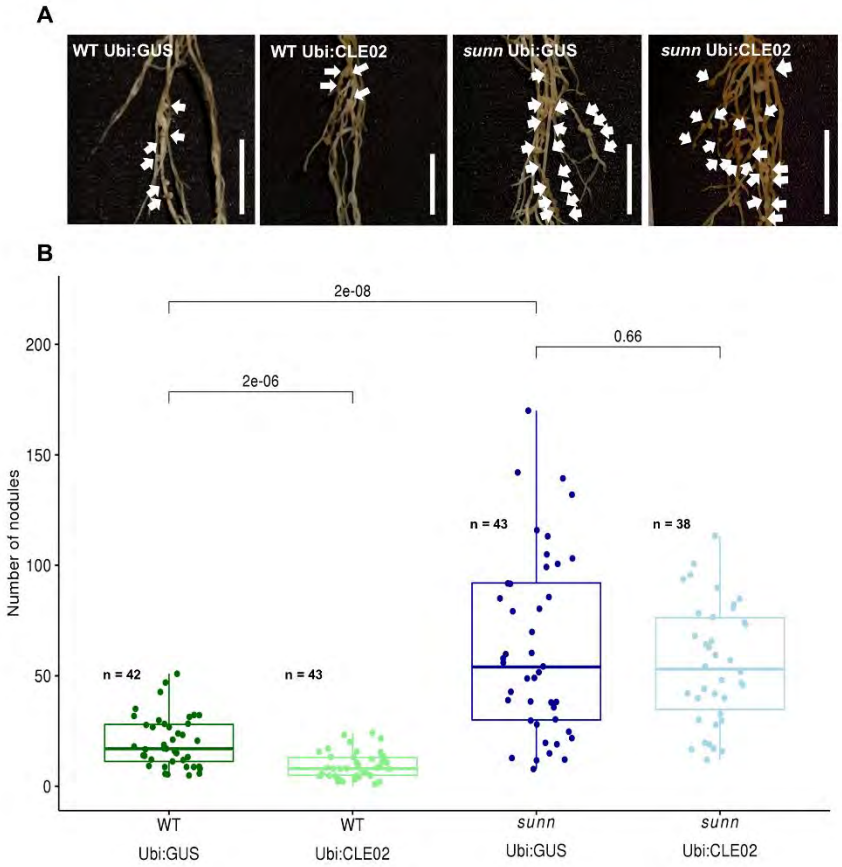
(**A**) Neighbor-Joining tree inferred from the molecular kinship matrix based on 431 951 SNPs from the HGDP-CEPH database showing the global human population structure. (**B**) Same tree as in (A) showing the clustering of the different sub-populations or ethnic groups sampled. (**C**) Bar plots depicting geographic distributions of genotypes at SNP 2_108973688 (*EDAR*) and SNP 15_46172199 (*SLC24A5*), highlighting LD patterns mainly due to the global population structure and to the selection of the derived alleles (coded 1) at the two genes. (**D**) Average and standard error of LD significances based on T_{r_v} statistics between SNPs of *SLC24A5* (SNP 15_46172199, SNP 15_46179457) and *EDAR* (SNP 2_108962124, SNP 2_108973688, SNP 2_108982808) within each human sub-population. The bar plot pinpoints Central South Asia as the main source of within population LD probably due to coselection of alleles in Pakistan ethnic groups (mainly from Burusho, delimited by two vertical dotted lines in (B) and (C))

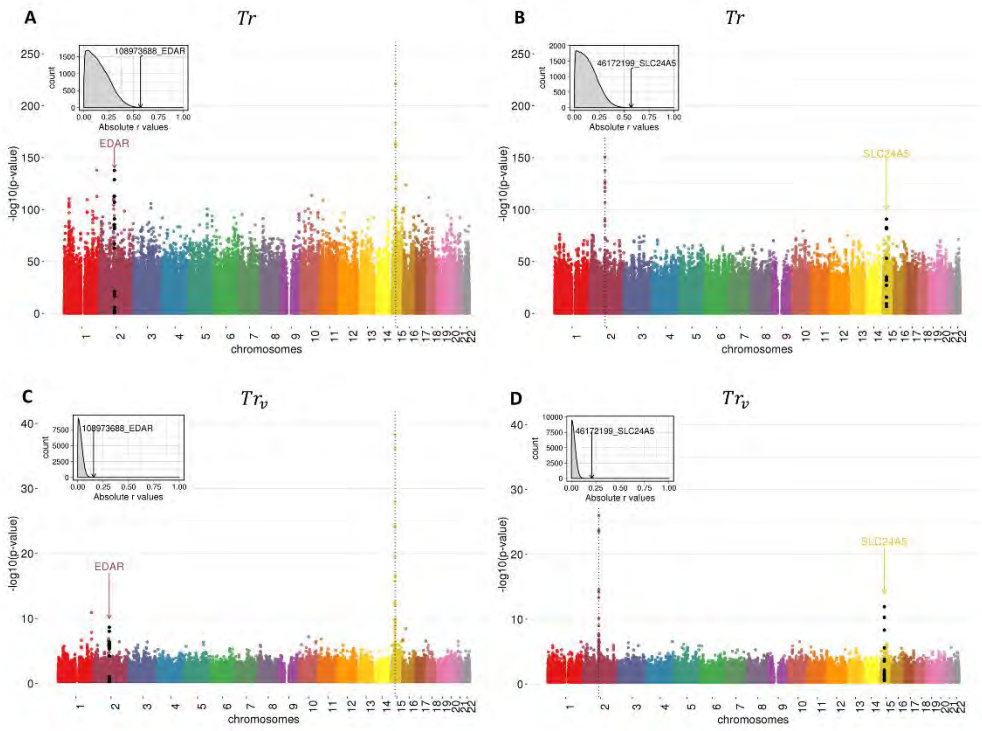

A
B
C
D
Self − mating
Random − mating
LD
Generations
Fixation rate
statistic
r
rv
corPC1
corPC1v
Mutations
additive
coadapted
compensatory
neutrality


t = 140
t = 300
A
B
C
D
E
F
r .. r_v —
cor_{PC1} .. cor_{PC1v} —
Self − mating
Random − mating
1-β
alpha
0.00
0.05
0.10
0.15
0.20
0.25
0.50
0.75
1.00
Mutations
additive
coadapted
compensatory

1


Two populations (Far West and Circum)
Far West population
A
B
C
D
$Tcor_{PC1}$
$Tcor_{PC1v}$
MtSUNN
MtCLE02
-log10(p-value)
chromosomes


A
WT Ubi:GUS
WT Ubi:CLE02
sunn Ubi:GUS
sunn Ubi:CLE02
B
2e-08
0.66
2e-06
200
150
100
50
0
Number of nodules
n = 42
n = 43
n = 43
n = 38
WT
Ubi:GUS
WT
Ubi:CLE02
sunn
Ubi:GUS
sunn
Ubi:CLE02


A
Tr
108973688_EDAR
count
Absolute r values
EDAR
-log10(p-value)
chromosomes
B
Tr
46172199_SLC24A5
SLC24A5
C
Tr_v
D
Tr_v

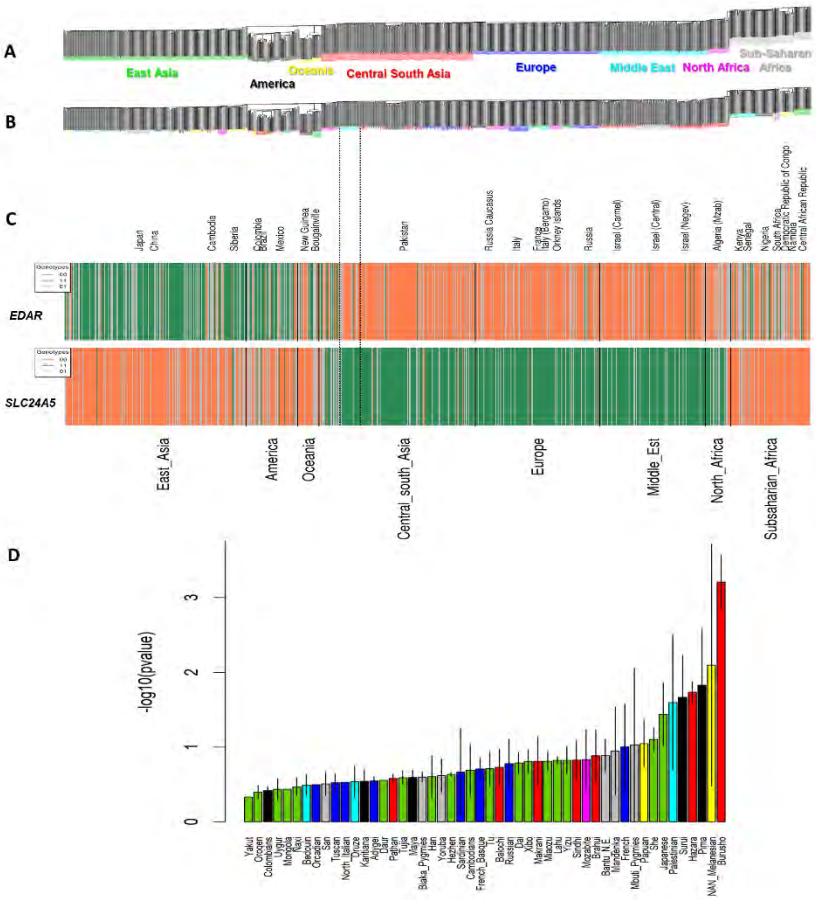

A
East Asia
America
Oceania
Central South Asia
Europe
Middle East
North Africa
Sub-Saharan Africa
B
C
Japan
China
Cambodia
Siberia
Papuan
Melano
New Guinea
Bougainville
Pakistan
Russia Caucasus
Italy
Tuscan
Italy (Bergamo)
Orkney islands
Russia
Israel (Carmel)
Israel (Central)
Israel (Negev)
Algeria (Mzab)
Kenya
Senegal
Nigeria
South Africa
Democratic Republic of Congo
Central African Republic
EDAR
SLC24A5
East_Asia
America
Oceania
Central_south_Asia
Europe
Middle_Est
North_Africa
Subsaharian_Africa
D
-log10(pvalue)
0
1
2
3

Table 1 – Two-locus epistatic selection models under coadaptation or compensation in a haploid population.

Allele at locus A	Allele at locus B	Allelic combination (haplotype)	Fitness value	
			Coadaptation	**Compensation**
A	*B*	*AB*	1	1
a	*B*	*aB*	1	$1-s$
A	*b*	*Ab*	1	$1-s$
a	*b*	*ab*	$1+s$	1

Specific fitness values are assigned to individuals depending on the allelic combination they carry at the two loci (e.g. SNP) in each epistatic selection model. Alleles *A* and *B* correspond to ancestral alleles, *a* and *b* correspond to mutated (or derived) alleles.

Table 2 – False positive (FP) proportions for T_r, T_{corPC1}, T_{r_v} and T_{corPC1_v} statistics in comparisons with the Student distribution ($\tau_{(n-2)}$) used for testing the significance of the correlation coefficient.

Generation	Mating scheme	Statistics	FP proportions		
			10%	**5%**	**1%**
140	self-mating	T_r	85 %	82 %	74 %
		T_{corPC1}	89 %	86 %	81 %
		$\boldsymbol{T_{r_v}}$	**13 %**	**8 %**	**3 %**
		$\boldsymbol{T_{corPC1_v}}$	**13 %**	**7 %**	**3 %**
	random-mating	T_r	72 %	66 %	55 %
		T_{corPC1}	83 %	78 %	70 %
		$\boldsymbol{T_{r_v}}$	**2.8 %**	**0.6 %**	**0.2 %**
		$\boldsymbol{T_{corPC1_v}}$	**2.5 %**	**0.6 %**	**0.1 %**
300	self-mating	T_r	92 %	91 %	87 %
		T_{corPC1}	95 %	93 %	91 %
		$\boldsymbol{T_{r_v}}$	**31 %**	**26 %**	**20 %**
		$\boldsymbol{T_{corPC1_v}}$	**37 %**	**31 %**	**22 %**
	random-mating	T_r	85 %	81 %	74 %
		T_{corPC1}	93 %	91 %	87 %
		$\boldsymbol{T_{r_v}}$	**4.6 %**	**2.3 %**	**1.2 %**
		$\boldsymbol{T_{corPC1_v}}$	**5.2 %**	**3.1 %**	**1.1 %**

False positive proportions are calculated as the proportion of simulations in which the statistics has a value greater than the defined rejection quantile of the $\tau_{(n-2)}$ distribution, for different type I errors: 10%, 5% and 1%. In our simulations, the sample size *n* was equal to 500. Since the sign of the correlation coefficient is not interpretable, especially for T_{corPC1} and T_{corPC1_v} the absolute values of T_r, T_{corPC1}, T_{r_v} and T_{corPC1_v} and of the Student distribution $\tau_{(n-2)}$ were used for false positive proportion calculation.

WILEY

Author Query Form

Journal: MPP

Article: 12964

Dear Author,

During the copyediting of your manuscript, the following queries arose.

Please refer to the query reference callout numbers in the page proofs and respond to each by marking the necessary comments using the PDF annotation tools.

Please remember illegible or unclear comments and corrections may delay publication.

Many thanks for your assistance.

AUTHOR: Please note that missing content in references have been updated where we have been able to match the missing elements without ambiguity against a standard citation database, to meet the reference style requirements of the journal. It is your responsibility to check and ensure that all listed references are complete and accurate.

Query reference	Query	Remarks
1	AUTHOR: Please confirm that given names (blue) and surnames/family names (vermilion) have been identified correctly.	
2	AUTHOR: Please verify that the linked ORCID identifiers are correct for each author.	
3	AUTHOR: Please check whether all affiliations have been set correctly.	
4	AUTHOR: 'Birker et al. (2009)' has not been included in the Reference List, please supply full publication details.	
5	AUTHOR: 'Narusaka (2009)' has not been included in the Reference List, please supply full publication details.	
6	AUTHOR: 'Williams et al. (2014)' has not been included in the Reference List, please supply full publication details.	
7	AUTHOR: Please give LRR in full.	
8	AUTHOR: Does TOU-A need to be given in full ?	
9	AUTHOR: Please give SNP in full.	
10	AUTHOR: Please check sense here 'a playful dynamics'.	
11	AUTHOR: Please check sense here 'detected over the kinetic of infection'.	
12	AUTHOR: Please check sense here 'top SNPs were falling in or in the vicinity'.	

13	AUTHOR: Johannes et al., 2007 has been changed to Johannes 2007 so that this citation matches the Reference List. Please confirm that this is correct.	
14	AUTHOR: Atkinson et al., 2012 has been changed to Atkinson & Urwin 2012 so that this citation matches the Reference List. Please confirm that this is correct.	
15	AUTHOR: Please check editing OK here 'Although computationally intensive, a complementary step will be to test the significance of all pairwise interactions among the 981,617 SNPs used in this study, which will require controlling the individual and joint effect of population structure on both SNPs tested in interaction'.	
16	AUTHOR: Poueymiro et al., 2009 has been changed to Poueymiro et al., 2014 so that this citation matches the Reference List. Please confirm that this is correct.	
17	AUTHOR: Please give BG in full.	
18	AUTHOR: Please provide manufacturer information for this manufacturer: town, state (if applicable), and country.	
19	AUTHOR: References 'Bartoli and Roux, 2017, Dodds and Rathjen, 2010, Jones and Dangl, 2006, Le Roux et al., 2015, Nürnberger, 2005, Osbourn, 1996.' have not been cited in the text. Please indicate where it should be cited; or delete from the Reference List and renumber the References in the text and Reference List.	

Funding Info Query Form

Please confirm that the funding sponsor list below was correctly extracted from your article: that it includes all funders and that the text has been matched to the correct FundRef Registry organization names. If a name was not found in the FundRef registry, it may not be the canonical name form, it may be a program name rather than an organization name, or it may be an organization not yet included in FundRef Registry. If you know of another name form or a parent organization name for a "not found" item on this list below, please share that information.

FundRef name	FundRef Organization Name
Occitanie Regional Council	
INRA Plant Health and Environment division	
Agence Nationale de la Recherche	Agence Nationale de la Recherche
Syngenta seeds	

Received: 21 November 2019 | Revised: 18 May 2020 | Accepted: 19 May 2020

DOI: 10.1111/mpp.12964

ORIGINAL ARTICLE

Molecular Plant Pathology WILEY

A complex network of additive and epistatic quantitative trait loci underlies natural variation of *Arabidopsis thaliana* quantitative disease resistance to *Ralstonia solanacearum* under heat stress

Nathalie Aoun[1] | Henri Desaint[1,2] | Léa Boyrie[3] | Maxime Bonhomme[3] | Laurent Deslandes[1] | Richard Berthomé[1] | Fabrice Roux[1]

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

[2]SYNGENTA seeds, Sarrians

[3]LRSV, Université de Toulouse, CNRS, Université Paul Sabatier, Castanet-Tolosan, France

Correspondence
Fabrice Roux, LIPM, Université de Toulouse, INRAE, CNRS, Castanet-Tolosan, France.
Email: fabrice.roux@inrae.fr

Funding information
Occitanie Regional Council; INRA Plant Health and Environment division; Agence Nationale de la Recherche, Grant/Award Number: DeCoD; Syngenta seeds

Abstract

Plant immunity is often negatively impacted by heat stress. However, the underlying molecular mechanisms remain poorly characterized. Based on a genome-wide association mapping approach, this study aims to identify in *Arabidopsis thaliana* the genetic bases of robust resistance mechanisms to the devastating pathogen *Ralstonia solanacearum* under heat stress. A local mapping population was phenotyped against the *R. solanacearum* GMI1000 strain at 27 and 30 °C. To obtain a precise description of the genetic architecture underlying natural variation of quantitative disease resistance (QDR), we applied a genome-wide local score analysis. Alongside an extensive genetic variation found in this local population at both temperatures, we observed a playful dynamics of quantitative trait loci along the infection stages. In addition, a complex genetic network of interacting loci could be detected at 30 °C. As a first step to investigate the underlying molecular mechanisms, the atypical meiotic cyclin *SOLO DANCERS* gene was validated by a reverse genetic approach as involved in QDR to *R. solanacearum* at 30 °C. In the context of climate change, the complex genetic architecture underlying QDR under heat stress in a local mapping population revealed candidate genes with diverse molecular functions.

KEYWORDS

epistasis, GWA mapping, heat stress, local score, natural accessions, *Ralstonia solanacearum*, SOLO DANCER

1 | INTRODUCTION

Climate scenarios predict that extreme climate events will become more frequent by the end of the century (IPCC, 2018), alongside an expected increase in global surface temperature from 1.5 to 4.8 °C (IPCC, 2018). In such a context of climate warming, global food security is at risk, with crop yields threatened by both the direct effect of increased temperature on plant development (Hatfield *et al.*, 2011; Saidi *et al.*, 2011; Bita and Geratz, 2013; Gray and Brady, 2016) and the indirect effect of increased temperature on the emergence of new pathogens and

the number and severity of epidemics (Garett *et al.*, 2006; Evans *et al.*, 2008; Bebber *et al.*, 2013; Bita and Gerats, 2013; Suzuki *et al.*, 2014). Unravelling the genetic and molecular mechanisms allowing plants to face pathogen attacks under elevated temperature therefore represents a promising strategy for sustainable disease resistance.

Being sessile, plants have developed a wide range of immune responses to face simultaneous and/or sequential stresses caused by various bioaggressors (Roux and Bergelson, 2016). Plant immunity relies on a surveillance system involving plasma membrane-anchored pattern recognition receptors (PRRs) that perceive microbial elicitors, called pathogen- or microbe-associated molecular patterns (PAMPs or MAMPs). PRR-triggered immunity (PTI) is efficient against a broad spectrum of pathogens (Cook *et al.*, 2015). Adapted pathogens such as phytopathogenic bacteria trigger susceptibility thanks to secreted virulence factors called effectors that can inhibit PTI and promote pathogen invasion (effector-triggered susceptibility, ETS). The specific recognition of pathogen effectors by plant intracellular nod-like receptors (NLRs) triggers a more robust immune response called effector-triggered immunity (ETI), often associated with a cell death response or hypersensitive response (HR) that restricts pathogen invasion to the infection site. In general, ETI is specific to a single pathogenic species, and even to a single pathogenic strain. This specificity causes a strong selective pressure on virulent strains to bypass ETI, making in most cases this form of immunity not durable in crop field conditions (Roux *et al.*, 2014). Another form of resistance represented by a reduction rather than an absence of disease refers to quantitative disease resistance (QDR) (St Clair, 2010; Mundt, 2014; Roux *et al.*, 2014; French *et al.*, 2016). QDR is generally polygenic, durable and broad spectrum (Young, 1996; Poland *et al.*, 2009). Unlike PTI and ETI, molecular mechanisms underlying QDR remain largely unknown (Roux *et al.*, 2014). Noteworthy is the alteration of all these major forms of immunity by heat stress. Numerous studies involving various pathosystems reported inhibition of ETI responses by a temperature increase (3–7 °C) (de Jong *et al.*, 2002; Xiao *et al.*, 2003; Yang and Hua, 2004; Wang *et al.*, 2009; Cheng *et al.*, 2013; Menna *et al.*, 2015; Aoun *et al.*, 2017; Venkatesh and Kang, 2019).

Bacterial wilt, caused by the gram-negative bacteria *Ralstonia solanacearum*, is one of the most devastating bacterial diseases in the world. Indeed, this soil-borne pathogen affects more than 200 species, including members of Solanaceae and Brassicaceae, and is responsible for dramatic yield losses not only in tropical and subtropical areas, but also in warm temperate regions (Elphinstone, 2005). In the model plant *Arabidopsis thaliana*, a broad-spectrum resistance response to *R. solanacearum* is conferred by the RPS4/RRS1-R locus that encodes a pair of NLR receptors cooperating molecularly to form homodimers (Deslandes *et al.*, 2002; Birker *et al.*, 2009; Narusaka 2009; Williams *et al.*, 2014). In addition, the LRR receptor-like kinase ERECTA was identified as underlying one of the three quantitative trait loci (QTLs) detected against the *R. solanacearum* strain 14.25 (Godiard *et al.*, 2003).

The genetic architecture and the molecular mechanisms of plant responses to *R. solanacearum* in changing abiotic environments, and more particularly under elevated temperature conditions, remain elusive. Recently, a genome-wide association study (GWAS) performed in *A. thaliana* and aimed at exploring the genetic bases associated with the natural variation of plant response to strain GMI1000 at 30 °C led to the identification of the *Strictosidine Synthase-Like protein 4* (*SSL4*) gene, although the underlying molecular mechanisms are still unknown (Aoun *et al.*, 2017). This study was based on 176 accessions of *A. thaliana* from a worldwide collection. While being informative, a limitation of this mapping population-based approach resides in an increased effect of the demographic history on genotype–phenotype association at large geographical scales. Statistical methods controlling for confounding by population structure can reduce the rate of false-positive associations, but to the detriment of a loss of detection power (i.e. markers linked to causative genes that are lost after correcting for population structure; Bergelson and Roux 2010, Brachi *et al.*, 2010). In addition, because different QTLs and/or different alleles at the same QTL can be responsible for the same phenotypic values, the power of GWAS can be strongly reduced by the effects of genetic and allelic heterogeneity due to the increased probability of the presence of rare alleles at large geographical scales (Bergelson and Roux, 2010). To limit these drawbacks, GWA mapping can be combined with traditional linkage mapping (based on the use of experimental populations such as recombinant inbred lines, RILs), which is prone to identifying rare alleles and not subjected to the effect of population structure (Bergelson and Roux, 2010). Combining GWA mapping and traditional linkage mapping has been demonstrated to reduce the rates of false positives and negatives when applied to flowering time data in *A. thaliana* (Brachi *et al.*, 2010), but remains time-consuming due to the need to phenotype thousands of experimental lines. To limit the drawbacks of GWA mapping performed at a worldwide scale, an alternative approach is to work at a small geographical scale (Bergelson and Roux, 2010). As reported in a GWAS performed on flowering in *A. thaliana* from a worldwide to a local scale (by using two highly polymorphic French mapping populations), a great reduction of confounding by population structure was observed at the smaller geographical scales (Brachi *et al.*, 2013). In addition, the genetic architecture was highly specific on the considered geographical scale (Brachi *et al.*, 2013).

In the present study, we therefore investigated the genetic bases of QDR to *R. solanacearum* under elevated temperature by performing a GWAS at a small geographical scale using the TOU-A local mapping population. This local mapping population offers several advantages, including (a) the detection of more than 1.9 million SNPs, only 5.6 times less than observed in a panel of 1,135 accessions collected at the worldwide scale (Frachon *et al.*, 2017), (b) an extensive genetic variation for a large range of phenotypic traits, including QDR, to the bacterial vascular pathogen *Xanthomonas campestris* pv *campestris*, (c) a linkage disequilibrium (LD) decay below 3 kb allowing fine-mapping of genomic regions associated with phenotypic variation, (d) a strongly reduced confounding effect by population structure, and (e) an adaptation to local warming in fewer than eight generations (Brachi *et al.*, 2013; Huard-Chauveau *et al.*, 2013; Baron *et al.*, 2015; Debieu *et al.*, 2016; Frachon *et al.*, 2017).

Interestingly, this work revealed a genetic architecture of natural variation of QDR to *R. solanacearum* that totally differs from the one previously described at the worldwide scale (Aoun *et al.*, 2017). In particular, at 30 °C, we observed a playful dynamics of 12 QTLs along the disease symptom progression, with most QTLs displaying complex

epistatic relationships. Using a reverse genetic approach, we identified *SOLO DANCERS* (*SDS*) encoding for a cyclin-like protein as the gene underlying one of the two additive QTLs detected at 30 °C.

2 | RESULTS

2.1 | Impact of temperature on genetic variation for QDR to *R. solanacearum* among local *A. thaliana* accessions

In this study, we tested 192 whole-genome sequenced local accessions of *A. thaliana* in response to the *R. solanacearum* GMI1000 reference strain, under growth chamber conditions. No germination was observed for six accessions that were therefore discarded from the study (Table S1). The remaining 186 local accessions from the TOU-A population were challenged with GMI1000 at 27 and 30 °C by cutting the roots. The accessions were on average more susceptible at 30 °C than at 27 °C (Figure 1a,b). For each temperature treatment, we observed a large genetic variation at most infection stages, that is, 5, 6, and 7 days after inoculation (dai) at 27 °C and 4, 5, 6, and 7 dai at 30 °C (Table 1 and Figure 1), with broad-sense heritability estimates ranging from 0.34 to 0.41 at 27 °C and from 0.29 to 0.39 at 30 °C (Table 1). Based on genotypic values estimated for the 186 TOU-A accessions, cross-temperature genetic correlation was weak, albeit significant (5 dai: Spearman's rho = 0.23, p = .003, 6 dai; Spearman's rho = 0.16, p = .033, 7 dai; Spearman's rho = 0.20, p = .008; Figure 1c), suggesting a flexible genetic architecture of *A. thaliana* response to the GMI1000 strain between 27 °C and 30 °C.

2.2 | Playful dynamics of QTLs at 27 and 30 °C

To increase the probability to discover QTLs with additive effects conferring QDR to *R. solanacearum* along the infection stages, we combined a genome-wide association mapping approach with a local score analysis (with tuning parameter ξ = 2) (Bonhomme *et al.*, 2019). We detected over the kinetic of infection 215 and 738 significant unique SNPs (i.e. top SNPs) at 27 and 30 °C, respectively (Figure 2). In agreement with weak cross-temperature genetic correlation, no single top SNP was common to both temperatures, indicating a contrasted genetic architecture for natural variation of response to *R. solanacearum* GMI1000 between 27 and 30 °C. Next, we focused on the 14 most highly significant additive QTLs (i.e. top QTL with a Lindley process >10, Figures 2 and 3). Two top QTLs were detected at 27 °C while the remaining top QTLs were detected at 30 °C (Figure 2). Interestingly, all these top QTLs displayed playful dynamics, with two QTLs (i.e. QTL1 at 27 °C and QTL3 at 30 °C) and 12 QTLs showing a decrease and increase in significance with advanced infection stages, respectively (Figure 2).

Based on LD calculation among the 14 top QTLs, both QTLs detected at 27 °C only present additive (i.e. independent) effects (Figures 4, S1, and S2). By contrast, at 30 °C, nine out of the 12 top QTLs also displayed epistatic interactions (Figures 4 and S2) with the identification of two groups of epistatic QTLs (Figure 5). The first one regroups seven QTLs (QTL5 + QTL6 + QTL7A + QTL7B + QTL 9 + QTL10 + QTL11) with highly significant pairwise LD values (p < .001); in particular at the interchromosomal level (Figures 4, 5, and S2). Based on the representative SNPs of the seven QTLs, 47.1% of disease index variation was explained by the cumulative number of resistance alleles at these QTLs (Figure 5c). It should be noted that c.80% of the accessions have a susceptible allele at each of the seven QTLs (Figure 5c), precluding testing with sufficient power any pairwise interactions among these QTLs. The second group contains QTL12A + QTL12B (Figure 5), with a clear disequilibrium in the number of accessions among the four expected haplotypes (SS = 88.9%, RR = 9%, RS = 0.7%, SR = 1.4%; Figure S1). QTL8 showed weak epistatic relationships with QTL5 and QTL9 (Figure 5a), whereas no significant epistatic relationship was detected for the two remaining QTLs detected at 30 °C (i.e. QTL3 and QTL4) (Figures 5, S1, and S2).

COLOUR online, B&W in print

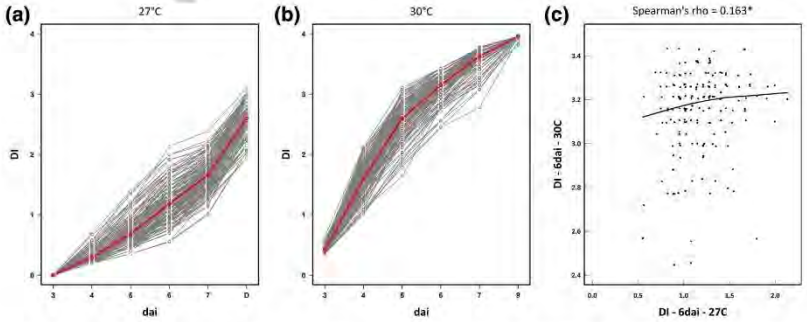


FIGURE 1 Genetic diversity of plant response to *R. solanacearum* GMI1000 strain in the local TOU-A mapping population. (a) Genetic variation of response dynamics at 27 °C. (b) Genetic variation of response dynamics at 30 °C. The red line represents the mean of disease index over all the accessions in (a) and (b). (c) Relationship between disease index at 6 dai scored at 27 and 30 °C. dai, days after inoculation. DI, disease index. The black line represents the locally weighted polynomial regression

TABLE 1 Natural variation among TOU-A natural accessions for disease index at 27 and 30 °C

Temperature (°C)	Model terms	Symptoms 3 dpi		Symptoms 4 dpi		Symptoms 5 dpi		Symptoms 6 dpi		Symptoms 7 dpi		Symptoms 9 dpi	
		F or LRT	*p*	*F* or LRT	*p*	*F* or LRT	*p*	*F* or LRT	*p*	*F* or LRT	*p*	*F* or LRT	*p*
27	Block	0.7	0.5262	1.7	0.2049	6.3	0.0036	36.4	0.0002	58.4	0.0002	50.9	0.0002
	Accession	0.0	1.0000	2.9	0.1076	9.0	0.0042	14.3	0.0004	7.9	0.0069	3.3	0.0906
	Control Col-0	ne	ne	54.8	0.0002	105.9	0.0002	122.8	0.0002	110.0	0.0002	87.1	0.0002
	H^2	0.00^{ns}		0.21^{ns}		0.35^{**}		0.41^{***}		0.34^{**}		0.37^{ns}	
30	Block	3.7	0.0382	20.9	0.0004	18.2	0.0004	9.0	0.0004	5.8	0.0064	4.2	0.0602
	Accession	1.4	0.2874	9.5	0.0051	18.4	0.0004	8.7	0.0064	12.4	0.0014	0.2	0.6956
	Control Col-0	0.1	0.7862	12.2	0.0014	6.8	0.0158	0.4	0.5967	2.4	0.1595	ne	ne
	H^2	0.13^{ns}		0.29^{**}		0.39^{***}		0.3^{**}		0.37^{**}		0.11^{ns}	

F, *F* value resulting from the test of fixed effect; LRT, LRT value resulting from the likelihood ratio test; H^2, broad-sense heritability values; ne, not estimated due to the absence of variation in disease symptoms among Col-0 control plants; ns, not significant. Italic terms indicate random effects.

$^{**}p < .01$; $^{***}p < .001$.

2.3 | Molecular functions of the candidate gene products underlying the QTLs identified

In agreement with the LD decay below 3 kb observed in the TOU-A population, the average size of the 14 QTLs was 4,589 bp (min = 141 bp, max = 26.07 kb, Figure 3), thereby limiting the number of candidate genes underlying each QTL. The AGI locus code and the corresponding predicted molecular function(s) of the candidate genes are indicated in Table 2. At 27 °C, the top SNP of QTL1 (SNP-1-22180112) is located in the coding region of *TREHALOSE PHOSPHATE SYNTHASE 10* (*At1g60140*) (Figure 3a). QTL2 covers a short region of c.1.77 kb, with the top SNP SNP-2-8253743 located in a gene (*At2g19050*) belonging to a GDSL-like Lipase/Acylhydrolase protein superfamily (Figure 3b).

At 30 °C, the candidate genes underlying the 12 top QTLs encode for various molecular functions. Functional classification was performed with the Classification Superviewer Tool on the university of Toronto website (http://bar.utoronto.ca/ntools/cgi-bin/ntools_classification_superviewer.cgi) using the MAPMAN classification as source (Provart and Zhu, 2003) and the list of genes in which top SNPs were falling in or in the vicinity. In particular, two QTLs correspond to genes involved in abiotic stress signalling pathways, that is, QTL3 with the top SNP SNP-4-6463310 and QTL7B with the top SNP SNP-1-26718947 falling within the *At4g10450* and *At1g70860* encoding for a ribosomal protein L6 family and a cytokinin responsive lipid transport protein, respectively (Figure 3c,h and Table 2). QTL4 covers a small region of 730 bp, with the top SNP SNP-1-5082790 falling within the promoter region of *SOLO DANCERS* (*At1g14750*) that encodes for an atypical meiotic cyclin-like protein (Figure 3d and Table 2). The top SNPs SNP-1-29319094 and SNP-4-10422518 of QTL8 and QTL11 were located in the genomic region of *At1g77990* (*SULPHATE TRANSPORTER 2;2*) and *At4g19030* (*NOD26-LIKE INTRINSIC PROTEIN 1,1*), respectively, encoding for a sulphate and an aquaporin transporter, respectively (Figure 3i,l and Table 2). QTL12A and QTL12B with the top SNP SNP-5-1569170 and SNP-5-1596241 are located within genes encoding for an EXPANSIN A2 (*At5g05290*) involved in cell wall modification and for a lignin peroxidase (*At5g05390*) involved in vascular development, respectively (Figure 3m,n and Table 2). However, among the different biological pathways represented by the 18 candidate genes identified, only the hormonal metabolism was significantly over-represented ($p < .01$). For instance, the top SNPs of QTL7A (SNP-1-26655520) and QTL10 (SNP-2-13134129) are located in genes *At1g70700* (*JAZ9*) and *At2g30830* (or alongside *At2g30810*), which are involved in jasmonate (JA), ethylene (ET), and gibberellin (GA) hormonal metabolisms, respectively (Figure 3g,h,k and Table 2).

2.4 | *SOLO DANCERS* is the gene underlying QTL4 involved in QDR to *R. solanacearum* GMI1000 strain at 30 °C

Next we investigated the molecular mechanisms underlying plant response to *R. solanacearum* at 30 °C in the TOU-A population. For this, we focused on the additive QTL with the highest

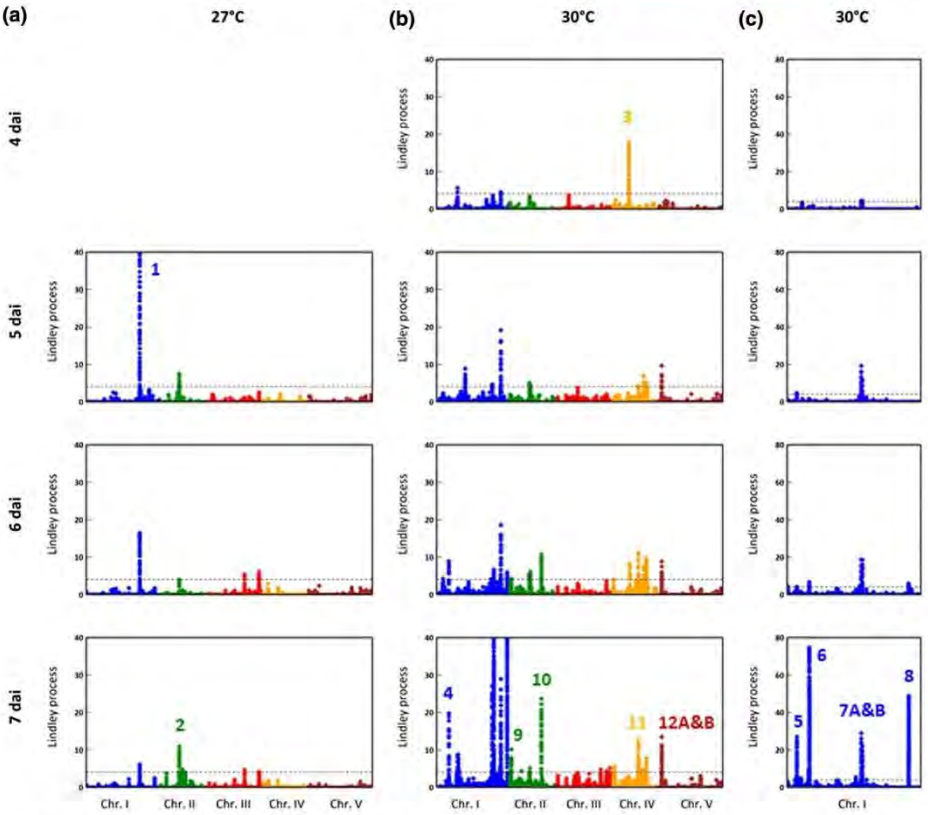


FIGURE 2 The genetics of QDR to *R. solanacearum* GMI1000 strain in the TOU-A population. (a) Manhattan plot of the Lindley process (local score method with a tuning parameter ξ = 2) at 5, 6, and 7 dai at 27 °C. (b) Manhattan plot of the Lindley process (ξ = 2) at 4, 5, 6, and 7 dai at 30 °C. (c) Zoom spanning a genomic region at the end of chromosome I from 23 Mb to 29.3 Mb containing five QTLs. The dashed line indicates the maximum of the five chromosome-wide significance thresholds. To better highlight minor QTLs in (a) and (b), Lindley process values on the y axis range from 0 to 40. Note that for the main association peak detected on chromosome 1 at 5 dai and 27 °C, the highest local score value of is 58.6. The number close to association peaks correspond to the 14 QTLs with a Lindley process value above 10. "7AandB" corresponds to two QTLs on chromosome 1 separated by c.63.5 kb. "12AandB" corresponds to two QTLs on chromosome 5 separated by c.27.1 kb

allelic effect, that is, QTL4 located on the top of chromosome I and that encompasses the *SOLO DANCERS* locus (*SDS*, *At1g14750*) (Figures 2b and S1). To check whether *SDS* was involved in this QDR, we monitored at 27 and 30 °C the phenotypical response of *sds-2* and *sds-3* null mutants (in Col-0 and Ws-4 genetic backgrounds, respectively). Col-0 and Ws-4 are both susceptible to GMI1000 at 30 °C while Ws-4, but not Col-0, is resistant at 27 °C. At 27 °C, the *sds-2* mutant response was not significantly different from Col-0 except at 6 dai (Figure 6a and Table S2). At 30 °C, the wilting of *sds-2* was significantly delayed compared to that of Col-0 from 3 to 5 dai (Figure 6b and Table S2). As expected, *sds-3* mutant and Ws-4 plants remained symptomless at 27 °C, from 3 to 7 dai (Figure 6c and Table S2). By contrast, at 30 °C, wilting of *sds-3* plants was strongly reduced during all infection stages (Figure 6d and Table S2). Altogether, these data suggest that *SDS* plays a role in wilting disease development on infection with the *R. solanacearum* GMI1000 strain at 30 °C.

3 | DISCUSSION

3.1 | Worldwide versus local genetic variation in *A. thaliana* facing *R. solanacearum*

In comparison with a temperature of 27 °C, local *Arabidopsis* accessions exposed at 30 °C were on average more susceptible to

COLOUR online, B&W in print

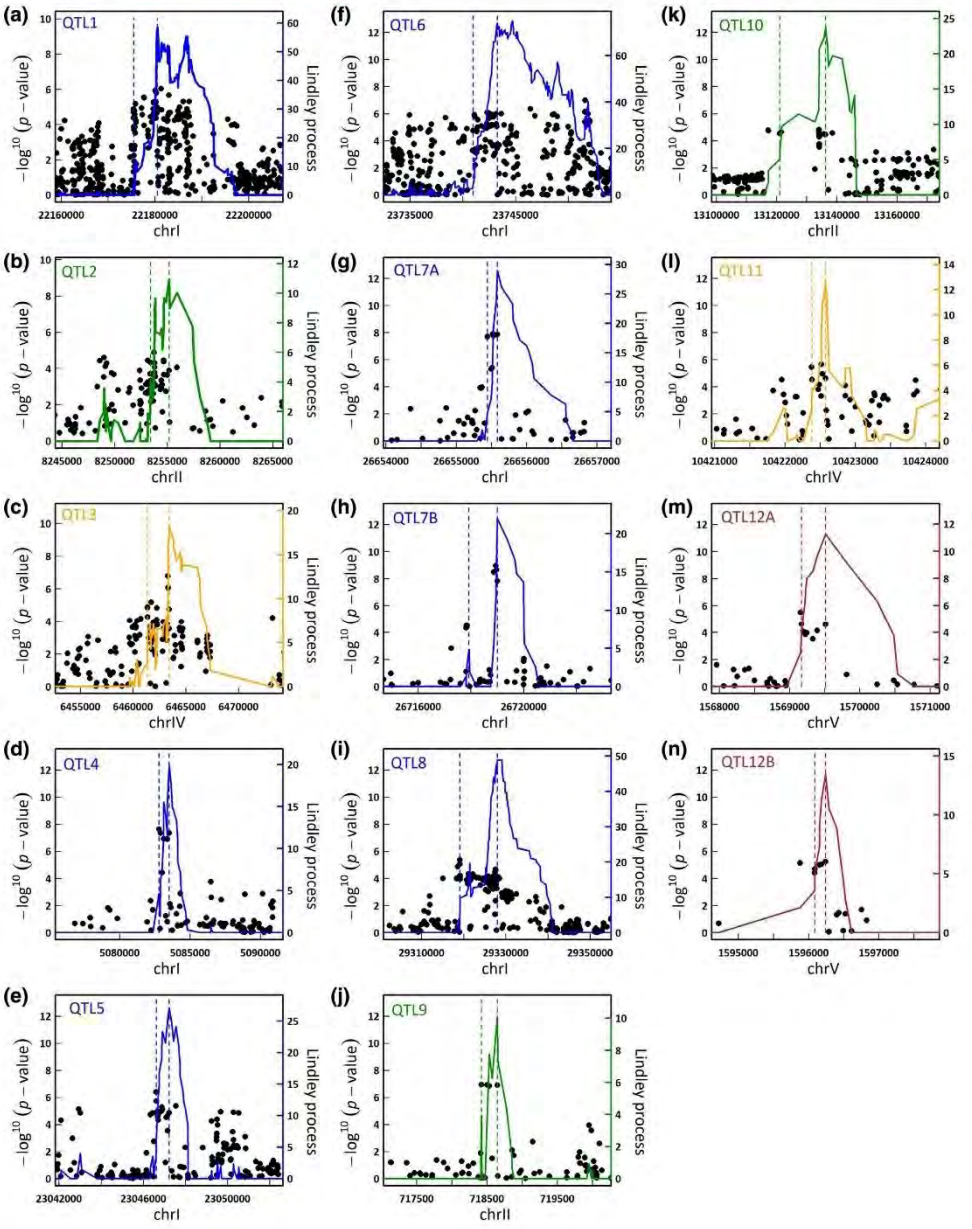


FIGURE 3 Zoom spanning the 14 QTLs with a Lindley process value above 10. Each of the 14 QTLs highlighted in Figure 2 are depicted from (a) to (n). The *x* axis corresponds to the physical position of the SNPs. The dots correspond to the $-\log_{10} p$ values of the SNPs obtained with the mixed model implemented in the EMMAX software (*y* axis on the left). The solid coloured curve indicates the Lindley process (local score method with $\xi = 2$) calculated from left to right (*y* axis on the right). The two coloured dashed vertical lines indicate the QTL intervals detected, without taking into account the right part of the curve (Fariello *et al.*, 2017; Bonhomme *et al.*, 2019)

COLOUR online, B&W in print

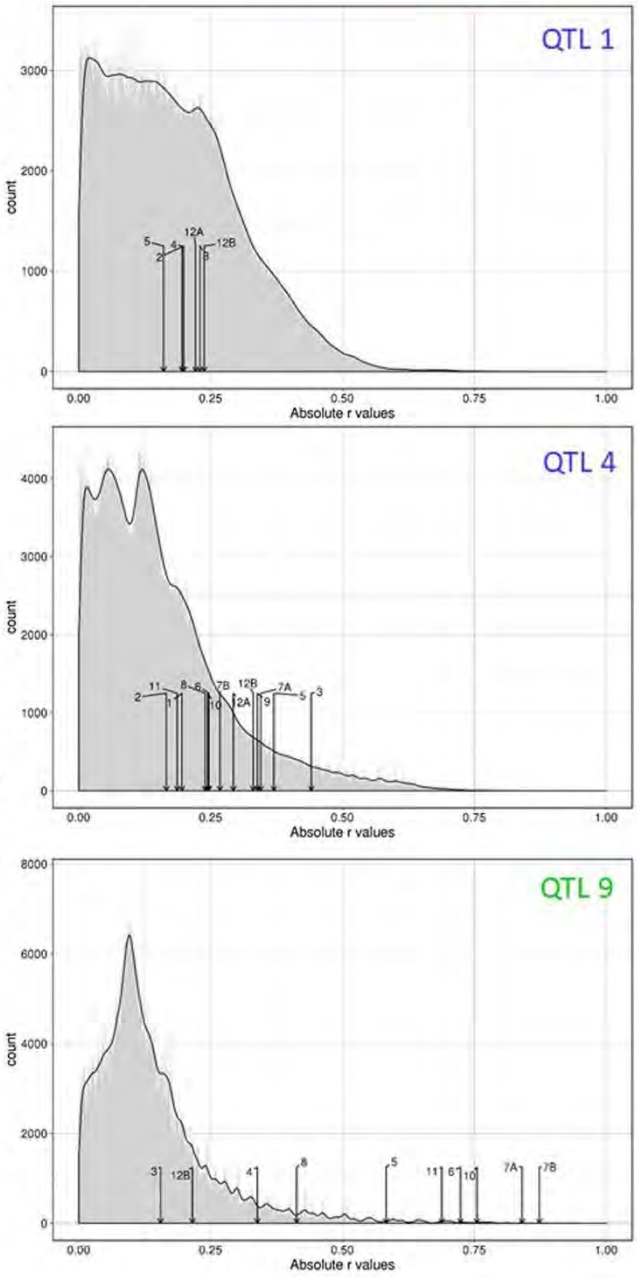


FIGURE 4 Detection of inter-QTLs epistasis for QTL1 (a), QTL4 (b), and QTL9 (c). For each QTL, a genome-wide distribution (grey area) was established by calculating LD values between the bait top SNP and all the other SNPs across the genome (with the exception of the SNPs located in a 100 kb window surrounding the bait top SNPs). Only SNPs with a MARF > 0.07 were considered. In addition, LD values (above 0.1) between the bait top SNP for the corresponding QTL and the bait top SNPs from the other QTLs are represented by arrows. The *x* axis corresponds to the LD estimates expressed as the absolute value of the *r* correlation coefficient. The black line corresponds to the density curve

the GMI1000 strain as indicated with a faster wilting disease progression than that observed at 27 °C. This observation is in line with the drastic impact of heat stress on *Arabidopsis* response to the GMI1000 strain previously monitored in a worldwide collection of *A. thaliana* (Aoun *et al.*, 2017). This is also in accordance with a growing number of studies performed on crops infected

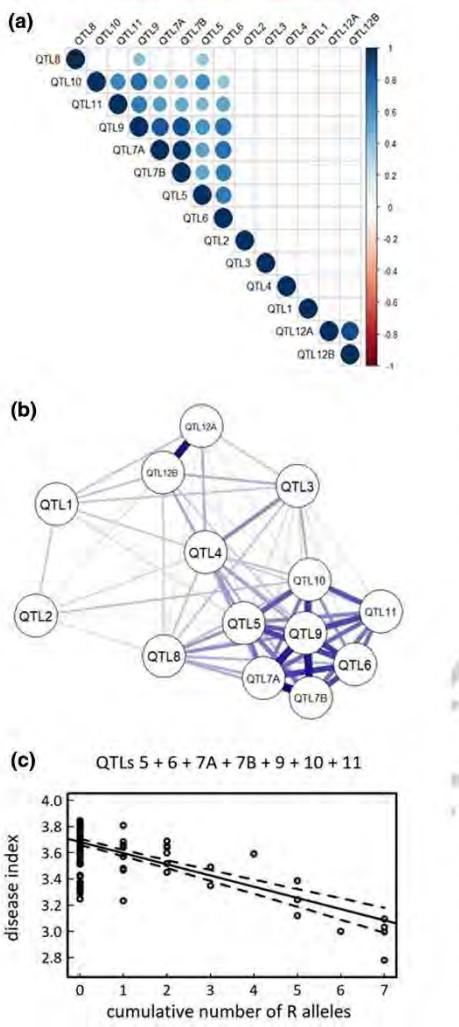


FIGURE 5 LD patterns among the 14 QTLs with a Lindley process above 10. (a) Graphical display of the LD matrix. (b) Weighted network visualization of the LD matrix (*qgraph* library implemented in the *R* environment). (c) Relationship between disease index and the cumulative number of resistant R alleles at the seven QTLs of the first group of epistatic QTLs. Dots correspond to the kinship adjusted genotypic values of the TOU-A accessions. A total amount of 47.1% of disease index variation was explained by the cumulative number of resistance alleles at the seven QTLs

by different pathogenic species that have described the drastic impact of heat stress on resistance response (Moury *et al.*, 1998; Jablonska *et al.*, 2007; Wang *et al.*, 2009; Webb *et al.*, 2010). The respective impact of heat stress on host and pathogen is still a matter of debate. Alteration of plant immunity under heat stress has been reported in several studies. For instance, a temperature of 28 °C inhibits "spontaneous lesion" phenotypes or autoimmune responses linked to autoactive alleles of genes encoding NLR resistance proteins (Zhu *et al.*, 2010; Negeri *et al.*, 2013). While heat stress increases *A. thaliana* susceptibility to infection with *Pseudomonas syringae* pv. *tomato* (Pst) DC3000, it also promotes the plant-dependent bacterial multiplication (Huot *et al.*, 2017). In this study, *R. solanacearum* grows faster at 30 °C than at 28 °C in in vitro conditions (Figure S3a) and a significant increase in bacterial multiplication was observed in planta at 30 °C (Figure S3b), suggesting an effect of heat stress on bacterial multiplication and its pathogenicity.

In previous studies, the level of genetic variation for diverse phenotypic traits such as flowering time and QDR to the bacterial pathogen *X. campestris* was similar between the TOU-A population and a set of worldwide accessions (Brachi *et al.*, 2013; Huard-Chauveau *et al.*, 2013; Debieu *et al.*, 2016). Here, the level of phenotypic and genetic variation for QDR to *R. solanacearum* in the TOU-A population was limited compared to that of the worldwide (WW) collection (mean standard deviation of phenotypic values between 5 and 7 dai: WW 27 °C = 1.61, TOU-A 27 °C = 1.38, WW 30 °C = 1.08, TOU-A 30 °C = 0.89; mean H^2 estimates between 5 and 7 dai: WW 27 °C = 0.77, TOU-A 27 °C = 0.37, WW 30 °C = 0.74, TOU-A 30 °C = 0.35; Aoun *et al.*, 2017). This may be explained by the absence in the TOU-A population of fully resistant accessions at 27 °C, which is consistent with the absence of any association peak located around the *RPS4/RRS1-R* locus on chromosome V (Figure 2). By contrast, this locus was detected as the major association peak at 27 °C in a set of worldwide accessions (Aoun *et al.*, 2017).

Given that most of the *A. thaliana* natural populations located in France are genetically diverse (Le Corre, 2005; Platt *et al.*, 2010; Brachi *et al.*, 2013; Frachon *et al.*, 2018; Frachon *et al.*, 2019), it would be interesting to investigate the level of genetic variation of QDR to *R. solanacearum* GMI1000 strain (or to other strains) within those populations. This would provide valuable information on the local dynamics of QDR to a bacterial pathogen in *A. thaliana* at the metapopulation level (Ding *et al.*, 2007; Vetter *et al.*, 2012; Karasov *et al.*, 2014; Roux and Bergelson, 2016).

3.2 | Complex genetic architecture of QDR to *R. solanacearum* at 27 and 30 °C

Combining GWA mapping related mixed models and genome-wide local score analysis increases the probability of discovering minor QTLs with additive effects (Fariello *et al.*, 2017; Bonhomme *et al.*, 2019). This is also well exemplified here with a more detailed characterization of the genetic determinants responsible for QDR to *R. solanacearum*. Since no top SNPs were common to both temperature treatments, our data illustrate how only a weak temperature increase of 3 °C can drastically affect the genetic architecture of QDR to *R. solanacearum*. We next evaluated the effect of the geographic

TABLE 2 List of candidate genes underlying the 14 QTLs with a Lindley process value above 10 at 27 and 30 °C

Temperature (°C)	dai	QTL id	Chromosome	Bait SNP	*p* value	Gene	Gene description
27	5	QTL1	1	22 180 112	1.20×10^{-6}	**AT1G60140**	**TREHALOSE PHOSPHATE SYNTHASE 10**
	7	QTL2	2	8 253 743	1.27×10^{-5}	**AT2G19050**	**GDSL-like Lipase/Acylhydrolase superfamily protein**
30	4	QTL3	4	6 463 310	1.63×10^{-7}	AT4G10440	S-adenosyl-L-methionine-dependent methyltransferases superfamily protein
						AT4G10450	**Ribosomal protein L6 family**
	7	QTL4	1	5 082 790	2.30×10^{-8}	**AT1G14750**	**SOLO DANCERS**
	7	QTL5	1	23 021 261	2.28×10^{-8}	AT1G62305	Core-2/I-branching beta-1,6-N-acetylglucosaminyltransferase family protein
						AT1G62330	O-fucosyltransferase family protein
	7	QTL6	1	23 742 319	8.08×10^{-7}	**AT1G63980**	**D111/G-patch domain-containing protein**
	7	QTL7A	1	26 655 520	1.31×10^{-7}	**AT1G70700**	**JASMONATE-ZIM-DOMAIN PROTEIN 9**
	7	QTL7B	1	26 718 947	1.07×10^{-9}	**AT1G70860**	**Polyketide cyclase/dehydrase and lipid transport superfamily protein**
	7	QTL8	1	29 319 094	4.28×10^{-6}	**AT1G77990**	**SULPHATE TRANSPORTER 2**
	7	QTL9	2	718 424	1.15×10^{-7}	**AT2G02620**	**Cysteine/Histidine-rich C1 domain family protein**
	7	QTL10	2	13 134 129	1.51×10^{-5}	AT2G30800	HELICASE IN VASCULAR TISSUE AND TAPETUM
						AT2G30810	Gibberellin-regulated family protein
						AT2G30820	aspartyl/glutamyl-tRNA(Asn/Gln) amidotransferase subunit
						AT2G30830	**2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein**
						AT2G30840	2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein
	7	QTL11	4	10 422 518	2.21×10^{-6}	**AT4G19030**	**NOD26-like intrinsic protein 1**
	7	QTL12A	5	1 569 170	2.41×10^{-5}	**AT5G05290**	**EXPANSIN A2**
	7	QTL12B	5	1 596 241	5.93×10^{-6}	**AT5G05390**	**LACCASE 12**

For each QTL, the candidate genes corresponding to the top SNP and the flanking gene are in bold and normal text, respectively.

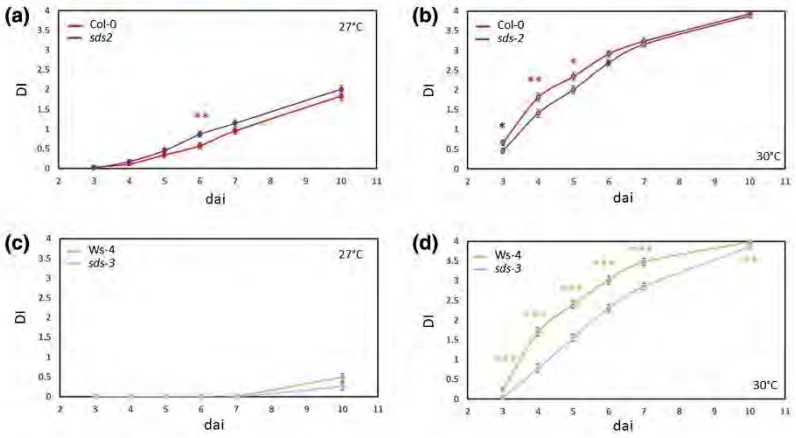


FIGURE 6 Effects of knockdown of *SDS* expression on the dynamics of disease symptoms after inoculation with the *R. solanacearum* GMI1000 strain in two genetic backgrounds at 27 and 30 °C. Dynamics of disease symptoms in Col-0 and *sds-2* mutant at 27 °C (a) and 30 °C (b). Dynamics of disease symptoms in Ws-4 and *sds-3* mutant at 27 °C (c) and 30 °C (d). Least-square means ± SE of the LS means from three independent inoculations (n = 72 plants per "genetic line * temperature" combination. Symbols *, **, and *** denote significant difference observed between each wild-type background and its corresponding mutant at $p < .05$, $p < .01$, and $p < .001$, respectively. Coloured stars indicate significant differences after a false-discovery rate correction. DI, disease index; dai, days after inoculation

scale on the genetic architecture of this QDR by applying a genome-wide local score approach to the EMMAX results previously obtained on a set of 176 worldwide *A. thaliana* accessions (Aoun *et al.*, 2017). No genomic regions containing top SNPs were shared between the local and worldwide scales (Data S1). Interestingly, similar results were obtained in a study investigating the genetic determinants of flowering time scored on both local and worldwide mapping populations of *A. thaliana* in two environmental conditions simulating two seasonal germination cohorts (Brachi *et al.*, 2013). The genetic architecture of flowering was highly dependent on both the geographical scale and the considered season (Brachi *et al.*, 2013). Together, the data obtained from various phenotypic traits reinforce the need to account for the geographical scale of phenotypic variation when choosing accession panels for GWAS (Bergelson and Roux, 2010). Consequently, this would help to get a better view of the genetic architecture flexibility of phenotypic traits.

Theoretical predictions suggest that phenotypic changes in ontogenetic time (typically time-to-event or time-to-failure traits such as flowering time or death time) are often driven by the temporal regulation of QTLs (Johannes, 2007). In this study, all the 14 top QTLs control QDR to *R. solanacearum* GMI1000 strain in a playful manner at both 27 and 30 °C, suggesting that disease progression to *R. solanacearum* highly depends on the time specificity of the genetic effects. At 30 °C, disease progression also resulted from a complex genetic network of interacting loci. The more complex genetic architecture observed at 30 °C suggests that under heat stress, *A. thaliana* responses to *R. solanacearum* involve a specific and complex network of mechanisms associated with the time specificity of the genetic effects. Few studies analysed the specificities of plant responses to individual or combined stresses, and mostly through transcriptome analyses. Interestingly, it turns out that transcriptional responses of plants to combined stresses are unique and cannot be predicted from that of individual stress (Atkinson and Urwin 2012; Suzuki *et al.*, 2014; Pandey *et al.*, 2015). In addition, combined stresses induce a major transcriptional reprogramming characterized by the regulation of the expression of a greater number of genes than observed with individual stresses (Rasmussen *et al.*, 2013; Suzuki *et al.*, 2014). For instance, more *Arabidopsis* genes were differentially expressed when nematode infection was combined with water stress compared to plants only subjected to nematode infection (Atkinson *et al.*, 2013). Similar observations were made in other pathosystems, including *A. thaliana* exposed to the simultaneous application of virus and heat or virus and drought (Prasch and Sonnewald, 2013; Pandey *et al.*, 2015). This may indicate that the more complex the environment is, the more the plants establish a response with a polygenetic architecture involving different genetic pathways.

Epistatic networks involving long-distance LD among physically unlinked loci were reported to represent the main fraction of phenotypic variance for herbivore resistance in *A. thaliana* at a worldwide scale (Brachi *et al.*, 2015), yeast growth (Forsberg *et al.*, 2017) or body weight and abdominal fat content in chicken (Carlborg *et al.*, 2006; Li *et al.*, 2013). While complex epistatic relationships among QTLs, including higher-order epistasis, may be therefore more frequent than anticipated (Carlborg and Haley, 2004; Roux *et al.*, 2005; Pettersson *et al.*, 2011), the functional validation of epistatic QTLs remains challenging but feasible if we consider a multi-CRISPR-Cas9 approach to create double, triple, quadruple, etc. mutants. Nonetheless, it would be interesting to determine whether such an epistatic network is

restricted to the TOU-A population by estimating LD between these five QTLs in other local highly polymorphic populations or at a larger geographical scale.

However, we should stress that some limitations in this study preclude a full description and understanding of the epistatic network underlying QDR to *R. solanacearum*. First, in contrast to traditional mapping population such as F2 populations or RILs, the number of accessions among the haplotypes expected between two (or more) epistatic QTLs was clearly unbalanced in the TOU-A population. While it may reflect the maintenance of haplotypes with extreme phenotypes (i.e. susceptible versus resistant) by selective processes (Brachi *et al.*, 2015), it precludes testing with sufficient power the magnitude and type of epistasis. Second, epistatic relationships were only tested on QTLs with additive effects that were first identified by combining a GWA mapping approach with a local score analysis. Although computationally intensive, a complementary step will be to test the significance of all pairwise interactions among the 981,617 SNPs used in this study, which will require controlling the individual and joint effect of population structure on both SNPs tested in interaction (Wang *et al.*, 2018).

3.3 | Various molecular functions are involved in QDR to the GMI1000 strain at 30 °C

Consistent with the molecular functions of previously cloned QDR genes (Roux *et al.*, 2014), the nature of most candidate genes underlying the 14 major QTLs identified here is quite diverse and they do not correspond to typical resistance genes encoding for NLRs. Indeed, unlike a previous GWAS performed on worldwide *A. thaliana* accessions that led to the detection of the *RPS4/RRS1-R* NLR locus as the main genetic determinant for full resistance to GMI1000 at 27 °C (Aoun *et al.*, 2017), the two main QTLs identified at 27 °C in the local TOU-A population do not correspond to any *NLR* genes. For QTL1, the top SNP fall in *TREHALOSE PHOSPHATE SYNTHASE 10* (*At1g60140*), suggesting that the regulation of trehalose-6-phosphate synthesis participates in the plant response. This is consistent with previous data showing that the production of this metabolite by the *R. solanacearum* effector RipTPS plays an important role in pathogen virulence (Poueymiro *et al.*, 2014). For QTL2, the top SNP fall in the *At2g19050* gene encoding a GDSL-like Lipase/Acylhydrolase superfamily protein. Interestingly, overexpression of *GLIP1* that also belongs to the *Arabidopsis GDSL LIPASE-LIKE* gene family was shown to confer enhanced resistance to several pathogens, including *Alternaria brassicicola*, *Erwinia carotovora* and *P. syringae* (Pst) (Kwon *et al.*, 2009). Therefore, these proteins might also play a role in plant immunity against *R. solanacearum*.

The molecular functions of the candidate genes underlying the 12 major QTLs detected at 30 °C are even more diverse. Interestingly, these functions may reflect different plant responses to face virulence strategies developed by the bacteria to colonize plant tissues and promote its multiplication within the xylem vessels. For instance, candidate genes underlying QTL7A and QTL10 are involved in the synthesis or signalling of hormones that may contribute positively to pathogen resistance and in plant response to combined biotic and abiotic stress. In particular, JA is known to interfere with GA signalling through the degradation of transcriptional repressors such as JAZ9 (the candidate gene underlying QTL7A) to balance plant defence response and growth (Yang *et al.*, 2012).

From 4 to 7 dai, QTL4 was detected with increasing significance on chromosome I. The corresponding candidate gene, *SDS*, encodes an atypical meiotic cyclin-like protein related to A- and B-type cyclins, previously described as being required for DNA double-strand break (DSB) repair (Azumi *et al.*, 2002; De Muyt *et al.*, 2009). To our knowledge, *SDS* has never been associated with plant disease susceptibility. Interestingly, two allelic null *sds* mutants were found to be more resistant at both 27 and 30 °C to GMI1000, albeit the allelic effect was different between the two genetic backgrounds. The functional validation of *SDS* as a susceptibility gene represents the first demonstration of its involvement in plant defence response to a bacterial pathogen under heat stress. It is noteworthy that (a) SDS acts together with CYCB3;1 in suppressing unscheduled cell wall synthesis (Bulankova *et al.*, 2013) and (b) the two candidate genes underlying QTL12A and QTL12B encode, respectively, proteins involved in cell wall and lignin polymerization. Two cyclin-L type proteins, MOS12 (Modifier of SNC1, 12) and MOS4-associated complex (Modifier of SNC1, 4), have also been shown to participate in the alternative splicing of *SNC1* and *RPS4* genes, thereby enabling the fine-tuning of *NLR* gene expression (Xu *et al.*, 2012). As several *NLR* genes have been described to be alternatively spliced without knowing the regulatory mechanism (Xu *et al.*, 2012), it is tempting to hypothesize that SDS would participate in the regulation of NLR functions under combined *R. solanacearum* and elevated temperature conditions through the production of splicing variants. Because the top SNPs are located in the promoter region of *SDS*, the next step to decipher the underlying molecular mechanisms would be to investigate the natural variation of *SDS* expression in the TOU-A population and its link to the QDR.

4 | EXPERIMENTAL PROCEDURES

4.1 | Bacterial strain, plant material, and growth conditions

The wild-type *R. solanacearum* GMI1000 strain was grown on complete BG medium as previously described (Plener *et al.*, 2010). GWAS was performed using 192 whole-genome sequenced natural accessions of the TOU-A population (France, Burgundy, 46°38′57.302″N, 04°07′16.892″E; Frachon *et al.*, 2017) (Table S1). Around five seeds of each accession were directly sown on Jiffy pots (Jiffy Products International AS) and left for 48 hr at 4 °C for stratification. Afterwards, plants were grown under controlled conditions for 4 weeks (22 °C, 70% relative humidity [RH], 9 hr of light) before inoculation. The two homozygous *sds-2* and *sds-3* mutants (SAIL and FAG105 T-DNA insertion mutants in Columbia-0 [Col-0]

and Wassilewskija [Ws-4] genetic backgrounds, respectively) were kindly provided by Raphaël Mercier (INRA, Versailles, France) (De Muyt *et al.*, 2009). An altered expression of *SDS* in these two mutants was confirmed in De Muyt *et al.* (2009). The two null mutants were grown as described above.

4.2 | Plant inoculation and phenotyping

Four-week-old plants were root-inoculated with the *R. solanacearum* GMI1000. The *R. solanacearum* GMI1000 reference strain was grown in complete BG medium, supplemented with 6 ml of glucose (20%) and 1 ml of triphenyltetrazolium chloride (1%), and incubated at 28 °C for 48 hr then left at room temperature for 24 hr. One day before inoculation, one colony was grown in liquid BG medium and grown overnight at 28 °C under shaking. Plants were inoculated with a bacterial suspension at $OD_{600\ nm}$ between 0.8 and 1. Before inoculation, roots were cut with scissors 1 cm from the bottom of the Jiffy pot (Deslandes *et al.*, 1998). This method gives the bacteria a direct access to the xylem vessels. During inoculation, plants were soaked in a bacterial suspension at 10^7 bacteria/ml for 15 min. Inoculated plants were incubated in growth chambers at 27 or 30 °C (75% RH, 12 hr light, 100 $\mu mol \cdot m^{-2} \cdot s^{-1}$). Disease symptoms were scored daily from 3 to 9 dai using a disease index scale from 0 to 4 as previously described (Deslandes *et al.*, 1998) with the scores from 0 to 4 corresponding to healthy and fully wilted plants, respectively.

4.3 | Natural variation of QDR in the TOU-A population

4.3.1 | Experimental design

For each temperature treatment, 624 plants were used and arranged by following a randomized complete block design (RCBD) of three temporal experimental blocks. Each block was represented by two trays of 104 positions, corresponding to one replicate per accession (n = 192 accessions) and the susceptible Col-0 accession was placed in the same three positions within each tray (n = 6). In each block, the remaining 10 positions in the trays were kept empty. Note that plants of the third block were not scored at 9 dai.

4.3.2 | Statistical analyses

For each temperature treatment, a mixed model (MIXED procedure in SAS v. 9.4; SAS Institute Inc., Cary, NC, USA) was used to explore the natural genetic variation of the disease index at each time point of phenotyping, as follows:

$$\text{disease index}_{ijc} = \mu + \text{block}_i + \text{accession}_j + \text{covCol}_c + \varepsilon_{ijc} \quad (1)$$

where μ is the overall mean of the phenotypic data, "block" accounts for differences in microenvironmental conditions between the three experimental blocks, "accession" corresponds to the genetic differences among the TOU-A natural accessions, covCol is a covariate accounting for tray effects within blocks (phenotypic mean of the three Col-0 replicates per tray was used as a covariate), and "ε" is the residual term. The factor "block" was considered as a fixed factor and the factor "accession" as a random factor. The significance of the random effect was determined by likelihood ratio tests of model with and without this effect. Residuals were normally distributed so no transformation was applied on raw phenotypic data. For GWA mapping analyses, we used best linear unbiased predictors (BLUPs) obtained for each natural accession. Because *A. thaliana* is a highly selfing species (Platt *et al.*, 2010), BLUPs correspond to genotypic values. Using a formula adapted from Gallais (1990), broad-sense heritabilities (H^2) at each time point of phenotyping were estimated from the variance component estimates of the "block" and "accession" terms obtained with the VARCOMP procedure in SAS v. 9.4 (SAS Institute Inc., Cary, NC, USA).

4.3.3 | GWA mapping with local score analysis

To fine map the genomic regions with additive effects associated with natural disease index variation at each time of phenotyping for each temperature treatment, a mixed model implemented in the software EMMAX was adopted (Efficient Mixed-Model Association eXpedited; Kang *et al.*, 2010). To control for the effect of population structure, we included as a covariate an identity-by-state kinship matrix K based on the 1,902,592 SNPs identified in the TOU-A population (Frachon *et al.*, 2017). Because rare alleles increase the rate of false positives when included in mixed models, we considered a threshold of minor allele relative frequency (MARF) >7% and ended up with 981,617 SNPs (Brachi *et al.*, 2010; Kang *et al.*, 2010). As previously described in Frachon *et al.* (2017), a threshold of 7% corresponds to the MARF value above which the p value distribution obtained from the mixed model is not dependent on MARF values in the TOU-A population.

Thereafter, we implemented a local score approach on the set of p values provided by EMMAX. The local score allows detection of significant genomic segments by accumulating the statistical signals from contiguous markers such as SNPs (Fariello *et al.*, 2017). In a given QTL region, the association signal, through the p values, will cumulate locally due to LD between SNPs, which will then increase the local score (Bonhomme *et al.*, 2019). Briefly, a sequence of scores is calculated along the chromosome as $X_i = -\log_{10}(p_i) - \xi$, where p_i is the p value of marker i and ξ is a tuning parameter with an optimal value that can be fixed at 2 or 3 in a GWAS context (Bonhomme *et al.*, 2019). Then, finding segments that accumulate strong signals is equivalent to finding peaks along a Lindley process defined as $h_i = \max(0, h_{i-1} + X_i)$ along the chromosome, with $h_0 = 0$. Significant SNP-phenotype associations were identified by estimating a chromosome-wide significance threshold for each chromosome (Bonhomme *et al.*, 2019).

4.3.4 | Detecting QTL epistasis

In order to detect epistatic interactions among our set of 14 top QTLs identified with additive effects by combining a GWA mapping approach with a local score analysis, we followed the procedure adopted in Brachi *et al.* (2015). We first identified within each QTL region the SNP with the highest association score estimated by EMMAX, hereafter named bait top SNP. For each of the 14 QTLs, we then computed LD estimates between the bait top SNP and all the other SNPs in the TOU-A population. LD between two biallelic (homozygous) SNPs was calculated using the absolute value of r statistic (correlation coefficient) between two SNP genotype vectors. For each QTL, we obtained a distribution of LD estimates between the bait top SNP and the other 981,616 SNPs of the population (MARF > 7%). In order to exclude strong LD values due to physical proximity, SNPs located in a 100 kb window surrounding a bait top SNP were not included in the calculation. To estimate whether the bait top SNP of a given QTL (i.e. focal bait top SNP) was significantly in LD with the bait top SNPs of the remaining 13 QTLs, we estimated in the LD distribution (conditional on each focal bait top SNP) the quantile q for each bait top SNP of the 13 QTLs. To be conservative, an LD estimate between a focal bait top SNP and another bait top SNP was declared significant if $(1 - q) < .01$.

4.3.5 | Estimates of allelic effect

To display the allelic effect of the bait SNPs after controlling for the effects of population structure, BLUPs estimated from model (1) were adjusted by fitting them with a kinship matrix. Kinship adjusted BLUPs were computed under the *R* environment 3.6.1 (R_Core_Team, 2019). In order to avoid pseudoreplication due to the presence of SNPs in stretches of LD, we first pruned the SNP data set with the snpgdspLD command using the following parameters 'ld.threshold = 0.8, slide.max.bp = 500, maf = 0.07 ("gdsfmt" and "SNPRelate" packages), leaving 365,952 SNPs for the estimation of the kinship matrix. The kinship matrix was then estimated using the popkin function (allowing missing data in the SNP matrix) in the popkin package, with the subpopulation vector set to NULL. Because the resulting matrix was not positive semi-definite, the function make.positive.definite() from the package lqmm was used. Finally, the kinship adjusted BLUPs were calculated with the function kin.blup from package rrBLUP. Keeping the notations from model (1), the parameters were: accession as geno, disease index as pheno, the above mentioned kinship matrix as K, GAUSS = F, indicating that the genotypes are not independent and follow $G = K V_G$, block as fixed effect and covCol as covariate. The kinship adjusted BLUPs were then extracted using the command $pred.

4.4 | Analysis of the *SDS* candidate gene

For each temperature treatment, an experiment of 288 plants was set up according to a RCBD of three temporal experimental blocks. Each block was represented by one tray of 96 positions, corresponding to 24 replicates of each genotype, that is, the *sds-2* and *sds-3* mutants with their corresponding wild-type background Col-0 and Ws-4, respectively.

For each temperature treatment, we tested whether each null mutant differs from its corresponding wild-type background along the infection stages by using the following model (MIXED procedure in SAS9.4; SAS Institute Inc.):

$$\text{disease index}_{ij} = \mu + \text{block}_i + \text{genotype}_j + \text{block}_i \times \text{genotype}_j + \varepsilon_{ij} \quad (2)$$

where μ is the overall mean of the phenotypic data, "block" accounts for differences in micro-environmental conditions between the three experimental blocks, "genotype" corresponds to the genetic differences between the T-DNA mutant and its corresponding wild-type background, "block × genotype" accounts for variation between genotype differences among blocks, and "ε" is the residual term. All factors were considered as fixed.

ACKNOWLEDGMENTS

We are grateful to Raphaël Mercier for providing T-DNA insertion mutant seeds (INRA, Versailles, France). N.A. benefited from a PhD grant co-financed by the Occitanie Regional Council and the INRA Plant Health and Environment Division (SPE). H.D. was funded by a grant from SYNGENTA seeds (Sarrians, France). L.B. benefited from a PhD grant financed by the ANR project DeCoD (ANR-16-CE20-0017-01). This study was performed at the LIPM belonging to the Laboratoire d'Excellence (LABEX) entitled TULIP (ANR-10-LABX-41).

AUTHOR CONTRIBUTIONS

R.B. and F.R. supervised the project. N.A., R.B., and F.R. designed the experiments. N.A. conducted the phenotyping experiments. N.A. and F.R. analyzed the phenotypic traits. F.R. performed the GWA mapping. M.B. performed the genome-wide local score analysis. L.B. performed the LD analyses. H.D. estimated the allelic effects of the QTLs. N.A., L.D., R.B., and F.R. wrote the manuscript. All authors contributed to the revisions.

DATA AVAILABILITY STATEMENT

Phenotypic data and GWA mapping results will be available in the Dryad database upon acceptance of the manuscript: https://doi.org/10.5061/dryad.XXX.

ORCID

Fabrice Roux https://orcid.org/0000-0001-8059-5638

REFERENCES

Aoun, N., Tauleigne, L., Lonjon, F., Deslandes, L., Vailleau, F., Roux, F. *et al.* (2017) Quantitative disease resistance under elevated temperature: genetic basis of new resistance mechanisms to *Ralstonia solanacearum*. *Frontiers in Plant Science*, 8, 1387.

Atkinson, N.J. and Urwin, P.E. (2012) The interaction of plant biotic and abiotic stresses: from genes to the field. *Journal of Experimental Botany*, 63, 3523–3543.

Atkinson, N.J., Lilley, C.J. and Urwin, P.E. (2013) Identification of genes involved in the response of Arabidopsis to simultaneous biotic and abiotic stresses. *Plant Physiology*, 162, 2028–2041.

Azumi, Y., Liu, D., Zhao, D., Li, W., Wang, G., Hu, Y. *et al.* (2002) Homolog interaction during meiotic prophase I in Arabidopsis requires the *SOLO DANCERS* gene encoding a novel cyclin-like protein. *The EMBO Journal*, 21, 3081–3095.

Baron, E., Richirt, J., Villoutreix, R., Amsellem, L. and Roux, F. (2015) The genetics of intra- and interspecific competitive response and effect in a local population of an annual plant species (A Bennett, Ed.). *Functional Ecology*, 29, 1361–1370.

Bartoli, C. and Roux, F. (2017) Genome-Wide Association studies in plant pathosystems: toward an ecological genomics approach. *Frontiers in Plant Science*, 8, 763.

Bita, C.E. and Gerats, T. (2013) Plant tolerance to high temperature in a changing environment: scientific fundamentals and production of heat stress-tolerant crops. *Frontiers in Plant Science*, 4, 273.

Brachi, B., Faure, N., Bergelson, J., Cuguen, J. and Roux, F. (2013) Genome-wide association mapping of flowering time in *Arabidopsis thaliana* in nature: genetics for underlying components and reaction norms across two successive years. *Acta Botanica Gallica*, 160, 205–219.

Brachi, B., Villoutreix, R., Faure, N., Hautekèete, N., Piquot, Y., Pauwels, M. *et al.* (2010) Investigation of the geographical scale of adaptive phenological variation and its underlying genetic bases in *Arabidopsis thaliana*. *Molecular Ecology*, 22, 4222–4240.

Brachi, B., Meyer, C.G., Villoutreix, R., Platt, A., Morton, T.C., Roux, F. *et al.* (2015) Coselected genes determine adaptive variation in herbivore resistance throughout the native range of *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences of the United States of America*, 112, 4032–4037.

Bebber, D.P., Ramotowski, M.A.T. and Gurr, S.J. (2013) Crop pests and pathogens move polewards in a warming world. *Nature Climate Change*, 3, 985–988.

Bergelson, J. and Roux, F. (2010) Towards identifying genes underlying ecologically relevant traits in *Arabidopsis thaliana*. *Nature Reviews Genetics*, 11, 867–879.

Bonhomme, M., Fariello, M.I., Navier, H., Hajri, A., Badis, Y., Miteul, H. *et al.* (2019) A local score approach improves GWAS resolution and detects minor QTL: application to *Medicago truncatula* quantitative disease resistance to multiple *Aphanomyces euteiches* isolates. *Heredity*, 123, 517–531.

Bulankova, P., Akimcheva, S., Fellner, N. and Riha, K. (2013) Identification of Arabidopsis meiotic cyclins reveals functional diversification among plant cyclin genes. *PLoS Genetics*, 9, e1003508.

Carlborg, O., Jacobsson, L., Ahgren, P., Siegel, P. and Andersson, L. (2006) Epistasis and the release of genetic variation during long-term selection. *Nature Genetics*, 38, 418–420.

Carlborg, H.O. and Haley, C.S. (2004) Epistasis: too often neglected in complex trait studies? *Nature Reviews Genetics*, 5, 618–625.

Cheng, C., Gao, X., Feng, B., Sheen, J., Shan, L. and He, P. (2013) Plant immune response to pathogens differs with changing temperatures. *Nature Communications*, 4, 2530.

Cook, D.E., Mesarich, C.H. and Thomma, B.P.H.J. (2015) Understanding plant immunity as a surveillance system to detect invasion. *Annual Review of Phytopathology*, 53, 541–563.

de Jong, C.F., Takken, F.L.W., Cai, X., de Wit, P.J.G.M. and Joosten, M.H.A.J. (2002) Attenuation of Cf-mediated defense responses at elevated temperatures correlates with a decrease in elicitor-binding sites. *Molecular Plant-Microbe Interactions*, 15, 1040–1049.

De Muyt, A., Pereira, L., Vezon, D., Chelysheva, L., Gendrot, G., Chambon, A. *et al.* (2009) A high throughput genetic screen identifies new early meiotic recombination functions in *Arabidopsis thaliana*. *PLoS Genetics*, 5, e1000654.

Debieu, M., Huard-Chauveau, C., Genissel, A., Roux, F. and Roby, D. (2016) Quantitative disease resistance to the bacterial pathogen *Xanthomonas campestris* involves an Arabidopsis immune receptor pair and a gene of unknown function. *Molecular Plant Pathology*, 17, 510–520.

Deslandes, L., Olivier, J., Theulieres, F., Hirsch, J., Feng, D.X., Bittner-Eddy, P. *et al.* (2002) Resistance to *Ralstonia solanacearum* in *Arabidopsis thaliana* is conferred by the recessive RRS1-R gene, a member of a novel family of resistance genes. *Proceedings of the National Academy of Sciences of the United States of America*, 99, 2404–2409.

Deslandes, L., Pileur, F., Liaubet, L., Camut, S., Can, C., Williams, K. *et al.* (1998) Genetic characterization of *RRS1*, a recessive locus in *Arabidopsis thaliana* that confers resistance to the bacterial soilborne pathogen *Ralstonia solanacearum*. *Molecular Plant-Microbe Interactions*, 11, 659–667.

Ding, J., Zhang, W., Jing, Z., Chen, J.-Q. and Tian, D. (2007) Unique pattern of R-gene variation within populations in Arabidopsis. *Molecular Genetics and Genomics*, 277, 619–629.

Dodds, P.N. and Rathjen, J.P. (2010) Plant immunity: towards an integrated view of plant-pathogen interactions. *Nature Reviews Genetics*, 11, 539–548.

Elphinstone, J.G. (2005) *The Current Bacterial Wilt Situation: A Global Overview*. St. Paul, MN: APS Press.

Evans, N., Baierl, A., Semenov, M.A., Gladders, P. and Fitt, B.D. (2008) Range and severity of a plant disease increased by global warming. *Journal of the Royal Society Interface*, 5, 525–531.

Fariello, M.I., Boitard, S., Mercier, S., Robelin, D., Faraut, T., Arnould, C. *et al.* (2017) Accounting for linkage disequilibrium in genome scans for selection without individual genotypes: the local score approach. *Molecular Ecology*, 26, 3700–3714.

Forsberg, S.K.G., Bloom, J.S., Sadhu, M.J., Kruglyak, L. and Carlborg, O. (2017) Accounting for genetic interactions improves modeling of individual quantitative trait phenotypes in yeast. *Nature Genetics*, 49, 497–503.

Frachon, L., Bartoli, C., Carrère, S., Bouchez, O., Chaubet, A., Gautier, M. *et al.* (2018) A genomic map of climate adaptation in *Arabidopsis thaliana* at a micro-geographic scale. *Frontiers in Plant Science*, 9, 967.

Frachon, L., Libourel, C., Villoutreix, R., Carrère, S., Glorieux, C., Huard-Chauveau, C. *et al.* (2017) Intermediate degrees of synergistic pleiotropy drive adaptive evolution in ecological time. *Nature Ecology and Evolution*, 1, 1551–1561.

Frachon, L., Mayjonade, B., Bartoli, C., Hautekèete, N.-C. and Roux, F. (2019) Adaptation to plant communities across the genome of *Arabidopsis thaliana*. *Molecular Biology and Evolution*, 36, 1442–1456.

French, E., Kim, B.S. and Iyer-Pascuzzi, A.S. (2016) Mechanisms of quantitative disease resistance in plants. *Seminars in Cell and Developmental Biology*, 56, 201–208.

Gallais, A. (1990) *Théorie de la Sélection en Amélioration des Plantes*. Paris: Masson.

Garrett, K.A., Dendy, S.P., Frank, E.E., Rouse, M.N. and Travers, S.E. (2006) Climate change effects on plant disease: genomes to ecosystems. *Annual Review of Phytopathology*, 44, 489–509.

Godiard, L., Sauviac, L., Torii, K.U., Grenon, O., Mangin, B., Grimsley, N.H. *et al.* (2003) ERECTA, an LRR receptor-like kinase protein controlling development pleiotropically affects resistance to bacterial wilt. *The Plant Journal*, 36, 353–365.

Gray, S.B. and Brady, S.M. (2016) Plant developmental responses to climate change. *Developmental Biology*, 419, 64–77.

Hatfield, J.L., Boote, K.J., Kimball, B.A., Ziska, L.H., Izaurralde, R.C., Ort, D. *et al.* (2011) Climate impacts on agriculture: implications for crop production. *Agronomy Journal*, 103, 351–370.

Huard-Chauveau, C., Perchepied, L., Debieu, M., Rivas, S., Kroj, T., Kars, I. *et al.* (2013) An atypical kinase under balancing selection confers broad-spectrum disease resistance in Arabidopsis. *PLoS Genetics*, 9, e1003766.

Huot, B., Castroverde, C.D.M., Velasquez, A.C., Hubbard, E., Pulman, J.A., Yao, J. *et al.* (2017) Dual impact of elevated temperature on plant defence and bacterial virulence in Arabidopsis. *Nature Communication*, 8, 1808.

IPCC. (2018) Summary for policy makers. In: *Global warming of 1.5 °C*. An IPCC Special Report on the impacts of global warming of 1.5 °C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty. Geneva, Switzerland: World Meteorological Organization, pp. 32.

Jablonska, B., Ammiraju, J.S.S., Bhattarai, K.K., Mantelin, S., de Ilarduya, O.M., Roberts, P.A. *et al.* (2007) The Mi-9 gene from *Solanum arcanum* conferring heat-stable resistance to root-knot nematodes is a homolog of Mi-1. *Plant Physiology*, 143, 1044–1054.

Johannes, F. (2007) Mapping temporally varying quantitative trait loci in time-to-failure experiments. *Genetics*, 175, 855–865.

Jones, J.D.G. and Dangl, J.L. (2006) The plant immune system. *Nature*, 444, 323–329.

Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.-Y., Freimer, N.B. *et al.* (2010) Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42, 348–354.

Karasov, T.L., Kniskern, J.M., Gao, L., DeYoung, B.J., Ding, J., Dubiella, U. *et al.* (2014) The long-term maintenance of a resistance polymorphism through diffuse interactions. *Nature*, 512, 436–440.

Kwon, S.J., Jin, H.C., Lee, S., Nam, M.H., Chung, J.H., Il, K.S. *et al.* (2009) GDSL lipase-like 1 regulates systemic resistance associated with ethylene signaling in Arabidopsis. *The Plant Journal*, 58, 235–245.

Le Corre, V. (2005) Variation at two flowering time genes within and among populations of *Arabidopsis thaliana*: comparison with markers and traits. *Molecular Ecology*, 14, 4181–4192.

Le Roux, C., Huet, G., Jauneau, A., Camborde, L., Trémousaygue, D., Kraut, A. *et al.* (2015) A receptor pair with an integrated decoy converts pathogen disabling of transcription factors to immunity. *Cell*, 161, 1074–1088.

Li, F., Hu, G., Zhang, H., Wang, S., Wang, Z. and Li, H. (2013) Epistatic effects on abdominal fat content in chickens: results from a genome-wide SNP-SNP interaction analysis. *PLoS One*, 8, e81520.

Menna, A., Nguyen, D., Guttman, D.S. and Desveaux, D. (2015) Elevated temperature differentially influences effector-triggered immunity outputs in Arabidopsis. *Frontiers in Plant Science*, 6, 995.

Moury, B., Selassie, K.G., Marchoux, G., Daubèze, A.-M. and Palloix, A. (1998) High temperature effects on hypersensitive resistance to Tomato Spotted Wilt Tospovirus (TSWV) in pepper (*Capsicum chinense* Jacq.). *European Journal of Plant Pathology*, 104, 489–498.

Mundt, C.C. (2014) Durable resistance: a key to sustainable management of pathogens and pests. *Infection, Genetics and Evolution*, 27, 446–455.

Negeri, A., Wang, G.-F., Benavente, L., Kibiti, C.M., Chaikam, V., Johal, G. *et al.* (2013) Characterization of temperature and light effects on the defense response phenotypes associated with the maize Rp1-D21 autoactive resistance gene. *BMC Plant Biology*, 13, 106.

Nürnberger, L.V. (2005) Non-host resistance in plants: new insights into an old phenomenon. *Molecular Plant Pathology*, 6, 335–345.

Osbourn, A.E. (1996) Preformed antimicrobial compounds and plant defense against fungal attack. *The Plant Cell*, 8, 1821–1831.

Pandey, P., Ramegowda, V. and Senthil-Kumar, M. (2015) Shared and unique responses of plants to multiple individual stresses and stress combinations: physiological and molecular mechanisms. *Frontiers Plant Science*, 16, 723.

Pettersson, M., Besnier, F., Siegel, P.B. and Carlborg, O. (2011) Replication and explorations of high-order epistasis using a large advanced intercross line pedigree. *PLoS Genetics*, 7, e1002180.

Platt, A., Horton, M., Huang, Y.S., Li, Y., Anastasio, A.E., Mulyati, N.W. *et al.* (2010) The scale of population structure in *Arabidopsis thaliana*. *PLoS Genetics*, 6, e1000843.

Plener, L., Manfredi, P., Valls, M. and Genin, S. (2010) PrhG, a transcriptional regulator responding to growth conditions, is involved in the control of the type III secretion system regulon in *Ralstonia solanacearum*. *Journal of Bacteriology*, 192, 1011–1019.

Poland, J.A., Balint-Kurti, P.J., Wisser, R.J., Pratt, R.C. and Nelson, R.J. (2009) Shades of gray: the world of quantitative disease resistance. *Trends in Plant Science*, 14, 21–29.

Poueymiro, M., Cazalé, A.C., François, J.M., Parrou, J.L., Peeters, N. and Genin, S. (2014) A Ralstonia solanacearum type III effector directs the production of the plant signal metabolite trehalose-6-phosphate. *mBio*, 5.

Prasch, C.M. and Sonnewald, U. (2013) Simultaneous application of heat, drought, and virus to Arabidopsis plants reveals significant shifts in signaling networks. *Plant Physiology*, 162, 1849–1866.

Provart, N. and Zhu, T. (2003) A browser-based functional classification superviewer for *Arabidopsis* genomics. *Current Protocols in Molecular Biology*, 271–272.

R_Core_Team (2019) *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rasmussen, S., Barah, P., Suarez-Rodriguez, M.C., Bressendorff, S., Friis, P., Costantino, P. *et al.* (2013) Transcriptome responses to combinations of stresses in Arabidopsis. *Plant Physiology*, 161, 1783–1794.

Roux, F. and Bergelson, J. (2016) The genetics underlying natural variation in the biotic interactions of *Arabidopsis thaliana*. *Current Topics in Developmental Biology*, 119, 111–156.

Roux, F., Camilleri, C., Giancola, S., Brunel, D. and Reboud, X. (2005) Epistatic interactions among herbicide resistances in *Arabidopsis thaliana*: the fitness cost of multiresistance. *Genetics*, 171, 1277–1288.

Roux, F., Voisin, D., Badet, T., Balagué, C., Barlet, X., Huard-Chauveau, C. *et al.* (2014) Resistance to phytopathogens e tutti quanti: placing plant quantitative disease resistance on the map. *Molecular Plant Pathology*, 15, 427–432.

Saidi, Y., Finka, A. and Goloubinoff, P. (2011) Heat perception and signalling in plants: a tortuous path to thermotolerance. *New Phytologist*, 190, 556–565.

St Clair, D.A. (2010) Quantitative disease resistance and quantitative resistance loci in breeding. *Annual Review of Phytopathology*, 48, 247–268.

Suzuki, N., Rivero, R.M., Shulaev, V., Blumwald, E. and Mittler, R. (2014) Abiotic and biotic stress combinations. *New Phytologist*, 203, 32–43.

Venkatesh, J. and Kang, B.C. (2019) Current views on temperature-modulated *R* gene-mediated plant defenses responses and tradeoffs between plant growth and immunity. *Current Opinion in Plant Biology*, 50, 9–17.

Vetter, M.M., Kronholm, I., He, F., Haweker, H., Reymond, M., Bergelson, J. *et al.* (2012) Flagellin perception varies quantitatively in *Arabidopsis thaliana* and its relatives. *Molecular Biology and Evolution*, 29, 1655–1667.

Webb, K.M., Oña, I., Bai, J., Garrett, K.A., Mew, T., Vera Cruz, C.M. *et al.* (2010) A benefit of high temperature: increased effectiveness of a rice bacterial blight disease resistance gene. *New Phytologist*, 185, 568–576.

Wang, Y., Bao, Z., Zhu, Y. and Hua, J. (2009) Analysis of temperature modulation of plant defense against biotrophic microbes. *Molecular Plant-Microbe Interactions*, 22, 498–506.

Wang, M., Roux, F., Bartoli, C., Huard-Chauveau, C., Meyer, C., Lee, H. *et al.* (2018) Two-way mixed-effects methods for joint association analysis using both host and pathogen genomes. *Proceedings of the*

National Academy of Sciences of the United States of America, 115, E5440–E5449.

Xiao, S., Brown, S., Patrick, E., Brearley, C. and Turner, J.G. (2003) Enhanced transcription of the Arabidopsis disease resistance genes RPW8.1 and RPW8.2 via a salicylic acid–dependent amplification circuit is required for hypersensitive cell death. *The Plant Cell*, 15, 33–45.

Xu, F., Xu, S., Wiermer, M., Zhang, Y. and Li, X. (2012) The cyclin L homolog MOS12 and the MOS4-associated complex are required for the proper splicing of plant resistance genes. *The Plant Journal*, 70, 916–928.

Yang, D.-L., Yao, J., Mei, C.-S., Tong, X.-H., Zeng, L.-J., Li, Q. *et al.* (2012) Plant hormone jasmonate prioritizes defense over growth by interfering with gibberellin signaling cascade. *Proceedings of the National Academy of Sciences of the United States of America*, 109, E1192–E1200.

Yang, S. and Hua, J. (2004) A haplotype-specific resistance gene regulated by BONZAI1 mediates temperature-dependent growth control in Arabidopsis. *The Plant Cell*, 16, 1060–1071.

Young, N.D. (1996) QTL mapping and quantitative disease resistance in plants. *Annual Review of Phytopathology*, 34, 479–501.

Zhu, Y., Qian, W., and Hua, J. (2010) Temperature modulates plant defense responses through NB-LRR proteins. *PLoS Pathogens*, 6, e1000844.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

FIGURE S1 Box-plots illustrating the additive effect of each top SNP located in five QTLs not involved in epistatic interactions as well as the epistatic effects between QTLs 12A and12B. Blue, green, orange, and red boxes indicate that the QTLs are located on chromosomes I, II, IV, and V, respectively. White boxes illustrate epistatic interactions among QTLs

FIGURE S2 Detection of inter-QTLs epistasis for 11 out of the 14 QTLs with a Lindley process value above 10. For each QTL, a genome-wide distribution (grey area) was established by calculating LD values between the bait top SNP and all the other SNPs across the genome (with the exception of the SNPs located in a 100-kb window surrounding the bait top SNP). Only SNPs with a MARF > 0.07 were considered. In addition, LD values (above 0.1) between the bait top SNP for the corresponding QTL and the bait top SNPs from the other QTLs are represented by arrows. The *x* axis corresponds to the LD estimates expressed as the absolute value of the *r* correlation coefficient. The black line corresponds to the density curve

FIGURE S3 Growth dynamics and internal growth curve of GMI1000 reference strain at 27 and 30 °C. (a) In vitro growth dynamics of the GMI1000 reference strain. Bacterial cultures were grown at 27 and 30 °C in complete liquid medium starting from a single colony. The coloured lines represent the mean of two independent biological repeats, each composed of 10 technical repeats. Standard deviation at each time point is represented by a vertical bar. (b) Box plot illustrating in planta bacterial growth at 27 and 30 °C using *A. thaliana Col-0 susceptible accession* plants. Each dot represents one of the 14 plants. In planta bacterial multiplication of the GMI1000 strain in Col-0 was significantly different between 27 and 30 °C ($F = 5.69$, $p = .0250$). *$p < .05$

TABLE S1 List of the 192 natural accessions of the local TOU-A mapping population used in this study. This list shows the ecotype ID and name of the accessions, as well as their germination status at 27 and 30 °C

TABLE S2 Effects of knockdown of *SDS* expression at 3, 4, 5, 6, 7, and 10 dai against the *Ralstonia solanacearum* GMI1000 strain in two wild-type genetic backgrounds at 27 and 30 °C. *F*, *F* value resulting from the test of fixed effect. *$p < .05$, **$p < .01$, ***$p < .001$; ne, not estimated due to the absence of phenotypic variation. Significant differences between each wild-type background and its corresponding mutant after an FDR correction are indicated in bold

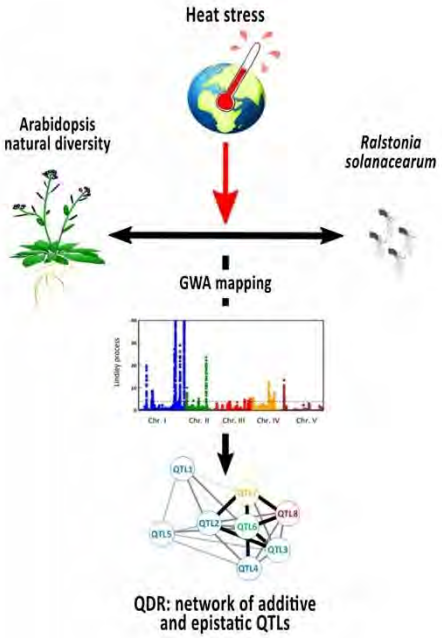
METHOD S1 Supplementary method

DATA S1 List of top SNPs identified at the local and worldwide scales at 27 and 30 °C

How to cite this article: Aoun N, Desaint H, Boyrie L, et al. A complex network of additive and epistatic quantitative trait loci underlies natural variation of *Arabidopsis thaliana* quantitative disease resistance to *Ralstonia solanacearum* under heat stress. *Molecular Plant Pathology*. 2020;00:1–16. https://doi.org/10.1111/mpp.12964

Graphical Abstract

The contents of this page will be used as part of the graphical abstract of html only. It will not be published as part of main article.



A genome-wide association study of the natural variation of *Arabidopsis thaliana* response to *Ralstonia solanacearum* under heat stress revealed resilient quantitative disease resistance underlain by a complex genetic network of additive and epistatic quantitative trait loci.

Bibliographie

Afzal, A. J., Wood, A. J., & Lightfoot, D. A. (2008). Plant receptor-like serine threonine kinases : Roles in signaling and plant defense. *Molecular Plant-Microbe Interactions: MPMI*, *21*(5), 507-517.

Arrighi, J.-F., Barre, A., Ben Amor, B., Bersoult, A., Soriano, L. C., Mirabella, R., de Carvalho-Niebel, F., Journet, E.-P., Ghérardi, M., Huguet, T., Geurts, R., Dénarié, J., Rougé, P., & Gough, C. (2006). The Medicago truncatula Lysine Motif-Receptor-Like Kinase Gene Family Includes NFP and New Nodule-Expressed Genes. *Plant Physiology*, *142*(1), 265-279.

Arrighi, J.-F., Godfroy, O., Billy, F. de, Saurat, O., Jauneau, A., & Gough, C. (2008). The RPG gene of Medicago truncatula controls Rhizobium-directed polar growth during infection. *Proceedings of the National Academy of Sciences*, *105*(28), 9817-9822.

Assis, R. (2014). Strong epistatic selection on the RNA secondary structure of HIV. *PLoS Pathogens*, *10*(9), e1004363.

Atwell, S., Huang, Y. S., Vilhjálmsson, B. J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A. M., Hu, T. T., Jiang, R., Muliyati, N. W., Zhang, X., Amer, M. A., Baxter, I., Brachi, B., Chory, J., Dean, C., Debieu, M., … Nordborg, M. (2010). Genome-wide association study of 107 phenotypes in a common set of Arabidopsis thaliana inbred lines. *Nature*, *465*(7298), 627-631.

Avery, L., & Wasserman, S. (1992). Ordering gene function : The interpretation of epistasis in regulatory hierarchies. *Trends in Genetics: TIG*, *8*(9), 312-316.

Baryshnikova, A., Costanzo, M., Myers, C. L., Andrews, B., & Boone, C. (2013). Genetic Interaction Networks : Toward an Understanding of Heritability. *Annual Review of Genomics and Human Genetics*, *14*(1), 111-133.

Basu Mallick, C., Iliescu, F. M., Möls, M., Hill, S., Tamang, R., Chaubey, G., Goto, R., Ho, S. Y. W., Gallego Romero, I., Crivellaro, F., Hudjashov, G., Rai, N., Metspalu, M., Mascie-Taylor, C. G. N., Pitchappan, R., Singh, L., Mirazon-Lahr, M., Thangaraj, K., Villems, R., & Kivisild, T. (2013). The Light Skin Allele of SLC24A5 in South Asians and Europeans Shares Identity by Descent. *PLoS Genetics*, *9*(11).

Bateson, W. (1909). *Mendel's principles of heredity, by W. Bateson*. Cambridge [Eng.] University Press. http://archive.org/details/mendelsprinciple00bate

Beaumont, M. A., & Nichols, R. A. (1996). Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London. Series B: Biological*

Sciences, *263*(1377), 1619-1626.

Begun, D. J., Holloway, A. K., Stevens, K., Hillier, L. W., Poh, Y.-P., Hahn, M. W., Nista, P. M., Jones, C. D., Kern, A. D., Dewey, C. N., Pachter, L., Myers, E., & Langley, C. H. (2007). Population Genomics : Whole-Genome Analysis of Polymorphism and Divergence in Drosophila simulans. *PLOS Biology*, *5*(11), e310.

Beissinger, T. M., Gholami, M., Erbe, M., Weigend, S., Weigend, A., de Leon, N., Gianola, D., & Simianer, H. (2016). Using the variability of linkage disequilibrium between subpopulations to infer sweeps and epistatic selection in a diverse panel of chickens. *Heredity*, *116*(2), 158-166.

Beleza, S., Santos, A. M., McEvoy, B., Alves, I., Martinho, C., Cameron, E., Shriver, M. D., Parra, E. J., & Rocha, J. (2013). The Timing of Pigmentation Lightening in Europeans. *Molecular Biology and Evolution*, *30*(1), 24-35.

Bergelson, J., & Roux, F. (2010). Towards identifying genes underlying ecologically relevant traits in Arabidopsis thaliana. *Nature Reviews. Genetics*, *11*(12), 867-879.

Bodmer, W. F. (1972). Evolutionary significance of the HL-A system. *Nature*, *237*(5351), 139-145 passim.

Boitard, S., Schlötterer, C., & Futschik, A. (2009). Detecting selective sweeps : A new approach based on hidden markov models. *Genetics*, *181*(4), 1567-1578.

Bonhomme, M., André, O., Badis, Y., Ronfort, J., Burgarella, C., Chantret, N., Prosperi, J.-M., Briskine, R., Mudge, J., Debéllé, F., Navier, H., Miteul, H., Hajri, A., Baranger, A., Tiffin, P., Dumas, B., Pilet-Nayel, M.-L., Young, N. D., & Jacquet, C. (2014). High-density genome-wide association mapping implicates an F-box encoding gene in Medicago truncatula resistance to Aphanomyces euteiches. *New Phytologist*, *201*(4), 1328-1342.

Bonhomme, M., Boitard, S., San Clemente, H., Dumas, B., Young, N., & Jacquet, C. (2015). Genomic Signature of Selective Sweeps Illuminates Adaptation of Medicago truncatula to Root-Associated Microorganisms. *Molecular Biology and Evolution*, *32*(8), 2097-2110.

Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J., Blott, S., & Sancristobal, M. (2010). Detecting selection in population trees : The Lewontin and Krakauer test extended. *Genetics*, *186*(1), 241-262.

Bonhomme, M., Fariello, M. I., Navier, H., Hajri, A., Badis, Y., Miteul, H., Samac, D. A., Dumas,

B., Baranger, A., Jacquet, C., & Pilet-Nayel, M.-L. (2019). A local score approach improves GWAS resolution and detects minor QTL : Application to Medicago truncatula quantitative disease resistance to multiple Aphanomyces euteiches isolates. *Heredity*, *123*(4), 517-531.

Bonhomme, M., & Jacquet, C. (2020). Genome-wide association mapping and population genomic features in Medicago truncatula. In F. de Bruijn (Éd.), *The Model Legume Medicago truncatula* (1re éd., p. 870-881). Wiley.

Botchkarev, V. A., & Fessing, M. Y. (2005). Edar signaling in the control of hair follicle development. *The Journal of Investigative Dermatology. Symposium Proceedings*, *10*(3), 247-251.

Boucher, B., & Jenna, S. (2013). Genetic interaction networks : Better understand to better predict. *Frontiers in Genetics*, *4*.

Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An expanded view of complex traits : From polygenic to omnigenic. *Cell*, *169*(7), 1177-1186.

Brachi, B., Meyer, C. G., Villoutreix, R., Platt, A., Morton, T. C., Roux, F., & Bergelson, J. (2015). Coselected genes determine adaptive variation in herbivore resistance throughout the native range of Arabidopsis thaliana. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(13), 4032-4037.

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL : Software for association mapping of complex traits in diverse samples. *Bioinformatics (Oxford, England)*, *23*(19), 2633-2635.

Branca, A., Paape, T. D., Zhou, P., Briskine, R., Farmer, A. D., Mudge, J., Bharti, A. K., Woodward, J. E., May, G. D., Gentzbittel, L., Ben, C., Denny, R., Sadowsky, M. J., Ronfort, J., Bataillon, T., Young, N. D., & Tiffin, P. (2011). Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume Medicago truncatula. *Proceedings of the National Academy of Sciences*, *108*(42), E864-E870.

Bryk, J., Hardouin, E., Pugach, I., Hughes, D., Strotmann, R., Stoneking, M., & Myles, S. (2008). Positive selection in East Asians for an EDAR allele that enhances NF-kappaB activation. *PloS One*, *3*(5), e2209.

Burgarella, C., Chantret, N., Gay, L., Prosperi, J.-M., Bonhomme, M., Tiffin, P., Young, N. D., & Ronfort, J. (2016). Adaptation to climate through flowering phenology : A case study in Medicago truncatula. *Molecular Ecology*, *25*(14), 3397-3415.

Burgarella, C., & Glémin, S. (2017). Population Genetics and Genome Evolution of Selfing Species. In John Wiley & Sons Ltd (Éd.), *ELS* (p. 1-8). John Wiley & Sons, Ltd.

Caicedo, A. L., Stinchcombe, J. R., Olsen, K. M., Schmitt, J., & Purugganan, M. D. (2004). Epistatic interaction between Arabidopsis FRI and FLC flowering time genes generates a latitudinal cline in a life history trait. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(44), 15670-15675.

Cann, H. M., de Toma, C., Cazes, L., Legrand, M.-F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W. F., Bonne-Tamir, B., Cambon-Thomsen, A., Chen, Z., Chu, J., Carcassi, C., Contu, L., Du, R., Excoffier, L., Ferrara, G. B., Friedlaender, J. S., Groot, H., … Cavalli-Sforza, L. L. (2002). A human genome diversity cell line panel. *Science (New York, N.Y.)*, *296*(5566), 261-262.

Chiu, C. H., & Paszkowski, U. (2020). Receptor-Like Kinases Sustain Symbiotic Scrutiny. *Plant Physiology*, *182*(4), 1597-1612.

Coop, G., Witonsky, D., Di Rienzo, A., & Pritchard, J. K. (2010). Using environmental correlations to identify loci underlying local adaptation. *Genetics*, *185*(4), 1411-1423.

Cordell, H. J. (2002). Epistasis : What it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, *11*(20), 2463-2468.

Crawford, N. G., Kelly, D. E., Hansen, M. E. B., Beltrame, M. H., Fan, S., Bowman, S. L., Jewett, E., Ranciaro, A., Thompson, S., Lo, Y., Pfeifer, S. P., Jensen, J. D., Campbell, M. C., Beggs, W., Hormozdiari, F., Mpoloka, S. W., Mokone, G. G., Nyambo, T., Meskel, D. W., … Tishkoff, S. (2017). Loci associated with skin pigmentation identified in African populations. *Science (New York, N.Y.)*, *358*(6365).

Creighton, H. B., & McClintock, B. (1931). A Correlation of Cytological and Genetical Crossing-Over in Zea Mays. *Proceedings of the National Academy of Sciences of the United States of America*, *17*(8), 492-497.

Darwin, C., Burndy Library, donor D., Henry Sotheran Ltd., bookseller D., & Edmonds & Remnants, binder D. (1859). *On the origin of species by means of natural selection, or, The preservation of favoured races in the struggle for life*. London : John Murray ... http://archive.org/details/onoriginofspec00darw

Daub, J. T., Dupanloup, I., Robinson-Rechavi, M., & Excoffier, L. (2015). Inference of Evolutionary Forces Acting on Human Biological Pathways. *Genome Biology and Evolution*, *7*(6), 1546-1558.

Daub, J. T., Hofer, T., Cutivet, E., Dupanloup, I., Quintana-Murci, L., Robinson-Rechavi, M., & Excoffier, L. (2013). Evidence for Polygenic Adaptation to Pathogens in the Human Genome. *Molecular Biology and Evolution*, *30*(7), 1544-1558.

De Mita, S., Chantret, N., Loridon, K., Ronfort, J., & Bataillon, T. (2011). Molecular adaptation in flowering and symbiotic recognition pathways : Insights from patterns of polymorphism in the legume Medicago truncatula. *BMC Evolutionary Biology*, *11*, 229.

De Mita, S., Ronfort, J., McKhann, H. I., Poncet, C., El Malki, R., & Bataillon, T. (2007). Investigation of the demographic and selective forces shaping the nucleotide diversity of genes involved in nod factor signaling in Medicago truncatula. *Genetics*, *177*(4), 2123-2133.

Deng, L., & Xu, S. (2017). Adaptation of human skin color in various populations. *Hereditas*, *155*.

Dievart, A., Gottin, C., Périn, C., Ranwez, V., & Chantret, N. (2020). Origin and Diversity of Plant Receptor-Like Kinases. *Annual Review of Plant Biology*, *71*, 131-156.

Ellegren, H., & Galtier, N. (2016). Determinants of genetic diversity. *Nature Reviews. Genetics*, *17*(7), 422-433.

Excoffier, L., Hofer, T., & Foll, M. (2009). Detecting loci under selection in a hierarchically structured population. *Heredity*, *103*(4), 285-298.

Fariello, M. I., Boitard, S., Mercier, S., Robelin, D., Faraut, T., Arnould, C., Recoquillay, J., Bouchez, O., Salin, G., Dehais, P., Gourichon, D., Leroux, S., Pitel, F., Leterrier, C., & SanCristobal, M. (2017). Accounting for linkage disequilibrium in genome scans for selection without individual genotypes : The local score approach. *Molecular Ecology*, *26*(14), 3700-3714.

Fariello, M. I., Boitard, S., Naya, H., SanCristobal, M., & Servin, B. (2013). Detecting Signatures of Selection Through Haplotype Differentiation Among Hierarchically Structured Populations. *Genetics*, *193*(3), 929-941.

Fay, J. C., & Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics*, *155*(3), 1405-1413.

Felsenstein, J. (1965). The effect of linkage on directional selection. *Genetics*, *52*(2), 349-363.

Ferrero-Serrano, Á., & Assmann, S. M. (2019). Phenotypic and genome-wide association with the local environment of Arabidopsis. *Nature Ecology and Evolution*, *3*(2), 274-285.

Fisher, R. A. (1918). The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Earth and Environmental Science Transactions of The Royal Society of Edinburgh*, *52*(2), 399-433.

Flood, P. J., & Hancock, A. M. (2017). The genomic basis of adaptation in plants. *Current Opinion in Plant Biology*, *36*, 88-94.

Frachon, L., Bartoli, C., Carrère, S., Bouchez, O., Chaubet, A., Gautier, M., Roby, D., & Roux, F. (2018). A Genomic Map of Climate Adaptation in Arabidopsis thaliana at a Micro-Geographic Scale. *Frontiers in Plant Science*, *9*.

Frachon, L., Mayjonade, B., Bartoli, C., Hautekèete, N.-C., & Roux, F. (2019). Adaptation to Plant Communities across the Genome of Arabidopsis thaliana. *Molecular Biology and Evolution*, *36*(7), 1442-1456.

François, O., Currat, M., Ray, N., Han, E., Excoffier, L., & Novembre, J. (2010). Principal component analysis under population genetic models of range expansion and admixture. *Molecular Biology and Evolution*, *27*(6), 1257-1268.

Gautier, M. (2015). Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates. *Genetics*, *201*(4), 1555-1579.

Gentzbittel, L., Ben, C., Mazurier, M., Shin, M.-G., Lorenz, T., Rickauer, M., Marjoram, P., Nuzhdin, S. V., & Tatarinova, T. V. (2019). WhoGEM : An admixture-based prediction machine accurately predicts quantitative functional traits in plants. *Genome Biology*, *20*(1), 106.

Gertz, J., Gerke, J. P., & Cohen, B. A. (2010). Epistasis in a quantitative trait captured by a molecular model of transcription factor interactions. *Theoretical Population Biology*, *77*(1), 1-5.

Glémin, S. (2007). Mating Systems and the Efficacy of Selection at the Molecular Level. *Genetics*, *177*(2), 905-916.

Glémin, S., & Ronfort, J. (2013). Adaptation and maladaptation in selfing and outcrossing species : new mutations versus standing variation: mating systems and adaptation. *Evolution*, *67*(1), 225-240.

Gomez-Roldan, V., Fermas, S., Brewer, P. B., Puech-Pagès, V., Dun, E. A., Pillot, J.-P., Letisse, F., Matusova, R., Danoun, S., Portais, J.-C., Bouwmeester, H., Bécard, G., Beveridge, C. A., Rameau, C., & Rochange, S. F. (2008). Strigolactone inhibition of shoot branching. *Nature*, *455*(7210), 189-194.

Gough, C., & Cullimore, J. (2011). Lipo-chitooligosaccharide signaling in endosymbiotic plant-microbe interactions. *Molecular Plant-Microbe Interactions: MPMI*, *24*(8), 867-878.

Gouy, A., Daub, J. T., & Excoffier, L. (2017). Detecting gene subnetworks under selection in biological pathways. *Nucleic Acids Research*, *45*(16), e149.

Griesmann, M., Chang, Y., Liu, X., Song, Y., Haberer, G., Crook, M. B., Billault-Penneteau, B., Lauressergues, D., Keller, J., Imanishi, L., Roswanjaya, Y. P., Kohlen, W., Pujic, P., Battenberg, K., Alloisio, N., Liang, Y., Hilhorst, H., Salgado, M. G., Hocher, V., … Cheng, S. (2018). Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis. *Science*, *361*(6398).

Grillo, M. A., De Mita, S., Burke, P. V., Solórzano-Lowell, K. L. S., & Heath, K. D. (2016). Intrapopulation genomics in a model mutualist : Population structure and candidate symbiosis genes under selection in Medicago truncatula. *Evolution; International Journal of Organic Evolution*, *70*(12), 2704-2717.

Grzeskowiak, L., Stephan, W., & Rose, L. E. (2014). Epistatic selection and coadaptation in the Prf resistance complex of wild tomato. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, *27*, 456-471.

Günther, T., & Coop, G. (2013). Robust Identification of Local Adaptation from Allele Frequencies. *Genetics*, *195*(1), 205-220.

Haasl, R. J., & Payseur, B. A. (2016). Fifteen years of genomewide scans for selection : Trends, lessons and unaddressed genetic sources of complication. *Molecular Ecology*, *25*(1), 5-23.

Haldane, J. B. S. (1927). A Mathematical Theory of Natural and Artificial Selection, Part V : Selection and Mutation. *Mathematical Proceedings of the Cambridge Philosophical Society*, *23*(7), 838-844.

Hancock, A. M., Brachi, B., Faure, N., Horton, M. W., Jarymowycz, L. B., Sperone, F. G., Toomajian, C., Roux, F., & Bergelson, J. (2011). Adaptation to climate across the Arabidopsis thaliana genome. *Science (New York, N.Y.)*, *334*(6052), 83-86.

Hansen, T. F. (2013). Why epistasis is important for selection and adaptation. *Evolution; International Journal of Organic Evolution*, *67*(12), 3501-3511.

Harris, E. E., & Meyer, D. (2006). The molecular signature of selection underlying human adaptations. *American Journal of Physical Anthropology*, *Suppl 43*, 89-130.

He, X., Qian, W., Wang, Z., Li, Y., & Zhang, J. (2010). Prevalent positive epistasis in Escherichia coli and Saccharomyces cerevisiae metabolic networks. *Nature Genetics*, *42*(3), 272-276.

Hill, W. G., & Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, *38*(6), 226-231.

Hu, X.-S., & Hu, Y. (2015). Genomic Scans of Zygotic Disequilibrium and Epistatic SNPs in HapMap Phase III Populations. *PLoS ONE*, *10*(6).

Hudson, R. R. (1990). Genetic Data Analysis. Methods for Discrete Population Genetic Data. *Science (New York, N.Y.)*, *250*(4980), 575.

Id-Lahoucine, S., Molina, A., Cánovas, A., & Casellas, J. (2019). Screening for epistatic selection signatures : A simulation study. *Scientific Reports*, *9*.

Izagirre, N., García, I., Junquera, C., de la Rúa, C., & Alonso, S. (2006). A scan for signatures of positive selection in candidate loci for skin pigmentation in humans. *Molecular Biology and Evolution*, *23*(9), 1697-1706.

Jagdishchandra Joshi, C., & Prasad, A. (2014). Epistatic interactions among metabolic genes depend upon environmental conditions. *Mol. BioSyst.*, *10*(10), 2578-2589.

Jombart, T., Pontier, D., & Dufour, A.-B. (2009). Genetic markers in the playground of multivariate analysis. *Heredity*, *102*(4), 330-341.

Josephs, E. B., Stinchcombe, J. R., & Wright, S. I. (2017). What can genome-wide association studies tell us about the evolutionary forces maintaining genetic variation for quantitative traits? *New Phytologist*, *214*(1), 21-33.

Kang, Y., Li, M., Sinharoy, S., & Verdier, J. (2016). A Snapshot of Functional Genetic Studies in Medicago truncatula. *Frontiers in Plant Science*, *7*.

Kang, Y., Sakiroglu, M., Krom, N., Stanton-Geddes, J., Wang, M., Lee, Y.-C., Young, N. D., & Udvardi, M. (2015). Genome-wide association of drought-related and biomass traits with HapMap SNPs in Medicago truncatula. *Plant, Cell & Environment*, *38*(10), 1997-2011.

Kelleher, J., Etheridge, A. M., & McVean, G. (2016). Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS Computational Biology*, *12*(5).

Kimura, M. (1968). Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genetical Research*, *11*(3), 247-269.

Kimura, M. (1971). Theoretical foundation of population genetics at the molecular level. *Theoretical Population Biology*, *2*(2), 174-208.

Kimura, Motoo. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press.

Koch, E., Ristroph, M., & Kirkpatrick, M. (2013). Long Range Linkage Disequilibrium across the Human Genome. *PLoS ONE*, *8*(12).

Lack, J. B., Cardeno, C. M., Crepeau, M. W., Taylor, W., Corbett-Detig, R. B., Stevens, K. A., Langley, C. H., & Pool, J. E. (2015). The Drosophila genome nexus : A population genomic resource of 623 Drosophila melanogaster genomes, including 197 from a single ancestral range population. *Genetics*, *199*(4), 1229-1241.

Laffont, C., Huault, E., Gautrat, P., Endre, G., Kalo, P., Bourion, V., Duc, G., & Frugier, F. (2019). Independent Regulation of Symbiotic Nodulation by the SUNN Negative and CRA2 Positive Systemic Pathways. *Plant Physiology*, *180*(1), 559-570.

Laloum, T., Baudin, M., Frances, L., Lepage, A., Billault-Penneteau, B., Cerri, M. R., Ariel, F., Jardinaud, M.-F., Gamas, P., de Carvalho-Niebel, F., & Niebel, A. (2014). Two CCAAT-box-binding transcription factors redundantly regulate early steps of the legume-rhizobia endosymbiosis. *The Plant Journal: For Cell and Molecular Biology*, *79*(5), 757-768

Lamason, R. L., Mohideen, M.-A. P. K., Mest, J. R., Wong, A. C., Norton, H. L., Aros, M. C., Jurynec, M. J., Mao, X., Humphreville, V. R., Humbert, J. E., Sinha, S., Moore, J. L., Jagadeeswaran, P., Zhao, W., Ning, G., Makalowska, I., McKeigue, P. M., O'donnell, D., Kittles, R., … Cheng, K. C. (2005). SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science (New York, N.Y.)*, *310*(5755), 1782-1786.

Lefebvre, B., Timmers, T., Mbengue, M., Moreau, S., Hervé, C., Tóth, K., Bittencourt-Silvestre, J., Klaus, D., Deslandes, L., Godiard, L., Murray, J. D., Udvardi, M. K., Raffaele, S., Mongrand, S., Cullimore, J., Gamas, P., Niebel, A., & Ott, T. (2010). A remorin protein interacts with symbiotic receptors and regulates bacterial infection. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(5), 2343-2348.

Lewontin, R. C., & Kojima, K. (1960). The Evolutionary Dynamics of Complex Polymorphisms. *Evolution*, *14*(4), 458-472.

Lewontin, R. C., & Krakauer, J. (1973). Distribution of gene frequency as a test of the theory

of the selective neutrality of polymorphisms. *Genetics*, *74*(1), 175-195.

Li, H., & Ralph, P. (2019). Local PCA Shows How the Effect of Population Structure Differs Along the Genome. *Genetics*, *211*(1), 289-304.

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M., Cavalli-Sforza, L. L., & Myers, R. M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science (New York, N.Y.)*, *319*(5866), 1100-1104.

Madsen, L. H., Tirichine, L., Jurkiewicz, A., Sullivan, J. T., Heckmann, A. B., Bek, A. S., Ronson, C. W., James, E. K., & Stougaard, J. (2010). The molecular network governing nodule organogenesis and infection in the model legume Lotus japonicus. *Nature Communications*, *1*(1), 1-12.

Mangin, B., Siberchicot, A., Nicolas, S., Doligez, A., This, P., & Cierco-Ayrolles, C. (2012). Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity*, *108*(3), 285-291.

Marjoram, P., & Wall, J. D. (2006). Fast « coalescent » simulation. *BMC Genetics*, *7*, 16.

Marsh, J. F., Rakocevic, A., Mitra, R. M., Brocard, L., Sun, J., Eschstruth, A., Long, S. R., Schultze, M., Ratet, P., & Oldroyd, G. E. D. (2007). Medicago truncatula NIN is essential for rhizobial-independent nodule organogenesis induced by autoactive calcium/calmodulin-dependent protein kinase. *Plant Physiology*, *144*(1), 324-335.

McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, *351*(6328), 652-654.

McVean, G. A. T., & Cardin, N. J. (2005). Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *360*(1459), 1387-1393.

McVean, G., Spencer, C. C. A., & Chaix, R. (2005). Perspectives on Human Genetic Variation from the HapMap Project. *PLoS Genetics*, *1*(4).

Mortier, V., De Wever, E., Vuylsteke, M., Holsters, M., & Goormachtig, S. (2012). Nodule numbers are governed by interaction between CLE peptides and cytokinin signaling. *The Plant Journal: For Cell and Molecular Biology*, *70*(3), 367-376.

Mortier, V., Den Herder, G., Whitford, R., Van de Velde, W., Rombauts, S., D'Haeseleer, K., Holsters, M., & Goormachtig, S. (2010). CLE peptides control Medicago truncatula nodulation locally and systemically. *Plant Physiology*, *153*(1), 222-237.

Nei, M. (1988). Relative roles of mutation and selection in the maintenance of genetic variability. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *319*(1196), 615-629.

Neuvial, P., Ambroise, C., Chaturvedi, S., Dehman, A., Koskas, M., Rigaill, G., & Vialaneix, N. (2017). *adjclust : Adjacency-Constrained Clustering of a Block-Diagonal Similarity Matrix (R package, available on CRAN)*. https://hal.inrae.fr/hal-02791370

Niel, C., Sinoquet, C., Dina, C., & Rocheleau, G. (2015). A survey about methods dedicated to epistasis detection. *Frontiers in Genetics*, *6*.

Nielsen, R. (2005). Molecular signatures of natural selection. *Annual Review of Genetics*, *39*, 197-218.

Nordborg, M. (2000). Linkage disequilibrium, gene trees and selfing : An ancestral recombination graph with partial self-fertilization. *Genetics*, *154*(2), 923-929.

Nordborg, Magnus, Hu, T. T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., Bakker, E., Calabrese, P., Gladstone, J., Goyal, R., Jakobsson, M., Kim, S., Morozov, Y., Padhukasahasram, B., Plagnol, V., Rosenberg, N. A., Shah, C., Wall, J. D., Wang, J., … Bergelson, J. (2005). The pattern of polymorphism in Arabidopsis thaliana. *PLoS Biology*, *3*(7), e196.

Nordborg, Magnus, & Weigel, D. (2008). Next-generation genetics in plants. *Nature*, *456*(7223), 720-723.

Ohta, T. (Mar 1982a). Linkage disequilibrium due to random genetic drift in finite subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America*, *79*(6), 1940-1944.

Ohta, T. (May 1982b). Linkage disequilibrium with the island model. *Genetics*, *101*(1), 139-155.

Oldroyd, G. E. D. (2013). Speak, friend, and enter : Signalling systems that promote beneficial symbiotic associations in plants. *Nature Reviews. Microbiology*, *11*(4), 252-263.

Otto, S. P., & Whitlock, M. C. (2009). The impact of epistatic selection on the genomic traces of selection. *Molecular Ecology*, *18*(24), 4985-4987.

Paape, T., Bataillon, T., Zhou, P., J Y Kono, T., Briskine, R., Young, N. D., & Tiffin, P. (2013). Selection, genome-wide fitness effects and evolutionary rates in the model legume Medicago truncatula. *Molecular Ecology*, *22*(13), 3525-3538.

Parniske, M. (2008). Arbuscular mycorrhiza : The mother of plant root endosymbioses.

Nature Reviews. Microbiology, *6*(10), 763-775.

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2*(12), e190.

Pavlidis, P., & Alachiotis, N. (2017). A survey of methods and tools to detect recent and strong positive selection. *Journal of Biological Research (Thessalonike, Greece)*, *24*, 7.

Pecrix, Y., Staton, S. E., Sallet, E., Lelandais-Brière, C., Moreau, S., Carrère, S., Blein, T., Jardinaud, M.-F., Latrasse, D., Zouine, M., Zahm, M., Kreplak, J., Mayjonade, B., Satgé, C., Perez, M., Cauet, S., Marande, W., Chantry-Darmon, C., Lopez-Roques, C., … Gamas, P. (2018). Whole-genome landscape of Medicago truncatula symbiotic genes. *Nature Plants*, *4*(12), 1017-1025.

Peng, B., & Amos, C. I. (2008). Forward-time simulations of non-random mating populations using simuPOP. *Bioinformatics (Oxford, England)*, *24*(11), 1408-1409.

Peng, B., & Kimmel, M. (2005). simuPOP : A forward-time population genetics simulation environment. *Bioinformatics (Oxford, England)*, *21*(18), 3686-3687.

Phillips, P. C. (1998). The Language of Gene Interaction. *Genetics*, *149*(3), 1167-1171.

Phillips, P. C. (2008). Epistasis—The essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, *9*(11), 855-867.

Phillips, P. C., Otto, S. P., & Whitlock, M. C. (1997). *The Evolutionary Importance of Gene Interactions and Variability o f Epistatic Effects*. 11.

Pollak, E. (1987). On the Theory of Partially Inbreeding Finite Populations. I. Partial Selfing. *Genetics*, *117*(2), 353-360.

Pool, J. E. (2015). The Mosaic Ancestry of the Drosophila Genetic Reference Panel and the D. melanogaster Reference Genome Reveals a Network of Epistatic Fitness Interactions. *Molecular Biology and Evolution*, *32*(12), 3236-3251.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904-909.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945-959.

Pritykin, Y., & Singh, M. (2013). Simple Topological Features Reflect Dynamics and Modularity in Protein Interaction Networks. *PLoS Computational Biology*, *9*(10), e1003243.

Qian, W., Zhou, H., & Tang, K. (2015). Recent Coselection in Human Populations Revealed by Protein–Protein Interaction Network. *Genome Biology and Evolution*, *7*(1), 136-153.

Radhakrishnan, G. V., Keller, J., Rich, M. K., Vernié, T., Mbadinga Mbadinga, D. L., Vigneron, N., Cottret, L., Clemente, H. S., Libourel, C., Cheema, J., Linde, A.-M., Eklund, D. M., Cheng, S., Wong, G. K. S., Lagercrantz, U., Li, F.-W., Oldroyd, G. E. D., & Delaux, P.-M. (2020). An ancestral signalling pathway is conserved in intracellular symbioses-forming plant lineages. *Nature Plants*, *6*(3), 280-289.

Rey, T., Bonhomme, M., Chatterjee, A., Gavrin, A., Toulotte, J., Yang, W., André, O., Jacquet, C., & Schornack, S. (2017). The Medicago truncatula GRAS protein RAD1 supports arbuscular mycorrhiza symbiosis and Phytophthora palmivora susceptibility. *Journal of Experimental Botany*, *68*(21-22), 5871-5881.

Rogers, A. R., & Huff, C. (2009). Linkage Disequilibrium Between Loci With Unknown Phase. *Genetics*, *182*(3), 839-844.

Ronfort, J., Bataillon, T., Santoni, S., Delalande, M., David, J. L., & Prosperi, J.-M. (2006). Microsatellite diversity and broad scale geographic structure in a model legume : Building a set of nested core collection for studying naturally occurring variation in Medicago truncatula. *BMC Plant Biology*, *6*, 28.

Ronfort, J., & Glemin, S. (2012). MATING SYSTEM, HALDANE'S SIEVE, AND THE DOMESTICATION PROCESS. *Evolution*, no-no.

Rosenberg, N. A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J. K., & Feldman, M. W. (2005). Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genetics*, *1*(6), e70.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., & Feldman, M. W. (2002). Genetic structure of human populations. *Science (New York, N.Y.)*, *298*(5602), 2381-2385.

Rosenberg, S., Templeton, A. R., Feigin, P. D., Lancet, D., Beckmann, J. S., Selig, S., Hamer, D. H., & Skorecki, K. (2006). The association of DNA sequence variation at the MAOA genetic locus with quantitative behavioural traits in normal males. *Human Genetics*, *120*(4), 447-459.

Sabeti, P. C., Schaffner, S. F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T. S., Altshuler, D., & Lander, E. S. (2006). Positive natural selection in the human lineage. *Science (New York, N.Y.)*, *312*(5780), 1614-1620.

Sabeti, Pardis C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., & Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, *419*(6909), 832-837.

Sabeti, Pardis C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., & Lander, E. S. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, *449*(7164), 913-918.

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., Hunt, S. E., Cole, C. G., Coggill, P. C., Rice, C. M., Ning, Z., Rogers, J., Bentley, D. R., Kwok, P. Y., Mardis, E. R., … International SNP Map Working Group. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, *409*(6822), 928-933.

Sadier, A., Viriot, L., Pantalacci, S., & Laudet, V. (2014). The ectodysplasin pathway : From diseases to adaptations. *Trends in Genetics: TIG*, *30*(1), 24-31.

Satgé, C. (2016). *Importance de l'ADN déméthylase DEMETER lors du développement nodulaire au cours de la symbiose Medicago truncatula/Sinorhizobium meliloti* [These de doctorat, Toulouse, INSA]. http://www.theses.fr/2016ISAT0004

Schauser, L., Roussis, A., Stiller, J., & Stougaard, J. (1999). A plant regulator controlling development of symbiotic root nodules. *Nature*, *402*(6758), 191-195.

Schork, N. J., Murray, S. S., Frazer, K. A., & Topol, E. J. (2009). Common vs. Rare allele hypotheses for complex diseases. *Current Opinion in Genetics & Development*, *19*(3), 212-219.

Schumer, M., & Brandvain, Y. (2016). Determining epistatic selection in admixed populations. *Molecular Ecology*, *25*(11), 2577-2591.

Segrè, D., DeLuna, A., Church, G. M., & Kishony, R. (2005). Modular epistasis in yeast metabolism. *Nature Genetics*, *37*(1), 77-83.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, *13*(11),

2498-2504.

Signor, C. L., Aimé, D., Bordat, A., Belghazi, M., Labas, V., Gouzy, J., Young, N. D., Prosperi, J.-M., Leprince, O., Thompson, R. D., Buitink, J., Burstin, J., & Gallardo, K. (2017). Genome-wide association studies with proteomics data reveal genes important for synthesis, transport and packaging of globulins in legume seeds. *New Phytologist*, *214*(4), 1597-1613.

Siol, M., Prosperi, J. M., Bonnin, I., & Ronfort, J. (2008). How multilocus genotypic pattern helps to understand the history of selfing populations : A case study in Medicago truncatula. *Heredity*, *100*(5), 517-525.

Slatkin, M. (2008). Linkage disequilibrium—Understanding the evolutionary past and mapping the medical future. *Nature Reviews. Genetics*, *9*(6), 477-485.

Speidel, L., Forest, M., Shi, S., & Myers, S. R. (2019). A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, *51*(9), 1321-1329.

Staab, P. R., Zhu, S., Metzler, D., & Lunter, G. (2015). scrm : Efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics*, *31*(10), 1680-1682.

Stanton-Geddes, J., Paape, T., Epstein, B., Briskine, R., Yoder, J., Mudge, J., Bharti, A. K., Farmer, A. D., Zhou, P., Denny, R., May, G. D., Erlandson, S., Yakub, M., Sugawara, M., Sadowsky, M. J., Young, N. D., & Tiffin, P. (2013). Candidate Genes and Genetic Architecture of Symbiotic and Agronomic Traits Revealed by Whole-Genome, Sequence-Based Association Genetics in Medicago truncatula. *PLoS ONE*, *8*(5), e65688.

Steiner, C. C., Weber, J. N., & Hoekstra, H. E. (2007). Adaptive variation in beach mice produced by two interacting pigmentation genes. *PLoS Biology*, *5*(9), e219.

Sung, W., Ackerman, M. S., Miller, S. F., Doak, T. G., & Lynch, M. (2012). Drift-barrier hypothesis and mutation-rate evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(45), 18488-18492.

Tadege, M., Wen, J., He, J., Tu, H., Kwak, Y., Eschstruth, A., Cayrel, A., Endre, G., Zhao, P. X., Chabaud, M., Ratet, P., & Mysore, K. S. (2008). Large-scale insertional mutagenesis using the Tnt1 retrotransposon in the model legume Medicago truncatula. *The Plant Journal: For Cell and Molecular Biology*, *54*(2), 335-347.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA

polymorphism. *Genetics*, *123*(3), 585-595.

Takahasi, K. R. (2007). Evolution of Coadaptation in a Subdivided Population. *Genetics*, *176*(1), 501-511.

Takahasi, K. R. (2009). Coalescent under the evolution of coadaptation. *Molecular Ecology*, *18*(24), 5018-5029.

Takahasi, K. R., & Innan, H. (2008). The Direction of Linkage Disequilibrium : A New Measure Based on the Ancestral-Derived Status of Segregating Alleles. *Genetics*, *179*(3), 1705-1712.

Takahasi, K. R., & Tajima, F. (2005). Evolution of coadaptation in a two-locus epistatic system. *Evolution; International Journal of Organic Evolution*, *59*(11), 2324-2332.

Takahata, N., & Nei, M. (1990). Allelic Genealogy under Overdominant and Frequency-Dependent Selection and Polymorphism of Major Histocompatibility Complex Loci. *Genetics*, *124*(4), 967-978.

Tan, S., Sanchez, M., Laffont, C., Boivin, S., Le Signor, C., Thompson, R. D., Frugier, F., & Brault, M. (2020). A cytokinin signalling type-B response regulator transcription factor acting in early nodulation. *Plant Physiology*.

Tang, H., Krishnakumar, V., Bidwell, S., Rosen, B., Chan, A., Zhou, S., Gentzbittel, L., Childs, K. L., Yandell, M., Gundlach, H., Mayer, K. F. X., Schwartz, D. C., & Town, C. D. (2014). An improved genome release (version Mt4.0) for the model legume Medicago truncatula. *BMC Genomics*, *15*, 312.

Tenaillon, M. I., Sawkins, M. C., Long, A. D., Gaut, R. L., Doebley, J. F., & Gaut, B. S. (2001). Patterns of DNA sequence polymorphism along chromosome 1 of maize (Zea mays ssp. Mays L.). *Proceedings of the National Academy of Sciences of the United States of America*, *98*(16), 9161-9166.

Valdés-López, O., Jayaraman, D., Maeda, J., Delaux, P.-M., Venkateshwaran, M., Isidra-Arellano, M. C., Reyero-Saavedra, M. del R., Sánchez-Correa, M. del S., Verastegui-Vidal, M. A., Delgado-Buenrostro, N., Van Ness, L., Mysore, K. S., Wen, J., Sussman, M. R., & Ané, J.-M. (2019). A Novel Positive Regulator of the Early Stages of Root Nodule Symbiosis Identified by Phosphoproteomics. *Plant & Cell Physiology*, *60*(3), 575-586.

Vitti, J. J., Grossman, S. R., & Sabeti, P. C. (2013). Detecting natural selection in genomic data. *Annual Review of Genetics*, *47*, 97-120.

Voight, B. F., Kudaravalli, S., Wen, X., & Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biology*, *4*(3), e72.

Wade, M. J., Winther, R. G., Agrawal, A. F., & Goodnight, C. J. (2001). Alternative definitions of epistasis : Dependence and interaction. *Trends in Ecology & Evolution*, *16*(9), 498-504.

Wakeley, J. (1999). Nonequilibrium migration in human history. *Genetics*, *153*(4), 1863-1871.

Wang, R. K., Lu, J. J., Xing, G. N., Gai, J. Y., & Zhao, T. J. (2011). Molecular evolution of two consecutive carotenoid cleavage dioxygenase genes in strigolactone biosynthesis in plants. *Genetics and Molecular Research: GMR*, *10*(4), 3664-3673.

Weigand, H., Weiss, M., Cai, H., Li, Y., Yu, L., Zhang, C., & Leese, F. (2018). Fishing in troubled waters : Revealing genomic signatures of local adaptation in response to freshwater pollutants in two macroinvertebrates. *The Science of the Total Environment*, *633*, 875-891.

Weinreich, D. M., Watson, R. A., & Chao, L. (2005). Perspective : Sign epistasis and genetic constraint on evolutionary trajectories. *Evolution; International Journal of Organic Evolution*, *59*(6), 1165-1174.

Wong, G. K.-S., Liu, B., Wang, J., Zhang, Y., Yang, X., Zhang, Z., Meng, Q., Zhou, J., Li, D., Zhang, J., Ni, P., Li, S., Ran, L., Li, H., Zhang, J., Li, R., Li, S., Zheng, H., Lin, W., … International Chicken Polymorphism Map Consortium. (2004). A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. *Nature*, *432*(7018), 717-722.

Wright, S. (1931). Evolution in Mendelian Populations. *Genetics*, *16*(2), 97-159.

Wright, S. (1978). *Evolution and the Genetics of Populations, Volume 4 : Variability Within and Among Natural Populations*. University of Chicago Press.

Yang, T., Deng, H.-W., & Niu, T. (2014). Critical assessment of coalescent simulators in modeling recombination hotspots in genomic sequences. *BMC Bioinformatics*, *15*, 3.

Ye, Y., Ding, Y., Jiang, Q., Wang, F., Sun, J., & Zhu, C. (2017). The role of receptor-like protein kinases (RLKs) in abiotic stress response in plants. *Plant Cell Reports*, *36*(2), 235-242.

Yoder, J. B., Stanton-Geddes, J., Zhou, P., Briskine, R., Young, N. D., & Tiffin, P. (2014). Genomic signature of adaptation to climate in Medicago truncatula. *Genetics*, *196*(4), 1263-1275.

Young, N. D., Debellé, F., Oldroyd, G. E. D., Geurts, R., Cannon, S. B., Udvardi, M. K.,

Benedito, V. A., Mayer, K. F. X., Gouzy, J., Schoof, H., Van de Peer, Y., Proost, S., Cook, D. R., Meyers, B. C., Spannagl, M., Cheung, F., De Mita, S., Krishnakumar, V., Gundlach, H., … Roe, B. A. (2011). The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature*, *480*(7378), 520-524.

Yukilevich, R., Lachance, J., Aoki, F., & True, J. R. (2008). Long-term adaptation of epistatic genetic networks. *Evolution; International Journal of Organic Evolution*, *62*(9), 2215-2235.

Zan, Y., Forsberg, S. K. G., & Carlborg, Ö. (2018). On the Relationship Between High-Order Linkage Disequilibrium and Epistasis. *G3: Genes|Genomes|Genetics*, *8*(8), 2817-2824.

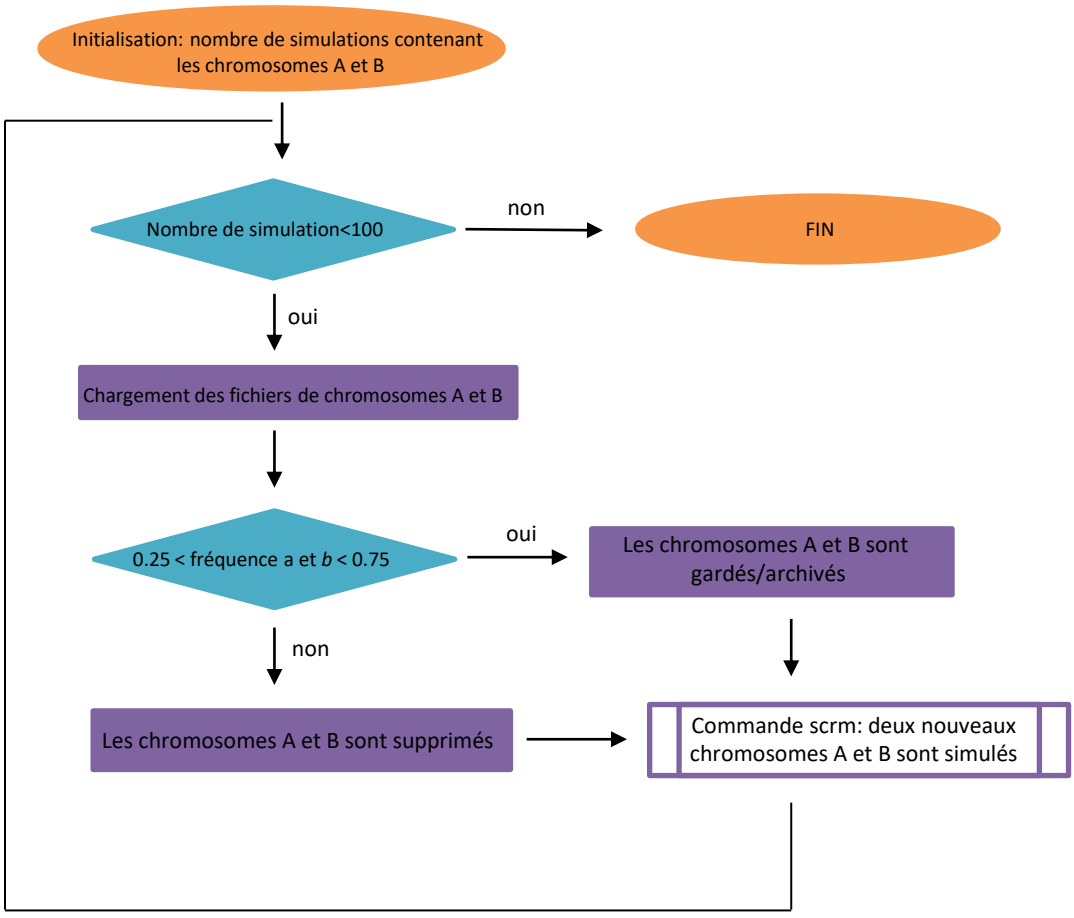
Zeng, K., Fu, Y.-X., Shi, S., & Wu, C.-I. (2006). Statistical Tests for Detecting Positive Selection by Utilizing High-Frequency Variants. *Genetics*, *174*(3), 1431-1439.

Annexes

Annexe 1. Méthode de simulation de la population ancestrale par coalescence.

Les simulations réalisées avec la méthode de coalescence ont été faites en deux temps:

1. Les chromosomes A et B qui portent les SNP qui seront co-sélectionnés dans la partie suivante sont simulés à l'aide d'un script python qui sélectionne uniquement les chromosomes simulés dont les futurs SNP co-sélectionnés sont en fréquences intermédiaires. Tous les chromosomes simulés qui ne remplissent pas cette condition ne sont pas gardés.



L'organigramme représente le script python *scrm_ancestral_population_A_B.py* permettant de simuler les chromosomes A et B. Afin d'optimiser les temps de calcul, ce script est exécuté 10 fois en parallèle et les simulations sont réalisées par groupe de 100 afin d'obtenir les 1000 chromosomes A et B. Les scripts sont exécutés en parallèle à l'aide de fonction *sarray* disponible sur le cluster de calcul de la Genotoul.

La commande python pour exécuter SCRM:

```
cmd1 = "scrm n nrep -t theta -r rho l -p 8 > /path/simul_1/file_chromosomesA.txt"
os.system(cmd1)
```

Commande sarray : ~ *$ sarray –mem=XG file_bash.sh*

a. Le *file_bash.sh* exécute le script python en fonction des argument fournis.

Exemple *file_bash.sh* pour 2 exécutions

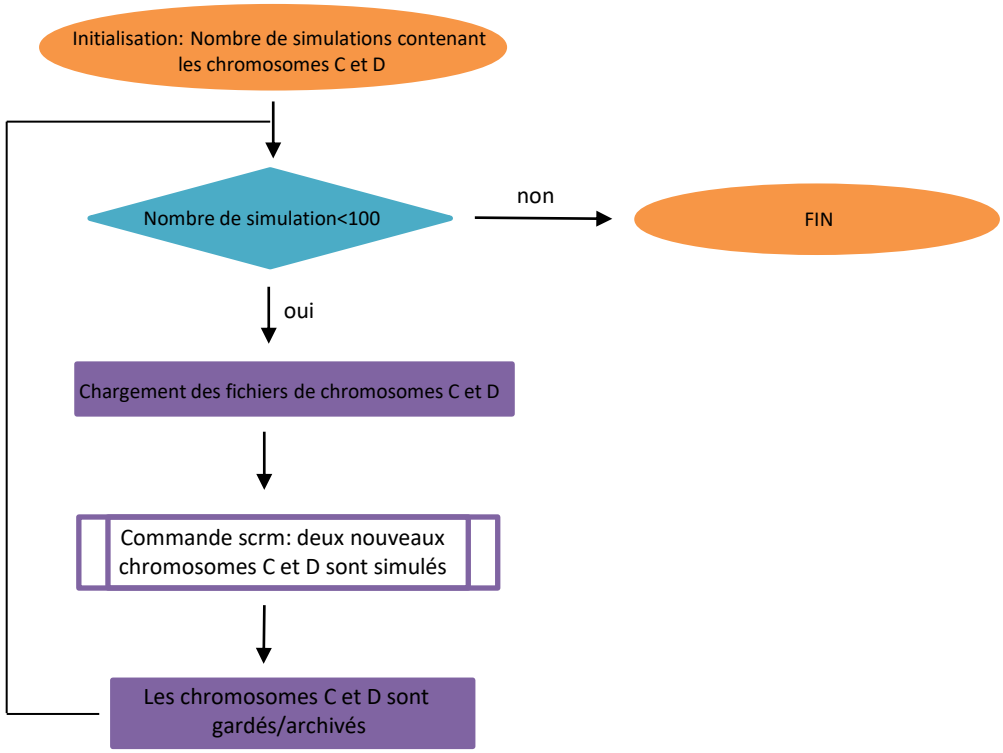
```
module load system/Python-3.4.3;module load bioinfo/scrm-1.7.2;python /path/scrm_ancestral_population _A_B.py 'argument1' 'argument2'
module load system/Python-3.4.3;module load bioinfo/scrm-1.7.2;python /path/scrm_ancestral_population _A_B.py 'argument1' 'argument2'
```

b. L'argument *--mem* correspond à la mémoire requise par nœud sachant que chaque exécution se fait sur une nœud différent (chaque ligne de *file_bash.sh*). Pour ce script, 35G ont été demandés.

Annexe 1. Méthode de simulation de la population ancestrale par coalescence.

Les simulations réalisées avec la méthode de coalescence ont été faites en deux temps:

2. Les chromosomes C et D sans appliquer de filtre.



L'organigramme représente le script python *scrm_ancestral_population _C_D.py* utilisé pour simuler les chromosomes C et D. Le script est également exécuté 10 fois en parallèle à l'aide la fonction *sarray.*

Commande sarray : ~ *$ sarray –mem=XG file_bash2.sh*

a. Le *file_bash.sh* exécute le script python en fonction des argument fournis.

Exemple *file_bash2.sh* pour 2 exécutions

module load system/Python-3.4.3;module load bioinfo/scrm-1.7.2;python /path/scrm_ancestral_population _C_D.py 'argument1' 'argument2'
module load system/Python-3.4.3;module load bioinfo/scrm-1.7.2;python /path/scrm_ancestral_population _C_D.py 'argument1' 'argument2'

Annexe 2. Méthode de simulation « forward in time ».

Les simulations « forward in time » sont réalisées en deux temps:

1. La population ancestrale est le point de départ des simulations « forward in time ». Les 100 premières générations sont simulées sous neutralité en panmixie et en autogamie, la population est structurée en deux sous-populations et évolue pendant 100 générations.

```
1  from simuPOP import *
2  pop  = importPopulation(format='ms', filename = "chromosomeA.txt",ploidy = 2)
3  pop2 = importPopulation(format='ms', filename = "chromosomeB.txt",ploidy = 2)
4  pop3 = importPopulation(format='ms', filename = "chromosomeC.txt",ploidy = 2)
5  pop4 = importPopulation(format='ms', filename = "chromosomeD.txt",ploidy = 2)
6  pop.addChromFrom(pop2)
7  pop.addChromFrom(pop3)
8  pop.addChromFrom(pop4)
9  pop.splitSubPop(0, [250, 250])
10 pop.setVirtualSplitter(sim.ProportionSplitter([0.05, 0.95]))
11
12 """ """ """ """ """ """ """
13 """ Population Evolved  """
14 """ """ """ """ """ """ """
15 pop.evolve(
16 initOps=[
17 sim.InitSex(),
18 sim.IdTagger()
19 ],
20 preOps=[sim.PyOperator(func=infSite)],
21 matingScheme=sim.HeteroMating(matingSchemes=[
22 sim.SelfMating(subPops=[(0, 1),(1, 1)],ops=[sim.Recombinator(intensity=rate_c)]),
23 sim.RandomMating(subPops=[(0, 0),(1, 0)],ops=[sim.Recombinator(intensity=rate_c)])
24 ]),
25 postOps=[
26 sim.Stat(numOfSegSites=sim.ALL_AVAIL, vars='segSites'),
27 sim.SavePopulation(output="!'snapshotN_%d.pop' % (gen)",step = 20)
28 ],
29 gen = 100
30 )
```

Cet exemple illustre comment une population structurée en deux sous-populations est simulée pendant 100 générations sous un modèle de sélection neutre en autogamie avec recombinaison et mutation.

- La ligne 1 importe la bibliothèque python Simpop.
- Lignes 2-5, la fonction `importPopulation()` charge les quatre chromosomes simulés avec « scrm ». Les fichiers de chromosome sont au format « ms » et les individus sont diploïdes; « `ploidy = 2` ».
- Lignes 6-8, la fonction `addChromFrom()`ajoute tous les chromosomes au sein d'une même population. La population totale « `pop` » est constituée de 500 individus diploïdes et 4 chromosomes.
- Ligne 9, la fonction `splitSubPop()` divise la population « `pop` » (appelée 0) en deux sous-populations de tailles égales de 250 individus chacune. Les deux nouvelle sous-populations s'appellent respectivement 0 et 1 (la reproduction ne se fera qu'entre individus d'une même sous-population).
- Ligne 10, la fonction `setVirtualSplitter(sim.ProportionSplitter())` divise la population en deux sous-populations virtuelles autrement appelée VSP (« Virtual subpopulation »). Les VPS permettent de définir des groupes d'individus qui partagent les mêmes propriétés. La fonction nous permettra d'appliquer deux modes de reproduction différents aux deux VPS aux proportions 5% et 95%.

Annexe 2. Méthode de simulation « forward in time ».

- Ligne 15, la dernière fonction `evolve()`fait évoluer la population pendant 100 générations sous condition de différents opérateurs:

Les deux premiers opérateurs sont appliqués avant l'évolution; paramètre `initOps` de la fonction `evolve()`. L'operateur `InitSex()` initialise le sexe des individus de manière aléatoire et l'opérateur `IdTagger()`attribue un identifiant unique à chaque individu qu'il conservent au cours de l'évolution. Le paramètre `preOps()` applique les opérateurs sur la population à chaque génération avant la reproduction. L'opérateur `sim.PyOperator()` appelle une fonction python et dans cet exemple, il appelle la fonction `func=infSite` qui définit le nombre de mutations appliquées à la population à chaque génération (voir partir I.3.2). Puis, `matingScheme` permet de définir les modes de reproduction et les fonctions qui génèrent la population de descendants à partir de la population parentale. Il y a plusieurs types de « mating schemes » disponibles et dans cet exemple, nous appliquons un mode de reproduction hétérogène avec `sim.HeteroMating()` qui applique deux modes de reproduction au deux sous-populations virtuelles créées en amont. Les deux systèmes de reproductions sont `sim.SelfMating()`et `sim.RandomMating()`. Les deux sous-populations réelles créées en ligne 9 (`pop.splitSubPop(0, [250, 250])`) sont appelées dans les premiers chiffres en parenthèse: « `subPops=[(0, 1),(1, 1)]` » et les sous-populations virtuelles (VPS) qui définissent les deux modes de reproduction sont appelées par les deuxièmes chiffres entre parenthèses: « `subPops=[(0, 1),(1, 1)]` » et les deux VPS sont en proportion 5% et 95%. Ainsi, dans chaque sous-population de 250 individus, il y a 5% de le reproduction qui se fait en panmixie et 95% qui se fait en autogamie (les 95% en autogamie sont désignés par 1 et les 5% panmixie sont désignés par 0). Pendant la reproduction, l'opérateur `sim.Recombinator()` applique la recombinaison à un taux « `rate_c` ». Enfin le paramètre `postOps` applique les opérateurs sur la population de descendants à chaque génération après la reproduction et l'opérateur `sim.SavePopulation()` permet d'extraire un fichier «`snapshot.pop` » contenant les chromosomes qui constituent l'ensemble de la population toutes les 20 générations.

La simulation réalisée avec un mode de reproduction à 100% panmictique (non représentée) est effectuée sans VPS (virtual subpopulation) et l'opérateur de reproduction est simplement « `randomMating()` »:

```
matingScheme=sim.RandomMating(subPopSize=[250, 250],ops=[sim.Recombinator(intensity=rate_c)]),
```

Annexe 2. Méthode de simulation « forward in time ».

Les simulations « forward in time » sont réalisées en deux temps:

2. Les différents modèles de sélection sont simulés pendant 200 générations.

```
1  from simuPOP import *
2  pop = sim.loadPopulation(path+'/snapshotN_100.pop')
3  pop.setVirtualSplitter(sim.ProportionSplitter([0.05, 0.95]))
4
5 """ """ """ """ """ """ """
6 """ Population Evolved  """
7 """ """ """ """ """ """ """
8
9  pop.evolve(
10 initOps=[
11 sim.InitSex(),
12 sim.IdTagger()
13 ],
14 preOps=[sim.PyOperator(func=infSite)], # mutation fonction call
15 matingScheme=sim.HeteroMating(matingSchemes=[ # mating scheme
16 sim.SelfMating(subPops=[(0, 1),(1, 1)],ops=[sim.Recombinator(intensity=rate_c)]),
17 sim.RandomMating(subPops=[(0, 0),(1, 0)],ops=[sim.Recombinator(intensity=rate_c)])
18 ]),
19 postOps=[
20 sim.Stat(numOfSegSites=sim.ALL_AVAIL, vars='segSites'),
21 sim.MaSelector(fitness=[1, 1, 1, 1, 1, 1, 1, 1, 1+s+s],loci=goodLoci, subPops=[(0,0), (0,1)]),
22 sim.MaSelector(fitness=[1, 1, 1, 1, 1, 1, 1, 1, 1+s+s],loci=goodLoci, subPops=[(1,0), (1,1)]),
23 sim.SavePopulation(output="!'path/snapshotN_%d.pop' % (gen)",step = 20)
24 ],
25 gen = 200
26 )
```

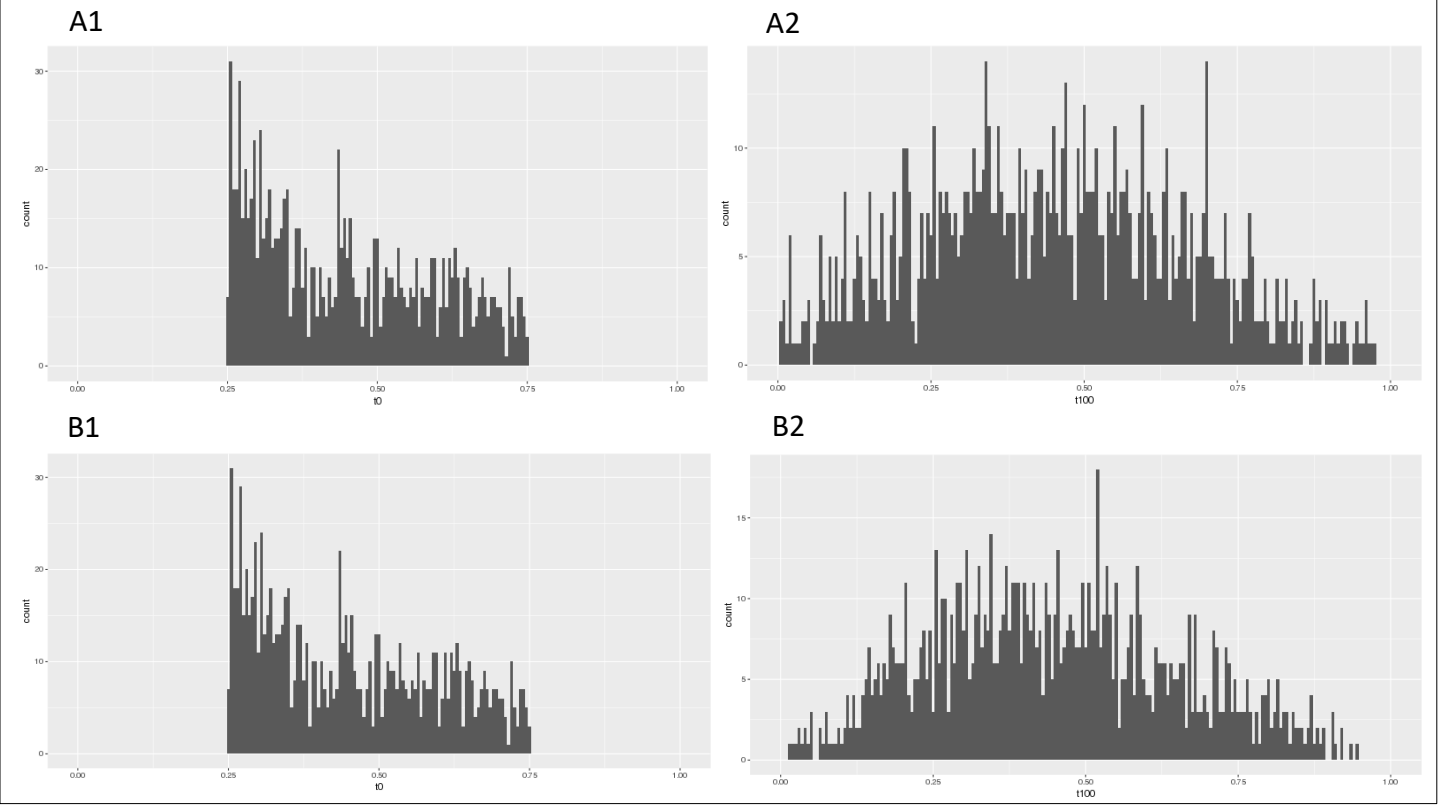
Cet exemple illustre comment une population structurée en deux sous-populations est simulée pendant 200 générations sous un modèle de sélection épistatique coadapté en autogamie avec recombinaison et mutation.

La plupart des opérateurs sont identiques à l'exemple précédent à l'exception de:

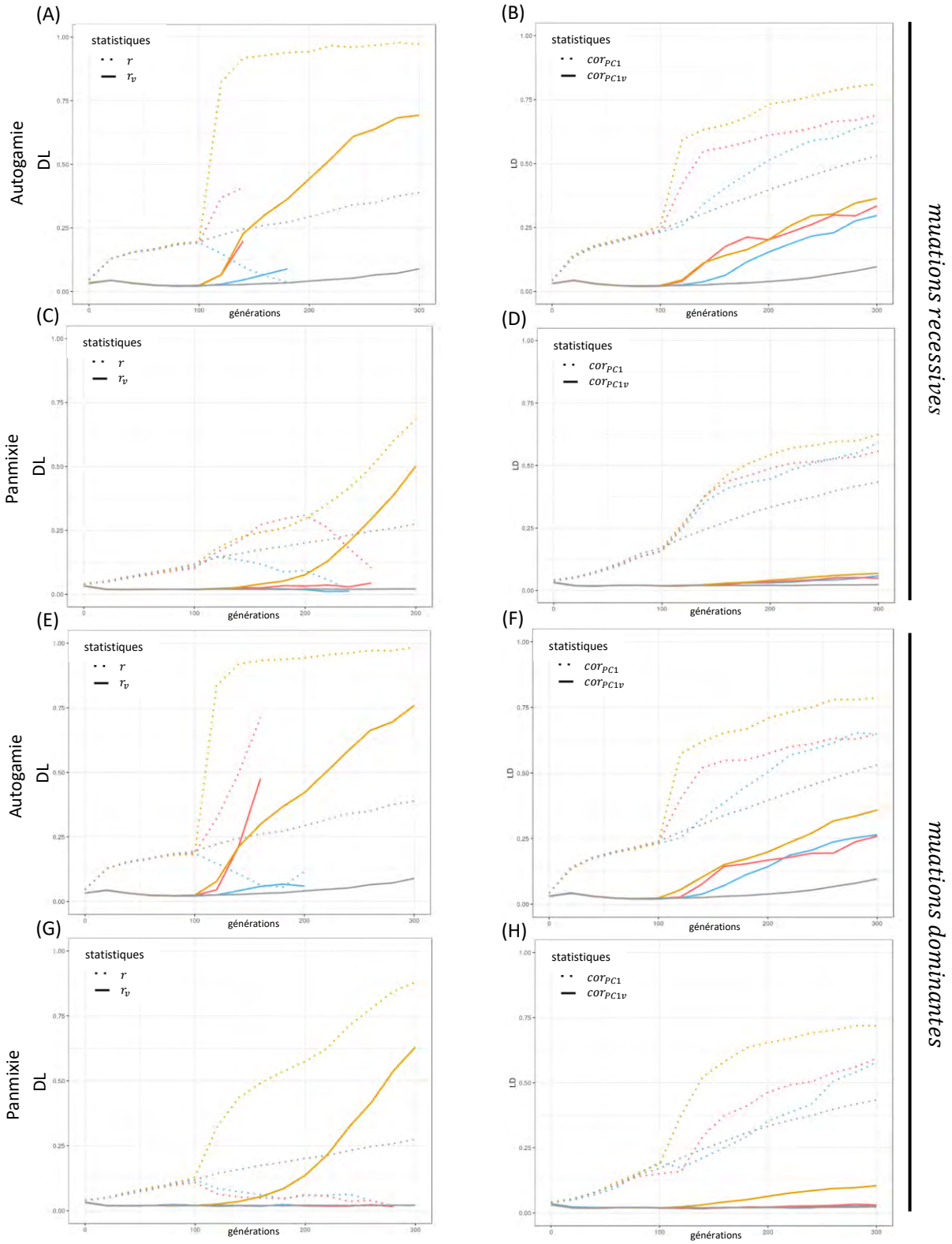
- La ligne2: la fonction `sim.loadPopulation()` qui permet de charger la population qui est extraite des dernières simulations à la génération 100.
- Lignes 21-22 fonction `pop.evolve()`et paramètre `postOps,` l'opérateur `sim.MaSelector()` appelé « `multi-alleles Selector` » permet de faire de la sélection sur différents allèles. Deux locus bi-alléliques dont les positions et les identifiants sont donnés dans la fonction `loci=goodLoci` sont les locus que nous désignons pour être cible de la sélection. Ces locus A et B portent les allèles ancestraux *A* et *B* et les allèles dérivés *a* et *b*. L'opérateur `sim.MaSelector()`attribue des *fitness* aux individus en fonction de leurs génotypes à ces deux locus dans l'ordre suivant: AABB, AABb, AAbb, AaBB, AaBb, Aabb, aaBB, aaBb, aabb. Sur l'exemple, seul les individus portant l'haplotype *ab* ont une *fitness* supérieur à 1: `fitness=[1, 1, 1, 1, 1, 1, 1, 1, 1+s+s]` avec `s` la valeur sélective qui est prédéfinie. Cet exemple correspond au modèle de sélection épistatique coadapté avec des mutations récessives. Enfin, la sélection est simulée séparément dans les deux sous-populations réelles de 250 individus: `subPops=[(0,0), (0,1)])` pour la première sous-population et `subPops=[(1,0), (1,1)])` pour la seconde sous-population.

Les 1000 simulations sont réalisées pour chaque combinaison de paramètres (reproduction – modèles de sélection et interaction entre les allèles), et en parallèle sur le cluster de calcul de la Genotoul à l'aide de la fonction *sarray.* Les commandes SimuPop qui ont été présentées sont intégrées dans un script python et un fichier bash permet d'exécuter simultanément les 1000 simulations en parallèle.

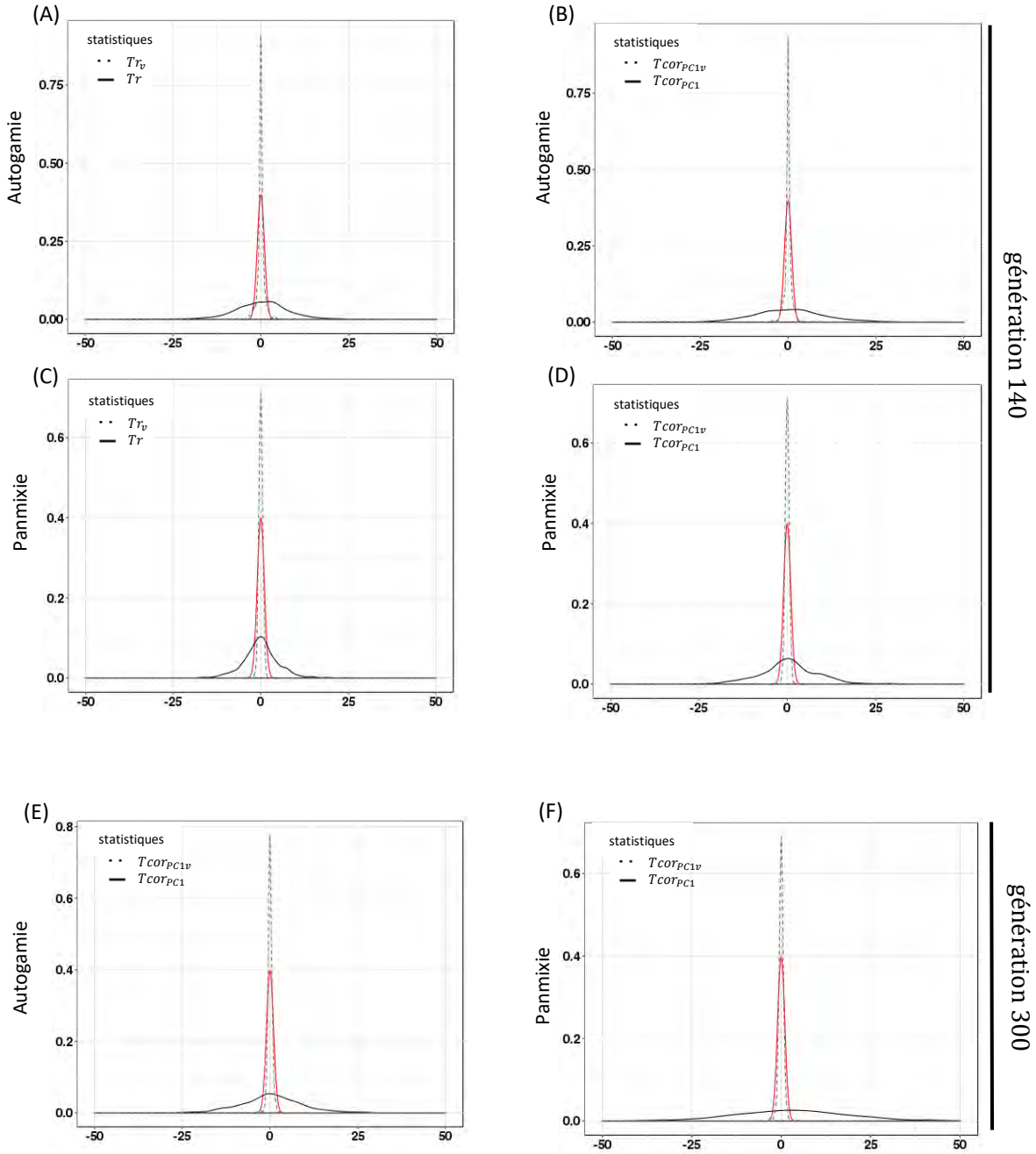
Annexe 3. Distribution des fréquences de l'allèle dérivé *a* du SNP soumis à la sélection sur le chromosome A aux générations t=0 et t=100. Données observées pour 1,000 simulations avec N=500 individus diploïdes. (A1, A2) système de reproduction autogame à 95% et (B1, B2) 100% panmixie. (A1, B1) montre la distribution des fréquences de l'allèle dérivé *a* au SNP du chromosome A à la génération zéro (A1, B1), et à la génération 100 (A2, B2) au moment où la sélection entre en jeu. A la génération 0, la médiane de la fréquence de l'allèle dérivé *a* sur 1000 simulations est de 0.43. A la génération 100, ces valeurs sont de 0,446 en autogamie et 0,435 en panmixie.



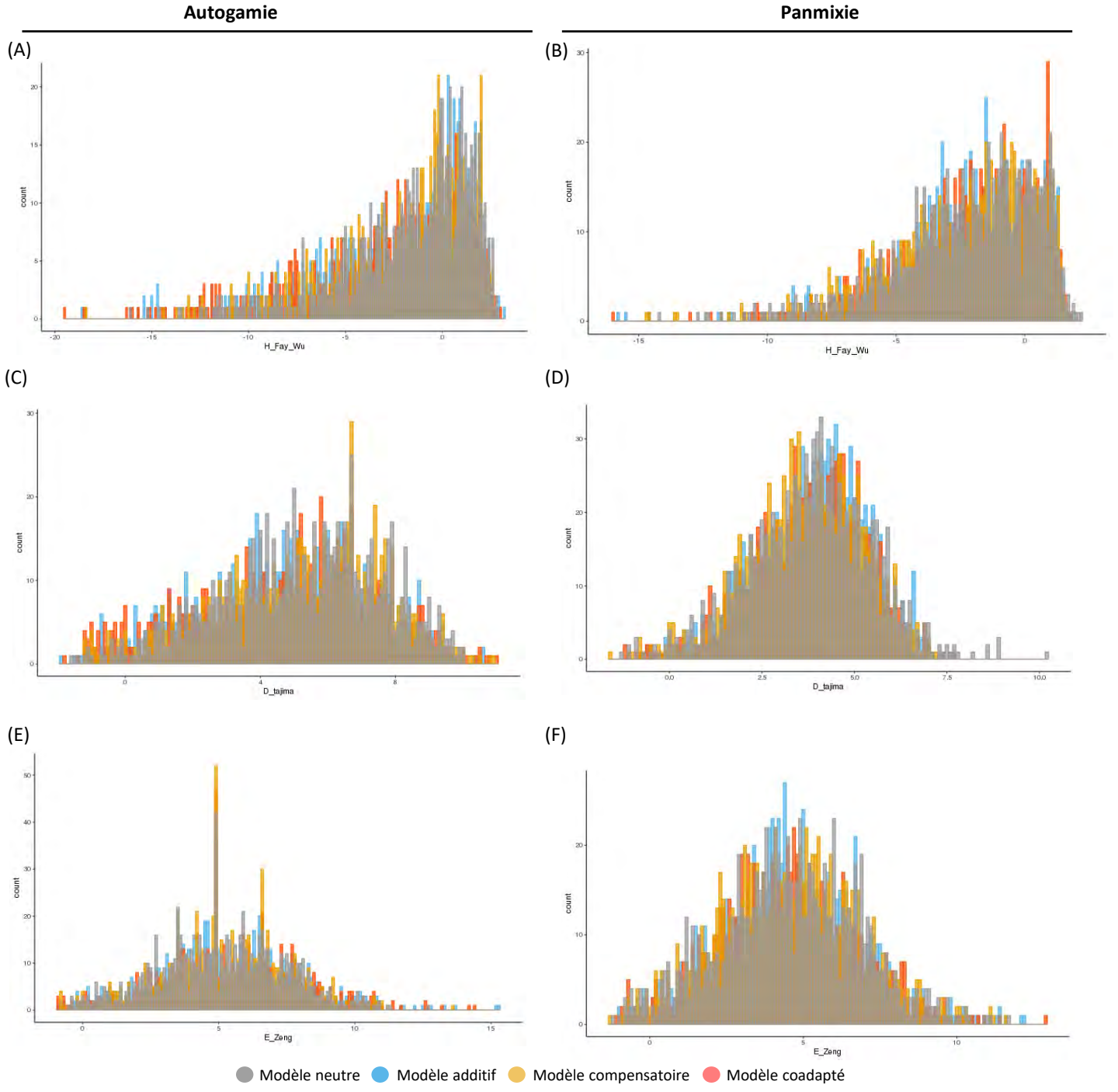
Annexe 4. Evolution du DL entre les locus simulés sous sélection épistatique, additive et sous neutralité pendant 300. Les mutations sous selection épistatique sont récessives (**A, B, C, D**) ou dominantes (**E, F, G, H**). En autogamie, le DL est calculé au niveau des SNP sous sélection avec les statistiques r et r_v (**A, E**) et sur des fenêtres génomiques avec les statistiques cor_{PC1} et cor_{PC1v} (**B, F**). En panmixie, le DL est calculé au niveau des SNP avec r et r_v (**C, G**) et sur des fenêtres génomiques avec cor_{PC1} et cor_{PC1v} (**D, H**).



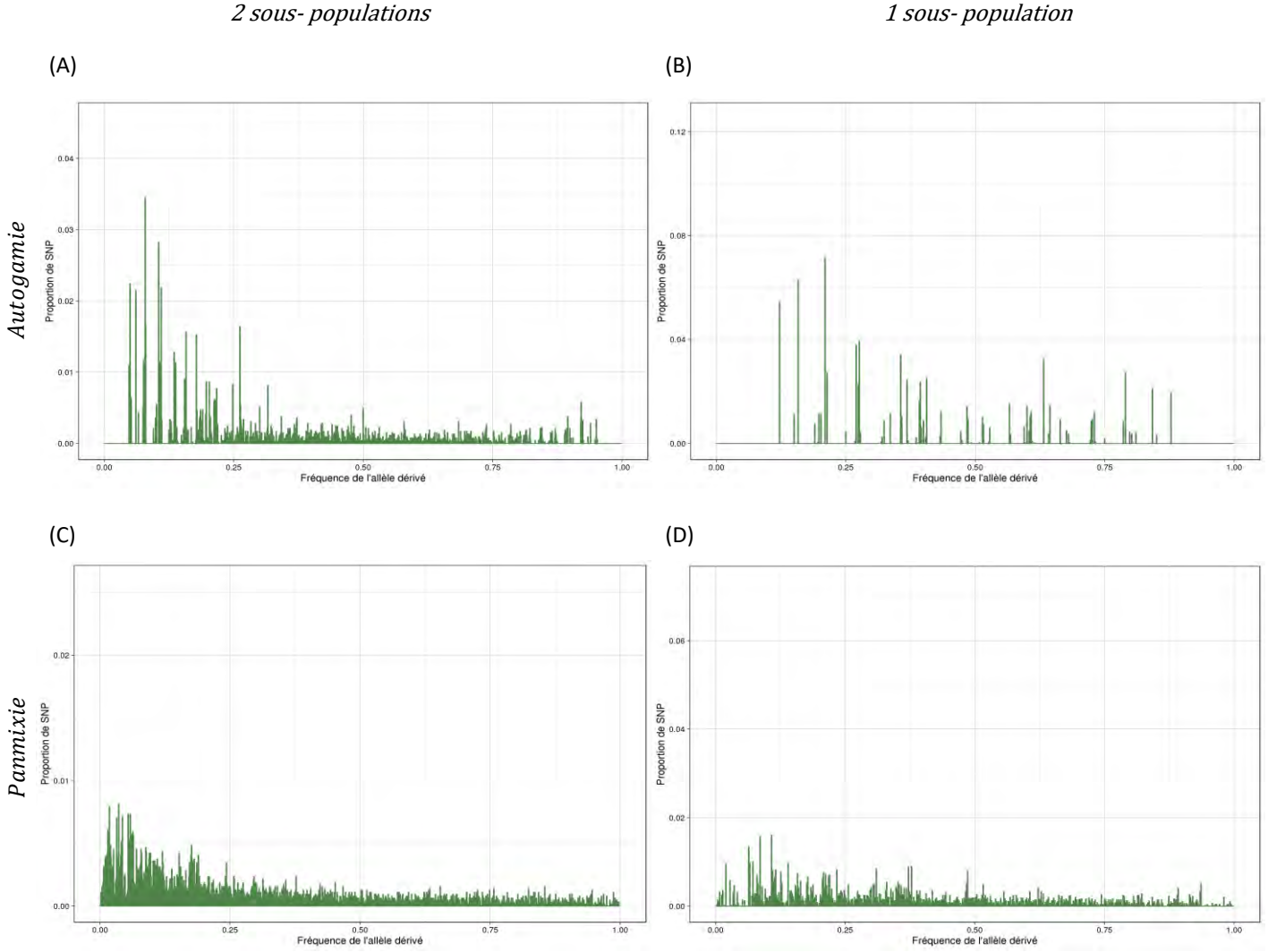
Annexe 5. Distribution des statistiques de corrélation et comparaison de *Tr* et Tr_v (respectivement $Tcor_{PC1}$ et $Tcor_{PC1v}$) avec la distribution théorique de Student $\tau_{(n-2)}$. Les figures représentent les distributions des statistiques *Tr* - Tr_v (A et C) et $Tcor_{PC1}$ - $Tcor_{PC1v}$ (B, D, E et F) en autogamie (A, B et E) et en panmixie (C, D et F). Les statistiques *T* sont calculées à la **génération 140** (A, B, C et D) ains qu'à la génération 300 (E et F) dans le modèle neutre et elles sont comparées à la distribution théorique de Student $\tau_{(n-2)}$ (trait rouge). Les courbes trait plein noir sont les distributions de *Tr* et $Tcor_{PC1}$ et les traits pointillés noirs sont les distributions de Tr_v et $Tcor_{PC1v}$. L'axe des abscisses correspond à la statistique *T* obtenue à partir des valeurs de corrélation (r, r_v, cor_{PC1}, cor_{PC1v}) et l'axe des Y correspond à la densité de probabilité.



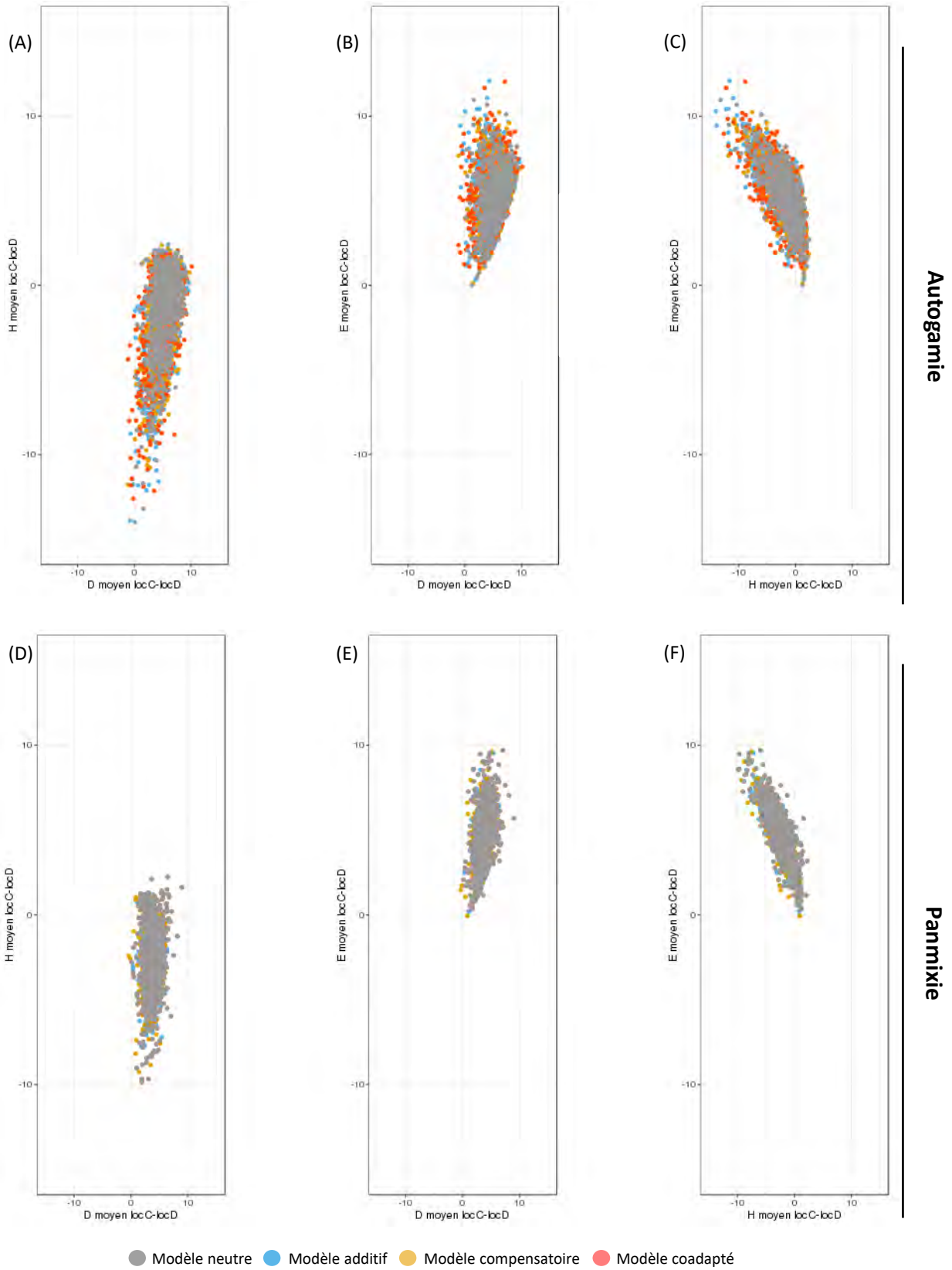
Annexe 6. Distribution des statistiques de neutralité calculées sur les locus simulés dans les différents modèles de sélection à la génération 140, dans les modèles de reproduction en autogamie et en panmixie, et sur les chromosomes C et D neutres. Les statistiques ***H*** de Fay & Wu, ***D*** de Tajima et ***E*** de Zeng sont calculées sur les fenêtres génomiques de 10kb situées au milieu des chromosomes C et D dans les quatre modèles de sélection; les modèles de sélection épistatiques compensatoire et coadapté, le modèle de sélection additif et le modèle neutre. Les figures (A, B) représentent les distributions de ***H*** de Fay & Wu, les figures (C, D) représentent les distributions du ***D*** de Tajima et les figures (E,F) représentent les distributions du ***E*** de Zeng. (A, C, E) sont extraites du modèle de reproduction en autogamie et (B, D, F) du modèle panmictique. L'ensemble des ces statistiques sont calculées à l'échelle d'une sous-population simulée de 250 individus. L'axe des abscisses correspond aux valeurs de statistiques calculées sur les données simulées et l'axe des ordonnées correspond à la courbe de densité. Les mutations sous sélection sont codominantes.



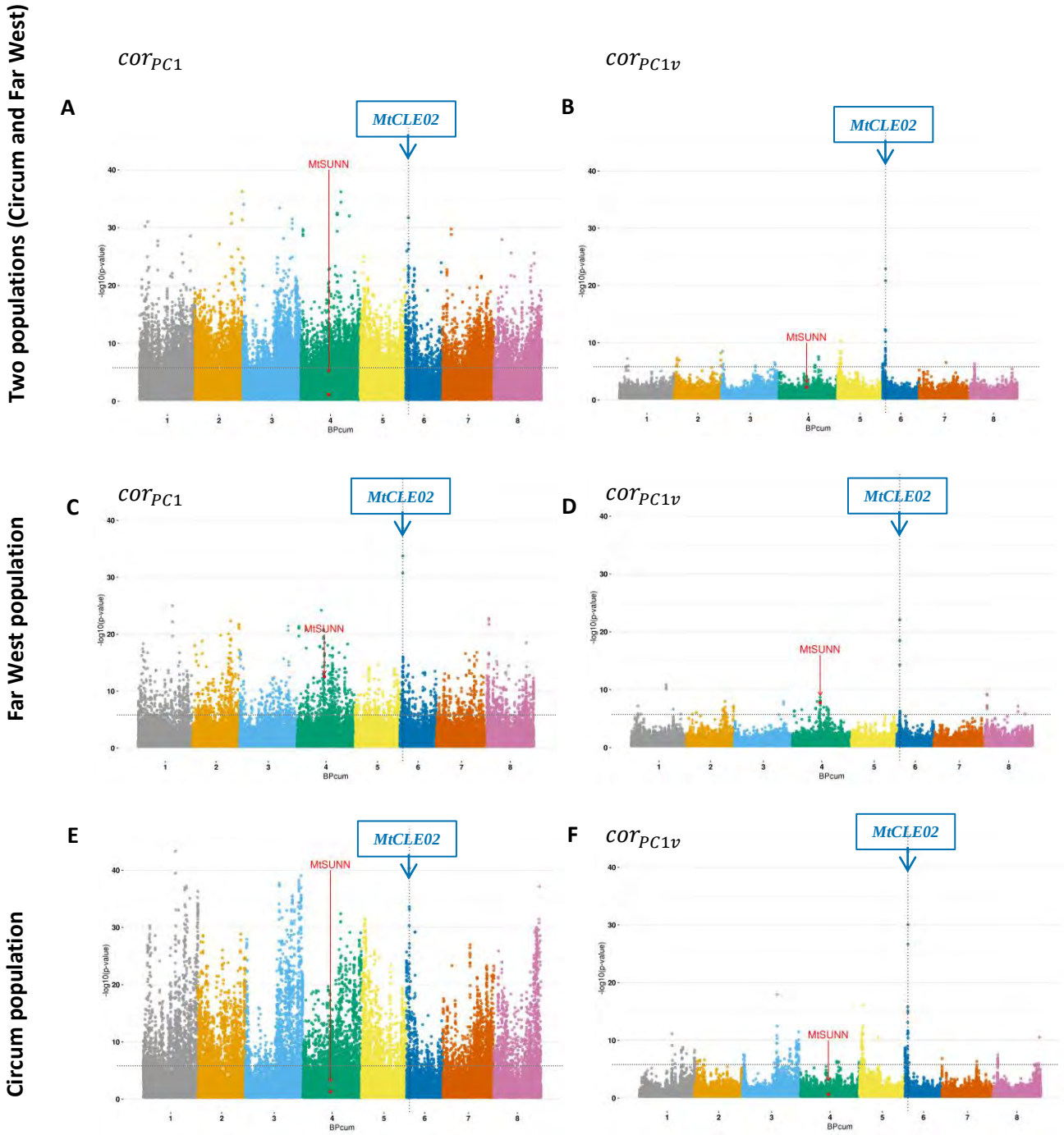
Annexe 7. Spectre des fréquences alléliques obtenu sur un chromosome simulé à la génération 140, sous neutralité. L'axe des X représente les classes de fréquence *i* de l'allèle dérivé des SNP simulés. L'axe des Y indique la proportion de SNP appartenant à chaque classe de fréquence de l'allèle dérivé. Les figures (A, B) sont extraites du modèle en autogamie, les figures (C, D) sont extraites du modèle en panmixie. Les figures (A, C) intègrent les données des deux sous-populations (500 individus diploïdes - $1 \leq i \leq 1000$), et les figures (B, D) se basent sur les données d'une sous-population (250 individus diploïdes - $1 \leq i \leq 500$). Les spectres de fréquences sont réalisés sur l'un des chromosomes neutres (chromosome C ou D).



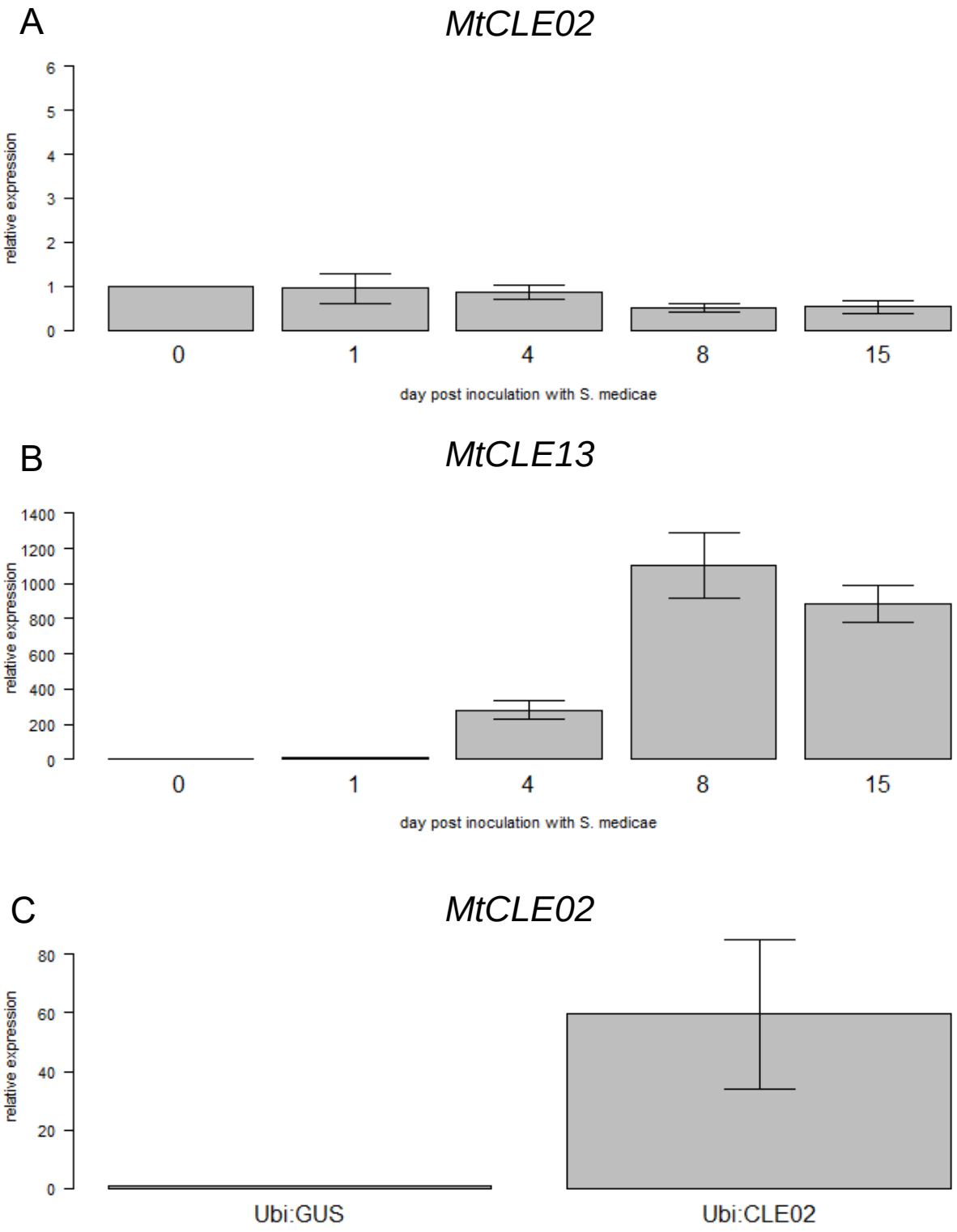
Annexe 8. Distributions conjointes des statistiques de neutralité: $DH_{\overline{AB}}$, $DE_{\overline{AB}}$ et $HE_{\overline{AB}}$ calculées entre les chromosomes C et D neutres. Les distributions conjointes des statistiques sont obtenues à partir des valeurs moyennes aux deux locus C et D pour chaque simulation, pour chaque modèles de sélection et pour chacune des statistiques. Les Figures (A, B et C) représentent les distributions en autogamie et les figures (D,E,F) les distributions en panmixie.



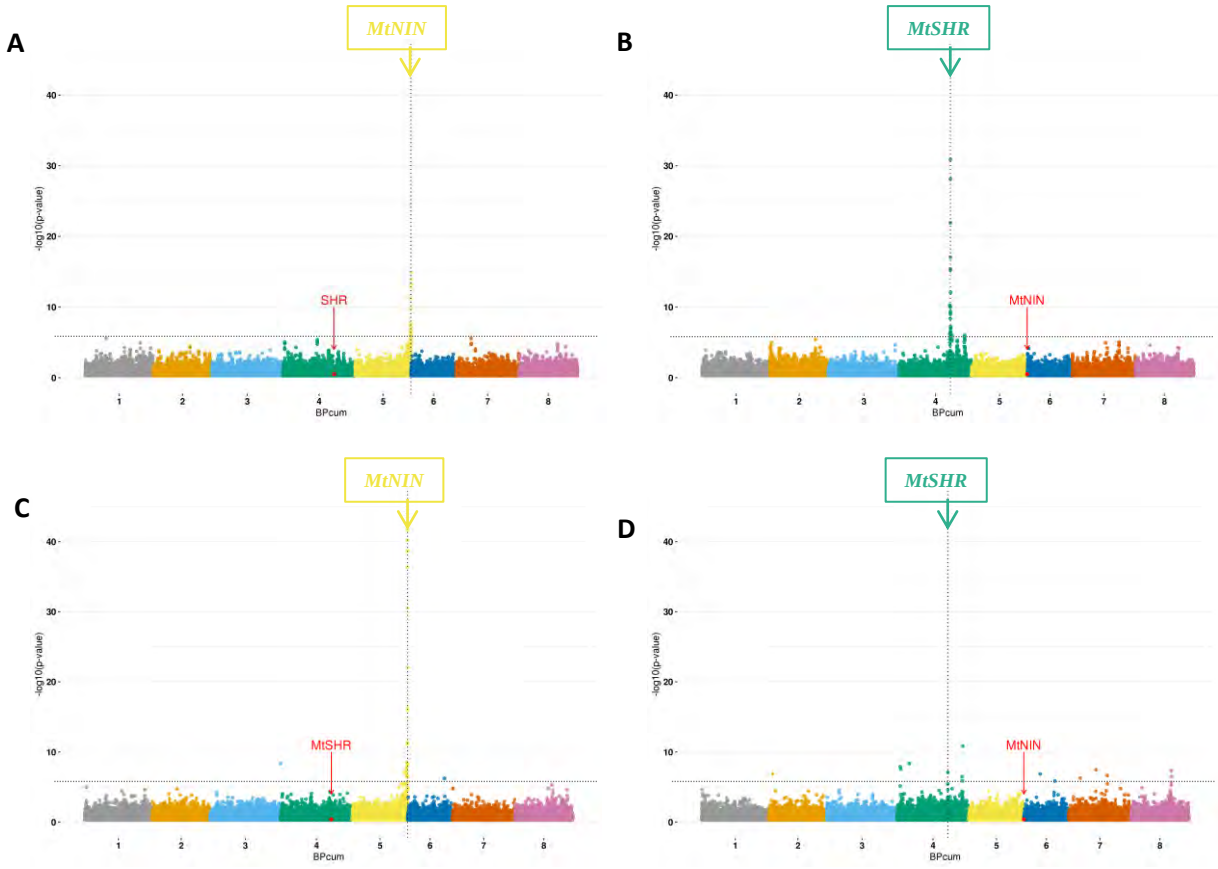
Annexe 9. Distribution du DL entre le gène appât *MtCLE02* et tous les autres gènes du génome de *M. truncatula*. Le DL entre le gène *MtCLE02* et les autres gènes de *M. truncatula* est calculé dans la population entière (**A** - **B**), dans la population Far-West (**C** - **D**) et dans la population Circum (**E** - **F**). Les p-valeurs des tests de corrélation sont calculées à partir des statistiques $Tcor_{PC1}$ (**A** – **C** – **E**) et $Tcor_{PC1v}$ **(B** - **D** - **F**) qui prend en compte la matrice kinship. L'axe des x correspond aux positions des gènes répartis sur les 8 chromosomes de *Medicago truncatula*. L'axe des y est le -$\log_{10}$(p-valeur) des tests de corrélation.



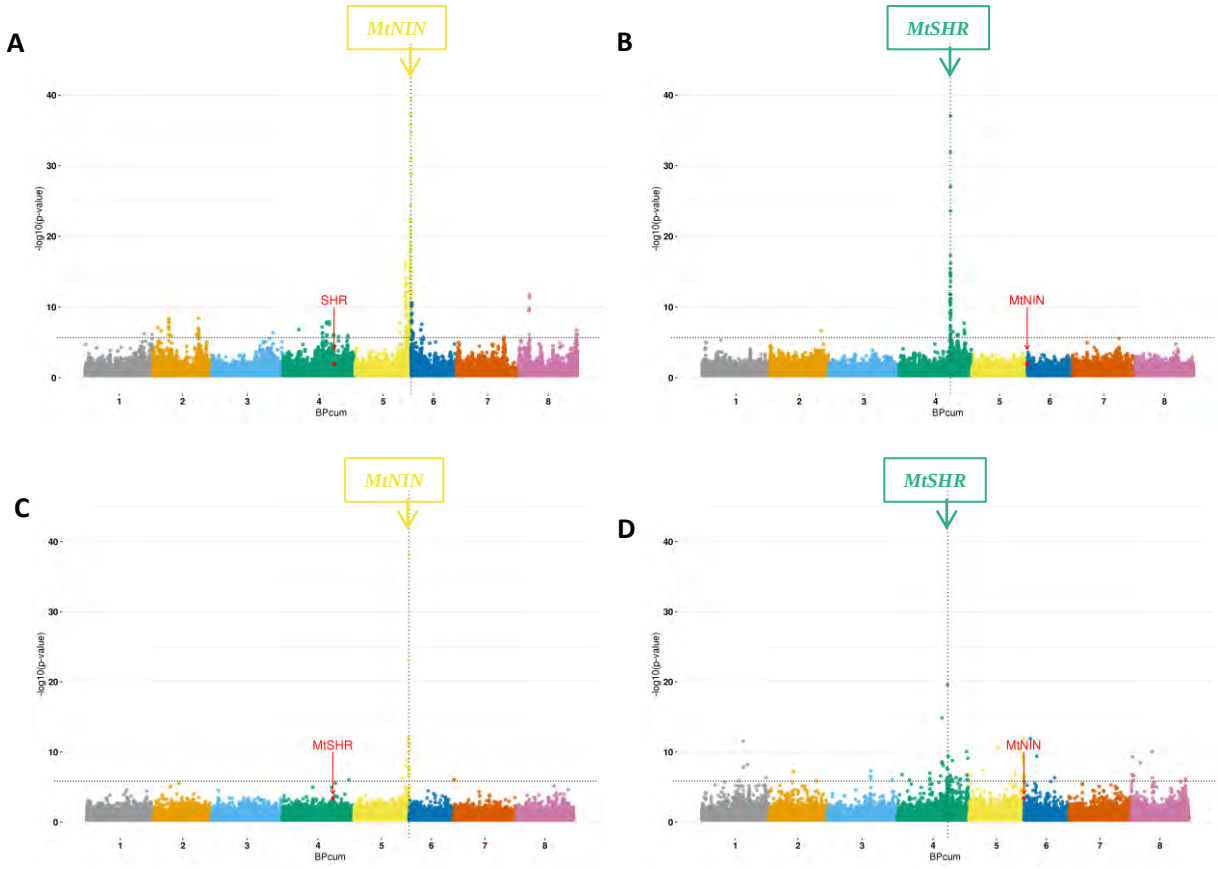
Annexe 10. Expression of *MtCLE02* in a nodulation kinetic and in overexpressing roots. (A) Real time RT-PCR analysis of *MtCLE02* expression in a nodulation kinetic, from 0 (non-inoculated roots), 1, 4, 8 or 15 days post rhizobium inoculation (dpi). 1 and 4 dpi corresponds to roots and 8 and 15 dpi to nodules. **(B)**. Real time RT-PCR analysis of *MtCLE13* expression in the same conditions as described in **A**. **(C)** Real time RT-PCR analysis of *MtCLE02* expression in roots overexpressing *MtCLE02* (Ubi:CLE02) or the *GUS* control (Ubi:GUS). Two independent transgenic roots were analyzed for each genotype.



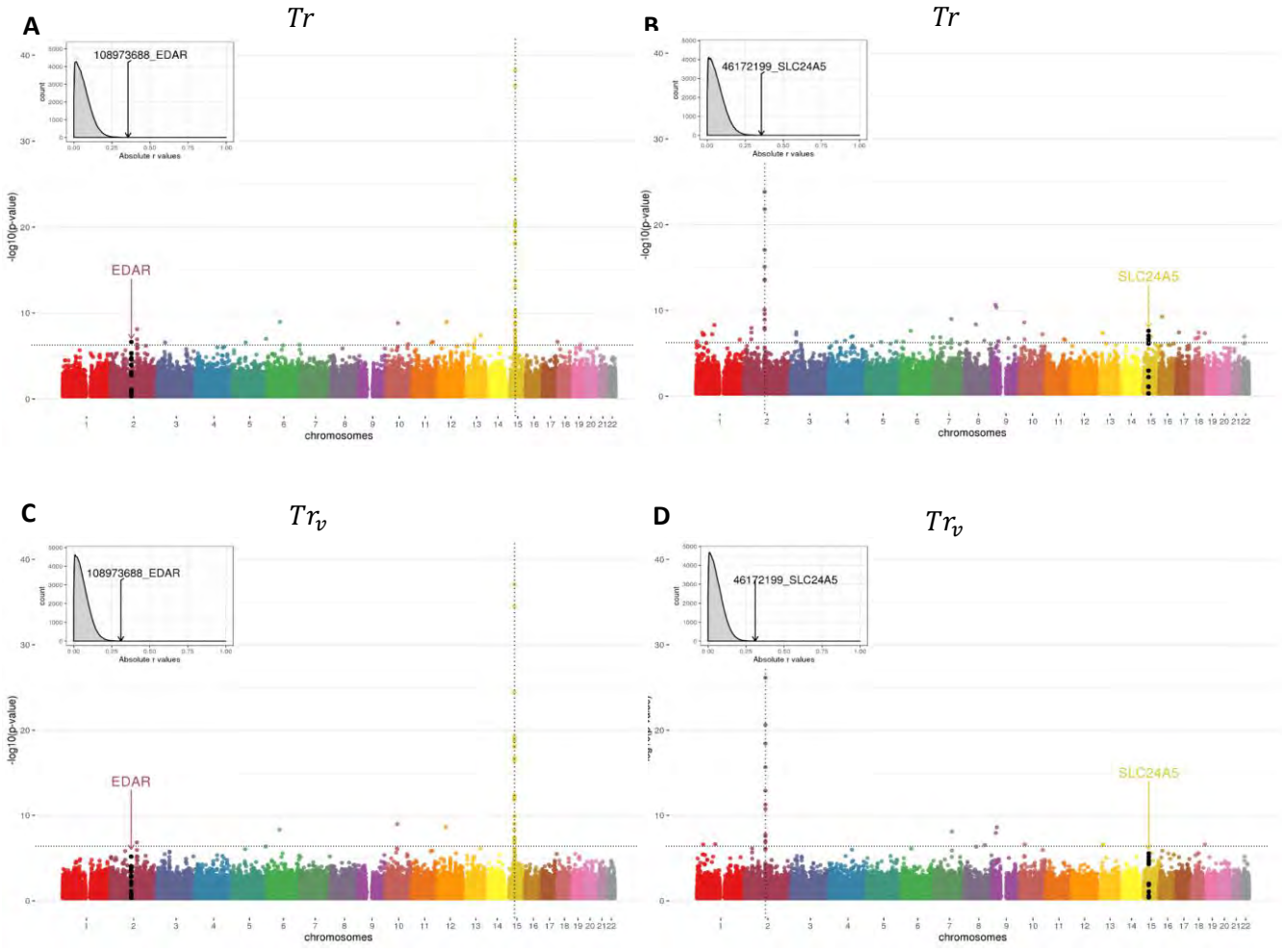
Annexe 11. Distribution du DL entre les gènes appâts *MtNIN* et *MtSHR* et tous les autres gènes du génome de *M. truncatula* dans la population Circum. Le DL entre les gènes *MtNIN* et *MtSHR* et tous les gènes de *M. truncatula* est calculé respectivement à partir de fenêtres génomiques de 10kb centrées sur chaque gène (**A - B**), ou à partir de fenêtres génomiques comprenant uniquement les SNP dans les gènes (séquence génique) (**C - D**). Les p-valeurs des tests de corrélation sont calculées à partir de la statistique $Tcor_{PC1v}$ qui prend en compte la matrice kinship. L'axe des x correspond aux positions des gènes répartis sur les 8 chromosomes de *Medicago truncatula*. L'axe des y est le -$\log_{10}$(p-valeur) des tests de corrélation.



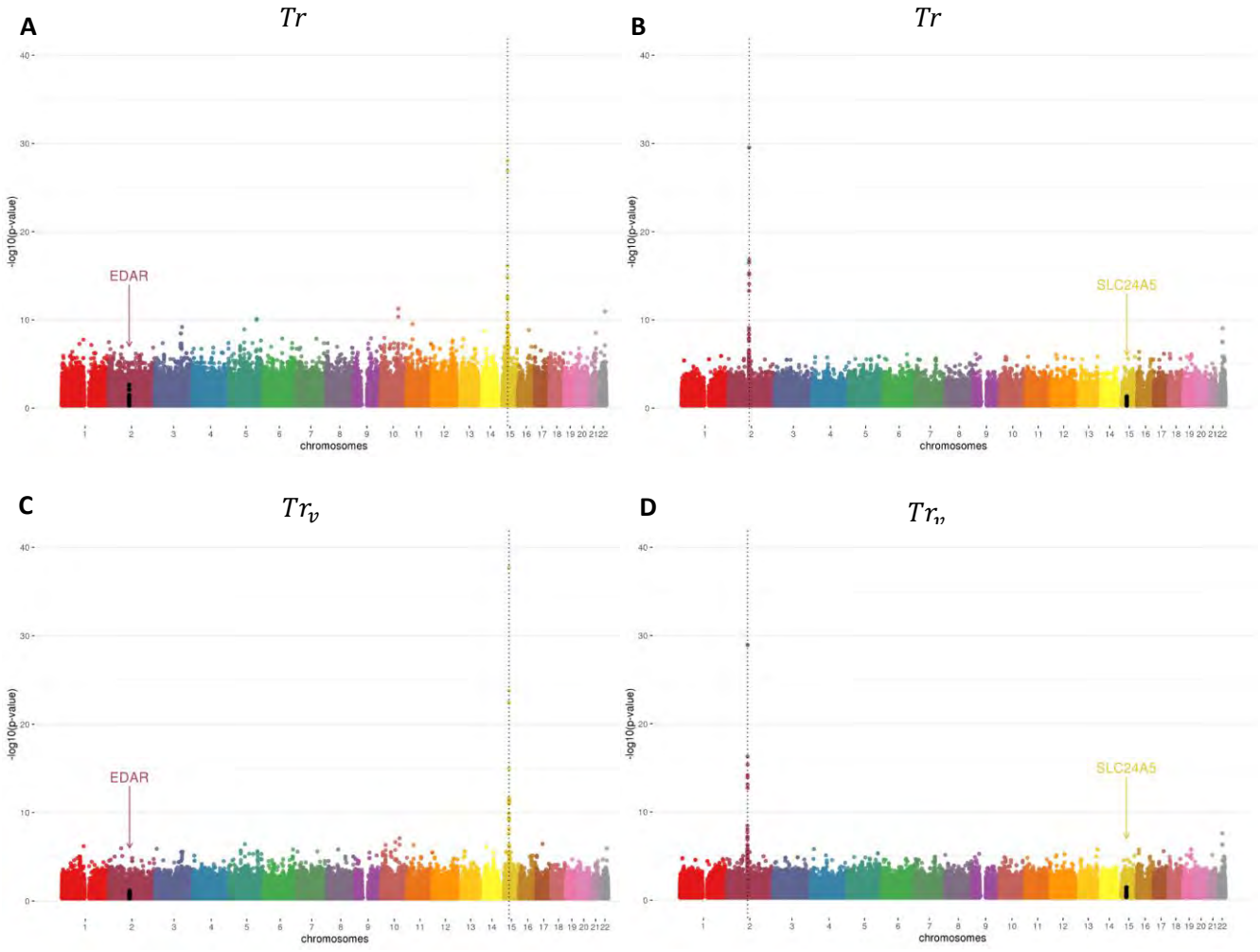
Annexe 12. Distribution du DL entre les gènes appâts *MtNIN* et *MtSHR* et tous les autres gènes du génome de *M. truncatula* dans la population entière. Le DL entre les gènes *MtNIN* et *MtSHR* et tous les gènes de *M. truncatula* est calculé respectivement à partir de fenêtres génomiques de 10kb centrées sur chaque gène (**A - B**), ou à partir de fenêtres génomiques comprenant uniquement les SNP dans les gènes (séquence génique) (**C - D**). Les p-valeurs des tests de corrélation sont calculées à partir de la statistique $Tcor_{PC1v}$ qui prend en compte la matrice kinship. L'axe des x correspond aux positions des gènes répartis sur les 8 chromosomes de *Medicago truncatula*. L'axe des y est le $-\log_{10}$(p-valeur) des tests de corrélation.



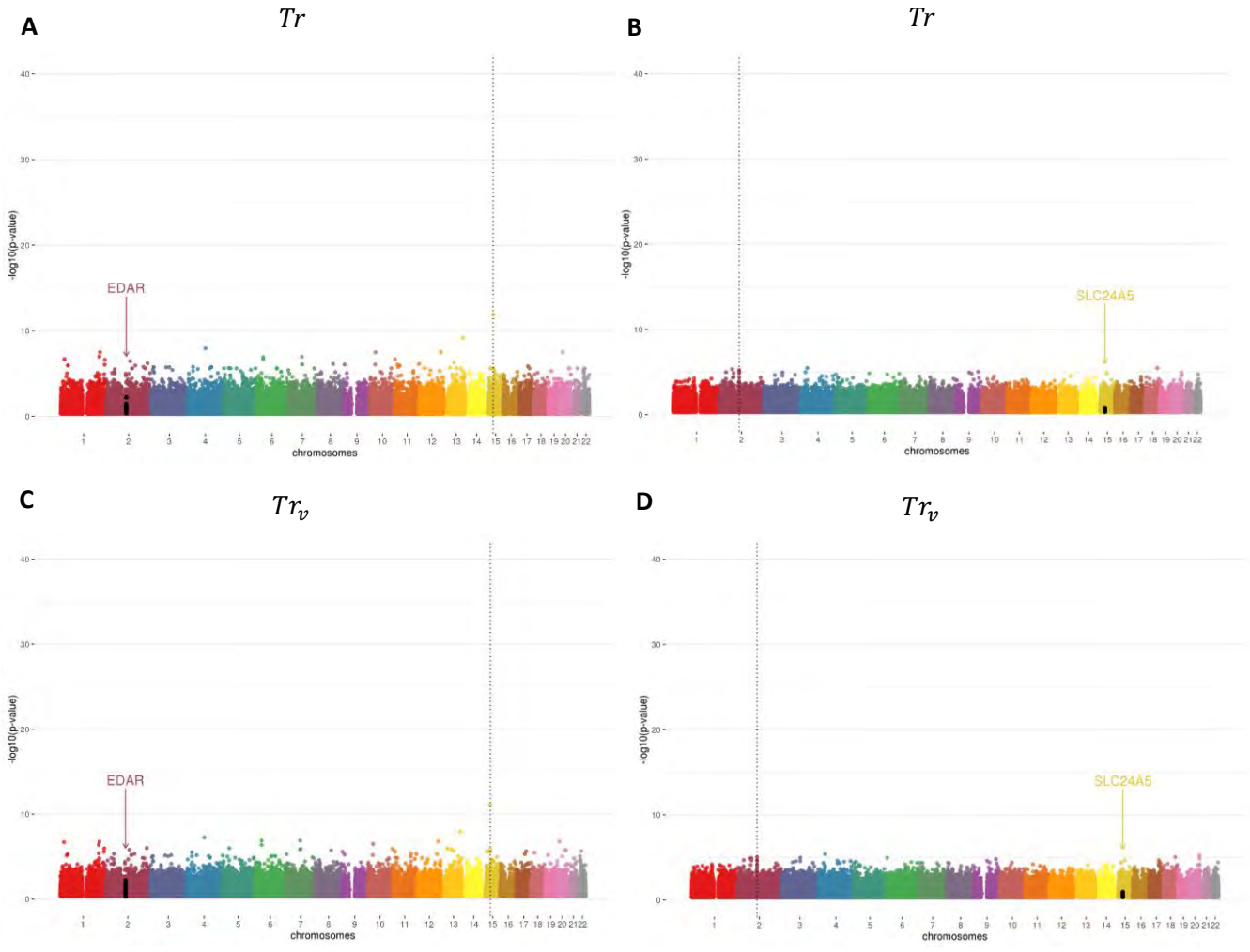
Annexe 13. Distribution du DL entre les SNP appâts des gènes *SLC24A5* et *EDAR* et tous les autres SNP des données humaines HGDP-CEPH de la population de l'Asie centrale du Sud (n=192). Le DL entre les SNP appâts 15_46172199 (*SLC24A5*) et 2_108973688 (*EDAR*) et tous les autres SNP du génome est calculé avec les statistiques T_r (**A, B**) ou T_{r_v} (**C, D**), qui prend en compte la matrice d'apparentement. L'axe des x correspond aux positions des SNP répartis sur les 22 autosomes humains, chaque point correspond à un SNP et les point noirs montrent les SNP des gènes candidats qui sont sous sélection épistatique avec le SNP appât de chaque figure (ligne verticale pointillée). L'axe des y est le -$\log_{10}$(p-value) du test de corrélation. Les graphiques en haut à gauche de chaque figure montrent les distributions du DL entre chaque SNP appât et tous les autres SNP du génome (à l'exception des SNP situés dans une fenêtre de 50Kb entourant les SNP appâts). La significativité de la corrélation entre le SNP appât de *SLC24A5* et le SNP cible de *EDAR* (respectivement le SNP appât de *EDAR* et le SNP cible de *SLC24A5*) est représentée par une flèche.



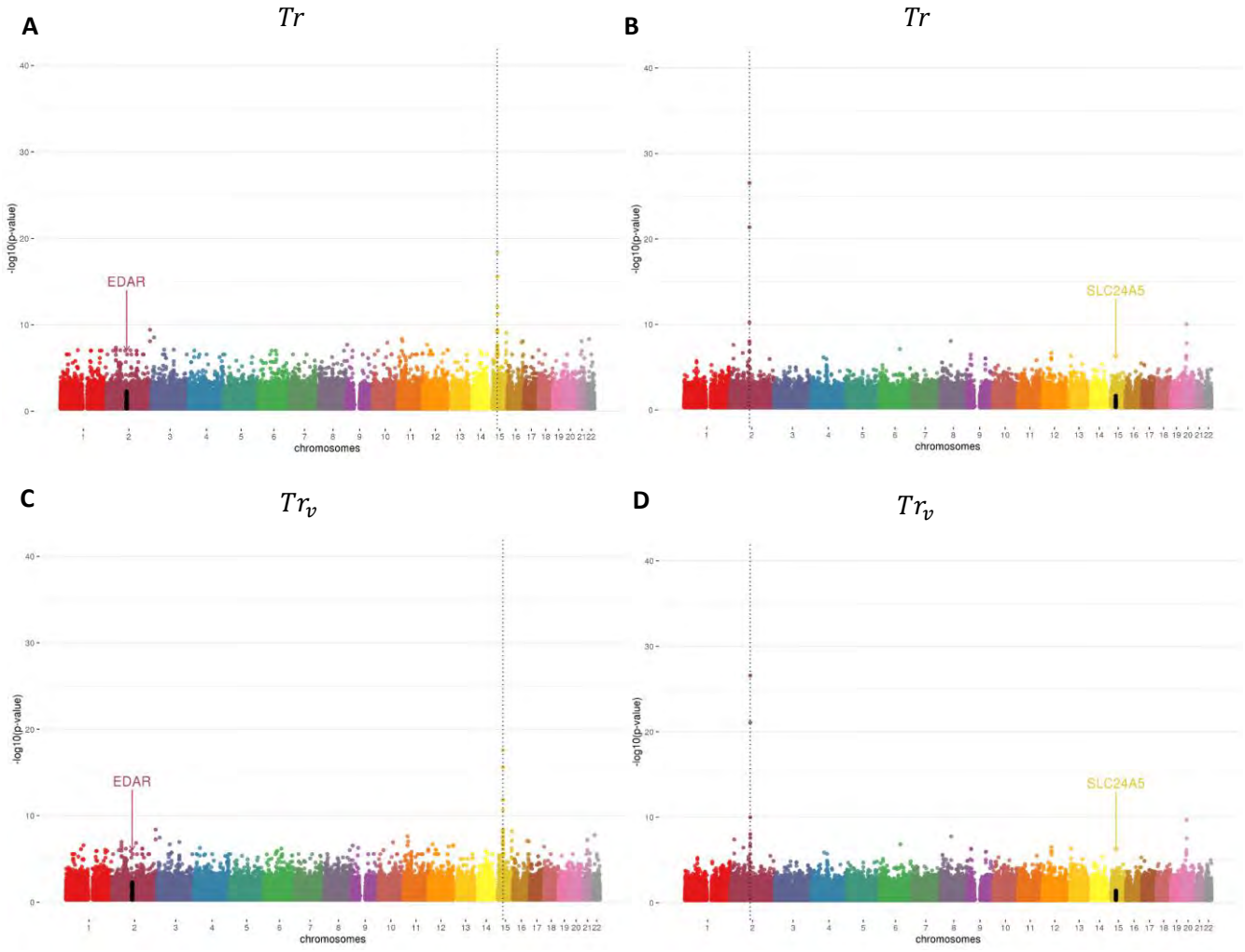
Annexe 14. Distribution du DL entre les SNP appâts des gènes *SLC24A5* et *EDAR* et tous les autres SNP des données humaines HGDP-CEPH de la population de l'Asie de l'Est (n=242). Le DL entre les SNP appâts 15_46172199 (*SLC24A5*) et 2_108973688 (*EDAR*) et tous les autres SNP du génome est calculé avec les statistiques T_r (**A, B**) ou T_{r_v} (**C, D**), qui prend en compte la matrice d'apparentement. L'axe des x correspond aux positions des SNP répartis sur les 22 autosomes humains, chaque point correspond à un SNP et les point noirs montrent les SNP des gènes candidats qui sont sous sélection épistatique avec le SNP appât de chaque figure (ligne verticale pointillée). L'axe des y est le -$\log_{10}$(p-value) du test de corrélation. Les graphiques en haut à gauche de chaque figure montrent les distributions du DL entre chaque SNP appât et tous les autres SNP du génome (à l'exception des SNP situés dans une fenêtre de 50Kb entourant les SNP appâts). La significativité de la corrélation entre le SNP appât de *SLC24A5* et le SNP cible de *EDAR* (respectivement le SNP appât de *EDAR* et le SNP cible de *SLC24A5*) est représentée par une flèche.



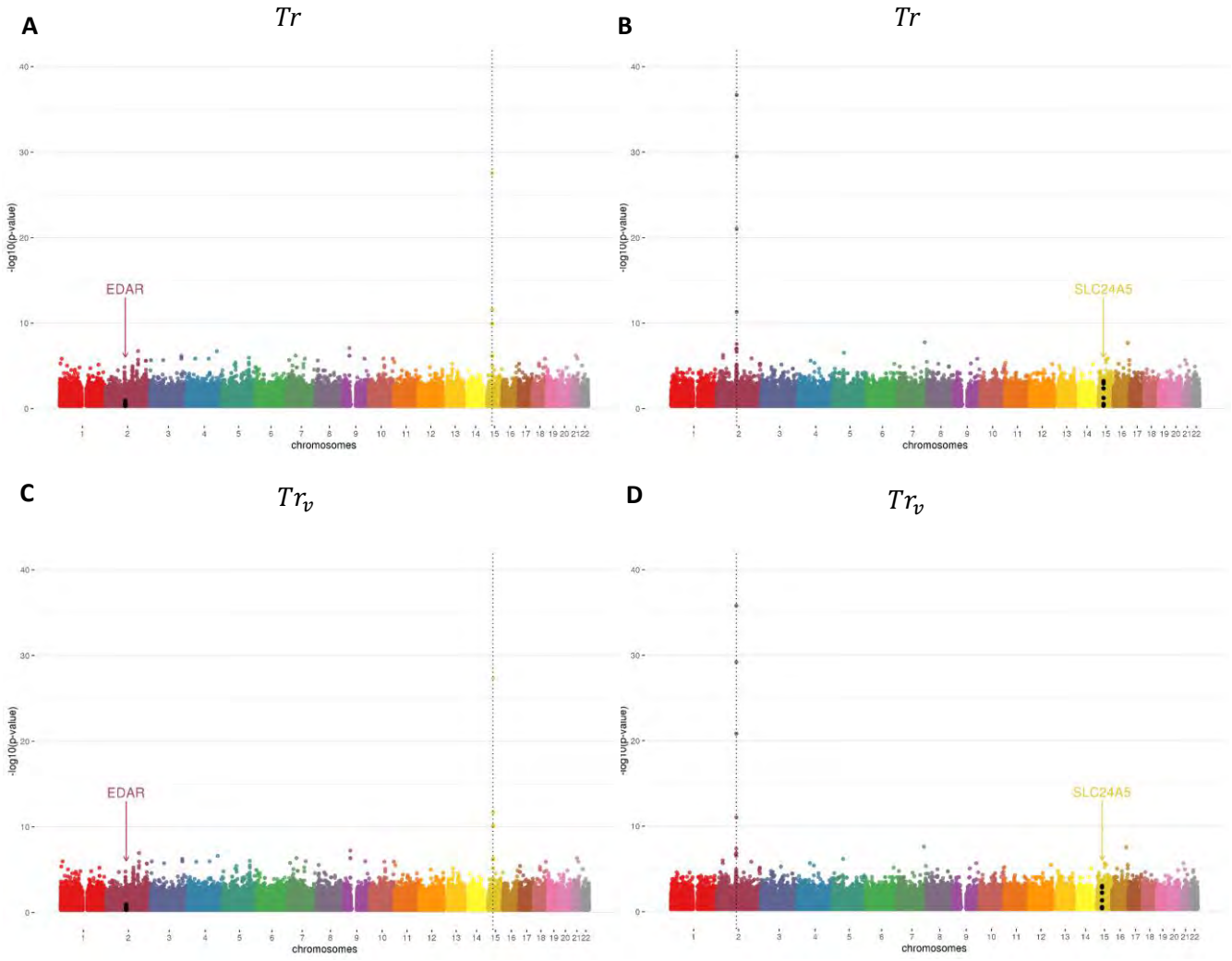
Annexe 15. Distribution du DL entre les SNP appâts des gènes *SLC24A5* et *EDAR* et tous les autres SNP des données humaines HGDP-CEPH de la population de l'Afrique Subsaharienne (n=105). Le DL entre les SNP appâts 15_46172199 (*SLC24A5*) et 2_108973688 (*EDAR*) et tous les autres SNP du génome est calculé avec les statistiques T_r (**A, B**) ou T_{r_v} (**C, D**), qui prend en compte la matrice d'apparentement. L'axe des x correspond aux positions des SNP répartis sur les 22 autosomes humains, chaque point correspond à un SNP et les point noirs montrent les SNP des gènes candidats qui sont sous sélection épistatique avec le SNP appât de chaque figure (ligne verticale pointillée). L'axe des y est le -$\log_{10}$(p-value) du test de corrélation. Les graphiques en haut à gauche de chaque figure montrent les distributions du DL entre chaque SNP appât et tous les autres SNP du génome (à l'exception des SNP situés dans une fenêtre de 50Kb entourant les SNP appâts). La significativité de la corrélation entre le SNP appât de *SLC24A5* et le SNP cible de *EDAR* (respectivement le SNP appât de *EDAR* et le SNP cible de *SLC24A5*) est représentée par une flèche.



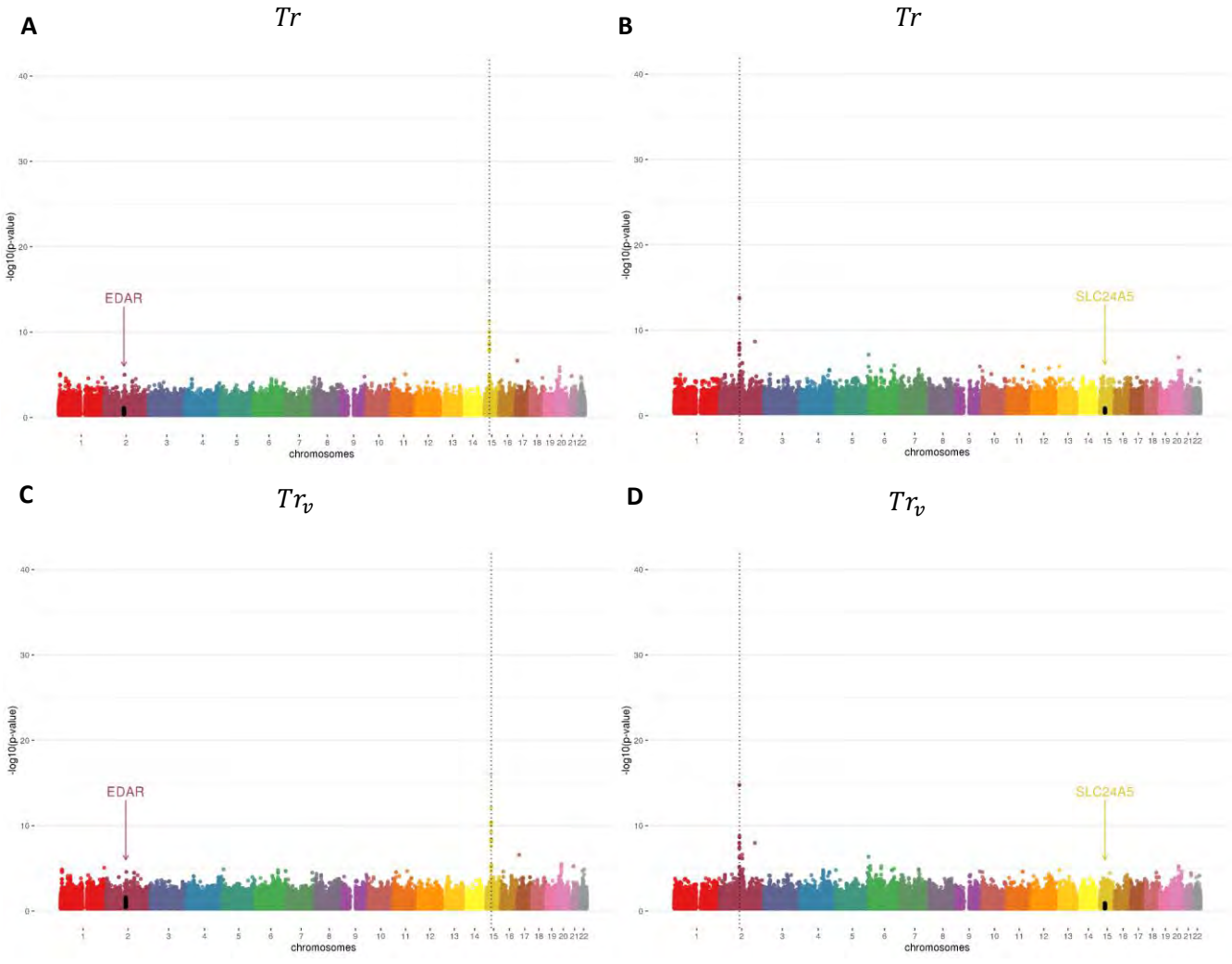
Annexe 16. Distribution du DL entre les SNP appâts des gènes *SLC24A5* et *EDAR* et tous les autres SNP des données humaines HGDP-CEPH de la population du Moyen Orient (n=134). Le DL entre les SNP appâts 15_46172199 (*SLC24A5*) et 2_108973688 (*EDAR*) et tous les autres SNP du génome est calculé avec les statistiques T_r (**A, B**) ou T_{r_v} (**C, D**), qui prend en compte la matrice d'apparentement. L'axe des x correspond aux positions des SNP répartis sur les 22 autosomes humains, chaque point correspond à un SNP et les point noirs montrent les SNP des gènes candidats qui sont sous sélection épistatique avec le SNP appât de chaque figure (ligne verticale pointillée). L'axe des y est le -log$_{10}$(p-value) du test de corrélation. Les graphiques en haut à gauche de chaque figure montrent les distributions du DL entre chaque SNP appât et tous les autres SNP du génome (à l'exception des SNP situés dans une fenêtre de 50Kb entourant les SNP appâts). La significativité de la corrélation entre le SNP appât de *SLC24A5* et le SNP cible de *EDAR* (respectivement le SNP appât de *EDAR* et le SNP cible de *SLC24A5*) est représentée par une flèche.



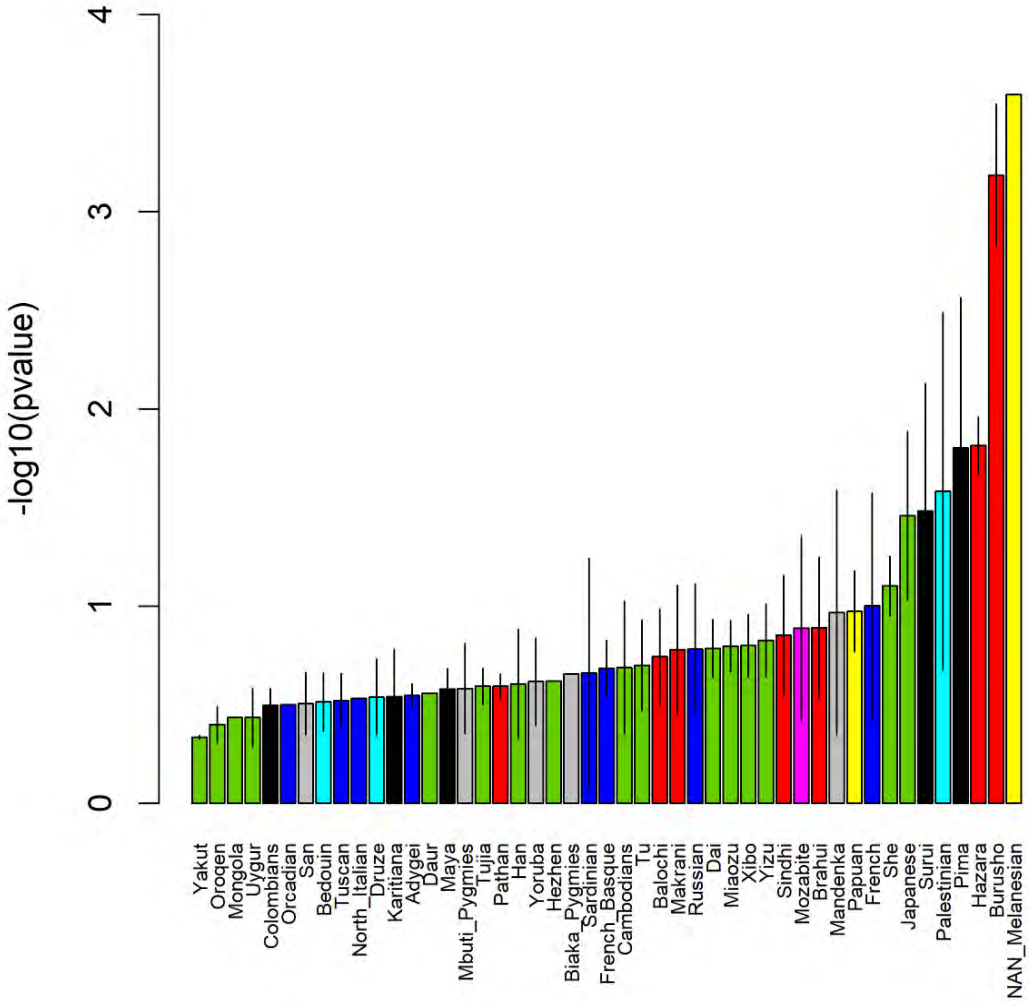
Annexe 17. Distribution du DL entre les SNP appâts des gènes *SLC24A5* et *EDAR* et tous les autres SNP des données humaines HGDP-CEPH de la population Européenne (n=158). Le DL entre les SNP appâts 15_46172199 (*SLC24A5*) et 2_108973688 (*EDAR*) et tous les autres SNP du génome est calculé avec les statistiques T_r (**A, B**) ou T_{r_v} (**C, D**), qui prend en compte la matrice d'apparentement. L'axe des x correspond aux positions des SNP répartis sur les 22 autosomes humains, chaque point correspond à un SNP et les point noirs montrent les SNP des gènes candidats qui sont sous sélection épistatique avec le SNP appât de chaque figure (ligne verticale pointillée). L'axe des y est le -$\log_{10}$(p-value) du test de corrélation. Les graphiques en haut à gauche de chaque figure montrent les distributions du DL entre chaque SNP appât et tous les autres SNP du génome (à l'exception des SNP situés dans une fenêtre de 50Kb entourant les SNP appâts). La significativité de la corrélation entre le SNP appât de *SLC24A5* et le SNP cible de *EDAR* (respectivement le SNP appât de *EDAR* et le SNP cible de *SLC24A5*) est représentée par une flèche.



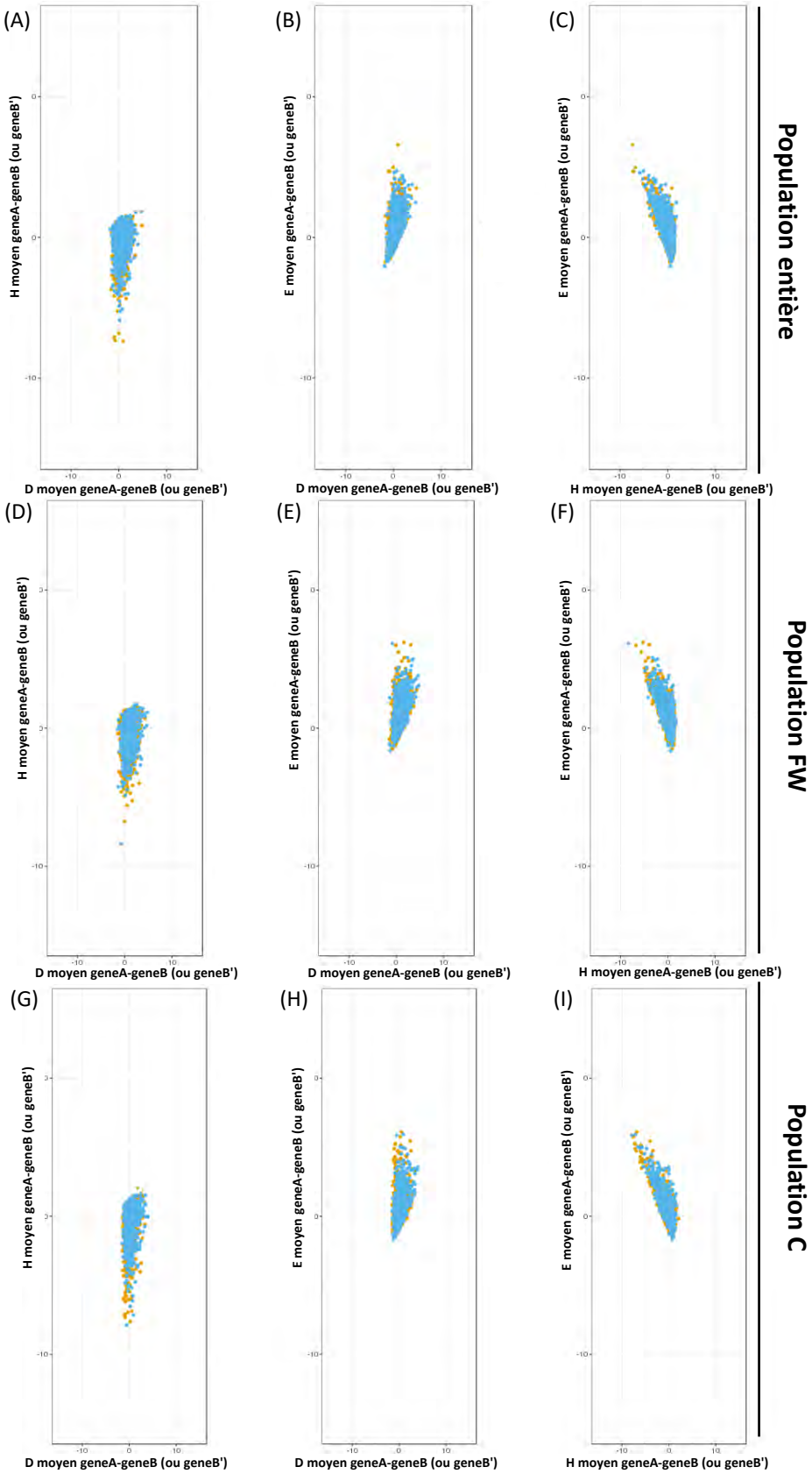
Annexe 18. Distribution du DL entre les SNP appâts des gènes *SLC24A5* et *EDAR* et tous les autres SNP des données humaines HGDP-CEPH de la population Américaine (n=64). Le DL entre les SNP appâts 15_46172199 (*SLC24A5*) et 2_108973688 (*EDAR*) et tous les autres SNP du génome est calculé avec les statistiques T_r (**A, B**) ou T_{r_v} (**C, D**), qui prend en compte la matrice d'apparentement. L'axe des x correspond aux positions des SNP répartis sur les 22 autosomes humains, chaque point correspond à un SNP et les point noirs montrent les SNP des gènes candidats qui sont sous sélection épistatique avec le SNP appât de chaque figure (ligne verticale pointillée). L'axe des y est le -$\log_{10}$(p-value) du test de corrélation. Les graphiques en haut à gauche de chaque figure montrent les distributions du DL entre chaque SNP appât et tous les autres SNP du génome (à l'exception des SNP situés dans une fenêtre de 50Kb entourant les SNP appâts). La significativité de la corrélation entre le SNP appât de *SLC24A5* et le SNP cible de *EDAR* (respectivement le SNP appât de *EDAR* et le SNP cible de *SLC24A5*) est représentée par une flèche.



Annexe 19. Distribution du DL entre les SNPs des gènes *SLC24A5* et *EDAR* dans les populations humaines.
La moyenne et l'écart-type des p-valeurs basées sur la statistique T_r entre les SNP de *SLC24A5* (SNP 15_46172199, SNP 15_46179457) et *EDAR* (SNP 2_108962124, SNP 2_108973688, SNP 2_108982808) sont représentés pour chaque population ou groupe ethnique.

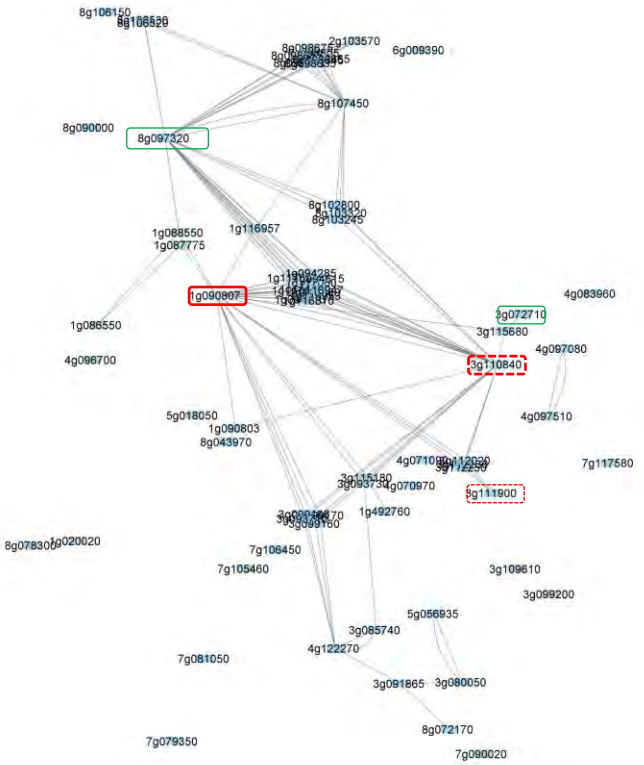


Annexe 20: Distributions conjointes des statistiques de neutralité $DH_{\overline{AB}}$, $DE_{\overline{AB}}$ et $HE_{\overline{AB}}$ calculées sur un ensemble de gènes sous sélection épistatique et sur un ensemble de gènes échantillonnés aléatoirement chez *Medicago truncatula*, dans la population entière et dans les sous-populations Far-West et Circum. Les distributions conjointes des statistiques sont obtenues à partir des valeurs moyennes entre les paires de gènes A-B (jaune), et des paires de gènes A-B' (bleu). Les Figures (A, B et C) représentent les distributions en dans la population entière, les figures (D,E,F) les distributions en population FW et les figures (G,H,I) les distributions en population C.



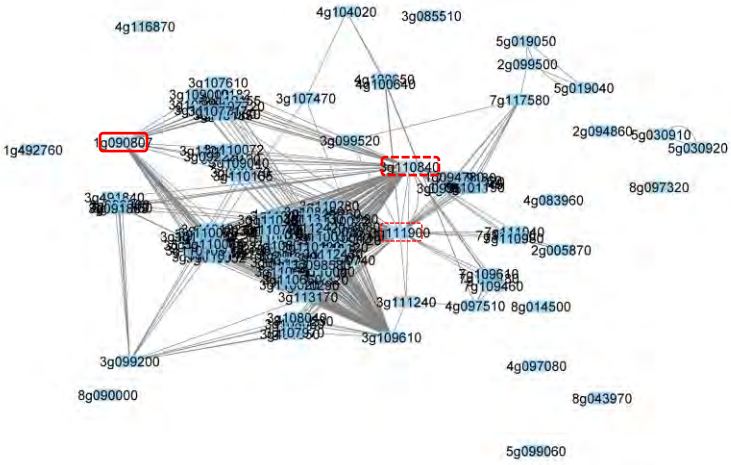
Annexe 21: Sous-réseaux génomiques d'interactions génétiques (au moins trois arêtes par gènes) pour des gènes impliqués dans les symbioses rhizobiennes et mycorhiziennes chez *M. truncatula* (A, B, C: population entière, FW et C).

A

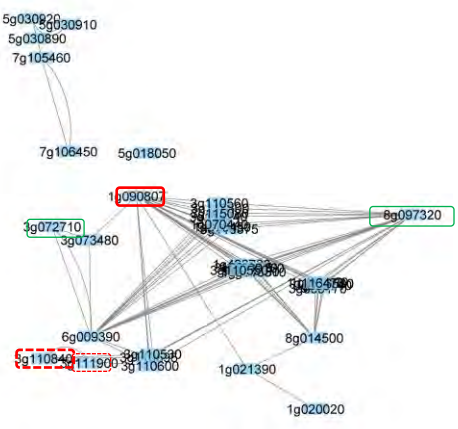


- interactions significatives au seuil de p-valeur de 10^{-8}
- interactions intra et inter-chromosomiques
- **chaque gène possède au moins <u>trois arêtes réseaux avec les identifiants des gènes</u>**
- les gènes caractérisés encadrés en rouges sont présents dans les 3 réseaux (les pointillés indiquent des gènes d'une même région sur le chromosome 3)
- les gènes caractérisés encadrés en vert sont présents dans 2 des 3 réseaux

B



C



Annexe 22: Sous-réseaux génomiques d'interactions génétiques (au moins deux arêtes par gènes) pour des gènes impliqués dans les symbioses rhizobiennes et mycorhiziennes chez *M. truncatula* (A, B, C: population entière, FW et C).