Collège Doctoral du Languedoc-Roussillon

THÈSE

Pour obtenir le grade de Docteur

Délivré par le
Centre international d'études supérieures en sciences agronomiques - Montpellier

Préparée au sein de l'école doctorale SIBAGHE (Systèmes Intégrés en Biologie, Agronomie, Géosciences, Hydrosciences, Environnement)

Et de l'unité de recherche AGAP (Amélioration Génétique et Adaptation des Plantes méditerranéennes et tropicales)

Spécialité : **Evolution, écologie, ressources génétiques, paléontologie**

Présentée par **Constance Vagne**

Rôle des effets épistatiques dans l'évolution d'une population à régime de reproduction mixte, et dans la régulation de l'expression des gènes homéologues chez une espèce autogame allotétraploïde, le blé dur (*Triticum turgidum*)

Soutenue le 19 décembre 2014 devant le jury composé de

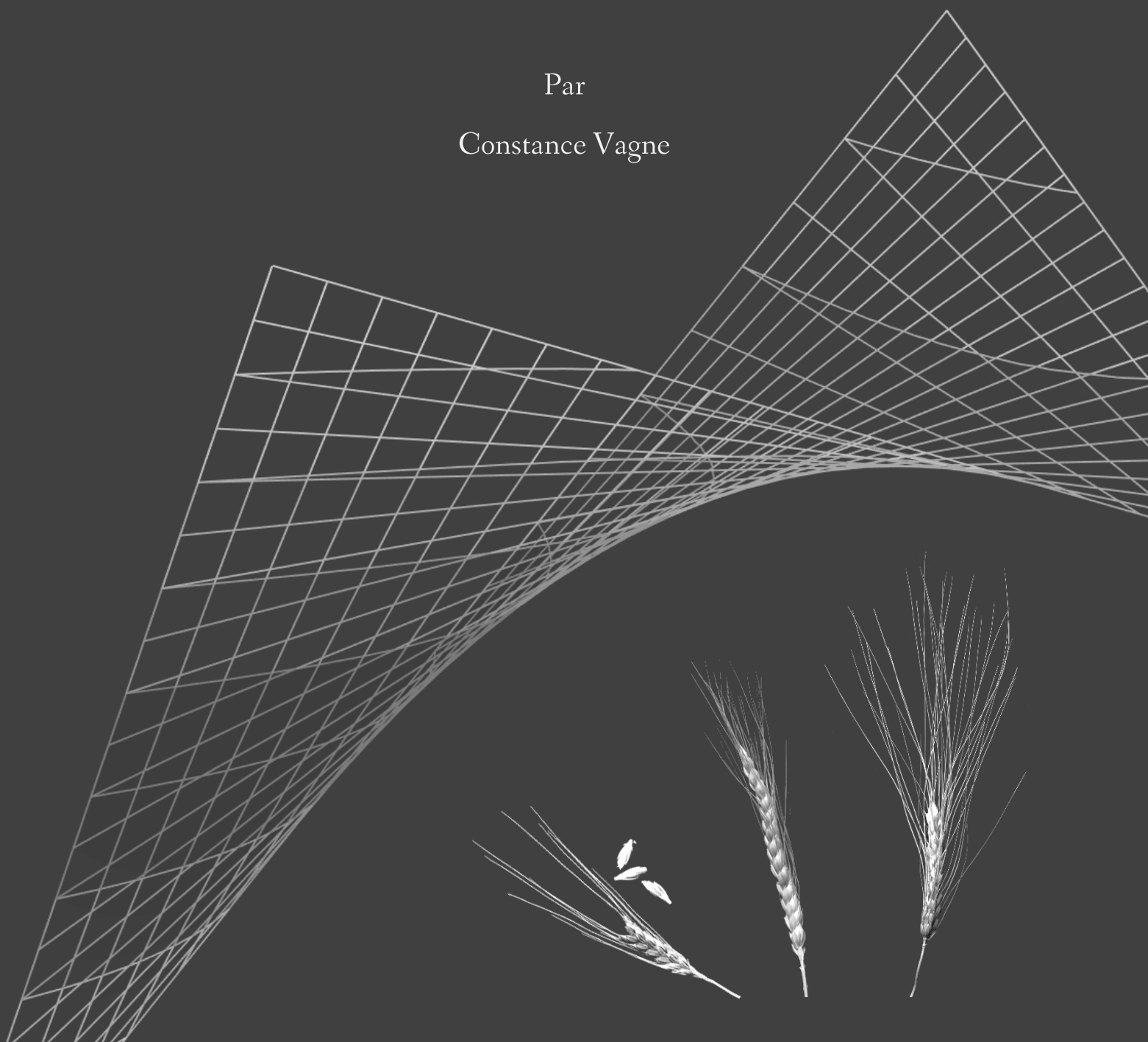
Brigitte MANGIN, Directrice de recherche, INRA Toulouse	Rapporteuse
Isabelle GOLDRINGER, Directrice de recherche, INRA du Moulon	Rapporteuse
Philippe LASHERMES, Directeur de recherche, IRD Montpellier	Examinateur
Christine DILLMANN, Professeur, Université Paris-Sud	Examinatrice
Eric JENCZEWSKI, Chargé de recherche, INRA de Versailles	Examinateur
Bénédicte FONTEZ, Maître de Conférences, Montpellier Supagro	Co-encadrante de thèse
Muriel TAVAUD, Maître de Conférences, Montpellier Supagro	Invitée
Jacques DAVID, Professeur, Montpellier Supagro	Directeur de thèse

Montpellier SupAgro

Rôle des effets épistatiques dans l'évolution d'une population à régime de reproduction mixte, et dans la régulation de l'expression des gènes homéologues chez une espèce autogame allotétraploïde, le blé dur (*Triticum turgidum*)

Par

Constance Vagne



REMERCIEMENTS

Je vais tenter l'art difficile des remerciements. Ils seront succincts mais les sentiments y sont.

Je voudrais tout d'abord remercier les membres du jury : Brigitte Mangin, Isabelle Goldringer, Christine Dillmann, Philippe Lashermes et Éric Jenczewski. Merci à vous d'avoir lu attentivement mon manuscrit et d'avoir accepté de l'évaluer.

Merci à mon directeur de thèse, Jacques David, pour m'avoir accueillie dans son équipe et pour m'avoir fait confiance tout au long de ce travail. Je remercie également chaleureusement mes encadrantes Bénédicte Fontez et Muriel Tavaud pour leur disponibilité, leur patience et leur gentillesse. Ça a été un véritable plaisir de travailler avec vous, malgré les difficultés que j'ai parfois rencontrées.

Je remercie les autres chercheurs de l'équipe GE2POP, Nathalie & Marie-Hélène, qui m'ont accompagnée aux Journées Jeunes Chercheurs, et avec qui on a parlé de films et de la vie associative à Montpellier ; Joëlle, Laurène, Pierre, Hélène, Vincent, Jean-Marie, Brigitte, Marie-France, Concetta, Vincent et Sylvain pour leur bon accueil dans l'équipe et leurs conseils avisés. Merci aussi aux membres de l'équipe MISTEA, notamment à Véronique, pour leur amabilité et le prêt de matériel. Merci également aux techniciennes du labo, Morgane, Muriel et Audrey pour leur travail et leur sympathie.

Merci à mes collègues thésards (du labo ou extérieurs) avec qui j'ai partagé mes inquiétudes et mes espoirs : Charlotte, Michel, Julie, Agota et Ingrid. J'ai aussi une pensée pour les stagiaires, post-doc et CDD du labo, avec qui j'ai eu de nombreuses discussions enjouées à la cantine, sur Minecraft, les tests de psychologie, les parcs d'attraction et d'autres sujets philosophiques : Pierre-Antoine, Julien, Maxime, Oscar, Iris, Jacques, Hajar et Martin. Parmi les CDD, je remercie fortement Yan pour son aide précieuse dans le projet EPO.

Merci à mes amis extérieurs au labo : Charlotte, Cécilia, Cécile, Rachel, Jean-Louis, Lauren, Julie, Marine, Silvère, David, Nico et Michel. Merci pour les soirées jeux de société, les sorties ciné, les discussions intéressantes, les actions de bénévolat faites ensemble, et tous les moments qui ont été si agréables et si forts. Je pense que grâce à vous, ces trois années ont été parmi les meilleures de ma vie, pleines de rebondissements.

Un merci tout spécial à Mathieu qui a toujours cru en moi, même dans les moments difficiles et qui m'a toujours encouragée. Ton soutien a été indispensable à cette thèse, et je sais que je pourrai encore longtemps compter sur toi.

Enfin, un grand merci aux membres de ma famille pour leurs encouragements et leur soutien.

PLAN

I. INTRODUCTION 11
I.1. Définitions de l'épistasie 11
I.1.1. L'épistasie en tant que non-indépendance des effets des gènes 13
I.1.2. L'épistasie en tant qu'interaction 13
I.1.3. Classifications de l'épistasie (en tant qu'interaction) 13
I.1.3.1. Epistasie statistique ou fonctionnelle 13
I.1.3.2. Epistasie concernant les homozygotes ou les hétérozygotes 14
I.1.3.3. Epistasie de magnitude ou de signe 16
I.2. L'estimation de l'épistasie et le renouveau de la génétique d'association 20
I.3. L'importance de l'épistasie et son rôle dans l'évolution 22
I.3.1. L'épistasie est quasiment omniprésente 23
I.3.2. L'épistasie peut contribuer à la variance additive 24
I.3.3. L'épistasie peut contraindre les trajectoires évolutives d'une population 25
I.3.4. L'épistasie peut générer de la dépression hybride et provoquer un isolement reproductif entre espèces 26
I.4. L'épistasie chez les autogames 27
I.5. L'épistasie et la polyploïdie : l'hétérosis fixé 28
I.6. L'épistasie dans la régulation de l'expression génétiques 30
I.7. L'épistasie et la gestion des ressources génétiques 33
I.8. Objectifs de la thèse 35
I.9. Notre modèle d'étude : le blé dur 37
II. ESTIMATION DE L'EPISTASIE DANS UN MODELE BILOCUS HAPLOÏDE 43
II.1. Préambule : l'orthogonalité, une propriété importante des modèles d'estimation des effets génétiques 44
II.2. Ecriture unilocus diploïde (estimation de l'effet additif et de dominance) 45
II.2.1 Ecriture générale des modèles d'estimation des effets génétiques unilocus diploïdes 45
II.2.2 Présentation de plusieurs modèles d'estimation des effets génétiques sous forme unilocus diploïde 47
II.2.2.1 Modèle G2A : modèle statistique où les effets additifs correspondent à des différences de moyennes 47
I.2.2.2 Modèle UWR : modèle d'estimation des effets fonctionnels 49
II.2.2.3 Modèle NOIA : un modèle orthogonal en présence de déséquilibre de Hardy-Weinber. 50
II.2.2.4 Autres modèles 52
II.3. Généralisation à plusieurs locus 52
II.4. Estimation de l'épistasie additive x additive : modèles haploïdes bilocus 53

II.4.1. Equivalence entre l'écriture diploïde et l'écriture haploïde 53
II.4.2. Ecriture des modèles présentés ci-dessus sous forme haploïde 55
II.4.2.1. Modèle G2A et modèle NOIA sous forme haploïde 55
II.4.2.2. Modèles UWR, F2 et F∞ haploïdes 56
II.4.2.3 Modèle centré G2A-D : modification du modèle G2A pour tenir compte du déséquilibre de liaison 56
II.4.2.4. Conclusion sur l'écriture haploïde 57
II.4.3. Calculs des variances et covariances génétiques dans un modèle dilocus haploïde. 59
II.5. Utilisation de ces modèles ultérieurement 61
III. RECHERCHE D'EPISTASIE DANS LA REGULATION DE L'EXPRESSION CHEZ *TRITICUM TURGIDUM* 65
III.1 Description du matériel végétal 65
III.1.1. Objectif : l'utilisation d'une population de pre-breeding pour une étude de d'association pangénomique (GWAS) 65
III.1.1.1. Les populations de pre-breeding 65
III.1.1.2. L'utilisation en association pangénomique 67
III.1.2 Description de la population étudiée 70
III.1.2.1. Fonctionnement de la population étudiée 70
III.1.2.2. Echantillonnage et marqueurs génétiques utilisés 70
III.1.3 Structure de la population 72
III.1.3.1. Détermination de la structure de la population en 2010 à l'aide d'une analyse factorielle des correspondances (AFC) et d'un dendrogramme 72
III.1.3.2. A l'origine de la structuration : la migration 72
III.1.4. Régime de reproduction 86
III.1.4.1. Calcul de la probabilité d'appartenir à une classe de consanguinité 86
III.1.4.2. Risque d'erreur dans l'assignation 87
III.1.4.3. Classe de consanguinité des individus de la population 88
III.1.5. Taille efficace 90
III.1.6. Conclusion et définition d'un échantillon sans structure pour l'étude sur les effets régulateurs de l'expression 91
III.2 Asymmetry of A and B genomes of tetraploid wheat *Triticum turgidum* in *cis* and *trans* regulations of the gene expression 93
III.2.1 Introduction 93
III.2.2 Materials and Methods 99
III.2.2.1 Plant material 99
III.2.2.2 RNA Sequencing (RNA-Seq) 100
III.2.2.3 Reads mapping 100
III.2.2.4 Genotyping 101

III.2.2.5 Data Analysis 102
III.2.3 Results 113
III.2.3.1 Comparison of the expression level and fractionation process between the two genomes 113
III.2.3.2 Regulatory effects 114
III.2.3.3 Relation between the expression of two homoelogous genes 119
III.2.3.4 Regulation among homoelogous pairs of genes 126
III.2.4. Discussion 127
III.2.4.1. Originality of this work 127
III.2.4.2. Technical discussion 127
III.2.4.3. Interpretation of results 129
III.2.4.4. Perspectives 135
III.2.5. References 135
III.3. Limites du modèle utilisé et perspectives 145
III.3.1 Indépendance entre les effets génétiques 145
III.3.2 Standardisation des effets 150
III.3.3 Résidus 152
III.3.2 Corrélation entre niveaux d'expression 155
IV. ROLE DE L'EPISTASIE DANS L'EVOLUTION 159
IV.1. Prédire l'évolution à court terme avec l'épistasie statistique 159
IV.1.1. Calcul de la covariance parent-enfant dans le cas haploïde bi-locus 159
IV.1.2 Prédiction du progrès génétique à l'aide de la covariance parent-enfant 161
IV.1.2.1. Le phénotype moyen des descendants en fonction du phénotype moyen des parents 161
IV.1.2.2. Prédiction de la réponse à la sélection en l'absence d'épistasie 162
IV.1.2.3. Prédiction de la réponse à la sélection en l'absence d'épistasie 164
IV.2 Prédire l'évolution à long terme avec l'épistasie fonctionnelle 166
IV.2.2 Théorie des paysages adaptatifs 166
IV.2.3 Epistasie et sélection par troncation 169
IV.3 Reciprocal sign epistasis and truncation selection: when is recombination favorable in a pre-breeding program with a selfing species? 173
IV.3.1 Introduction 175
IV.3.2 Materials and Methods 178
IV.3.2.1 Model 178
IV.3.2.2 Step 1: Identification of the different population configurations 179
IV.3.2.3 Step 2: Numerical calculations of times to fixation in populations of infinite size 184
IV.3.2.4 Step 3: Simulation of the evolution of a finite population 184

IV.3.2.5 Rationale of the model 185
IV.3.3 Results 186
IV.3.3.1 Evolutionary paths and final equilibria 186
IV.3.3.2 Fixation according to the recombination rate and strength of selection 190
IV.3.4 Discussion 197
IV.3.5 Conflict of interest 200
IV.3.6 Data Archiving 200
IV.3.7 Acknowledgements 201
IV.3.8 References 202
V. DISCUSSION GENERALE ET PERSPECTIVES 209
V.1. Difficulté de la détection de l'épistasie 209
V.2. Epistasie et pléiotropie sur le niveau d'expression de paires de gènes chez le blé dur 210
V.2.1. Détection d'épistasie SNP x fond génétique 210
V.2.2. Détection de pléiotropie 212
V.2.3. Les réseaux de gènes, la prochaine étape ? 213
V.3. Rôle de l'épistasie fonctionnelle dans l'adaptation et application au pre-breeding 216
V.4. Conclusion 217
VI. REFERENCES 220
VII. ANNEXES 247
VII.1 Calculs : types d'épistasie et effets génétiques 247
VII.2 Broadening the genetic basis of durum wheat 249
VII.3. Amorces et sources des microsatellites utilisés dans la partie III.1. 266
VII.4. Détails des calculs du modèle d'évolution d'une population connaissant un épisode de migration 267
VII.4.1. Expression analytique de l'évolution des fréquences génotypiques au locus responsable de la stérilité mâle 267
VII.4.1.1 Composition génétique des graines récoltées sur les plantes femelles (allofécondation), pour le locus F 267
VII.4.1.2 Composition génétique des graines récoltées sur les plantes femelles (autofécondation), pour le locus F 269
VII.4.1.3 Evolution des fréquences génotypique au locus F d'une génération à l'autre 270
VII.4.1.4 Fréquences génotypiques au locus F à l'équilibre 270
VII.4.2. Expression analytique de l'évolution des fréquences génotypiques à un locus A 271
VII.4.2.1 Composition génétique des graines récoltées sur les plantes femelles (allofécondation), pour le locus A 271
VII.4.2.2 Composition génétique des graines récoltées sur les plantes hermaphrodites (autofécondation), pour le locus A 272

VII.4.2.3 Evolution des fréquences génotypique au locus A d'une génération à l'autre 272

VII.4.3. Expression analytique de l'évolution des fréquences génotypiques pour les locus F et A 272

VII.4.3.1 Composition génétique des graines récoltées sur les plantes femelles (allofécondation), pour les locus F et A 272

VII.4.3.2 Composition génétique des graines récoltées sur les plantes hermaphrodites (autofécondation), pour les locus F et A 273

VII.4.3.3 Evolution des fréquences génotypique aux locus A et F d'une génération à l'autre ... 274

VII.5. Principe du RPDS (Reads per Divergent Sites) 275

VII.6. Fichiers supplémentaires de l'article *Reciprocal sign epistasis and truncation selection: when is recombination favorable in a pre-breeding program with a selfing species?* 276

VII.6.1. Supplementary File 1: Conditions to move from one configuration to another or to fix a genotype 276

VII.6.1.1. Proposition 1: Frequency differential between the two competitor genotypes 276

VII.6.1.2. Proposition 2: Evolution of the frequency of the most efficient genotype *AB* 276

VII.6.1.3. Conditions for the evolution of the population 276

VII.6.1.4. Possible evolutionary paths from configuration 1 278

VII.6.1.5. Possible evolutionary paths from configuration 2 282

VII.6.2. Supplementary File 2: Results of simulations in finite populations 288

VII.6.3. Supplementary File 3: Initial State 1, when *aB* is not at the same frequency as *Ab* 290

VII.7. Commentaires des reviewers sur l'article *Reciprocal sign epistasis and truncation selection: when is recombination favorable in a pre-breeding program with a selfing species?* 291

PARTIE I

INTRODUCTION

I. INTRODUCTION

Longtemps négligée, l'épistasie – l'interaction entre les gènes - connaît un renouveau avec les possibilités offertes par les méthodes de séquençage de nouvelle génération (NGS : Next Generation Sequencing) (Davey et al., 2011). Etant quasiment omniprésente, et contribuant de manière substantielle à un grand nombre de caractères phénotypes chez diverses espèces, elle peut jouer un rôle fondamental dans l'évolution des populations.

Dans son sens le plus large, l'épistasie désigne le fait que l'effet d'un allèle dépende de son fond génétique – c'est-à-dire de l'ensemble des allèles du génome. Dans son sens le plus courant, on appelle « épistasie » l'interaction entre deux gènes ou quelques gènes. De manière générale, l'effet d'un allèle peut dépendre d'un ou plusieurs autres gènes, notamment si ces gènes constituent un réseau.

En termes évolutifs, l'épistasie peut présenter un inconvénient, car elle peut contraindre les chemins adaptatifs que prend une population soumise à sélection. Elle peut également présenter des avantages, en évitant la ségrégation d'allèles coadaptés, notamment chez les espèces polyploïdes. Dans cette thèse, nous avons voulu approfondir le rôle de l'épistasie dans l'évolution et la régulation de l'expression, en lien avec le régime de reproduction et la polyploïdie.

I.1. Définitions de l'épistasie

Il existe plusieurs définitions de l'épistasie (Cordell, 2002; Wade et al., 2001). Le terme « épistasie » a été employé pour la première fois en 1909 par Bateson (Bateson and Mendel, 1909) pour décrire l'effet de masquage d'un allèle, c'est-à-dire le fait qu'un allèle à un locus empêche un allèle à autre locus de s'exprimer. Par exemple, chez le chat, la couleur rousse du pelage est déterminée par le gène co-dominant *orange* (locus O), situé sur le chromosome X (Schmidt-Küntzel et al., 2009). Si le chat porte l'allèle O en une (pour les mâles) ou deux copies (pour les femelles), il sera roux. Sinon, il aura un pelage foncé (noir ou brun), ou éventuellement bicolore pour les femelles. D'autres gènes agissent sur cette couleur de base pour former le pelage : ajout de taches blanches, de rayures, etc. Le gène dominant *white* (locus W) a également un effet sur la couleur du pelage du chat (Geigy et al., 2007). En effet, si un chat est porteur d'au moins un allèle W (génotype W/W ou W/w) à ce locus, il sera complètement blanc, quel que soit son génotype au locus O. On peut donc dire que l'allèle W du gène W est épistatique au locus O (et aux autres gènes ayant un effet sur le pelage), car il masque son effet. Ceci est illustré en **figure 1**.

L'épistasie a pris ensuite un sens plus général. Elle désigne le fait que l'action d'un gène dépende du génotype à un autre gène ou à plusieurs autres gènes. On peut aussi parler d'épistasie quand l'effet

		Locus W		
		ww	*Ww*	*WW*
Locus O	*OO* (♀) ou *O* (♂)			
	oo (♀) ou *o* (♂)			
	Oo (♀)			

Figure 1 : Exemple d'un phénotype (couleur du pelage du chat) soumis à épistasie selon la définition de Bateson (1909)

La couleur du pelage du chat est déterminée par de nombreux gènes. Ici deux sont illustrés : locus *orange* O (situé sur le chromosome X) et locus *white* W

Le locus W est dit épistatique car il empêche l'expression du locus O.

d'un allèle dépend de l'ensemble du fond génétique. Il existe deux grands types de définitions utilisés en génétique.

I.1.1. L'épistasie en tant que non-indépendance des effets des gènes

C'est le concept utilisé en génétique des populations, notamment pour modéliser la valeur sélective (Otto and Feldman, 1997). Une multitude de définitions de l'indépendance existent dans ces disciplines, mais le modèle plus couramment utilisé est le modèle multiplicatif (Wade et al., 2001). Dans ce cas, l'indépendance est définie de manière probabiliste : deux gènes sont considérés comme indépendants (et donc, non épistatiques) quand la valeur sélective d'un génotype est égale au produit des valeurs sélective de chacun des allèles. Par exemple, si l'on prend un modèle haploïde simple : si la valeur sélective d'un allèle A à un locus A est égale à $W_A = (1 - s)$, et celle d'un allèle B (locus B) à $W_B = (1 - t)$, (où $0 < s, t < 1$ et où s et t sont des coefficients de sélection) alors, s'il y a indépendance, la valeur sélective du génotype AB est égale à $W_{AB} = (1 - s)(1 - t)$. En présence d'épistasie, la valeur sélective s'écrira $W_{AB} = (1 - s)(1 - t) + \varepsilon$, avec ε la déviation épistatique.

I.1.2. L'épistasie en tant qu'interaction

Fisher a défini l'épistasie comme étant l'interaction entre locus différents (Fisher, 1918). Elle est distincte de la dominance, qui est l'interaction entre allèles à un même locus. Il s'agit de la définition utilisée en génétique quantitative, et c'est celle que nous emploierons dans ce document. Tous les modèles de génétique quantitative sont basés sur la régression linéaire (Falconer and Mackay, 1996; Zeng et al., 2005), dans lesquels l'interaction est définie comme étant un écart à l'additivité. Prenons un exemple : soit une population haploïde où la valeur sélective moyenne est égale à 1. Soit un allèle délétère A à un premier locus qui a un effet -s sur la valeur sélective, et un allèle B à un second locus qui a un effet -t. Alors, en l'absence d'interaction, la valeur sélective du génotype AB sera égale à 1-s-t. En présence d'épistasie, la valeur sélective s'écrira $W_{AB} = 1 - s - t + e$, avec e l'effet épistatique.

I.1.3. Classifications de l'épistasie (en tant qu'interaction)

L'épistasie existe sous plusieurs formes et différentes catégorisations permettent de déterminer sa nature. Ici, nous détaillerons les différentes formes qu'elle peut prendre, quand elle est définie comme « interaction entre les gènes », c'est-à-dire estimée à partir de modèles de génétique quantitative, basés sur la régression linéaire.

I.1.3.1. Epistasie statistique ou fonctionnelle

Historiquement, les modèles de génétique quantitative estiment des effets génétiques (additifs, de dominance ou épistatiques) qui découlent à la fois de la propriété des gènes étudiés, mais aussi de la population dans laquelle l'étude est menée. C'est notamment le cas du premier modèle de génétique

quantitative qui a été développé, celui de Fisher (Fisher, 1918). Ces modèles sont appelés « modèles statistiques ». Dans ce cas, les effets estimés sont généralement fonction des fréquences alléliques et/ou génotypiques. Quand l'épistasie est estimée avec ce type de modèle, on parle alors « d'épistasie statistique ». Ce type d'épistasie est celle qui intéresse les sélectionneurs, car c'est celle qui permet de déterminer les sources de variations génétiques au sein de la population.

D'autres modèles cherchent à estimer des effets génétiques indépendamment de la population étudiée. Ces effets sont donc censés refléter le fonctionnement des gènes uniquement. Ces modèles mettent l'accent sur les valeurs génotypiques (la valeur attendue du phénotype en l'absence d'effets environnementaux), car elles représentent les effets des allèles indépendamment des fréquences alléliques dans la population. On parle de modèles fonctionnels. L'un des plus connus est celui développé par Falconer and Mackay, 1996. (Cheverud and Routman, 1995) en ont proposé un, nommé UWR (UnWeighted Regression model), qui permet d'estimer l'épistasie. L'épistasie détectée par leur modèle a été nommé « épistasie fonctionnelle ». D'autres auteurs ont par la suite développé des modèles pour estimer cette épistasie fonctionnelle (Álvarez-Castro and Carlborg, 2007; Hansen and Wagner, 2001).

La **figure 2** illustre la différence entre un modèle fonctionnel (Figure 2.A) et un modèle statistique (Figure 2.B). Il s'agit de modèles unilocus diploïde, et donc ne permettant pas l'estimation de l'épistasie. Néanmoins, la figure permet de comprendre la différence entre les deux types de modèles. Soit B un locus portant deux allèles b et B. Les valeurs génotypiques (Figure 2.A) sont égales à :

$$G_{BB} = c + a$$

$$G_{bB} = c + d$$

$$G_{bb} = c - a$$

Avec a l'effet additif, et d l'effet de dominance.

Dans le modèle statistique (Figure2.B), l'effet additif correspond à la pente de la droite de régression $G(x)$, avec x le nombre d'allèles B, et $G(x)$ la valeur génotypique. Cette régression est pondérée en fonction des fréquences génotypiques dans la population. Dans le modèle statistique, l'effet de dominance est estimé de la même façon que dans le modèle fonctionnel : $d = G_{bB} - \frac{G_{bb}+G_{BB}}{2}$

I.1.3.2. Epistasie concernant les homozygotes ou les hétérozygotes

Dans un modèle bi-locus, bi-allélique et diploïde, l'épistasie peut-être partitionnée en quatre composantes:

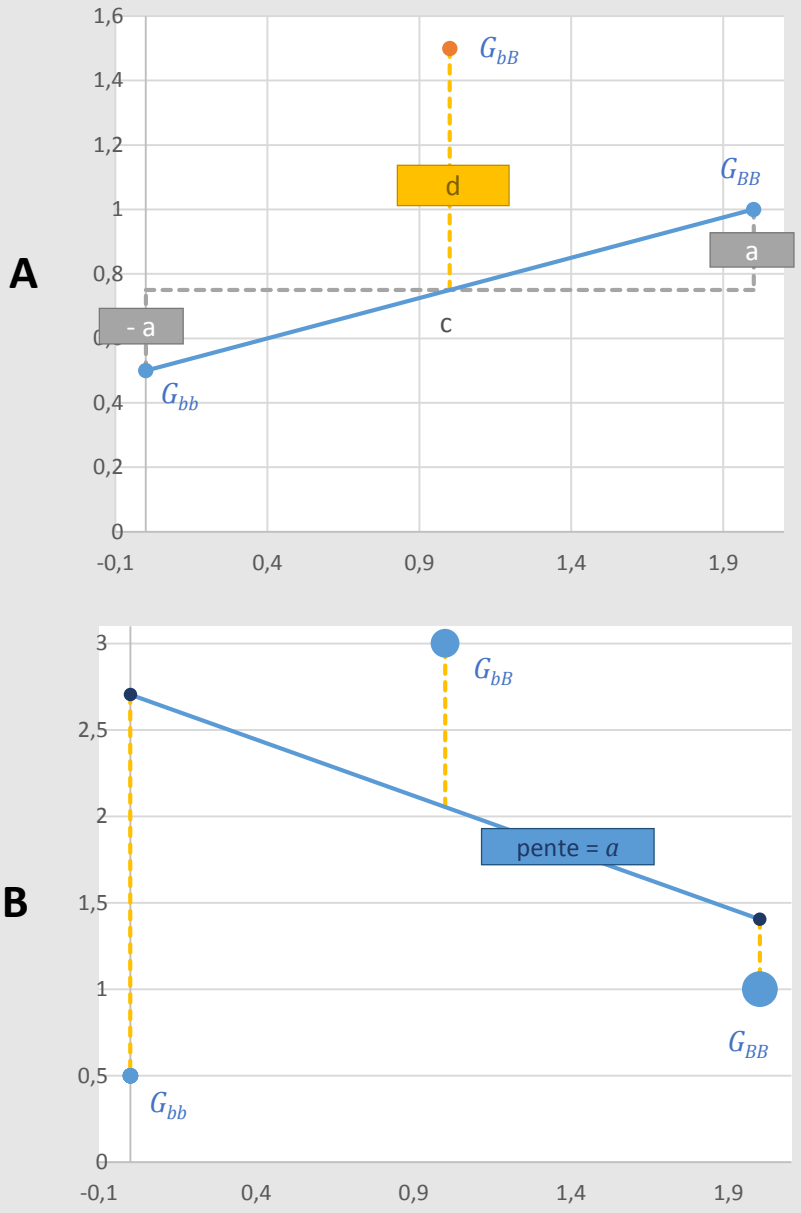


Figure 2 : Interprétation graphique d'un modèle fonctionnel (A) et statistique (B)
En abscisse : nombre d'allèles B pour chaque génotype (0, 1, 2)
En ordonnée : valeurs génotypiques G_{bb}, G_{bB} et G_{BB}
Points bleus : les valeurs génotypiques.
Droite bleue : la droite de régression $G(x)$, avec x le nombre d'allèle B. Dans le graphique A, cette droite n'est pas pondérée. Dans le graphique B, elle est pondérée fonction de la fréquence génotypique.
a: effet additif
d: effet de dominance

- l'épistasie additive x additive. Il s'agit de l'interaction entre un allèle à un locus et un autre allèle au deuxième locus
- l'épistasie additive x dominante ou dominante x additive. C'est l'interaction entre un allèle à un locus et le couple d'allèles au deuxième locus
- l'épistasie dominante x dominante. C'est l'interaction entre les deux couples d'allèles

Ces types d'épistasie sont illustrés dans le **tableau 1**, d'un point de vue fonctionnel (en examinant les effets intrinsèques des gènes).

Historiquement, les premiers à avoir ainsi décomposé l'épistasie en quatre éléments estimables par un modèle statistique ont été Cockerham et Kempthorne (Cockerham, 1954; Kempthorne, 1954). Il s'agit des quatre composantes de l'épistasie de premier ordre, c'est-à-dire impliquant deux locus. Si plus de deux locus sont inclus dans le modèle, il est possible d'estimer de l'épistasie de deuxième ordre ou plus. Dans la pratique, seule l'épistasie de premier ordre (plus rarement, de deuxième ordre) est estimée, car plus les effets d'interaction sont complexes, plus leur estimation devient difficile et nécessite des échantillons de grande taille. Néanmoins, depuis quelques années, plusieurs méthodes, se basant notamment sur les réseaux de gènes, ont été développées pour tenter de détecter de l'épistasie d'ordre supérieure (Imielinski and Belta, 2008; Pettersson et al., 2011; Stich et al., 2007) .

De l'épistasie additive x additive et dominante x dominante a été détectée pour la résistance à la rouille chez le blé dur (Bnejdi and Gazzah, 2008). Il a été signalé que l'épistasie impliquant des effets de dominance (additive x dominante, dominante x additive ou dominante x dominante) pouvait entraîner de la dépression de consanguinité chez le blé dur, pour le nombre de grains par épi (Sharma and Sain, 2004). Enfin, toujours sur le blé dur, la concentration en β-carotène est soumise à diverses composantes de l'épistasie : additive x additive, additive x dominante, et dominante x dominante (Santra et al., 2005).

Dans notre travail, nous avons utilisé le blé dur (*Triticum turgidum*) comme modèle d'étude. Comme il s'agit d'une espèce autogame, l'hétérozygotie y est relativement réduite, et nous avons supposé que les effets de dominance étaient alors négligeables. Nous avons donc utilisé des modèles haploïdes (équivalents à des modèles diploïdes sans hétérozygotes) et la composante de l'épistasie sur laquelle notre travail a essentiellement porté a été l'épistasie additive x additive.

I.1.3.3. Epistasie de magnitude ou de signe

L'épistasie additive x additive peut être soit « de magnitude », soit « de signe » (Weinreich et al., 2005). La **figure 3** détaille la différence entre ces deux formes d'épistasie dans le cas haploïde. L'épistasie de signe (Figure 3.D-E) désigne le fait qu'une mutation favorable dans un fond génétique devienne

Tableau 1 : les différentes composantes de l'épistasie (effets physiologiques)
Dans les cases : les valeurs génotypiques.

A. Tous les effets génétiques sont nuls, à l'exception de l'épistasie additive x additive ($a \times a$)
B. Tous les effets génétiques sont nuls, à l'exception de l'épistasie additive x dominante ($a \times d$)
C. Tous les effets génétiques sont nuls, à l'exception de l'épistasie dominante x dominante ($d \times d$)

L'épistasie dominante x additive se manifeste de la même façon que l'épistasie additive x dominante (B), sauf que les lignes et les colonnes sont inversées.

A. Tous les effets génétiques fonctionnels sont nuls, sauf l'épistasie additive x additive

	BB	***Bb***	***bb***
AA	1	0	-1
Aa	0	0	0
aa	-1	0	1

$$a \times a = \frac{G_{AABB} + G_{aabb} - G_{aaBB} - G_{AAbb}}{4} = 1$$

B. Tous les effets génétiques fonctionnels sont nuls, sauf l'épistasie additive x dominante

	BB	***Bb***	***bb***
AA	-1	0	1
Aa	2	0	-2
aa	-1	0	1

$$a \times d = \frac{(2G_{AaBB} - G_{AABB} - G_{aaBB})}{4} - \frac{(2G_{Aabb} - G_{AAbb} - G_{aabb})}{4} = 1$$

C. Tous les effets génétiques fonctionnels sont nuls, sauf l'épistasie dominante x dominante

	BB	***Bb***	***bb***
AA	1	-2	1
Aa	-2	4	-2
aa	1	-2	1

$$d \times d = \frac{(2G_{AaBb} - G_{AABb} - G_{aaBb})}{2} - \frac{(2G_{AaBB} - G_{AABB} - G_{aaBB})}{4} - \frac{(2G_{Aabb} - G_{AAbb} - G_{aabb})}{4} = 9$$

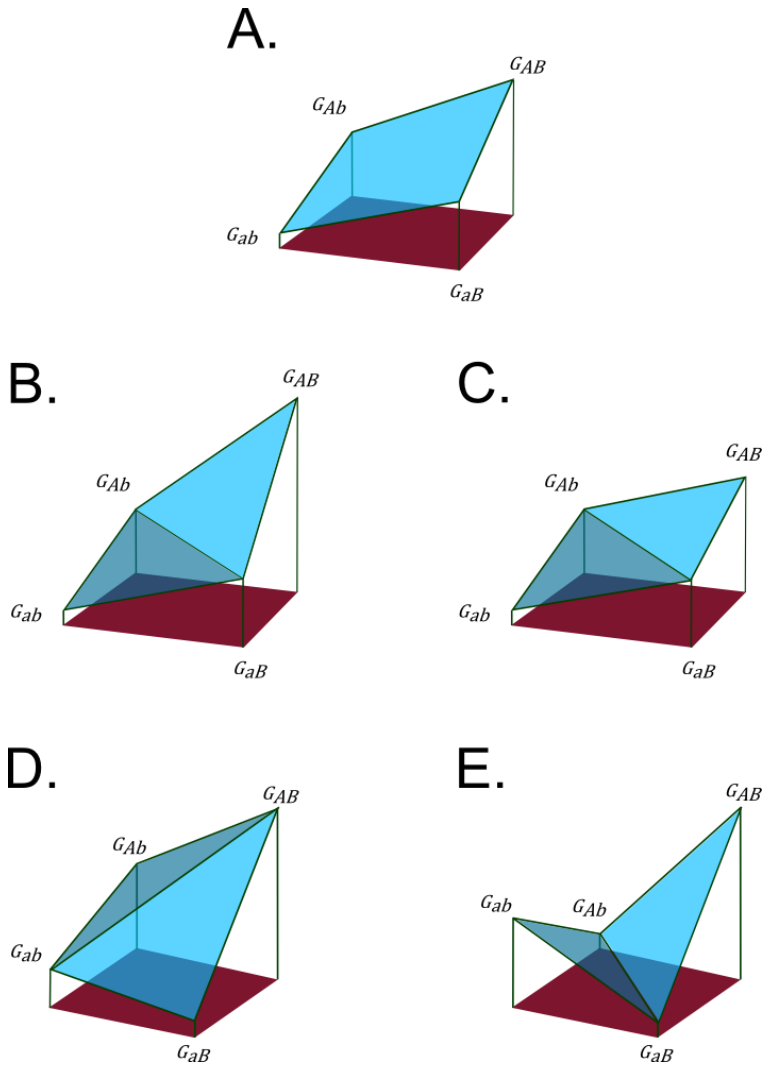


Figure 3 : Epistasie de magnitude et épistasie de signe dans le cas haploïde bi-locus
Abscisse : allèle au locus B (b ou B), ordonné : allèle au locus A (a ou A), cote : G_{ij} = valeur génotypique du génotype ij (i = allèle au locus A, j = allèle au locus B) = phénotype attendu en l'absence d'effets environnementaux

- **A. Additivité**
- **B. Epistasie positive** (un cas d'épistasie de magnitude). La valeur génotypique G_{AB} est plus forte que dans l'attendu additif. L'allèle A a un effet plus positif sur le phénotype quand il est associé à l'allèle B que quand il est associé à l'allèle b ($G_{AB} - G_{aB} > G_{Ab} - G_{ab}$). De la même façon, l'allèle B a un effet plus positif sur le phénotype quand il est associé à A quand il est associé à a ($G_{AB} - G_{Ab} > G_{aB} - G_{ab}$).. Cependant, A est toujours plus favorable que a, et B toujours plus favorable que b.
- **C. Epistasie négative** (un cas d'épistasie de magnitude). La valeur génotypique G_{AB} est plus faible que l'attendu additif. L'allèle A a un effet moins positif sur le phénotype quand il est associé à l'allèle B que quand il est associé à l'allèle b ($G_{AB} - G_{aB} < G_{Ab} - G_{ab}$). De la même façon, l'allèle B a un effet moins positif sur le phénotype quand il est associé à A quand il est associé à a ($G_{AB} - G_{Ab} < G_{aB} - G_{ab}$). Cependant, A est toujours plus favorable que a, et B toujours plus favorable que b.
- **D. Epistasie de signe**. L'allèle B est plus favorable que b quand l'allèle au locus A est A ($G_{AB} - G_{Ab} > 0$). Cependant, si le locus A porte l'allèle a, son sens s'inverse : l'allèle b est plus favorable que l'allèle B ($G_{aB} - G_{ab} < 0$). L'effet des allèles du locus B change donc de signe en fonction de l'allèle porté par le locus A. Cependant, l'allèle A est toujours plus favorable que l'allèle a.
- **E. Epistasie de signe réciproque**. L'allèle B est favorable quand il est associé à l'allèle A ($G_{AB} - G_{Ab} > 0$), et défavorable quand il est associé à l'allèle a ($G_{aB} - G_{ab} < 0$). L'inverse est vrai aussi : l'allèle A est favorable quand il est associé à l'allèle B ($G_{AB} - G_{aB} > 0$), et défavorable quand il est associé à l'allèle b ($G_{Ab} - G_{ab} < 0$). L'effet des allèles aux deux locus change donc de signe en fonction de l'allèle porté par l'autre locus.

défavorable dans un autre fond génétique. L'épistasie de magnitude (Figure 3.B-C) désigne un écart à l'additivité, sans pour autant que l'effet des allèles ne change de sens : un allèle favorable reste favorable, quel que soit le fond génétique. Il en va de même pour un allèle défavorable.

Il existe deux sortes d'épistasie de magnitude : l'épistasie positive et l'épistasie négative. L'épistasie positive se manifeste quand deux allèles ayant un effet individuel favorable, sont encore plus bénéfiques quand ils sont associés (Figure 3.B). Elle se manifeste aussi quand deux allèles négatifs ont un effet moins défavorable qu'attendu quand ils sont combinés. A l'inverse, on parle d'épistasie négative si, lorsque deux allèles bénéfiques sont associés, ils ont un effet moins favorable que la somme de leurs effets (Figure 3.C). Elle apparait également quand deux allèles négatifs combinés ont un effet particulièrement défavorable.

De la même façon, l'épistasie de signe peut-être simple, ou de signe réciproque. Il y a épistasie de signe simple (Figure 3.D) lorsque seulement l'un des deux locus a un effet dont le signe change en fonction de l'autre locus. A l'inverse, l'épistasie de signe réciproque (Figure 3.E) se manifeste si les deux locus ont un effet qui dépend du génotype à l'autre locus.

Si l'on se place dans un modèle fonctionnel simple, haploïde et bilocus, on peut constater que le type d'épistasie peut être défini d'un point de vue mathématique. Nous prendrons une version haploïde du modèle fonctionnel de Cheverud & Routman (Cheverud and Routman, 1995), qui sera présenté ultérieurement (II.2.2.2 pour la version diploïde, II.4.2.1. pour la version haploïde). Soit un individu k porteur de l'allèle i au locus A et de l'allèle j au locus B. Son phénotype Y_{ijk} d'est égal à :

$$Y_{ijk} = G_{ij} + \varepsilon_{ijk} = \mu + x_i \cdot a_A + x_j \cdot a_B + x_j \cdot x_i \cdot aa + \varepsilon_{ijk}$$

Avec $x_i = \begin{cases} 1/2 & \text{si } i = A \\ -1/2 & \text{si } i = a \end{cases}$ et $x_j = \begin{cases} 1/2 & \text{si } j = B \\ -1/2 & \text{si } j = b \end{cases}$

Avec G_{ij} la valeur génotypique associé au génotype ij ; a_A l'effet additif au locus A ; a_B l'effet additif au locus B ; aa l'effet de l'épistasie additive x additive ; ε_{ijk} l'effet de l'environnement.

Les effets génétiques sont égaux à :

$$a_A = \frac{(G_{AB} + G_{Ab})}{2} - \frac{(G_{aB} + G_{ab})}{2}$$

$$a_B = \frac{(G_{AB} + G_{aB})}{2} - \frac{(G_{Ab} + G_{ab})}{2}$$

$$aa = G_{AB} - G_{Ab} - G_{aB} + G_{ab}$$

A partir de ces effets génétiques, il est possible de définir le type d'épistasie (magnitude ou de signe) auquel on est confronté.

On oriente les effets additifs de sorte qu'ils soient positifs ($a_A \geq 0$ et $a_B \geq 0$), c'est-à-dire, de sorte que l'allèle A ait un effet plus fort que l'allèle a, et que l'allèle B ait un effet plus fort que l'allèle b. On suppose aussi que dans tous les cas, le génotype AB a la plus forte valeur génotypique (G_{AB}). En posant cela, on peut en déduire que :

- Il n'y a pas d'épistasie si $aa = 0$
- Il y a de l'épistasie de magnitude, positive si $0 < aa \leq 2a_B$ et $0 < aa \leq 2a_A$
- Il y a de l'épistasie de magnitude, négative si $0 > aa \geq -2a_B$ et $0 > aa \geq -2a_A$
- Il y a de l'épistasie de signe simple si $0 < 2a_A < aa \leq 2a_B$ ou $0 < 2a_B < aa \leq 2a_A$
- Il y a de l'épistasie de signe réciproque si $0 < 2a_A, 2a_B < aa$

L'ordre des valeurs génotypiques et des effets génétiques fonctionnels dans chaque cas d'épistasie, est illustré dans le **tableau 2**. Les détails des calculs sont présentés en **Annexe VII.1**.

Tableau 2 : traduction mathématique des différents types d'épistasie

aa: effet épistatique
a_A : effet additif du locus A, > 0
a_B : effet additif du locus B, > 0

Type d'épistasie	Conditions	Ordre des valeurs génotypiques	Graphique de la figure 3
Pas d'épistasie	$aa = 0$	$G_{AB} \geq G_{Ab}, G_{aB} \geq G_{ab}$	A
Epistasis de magnitude, positive	$aa > 0$ $aa \leq 2a_A, 2a_B$	$G_{AB} \geq G_{Ab}, G_{aB} \geq G_{ab}$	B
Epistasis de magnitude, négative	$aa < 0$ $-aa \leq 2a_A, 2a_B$	$G_{AB} \geq G_{Ab}, G_{aB} \geq G_{ab}$	C
Epistasie de signe simple	$0 < 2a_A < aa \leq 2a_B$ ou $0 < 2a_B < aa \leq 2a_A$	$G_{AB} > G_{aB} > G_{ab} > G_{Ab}$ ou $G_{AB} > G_{Ab} > G_{aB} > G_{ab}$	D
Epistasie de signe réciproque	$0 < 2a_A, 2a_B < aa$	$G_{AB} \geq G_{ab} > G_{Ab}, G_{aB}$	E

I.2. L'estimation de l'épistasie et le renouveau de la génétique d'association

L'épistasie peut être estimée de plusieurs façons. Une méthode qui a longtemps été utilisée est celle de la méthode des plans de croisement (Gallais, 1990). L'objectif est de constituer des familles de plein-frères ou de demi-frères, avec un ou deux parents contrôlés. En effectuant une analyse de la variance, on peut en déduire des variances des effets de la famille (quand un seul parent est contrôlé), ou des effets de chacun deux parents (quand les deux parents sont contrôlés). A partir de là, les variances

génétiques peuvent être déduites. Cette méthode a par exemple été utilisée pour estimer les variances additive et additive x additive sur des haploïdes doublés de blé tendre, pour différents caractères (taille de la plante, composants du rendement, etc.), par Goldringer et al., 1997.

Il est également possible de déterminer les effets et variances génétiques en s'intéressant directement aux gènes causaux, quand ils sont connus. Par exemple une étude a tâché d'identifier les effets des gènes Glu-A1, Glu-B1 et Glu-D1 dans la qualité de la panification chez le blé tendre (Kolster et al., 1991). Ces gènes produisent des protéines, qui constituent des sous-unités des gluténines de haut poids moléculaires. Les gluténines sont les protéines qui confèrent une propriété d'élasticité à la pâte du pain, et elles sont donc d'une importance capitale pour la qualité de panification du pain. Les allèles des gènes ont été identifiés en examinant directement leur produit (les protéines) à l'aide d'une électrophorèse. A partir de ces données, il a été possible d'estimer des effets génétiques et d'en déduire des variances. Il a été démontré que les interactions entre les gènes Glu-A1, Glu-B1 et Glu-D1 n'étaient pas négligeables. Ce genre d'étude nécessite néanmoins de connaître les gènes impliqués dans le caractère étudié.

Une autre méthode consiste à effectuer de la génétique d'association. Ces études utilisent de nombreux marqueurs génétiques, qui couvrent l'ensemble du génome, et font l'hypothèse qu'un (ou plusieurs) des marqueurs étudiés est à l'origine de la variation phénotypique, ou bien est lié physiquement au locus en cause (Rafalski, 2010). Ainsi, si les individus présentant un trait particulier sont tous porteurs d'un allèle particulier à un marqueur, il est fort probable qu'il y a un déséquilibre de liaison entre cet allèle du marqueur, et l'allèle responsable du trait au locus recherché. Certaines études d'association utilisaient à peine quelques dizaines ou une centaine de marqueurs. Une étude utilisant 115 microsatellites (Reif et al., 2011) a permis de découvrir plus d'une vingtaine de QTL épistatiques pour des traits liés à la qualité du grain (volume de sédimentation et poids spécifique) chez le blé tendre. L'épistasie comptait pour environ 13-14% de la variance génétique de ces caractères. Une autre étude, utilisant également 115 microsatellites chez le blé tendre, a pu démontrer le rôle substantiel de l'épistasie dans la résistance à la fusariose (Miedaner et al., 2011)

A partir de 2004, le développement des techniques de séquençage haut-débit a permis le démarrage des études d'association pangénomique (GWAS : Genome-Wide Association Studies). Ces études sont des études d'association qui se caractérisent par l'utilisation d'un très grand nombre de marqueurs (en l'occurrence, des SNPs), distribués tout le long du génome. L'objectif est de balayer l'ensemble des marqueurs, afin de trouver des variants génétiques associés au phénotype d'intérêt. Les GWAS ont permis la découverte de nombreux variants responsables de certains caractères phénotypiques, comme des maladies.

Néanmoins, la plupart des QTL détectés par les études de GWAS n'ont qu'un très faible effet sur le phénotype (Manolio et al., 2009; Zuk et al., 2012). Les variances additives trouvées par des plans de croisement représentaient une part bien plus importante de la variance génétique, que celle estimées à partir des marqueurs moléculaires. On a donc parlé « d'héritabilité manquante ». Par exemple, les études basées sur les ressemblances entre apparentés ont montré que la taille humaine était à 75-80% héritable (Macgregor et al., 2006; Silventoinen et al., 2003). Or plus de 180 locus associés avec la taille ont été détecté dans une étude portant sur plus de 180 000 individus, et ces 180 locus ne représentent qu'à peine 10% de la variation de la taille (Lango Allen et al., 2010). Une grande part de l'héritabilité reste donc à expliquer. L'épistasie pourrait expliquer une part de cette héritabilité manquante (Zuk et al., 2012). Une autre cause de cette héritabilité manquante pourrait être la présence de variants rares (Zuk et al., 2014).

La détection de l'épistasie dans les études d'association pangénomique est complexe. Il existe plusieurs méthodes qui se fondent sur l'idée des réseaux de gènes (McKinney and Pajewski, 2012). Elles suggèrent que les facteurs génétiques d'un trait phénotype seront mieux compris si on conceptualise la carte génotype-phénotype (= le lien entre le génotype et le phénotype) comme un réseau d'interactions gène à gène, couplées à des effets additifs. Ces réseaux sont appelés « réseaux épistatiques ». Certains gènes ont des effets individuels qui sont plus ou moins importants : ce sont eux qui ont détectés par des analyses classiques. D'autres n'ont pas d'effets individuels, mais interagissent avec un très grand nombre de gènes et modulent leur effet sur le phénotype d'intérêt. Il est donc important d'être capable de les détecter pour comprendre l'architecture génétique d'un trait.

I.3. L'importance de l'épistasie et son rôle dans l'évolution

Les études de génétique quantitative se sont longtemps concentrées uniquement sur les effets additifs des gènes et sur la variance additive. Cela est dû au fait que, dans le cadre d'un programme de sélection, la réponse à la sélection a été et demeure le principal intérêt des sélectionneurs. Comme cette réponse à la sélection dépend essentiellement de la variance additive, tout ce qui n'était pas additif (épistasie, dominance, interaction génotype x environnement…) a donc longtemps été considéré comme étant une nuisance.

A l'heure actuelle, l'importance de l'épistasie est encore controversée (Carlborg and Haley, 2004; Hill et al., 2008). En réalité, l'importance qu'on accorde à l'épistasie dépend du point de vue que l'on adopte. Si l'on examine la part de variance génétique à laquelle contribue la variance épistatique, les données empiriques et théoriques démontrent qu'elle est souvent faible, et que la plupart du temps la majeure partie de la variance génétique est due à la variance additive (Crow, 2010; Hill et al., 2008;

Mäki-Tanila and Hill, 2014). Cela est particulièrement vrai si les fréquences alléliques sont déséquilibrées et/ou si le nombre de locus causaux est élevé.

Décomposer la variance génétique (variance additive, variance de dominance, variance épistatique) est utile dans un certain contexte : prédire l'amélioration d'une population végétale ou animale à cours termes (une ou quelques générations) dans le cadre d'un programme de sélection. Néanmoins, cela ne donne aucune information sur l'architecture génétique du trait étudié. Or, connaître l'architecture génétique est utile, par exemple pour les diagnostics médicaux, pour estimer l'évolution à long terme d'une population ou pour rechercher des combinaisons de gènes coadaptés dans le cadre de l'amélioration des plantes. L'épistasie peut être à l'origine d'un certain nombre de phénomènes importants d'un point de vue évolutif.

I.3.1. L'épistasie est quasiment omniprésente

De nombreuses études ont mis en évidence que l'épistasie joue un rôle important dans l'architecture génétique des caractères complexes (Flint and Mackay, 2009). Des QTL épistatiques ont pu être détectés sur un nombre très important d'espèces, tels que le poulet (Carlborg and Haley, 2004; Carlborg et al., 2003; Pettersson et al., 2011), la souris (Shimomura et al., 2001), la levure (Bloom et al., 2013) ou la drosophile (Anholt et al., 2003; Fedorowicz et al., 1998; Huang et al., 2012a). Chez les plantes, et notamment les poacées, on en a découvert chez le riz (Cao et al., 2001; Hu et al., 2004; Li et al., 1997; Yu et al., 1997), chez l'orge (Szűcs et al., 2007; Teulat et al., 1998; Xu and Jia, 2007) ou encore chez l'avoine (Wooten et al., 2008). Chez le petit épeautre, des interactions ont pu être détectées pour la vernalisation (Tranquilli and Dubcovsky, 2000). Chez le blé tendre, de l'épistasie a pu être détectée pour des traits tels que le rendement en grain (Goldringer et al., 1997), la précocité (Goldringer et al., 1997), la résistance à l'oïdium (Goldringer et al., 1997), la taille de la plante (Zhang et al., 2008) ou encore pour des caractères liés à la qualité du grain et de la farine (Carrillo et al., 1990; Rousset et al., 1992). Chez le blé dur, de nombreux caractères se sont également révélés épistatiques : des caractères morphologiques (Sayar et al., 2007; Sharma et al., 1998, 2003), des résistances à des maladies (Bnejdi and Gazzah, 2008; Haile et al., 2012; Somers et al., 2006), ou encore le rendement (Maccaferri et al., 2008; Sharma and Sain, 2004).

(Roff and Emerson, 2006) ont cherché à déterminer si certains types de traits phénotypiques étaient plus sujets à l'épistasie. Ils ont comparé les traits d'histoire de vie avec les traits morphologique. Les traits d'histoire de vie désignent les traits qui ont un lien direct avec la valeur sélective et le cycle reproductif : les traits phénologiques, le nombre de descendants, la durée du temps de reproduction, etc. Ils ont réussi a démontré que les effets épistatiques sont fréquemment détectés dans les deux types de traits, mais qu'ils sont tout de même significativement plus courants dans les traits d'histoire de vie. Environ 85% des traits d'histoire de vie présentaient des effets épistatiques significatifs, contre

74% des traits morphologiques. Cela est concordant avec le fait que, chez des lignées haploïdes doublées de blé tendre, une importante variance épistatique avait été trouvée pour le rendement en grains et la précocité, tandis que les traits morphologiques présentaient une part de variance additive importante (Goldringer et al., 1997). Il est donc probable que la valeur sélective soit elle-même fréquemment sujette à de l'épistasie.

I.3.2. L'épistasie peut contribuer à la variance additive

(Cheverud and Routman, 1995) ont été les premiers à démontrer que l'épistasie fonctionnelle peut contribuer à la variance additive : l'épistasie peut donc jouer un rôle dans l'adaptation d'une population en contribuant à sa variance additive (Barton and Turelli, 2004; Cheverud and Routman, 1995; Cheverud et al., 1999). La **Figure 4a** montre que, pour un trait physiologiquement soumis à l'épistasie additive x additive uniquement, la proportion de variance épistatique par rapport à la variance génétique totale, peut-être très faible si les fréquences alléliques sont déséquilibrées.

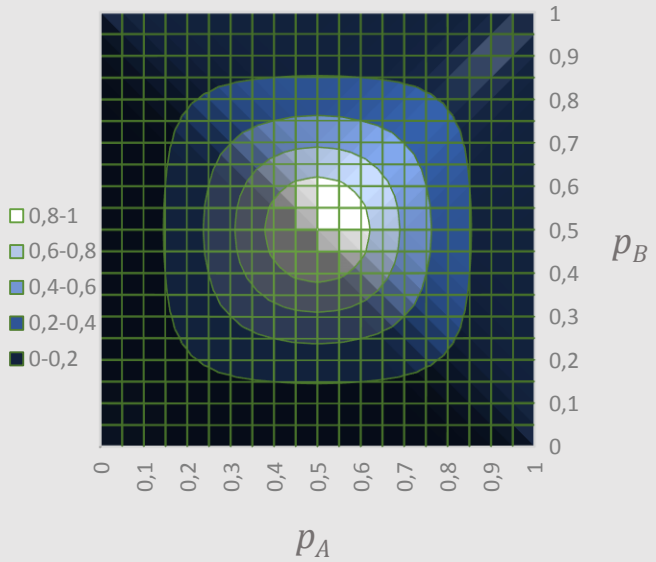


Figure 4a : Part de variance épistatique par rapport à la variance génétique totale (V_{aa}/V_G) en fonction des fréquences alléliques
Cas pour un modèle haploïde bilocus (locus A et locus B) biallélique. Ici $V_G = V_{aa} + V_{add}$ avec V_{add} la variance additive.
p_A = fréquence de l'allèle A au locus A
p_B = fréquence de l'allèle B au locus B
Déséquilibre de liaison nul
Valeurs génotypiques : G_{AB} = 1 ; G_{Ab} = - 1 ; G_{aB} = -1 ; G_{ab} = 1

Il a été démontré que, quand une population traverse un goulot d'étranglement, sa variance non-additive (notamment épistatique) peut être convertie en variance additive sous certaines conditions (Barton and Turelli, 2004; Cheverud and Routman, 1996; Cheverud et al., 1999; Goodnight, 1988; Lopez-Fanjul et al., 1999), ce qui peut augmenter sa capacité à évoluer à court terme. En présence de dérive, les fréquences alléliques s'éloignent de la valeur intermédiaire de 0,5 et deviennent plus

déséquilibrées (Falconer and Mackay, 1996; Luikart et al., 1998). Si les effets génétiques sont uniquement additifs, ce déséquilibrage des fréquences alléliques résultera en une réduction de la variance additive. Mais en présence d'épistasie, la variance additive peut parvenir à se maintenir, voire même à augmenter. Cela est dû au fait que quand les fréquences alléliques sont déséquilibrées, l'épistasie fonctionnelle est traduite en variance additive, comme le montre la figure 4a. Ceci est particulièrement vrai pour l'épistasie additive x additive. Ce phénomène de conversion est par ailleurs accentué en présence d'un grand nombre de locus (Naciri-Graven and Goudet, 2003), ou si la population est structurée en sous-populations (Whitlock et al., 1993).

L'augmentation ou le maintien de la variance additive suite à un goulot d'étranglement a été mise en évidence sur des données empiriques, notamment chez des souris (Cheverud and Routman, 1996; Cheverud et al., 1999), chez des mouches (Bryant et al., 1986; Heerwaarden et al., 2008) ou chez *Brassica rapa* (Briggs and Goldman, 2006).

I.3.3. L'épistasie peut contraindre les trajectoires évolutives d'une population

Les effets épistatiques peuvent jouer un rôle majeur dans la trajectoire évolutive d'une population. En effet, l'épistasie de signe réciproque (cf. I.1.3.3. Epistasie de magnitude ou de signe) peut générer des paysages adaptatifs rugueux (présentant des pics) (Poelwijk et al., 2011). Les paysages adaptatifs, développés par Wright, représentent la valeur sélective d'une population en fonction de sa composition génétique (Wright, 1932). Ils indiquent comment une population peut s'adapter, puisque celle-ci ne peut évoluer qu'en montant « en altitude », c'est-à-dire qu'en gagnant en valeur sélective. Un exemple est illustré en **Figure 4b**. L'épistasie de signe réciproque, en générant des pics et des vallées, peut donc contraindre l'évolution d'une population.

Il a été démontré que l'épistasie de signe et les paysages adaptifs rugueux n'étaient pas rares (Colegrave and Buckling, 2005; Lalić and Elena, 2012; Nahum et al., 2014; Poon and Chao, 2005; Schenk et al., 2013; Weinreich et al., 2005, 2006). Depuis longtemps, il a été suggéré que les phénomènes de radiation évolutive (évolution rapide d'un grand nombre d'espèces diversifiées, à partir d'un ancêtre commun) étaient dus à l'existence de paysages adaptatifs rugueux (Schluter, 2000; Simpson, 1984; Svensson and Calsbeek, 2012). Cela a été mis en évidence récemment chez des poissons du genre *Cyprinodon* (Martin and Wainwright, 2013).

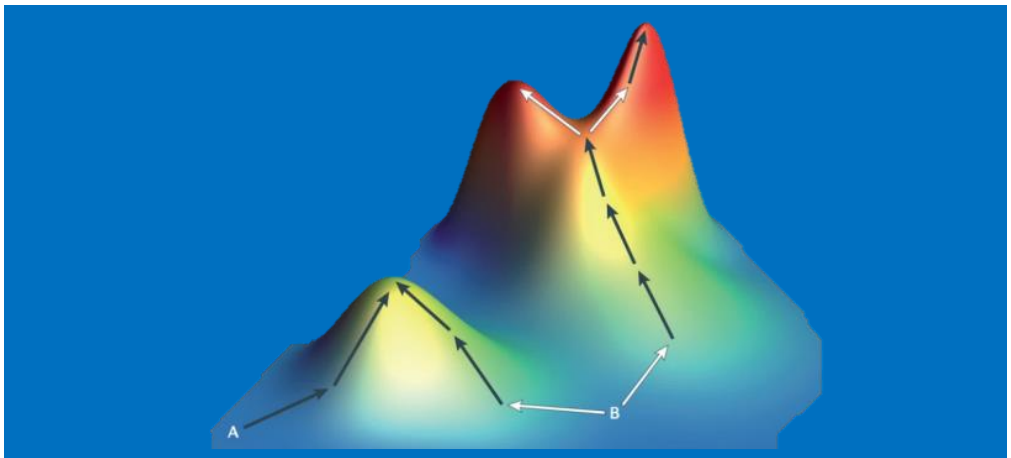


Figure 4b : Exemple de paysages adaptatifs (figure tirée de Palmer and Kishony, 2013)
Les flèches indiquent les chemins que peut emprunter une population à partir des points A et B.

I.3.4. L'épistasie peut générer de la dépression hybride et provoquer un isolement reproductif entre espèces

S'étant principalement intéressé aux forces qui façonnent les changements au sein des divers lignages (notamment la sélection), Darwin n'était pas parvenu a donné d'explication quant à l'isolement reproductif des espèces (Darwin, 1859). Comme il l'a admis, il n'était pas évident de comprendre comment la sélection naturelle avait pu provoquer la stérilité ou la létalité des génotypes hybrides.

L'explication est venue plus tard avec le modèle de Bateson-Dobzhansky-Müller (Bateson, 1909; Dobzhansky, 1934; Muller, 1939, 1940). Ils expliquèrent par quel mécanisme deux populations, descendant d'une même lignée, pouvaient ne plus se croiser, alors même qu'aucun gène ne provoquait de stérilité au sein de chacune des populations. Leur modèle est le suivant : deux populations proviennent d'une même lignée ; leur génotype d'origine, à deux locus A et B, est $AAbb$. Dans l'une des populations, la mutation a apparait au locus A, et est favorable. Les individus $aabb$ et $Aabb$ sont fertiles, et la mutation se fixe. Cette population se retrouve donc avec un génotype $aabb$. Dans l'autre population, une autre mutation favorable B apparait, cette fois-ci au locus B. Elle se fixe car, il n'y a pas de problème de fertilité avec les génotypes $AAbB$ et $AABB$. Cette population a donc pour génotype $AABB$. Le fait est que l'allèle B n'a pas pu « se confronter » à l'allèle a, avant d'être fixé. Il est donc possible que B et a soient incompatibles, et que leur combinaison ait une plus faible valeur sélective, voire soit stérile ou létale. Il s'agit donc d'épistasie de signe réciproque (cf. I.1.3.3. Epistasie de magnitude ou de signe).

Dans les cas les moins extrêmes, l'épistasie de signe provoque simplement de la dépression hybride, c'est-à-dire que le génotype hybride est moins performant que ses parents, sans qu'il soit pour autant stérile ou létal (Lynch, 1991). Dans certains cas, la dépression hybride n'apparait pas immédiatement,

et l'hybride F1 peut même surpasser ses parents en raison de l'hétérosis. C'est dans les générations suivantes que la dépression hybride se manifeste, une fois que l'effet d'heterosis disparait. Quand l'épistasie de signe réciproque devient extrême (létalité ou stérilité des hybrides), elle explique l'isolement reproductif complet d'espèces issues d'une même lignée. Dans tous les cas, elle engendre une vallée adaptative sur le paysage des valeurs sélectives (cf I.3.3. L'épistasie peut contraindre les trajectoires évolutives d'une population).

Ce phénomène a été très étudié chez des espèces animales, comme la drosophile (Brideau et al., 2006; Cabot et al., 1994; Orr and Irving, 2001; Palopoli and Wu, 1994; Wu and Beckenbach, 1983), mais a pu être également mis en évidence chez des plantes. Chez le riz (*Oryza sativa* L.), des cas de stérilité ou de dépression hybride, dus à des combinaisons alléliques incompatibles, sont fréquemment observés quand les deux sous-espèces, *indica* et *japonica*, sont croisées (Kubo et al., 2008, 2011; Li et al., 1997b). Des incompatibilités de Dobzhansky-Muller serait également à l'origine de l'isolement reproductif entre des espèces de coton (Song et al., 2009), des espèces d'eucalyptus (Costa e Silva et al., 2012), des espèces d'iris (Taylor et al., 2009), des espèces de légumineuses (Fenster and Galloway, 2000), et entre des espèce du genre *Mimulus* (Fishman and Willis, 2001). Mais l'un des exemples les plus parlants est la nécrose de certains descendants provenant de croisements entre différentes souches d'*Arabidopsis thaliana* (Bikard et al., 2009; Bomblies et al., 2007). Cette nécrose est due à l'activation inadaptée du système immunitaire de ces plantes, provoquée par des combinaisons nocives d'allèles. Des phénomènes de nécroses dus à l'épistasie ont également été observés entre deux cultivars de blé tendre (Hermsen, 1963a, 1963b; Heyne et al., 1943; Iehisa et al., 2014), et chez bien d'autres espèces végétales (Bomblies and Weigel, 2007).

L'accumulation de différences en termes de combinaisons épistatiques peut apparaître dans plusieurs cas. Elle peut être provoquée par l'isolement des populations l'une de l'autre pendant une longue période (Orr, 1995). C'est par exemple le cas de populations de moustiques de l'espèce *Wyeomyia smithii* aux Etats-Unis (Hard et al., 1992; Lair et al., 1997). Cette accumulation peut aussi être causée par un goulot d'étranglement (Fenster et al., 1997) ou par une adaptation à de nouvelles conditions environnemental (Dettman et al., 2007).

I.4. L'épistasie chez les autogames

On retrouverait plus de combinaisons alléliques coadaptées dans les espèces autogames, qui seraient donc plus sujettes à de l'épistasie de signe réciproque et à des incompatibilités de Dobzhansky-Müller (cf. I.3.4). Il existe plusieurs raisons à ceci. Premièrement, l'autofécondation réduit la recombinaison efficace (la recombinaison entre régions chromosomiques différentes, résultant en un brassage génétique) (Nordborg, 2000; Pollak, 1987). Par conséquent, les combinaisons alléliques sont plus

facilement maintenues. La sélection peut donc agir plus directement sur ces combinaisons de gènes, comme si une combinaison de gènes n'était qu'un seul gène (on parle parfois de « super gènes » dans ce cadre-là) (Allard et al., 1993; Fenster et al., 1997; Nordborg et al., 2002; Pollak and Sabran, 1999). L'autre raison est que chez les espèces autogames, le flux d'allèles est réduit : les populations autogames sont par conséquent généralement plus différenciées entre elles, que les populations allogames (Charlesworth and Pannell, 2001; Grundt et al., 2006; Hamrick and Godt, 1996). Cela favorise donc l'isolement entre populations, qui est l'un des facteurs d'apparition d'incompatibilités de Dobzhansky-Müller (Templeton, 1986). Les populations fortement autogames sont souvent composées d'un ensemble de génotypes multilocus, c'est-à-dire des génotypes marqués par des combinaisons alléliques bien particulières (Allard et al., 1993; Selander and Kaufman, 1973), qui correspondent à des pics sur le paysage adaptatif. Enfin, le fait que l'autogamie favorise la spéciation corrobore l'idée selon laquelle elle permet le développement de complexes de gènes coadaptés, générant des vallées adaptatives (Levin, 1971; Martin and Willis, 2007; Wendt et al., 2002).

Bien que toutes les espèces autogames ne présentent pas de forts déterminismes génétiques liés à l'épistasie, il semblerait néanmoins qu'elle soit plus fréquemment détectée chez ces espèces, que chez les espèces allogames (Holland, J. B., 2001). De la même façon, la dépression hybride y est plus fréquemment observée (Edmands and Timmerman, 2003; Frankham et al., 2011; Lynch, 1991). L'effet de la dépression hybride a notamment été étudié dans une population d'orge sauvage, très structurée par l'autofécondation (Volis et al., 2011). Il a été démontré que le croisement entre individus éloignés, spatialement et génétiquement, pouvait aboutir à une réduction de la valeur sélective, via probablement le cassage de complexes de gènes coadaptés. Dans une étude portant sur trois espèces autogames de Brassicacées arctiques, l'accumulation d'incompatibilités et l'isolement reproductif, ont pu être constatés (Grundt et al., 2006). Ces espèces sont relativement peu différenciées génétiquement, ce qui indique une radiation récente. Pour expliquer leur isolement reproductif malgré leur origine récente, les auteurs mettent en avant le régime autogame de ces espèces, pour expliquer l'accumulation d'incompatibilités hybrides. Enfin, les souches pures de l'espèce de nématode autogame *Caenorhabditis elegans* (autogame) sont plus performantes que les croisements (Dolgin et al., 2007). Cette dépression hybride n'est pas visible chez l'espèce apparentée allogame *Caenorhabditis remanei.*

I.5. L'épistasie et la polyploïdie : l'hétérosis fixé

La polyploïdisation est le phénomène par lequel le génome est dupliqué. Les espèces autopolyploïdes possèdent deux (ou plusieurs) génomes dérivant d'une même espèce. Les génomes des espèces allopolyploïdes, dérivent quant à eux d'espèces différentes : les allopolyploïdes sont donc issus d'un

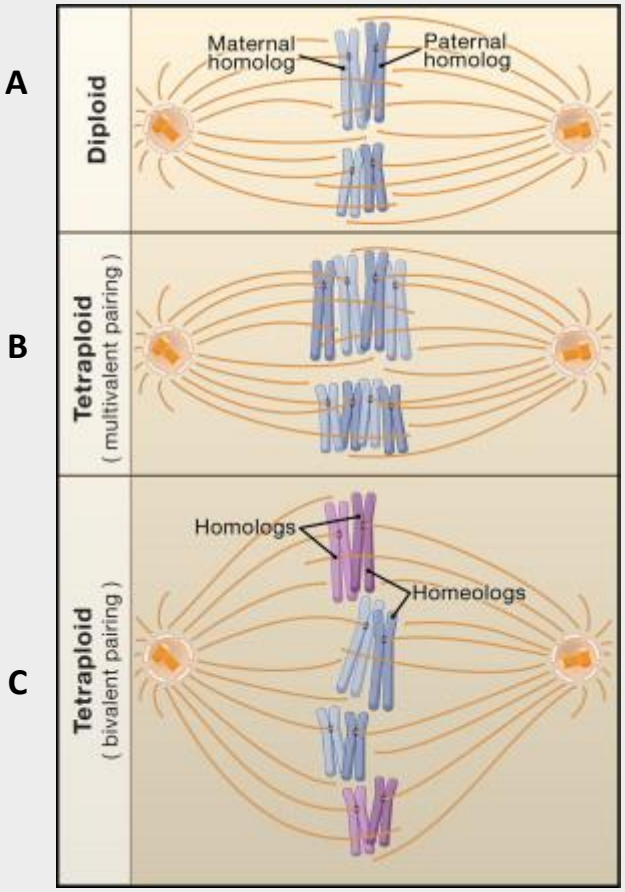


Figure 5 : Représentation de la méiose chez les diploïdes et polyploïdes (tirée de Otto, 2007)

(A) Diploïde. Deux paires de chromosomes sont représentés : une courte et une longue. Chaque couleur (bleu clair et bleu foncé) correspond à un parent. Les deux paires sont distinguées lors de la méiose.

(B) Chez un autotétraploïde, les chromosomes sont dupliqués. Les chromosomes dupliqués sont similaires, et il n'est pas possible de distinguer ceux provenant de la mère de deux provenant du père. Par conséquent, ils se regroupent aléatoirement lors de la méiose, soit par paire (non montré), soit tous ensemble. Dans tous les cas, les gamètes peuvent hériter de n'importe quelle paire de chromosome (deux du père, deux de la mère, un de la mère et l'autre du père, …) : l'hérédité est multisomique.

(C) Chez les allotétraploïdes, les paires parentales (= homéologues) sont bien distinguées. Les paires bleues proviennent du premier parent (bleu clair = chromosome hérité de la grand-mère, bleu foncé = chromosome hérité du grand-père). Les paires violettes du deuxième parent, d'une autre espèce (violet clair = grand-mère, violet foncé = grand-père). Les gamètes hériteront d'un chromosome paternel et d'un chromosome maternel. L'hérédité est disomique.

phénomène d'hybridation. La **figure 5** illustre la méiose chez les diploïdes et les polyploïdes. Chez la plupart des autopolyploïdes, il n'est pas possible de distinguer, parmi les 4 chromosomes homologues, ceux provenant du premier parent de ceux hérité de l'autre parent. En effet, ces quatre chromosomes proviennent de la même espèce. Lors de la méiose, les quatre chromosomes peuvent alors former des appariements bivalents ou quadrivalents, de manière aléatoire (Otto, 2007; Ramsey and Schemske, 2002). On parle d'hérédité multisomique (Figure 5.A).

En revanche, chez les allopolyploïdes, les deux paires de chromosomes homéologues (héritées de chacun des deux parents) sont clairement différenciées. Lors de la méiose, chaque chromosome formera alors un appariement bivalent avec son homologue (Otto, 2007; Ramsey and Schemske, 2002). L'hérédité, identique à celle des diploïdes, est dite disomique (Figure 5.A).

La polyploïdie disomique (donc, chez les allopolyploïdes, principalement) pourrait être un moyen de compenser les interactions entre homologues (dominance) par des interactions alléliques entre homéologues, chez les plantes autogames (Mac Key, 1970). Par conséquent, l'épistasie serait moins forte chez des polyploïdes multisomiques comme la luzerne tétraploïde (*Medicago sativa* L.). Des études portant sur l'espèce herbacée autogame tétraploïde *Avena barbata* ont démontré que la polyploïdisation permettait de fixer l'hétérosis (Allard et al., 1993; García et al., 1991). *Avena barbata* est une plante autotraploïde, provenant de l'hybridation entre deux écotypes d'une même espèce diploïde, nommés *Avena hirtula* (écotype méditerranéen) et *Avena wiestii* (écotype du désert). Comme les écotypes sont bien différenciés, les paires d'homéologues sont distinguées lors de la méiose, et son hérédité est disomique. Cette fixation de l'hétérosis rend *Avena barbata* plus performante que ses parents diploïdes. Le principe de l'hétérosis fixée, basé sur l'exemple d'*Avena barbata*, est illustré **figure 6**.

I.6. L'épistasie dans la régulation de l'expression génétiques

Les variations du niveau d'expression des gènes, dépendent d'un ensemble d'interactions dynamiques : interactions entre facteurs environnementaux, interactions entre cellules, interactions génétiques (épistasie), sans oublier les interactions génotype x environnement. Le niveau d'expression est donc un caractère très complexe.

La composante génétique de la variation est due aux différents effets régulateurs génétiques de l'expression. Les éléments de régulations *cis* (Cis Regulatory Elements - CREs) sont des régions d'ADN non codantes, situées à proximité des gènes qu'elles régulent, souvent dans la région promotrices de ceux-ci. Leur effet sur l'expression d'un gène voisin est appelé « effet *cis* » (Cowles et al., 2002). Des variations dans leur séquence peuvent entraîner des variations dans l'expression du gène voisin

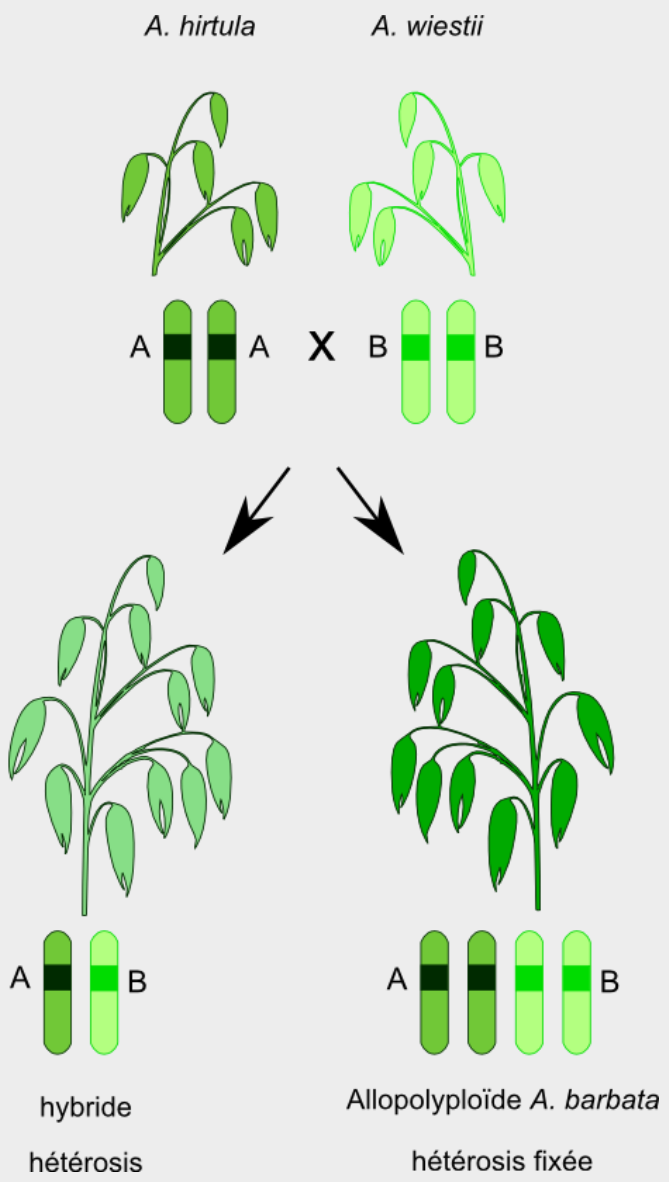


Figure 6 : L'hétérosis fixée grâce à la polyploïdie. Exemple d'*Avena barbata*

Les parents diploïdes (*Avena hirtula* et *Avena wiestii)*, autogames, sont deux écotypes d'une même espèce. Pour un locus donné, chacun a fixé un allèle (allèle A et allèle B). L'hybride « simple » (non polyploïde) porte le génotype AB, et est plus performant que ses parents (hétérosis). Néanmoins, cet effet d'hétérosis disparaîtra progressivement, puisque chaque génération d'autofécondation réduira de moitié le nombre d'hétérozygotes. L'hybride polyploïde, *Avena barbata*, porte le génotype AABB. Il est également sujet au phénomène de l'hétérosis. Néanmoins, chaque gamète qu'il produira portera le génotype hybride AB, car la transmission des chromosomes est disomique. L'hétérosis est donc fixée, et sera transmise à chaque descendant.

qu'elles régulent. Les éléments régulateurs en *trans* régulent l'expression de gènes à distance, en codant pour des protéines impliquées dans la régulation de l'expression et dans la dégradation de l'ARN : des facteurs de transcription principalement, mais aussi des ARN hélicases ou des ribozymes (Chesler et al., 2005). Des variations dans ces séquences codantes peut modifier ces protéines, et donc impacter leur liaison à l'ADN et leur action, et par conséquent l'expression du gène-cible. Des protéines, ne se liant pas à l'ADN, mais ayant un impact indirect sur l'expression d'autres gènes, via des cascades moléculaires complexes, peuvent également être à l'origine d'un effet *trans*. Par exemple, le polymorphisme des gènes codant pour des canaux calciques a diverses répercussions sur des cibles transcriptionnelles neurales, en contrôlant la concentration en calcium (Bading et al., 1993).

Il a été démontré que l'effet de l'expression d'un ou plusieurs gènes sur un autre gène (effets *trans* au sein d'un réseau de gène) pouvait se traduire par de l'épistasie sur le niveau d'expression de l'un de ces gènes. Dans une étude, plusieurs types de réseaux (à trois gènes) ont été simulés, et l'effet de l'épistasie sur le niveau d'expression de l'un des gènes, a été estimé (Gjuvsland et al., 2007). Le type de réseau générant le plus d'épistasie sont ceux qui constituent une rétroaction positive (**Figure 7**). Ce type de réseau de gènes permet l'existence de plusieurs états d'équilibre (Plahte et al., 1995; Thomas and Richelle, 1988). Ces systèmes permettant plusieurs états stables (multistabilité) sont d'une grande importance dans les systèmes biologiques, permettant aux cellules et aux organismes de passer d'un état discret à un autre (Angeli et al., 2004). Cette multistabilité serait par exemple à l'origine de la différenciation cellulaire (Thomas and Kaufman, 2001).

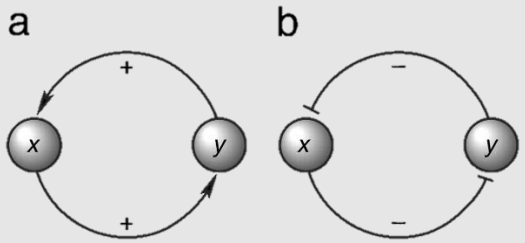


Figure 7 : réseaux de gènes à deux locus (*x* et *y*), générant une rétroaction positive (tirée de Angeli et al., 2004)

a. Une boucle avec une rétroaction mutuellement activatrice. L'expression du gène *y* active l'expression de *x*, qui elle-même active celle de *y*.
b. Une boucle avec une rétroaction mutuellement inhibitrice L'expression du gène *y* inhibe l'expression de *x*, qui elle-même inhibe celle de *y*.

Notons que chez les espèces de drosophiles, les incompatibilités de Dobzansky-Müller sont causées en partie par une expression inadéquate de certains gènes (Haerty and Singh, 2006). Autrement dit,

les combinaisons de gènes coadaptés du modèle de Bateson-Dobzansky-Müller sont probablement des réseaux de gènes régulant l'expression l'un de l'autre.

Sachant cela, il est donc intéressant d'estimer les effets génétiques régulant le niveau d'expression. Les effets *cis* correspondent aux effets additifs des SNPs sur l'expression du gène auquel ils appartiennent. Les effets *trans* peuvent être détectés via les effets additifs des SNPs sur l'expression de gènes auquel ils n'appartiennent pas. La détection d'épistasie peut être également le signe de régulations *trans*, et notamment de rétroactions positives.

I.7. L'épistasie et la gestion des ressources génétiques

Face au changement climatique et à d'autres enjeux environnementaux, l'amélioration variétale doit être capable de créer des variétés pouvant générer des rendements toujours aussi élevés dans de nouvelles conditions. Dans ce contexte, il est important de maintenir la capacité des espèces cultivées à évoluer en réponse à des modifications de leur environnement et à des sélections anthropiques. Or, chez beaucoup d'espèces, la domestication s'est traduite par une réduction de la diversité génétique comparativement à l'ancêtre sauvage, du fait d'un goulot d'étranglement (Buckler et al., 2001). L'utilisation quasi-exclusive de variétés élites depuis la révolution agricole a encore accentué cette perte de diversité chez les espèces cultivées (Haudry et al., 2007; Liu and Burke, 2006; Roussel et al., 2004, 2005; Thuillet et al., 2005; Wright et al., 2005). Mais ce n'est qu'à partir des années 1970 qu'elle a été perçue comme problématique, car réduisant le potentiel adaptatif de ces espèces. Depuis quelques décennies, la conservation des ressources génétiques est donc devenue un enjeu international (Hoisington et al., 1999).

Cette conservation de la diversité peut s'effectuer de différentes manières. La conservation ex-situ consiste en la conservation d'individus en dehors de leur habitat naturel (Frankel, 1995). Il s'agit le plus souvent de stocks de semences, mais la conservation ex-situ comprend également les stocks de pollen et de plantules in-vitro. Les vergers de conservation et les jardins botaniques sont aussi des formes de conservation ex-situ (Engelmann and Engels, 2002). Il s'agit donc d'une conservation dont l'objectif est de maintenir des génotypes existants. La diversité génétique peut également être conservée in situ, c'est-à-dire dans une réserve naturelle, autrement dit dans un écosystème en interaction avec d'autres espèces et avec des contraintes abiotiques et biotiques. Cette conservation, plus dynamique, a l'intérêt de maintenir les capacités d'évolution et d'adaptation des espèces concernées face aux changements de l'environnement (Engelmann and Engels, 2002; Enjalbert et al., 1998).

Plus récemment, l'idée d'une gestion dynamique des ressources génétiques a émergé : ce mode de conservation repose sur des populations artificielles, à base génétique large, cultivées en plein champs

et en interaction avec leur milieu. L'action des forces évolutives sur ces populations est donc maintenue (Rhoné et al., 2008). Contrairement à la conservation in-situ, ces populations sont gérées, c'est-à-dire que certains paramètres, tels que, par exemple, le taux d'allofécondation ou la taille de la population, sont contrôlés. Ces populations peuvent également servir au pre-breeding, c'est-à-dire à l'amélioration de plantes cultivées en utilisant des ressources génétiques sauvages ou provenant d'anciennes variétés (Goldringer et al., 2001). De populations de gestion dynamique de de blé tendre se sont montrées capables de s'adapter à divers stress, avec le développement de résistances à l'oïdium (Paillard et al., 2000a, 2000b) ou l'évolution de la précocité (Rhoné et al., 2008, 2010). Chez l'orge, des augmentations de rendement ont pu être obtenu dans un environnement compétitif (Suneson, 1956).

Le régime de reproduction est d'une grande importance pour la gestion dynamique (Enjalbert et al., 1998). La recombinaison n'est efficace que si elle a lieu entre des régions chromosomiques différentes. L'hétérozygotie – et donc l'allofécondation – augmentent donc le taux de recombinaison efficace (Pollak, 1987). A priori, un fort taux de recombinaison efficace présente plusieurs intérêts pour la capacité à s'adapter. Premièrement, la recombinaison facile le brassage génétique, en réduisant le déséquilibre de liaison. Quand la recombinaison est faible, on observe que de larges régions chromosomiques à fort déséquilibre de liaison entourent les allèles introgressées dans de nouvelles lignées. Une recombinaison efficace fractionne ces blocs et permet des introgressions plus ciblées (Feuillet et al., 2008). Dans le même ordre d'idée, la recombinaison réduit l'effet du fond génétique (Rice, 2002). En présence de recombinaison, une mutation favorable apparue dans un mauvais fond génétique, aura plus de chance d'être maintenue, puisqu'il y a plus de chance qu'elle soit transmise à des fonds génétiques moins mauvais. A l'inverse, les mutations délétères apparues dans de bons fonds génétiques pourront être plus facilement éliminées. De plus, la recombinaison réduit la perte de polymorphisme autour des locus sélectionnés, et par conséquent augmente le polymorphisme de ces régions (Marais and Charlesworth, 2003; Orr, 2005). Enfin, augmenter le taux d'allofécondation pourrait avoir un autre intérêt dans les populations de gestion dynamique : les espèces autogames ont une taille efficace plus petite que les espèces allogames, et par conséquent la diversité génétique y est plus rapidement perdue (Caballero and Hill, 1992). Introduire de l'allofécondation peut alors permettre un meilleur maintien de la diversité génétique. Cela a notamment été démontré sur des populations d'*Arabidopsis thaliana* (Porcher et al., 2004).

D'un autre côté, la recombinaison peut également être désavantageuse si elle casse des combinaisons de gènes coadaptées, et provoquent de la dépression hybride, notamment chez des espèces autogames (Costa e Silva et al., 2012; Lynch, 1991). Il est donc intéressant de déterminer quels effets

peut avoir la recombinaison en présence de complexes de gènes coadaptés (épistasie de signe réciproque).

Il est possible de gérer le taux d'allofécondation dans une population à base génétique large d'une espèce naturellement autogame, en y introduisant un gène de stérilité mâle (Goldringer et al., 2001; Jain and Suneson, 1963, 1966; Porcher et al., 2004). En effet, les graines récoltées sur des individus mâles stériles seront nécessairement issus d'allofécondation. Pour le semis annuel, il est donc possible de fixer une proportion de graines qui seront récoltées sur les mâles stériles et donc nées par allofécondation.

I.8. Objectifs de la thèse

Nous avons voulu approfondir le rôle de l'épistasie sur l'évolution d'une population et vérifier son importance dans l'expression de gènes homéologues d'une plante autogame allotétraploïde. Ces travaux ont comme finalité d'établir des outils et des connaissances utiles à une meilleure valorisation de la diversité génétique contenues dans les compartiments sauvages et primitifs des espèces des plantes cultivées. L'approche abordée ici est celle de l' « evolutionary pre breeding », à savoir la constitution de populations composites, soumises à une double sélection naturelle et anthropique.

Deux points nous ont plus particulièrement intéressés :

- La fixation de l'hétérosis chez une espèce autogame allopolyploïde. Dans la vision traditionnelle de l'hétérosis fixée, l'effet d'interaction entre les allèles des deux locus homéologues est fixé, c'est-à-dire qu'on ne considère pas qu'il puisse y avoir de polymorphisme aux locus homéologues. La variation de ces effets n'est donc pas mentionnée. Or, il peut exister un polymorphisme au niveau des gènes homéologues, et donc une variation de l'effet d'interaction entre ces deux locus (épistasie). Cette variation permet l'estimation de ces effets par des modèles de génétique quantitative. Si les interactions entre allèles homéologues (épistasie additive x additive) peuvent permettre de remplacer dans des espèces autogames les interactions avantageuses à l'état hétérozygotes entre allèles homologues (dominance détectées dans les espèces diploïdes allogames), alors pour les paires de locus polymorphes, on pourrait s'attendre à détecter plus d'épistasie entre gènes homéologues, qu'entre des gènes pris au hasard. Cette prédiction sera testée sur des données d'expression de gènes, générées sur une population pilote de pre-breeding de blé dur.
- L'effet de la recombinaison sur la sélection d'un trait soumis à de l'épistasie de signe réciproque. Nous nous sommes situés dans le cadre du pre-breeding, où une faible sélection anthropique est effectuée : généralement, une petite proportion d'individus présentant des phénotypes particulièrement peu favorables à la culture (sensibles aux malades, à la verse)

sont éliminés chaque année. Nous avons donc modélisé la sélection d'une population de pre-breeding comme étant le fait de garder chaque année une proportion p d'individus. Il s'agit donc d'un type particulier de sélection par troncation.

Pour répondre à ces deux objectifs, la thèse a été divisée en trois grandes parties :

1) Dans un premier temps, la bibliographie sur les modèles d'estimation des effets génétiques a été effectuée, afin de déterminer les plus adéquats en fonction des contextes, notamment celui de notre étude (une population d'une espèce autogame polyploïde, avec vraisemblablement du déséquilibre de liaison). Certains modèles ont été réécrits si besoin. Nous avons essayé d'illustrer les choix des différents modèles pour vulgariser des approches qui sont encore peu diffusées dans la communauté des généticiens s'intéressant aux analyses de GWAS. Nous avons aussi contribué à améliorer ces modèles sur différents aspects, notamment une meilleure prise en compte du déséquilibre de liaison entre locus pour estimer les variances génétiques. Ces modèles ont des philosophies différentes, et nous avons tâché de les mettre en évidence et à la portée du lecteur. Cette approche nous a permis de faire un choix pertinent et argumenté pour l'approche empirique menée sur la recherche d'effets épistatique dans la variation d'expression des gènes en situation homéologues, chez une espèce autogame polyploïde, le blé dur.
2) Nous avons donc utilisé l'un de ces modèles sur des données d'expression du blé dur, afin de détecter des effets régulateurs, et notamment de l'épistasie entre gènes homéologues.
    a) Pour ce faire, nous avons d'abord caractérisé une population de blé dur qui a été constituée par croisement entre une population ségrégant pour un gène de stérilité mâle génique, et des accessions issues de taxons sauvages et primitifs de l'espèce *Triticum turgidum*. Cette population ayant une structure particulière, nous avons analysé sa structure à l'aide d'une analyse en composantes factorielles (AFC) et d'une méthode de classification (dendrogramme). Cette étape, nous a permis de constituer un échantillon adapté de 177 lignées de blé dur.
    b) Par ailleurs, nous avons développé un modèle permettant de comprendre l'origine de cette structure. Plus exactement, nous avons voulu identifier des individus descendants de lignées migrantes. Pour cela, nous avons utilisé des formalismes basés sur des matrices de transition. Ce modèle a servi ici à comprendre l'origine de la structure de la population, mais peut également permettre de comprendre comment des migrants peuvent être intégrés dans une population, en fonction du régime de reproduction.

c) L'échantillon de 177 lignées de blé dur, constitué grâce à l'étape 2.a), a ensuite servi de base à une étude originale d'eQTL, utilisant les nouvelles technologies de séquençage (NGS) pour générer une importante masse de données d'expression en RNAseq. Cette analyse est une première pour le blé dur et probablement pour les espèces polyploïdes.

3) Enfin, nous avons modélisé la sélection d'un caractère épistatique (épistasie de signe réciproque) dans une population de pre-breeding, afin de tester l'effet de la recombinaison en présence de complexes de gènes coadaptés.

I.9. Notre modèle d'étude : le blé dur

Le matériel végétal d'étude de cette thèse a été le blé dur (*Triticum turgidum* ssp. *durum*). Il s'agit d'une espèce de blé tétraploïde d'intérêt économique, cultivée à grande échelle dans le bassin méditerranéen. Cependant, elle ne représente que 5 à 10% des surfaces cultivées en blé au niveau mondial, le reste étant emblavé en blé tendre (*T. aestivum*). Le blé dur est utilisé uniquement en consommation humaine, pour la confection de pâtes, de semoule et d'autres produits typiques de la culture méditerranéenne (boulgour, baklava, crêpes mille trous, etc.). A l'heure actuelle, le pourtour méditerranéen constitue la principale région productrice, et représente à l'échelle mondiale, plus de la moitié de la surface cultivée en blé dur (International Grains Council, 2012). La production mondiale s'élève à 35,1 millions de tonnes (FranceAgriMer, 2013). En Europe, la France est le deuxième producteur après l'Italie, avec une production de 2,4 millions de tonnes par an et un rendement moyen de 54,1 quintaux par hectare.

Pour notre étude, le blé dur correspond bien au modèle attendu, étant une céréale majoritairement autogame (environ 96-99% d'autofécondation (Tsegaye, 1996)) et allotétraploïde. Son génome, de 28 chromosomes, est constitué de deux sous-génomes de 14 chromosomes chacun, les génomes A et B. Il partage ces génomes avec le blé tendre, une espèce hexaploïde possédant un sous-génome supplémentaire, le génome D (Kihara, 1944; Mcfadden and Sears, 1946).

L'ancêtre sauvage du blé dur, *T. t. dicoccoides* provient de l'hybridation spontanée entre deux espèces diploïdes, à une période située entre 500 000 ans et 1 M d'années dans le passé (Huang et al., 2002). L'ancêtre de l'espèce contemporaine *Triticum urartu* est à l'origine du génome A (Dvořák et al., 1988, 1993; Huang et al., 2002; Petersen et al., 2006). Cette espèce est aujourd'hui très autogame. Le génome B proviendrait d'un apparenté de l'actuel *Aegilops speltoides* (Daud and Gustafson, 1996; Kilian et al., 2007; Petersen et al., 2006), qui devait posséder comme lui une auto-incompatibilité très forte faisant d'*Ae. speltoides* une espèce allogame obligatoire.

L'amidonnier sauvage fut l'une des premières plantes à avoir été domestiquée, il y a environ 10 000 ans dans la région du volcan Karaca dag, en Turquie (Kilian et al., 2009; Nesbitt and Samuel, 1998; Özkan et al., 2002). Cette domestication aboutit à une première forme domestique, l'amidonnier domestique (*T. turgidum spp. dicoccum*), qui garda quelques caractères primitifs de *T. t. dicoccoides*, comme des glumes indurées, un grain assez long qui reste enchâssé dans ces glumes dures (appelé grain vêtu, improprement car la glumelle n'est pas adhérente comme chez l'orge) et un faible nombre de fleurs par épillet. L'amidonnier domestique se caractérise également par un rachis non cassant et une floraison uniforme (Charmet, 2011).

L'amidonnier domestique fut rapidement remplacé, il y a 7000 ans environ, par des formes plus modernes, qui constituent réellement la base du blé dur actuel, *T. t.* ssp *durum*. Ces formes ont des glumes fines et des grains facilement extractibles, d'où leur dénomination de « grain nus » (Nesbitt and Samuel, 1998). L'histoire évolutive du blé dur a pris ensuite un nouveau tournant dans les années 1960, avec l'introgression de gènes de nanisme dans le blé dur, via des croisements avec le blé tendre. Cela engendra une nouvelle perte de diversité dans le compartiment élite (Haudry et al., 2007; Thuillet et al., 2005). Les variétés modernes de blé dur sont toutes issues de ces croisements.

Dans la **figure 8**, l'analyse de structuration et de diversité de *T. turgidum* réalisée au laboratoire (David et al., in press) (en **Annexe VII.2**) montre que le blé dur élite (groupe 4 sur la figure) a un faible niveau de diversité, et qu'il est très distant des formes primitives ou sauvages (groupes 6 et 3) à fort niveau de diversité.

La population d'étude de cette thèse est une population créée par hybridation entre ces fonds génétique (voir III.1). La remise en contact de fonds génétiques séparés depuis longtemps dans l'Histoire pose donc des questions intéressantes sur l'importance de l'épistasie dans le fonctionnement d'une population hybride. Elle donne à s'interroger sur les conséquences de son éventuelle contribution aux valeurs sélectives et agronomiques de telles populations, et sur l'utilisation des ressources génétiques du blé dur.

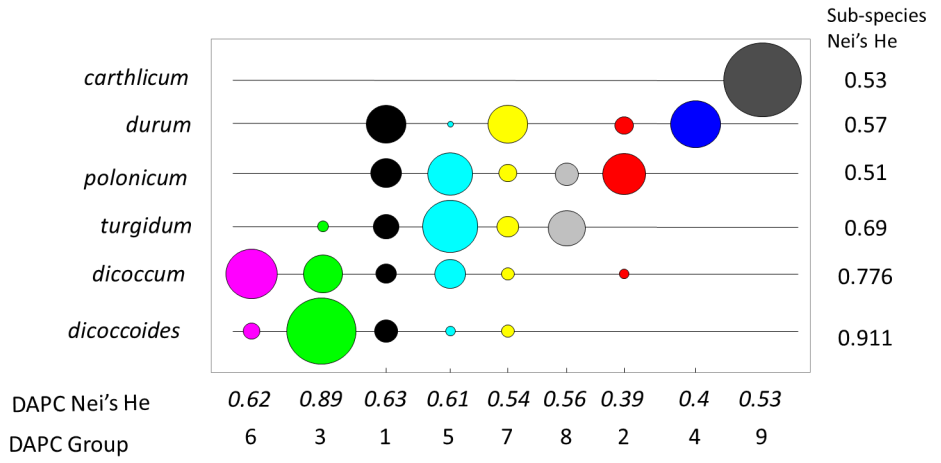


Figure 8. Distribution des taxons de 492 accessions de *Triticum turgidum* spp. au sein de 9 groupes obtenus avec une analyse discriminante en composantes principales (DAPC), réalisée avec 14 marqueurs microsatellites (tirée de David et al., in press).
En abscisse : le numéro de chaque groupe de DAPC, et l'hétérozygotie attendue de Nei (Nei, 1978) de chacun des groupes (mesure de diversité).
En ordonnées : à gauche, les différents taxons ; à droite, l'hétérozygotie attendue de chacun des taxons.
La surface de chaque cercle correspond à la proportion de chaque groupe au sein d'un taxon.

PARTIE II

ESTIMATION DE L'EPISTASIE DANS UN MODELE BILOCUS HAPLOÏDE

II. ESTIMATION DE L'EPISTASIE DANS UN MODELE BILOCUS HAPLOÏDE

Pour pouvoir prédire quel sera l'effet de l'épistasie sur l'évolution d'une population, il est nécessaire de pouvoir l'estimer correctement, y compris en présence de déséquilibres (déséquilibre de Hardy Weinberg, déséquilibre de liaison). Dans ce chapitre, plusieurs modèles d'estimation des effets génétiques seront examinés. Il s'agit de modèles de génétique quantitative, où l'épistasie est définie comme étant l'interaction entre gènes. La valeur génotypique (la valeur phénotypique qu'aurait un génotype en l'absence d'effets environnementaux) y est modélisée comme une régression linéaire. Ces modèles permettent d'estimer les effets additifs, de dominance et épistatiques, pour un caractère phénotypique donné. Ils sont tous bi-alléliques.

En préambule (section II.1), nous reviendrons rapidement sur une propriété importante des modèles de génétique quantitative : l'orthogonalité. Tous les modèles ne sont pas orthogonaux et cela peut poser quelques problèmes lors de l'estimation des effets génétiques.

Avant de présenter les modèles qui permettent d'estimer des effets d'épistasie, nous présenterons quelques modèles dans le cas unilocus diploïde (section II.2). Nous avons choisi de les présenter sous cette forme (unilocus diploïde) car il s'agit d'une des formes les plus basiques et celle sous laquelle ces modèles sont les plus fréquemment présentés. La forme unilocus diploïde permet d'estimer l'effet additif du locus étudié, et l'effet d'interaction entre ses deux allèles (dominance). Nous montrerons ensuite comment généraliser ce type de modèles à plusieurs locus, en l'absence de déséquilibre de liaison (section II.3).

Enfin, nous écrirons ces modèles sous forme bilocus haploïde (section II.4). Notre espèce d'étude est le blé tétraploïde, *Triticum turgidum,* autogame, et présentant par conséquent un fort taux d'homozygotie. Nous avons donc supposé que les effets de dominance, ainsi que les effets épistatiques incluant de la dominance, ne jouaient pas un rôle important dans cette espèce, les phases hétérozygotes créées par l'allogamie résiduelle étant transitoires dans une population. Nous acceptons que ces effets puissent être importants en valeur absolue, mais nous avons présumé que leur rôle dans l'adaptation des populations autogames était marginal. Nous avons donc cherché à estimer principalement les effets additifs et les effets de l'épistasie additive x additive dans une population de lignées fixées, que permettent d'estimer les modèles bilocus haploïdes.

II.1. Préambule : l'orthogonalité, une propriété importante des modèles d'estimation des effets génétiques

Plusieurs auteurs ont mis en avant l'importance de l'orthogonalité des modèles d'estimation (Álvarez-Castro and Carlborg, 2007; Kao and Zeng, 2002; Yang, 2004; Zeng et al., 2005). Un modèle est orthogonal, si les covariances entre les variables explicatives (ici, les effets génétiques) sont nulles.

L'orthogonalité est importante pour plusieurs raisons. Premièrement, l'estimation d'un effet génétique ne varie pas dans un modèle complet et dans un modèle réduit. Autrement dit, si un effet est enlevé ou ajouté à un modèle orthogonal, les estimations demeurent inchangées. Cela permet une comparaison directe entre le modèle complet et le modèle réduit. Deuxièmement, avec un modèle orthogonal, la variance génétique peut être décomposée simplement en plusieurs variances, chacune correspondant à un effet génétique.

Les déséquilibres (déséquilibre de Hardy-Weinberg, déséquilibre de liaison) induisent *de facto* des covariances entre effets (Zeng et al., 2005). Le déséquilibre de Hardy-Weinberg peut être interprété comme une corrélation entre allèles à un seul locus. Si l'on souhaitait créer un modèle différenciant les effets des deux allèles (en distinguant l'effet de l'allèle maternel de celui de l'allèle paternel), il ne serait pas possible de créer un modèle orthogonal en présence de déséquilibre de Hardy-Weinberg. Mais en l'occurrence, dans les modèles de génétique quantitative, les deux types d'allèles ne sont généralement pas différenciés. C'est l'effet d'un locus, sur les deux brins d'ADN, qui intéresse les généticiens. Par conséquent, la corrélation qu'induit le déséquilibre de Hardy-Weinberg peut être surmontée par les modèles de génétique quantitative classique. Ainsi, des auteurs sont parvenus à écrire des modèles orthogonaux, même en présence de déséquilibre de Hardy-Weinberg (Álvarez-Castro and Carlborg, 2007; Gallais, 1990). En revanche, le déséquilibre de liaison induit toujours une covariance entre effets, si l'on cherche à estimer les effets de chaque locus indépendamment, ce qui est très généralement le cas. Il en irait différemment si les locus en déséquilibre de liaison étaient intégrés au modèle comme étant un seul « super gene », ce qui n'est pas le cas des modèles que nous avons étudiés.

Le déséquilibre de liaison peut être créé par un lien physique entre deux locus situés dans la même région. Dans ce cas, le déséquilibre de liaison décroit avec la distance génétique entre locus, progressivement dans les espèces autogames comme le blé, plus rapidement dans les espèces allogames, du fait d'un plus fort taux de recombinaison efficace (Rafalski, 2010). En effet, la recombinaison efficace – plus forte chez les espèces allogames (Pollak, 1987)- a pour effet de réduire le déséquilibre de liaison. En panmixie, en l'absence de mutation, de séléction et de dérive, le déséquilibre de liaison à la génération g est égal à (Hill and Robertson, 1968) :

$$D_g = D_{g-1}\,(1-r)$$

Où D_g est le coefficient du déséquilibre de liaison entre deux locus à la génération g et r le taux de recombinaison entre les deux locus.

Ce qui revient à écrire : $D_g = D_0\,(1-r)^g$

Ainsi, le déséquilibre de liaison diminue au cours des générations en fonction du taux de recombinaison. Plus le taux de recombinaison sera fort, plus le déséquilibre de liaison diminuera rapidement. Le déséquilibre de liaison est donc plus important sur les espèces autogames. Le déséquilibre de liaison ne dépend cependant pas uniquement de la distance physique et du taux de recombinaison. Des phénomènes démographiques et évolutifs tels que la dérive ou la sélection) peuvent également engendrer du déséquilibre de liaison (Mackay and Powell, 2007).

II.2. Ecriture unilocus diploïde (estimation de l'effet additif et de dominance)

Dans cette partie, plusieurs modèles sont présentés, sous l'écriture la plus courante (modèle unilocus diploïdes). Dans ce cas-là, seul l'effet additif d'un locus, ainsi que son effet de dominance, peuvent être estimés.

II.2.1 Ecriture générale des modèles d'estimation des effets génétiques unilocus diploïdes

Le premier modèle de génétique quantitative a été développé par Fisher (Fisher, 1918), qui définissait les effets génétiques comme des différences de moyennes. C'est un modèle bi-allélique qui se situe dans une population panmictique et en équilibre de liaison.

Soit P_{ijk} le phénotype d'un individu k, portant un génotype ij. Ce phénotype est égal à la valeur génotypique de ij, plus un effet environnemental propre à k, $\varepsilon_k\,(\varepsilon_k \sim N(0, \sigma_E^2) iid)$:

$$P_{ijk} = G_{ij} + \varepsilon_{ijk} \qquad \textbf{(1)}$$

La valeur d'un génotype ij se décompose en plusieurs effets génétiques :

$$G_{ij} = \mu + a_i + a_j + d_{ij}$$

Avec μ la moyenne, a_i l'effet additif de l'allèle i, a_j l'effet additif de l'allèle j, et d_{ij} l'effet de déviation due à la dominance (interaction entre les allèles).

Cette régression peut s'écrire aussi :

$$G_{ij} = \mu + x_i a + x_j a + x_{ij} d$$

Où $x_i a = a_i$, $x_j a = a_j$ et $x_{ij} d = d_{ij}$. Les coefficients x_i, x_j et x_{ij} sont variables et prennent une valeur dépendant des allèles i et j. Fisher a appelé a l'effet de substitution d'un allèle, mais d'autres auteurs (Álvarez-Castro and Carlborg, 2007; Yang, 2004; Zeng et al., 2005) l'appellent simplement « effet additif ». C'est ce terme que nous emploierons. Quant à d, il est appelé effet de dominance. La confusion des termes utilisés par les auteurs peut entraîner la confusion entre deux notions : les effets additifs des allèles (a_i et a_j) qui sont des variables aléatoires, et le paramètre d'effet additif (a) qui est fixe. Cette confusion existe aussi pour la dominance.

Ce modèle linéaire peut également s'écrire de manière matricielle. Soit un locus porteur de deux allèles : b et B, en fréquence p_b et p_B dans la population. Il y a donc trois génotypes possibles : bb, bB et BB, dont les fréquences seront notées p_{bb}, p_{bB} et p_{BB}. Les trois valeurs génotypiques G_{bb}, G_{bB} et G_{BB} sont égales à :

$$\begin{pmatrix} G_{BB} \\ G_{bB} \\ G_{bb} \end{pmatrix} = \begin{pmatrix} 1 & w_{BB} & x_{BB} \\ 1 & w_{bB} & x_{bB} \\ 1 & w_{bb} & x_{bb} \end{pmatrix} \begin{pmatrix} \mu \\ a \\ d \end{pmatrix}$$

Avec $w_{ij} = x_i + x_j$. Les coefficients w_{ij} et x_{ij} dépendent du modèle utilisé.

On abrègera cette formule en :

$$\boldsymbol{G} = \boldsymbol{S} \cdot \boldsymbol{E} \qquad \textbf{(2)}$$

Avec $\boldsymbol{G}$ le vecteur des valeurs génotypiques, $\boldsymbol{S}$ la matrice de « design » et $\boldsymbol{E}$ le vecteur des effets génétiques.

Pour pouvoir estimer les effets génétiques à partir du phénotype et du génotype des n individus de l'échantillon, on peut écrire l'équation **(1)** sous forme matricielle. Soit $\boldsymbol{P}$ le vecteur des phénotypiques de n individus :

$$\boldsymbol{P} = \boldsymbol{Z} \cdot \boldsymbol{G} + \boldsymbol{\varepsilon} \quad \textbf{(3)}$$

$$\begin{pmatrix} P_1 \\ P_2 \\ . \\ . \\ P_n \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ \dots & \dots & \dots \\ 0 & 1 & 0 \\ \dots & \dots & \dots \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} G_{BB} \\ G_{bB} \\ G_{bb} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_n \end{pmatrix}$$

Où $\boldsymbol{Z}$ est une matrice de dimension (nx3). Les n lignes de la matrice $\boldsymbol{Z}$ indique le génotype de chaque individu. Le vecteur $\boldsymbol{\varepsilon}$ est le vecteur des effets environnementaux.

En remplaçant dans l'équation **(3)** $\boldsymbol{G}$ par $\boldsymbol{S} \cdot \boldsymbol{E}$ (voir équation **(2)**), on peut écrire

$$\boldsymbol{P} = \boldsymbol{Z} \cdot \boldsymbol{S} \cdot \boldsymbol{E} + \boldsymbol{\varepsilon}$$

Où l'on définit la matrice de design comme $\boldsymbol{X} = \boldsymbol{Z} \cdot \boldsymbol{S}$, et où l'on utilise les résultats classiques d'estimation par les moindres carrés pour l'estimation de $\boldsymbol{E}$:

$$\widehat{\boldsymbol{E}} = ((\boldsymbol{Z}\,\boldsymbol{S})^T \boldsymbol{Z}\boldsymbol{S})^{-1} (\boldsymbol{Z}\boldsymbol{S})^T \boldsymbol{P}$$
$$\rightarrow \widehat{\boldsymbol{E}} = (\boldsymbol{S}^T \boldsymbol{Z}^T \boldsymbol{Z}\boldsymbol{S})^{-1} \boldsymbol{S}^T \boldsymbol{Z}^T \boldsymbol{P}$$

Si $\boldsymbol{S}$ est une matrice inversible alors, on peut simplifier :

$$\widehat{\boldsymbol{E}} = \boldsymbol{S}^{-1} \begin{pmatrix} \frac{\sum P_{BBk}}{n_{BB}} \\ \frac{\sum P_{bBk}}{n_{bB}} \\ \frac{\sum P_{bbk}}{n_{bb}} \end{pmatrix}$$

La valeur génotypique d'un génotype ij est estimée en moyennant le phénotype de tous les individus portant ce génotype. Par conséquent $\widehat{G_{ij}} = \frac{\sum P_{ijk}}{n_{ij}}$.

Donc $\widehat{\boldsymbol{E}} = \boldsymbol{S}^{-1}\widehat{\boldsymbol{G}}$

Par conséquent, on retrouve bien l'équation **(2)** en espérance. On peut donc écrire que : $\boldsymbol{E} = \boldsymbol{S}^{-1}\boldsymbol{G}$

Dans la suite de ce paragraphe nous nous intéressons à la quantité $\boldsymbol{E}$ ($= \boldsymbol{S}^{-1}\boldsymbol{G}$) qui est le vecteur des effets génétiques que nous cherchons à estimer.

II.2.2 Présentation de plusieurs modèles d'estimation des effets génétiques sous forme unilocus diploïde

II.2.2.1 Modèle G2A : modèle statistique où les effets additifs correspondent à des différences de moyennes

L'un des modèles les plus généraux est le modèle G2A (general two-allele), développé par Zeng et al., 2005. Dans l'esprit, il est quasiment similaire à celui de Fisher. Il s'écrit :

$$G_{ij} = \mu + w_{ij}a + x_{ij}d$$

Avec $w_{ij} = \begin{cases} 2p_b & \text{si } ij = BB \\ p_b - p_B & \text{si } ij = bB \\ -2p_B & \text{si } ij = bb \end{cases}$ et $x_{ij} = \begin{cases} -2p_b^2 & \text{si } ij = BB \\ 2p_b p_B & \text{si } ij = bB \\ -2p_B^2 & \text{si } ij = bb \end{cases}$

Les coefficients w_{ij} et x_{ij} sont centrés. On note N_i la variable aléatoire de l'allèle maternel i (N_j est celle de l'allèle paternel j). La variable N_i ne peut prendre que deux valeurs dans le cas bi-allélique :

$$N_i = \begin{cases} 1 & \text{si } i = B \\ 0 & \text{si } i = b \end{cases}$$

Ainsi, $N_i = \{0,1\}$ et est modélisée par une loi de Bernoulli de paramètre p_B (fréquence de l'allèle B). Zeng propose de travailler sur la variable centrée $x_i = N_i - p_B$. Notons qu'en pratique, nous n'aurons à disposition qu'un échantillon de n individus, et que les fréquences alléliques dans la population ne sont pas connues mais estimées à partir de l'échantillon. Pour chaque allèle, le coefficient x_i est égal à $x_i = \begin{cases} p_b & \text{si } i = B \\ -p_B & \text{si } i = b \end{cases}$

Les coefficients du modèles sont donc égaux à $w_{ij} = x_i + x_j$ et $x_{ij} = -2x_i x_j$

La matrice $\boldsymbol{S}_{G2A}$ est égale à :

$$\boldsymbol{S}_{G2A} = \begin{pmatrix} 1 & 2p_b & -2p_b^2 \\ 1 & p_b - p_B & 2p_b p_B \\ 1 & -2p_B & -2p_B^2 \end{pmatrix}$$

Et la matrice $\boldsymbol{S}_{G2A}^{-1}$ à :

$$\boldsymbol{S}_{G2A}^{-1} = \begin{pmatrix} p_B^2 & 2p_b p_B & p_b^2 \\ p_B & p_b - p_B & -p_b \\ -\frac{1}{2} & 1 & -\frac{1}{2} \end{pmatrix}$$

Dans ce modèle, où les variables sont centrées, la moyenne estimée correspond donc à une moyenne dans une population en équilibre de Hardy-Weinberg:

$$\mu = p_B^2 G_{BB} + 2p_b p_B G_{bB} + p_b^2 G_{bb}$$

L'effet additif est estimé par la différence entre la moyenne phénotypique des individus porteur de l'allèle B, et celle des individus porteur de l'allèle b :

$$a = \mu_B - \mu_b = (p_B G_{BB} + p_b G_{bB}) - (p_B G_{bB} + p_b G_{bb}) = p_B G_{BB} + (p_b - p_B) G_{bB} - p_b G_{bb}$$

Enfin, l'effet de dominance, en tant qu'effet d'interaction, est indépendant des fréquences alléliques. Il est estimé par la valeur génotypique de l'hétérozygote moins la moyenne des valeurs génotypiques des deux homozygotes :

$$d = G_{bB} - \frac{G_{bb} + G_{BB}}{2}$$

L'intérêt de ce modèle est qu'il permet d'estimer facilement des variances et covariances génétiques. Néanmoins, certaines modifications sont nécessaires si la population est en déséquilibre de liaison. Tout cela sera détaillé dans les paragraphes **II.4.2.3 et II.4.3**.

Un autre avantage est que le modèle est orthogonal en présence d'équilibre de liaison et de Hardy-Weinberg : dans ce cas-là, les effets génétiques, additifs ($w_{ij}a$) et dominance ($x_{ij}d$), sont indépendants, et il n'y a pas de covariances génétiques entre eux. Ainsi, sous ces conditions, l'estimation des effets demeurera inchangée si un nouveau locus est rajouté au modèle. En présence

de déséquilibre de Hardy-Weinberg, le modèle n'est plus orthogonal : les effets génétiques deviennent dépendants, et les estimations changent si un effet est rajouté ou enlevé du modèle.

I.2.2.2 Modèle UWR : modèle d'estimation des effets fonctionnels

Le modèle UWR (unweighted-regression model) a été développé par Cheverud et Routman en 1995 (Cheverud, 2000; Cheverud and Routman, 1995). Son objectif est de mesurer des effets fonctionnels et non pas des effets statistiques. En effet, le modèle de Fisher et le modèle G2A estiment des effets qui dépendent à la fois de la physiologie du gène étudié, mais également de la population dans laquelle il est estimé. Le modèle UWR permet d'estimer des effets indépendamment des fréquences génotypiques. Son modèle est le suivant :

$$G_{ij} = \mu + w_{ij}a + x_{ij}d$$

Les coefficients du modèle sont égaux à :

$$w_{ij} = x_i + x_j = \begin{cases} 1 & \text{si } ij = BB \\ 0 & \text{si } ij = bB \\ -1 & \text{si } ij = bb \end{cases} \text{ et } x_{ij} = \begin{cases} -1/3 & \text{si } ij = BB \\ 2/3 & \text{si } ij = bB \\ -1/3 & \text{si } ij = bb \end{cases}$$

La matrice $\boldsymbol{S}_{UWR}$ est donc égale à :

$$\boldsymbol{S}_{UWR} = \begin{pmatrix} 1 & 1 & -1/3 \\ 1 & 0 & 2/3 \\ 1 & -1 & -1/3 \end{pmatrix}$$

Et la matrice $\boldsymbol{S}_{UWR}^{-1}$ à :

$$\boldsymbol{S}_{UWR}^{-1} = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1/2 & 0 & -1/2 \\ -\frac{1}{2} & 1 & -\frac{1}{2} \end{pmatrix}$$

La moyenne μ correspond donc à une moyenne ajustée :

$$\mu = \frac{G_{BB} + G_{bB} + G_{bb}}{3}$$

L'effet additif est simplement estimé par la différence de valeur génotypique entre les deux homozygotes, divisée par deux :

$$a = \frac{G_{BB} - G_{bb}}{2}$$

Enfin, l'effet de dominance est, comme pour le modèle G2A, estimé par la valeur génotypique de l'hétérozygote moins la moyenne des valeurs génotypiques des homozygotes :

$$d = G_{bB} - \frac{G_{bb} + G_{BB}}{2}$$

L'intérêt de ce modèle est d'estimer des effets fonctionnels indépendants des caractéristiques de la population. Il est similaire à beaucoup de modèles de génétique quantitative, comme celui par exemple de Falconer et Mackay (Falconer and Mackay, 1996), qui écrivent : $G_{bb} = \mu - a$, $G_{bB} = \mu + d$ et $G_{BB} = \mu + a$ (cf. Figure 2.A). Dans ce cas, les effets a et d sont estimés de la même façon qu'avec le modèle UWR.

L'inconvénient de ce modèle est qu'il n'est pas orthogonal, sauf dans une population où les génotypes sont équifréquents. Par conséquent, rajouter un nouveau locus modifiera les estimations des effets génétiques.

II.2.2.3 Modèle NOIA : un modèle orthogonal en présence de déséquilibre de Hardy-Weinber

Le modèle NOIA (natural and orthogonal interactions model) (Álvarez-Castro and Carlborg, 2007) comprend deux formulations : l'une fonctionnelle, l'autre statistique.

II.2.2.3.1 Formulation fonctionnelle

L'objectif est le même que pour le modèle UWR : estimer des effets indépendamment des fréquences génotypiques.

La matrice $\boldsymbol{S}_{NOIA,F}$ est égale à :

$$\boldsymbol{S}_{NOIA,F} = \begin{pmatrix} 1 & 2p_b & -p_{bB} \\ 1 & p_b - p_B & 1 - p_{bB} \\ 1 & -2p_B & -p_{bB} \end{pmatrix}$$

La deuxième colonne est identique à celle de la matrice $\boldsymbol{S}_{G2A}$. La matrice $\boldsymbol{S}_{NOIA,F}^{-1}$ quant à elle est égale à:

$$\boldsymbol{S}_{NOIA,F}^{-1} = \begin{pmatrix} p_{BB} & p_{bB} & p_{bb} \\ 1/2 & 0 & -1/2 \\ -\frac{1}{2} & 1 & -\frac{1}{2} \end{pmatrix}$$

Les effets additifs et de dominance sont ainsi formulés de la même manière que dans le modèle UWR : $a = \frac{G_{BB} - G_{bb}}{2}$ et $d = G_{bB} - \frac{G_{bb} + G_{BB}}{2}$. En revanche, la moyenne est celle de la population $\mu = p_{BB}G_{BB} + p_{bB}G_{bB} + p_{bb}G_{bb}$.

Ce modèle est orthogonal si $p_{bB} = 0$ ou si $p_{BB} = p_{bb}$.

II.2.2.3.2 Formulation statistique

Dans la formulation statistique du modèle NOIA, la matrice $\boldsymbol{S}_{NOIA,S}$ est égale à :

$$\boldsymbol{S}_{NOIA,S} = \begin{pmatrix} 1 & 2p_b & \frac{-2p_{bB}p_{bb}}{f(p_{bb},p_{BB})} \\ 1 & p_b - p_B & \frac{4p_{bb}p_{BB}}{f(p_{bb},p_{BB})} \\ 1 & -2p_B & \frac{-2p_{bB}BB}{f(p_{bb},p_{BB})} \end{pmatrix}$$

Avec $f(p_{bb},p_{BB}) = p_{bb} + p_{BB} - (p_{bb} - p_{BB})^2$. On remarque que si $p_{bb} = p_b^2$, $p_{BB} = p_B^2$ et $p_{bB} = 2p_b p_B$ (équilibre de Hardy-Weinberg), alors le modèle est identique au modèle G2A.

La matrice $\boldsymbol{S}_{NOIA,S}^{-1}$ est égale à :

$$\boldsymbol{S}_{NOIA,S}^{-1} = \begin{pmatrix} p_{BB} & p_{bB} & p_{bb} \\ \frac{2p_{bb}p_B}{f(p_{bb},p_{BB})} & \frac{p_{bB}(p_{bb} - p_{BB})}{f(p_{bb},p_{BB})} & \frac{-2p_{BB}p_b}{f(p_{bb},p_{BB})} \\ -\frac{1}{2} & 1 & -\frac{1}{2} \end{pmatrix}$$

Comme pour le modèle G2A, la valeur des effets additifs dépend des fréquences génotypiques. Si $p_{bB} = 0$ ou si $p_{BB} = p_{bb}$, alors les formulations fonctionnelle et statistique du modèle NOIA sont équivalentes.

Ce modèle est une construction mathématique pour être orthogonal, même en présence de déséquilibre de Hardy-Weinberg. Une représentation graphique de la formulation statistique du modèle NOIA est illustrée en **figure 9**. Les effets génétiques se déduisent d'une pente de régression, de la valeur génotypique en fonction du nombre d'allèles B, pondérée par leur fréquence génotypique. Appelons cette régression $G(x)$, avec x le nombre d'allèles B. Alors les effets additifs se modélisent comme :

$$a_{ij} = w_{ij}a = G(x_{ij}) - \mu.$$

Où x_{ij} est le nombre d'allèles B compris dans le génotype ij.

Quant aux effets de dominance, on peut les déduire par :

$$d_{ij} = x_{ij}d = G_{ij} - G(x_{ij})$$

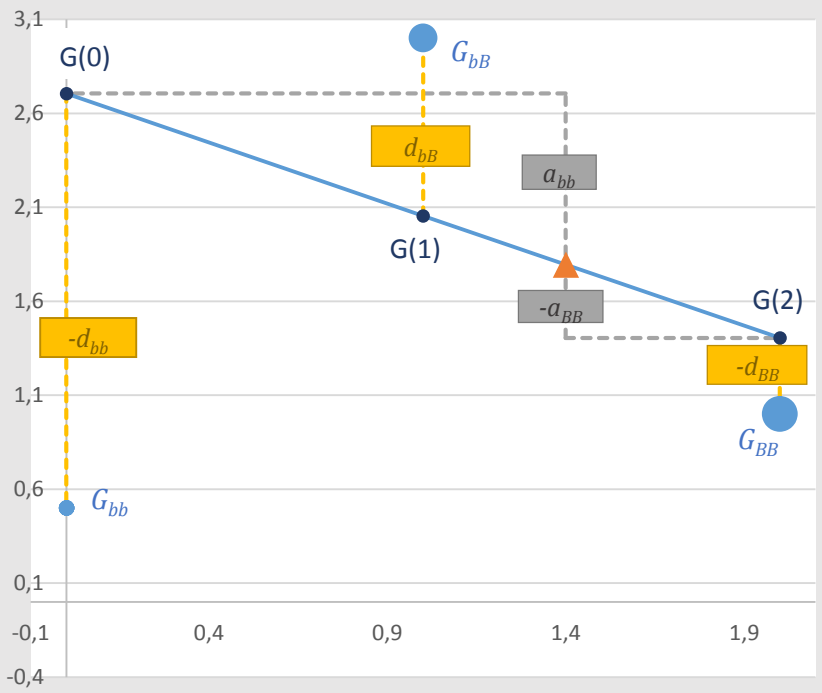


Figure 9 : Interprétation graphique de la formulation statistique du modèle NOIA
En abscisse : nombre d'allèles B pour chaque génotype (0, 1, 2)
En ordonnée : valeurs génotypiques G_{bb}, G_{bB} et G_{BB}
Ici les fréquences génotypiques sont égales à $p_{bb} = 0.9$, $p_{bB} = 0.42$ et $p_{BB} = 0.49$.
Points bleus : les valeurs génotypiques.
Droite bleue : la droit de régression $G(x)$, avec x le nombre d'allèle B, pondérée en fonction de la fréquence génotypique.
Triangle orange : moyenne phénotypique de la population en fonction du nombre moyen d'allèle B dans la population.
a_{ij}: effet additif associé au génotype ij $a_{ij} = w_{ij}a$.
d_{ij} : effet additif associé au génotype ij $d_{ij} = x_{ij}a$

II.2.2.4 Autres modèles

Deux autres modèles sont fréquemment utilisés pour détecter des effets génétiques, notamment dans des études de QTL : le modèle F2 (Cockerham, 1954) et le modèle F∞ (Hayman, 1957). Ces modèles apparaissent comme des cas particuliers des deux formulations du modèle NOIA, avec $p_{bb} = 1/4$, $p_{bB} = 1/2$ et $p_{BB} = 1/4$ pour le modèle F2 et $p_{bb} = 1/2$, $p_{bB} = 0$ et $p_{BB} = 1/2$ pour le modèle F∞. Ces modèles sont orthogonaux dans des populations F2 et F∞, respectivement.

II.3. Généralisation à plusieurs locus

Considérons deux locus, A et B. La matrice de « design » propre à chacun de ces locus est appelée $\boldsymbol{S}_A$ et $\boldsymbol{S}_B$, respectivement. Pour généraliser un modèle à plusieurs locus, il suffit d'appliquer le modèle de Kronecker. Par exemple :

$$\boldsymbol{G} = \boldsymbol{S}_B \otimes \boldsymbol{S}_A \cdot \boldsymbol{E}$$

$$\begin{pmatrix} G_{aabb} \\ G_{aabB} \\ G_{aaBB} \\ G_{aAbb} \\ G_{aAbB} \\ G_{aABB} \\ G_{AAbb} \\ G_{AAbB} \\ G_{AABB} \end{pmatrix} = \begin{pmatrix} 1 & w_{BB} & x_{BB} \\ 1 & w_{bB} & x_{bB} \\ 1 & w_{bb} & x_{bb} \end{pmatrix} \otimes \begin{pmatrix} 1 & w_{AA} & x_{AA} \\ 1 & w_{aA} & x_{aA} \\ 1 & w_{aa} & x_{aa} \end{pmatrix} \begin{pmatrix} \mu \\ a_A \\ d_A \\ a_B \\ d_B \\ aa \\ ad \\ da \\ dd \end{pmatrix}$$

$$\begin{pmatrix} G_{aabb} \\ G_{aabB} \\ G_{aaBB} \\ G_{aAbb} \\ G_{aAbB} \\ G_{aABB} \\ G_{AAbb} \\ G_{AAbB} \\ G_{AABB} \end{pmatrix} = \begin{pmatrix} 1 & w_{aa} & x_{aa} & w_{bb} & x_{bb} & w_{aa} \cdot w_{bb} & w_{aa} \cdot x_{bb} & x_{aa} \cdot w_{bb} & x_{aa} \cdot x_{bb} \\ 1 & w_{aa} & x_{aa} & w_{bB} & x_{bB} & w_{aa} \cdot w_{bB} & w_{aa} \cdot x_{bB} & x_{aa} \cdot w_{bB} & x_{aa} \cdot x_{bB} \\ 1 & w_{aa} & x_{aa} & w_{BB} & x_{BB} & w_{aa} \cdot w_{BB} & w_{aa} \cdot x_{BB} & x_{aa} \cdot w_{BB} & x_{aa} \cdot x_{BB} \\ 1 & w_{aA} & x_{aA} & w_{bb} & x_{bb} & w_{aA} \cdot w_{bb} & w_{aA} \cdot x_{bb} & x_{aA} \cdot w_{bb} & x_{aA} \cdot x_{bb} \\ 1 & w_{aA} & x_{aA} & w_{bB} & x_{bB} & w_{aA} \cdot w_{bB} & w_{aA} \cdot x_{bB} & x_{aA} \cdot w_{bB} & x_{aA} \cdot x_{bB} \\ 1 & w_{aA} & x_{aA} & w_{BB} & x_{BB} & w_{aA} \cdot w_{BB} & w_{aA} \cdot x_{BB} & x_{aA} \cdot w_{BB} & x_{aA} \cdot x_{BB} \\ 1 & w_{AA} & x_{AA} & w_{bb} & x_{bb} & w_{AA} \cdot w_{bb} & w_{AA} \cdot x_{bb} & x_{AA} \cdot w_{bb} & x_{AA} \cdot x_{bb} \\ 1 & w_{AA} & x_{AA} & w_{bB} & x_{bB} & w_{AA} \cdot w_{bB} & w_{AA} \cdot x_{bB} & x_{AA} \cdot w_{bB} & x_{AA} \cdot x_{bB} \\ 1 & w_{AA} & x_{AA} & w_{BB} & x_{BB} & w_{AA} \cdot w_{BB} & w_{AA} \cdot x_{BB} & x_{AA} \cdot w_{BB} & x_{AA} \cdot x_{BB} \end{pmatrix} \begin{pmatrix} \mu \\ a_A \\ d_A \\ a_B \\ d_B \\ aa \\ ad \\ da \\ dd \end{pmatrix}$$

En plus des effets simples (moyenne μ, effets additifs a_A et a_B, et effets de dominance d_A et d_B), les différents types d'épistasie peuvent être estimés :

- l'épistasie additive x additive, notée aa.
- l'épistasie additive x dominante (noté ad) et dominante x additive (noté da).
- l'épistasie dominante x dominante, notée dd.

II.4. Estimation de l'épistasie additive x additive : modèles haploïdes bilocus

Dans cette partie, nous présenterons des modèles haploïdes, permettant d'estimer les effets additifs et l'épistasie additive x additive, et donc adaptés à l'étude de l'espèce *Triticum turgidum*, qui est autogame et présente donc un fort taux d'homozygotie. Par conséquent, seuls les génotypes homozygotes ont donc été pris en compte.

II.4.1. Equivalence entre l'écriture diploïde et l'écriture haploïde

Considérons un modèle diploïde à deux locus : les locus A et B. Alors, la valeur génotypique du génotype *ikjm* peut s'écrire comme :

$$G_{ikjm} = \mu + a_{A,i}^{d} + a_{A,k}^{d} + d_{A,ik}^{d} + a_{B,j}^{d} + a_{B,m}^{d} + d_{B,jm}^{d} + aa_{ij}{}^{d} + aa_{kj}{}^{d} + aa_{im}{}^{d} + aa_{km}{}^{d} + ad_{ijm}{}^{d} + ad_{kjm}{}^{d} + da_{ikj}{}^{d} + da_{ikm}{}^{d} + dd_{ikjm}{}^{d}$$

Où :

- μ est la moyenne

- $a^d_{A,i}$ l'effet additif de l'allèle i au locus A, $a^d_{A,k}$ l'effet additif de l'allèle k au locus A, $a^d_{B,j}$ l'effet de l'allèle j au locus B et $a^d_{B,m}$ l'effet de l'allèle m au locus B (effets additifs)
- $d^d_{B,jm}$ l'effet de dominance du couple d'allèle ik au locus A et $d^d_{B,jm}$ l'effet de dominance du couple d'allèle jm au locus B (effets de dominance)
- ${aa_{ij}}^d$ l'effet de l'interaction entre les allèles i et j, ${aa_{kj}}^d$ l'effet de l'interaction entre les allèles k et j, ${aa_{im}}^d$ l'effet de l'interaction entre les allèles i et m , ${aa_{km}}^d$ l'effet de l'interaction entre les allès k et m (effets de l'épistasie additive x additive)
- ${ad_{ijm}}^d$ l'effet de l'interaction entre l'allèle i au locus A et le couple d'allèles jm au locus B, et ${ad_{kjm}}^d$ l'effet de l'interaction entre l'allèle k au locus A et le couple d'allèles jm au locus B (effets de l'épistasie additive x dominante)
- ${da_{ikj}}^d$ l'effet de l'interaction entre le couple d'allèles ik au locus A et l'allèle j au locus B, et ${da_{ikm}}^d$ l'effet de l'interaction entre le couple d'allèles ik au locus A et l'allèle m au locus B (effets de l'épistasie dominante x additive)
- ${dd_{ikjm}}^d$ l'effet de l'interaction entre le couple d'allèles ik au locus A et le couple d'allèles jm au locus B (effet de l'épistasie dominante x dominance)

L'indice d sert à indiquer que l'on est dans un modèle diploïde.

Si seuls les génotypes homozygotes sont donc été pris en compte quatre génotypes sont donc possibles pour deux locus : $aabb$, $aaBB$, $AAbb$ et $AABB$. Les effets de dominance ou d'épistasie incluant de la dominance, disparaissent. Le modèle, sous forme diploïde, peut donc s'écrire comme :

$$G_{iijj} = \mu + 2a^d_{A,i} + 2a^d_{B,j} + 4{aa_{ij}}^d$$

Où $a^d_{A,i}$ est l'effet de l'allèle i au locus A, $a^d_{B,j}$ l'effet de l'allèle j au locus A, et ${aa_{ij}}^d$ l'effet d'interaction entre les allèles i et j. L'indice d sert à indiquer que l'on est dans un modèle diploïde.

Remarquons que dans le cas où les effets incluant de la dominance sont nuls, la moyenne dans une population consanguine et dans une population présentant des hétérozygotes est identique. En revanche, en présence d'effets de dominance et d'épistasie incluant de la dominance, la moyenne changera (Gallais, 1990).

Il est plus simple d'écrire ce modèle sous une forme haploïde, c'est-à-dire de modéliser chaque génotype homozygote à un locus comme un allèle. Les génotypes sont donc notés ab, aB, Ab et AB, et la valeur génotypique d'un génotype ij est égale à :

$$G_{ij} = \mu + a^h_{A,i} + a^h_{B,j} + {aa_{ij}}^h$$

Où $a_{A,i}^h$ est l'effet de l'allèle i au locus A, $a_{B,j}^h$ l'effet de l'allèle j au locus B, et $aa_{ij}{}^h$ l'effet d'interaction entre les allèles i et j. L'indice h sert à indiquer que l'on est dans un modèle haploïde.

On peut remarquer que $a_{A,i}^h = 2a_{A,i}^d$, $a_{B,j}^h = 2a_{B,j}^d$ et $aa_{ij}{}^h = 4aa_{ij}{}^d$.

Le modèle haploïde peut aussi s'écrire comme :

$$G_{ij} = \mu + x_i a_A + x_j a_B + x_{ij} aa$$

Par la suite, nous emploierons cette écriture.

II.4.2. Ecriture des modèles présentés ci-dessus sous forme haploïde

Les allèles i et j sont situés à deux locus différents, A et B. Les fréquences alléliques sont égales à p_a, p_A, p_b et p_B, pour les allèles a, A, b et B, respectivement. Les effets a_A et a_B sont les effets additifs de ces deux locus. L'effet aa désigne l'épistasie additive x additive.

II.4.2.1. Modèle G2A et modèle NOIA sous forme haploïde

Quand les effets de dominance sont éliminés du modèle, les modèles G2A et NOIA (les deux formulations : fonctionnelle et statistique) sont identiques.

Le modèle G2A (ou NOIA) haploïde à deux locus peut-être écrit comme :

$$\begin{pmatrix} G_{AB} \\ G_{aB} \\ G_{Ab} \\ G_{ab} \end{pmatrix} = \begin{pmatrix} 1 & p_b & p_a & p_a p_b \\ 1 & p_b & -p_A & -p_A p_b \\ 1 & -p_B & p_a & -p_a p_B \\ 1 & -p_B & -p_A & p_A p_B \end{pmatrix} \begin{pmatrix} \mu \\ a_A \\ a_B \\ aa \end{pmatrix}$$

La matrice inverse est $\boldsymbol{S}_{G2A}^{-1}$, elle, égale à :

$$\boldsymbol{S}_{G2A}^{-1} = \begin{pmatrix} p_A p_B & p_a p_B & p_A p_b & p_a p_b \\ p_A & p_a & -p_A & -p_a \\ p_B & -p_B & p_b & -p_b \\ 1/4 & -1/4 & -1/4 & 1/4 \end{pmatrix}$$

On peut remarquer que quand plusieurs locus sont considérés, la formulation fonctionnelle du modèle NOIA n'estime plus les effets additifs indépendamment des fréquences alléliques. Contrairement au modèle UWR, ce modèle n'est donc pas véritablement fonctionnel.

La moyenne estimée μ ne correspond pas à la moyenne réelle de la population, mais à la moyenne qu'aurait la population en l'absence de déséquilibre de liaison. Par ailleurs, comme le déséquilibre de liaison n'est pas pris en compte dans la matrice $\boldsymbol{S}_{G2A}^{-1}$, les effets ne sont pas centrés.

Ce modèle ne peut pas être généralisé par le produit de Kronecker, car contrairement aux autres modèles, le coefficient de l'effet épistatique (x_{ij}) n'est pas égal au produit des coefficients des effets additifs ($x_i x_j$).

II.4.2.2. Modèles UWR, F2 et F∞ haploïdes

Sous forme haploïde, les modèles UWR, F2 et F∞ sont équivalents. Ils s'écrivent :

$$\begin{pmatrix} G_{AB} \\ G_{aB} \\ G_{Ab} \\ G_{ab} \end{pmatrix} = \begin{pmatrix} 1 & 1/2 & 1/2 & 1/4 \\ 1 & 1/2 & -1/2 & -1/4 \\ 1 & -1/2 & 1/2 & -1/4 \\ 1 & -1/2 & -1/2 & 1/4 \end{pmatrix} \begin{pmatrix} \mu \\ a_A \\ a_B \\ aa \end{pmatrix}$$

La matrice inverse est $\boldsymbol{S}_{UWR}^{-1}$, elle, égale à :

$$\boldsymbol{S}_{UWR}^{-1} = \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/2 & 1/2 & -1/2 & -1/2 \\ 1/2 & -1/2 & 1/2 & -1/2 \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

Ce modèle est un cas particulier du modèle G2A, quand tous les génotypes sont équifréquents.

II.4.2.3 Modèle centré G2A-D : modification du modèle G2A pour tenir compte du déséquilibre de liaison

Le modèle G2A, sous forme haploïde, s'écrit comme :

$$G_{ij} = \mu + x_i a_A + x_j a_B + x_i x_j aa$$

Avec $x_i = \begin{cases} p_a & \text{si } i = A \\ -p_A & \text{si } i = a \end{cases}$ et $x_j = \begin{cases} p_b & \text{si } i = A \\ -p_B & \text{si } i = a \end{cases}$

x_i et x_j sont centrés. En effet $\overline{x_i} = p_A x_A + p_a x_a = p_A p_a - p_a p_A = 0$. De même $\overline{x_j} = 0$.

A l'inverse, $x_i x_j$ n'est pas une variable centrée :

$$\begin{aligned} \overline{x_i x_j} &= p_{AB} x_A x_B + p_{Ab} x_A x_b + p_{aB} x_a x_B + p_{ab} x_a x_b \\ &= [p_A p_B + D](1 - p_A)(1 - p_B) + [p_A(1 - p_B) - D](1 - p_A)(-p_B) \\ &\quad + [(1 - p_A)p_B - D](-p_A)(1 - p_B) + [(1 - p_A)(1 - p_B) + D](-p_A)(-p_B) = D \end{aligned}$$

Avec D le coefficient du déséquilibre de liaison.

Avec un modèle centré, il est très simple de calculer la matrice de variance-covariance. Le fait que $x_i x_j$ ne soit pas centré en présence de déséquilibre de liaison pose donc problème pour estimer les variances et covariances génétiques. J'introduis donc un nouveau modèle, G2A-D, qui est une forme centrée du modèle G2A.

Pour obtenir une valeur centrée, il faut donc soustraire D à $x_i x_j$: $x_{ij} = x_i x_j - D$; $\overline{x_{ij}} = 0$

Dans ce cas, la nouvelle matrice $\boldsymbol{S}_{G2A-D}$ est égale à :

$$\boldsymbol{S}_{G2A-D} = \begin{pmatrix} 1 & p_b & p_a & p_a p_b - D \\ 1 & p_b & -p_A & -p_A p_b - D \\ 1 & -p_B & p_a & -p_a p_B - D \\ 1 & -p_B & -p_A & p_A p_B - D \end{pmatrix}$$

La matrice $\boldsymbol{S}_{G2A-D}^{-1}$ modifiée, quant à elle, est égale à :

$$\begin{pmatrix} p_A p_B + D & p_a p_B - D & p_A p_b - D & p_a p_b + D \\ p_A & p_a & -p_A & -p_a \\ p_B & -p_B & p_b & -p_b \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

Le paramètre µ représente la valeur moyenne dans la population. Il est important d'utiliser ces matrices $\boldsymbol{S}_{G2A-D}$ et $\boldsymbol{S}_{G2A-D}^{-1}$ pour estimer directement et correctement les variances et covariances génétiques, en présence de déséquilibre de liaison (paragraphe II.4.3).

Par rapport au modèle G2A, les effets additifs peuvent toujours s'interpréter en termes de différence de moyennes phénotypiques entre porteurs des deux allèles d'un locus. Par exemple :

$$a_B = (p_A G_{AB} + p_a G_{aB}) - (p_A G_{Ab} + p_a G_{ab})$$

Enfin, l'effet d'interaction (ici l'épistasie) est indépendante des fréquences alléliques, comme l'était l'effet de dominance dans le modèle diploïde unilocus :

$$aa = G_{AB} - G_{aB} - G_{Ab} + G_{ab}$$

Les estimations des effets additifs et epistatiques ne changent donc pas quand on utilise le modèle G2A-D par rapport au modèle G2A. Seule l'estimation de la moyenne µ varie, puisqu'elle tient dorénavant compte du déséquilibre de liaison.

II.4.2.4. Conclusion sur l'écriture haploïde

Dans le cas haploïde, les effets de dominance sont non estimables. Par conséquent, des modèles différents dans leur écriture diploïde deviennent identiques lorsqu'ils sont écrits dans le cas haploïde. Deux modèles existent donc dans le cas haploïde :

- un modèle fonctionnel, indépendant des fréquences, qui correspond aux modèles UWR, F2 et F∞
- un modèle statistique, dépendants des fréquences. Deux sous-modèles existent :
    - Un modèle ne tenant pas compte du déséquilibre de liaison. Il correspond au modèle G2A et NOIA.
    - Un modèle tenant compte du déséquilibre de liaison. Il correspond au modèle G2A-D.

L'épistasie additive x additive est estimée de la même façon par tous ces modèles ($aa = G_{AB} - G_{aB} - G_{Ab} + G_{ab}$). L'estimation de la moyenne varie entre le modèle fonctionnel, le modèle statistique ne tenant pas compte du déséquilibre de liaison, et le modèle statistique tenant compte du déséquilibre de liaison. Enfin les effets additifs ne sont pas estimés de la même manière par le modèle fonctionnel ($a_B = \frac{1}{2}(G_{AB} - G_{Ab} + G_{aB} - G_{ab})$ et par le modèle statistique ($a_B = (p_A G_{AB} + p_a G_{aB}) - (p_A G_{Ab} + p_a G_{ab})$). Nous avons représenté, sur la **figure 10**, l'estimation de a_B par les deux modèles en fonction de p_A, en absence et en présence d'épistasie.

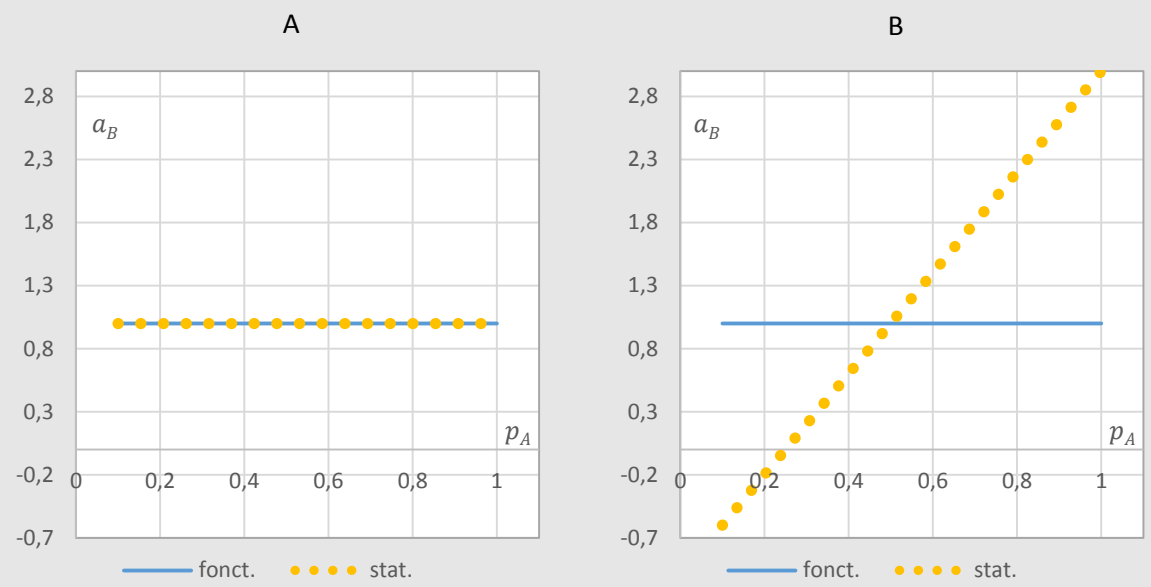


Figure 10 : estimation de a_B en fonction de p_A par les deux modèles, en l'absence (A) et en la présence d'épistasie (B).

a_B^f : estimation de a_B par le modèle fonctionnel, représentée par une ligne bleue ; a_B^s estimation de a_B par le modèle statistique, représentée par des pointillés jaunes.

(A) Absence d'épistasie : $G_{AB} = 1$; $G_{aB} = 0$; $G_{Ab} = 0$; $G_{ab} = -1$
Dans ce cas-là, quelle que soit p_A, $a_B^f = 1$ et $a_B^s = 1$ aussi car $G_{AB} - G_{Ab} = G_{aB} - G_{ab}$.

(B) Présence d'épistasie : $G_{AB} = 2$; $G_{aB} = -1$; $G_{Ab} = -1$; $G_{ab} = 0$
Dans ce cas-là, quelle que soit p_A, $a_B^f = 1$
En revanche a_B^s varie en fonction de p_A car $G_{AB} - G_{Ab} \neq G_{aB} - G_{ab}$.

La figure 10 montre qu'en l'absence d'épistasie (Figure 10.A), les modèles fonctionnel et statistique estiment de la même façon a_B. En effet, en l'absence d'épistasie $G_{AB} - G_{Ab} = G_{aB} - G_{ab}$, donc $a_B^s = p_A(G_{AB} - G_{Ab}) + (1 - p_A)(G_{aB} - G_{ab}) = G_{AB} - G_{Ab}$. Par ailleurs $a_B^f = \frac{1}{2}(G_{AB} - G_{Ab} + G_{aB} - G_{ab}) = G_{AB} - G_{Ab}$. Donc $a_B^s = a_B^f$.

En revanche, en présence d'épistasie (Figure 10.B), les estimations ne sont pas les mêmes, sauf pour $p_A = 0{,}5$. En effet, $G_{AB} - G_{Ab} \neq G_{aB} - G_{ab}$ et donc a_B^s devient sensible aux fréquences alléliques.

Nous pouvons constater que dans le cas haploïde bilocus, l'épistasie est toujours estimée indépendamment des fréquences alléliques, y compris dans les modèles statistiques. Cela n'est pas vrai pour le cas diploïde. La distinction entre « épistasie fonctionnelle » et « épistasie statistique », présentée en I.1.3.1. est donc caduque dans le cas des modèles haploïdes bilocus. Néanmoins, il reste à distinguer l'effet épistatique (le paramètre estimé par les modèles) de la variance épistatique. Comme l'a montré la Figure 4a en l'introduction, la variance épistatique peut être faible même si l'effet épistatique est fort.

II.4.3. Calculs des variances et covariances génétiques dans un modèle dilocus haploïde.

Le modèle G2A-D peut permettre l'estimation des variances et covariances génétiques d'une population. Nous chercherons à calculer :

- la variance additive au locus B : $V_B = V(x_j a_B) = {a_B}^2 \times V(x_j)$
- la variance additive au locus A : $V_A = V(x_i a_A) = {a_A}^2 \times V(x_i)$
- la variance épistatique : $V_{aa} = V([x_i x_j - D]aa) = aa^2 \times V(x_i x_j - D)$
- la covariance entre effets additifs : $\mathrm{cov}_{AB} = \mathrm{cov}(x_i a_A, x_j a_B) = a_A \times a_B \times \mathrm{cov}(x_i, x_j)$
- la covariance entre l'épistasie et l'effet additif au locus B : $\mathrm{cov}_{Baa} = \mathrm{cov}(x_j a_B, [x_i x_j - D]aa) = a_B \times aa \times \mathrm{cov}(x_j, x_i x_j - D)$
- la covariance entre l'épistasie et l'effet additif au locus A : $\mathrm{cov}_{Aaa} = \mathrm{cov}(x_i a_A, [x_i x_j - D]aa) = a_A \times aa \times \mathrm{cov}(x_i, x_i x_j - D)$

Soit $\boldsymbol{P}$ le vecteur des phénotypiques de n individus. Comme nous l'avons vu dans le paragraphe II.2.1, il peut se modéliser comme :

$$\boldsymbol{P} = \boldsymbol{Z} \cdot \boldsymbol{G} + \boldsymbol{\varepsilon} = \boldsymbol{Z} \cdot \boldsymbol{S}_{G2A-D} \cdot \boldsymbol{E}_{G2A-D} + \boldsymbol{\varepsilon}$$

Avec $\boldsymbol{Z}$ est une matrice de design de dimension (nx4) ; $\boldsymbol{\varepsilon}$ le vecteur des effets environnementaux ($\boldsymbol{\varepsilon} \sim \boldsymbol{N}(0, \sigma^2 I)$) ; $\boldsymbol{G}$ le vecteur des valeurs génotypique ; $\boldsymbol{S}_{G2A-D}$ matrice de design du modèle G2A-D ; $\boldsymbol{E}_{G2A-D}$ le vecteur des effets génétiques estimés avec le modèle centré.

Par définition, la matrice de variance covariance des variables explicatives (hors moyenne) vaut :

$$\begin{pmatrix} V(x_j) & \mathrm{cov}(x_i, x_j) & \mathrm{cov}(x_j, x_i x_j - D) \\ \mathrm{cov}(x_i, x_j) & V(x_i) & \mathrm{cov}(x_i, x_i x_j - D) \\ \mathrm{cov}(x_j, x_i x_j - D) & \mathrm{cov}(x_i, x_i x_j - D) & V(x_i x_j - D) \end{pmatrix}$$

Elle est égale à la sous matrice suivante (où on a enlèvé 1ère ligne et colonne concernant l'intercept) :

$$\left[\boldsymbol{S}_{G2A-D}{}^T \begin{pmatrix} p_{BA} & 0 & 0 & 0 \\ 0 & p_{Ba} & 0 & 0 \\ 0 & 0 & p_{bA} & 0 \\ 0 & 0 & 0 & p_{ba} \end{pmatrix} \boldsymbol{S}_{G2A-D} \right] [2:4, 2:4]$$

Soit :

$$\begin{pmatrix} p_B p_b & D & D(p_B - p_b) \\ D & p_A p_a & D(p_A - p_a) \\ D(p_B - p_b) & D(p_A - p_a) & p_B p_b p_A p_a + D(p_B - p_b)(p_A - p_a) - D^2 \end{pmatrix}$$

Elle est estimée par $\left[\boldsymbol{S}_{G2A-D}{}^T (\frac{1}{n} \boldsymbol{Z}^T \boldsymbol{Z}) \boldsymbol{S}_{G2A-D} \right] [2:4, 2:4]$

Pour obtenir les variances et covariances des effets génétiques (additifs : $a_{A,i} = x_i a_A$ et $a_j = x_j a_B$; épistatique : $aa_{ij} = (x_i x_j - D)aa$), il suffit de multiplier la matrice précédente à droite et à gauche par la matrice suivante :

$$\boldsymbol{E}_{diag} = \begin{pmatrix} a_B & 0 & 0 \\ 0 & a_A & 0 \\ 0 & 0 & aa \end{pmatrix}$$

On obtient bien le résultat cherché :

$$\begin{pmatrix} {a_B}^2 \times V(x_j) & a_A \times a_B \times \mathrm{cov}(x_i, x_j) & a_B \times aa \times \mathrm{cov}(x_j, x_i x_j - D) \\ a_A \times a_B \times \mathrm{cov}(x_i, x_j) & {a_A}^2 \times V(x_i) & a_A \times aa \times \mathrm{cov}(x_i, x_i x_j - D) \\ a_B \times aa \times \mathrm{cov}(x_j, x_i x_j - D) & a_A \times aa \times \mathrm{cov}(x_i, x_i x_j - D) & aa^2 \times V(x_i x_j - D) \end{pmatrix}$$

$$= \begin{pmatrix} V_B & \mathrm{cov}_{AB} & \mathrm{cov}_{Baa} \\ \mathrm{cov}_{AB} & V_A & \mathrm{cov}_{Aaa} \\ \mathrm{cov}_{Baa} & \mathrm{cov}_{Aaa} & V_{aa} \end{pmatrix}$$

On constate avec ce calcul que les effets ne sont indépendants qu'en l'absence de déséquilibre de liaison entre locus ($D = 0$).

La **figure 11** montre la variation de V_{aa} en fonction des fréquences alléliques et du déséquilibre de liaison.

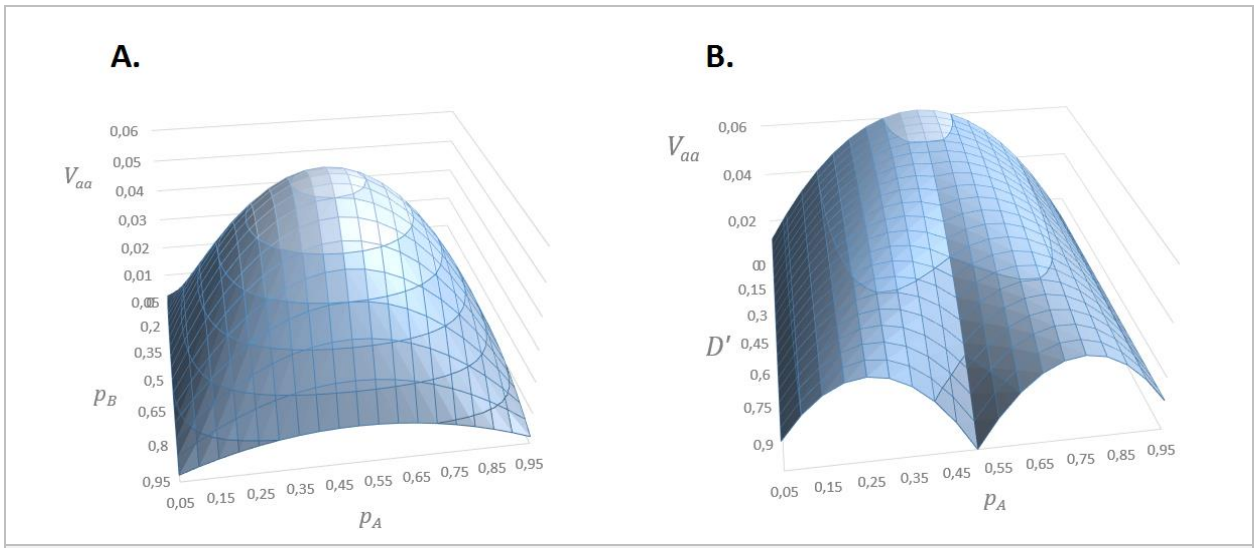


Figure 11 : Variance épistatique V_{aa} en fonction des fréquences alléliques et du déséquilibre de liaison

L'effet épistatique additif x additif aa ($= \boldsymbol{G_{AB} + G_{ab} - G_{Ab} - G_{aB}}$ dans tous les modèles) est égal à 1.
$\boldsymbol{V_{aa} = [p_B p_b p_A p_a + D(p_B - p_b)(p_A - p_a) - D^2]aa^2}$

- **A.** V_{aa} en fonction de la fréquence de l'allèle A (p_A) et de l'allèle B (p_B) quand le déséquilibre de liaison est nul ($D' = 0$). Notez que $p_A = 1 - p_a$ et $p_B = 1 - p_b$.
- **B.** V_{aa} en fonction des fréquences alléliques au locus A (p_A) et du déséquilibre de liaison normalisé D'. La fréquence de l'allèle B est fixé à $p_B = 0.5$.
  $\boldsymbol{D' = D/D_{max}}$ avec $\boldsymbol{D_{max} = \min(p_A p_b, p_a p_B)}$ et $D > 0$.

On voit que quand le déséquilibre de liaison est nul (Figure 11.A), la variance épistatique est maximale pour des fréquences intermédiaires. La Figure 11.B illustre la variance épistatique en fonction du déséquilibre de liaison normalisé (D') de la fréquence de l'allèle A (p_A) dans le cas où la fréquence de l'allèle B est fixé à p_B = 0.5. Il s'agit donc d'un cas particulier, où $V_{aa} = \frac{p_A p_a}{4} - D^2$. Dans ce cas, le déséquilibre de liaison a tendance à réduire la variance épistatique. Par ailleurs si p_A = 0.5, alors le déséquilibre de liaison maximale peut atteindre la valeur maximale possible : 0.25. En effet, $D_{max} = \min(p_A p_b, p_a p_B) = 0.25$. Cela explique le creux en p_A = 0.5.

II.5. Utilisation de ces modèles ultérieurement

Dans cette thèse, seuls des modèles bilocus haploïdes ont été utilisés. Dans le chapitre suivant (chapitre III), nous avons cherché à estimer les effets additifs et épistatiques de marqueurs SNPs sur le niveau d'expression de gènes dans des lignées fixées de blé dur, afin de détecter des eQTL. Nous voulions estimer des effets fonctionnels, ne dépendant pas des fréquences génotypiques. Nous avons donc utilisé le modèle UWR (cf. II.4.2.2.), en dépit du fait qu'il ne soit pas orthogonal.

Dans le chapitre IV, nous avons voulu approfondir le rôle de l'épistasie dans l'évolution d'une population. Dans un premier temps, nous nous sommes intéressés à l'évolution à court terme (section IV.1). Nous avons donc estimé la contribution de l'épistasie à la covariance parents-enfants, qui permet de prédire la réponse à la sélection. Dans ce cas-là, nous avons dû utiliser les variances et covariances génétiques, et avons donc utilisé le modèle G2A-D (présenté dans le paragraphe II.4.2.3). Dans une deuxième partie (section IV.2), avons étudié le rôle de l'épistasie fonctionnelle dans l'évolution à long terme. Nous n'avons pas utilisé de modèle pour estimer des effets génétiques, puisqu'il s'agissait d'un travail fondé sur des simulations et des calculs numériques et formels. Cependant, on peut noter que nous nous sommes intéressés aux effets génétiques fonctionnels, c'est-à-dire à ceux définis par un modèle fonctionnel (comme le modèle UWR) et ne dépendant que des valeurs génotypiques.

PARTIE III

RECHERCHE D'EPISTASIE DANS LA REGULATION DE L'EXPRESSION CHEZ *TRITICUM TURGIDUM*

III. Recherche d'epistasie dans la regulation de l'expression chez *Triticum turgidum*

Dans cette partie, nous avons étudié les régulations de l'expression génétique chez le blé tétraploïde, *Triticum turgidum*. Les lignées étudiées dans ce cadre proviennent d'une population de pre-breeding de blé dur. Avant de commencer l'étude, nous avons voulu étudier en détails les caractéristiques de notre échantillon.

III.1 Description du matériel végétal

III.1.1. Objectif : l'utilisation d'une population de pre-breeding pour une étude de d'association pangénomique (GWAS)

Nous possédons au laboratoire une population de pre-breeding de blé dur, à base génétique large, ségrégeant pour un gène de stérilité mâle. L'objectif de la partie III.1 est de l'étudier avec des marqueurs microsatellites afin de déterminer si elle est utilisable pour une étude de GWAS.

III.1.1.1. Les populations de pre-breeding

Alors que les espèces domestiques ont souvent perdu une grande partie de leur diversité génétique (Buckler et al., 2001), l'introduction de diversité ancestrale ou sauvage dans le compartiment cultivé peut être une solution pour créer un matériel génétique original, de base génétique élargie. On appelle pre-breeding ce transfert de diversité, d'apparentés sauvages ou anciens, vers du matériel élite (FAO, 1996). Plusieurs méthodes existent pour valoriser la diversité génétique sauvage ou ancestrale : l'introgression et l'incorporation (Simmonds, 1993). L'introgression désigne le transfert d'un ou deux allèles de génotypes sauvages vers des variétés élites par des rétrocroisements. Cette approche nécessite une connaissance *a priori* de l'architecture génétique du trait à améliorer. L'incorporation, quant à elle, désigne le fait d'élargir la base génétique du matériel élite, en utilisant des génotypes exotiques. Elle ne nécessite pas de connaître l'architecture génétique, parce que l'incorporation est massive et non ciblée. Cette incorporation peut être effectuée en faisant des croisements entre des individus cultivés et des individus sauvages ou élites, mais cette méthode est limitée par la capacité technique à effectuer des croisements. Une solution plus adaptée est de créer une population de pre-breeding, comme celle cultivée au laboratoire. Les populations de pre-breeding sont des populations à base génétique large, qui subissent une légère sélection anthropique pour des traits d'intérêt agronomique, en plus de la sélection naturelle due à l'interaction avec l'environnement. Il y a déjà

plusieurs décennies, il a été démontré que la diversité génétique pouvait être valorisée pour l'amélioration variétale, en créant des populations issues de croisements entre un grand nombre de parents, et en favorisant la recombinaison (Suneson, 1956). Ce type de population a été nommé « populations de croisements composites » et leur principe est globalement le même que celui des populations de pre-breeding : rassembler divers stocks de semence, les croiser, et en sélectionner le produit (Phillips and Wolfe, 2005). La différence est que dans le cas des populations de pre-breeding, l'accent est mis sur la provenance exotique de certains parents. Ces populations à base génétique large peuvent aussi servir à conserver les ressources génétiques : on parle alors de « gestion dynamique » (David et al., in press; Enjalbert et al., 1998; Goldringer et al., 2001, 2001; Henry et al., 1991; Porcher et al., 2004). Contrairement aux populations de pre-breeding, les populations de gestion dynamique ne sont pas nécessairement soumises à sélection anthropique. Si elles sont utilisées dans un but de conservation des ressources, la sélection anthropique n'est pas nécessaire. En revanche, si elles sont orientées également vers un objectif d'amélioration variétale (pre-breeding), un minimum de sélection peut s'avérer nécessaire, chez certaines espèces. Par exemple, chez le blé tendre, les populations sans sélection anthropique tendent à évoluer vers des grandes tailles, ce qui est un phénotype peu adapté à la culture (Boulc'h et al., 1994; Goldringer et al., 1998). Par ailleurs, les populations de gestion dynamique sont cultivées sur plusieurs sites, présentant des conditions environnementales diverses, afin de permettre une plus grande diversité via la présence de différentes forces de sélection locales (David et al., 1997; Goldringer et al., 1998, 2001; Lavigne et al., 2001; Paillard et al., 2000a, 2000b; Porcher et al., 2004). Enfin, en plus de servir à l'amélioration variétale et à la conservation de la diversité cultivée, ces populations à bases génétiques larges, peuvent être des outils scientifiques pour étudier des phénomènes évolutifs (Lavigne et al., 2001; Porcher et al., 2004, 2006). En résumé, les populations à base large peuvent être utilisées à différentes fins : amélioration variétale (population composite et de pre-breeding), en utilisant ou non des ressources génétiques ancestrales ou sauvages, conservation de la diversité ou les deux à la fois.

Dans une population de pre-breeding d'une espèce autogame, l'utilisation d'un gène de stérilité mâle peut s'avérer nécessaire pour augmenter artificiellement le taux d'allofécondation (Enjalbert et al., 1998), et ceci pour plusieurs raisons. Premièrement, en l'absence d'allofécondation, le flux d'allèles entre individus d'une population sera très faible, et l'efficacité de l'incorporation d'allèles sauvages ou ancestraux dans le comportement élite sera limitée. Deuxièmement, l'allofécondation augmente le taux de recombinaison efficace, c'est-à-dire le taux de recombinaison entre régions chromosomiques différentes (Nordborg, 2000; Pollak, 1987). Or, l'introgression massive de diversité exotique dans le compartiment élite induit la formation de longs blocs de déséquilibre de liaison (Feuillet et al., 2008). La recombinaison est capable de casser ces blocs et de permettre une introgression plus ciblée. La

recombinaison a d'autres avantages. Elle permet de réduire l'effet du fond génétique (Rice, 2002) : ainsi, un allèle bénéfique, apparu dans un fond génétique mauvais, aura plus de chances d'être conservé en présence de recombinaison ; à l'inverse, une mutation défavorable apparue dans un bon fond génétique aura plus de chance d'être éliminée s'il y a de la recombinaison. Enfin, la recombinaison limite la perte de diversité autour de locus sélectionnés (effet de balayage sélectif) et donc augmente le polymorphisme de ces régions (Marais and Charlesworth, 2003; Orr, 2005). Augmenter le taux d'allogamie a un autre intérêt pour les populations à base génétique large : en effet, les espèces autogames ont en générale une taille efficace plus faible que les espèces allogames (Pollak, 1987), ce qui induit une plus grande perte de diversité par dérive (Wang and Caballero, 1999). Enfin, l'allofécondation permet de générer un plus grand nombre de combinaisons alléliques, générant une plus grande variation génotypique, utilisable pour augmenter le potentiel adaptatif de la population.

III.1.1.2. L'utilisation en association pangénomique

Plusieurs éléments doivent être pris en compte pour l'utilisation en association pangénomique (GWAS : Genome-wide association study) : le déséquilibre de liaison, la structuration et le polymorphisme de l'échantillon.

Le déséquilibre de liaison est au cœur des études d'association, puisque leur principe est de détecter des variants en déséquilibre de liaison avec le polymorphisme au locus causal. Si l'on ne dispose pas d'un grand nombre de marqueurs, un déséquilibre de liaison fort peut être avantageux, pour permettre de détecter des QTL dans des zones de grandes taille (Rafalski, 2010). Mais dans le cadre des études d'association pangénomique, un grand nombre de marqueurs est utilisé. Dans ce cas, un fort déséquilibre de liaison est désavantageux, puisqu'il réduit la précision de l'étude d'association. En effet, cela implique que des marqueurs éloignés du locus causal, mais appartenant au même bloc de déséquilibre de liaison, seront détectés comme ayant un effet significatif ; autrement dit, tous les marqueurs appartenant au bloc de déséquilibre de liaison auront statistiquement un effet significatif, et il ne sera pas possible de déterminer la position exacte du locus causal. Un déséquilibre de liaison réduit permet au contraire de localiser avec une plus grande précision les QTL.

Le déséquilibre de liaison est utilisé en génétique d'association pour permettre de localiser un locus causal. C'est donc le déséquilibre de liaison dû à une liaison physique qui est pertinent. Or un déséquilibre de liaison peut être induit par d'autres phénomènes : structuration, sélection, dérive, etc. (voir l'**encadré 1**). Ces sources de déséquilibre de liaison autre que la liaison physique peuvent donc créer des erreurs et doivent être prises en compte dans l'analyse.

Ainsi, la structuration peut perturber les résultats d'une étude d'association pangénomique, en créant un tel déséquilibre de liaison (Mackay and Powell, 2007). Deux types de structurations peuvent exister :

Encadré 1 : Phénomènes démographiques et évolutifs à l'origine du déséquilibre de liaison (Mackay & Powell 2007)

Mutation
L'apparition d'une mutation engendre du déséquilibre de liaison, puisqu'elle n'est présente que dans un seul haplotype. Au cours des générations, la recombinaison réduira ce déséquilibre de liaison, mais cela prendra un certain temps, en particulier pour les locus proches de la mutation.

Dérive génétique
Toute population finie génère un certain degré de déséquilibre de liaison. Par échantillonnage, certains allèles d'un locus seront plus souvent associés à certains allèles d'un autre locus.

Sélection
La sélection peut créer du déséquilibre de liaison de deux façons :

- par balayage sélectif : la région entourant le locus à l'origine du trait sélectionné est elle aussi sélectionnée par entraînement. Cela entraine une perte de polymorphisme dans l'ensemble de la région et une augmentation du déséquilibre de liaison.
- par épistasie : quand plusieurs locus interagissent ensemble pour former un phénotype avec la meilleure valeur sélective, c'est le génotype portant la combinaison de ces allèles qui sera sélectionné, créant ainsi un déséquilibre de liaison entre ces locus.

Structure
Il y aura du déséquilibre de liaison s'il y a une structure dans la population, avec deux sous-populations aux fréquences alléliques différentes, les croisements ayant préférentiellement lieu entre individus de la même sous-population.

une « macro-structure » qui correspond à une structuration en quelques grandes sous-populations, et une « micro-structure » qui correspond à l'apparentement entre individus (Yu et al., 2006). Les conséquences de ces structurations – lorsqu'elles ne sont pas prises en compte – ont bien été documentées (Helgason et al., 2005; Newman et al., 2001) : de fausses associations peuvent être détectées, ou à l'inverse, de véritables QTL peuvent ne pas être détectés. Un moyen de résoudre ce problème est d'intégrer des matrices de structures et d'apparentement au modèle, afin de définir la covariance génétique entre individus (Yu et al., 2006).

Dans le même ordre d'idée : la dérive peut également engendrer un déséquilibre de liaison entre locus non liés (Mackay and Powell, 2007). Une étude d'association pangénomique nécessite donc d'utiliser une population avec une taille efficace raisonnable.

Enfin, le niveau de polymorphisme en ségrégation dans la population est également important. Pendant longtemps, la recherche de QTL s'est cantonnée à des populations bi-parentales (généralement des F2, des rétrocroisements, ou des lignées recombinantes). Augmenter le nombre de parents permet d'augmenter la diversité génétique, et donc, les chances qu'un QTL soit polymorphe (et donc détectable). C'est l'idée des populations MAGIC, dont le concept a été développé il y a quelques années (Cavanagh et al., 2008). Par ailleurs, dans les populations biparentales de type F2 ou dérivant de rétrocroisements, le déséquilibre de liaison est élevé du fait d'un faible niveau de recombinaison. La recherche de QTL y est donc peu précise. Dans les populations MAGIC, le déséquilibre de liaison est plus faible, permettant de localiser de manière plus précise les QTL. Ces population MAGIC constituent donc un matériel de choix pour les études d'association pangénomique : un faible déséquilibre de liaison, pas de structure, et une forte diversité génétique (Cavanagh et al., 2008; Mackay and Powell, 2007).

La population de blé dur cultivée au laboratoire ressemble aux populations MAGIC. En effet, elle est issue de parents divers (blé dur *T. t. durum*, amidonnier sauvage *T. t. dicoccoides* et amidonnier domestique *T. t. dicoccum*), et son taux d'allofécondation a été fixé – à l'aide d'un gène de stérilité mâle - à 20%, soit à un niveau bien plus élevé que dans les populations naturelles, qui est de l'ordre de 1-4% (Tsegaye, 1996). Nous supposons que ce taux d'allofécondation relativement élevé a permis de réduire le niveau de déséquilibre de liaison. Dans des espèces autogames, comme le blé ou l'orge, il a été montré que l'allofécondation gérée par la stérilité mâle, pouvait permettre de réduire le déséquilibre de liaison.

L'objectif de la section III.1 est d'étudier en détails cette population afin de déterminer si elle est adaptée à une étude de recherche d'association pangénomique. Nous avons donc eu plusieurs objectifs :

1. L'étude de la macro-structure de la population
2. L'étude du régime de reproduction
    a. Pour comprendre l'origine de la structure, s'il y en a une
    b. Pour déterminer si le taux d'allofécondation est bien fixé à 20%
3. L'étude de la dérive : estimation de la taille efficace de la population

III.1.2 Description de la population étudiée

III.1.2.1. Fonctionnement de la population étudiée

En 1997, une population à base génétique large a été créée en intégrant dans une population de blé dur (*Triticum turgidum* ssp *durum*) fournie par F. Kaan (INRA), des variétés anciennes et sauvages de *Triticum turgidum,* notamment de l'amidonnier sauvage (*Triticum turgidum* ssp. *dicoccoides*) et domestique (*Triticum turgidum* ssp. *dicoccum*). D'autres sous-espèces de *Triticum turgidum* ont également été incorporées, notamment *Triticum turgidum turgidum*. Un gène de stérilité mâle, nucléaire et récessif, est présent dans la population car il ségrégerait dans la population de blé dur d'origine. Cette population a un effectif démographique d'environ 4 000 individus.

A floraison, les plantes mâles stériles sont repérées par des glumes bâillantes et la stérilité est confirmée visuellement par un l'état plumeux du stigmate, caractéristique d'une pollinisation qui n'a pas encore eu lieu. La récolte a lieu indépendamment sur les plantes mâles stériles et sur les hermaphrodites. Pour former la population de l'année suivante, 20% des graines semées proviennent des plantes mâles stériles (environ 150 plantes), le reste provenant des individus hermaphrodites (environ 2000-3000 individus). Depuis 2002, 10% des graines proviennent de lignées performantes tirées d'un programme de sélection INRA. Ces lignées, encore en cours de fixation, sont issues de croisements impliquant différentes sous-espèces de *T. turgidum*, suivis de 4 ou 5 générations d'autofécondation ; elles sont donc fortement homozygotes, mais ne sont pas encore complètement fixées. Les graines provenant de la même lignée ont donc des génotypes très proches. Chaque année, la lignée en provenance du programme de sélection inclue dans la population est différente et provient d'un croisement différent. Cette incorporation de lignées INRA correspond à une migration dans la population, et ces lignées seront dorénavant appelées « lignées migrantes ».

III.1.2.2. Echantillonnage et marqueurs génétiques utilisés

La population est reconduite d'année en année suivant le même protocole. Pour notre expérimentation, une extraction de matériel a été tirée des populations en 2004 et 20009. Un épi a été récolté sur 45 plantes hermaphrodites en 2004, et sur 192 plantes hermaphrodites en 2009. Une graine par épi a ensuite été mise à germer et son ADN a été extrait selon un protocole adapté de Tai and Tanksley, 1990. Nous avons donc deux échantillons : l'un représentatif de la génération de 2005,

Nom du microsatellite	Groupe de liaison	Publication	Génotypage de la population de 2010	Génotypage de la population de 2005 et des migrants
Xgpw3142	1A	Sourdille et al, 2009	X	X
Xgwm164	1A	Korzun V et al, 1997/ Roder MS et al, 1998	X	
Xgpw7577	1B	Sourdille et al, 2009	X	X
Xgwm413	1B	Roder MS et al, 1998	X	
Xgpw7101	2A	Sourdille et al, 2009	X	X
Xgwm312	2A	Roder MS et al, 1998	X	
Xgmw614a	2A	Roder MS et al, 1998	X	X
Xgwm374	2B	Roder MS et al, 1998	X	X
Xgwm2	3A	Roder MS et al, 1995 / Roder MS et al, 1998	X	
Xgwm299	3B	Roder MS et al, 1998	X	X
Xgwm285	3B	Roder MS et al, 1998 / Somers DL et al, 2003 / Somers DL et al, 2004	X	
Xgpw2239	4A	Sourdille et al, 2009	X	X
Xgwm601	4A	Roder MS et al, 1998	X	
Xgwm495	4B	Roder MS et al, 1998	X	
Xwmc710	4B	Somers DJ et al, 2004	X	
Xgwm126	5A	Roder MS et al, 1998	X	
Xgwm213	5B	Roder MS et al, 1998	X	
Xgwm234	5B	Korzun V et al, 1997/ Roder MS et al, 1998	X	
Xgpw3101	6A	Sourdille et al, 2009	X	
Xgmw193	6B	Korzun V et al, 1997/ Roder MS et al, 1998	X	X
Xgpw2103	7A(-7B)	Sourdille et al, 2004	X	X
Xgwm297	7B	Korzun V et al, 1997/ Roder MS et al, 1998	X	
Xgwm537	7B	Roder MS et al, 1998	X	

Tableau 3 : Nom, groupe de liaison et source des microsatellites utilisés

Une croix X indique que le microsatellite a été utilisé pour le génotypage. Les amorces correspondantes sont présentées en Annexe VII.3.

et l'autre de la génération de2010. L'échantillon de 2010 a été génotypé avec 23 microsatellites non liés ; celui de 2005 avec seulement 9 microsatellites (appartenant aux 23 utilisés pour l'échantillon de 2010). Les noms et caractéristiques des microsatellites utilisés pour l'étude sont présentés dans le **tableau 3**. Par ailleurs, des lignées migrantes F6 ont également été génotypées notamment la lignée 98PO26IN, introduite en 2003, qu'on appellera Migrant 1; et la lignée 01PODD41.5, introduite en 2007, qu'on nommera Migrant 2. Ces lignées migrantes ont été génotypées avec les neuf microsatellites utilisés pour l'analyse de la population de 2005.

III.1.3 Structure de la population

III.1.3.1. Détermination de la structure de la population en 2010 à l'aide d'une analyse factorielle des correspondances (AFC) et d'un dendrogramme

Deux méthodes ont été utilisées pour déterminer la structure de la population en 2010, en utilisant les 23 marqueurs microsatellites avec lesquels elle a été génotypée. Premièrement, une AFC (Analyse Factorielle des Correspondances) sur trois axes a été effectuée à l'aide du logiciel Genetix (Belkhir et al., 1996). Ensuite, un dendrogramme a été réalisé selon la méthode de Ward (Ward, 1963), à partir des distances euclidiennes entre individus, calculées grâces aux données génotypiques. La réalisation de ce dendrogramme a nécessité l'utilisation du package pvclust de R (Suzuki and Shimodaira, 2006). Ce même package permet de calculer des p-values pour chaque embranchement du dendrogramme. Le calcul de ces p-values, approximativement sans biais, s'est effectué par un ré-échantillonnage bootstrap multi-échelle. Les résultats de ces deux méthodes sont présentés sur la **Figure 12.**

Les deux premiers embranchements du dendrogramme (Figure 12.A) sont stables après rééchantillonnage (présents dans au moins 95% des bootstraps). La population a donc été divisée en trois sous-groupes, correspondants aux premier et deuxième embranchements, et au reste de la population. L'AFC (Figure 12.B) confirme ces groupes, qui sont bien séparés du reste de la population selon les trois premiers axes. L'un de ces groupes comprend 9 individus. Il est figuré en bleu sur la figure 12, et sera nommé « groupe 1 ». L'autre groupe comprend 23 individus. Il est figuré en jaune sur la figure 12 et sera appelé « groupe 2 ». Le reste de la population comprend 160 individus et sera dorénavant appelé « cœur de population ».

III.1.3.2. A l'origine de la structuration : la migration

Nous avons accès, du fait que la population ait été génotypées à deux époques (2005 et 2010), à des données de fréquences alléliques séparées de 5 années. Ces données nous ont permis d'établir un modèle dynamique d'évolution, plus riche qu'une classification à une année donnée. Ce modèle peut servir à différentes fins, comme par exemple, prédire l'évolution d'une population suite à un

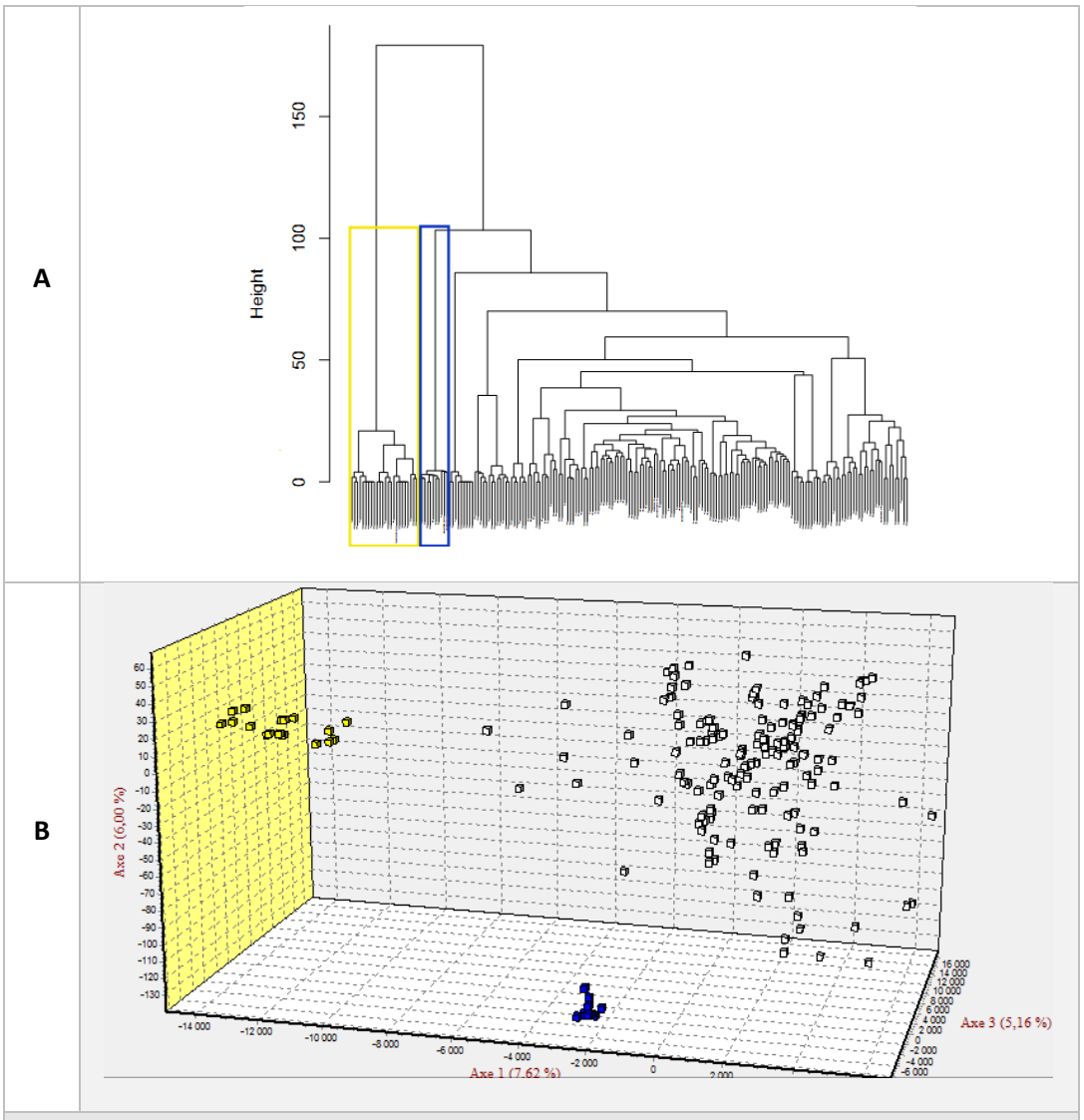


Figure 12 : Structure de l'échantillon de 2010, réalisée avec 23 marqueurs microsatellites

A. Dendrogramme
B. AFC. Le premier axe représente 7,62% de la variance, le second 6,00% et le troisième 5,16% de la variance.

Les couleurs correspondent aux groupes définis par le dendrogramme. Jaune = groupe 2 (23 individus) ; bleu = groupe 1 (9 individus) ; blanc = reste de la population (160 individus).

évènement de migration, en fonction de plusieurs paramètres. Ainsi, dans notre cas, ce modèle nous a servi à identifier des individus descendant d'une lignée migrante.

Lorsque les génotypes microsatellites des migrants 1 et 2 sont inclus dans l'AFC, ils sont assignés chacun à l'un des deux petits groupes : le migrant 1 est assigné au groupe 1, et le migrant 2 au groupe 2. Cela suggère que la structuration de la population est due à l'introduction par migration des lignées du programme d'amélioration INRA. Les individus proches de ces migrants correspondent donc aux descendants par autofécondation des graines qui ont été incorporées quelques années auparavant. Pour pouvoir détecter et suivre avec plus de précision les descendants des migrants par autofécondation ou intégré par allofécondation grâce au gène de stérilité mâle, nous avons développé un modèle permettant de suivre l'évolution d'une population au régime de reproduction mixte après un épisode de migration. Néanmoins, nous n'avons pu appliquer ce modèle qu'au migrant 2, introduit en 2007, car il est nécessaire d'avoir une estimation des fréquences alléliques dans la population, avant introduction du migrant, comme nous l'expliquerons ci-dessous.

IIII.1.3.2.1 Déterminer si un individu descend d'un migrant ou non

Soit G_i le génotype d'un individu i de l'échantillon. Un individu descendant d'un migrant est appelé « M ». S'il ne descend pas d'un migrant, il est nommé « NM ». Les probabilités $P(G_i/M)$ et $P(G_i/NM)$ correspondent aux probabilités qu'un individu porte le génotype G_i sachant qu'il descend d'un migrant, ou sachant qu'il ne descend pas d'un migrant. En utilisant la formule de Bayes, il est possible de déterminer la probabilité qu'un individu descende d'un migrant, sachant son génotype :

$$P(M/G_i) = \frac{P(G_i/M)\times p_M}{P(G_i)} = \frac{P(G_i/M)\times p_M}{P(G_i/M)\times p_M + P(G_i/NM)\times(1-p_M)} \quad \textbf{(1)}$$

Avec p_M la probabilité de descendre d'un migrant et $P(G_i)$ la probabilité du génotype.

IIII.1.3.2.2 Modèle d'estimation de la probabilité d'un génotype sachant qu'il descend d'un migrant

$P(M/G_i)$ ne peut être calculé que si l'on connait $P(G_i/M)$ et $P(G_i/NM)$. Pour estimer ces probabilités, un modèle a été développé pour inférer ces probabilités.

IIII.1.3.2.2.1 Principe du modèle

L'idée générale du modèle est d'estimer la probabilité (la fréquence) d'un génotype dans deux sous-populations : la sous-population « M », constituée d'individus ayant au moins un migrant dans leurs ancêtres ; et la population « NM » constituée d'individus sans aucun migrant dans leur généalogie.

Pour cela, nous modélisons un phénomène de migration, survenant à la génération 1 dans la population. Par soucis de simplification, on suppose que cette introduction de migrants, à une génération initiale, est le seul phénomène de migration existant dans l'évolution de la population.

Le modèle permet d'estimer l'évolution des fréquences génotypiques au cours du temps, en connaissant la composition génotypique des migrants, et la composition génétique de la population avant l'évènement de migration. D'une part, on connait les fréquences initiales de la sous-population « M » si la lignée migrante a été génotypée : comme il s'agit de lignées majoritairement fixée, nous considérons que la sous-population migrante n'est constituée que d'un seul génotype. D'autre part, un échantillon de la population, datant d'avant l'évènement de migration qui nous intéresse, sert à estimer les fréquences génotypiques de la population avant l'introduction du migrant.

Nous nous sommes intéressés au migrant 2, introduit lors du semis de 2006. Les plantes semées en 2006 ont été récoltées en 2007. La génération 1 correspond donc à l'année de récolte 2007. L'échantillon de la population récolté en 2005 nous a ainsi servi à estimer les fréquences génotypiques dans la population avant son introduction. L'échantillon de la population récoltée 2010 correspond donc à la population à la génération 4 après introduction.

Nous n'avons pas pu utiliser ce modèle pour le migrant 1 (introduit en 2003) car la population n'a pas été échantillonnée et génotypée avant cette date. Nous n'avions donc aucun moyen d'estimer les fréquences génotypiques dans la population avant son introduction.

Un individu « M » ne peut avoir que des descendants « M ». A l'inverse un individu « NM » peut donner des descendants « M » s'il se croise avec un « M ».

IIII.1.3.2.2.2 Evolution des fréquences génotypiques à deux locus (le locus de stérilité mâle et un locus indépendant)

Ici nous détaillerons le modèle pour deux locus : le locus de stérilité mâle, et un locus indépendant représentant un marqueur. Le modèle nécessite de calculer les fréquences génotypiques dans la population à la génération $g+1$, connaissant les fréquences à la génération g.

Soit F, le locus du gène de stérilité mâle. Comme il est récessif, les plantes portant les génotypes FF et Ff sont hermaphrodites, tandis que les plantes portant le génotype ff sont des femelles. Les plantes hermaphrodites se reproduisent majoritairement par autogamie, tandis que les femelles ne se reproduisent que par allofécondation. Par simplification, nous considérerons que le taux d'allofécondation résiduelle des hermaphrodites est nul.

Soit A un autre locus, représentant un marqueur. Dans notre exemple, il ne porte que deux allèles : A et a. On suppose que les fréquences alléliques au locus A sont identiques chez les femelles et les

hermaphrodites. Par ailleurs, on suppose que les deux locus, A et F, ne sont pas liés physiquement et sont en équilibre de liaison

Soit α la proportion de graines récoltées sur des femelles, parmi celles semées chaque année. Ainsi, une proportion α d'individus dans la population provient d'allofécondation, et une proportion 1 - α d'autofécondation. Les migrations annuelles ne sont pas inclues dans le calcul des fréquences génotypiques.

Soient :

- $P_{jklm}^{(g)}$ la fréquence du génotype bilocus $jklm$ dans la population, à la $g^{\text{ième}}$ génération, avec $j, k = \{F, f\}$ et $l, m = \{A, a\}$.
- $P_{jk}^{(g)}$ la fréquence du génotype jk au locus F dans la population, à la $g^{\text{ième}}$ génération, avec $j, k = \{F, f\}$
- $P_{lm}^{(g)}$ la fréquence du génotype lm au locus A dans la population, à la $g^{\text{ième}}$ génération, avec $l, m = \{A, a\}$
- $P_{j}^{(g)}$ la fréquence de l'allèle j dans la population, avec $j = \{F, f\}$
- $P_{l}^{(g)}$ la fréquence de l'allèle l dans la population, avec $l = \{A, a\}$
- $P_{j,o}^{(g)}$ la fréquence de l'allèle j parmi les gamètes femelles (oosphères) de la population, avec $j = \{F, f\}$
- $P_{l,o}^{(g)}$ la fréquence de l'allèle l parmi les gamètes femelles (oosphères) de la population, avec $l = \{A, a\}$

Les fréquences alléliques chez les gamètes sont identiques aux fréquences alléliques dans la population : $P_{l,o}^{(g)} = P_{l}^{(g)}$. Soient $\boldsymbol{\Pi}$ le vecteur des fréquences génotypiques. Comme il y a neuf génotypes possibles dans la population, il est de dimension (9 x 1). Les fréquences génotypes à la génération g+1 sont égales à :

$$\boldsymbol{\Pi}^{(g+1)} = \frac{\alpha \boldsymbol{O} + (1-\alpha)\boldsymbol{S}}{P_{FF}^{(g)} + P_{Ff}^{(g)}} \boldsymbol{\Pi}^{(g)}$$

Avec

$$\mathbf{\Pi}^{(g)} = \begin{bmatrix} P_{ffaa} \\ P_{ffaA} \\ P_{ffAA} \\ P_{fFaa} \\ P_{fFaA} \\ P_{fFAA} \\ P_{FFaa} \\ P_{FFaA} \\ P_{FFAA} \end{bmatrix}^{(g)}$$

$$\boldsymbol{O} = \begin{bmatrix} 0 & 0 & 0 & P_{a,o}^{(g)}/2 & P_{a,o}^{(g)}/4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & P_{A,o}^{(g)}/2 & 1/4 & P_{a,o}^{(g)}/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & P_{A,o}^{(g)}/4 & P_{A,o}^{(g)}/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & P_{a,o}^{(g)}/2 & P_{a,o}^{(g)}/4 & 0 & P_{a,o}^{(g)} & P_{a,o}^{(g)}/2 & 0 \\ 0 & 0 & 0 & P_{A,o}^{(g)}/2 & 1/4 & P_{a,o}^{(g)}/2 & P_{a,o}^{(g)} & 1/2 & P_{a,o}^{(g)} \\ 0 & 0 & 0 & 0 & P_{A,o}^{(g)}/4 & P_{A,o}^{(g)}/2 & 0 & P_{A,o}^{(g)}/2 & P_{A,o}^{(g)} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

et

$$\boldsymbol{S} = \begin{bmatrix} 0 & 0 & 0 & 1/4 & 1/16 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/16 & 1/4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/8 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 1/16 & 0 & 1 & 1/4 & 0 \\ 0 & 0 & 0 & 0 & 1/8 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 1/16 & 1/4 & 0 & 1/4 & 1 \end{bmatrix}$$

$\boldsymbol{O}$ et $\boldsymbol{S}$ sont des matrices de transition : $\boldsymbol{O}$ correspond à l'allofécondation, et $\boldsymbol{S}$ à l'autofécondation. Le détail des calculs de ces matrices est présenté en **Annexe VII.4**. La matrice de transition $\boldsymbol{O}$ donne les probabilités pour un génotype paternel donné, de produire par allofécondation un descendant parmi les 9 génotypes possibles. Ainsi pour les génotypes paternels $ffaa$, $ffaA$ et $ffAA$ qui correspondent

aux trois premières colonnes de la matrice de transition, on constate que toutes les probabilités sont nulles car ces génotypes sont celui d'une femelle. Prenons l'exemple du génotype $fFaa$ (4ième colonne). Ce génotype a une probabilité 1/2 de former un grain de pollen porteur du génotype fa. Ce grain de pollen a ensuite une probabilité $P_{a,o}^{(g)}$ de rencontrer un oosphère porteur d'un génotype fa (les femelles donnant forcément des gamètes avec l'allèle f au locus F : $P_{fa,o}^{(g)} = P_{f,o}^{(g)} P_{a,o}^{(g)} = P_{a,o}^{(g)}$). Par conséquent, la probabilité qu'un père $fFaa$ forme par allofécondation un descendant $ffaa$ (1ère ligne de la matrice de transition O) est égale à $P_{a,o}^{(g)}/2$: c'est donc ce coefficient que l'on va trouver à la 1ère ligne et 4ème colonne. La probabilité pour ce même génotype paternel $fFaa$ de former un individu $ffaA$ (deuxième ligne) est égale à $P_{A,o}^{(g)}/2$, celle de former un individu $fFaa$ (quatrième ligne) est égale à $P_{a,o}^{(g)}/2$, et enfin, celle de former un individu $fFaA$ (cinquième ligne) est égale à $P_{A,o}^{(g)}/2$. La somme des probabilités par colonne vaut bien 1. Il en est de même pour la matrice $\boldsymbol{S}$, mais dans un contexte d'autofécondation, donc avec des fréquences alléliques dans les oosphères qui ne dépendent pas de la composition de la population. Comme les individus femelles ne produisent pas de pollen, il faut normaliser par la proportion d'individus hermaphrodites ($P_{FF}^{(g)} + P_{Ff}^{(g)}$) pour obtenir le vecteur des fréquences génotypiques à la génération suivante.

IIII.1.3.2.2.3 Evolution des fréquences dans les sous-populations « M » et « NM »

Une fois que l'on sait comment calculer les fréquences d'une génération à l'autre, le modèle peut être utilisé pour regarder comment ces fréquences évoluent dans les deux sous-populations « M » et « NM ». Un individu sera considéré comme « M » si au moins l'un de ses deux parents est « M » ; sinon il sera « NM ». Appelons $P_{jklm,M}^{(g)}$ la fréquence du génotype $jklm$ parmi la sous-population M ; et appelons $P_{jklm,NM}^{(g)}$ la fréquence du génotype $jklm$ parmi la sous-population NM.

IIII.1.3.2.2.3.1 Les vecteurs initiaux

Les vecteurs $\mathbf{\Pi}_M^{(g)}$ et $\mathbf{\Pi}_{NM}^{(g)}$constituent les vecteurs des fréquences génotypiques au sein de chaque sous-population. Soit $p_M^{(g)}$ la probabilité d'appartenir à la population « M » à la génération g, et soit $1 - p_M^{(g)}$ la probabilité d'appartenir à la population « NM ». Remarquez que les fréquences génotypiques dans la population globale sont égales à :

$$\mathbf{\Pi}^{(g)} = p_M^{(g)} \times \mathbf{\Pi}_M^{(g)} + (1 - p_M^{(g)}) \times \mathbf{\Pi}_{NM}^{(g)}$$

On note « 0 » la génération juste avant l'introduction du migrant dans la population. A ce moment-là, les sous-populations « M » et « NM » sont deux populations indépendantes : « NM » est la population de pre-breeding, « M » est une lignée INRA cultivée à part.

Les fréquences génotypiques au locus A, dans la population de pre-breeding (« NM »), à la génération 0, sont estimées à partir de l'échantillon de 2005 . On en déduit $P_{aa,NM}^{(0)}$, $P_{aA,NM}^{(0)}$ et $P_{AA,NM}^{(0)}$. Par ailleurs, après plusieurs générations, les fréquences génotypiques au locus F se stabilisent. On suppose que lors de l'introduction du migrant, la population est à l'équilibre quant au gène de stérilité mâle. En l'absence d'allofécondation résiduelle et de migration préalable, les fréquences génotypiques au locus F à l'équilibre sont égales à (voir **Annexe VII.4**, paragraphe VII.4.1.4):

$$P_{ff}^{\infty} = \frac{\alpha}{2}$$

$$P_{Ff}^{\infty} = \frac{(2-\alpha)\alpha}{(\alpha+1)}$$

$$P_{FF}^{\infty} = \frac{(2-\alpha)(1-\alpha)}{2(\alpha+1)}$$

Par conséquent $P_{ff,NM}^{(0)} = \frac{\alpha}{2}$, $P_{Ff,NM}^{(0)} = \frac{(2-\alpha)\alpha}{(\alpha+1)}$ et $P_{FF,NM}^{(0)} = \frac{(2-\alpha)(1-\alpha)}{2(\alpha+1)}$

Les fréquences génotypiques aux quatre locus peuvent être déduites en faisant un produit, car les deux locus sont supposés non liés : $P_{jklm,NM}^{(g)} = P_{jk,NM}^{(g)} \times P_{lm,NM}^{(g)}$

Les individus migrants portent l'allèle sauvage F au locus de stérilité mâle. Si la lignée migrante introduite dans la population porte l'allèle A au locus A, alors, au moment de l'introduction des individus migrants :

$$\mathbf{\Pi}_M^{(0)} = \begin{bmatrix} P_{ffaa} \\ P_{ffaA} \\ P_{ffAA} \\ P_{fFaa} \\ P_{fFaA} \\ P_{fFAA} \\ P_{FFaa} \\ P_{FFaA} \\ P_{FFAA} \end{bmatrix}_M^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

L'introduction du migrant est modélisée lors de la génération 1. Soit β le pourcentage d'individus migrants introduits dans la population. Par conséquent :

$$\mathbf{\Pi}^{(1)} = (1-\beta)\mathbf{\Pi}_{NM}^{(0)} + \beta\mathbf{\Pi}_M^{(0)}$$

On suppose qu'entre la génération 0 et la génération 1, les fréquences génotypiques au sein de la population « NM » n'ont pas changé. Par conséquent :$\mathbf{\Pi}_M^{(1)} = \mathbf{\Pi}_M^{(0)}$ et $\mathbf{\Pi}_{NM}^{(1)} = \mathbf{\Pi}_{NM}^{(0)}$. Notons que $p_M^{(1)} = \beta$.

IIII.1.3.2.2.3.2 Différents cas de reproduction

Pour simplifier les notations dans les calculs, nous noterons $P_{♀/M}^{(g)}$ et $P_{♀/NM}^{(g)}$ la proportion de femelles parmi les individus « M » et « NM », respectivement. De la même façon $P_{♂/M}^{(g)}$ et $P_{♂/NM}^{(g)}$ désignent la proportion d'hermaphrodites dans les deux sous-populations. Remarquez que : $P_{♀/M}^{(g)} = P_{ff,M}^{(g)}$; $P_{♀/NM}^{(g)} = P_{ff,NM}^{(g)}$; $P_{♂/M}^{(g)} = P_{fF,M}^{(g)} + P_{FF,M}^{(g)}$; $P_{♂/NM}^{(g)} = P_{fF,NM}^{(g)} + P_{FF,NM}^{(g)}$.

Notons aussi que la proportion totale de femelles dans l'ensemble de population est égale à :

$$P_{♀}^{(g)} = P_{ff}^{(g)} = p_M^{(g)} \times P_{♀/M}^{(g)} + (1 - p_M^{(g)}) \times P_{♀/NM}^{(g)}$$

De la même façon :

$$P_{♂}^{(g)} = p_M^{(g)} \times P_{♂/M}^{(g)} + (1 - p_M^{(g)}) \times P_{♀/NM}^{(g)}$$

Par ailleurs :

$$P_{M/♀}^{(g)} = \frac{P_{♀\cap M}^{(g)}}{P_{♀}^{(g)}} = \frac{p_M^{(g)} \times P_{♀/M}^{(g)}}{P_{♀}^{(g)}}$$

De la même façon :

$$P_{NM/♀}^{(g)} = \frac{(1 - p_M^{(g)}) \times P_{♀/NM}^{(g)}}{P_{♀}^{(g)}};\ P_{M/♂}^{(g)} = \frac{p_M^{(g)} \times P_{♂/M}^{(g)}}{P_{♂}^{(g)}}; P_{NM/♂}^{(g)} = \frac{(1 - p_M^{(g)}) \times P_{♂/NM}^{(g)}}{P_{♂}^{(g)}}$$

Un individu sera « M » dans deux cas :

- Il descend par autofécondation d'un individu M, forcément hermaphrodite
- Il descend par allofécondation de deux parents, dont l'un au moins est « M »

Les différents cas de reproduction possible (autofécondation d'un hermaphrodite « M », allofécondation d'une femelle « M » par un grain de pollen « NM » etc.) sont schématisés sur la **figure 13**, et détaillés ci-dessous :

Cas 1 : Le descendant est un « M » qui a été récolté sur une femelle « M » allofécondée par un grain de pollen quelconque. La probabilité du cas 1 (probabilité de descendre d'une femelle « M ») est égale à la probabilité d'être M quand on est femelle, fois la probabilité qu'une femelle soit récoltée pour renouveler la population :

$$P_1^{(g)} = \alpha P_{M/♀}^{(g)}$$

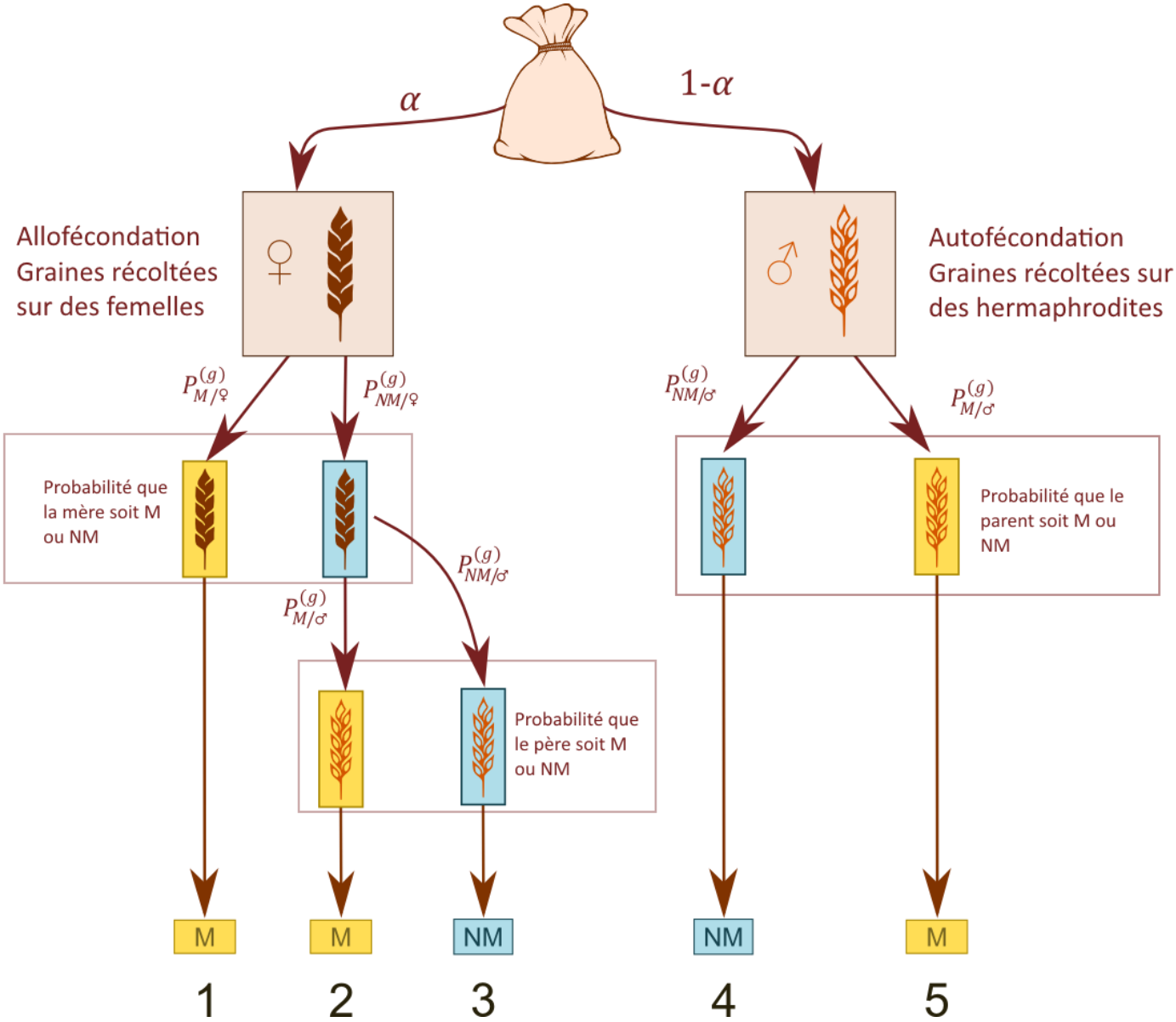


Figure 13 : Les différents cas de reproduction possible

Jaune = individus « M » ; bleu = individu « NM »

<u>Cas 1</u> : une femelle « M » donne un descendant « M ». Probabilité : $P_1^{(g)} = \alpha P_{M/♀}^{(g)}$

<u>Cas 2</u> : une femelle « NM » donne un descendant « M » avec un hermaphrodite « M ».

Probabilité : $P_2^{(g)} = \alpha P_{NM/♀}^{(g)} \times P_{M/♂}^{(g)}$

<u>Cas 3</u> : une femelle « NM » donne un descendant « NM » avec un hermaphrodite « NM ». Probabilité : $P_3^{(g)} = \alpha P_{NM/♀}^{(g)} \times P_{NM/♂}^{(g)}$

<u>Cas 4</u> : un hermaphrodite « NM » donne un descendant « NM ».

Probabilité : $P_4^{(g)} = (1-\alpha) P_{NM/♂}^{(g)}$

<u>Cas 5</u> : un hermaphrodite « M » donne un descendant « M ». Probabilité : $P_5^{(g)} = (1-\alpha) P_{M/♂}^{(g)}$

Cas 2 : Le descendant est un « M » qui a été récolté sur une femelle « NM » allofécondée par un grain de pollen « M ». La probabilité du cas 2 (probabilité de descendre d'une femelle NM et d'un père M) est égale à la probabilité qu'un père hermaphrodite soit « M », multiplié par la probabilité qu'une femelle soit « NM », multipliée à la probabilité qu'une femelle soit récoltée pour renouveler la population :

$$P_2^{(g)} = \alpha P_{NM/♀}^{(g)} \times P_{M/♂}^{(g)}$$

Cas 3 : Le descendant est un « NM » qui a été récolté sur une femelle « NM » allofécondée par un grain de pollen « NM ». La probabilité du cas 3 (probabilité de descendre d'une femelle « NM » et d'un père « NM ») est égale à la probabilité qu'un père hermaphrodite soit « NM », multiplié par la probabilité qu'une femelle soit « NM », multipliée par la probabilité qu'une femelle soit récoltée pour renouveler la population :

$$P_3^{(g)} = \alpha P_{NM/♀}^{(g)} \times P_{NM/♂}^{(g)}$$

Cas 4 : Le descendant est un « NM » qui a été récolté sur un hermaphrodite « NM », qui s'est autofécondé. La probabilité du cas 4 (probabilité de descendre d'un hermaphrodite « NM ») est égale à la probabilité qu'un hermaphrodite soit NM, multiplié par la probabilité qu'un hermaphrodite soit récolté pour renouveler la population :

$$P_4^{(g)} = (1-\alpha) P_{NM/♂}^{(g)}$$

Cas 5 : Le descendant est un « M » qui a été récolté sur un hermaphrodite « M », qui s'est autofécondé. La probabilité du cas 5 (probabilité de descendre d'un hermaphrodite « M ») est égale à la probabilité qu'un hermaphrodite soit « M », multiplié par la probabilité qu'un hermaphrodite soit récolté pour renouveler la population :

$$P_5^{(g)} = (1-\alpha) P_{M/♂}^{(g)}$$

On trouve donc que $p_M^{(g+1)} = P_1^{(g)} + P_2^{(g)} + P_5^{(g)} = \alpha(P_{M/♀}^{(g)} + P_{NM/♀}^{(g)} \times P_{M/♂}^{(g)}) + (1-\alpha) P_{M/♂}^{(g)}$

IIII.1.3.2.2.3.3 Matrices de transition

Comme les deux sous-populations sont étudiées indépendamment, les matrices de transitions ne sont pas toujours celles présentées dans le paragraphe IIII.1.3.2.2.2. Soient $\mathbf{\Pi}_x^{(g)}$ le vecteur des fréquences génotypiques des individus nés via le cas *x*.

Dans le cas d'autofécondation (cas 4 et 5), la matrice de transition $\boldsymbol{S}$ est maintenue.

$$\mathbf{\Pi}_4^{(g)} = \frac{\boldsymbol{S}}{P_{♂,NM}^{(g)}} \mathbf{\Pi}_{NM}^{(g)}$$

$$\mathbf{\Pi}_5^{(g)} = \frac{\boldsymbol{S}}{P_{♂,M}^{(g)}} \mathbf{\Pi}_{M}^{(g)}$$

Dans les deux cas, on divise par la fréquence des hermaphrodites ayant participé à la reproduction en produisant du pollen.

En revanche, dans les cas 1, 2 et 3, il faut utiliser une matrice d'allofécondation $\boldsymbol{O}$ modifiée. Dans le cas 1, une femelle « M » est fécondée par un hermaphrodite quelconque de la population. La matrice de transition, dénommée $\boldsymbol{O}_M$ est semblable à la matrice $\boldsymbol{O}$, exceptée qu'au lieu de contenir les fréquences alléliques des oosphères de toutes les femelles de la population, elle contient celles des femelles « M»

$$\boldsymbol{O}_M = \begin{bmatrix} 0 & 0 & 0 & P_{a,o/M}^{(g)}/2 & P_{a,o/M}^{(g)}/4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & P_{A,o/M}^{(g)}/2 & 1/4 & P_{a,o/M}^{(g)}/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & P_{A,o/M}^{(g)}/4 & P_{A,o/M}^{(g)}/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & P_{a,o/M}^{(g)}/2 & P_{a,o/M}^{(g)}/4 & 0 & P_{a,o/M}^{(g)} & P_{a,o/M}^{(g)}/2 & 0 \\ 0 & 0 & 0 & P_{A,o/M}^{(g)}/2 & 1/4 & P_{a,o/M}^{(g)}/2 & P_{A,o/M}^{(g)} & 1/2 & P_{a,o/M}^{(g)} \\ 0 & 0 & 0 & 0 & P_{A,o/M}^{(g)}/4 & P_{A,o/M}^{(g)}/2 & 0 & P_{A,o/M}^{(g)}/2 & P_{A,o/M}^{(g)} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{\Pi}_1^{(g)} = \frac{\boldsymbol{O}_M}{P_{♂}^{(g)}} \mathbf{\Pi}^{(g)}$$

Ici, on divise par la fréquence des hermaphrodites ($P_{♂}^{(g)} = P_{FF}^{(g)} + P_{Ff}^{(g)}$) au sein de la population totale. En effet, l'ensemble des hermaphrodites (« M » ou « NM ») participent à la reproduction dans le cas 1.

Dans les cas 2 et 3, une femelle « NM » est fécondé, soit par un hermaphrodite « M » dans le cas 2, soit par un hermaphrodite « NM » dans le cas 3. La matrice de transition, dénommée $\boldsymbol{O}_{NM}$ contient les fréquences alléliques des oosphères des femelles « NM»

$$\mathbf{\Pi}_2^{(g)} = \frac{\boldsymbol{O}_{NM}}{P_{♂,M}^{(g)}} \mathbf{\Pi}_{M}^{(g)}$$

$$\mathbf{\Pi}_3^{(g)} = \frac{\boldsymbol{O}_{NM}}{P_{♂,NM}^{(g)}} \mathbf{\Pi}_{NM}^{(g)}$$

Au final, on trouve que :

$$\mathbf{\Pi}^{(g+1)} = P_1^{(g)}\mathbf{\Pi}_1^{(g)} + P_2^{(g)}\mathbf{\Pi}_2^{(g)} + P_3^{(g)}\mathbf{\Pi}_3^{(g)} + P_4^{(g)}\mathbf{\Pi}_4^{(g)} + P_5^{(g)}\mathbf{\Pi}_5^{(g)}$$

$$\mathbf{\Pi}_M^{(g+1)} = \frac{P_1^{(g)}\mathbf{\Pi}_1^{(g)} + P_2^{(g)}\mathbf{\Pi}_2^{(g)} + P_5^{(g)}\mathbf{\Pi}_5^{(g)}}{P_1^{(g)} + P_2^{(g)} + P_5^{(g)}} = \frac{P_1^{(g)}\mathbf{\Pi}_1^{(g)} + P_2^{(g)}\mathbf{\Pi}_2^{(g)} + P_5^{(g)}\mathbf{\Pi}_5^{(g)}}{p_m^{(g+1)}}$$

$$\mathbf{\Pi}_{NM}^{(g+1)} = \frac{P_3^{(g)}\mathbf{\Pi}_3^{(g)} + P_4^{(g)}\mathbf{\Pi}_4^{(g)}}{P_3^{(g)} + P_4^{(g)}} = \frac{P_3^{(g)}\mathbf{\Pi}_3^{(g)} + P_4^{(g)}\mathbf{\Pi}_4^{(g)}}{1 - p_m^{(g+1)}}$$

IIII.1.3.2.2.4 Application aux données de la génération 2010

Le modèle a été appliqué pour déterminer quels étaient les descendants du migrant 2 parmi les individus de l'échantillon de 2010. Le migrant a été introduit en 2007 (génération 1) et l'échantillon prélevé en 2010. Nous nous sommes donc intéressés aux estimations des fréquences génotypiques à la génération 4, qui correspond à l'année 2010 : $\mathbf{\Pi}_{NM}^{(4)}$ et $\mathbf{\Pi}_M^{(4)}$.

Le migrant 2 a été génotypé sur neuf marqueurs microsatellites. Ainsi les fréquences génotypiques de neuf marqueurs A_1, A_2…A_9 ont pu être estimées, au sein des « M » et des « NM », au cours des 4 générations comprenant et suivant l'introduction du migrant. Ces marqueurs microsatellites possédaient aux moins trois allèles. Les calculs numériques effectués à partir des données de génotypages ne correspondent donc pas exactement aux calculs formels présentés ci-dessus, qui considéraient le cas d'un marqueur A bi-allélique. Cependant, le raisonnement est identique, et nous avons utilisé un modèle identique mais généralisé à k allèles.

Les paramètres utilisés dans nos calculs ont été α = 0,2 et β = 0,1. Nous avons trouvé qu'au bout de 4 générations, $p_M^{(4)} = 0{,}11$.

Pour calculer la fréquence d'un génotype G au sein d'une des sous populations « M » ou « NM », nous avons effectué le produit des fréquences en supposant que les différents marqueurs étaient en équilibre de liaison. Par exemple, soit un génotype trilocus A_1a_1 - A_2A_2 - a_3a_3. Sa fréquence parmi la sous-population « M » est égale à :

$$P_{G,M} = P_{A_1a_1,M} \times P_{A_2a_2,M} \times P_{a_3a_3,M}$$

Avec par exemple, $P_{A_1a_1,M} = P_{ffA_1a_1,M} + P_{fFA_1a_1,M} + P_{FFA_1a_1,M}$

Ainsi, pour chaque individu i de l'échantillon de 2010, nous avons pu estimer quelle était la fréquence de son génotype G_i chez les descendants de migrants (« M ») : $P^{(4)}_{G_i,M} = P(G_i/M)$, et dans la population globale $P^{(4)}_{G_i} = P(G_i)$. On a pu donc en déduire $P(M/G_i)$:

$$P(M/G_i) = \frac{P(G_i/M) \times p^{(4)}_M}{P(G_i)}$$

Les individus qui avaient une probabilité $P(M/G_i)$ supérieure à 0,6 furent considérés comme étant des descendants du migrant 2 (population « M »), tandis que les autres furent assignées à la catégorie « NM » (non descendants du migrant 2).

Nous avons ainsi pu constater que 29 individus de l'échantillon de 2010 pouvaient être considérés comme des descendants du migrant 2. Les 23 individus constituant le groupe 2 défini par le dendrogramme ont tous étaient assignés dans la catégorie « M ». Les six autres individus « M » appartiennent au cœur de population. Ces résultats semblent donc confirmer que la structuration de la population provient d'évènements de migration. La « population cœur », comprenant 160 individus, semblent bien représenter la population de base, dans laquelle figurent des descendants de lignées migrantes obtenues par allofécondation avec la population résidente.

Néanmoins, le fait que certains descendants de migrants semblent être restés des lignées indépendantes, tandis que d'autres se sont visiblement intégré dans la population, suggèrent également un rôle du régime de reproduction dans cette structuration, qui sera précisé en III.1.4.

III.1.3.2.2.5 Limites et intérêts du modèle

Le modèle comprend des limites. Premièrement l'échantillon de 2005, servant pour estimer les fréquences avant l'introduction du migrant 2, est très petit : seulement 45 individus ont été génotypés. Le nombre de marqueurs utilisés dans ce modèle n'est également pas très élevé : neuf seulement. Par ailleurs, le modèle est très simplifié, puisqu'on suppose que les fréquences génotypiques au locus F sont à l'équilibre et que les fréquences alléliques sont les mêmes chez les hermaphrodites et chez les femelles. On n'a pas modélisé d'évènement migratoire après l'introduction du migrant 2, et pourtant en 2008, en 2009 et en 2010, des migrants ont été introduits ; ils n'ont cependant pas provoqué de structure visible dans l'échantillon de 2010. De plus, l'allofécondation résiduelle chez les hermaphrodites a été négligée. Enfin, le modèle est déterministe, et donc valable sur de courtes périodes stables et pour une population de grande taille. Un modèle stochastique permettrait plus de souplesse, et de calculer, par exemple, une précision d'estimation pour la probabilité de provenir d'un migrant sachant un génotype.

Néanmoins, le modèle nous a donné des résultats cohérents puisque les génotypes les plus proches du migrant 2 (ceux formant le groupe 2) ont tous été considérés comme étant ses descendants.

Ce type de modèle est intéressant car il peut avoir plusieurs applications. Outre le fait qu'il puisse permettre de repérer des descendants de migrant, il peut servir à faire des scénarios d'évolutions par simulation, en faisant varier les paramètres populationnels (taux d'allofécondation, fréquences initiales, etc.)

III.1.4. Régime de reproduction

Dans cette partie, le régime de reproduction sera étudié pour plusieurs raisons. D'abord, nous chercherons à élucider son rôle dans la structuration de la population. En effet, la migration est à l'origine de la structuration principale de la population, mais il est probable que le régime de reproduction joue un rôle dans le maintien ou non de cette structuration. Nous posons l'hypothèse que les groupes structurés (groupes 1 et 2) sont des descendants de migrants (des individus « M ») par autofécondation, tandis que la population cœur correspond aux descendants de la population d'origine (individus « NM »), par autofécondation et par allofécondation, et aux descendants des migrants (« M ») ayant intégré la population d'origine par allofécondation.

Ensuite, nous chercherons à estimer le taux d'allofécondation réel de la population, pour déterminer s'il est bien égal à 20%.

III.1.4.1. Calcul de la probabilité d'appartenir à une classe de consanguinité

Pour déterminer l'impact du régime de reproduction, la classe de consanguinité des 192 individus de l'échantillon 2010 a été estimée. Un individu appartient à la classe de consanguinité S_n s'il est issu de n générations successives d'autofécondation après la dernière allofécondation de sa généalogie.

La méthode utilisée est celle développée par Enjalbert and David, 2000. Elle se base sur l'idée qu'à chaque génération d'autofécondation, la probabilité d'être hétérozygote à un locus est divisée par deux. La probabilité d'être hétérozygote sachant la classe de consanguinité S_n est donc la suivante :

$$P(ht/S_n) = \frac{He}{2^n} \quad \textbf{(1)}$$

L'extension de cette relation à plusieurs locus, permet d'écrire la probabilité conditionnelle d'un génotype G_i appartenant à individu i :

$$P(G_i/S_n) = \left(\frac{He_l}{2^n}\right)^a \left(\frac{1 - He_l}{2^n}\right)^{1-a}$$

Où a est une variable indicatrice qui prend la valeur 1 lorsque l'individu est hétérozygote et 0 lorsque l'individu est homozygote ; He_l est l'hétérozygotie attendu au locus l.

Pour L locus indépendants, la vraisemblance du génotype G_i d'un individu i est le produit des probabilités conditionnelles des génotypes à chacun des locus :

$$P(G_i/S_n) = \prod_{l=1}^{L} \left(\frac{He_l}{2^n}\right)^a \left(\frac{1 - He_l}{2^n}\right)^{1-a}$$

Chaque individu de l'échantillon de 2010 a été assigné à la classe de consanguinité qui maximisait la vraisemblance pour le génotype observé sur l'ensemble des 23 marqueurs.

III.1.4.2. Risque d'erreur dans l'assignation

Nous avons voulu déterminer quel était le risque d'assigner un génotype à une mauvaise classe de consanguinité.

Des simulations ont été effectuées avec le logiciel R afin de connaitre le pourcentage d'erreur commis lors des assignations. Nous avons simulé, 100 fois, à partir de notre jeu de données, 6 classes de consanguinité : S0, S1, S2, S3, S4 et S5, contenant chacune 1 000 individus. Cela consistait au tirage du statut homozygotes/hétérozygotes pour chacun des locus, en fonction de la probabilité d'être hétérozygote ou homozygotes au locus sachant la classe de consanguinité S_n. Ces individus ont ensuite été réassignés dans des classes de consanguinité S_k en réutilisant la méthode d'Enjalbert et David. La proportion d'individus d'une classe S_n (classe vraie) assignés à la classe S_k correspondante nous permet d'estimer $P(S_k/S_n)$, la probabilité d'être assigné à S_k sachant que l'on est S_n. Les probabilités $P(S_k/S_n)$ sont présentées dans le **tableau 4**.

L'assignation des individus aux différentes classes de consanguinité est très bonne lorsque les individus sont S0. En revanche, pour les autres classes consanguinité, l'erreur est importante. Pour éviter une trop forte erreur, les classes S3, S4 et S5 ont été combinées dans l'étude.

Tableau 4 : Probabilité pour un individu d'une classe S_n d'être assigné dans chaque classe

	Probabilité d'être assigné à chaque classe S_n					
Classe de consanguinité réelle de l'individu	S0	S1	S2	S3	S4	S5
S0	**0.955**	0.045	0	0	0	0
S1	0.047	**0.789**	0.161	0.003	0	0
S2	0	0.167	**0.579**	0.173	0.068	0.013
S3	0	0.012	0.289	**0.289**	0.285	0.125
S4	0	0	0.078	0.179	**0.385**	0.358
S5	0	0	0.016	0.059	0.312	**0.6130**

III.1.4.3. Classe de consanguinité des individus de la population

Les 192 individus de la population de 2010 ont été assignés à la classe de consanguinité qui maximisait leurs vraisemblances. Le **tableau 5** répertorie le nombre d'individus assignés à chaque classe :

Tableau 5 : Pourcentage d'assignation des individus à différentes classes de consanguinité

Classe de consanguinité	S0	S1	S2	S3+
Effectif	6	12	25	149
% de la population	3,12	6,25	13,0	78,4

III.1.4.3.1. Classes de consanguinité et structure de la population

La répartition des classes de consanguinité indique que dans le cœur de la population, on retrouve toutes les classes de consanguinité. A l'inverse, on ne trouve pas d'individus S0 ou S1 dans les groupes 1 et 2 définis par le dendrogramme. Ces groupes correspondent donc probablement à des descendants de migrants ne s'étant reproduit que par autofécondation, se développant donc indépendament en tant que lignée. Les génotypes arrivés par migration contenaient visiblement une légère hétérozygotie résiduelle, générant une faible variation intra-groupe. Autrement dit, cette variation intra-groupe est probablement due à la ségrégation des locus initialement hétérozygotes dans les lignées migrantes. L'ensemble des descendants d'un migrant par autofécondation crée donc un groupe de lignées très apparentées, n'ayant pas échangé de gènes pour l'instant avec le reste de la population

Ces résultats confirment donc notre hypothèse. La migration crée des groupes ; l'allofécondation permet l'intégration des allèles migrants dans le cœur de la population, tandis que l'autofécondation maintient les groupes de diversité.

On constate ainsi que les 23 descendants du migrants 2 qui appartiennent au groupe 2, ont tous été assignés à des classes de consanguinité au moins S2. A l'inverse, les 6 autres sont S0 ou S1, à l'exception d'un individu qui est probablement S2. La classe d'autofécondation des descendants du migrant 2 est figurée sur une AFC (Analyse Factorielle des Correspondances) représentant les 192 individus de l'échantillon de 2010 (**Figure 14**).

III.4.3.2. Estimation du taux d'allofécondation

Comme les épis prélevés dans la population étaient tous issus de plantes hermaphrodites, les individus S0 sont, a priori, issus d'allofécondation résiduelle. Il y a cependant un risque d'erreur d'identification des hermaphrodites. Le pourcentage d'individus appartenant à la classe S0 (c'est-à-dire issus d'allofécondation) correspond au taux d'allofécondation entre hermaphrodites, qu'on nommera t. On

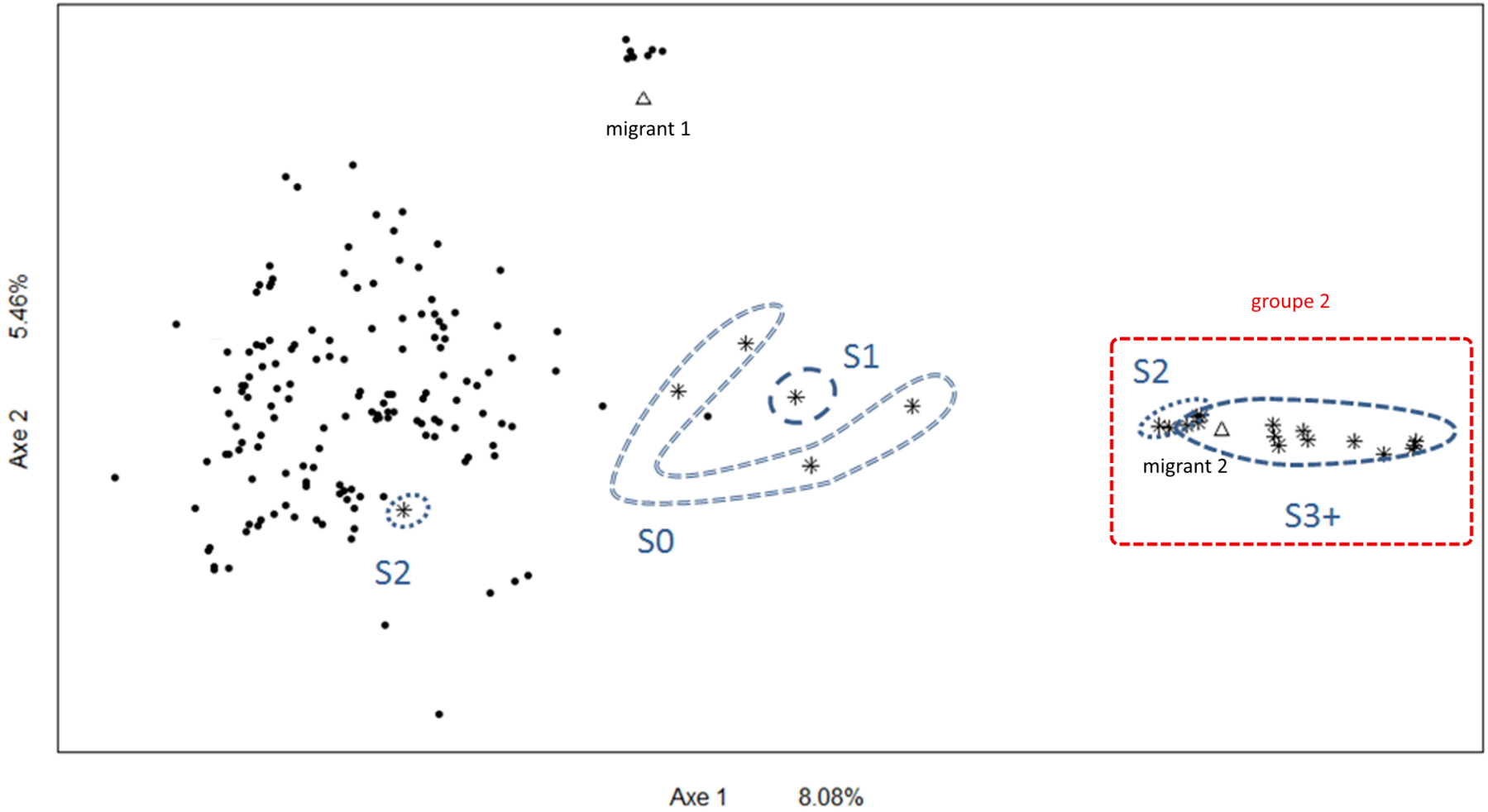


Figure 14 : Analyse Factorielle des Correspondances de l'échantillon de 2010, sur 23 marqueurs.

Les migrants 1 et 2 sont représentés par des triangles, et les descendants du migrant 2, par des étoiles. Les classes S0, S1, S2 et S3+ indique la classe de consanguinité à laquelle appartiennent les descendants du migrant 2. Le groupe en rouge correspond au groupe 2 défini par le dendrogramme.

s'attend à ce que cette valeur de t, qui comprend le taux d'erreur d'identification d'un hermaphrodite et le taux d'allofécondation, soit de faible valeur, chez le blé dur, les rares estimations disponibles sur marqueurs morphologiques ne dépassant pas 5% (Tsegaye, 1996). On trouve ainsi que le taux d'allofécondation des hermaphrodites est égal à 3,1% dans la population de 2010. Cela est en cohérence avec la littérature qui évoque des taux allant de 2% à 4% (Tsegaye, 1996).

La classe de consanguinité S1 correspond aux descendants par allofécondation à la génération précédente (puis qui ont subi une autofécondation). Leur probabilité est égal à :

$$P(S_1) = (1-t) \times [(1-t)\alpha + t]$$

$1-t$ correspond à la probabilité de comprendre un événement d'autofécondation quand on descend d'un hermaphrodite (cas à la dernière génération) ; $(1-t)\alpha + t$ est la probabilité de comprendre un évènement d'allofécondation dans l'ensemble de la population (cas à l'avant dernière génération). Connaissant $P(S_1)$ et connaissant t, nous avons pu en déduire α, qui est égal à :

$$\alpha = \frac{\left(\frac{P(S_1)}{(1-t)} - t\right)}{1-t} = 3{,}4\%$$

Ce taux d'allofécondation est très faible, bien plus faible que le niveau attendu (20%). Il se peut que l'échantillon ne soit pas représentatif de la population, ou que le fonctionnement mis en place, utilisant le gène de stérilité mâle, soit source d'erreur : peut-être que les mâles stériles ne sont pas correctement identifiés. Certains hermaphrodites sont peut-être confondus avec des mâles stériles, ce qui induit une réduction du taux d'allofécondation. De fait, des études morphologiques fines réalisée en 2014 par Muriel Tavaud montre que la stérilité mâle détectée sur le bâillement des glumes n'est pas très efficace (communication personnelle), et de nombreuses graines ont pu être observées sur des épis, considérés visuellement comme mâles stériles et ensachés avant floraison

III.1.5. Taille efficace

La dérive peut générer déséquilibre de liaison entre locus non liés physiquement, et donc provoquer des erreurs dans une étude d'association pangénomique. Nous avons cherché à estimer la taille efficace de la population, pour déterminer son effet sur le développement du déséquilibre de liaison .

La taille efficace peut-être estimée à partir des changements de fréquences alléliques au cours du temps (Goldringer and Bataillon, 2004; Waples, 1989). Cette méthode se base sur l'idée que, sous l'action de la dérive génétique uniquement, la variance des fréquences alléliques, entre une population à la génération 0 et la même population à la génération g, dépend de la taille efficace Ne :

$$V(p_g) = p_0(1-p_0)\left[1 - \left(1 - \frac{1}{2Ne}\right)^g\right] \text{ (Crow and Kimura, 1970)}$$

où p_0 est la fréquence allèliques dans la population à la génération 0 et p_g à la génération g.

Si le nombre g de générations n'est pas trop grand ($g \ll Ne$), alors Ne peut être approximé par :

$$Ne \approx \frac{p_0(1-p_0)g}{2V(p_g)}$$

(Nei and Tajima, 1981) ont proposé un estimateur de la variance standardisée entre deux générations, g_x et g_y, pour chaque locus l portant I_l allèles :

$$F_{c,l} = \left(\frac{1}{I_l}\right)\sum_{i=1}^{I_l}\frac{(p_{x(i,l)} - p_{y(i,l)})^2}{(p_{x(i,l)} - p_{y(i,l)})/2 - p_{x(i,l)}p_{y(i,l)}}$$

Où $p_{x(i,l)}$ représente la fréquence de l'allèle i au locus l dans un échantillon de taille n_x issu de la population à la génération g_x. Idem pour $p_{y(i,l)}$, pour un échantillon de taille n_y de la population à la génération g_y.

Le $\mathrm{F_{c,l}}$ de chacun des neuf marqueurs microsatellites génotypés à la fois sur la l'échantillon de 2005 et de 2010 a été calculé. La distribution des $\mathrm{F}_{c,l}$ a été visualisée afin de vérifier si certains avaient des valeurs extrêmes (ce qui peut être le signe qu'il est sélectionné) : aucun n'était dans ce cas. Un F_c multilocus a ensuite été estimé comme ci-dessous :

$$F_c = \frac{\sum_l I_l F_{c,l}}{\sum_l I_l}$$

Enfin, cet estimateur permet de calculer la taille efficace :

$$Ne = \frac{g_y - g_x}{2(\mathrm{F_c} - 1/(2n_x) - 1/(2n_y)}$$

La taille efficace de dérive a été estimée à partir du F_c multilocus, calculé sur les données de 2005 et les données du cœur de population de 2010 (160 individus). Elle a été estimée à 57,23, en se basant sur les locus non sélectionnés. Par rapport à l'effectif démographique (4 000 plantes), elle est donc très faible, ce qui indique une forte dérive génétique.

III.1.6. Conclusion et définition d'un échantillon sans structure pour l'étude sur les effets régulateurs de l'expression

Le taux d'allofécondation estimé sur la population est plus faible que prévu par notre système d'identification visuelle des mâles stérile, et la population a été soumise à une forte dérive. Ces deux éléments peuvent induire une structuration. Pour éviter d'avoir un échantillon trop structuré, comportant des individus très apparentés et qui soit source d'erreurs lors de notre travail sur les données d'expression, nous avons souhaité sélectionner les 160 lignées appartenant au cœur de

population, visiblement moins structuré. Malheureusement, une vingtaine de ces lignées ont été perdues.

Comme l'échantillon était relativement petit, 45 nouvelles lignées ont été génotypées sur 5 marqueurs microsatellites, parmi les 23 utilisés pour génotyper l'échantillon de 2010 (Xgpw3142, Xgwn148, Xgwn213, Xgwm234, Xgwm601 et Xwmc710). En calculant les fréquences génotypiques dans les trois groupes de l'échantillon de 2010, définis par l'AFC, il a été possible d'estimer la probabilité d'un génotype d'appartenir à un groupe. Pour chaque génotype G_i, on connaissait donc $P(G_i/\mathrm{groupe_j})$, la probabilité pour un génotype G_i d'appartenir au groupe j ($j = \{1,2,3\}$; le groupe 3 étant le coeur de population).

Par ailleurs, les probabilités des groupes 1, 2 et 3 ($P(\mathrm{groupe_1})$, $P(\mathrm{groupe_2})$ et $P(\mathrm{groupe_3})$) étaient connues, puisqu'elles correspondaient à la proportion d'individus assignés à ces trois groupes. Il a donc été possible de calculer $P(G_i/\mathrm{groupe_1})$, $P(G_i/\mathrm{groupe_1})$ et $P(G_i/\mathrm{groupe_3})$. Les nouvelles lignées ont été assignées au groupe qui maximisait la vraisemblance de leur génotype. Trente-cinq lignées ont été assignées au groupe 3, soit au cœur de population. Au final, nous avons obtenu un échantillon de 177 lignées appartenant au cœur de population.

Des analyses ultérieures, réalisées par Yan Holtz utilisant 721 SNPs (génotypage réalisé ultérieurement, cf. III.2.2.4) et portant sur cet échantillon ont démontré que globalement, il n'y avait pas de déséquilibre de liaison longue distance sur le chromosome 3B (David et al., in press). En moyenne, le déséquilibre de liaison diminuait rapidement après 10-20 Mb, et atteignait un plateau après 40 Mb. Quelques contigs étaient néanmoins impliqués dans du déséquilibre de liaison à moyenne et longue distance. Bien qu'il soit possible que cela soit dû à de la structuration, il est plus probable que ce soit un artefact causé par une assignation incorrecte de ces contigs sur la région télomérique du bras long du chromosome 3B.

Cela nous confirme que cet échantillon de 177 lignées issues du cœur de la population constitue un échantillon approprié pour une étude d'association pangénomique, puisqu'il est polymorphe, peu structuré, et avec un déséquilibre de liaison diminuant rapidement avec la distance physique, du fait du brassage par allofécondation de la population.

La suite est présentée en anglais comme un projet de publication et portera sur l'échantillon détaillé ci-dessus. L'extraction de l'ARN et la préparation des banques ADNc a été réalisé par le laboratoire de Sylvain Santoni. L'analyse bioinformatique (mapping, détection de SNP, validation des SNPs et normalisation des données d'expression) a été effectuée par Yan Holtz. Pour ma part, j'ai réalisé les analyses de données (détection d'un biais d'expression et surtout détection d'eQTL).

III.2 Asymmetry of A and B genomes of tetraploid wheat *Triticum turgidum* in *cis* and *trans* regulations of the gene expression

III.2.1 Introduction

Polyploidy, the phenomenon consisting in the whole genome duplication, is a major process of plant evolution. Indeed, most flowering plants have experienced at least an event of polyploidization once in their evolutionary history (Cui et al., 2006; Jiao et al., 2011; Van de Peer et al., 2009; Soltis et al., 2008). Polyploidization could have several evolutionary advantages by allowing the development of characters such as drought tolerance or resistance to pests (Chen, 2007). Several mechanisms could be involved in this fitness improvement: the fixation of heterozygosity (i.e., equivalence between two different and fixed homoelogous genes and two different homologous alleles), even in selfing species, the reduction of the impacts of a mutation by genome redundancy and the creation of new gene function via the neo and subfunctionalization (Chen, 2007; Comai, 2005; Udall and Wendel, 2006). It was demonstrated that, in autogamous allopolyploid plants, dominance could be converted to additive x additive epistasis, and that heterosis could be fixed (Allard et al., 1993; García et al., 1991). This was particularly observed in the herbaceous species *Avena barbata*, which is more vigorous than its diploid ancestors.

It has been thus suggested that polyploid plants have had a greater chance of survival during the Cretaceous–Tertiary extinction event (Fawcett et al., 2009). Moreover, the success of polyploid plants in domestication could also be a sign of the great adaptability of these plants (Dubcovsky and Dvorak, 2007; Feldman and Levy, 2009). For example, in the areas where tetraploid emmer (*Triticum turgidum* ssp. *diccocum*), the ancestor of durum wheat (*Triticum turgidum* ssp. *durum*), and diploid einkorn (*Triticum monococcum*), were both cultivated, emmer has always become the primary domesticated cereal (Dubcovsky and Dvorak, 2007). Similarly, bread wheat, *Triticum aestivum* (hexaploid), represents 95% of the wheat crop, while durum wheat, *Triticum turgidum* ssp. *durum* (tetraploid), represents only 5%. This is mainly due to the greater adaptability of the bread wheat, which can be grown in very diverse environments.

In new allopolyploid species, the assembling of two divergent genomes causes a major stress to the new plant. Barbara McClintock coined the term "genome shock" to describe the unexpected changes due to hybridization (McClintock, 1984). The plant must undergo several transformations in order to face the new challenges of allopolyploidization (Feldman and Levy, 2009). First, homoelogous chromosomes must be differentiated from each other to prevent the homoelogous pairing during meiosis. In wheat, this differentiation process may be achieved through elimination of DNA sequences (Feldman and Levy, 2009; Feldman et al., 1997); some sequences would therefore become homoelogous-specific and may facilitate homology recognition. Homologous pairing during meiosis is

also sustained by the *Ph1* and *Ph2* gene systems (Sears, 1976) which evolved at the polyploid level in wheat. Second, the polyploid plants must deal with redundant or unbalanced gene expression between both genomes; therefore there is a need to harmonize gene expression, via gene silencing or gene elimination. The elimination of one of the homoelogous copies is called "fractionation" and enables to return to a diploid state. This process is not rare. For example, in *Arabidopsis thaliana* – a species that underwent a polyploidization event 20-40 Mya (Woodhouse et al., 2014) – 28.9% of homoelogous genes have been retained as pairs, while one of two copies was lost in the remaining 71.4% (Thomas et al., 2006). In maize (Woodhouse et al., 2010) and *Brassica rapa* (Tang et al., 2012), it was shown that the fractionation mechanism consists in a intrachromosomal recombination followed by a deletion. Research on synthetic allopolyploids showed that genetic and epigenetic alterations occur immediately after allopolyploid formation (Adams et al., 2003, 2004; He et al., 2003; Kashkush et al., 2002; Shaked et al., 2001; Wang et al., 2004). These alterations involve modifications such as sequence deletions (Feldman et al., 1997; Gaeta et al., 2007; Kashkush et al., 2002; Kotseruba et al., 2003; Ma and Gustafson, 2006; Ozkan et al., 2001; Shaked et al., 2001), transposon activations (Comai et al., 2000; Kashkush et al., 2002; Madlung et al., 2005), chromosomal rearrangements (Gaeta et al., 2007; Madlung et al., 2005; Pires et al., 2004; Song et al., 1995; Weiss-Schneeweiss et al., 2007), gene conversions (Kovarik et al., 2004, 2005; Wendel et al., 1995), histone modifications (Chen and Pikaard, 1997) and cytosine methylations or demethylations (Houchins et al., 1997; Kashkush et al., 2002; Lukens et al., 2006; Madlung et al., 2005; Shaked et al., 2001; Xu et al., 2009).

These genetic and epigenetic changes modify gene expression in allopolyploids (Chen, 2007; Chen and Ni, 2006; Osborn et al., 2003). For example, chromosomal rearrangements altered gene expression in *Brassica napus* polyploids (Gaeta et al., 2007; Pires et al., 2004). It was as well shown that cytosine methylations silence genes in species such as wheat (Kashkush et al., 2002; Shaked et al., 2001), *B. napus* (Lukens et al., 2006; Xu et al., 2009) or *Arabidopsis* allopolyploids (Madlung et al., 2002, 2005). Moreover, genetic and epigenetic mechanisms can be interrelated. For example Madlung et al. 2005 noticed that transposable elements belonging to the *Sunfish* families were demethylated and reactivated in an *Arabidopsis* allotetraploid, which suggests that epigenetic changes activated transposons. Then, reactivating transposons may lead to chromosomal rearrangements. The importance of these genetic and epigenetic processes may change depending on the species (Chen, 2007). For example, sequence deletion is commonly related in wheat (Feldman et al., 1997; Shaked et al., 2001) and numerous chromosomal rearrangements were observed in *Brassica* species (Gaeta et al., 2007).

The two genomes involved in allopolyploidization event do not behave in the same way. One aspect of this non-equivalence is unbalanced expression level bias (Grover et al., 2012). In a pair of homoelogs,

one gene can contribute more to the total expression than the other, i.e. expression ratios between homoelogous genes can differ significantly from the 50/50 assumption.

If there are more pairs of genes which display expression bias toward one parent than toward the other one, the term "unbalanced expression level bias" can be used. For example, this was related in allotetraploid plants created by the hybridization between *Arabidopsis thaliana* and *Arabidopsis arenosa*: there were more pairs of homoelogous genes where the expression ratio was in favour of *A. arenosa*, than pairs displaying expression ratio in favour of *A. thaliana* (Wang et al., 2006). Unbalanced expression level bias was as well observed in *Brassica* (Chen and Pikaard, 1997), in maize (Schnable et al., 2011), in hexaploid wheat (Akhunova et al., 2010) and in the recent natural allotetraploid *Tragopogon miscellus* (Buggs et al., 2010). Unbalanced expression level bias should not be confused with unbalanced level expression dominance (Grover et al., 2012), which is another aspect of the non-equivalence between genomes. Expression level dominance is the phenomenon where the total expression of homoelogs for a given gene in an allopolyploid is statistically equivalent to the expression level of that gene in only one of the parents. This concept does not consider homoeologous genes separately, but rather refers to the total expression level of a pair of duplicated genes, when measured in the allopolyploid and when compared to its parents. Unbalanced expression dominance was related in polyploid species such as cotton (Flagel and Wendel, 2010) and bread wheat (Chagué et al., 2010; Qi et al., 2012). A final aspect of the non-equivalence between genomes is the fractionation bias, i.e., the fact that one genome retains more genes than the other during the fractionation process. This phenomena was observed in *Arabidopsis* polyploids (Wang et al., 2006), in *Brassicaceae* species (Tang et al., 2012; Thomas et al., 2006) and also in maize (Schnable et al., 2011, 2009; Woodhouse et al., 2010). Moreover, in polyploids such as *Arabidopsis suecica* (Chang et al., 2010), maize (Schnable et al., 2011) and *B. rapa* (Cheng et al., 2012), the genome which has lost the most genes is the one whose genes are expressed at relatively lower level compared to their homoelog. Indeed, genes which are less expressed than their homoelog are more likely to be removed, because there is less risk that their loss causes a decrease in fitness (Freeling et al., 2012). Woodhouse et al. (2014) calls genome dominance this inverse correlation between expression and gene loss. Most of the ancient polyploids exhibit genome dominance, while few do not (Garsmeur et al., 2014). The former may be derived from allopolyploidization process and the latter from autopolyploidization.

Differences in expression between homoelogs could be due to differences in gene regulations among parents. These regulations can be of several types. *Cis* regulations are performed by a DNA sequence located near or within the target gene. They could be due to changes in promoter or enhancer sequences, allele-specific epigenetic modifications or changes in the transcribed region which affects mRNA stability ARNm (Wittkopp et al., 2008). *Trans* regulations occur when a gene/sequence regulates

the expression of a target gene remotely located. (Shi et al., 2012) addressed the role of regulatory divergence between parents (i.e. differences in expression regulation) in a synthetic allotetraploid derived from a cross between *Arabidopsis thaliana* and *Arabidopsis arenosa.* They found that there was more *cis* than *trans* regulatory divergences. Furthermore, for genes subjected to both *cis* and *trans* divergence, *trans* effects could compensate expression differences due to *cis* effects, which maintained an expression divergence between the two parents. Finally, they noticed that *trans* regulators from *A. arenosa* had a stronger effect than those of *A. thaliana*. They supposed that the higher genetic diversity within *A. arenosa* (an outcrossing species) may confer plasticity to *trans* factors, which could then bind or interact with genes from a related species. In contrast, *A. thaliana* is an inbred, whose the low level of genetic diversity may have increased the specificity of interactions between *trans*-factors and DNA sequences. A similar work was undertaken by (Chaudhary et al., 2009) in an allopolyploid cotton (*Gossypium* L.). They found as well that that most variation in gene expression following genome merger is a consequence of *cis* regulatory divergence.

Variation in *cis* and *trans* regulations can be estimated by two methods: Allele-Specific Expression (ASE) analyses (Wittkopp et al., 2004) and detection of expression Quantitative Trait Loci (eQTL) with a linear model, similar to those used in association mapping (Cheung et al., 2005). The basic assumption of ASE analyses is: if two homologous alleles or two homoelogous genes have a different level of expression within the same individual, this is the reflection of a difference in *cis* regulation only. Indeed, it is assumed that in the same cellular environment (i.e. in a same individual: a hybrid, a F1 plant or, a newly synthetized allopolyploid, …) two alleles or two homoelogous genes are subjected to the same *trans* regulations. Conversely, a difference in expression level in different cellular environments (e.g., in the two parents of a hybrid) reflects all the regulatory divergences (*cis* and *trans* regulations). Logically, divergence in *trans* regulations can be estimated by subtracting the expression ratio between the two alleles in the same cellular environment, from the expression ratio between the two alleles in different cellular environments. Shi et al. 2012 and Chaudhary et al., 2009 used this method to assess regulatory divergence in the synthetic polyploid they studied. ASE method can allow as well the detections of eQTL (Sun and Hu, 2013). In this case, a polymorphism is considered as a *cis* eQTL if it alters gene expression in an allele-specific manner. Otherwise, it is considered as a *trans* eQTL. However, this method requires heterozygous genotypes, to distinguish *cis* eQTL from *trans* ones, and thus cannot be performed on homozygous lines.

Genome-wide association mapping approach (Hirschhorn and Daly, 2005), or other methods for QTL detection, can as well be used to detect eQTL. The goal is to identify markers correlated with the phenotype of interest, which is here the expression level. If a marker has a statistical effect on the expression level of a nearby gene, it will be considered as *cis* acting. If it is distant from the gene it

regulates, it will be considered as a *trans* eQTL. Genome-wide association studies (GWAS) are characterized by the use of a high density of SNPs (Single Nucleotide Polymorphisms) across the whole genome or the transcriptome. Their goal is to scan all the markers in order to find genetic variations associated with the phenotype of interest, and therefore constitute a powerful tool to reconnect one quantitative trait to its genetic determinants.

Several techniques can be used to estimate expression levels (Wang et al., 2009). Hybridization-based approaches consist in hybridizing fluorescently labelled cDNA sequences of the studies sample, with microarrays. These methods are high throughput and relatively inexpensive. The disadvantage of this method is that only the expression of known genes can be studied. In contrast, sequence-based approaches are based on the sequencing. The earliest are based on Sanger sequencing. The oldest ones involved sequencing of cDNA or EST libraries, but were low throughput, expensive and generally not quantitative. To remedy these problems, tag-based methods were then developed. These methods could provide accurate expression levels and were high throughput, but were expensive. In addition, a reference sequence was required. These disadvantages greatly limited the benefit of these techniques. In recent years, a new sequence-based technology has emerged: massively parallel RNA sequencing (RNA-seq). It used recently developed high-throughput sequencing technologies. Usually, RNA is converted to a cDNA libraries before sequencing. This technology has several advantages (Majewski and Pastinen, 2011; Wang et al., 2009) : it simultaneously maps transcribed regions and gene expression; it provides a very accurate assessment ; it is high throughput ; it requires a low amount of RNA ; it provides data of sequences which are not yet known and documented ; its cost is relatively low. For eQTL mapping, RNA sequencing (RNA-seq) appears to be a method of choice, because it provides both the expression and SNP data (Majewski and Pastinen, 2011).

Numerous studies of eQTL were performed on diploid plants, such as maize (*Zea mays*) (Holloway et al., 2011; Schadt et al., 2003; Shi et al., 2007; Stupar and Springer, 2006; Swanson-Wagner et al., 2009), *Arabidopsis thaliana* (DeCook et al., 2006; Gan et al., 2011; Keurentjes et al., 2007; West et al., 2007; Zhang et al., 2011), rice (*Oryza sativa*) (Wang et al., 2010), barley (*Hordeum vulgare*) (Chen et al., 2010; Moscou et al., 2011; Potokina et al., 2008) *B. rapa* (Hammond et al., 2011), and *Populus* (Drost et al., 2010). Most of these studies used biparental populations and Composite Interval Mapping (CIM) to detect eQTL. These studies have shown that regulations in *cis* are stronger than those in *trans* (Hammond et al., 2011; Holloway et al., 2011; Keurentjes et al., 2007; Schadt et al., 2003; Swanson-Wagner et al., 2009; Wang et al., 2006). However, *trans* effects are not negligible, as they can represent in certain studies about 80% of the detected eQTL (Hammond et al., 2011; Swanson-Wagner et al., 2009; Wang et al., 2010). Studies present important variations in the ratio between the number of detected *cis* eQTL and the number of detected *trans* eQTL, even within a same species. For example,

in maize, Swanson-Wagner et al. 2009 found that 78% of the detected eQTL acted in *trans* (using microarray hybridization and RNA from aboveground tissues, in hybrids), whereas Holloway et al., 2011 found a proportion of 10% of eQTL which act in *trans* (using microarray hybridization and RNA from roots, in a double haploid population). This may be due to a difference in plant materials, statistical methodology, or in the definition of *cis* and *trans* effects. In poplar, Drost et al 2010 found that eQTL detected in several organs act primarily in *cis*. In contrast, eQTL identified in a single organ largely regulate in *trans*. Trans regulatory eQTL are not distributed randomly along the genome, since many are concentrated in hotspots (Cubillos et al., 2012; Kliebenstein, 2009). It was observed that GWAS identified more *cis* eQTL than *trans* ones, compared to eQTL studies in structured populations, derived from crosses (Cubillos et al., 2012). In addition, *cis* eQTL found in GWAS population exhibit intermediate allele frequencies, whereas *trans* eQTL have more extreme allele frequencies. This could be the sign that *trans* eQTL are subjected to stronger selection (Cubillos et al., 2012) than *cis* eQTL.

So far, no eQTL study on plants aimed to detect epistatic effects. Epistatic effects between eQTL have nevertheless been detected on human (Becker et al., 2012; Hemani et al., 2014) and yeast (Brem and Kruglyak, 2005; Brem et al., 2005; Storey et al., 2005). The gene expression level is a complex trait. The epistasis is almost ubiquitous, and thus it is likely that gene expression in plants is also subject to epistasis. This could especially be true in autogamous allopolyploid species where epistatic effects could compensate for a lack of heterosis (Allard et al., 1993; García et al., 1991). Furthermore, it was demonstrated that *trans* effects within a network of genes could result in epistasis on the level of expression of one of these genes. In one study, where several types of networks (three genes) were simulated, the effect of epistasis on the expression level of a gene, was estimated (Gjuvsland et al., 2007). It was shown that gene network can generate epistasis, particularly by positive feedback loops. It is therefore expected to find epistasis effects on gene level expression.

In addition, no eQTL study has focused on polyploid plants. One can assume that homoelogous genes are more likely to have the same biological function and to belong to the same network of genes. Although gene networks are rewired following polyploidization, divergence between homoelogs is mainly achieved through spatiotemporal partitioning (subfunctionalization), and not neofunctionalization (De Smet and Van de Peer, 2012). Consequently, homoelogs may still tend to be part of the same gene network, even in ancient polyploid species. Therefore, presumably *trans* effects are more frequent between homoelogous genes. The same applies for epistasis, especially because *trans* effects can appear as epistasis, and because interactions between homoelogs can compensate for dominance (Gjuvsland et al., 2007).

It would be interesting to study the regulation of expression in a polyploid, especially in an ancient polyploid, which is polymorphic at both homoelogous loci. This is an ideal case to test whether the estimation models, presented in part II, are adapted to estimate *cis* effects, *trans* effects and interaction effects. It may be of special interest to work with polyploid species, whose genomes come from species with different properties, including different mating systems. It is the occasion to determine if these genomes may have different regulatory behavior, i.e. if they differ in terms of regulatory effects.

Tetraploid wheat (*Triticum turgidum*) is of special interest for this, since it is a spontaneous and relatively recent allopolyploid. It is an allopolyploid cereal, containing two genomes: the A and B genomes. It was one of the first plant to be domesticated, 10 000 years ago, in the region of the Karaca dag volcan, in Turkey (Charmet, 2011; Özkan et al., 2002; Salamini et al., 2002). Its wild form, *Triticum turgidum* ssp. *diccocoides*, arose 0.5 Million years ago, through hybridization between *T. urartu* (donor of the A genome) and a species related of the current species *Aegilops speltoides* (donor of the B genome) (Daud and Gustafson, 1996; Dvořák et al., 1988, 1993; Huang et al., 2002; Kilian et al., 2007; Petersen et al., 2006). The current descent of the A donor, *T. urartu* is a selfer ; it is difficult to determine the exact mating system of the B donor, but its relative *A. speltoides* is a outcrosser (Escobar et al., 2010). Therefore, we can assume that the two genomes of the tetraploid wheat come from two species with different mating system. Most hybrids *T. urartu* x *A. speltoides* are letal (Gill and Waines, 1978).

We worked with 175 highly homozygous lines issued from a population of pre-breeding, resulting from crosses between durum wheat (*Triticum turgidum* ssp. *durum*), domestic emmer (*Triticum turgidum* ssp. *diccocum*) and wild emmer (*Triticum turgidum* ssp. *diccocoides*). We were interested in determining how expression is regulated in *Triticum turgidum*. More specifically, we aimed to understand (i) if one of the two genomes of the tetraploid wheat contains more and/or stronger eQTL, (ii) if *trans* regulatory effects are more numerous and/or stronger between homoelogous genes than between non-homoelogous genes and (iii) if one could find evidence of epistastic effects between homoelogous loci.

III.2.2 Materials and Methods

III.2.2.1 Plant material

A population with a broad genetic base of tetraploid wheat (*Triticum turgidum*) was created in 1997 from composite crosses between durum wheat segregating for a male sterility gene plants (*Triticum turgidum* ssp. *durum*), wild emmer (*Triticum turgidum* ssp. *dicoccoides*) and domestic emmer (*Triticum turgidum* ssp. *dicoccum*) (David et *al*. submitted). Other subspecies of *Triticum turgidum* were also

incorporated in the crosses (*T. t. polonicum and T. t. turgidum*). The recessive male sterility gene was used to monitor the outcrossing rate, which was empirically set at 0.2. In 2009, 177 hermaphrodite plants, originating from an unstructured portion of the population (see paragraph III.1.4) were sampled randomly. They underwent three successive generations of selfing by single seed descent (SSD). Thus, the level of inbreeding is at least S3, but most of them have likely undergone several more selfing generations in the population before their sampling.

III.2.2.2 RNA Sequencing (RNA-Seq)

For each accession, seeds were harvested and germinated in petri dishes with 4 ml of purified water at constant temperature of 30°C in the dark in a growth chamber. Coleoptiles and primary leaves were sampled from 7 day-old seedlings and crushed in liquid nitrogen. Sequence data were following three steps: 1) mRNAs were extracted and purified. 2) cDNA libraries were built using Truseq adaptators. 3) cDNA libraries were sequenced using the Illumina mRNA-Seq, paired-end indexed protocol. Each cDNA library was sequenced on HiSeq2000 or HiSeq2500 by multiplex of 24 per lane.

Reads were preprocessed with cutadapt (Martin, 2011) to remove adaptor sequences and trim the end of reads with low quality scores (parameter -q 20) while keeping reads with a minimum length of 35 bp. We then filtered the reads on the basis of their mean quality score, keeping those with a mean quality higher than 30, and subsequently, we removed orphan reads (i.e. those for which the mate was discarded in the previous filtering steps) using a homemade script.

An approximate total of 3 billion read pairs was produced with a number of reads per accession ranging from 4.5 to 51 million. Reads had an average length of 100 bp. The cleaning steps resulted in 5.5 million read pairs per sample on average (min=3.4 10^6, max=10.1 10^6 and sd=1.1 10^6), with an average of 392 913 orphan reads (min=179 601, max=846 175 and sd=124 223). This resulted in an average of 11.5 million usable reads per accession (min=7.2 10^6, max=20.5 10^6, sd=2.2 10^6). See (David et al., in press).

III.2.2.3 Reads mapping

We used a reference sequence derived from the bread wheat chromosome survey sequence for the *Triticum aestivum* cv. Chinese Spring generated by the International Wheat Genome Sequencing Consortium (IWGSC) (Mayer et al., 2014a) to map the reads. The Biomart facilities of EnsEMBL (release 22, (Hubbard et al., 2002)) were used to collect the transcripts (CDS+UTRs) and the physical genomic location of the 66 307 indexed genes, belonging to either the A (32 081 genes) or B (34 226 genes) genome. As these sequences were obtained by separately sequencing each bread wheat chromosome arm, this reference enabled us to distinguish homoelogous durum wheat copies and to assign a physical position to our predicted tetraploid wheat SNPs. Pairs of genes were considered as

homoelogous pairs, when designated as such in the Ensembl database. Note finally that, since this is based on genomic data, no alternative transcripts have been documented so far in this database.

Reads were mapped on this reference using BWA (Li and Durbin 2009) while allowing 3 errors (-n 3 in the aln step). No read may be retained or detected in very weakly expressed genes, and thus there may be missing data for both the expression and the genotype of these genes. RNAseq does not allow to discriminate between missing data and non-expressed genes, as a situation in which no read is mapped can be considered as non-expression case. On the 66 307 genes of the reference, 62 166 (30 130 from the genome A and 32 036 from the genome B) were matched by our data, i.e. considered as expressed in at least one individual. Among them, 13 065 homoelogous pairs were identified.

Reads per kilobase per million mapped reads (RPKM) (Mortazavi et al., 2008) were estimated for each contig and each individual, using a custom Perl script to parse the mapping files (David et al., in press). RPKM was used as a gene expression proxy. RPKM is normalized relatively to the total number of reads per individual, and thus eliminates the bias that could be caused by a difference in sequencing coverage. Reads belonging to identical areas of the two homoelogous genes can be mapped to any of the *T. aestivum* homoelogous copies. Therefore, to allow a better discrimination between homoelogous genes, only reads per divergent sites (RPDS) were considered. As for RPKM measure, RPDS are weighted by the total number of reads per individual. A study which justifies the use of RPDS instead of RPKM for homeologous expression comparison can be found in **Annexe VII.5.**

III.2.2.4 Genotyping

Genotype calling was carried out using *read2SNP* (Gayral et al., 2013). Read2SNP considers the number of reads displaying each nucleotide at a given position and returns the most probable genotype given this nucleotide count and a *Fis* parameter characterizing the population mating system (Wright, 1950). The *Fis* was here considered as a quality parameter and SNPs showing a heterozygosity excess were eliminated, since they were likely coming from spurious mapping due to multigenic families or to close homoelogs. Genotypes inferred with a coverage of less than 10 reads, a *read2SNP* probability for the calling below 99% or a SNP having an observed *Fis* value below 0.8, were considered as missing data. Genotyping failed for two individuals; therefore genotype data were available for only 175 individuals.

On the 62 166 detected genes, only 18 303 were polymorphic. A total of 84 710 SNPs was identified, which corresponds to about 4.6 SNPs per polymorphic gene in average. Of these 84 710 SNPs, a second sorting was performed: only SNPs with two alleles were retained. In addition, all heterozygous genotypes (representing 11.8% of the data) were considered as missing data. Real heterozygosity was expected around 2%, which means that even after a sort of SNPs based on the *Fis* criterion, some spurious SNPs were retained by read2snp.

III.2.2.5 Data Analysis

III.2.2.5.1 Expression bias

We aimed to determine if there was an expression bias between the two genomes of the tetraploid durum wheat, i.e. if one of the two genomes was more expressed than the other.

First, we examined this bias in terms of proportion of expressed genes. We wanted to verify if a genome expressed significantly more genes than the other, or not. The proportion of genes of the A and B genomes, expressed in our sample (transcriptomic data) was compared to the gene number identified in the genomes of *Triticum aestivum* (Mayer et al., 2014a) with a Pearson's chi-squared test. Indeed, assuming the genes of A and B genomes are in the same proportion in both *T. turgidum* and *T. aestivum* genomes, then the *T. aestivum* genomic data are representative of the genome of *T. turgidum*.

Then, this bias was examined in terms of expression level, using three methods:

- Firstly, the expression level was compared between the A genome and the B genome. The means of the expression level (estimated by the RPKM) of all the 62 166 genes were calculated and they were submitted to a permutation test (10 000 permutations), to compare the mean of means of expression level between the two genomes.
- Secondly, the average level expression was again compared between the two genomes, using another method. A Wilcoxon test was used to compare the average expression levels of the 13 605 genes belonging to the A genome, with the averages ones of their 13 605 homeologs from the B genome. In other words, a sample of 13 605 values (the 13 605 average levels of expression of the genes from the genome A) was compared with a paired sample of 13 605 values (the 13 605 average levels of expression of the genes from the genome B).
- Finally, each pair of homoelogous genes was individually studied. For each one, a Wilcoxon test was applied on the expression data of the 177 lines, to asses if one copy was significantly more expressed than the other. If it was the case, the considered pair of homoelogs was regarded as subjected to an expression bias. P-values of the Wilcoxon test were corrected using the Holm correction. Then, at a global level, we examined if there were more homoelogous pairs of genes which display expression bias toward one genome than toward the other one. If it was true, then it could be stated that there was an unbalanced expression bias, as defined by (Grover et al., 2012).

III.2.2.5.2 Study of eQTL

In a second part, we aimed to detect eQTL, and to identify their characteristics, i.e. if they are mainly *cis* or *trans* acting, or if they are more often localized on one genome. Our aim was to assess for a pair

of SNPs (one belonging to the A genome, the other to the B genome), their additive *cis* and *trans* effects and the effect of epistasis on the expression of the two genes involved in the pair, separately. These SNPs were called SNP A and SNP B; the genes to which they belong, gene A and gene B.

III.2.2.5.2.1 Detection of eQTL

III.2.2.5.2.1.1 Formation of pairs of SNPs

First, we had to form pairs of SNPs. Two kinds of pairs were considered: pairs of SNPs belonging to homoelogous genes and random pairs of SNPs belonging to non-homologous genes. Expected heterozygosity (Nei heterozygosity He (Nei, 1978)) was calculated for each selected SNP.

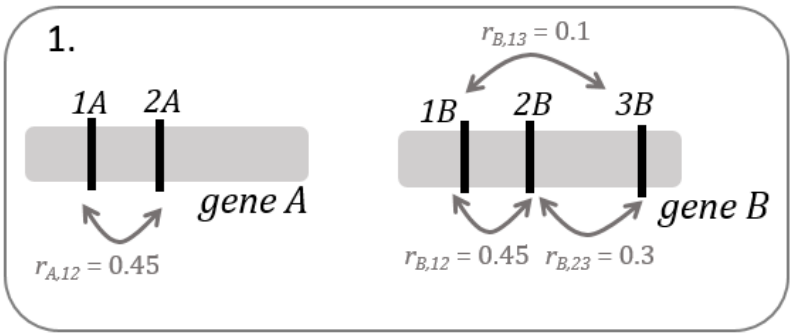
III.2.2.5.2.1.1.1 Formation of homoelogous pairs

For each one of the 13 065 pairs of homoelogous genes, all possible pairs of SNPs were examined. **Figure 1** details how these pairs of SNPs were sorted. As SNPs are biallelic and homozygous, four possible genotypes may exist for one pair of SNPs. Our sorting criterion was the number M of individuals carrying the rarest genotype. The pairs where $M < 5$ were eliminated. For each pair of homoelogous genes, pairs of SNPs were sorted from the best represented (largest M) to the least represented (smallest M). The pair of SNPs for which M was maximal was selected. Then, among the remaining pairs of SNPs, some could be selected if they were in a linkage disequilibrium inferior to 0.4 with the already selected pairs. Linkage disequilibrium between pairs of SNPs was defined as follows:

$$r_{AB} = \min(r_A, r_B)$$

With r_A and r_B being the linkage disequilibria, expressed as correlation coefficients, between the two SNPs from the gene A and the two SNPs from the gene B, respectively. As the linkage disequilibrium between the pairs is equal to the smallest linkage disequilibrium between SNPs, two SNPs in linkage disequilibrium may have been selected to form two different pairs. But as the other two SNPs with which they form a couple are not in linkage disequilibrium, the two pairs of SNPs are globally different. There are cases where one SNP is used to form two or several pairs, each being formed with SNPs that are not in linkage disequilibrium. This procedure allowed us to properly cover genes and form a larger number of pairs of SNPs.

Finally, 617 pairs of SNP were obtained, distributed over 548 pairs of homoelogous genes. On the 1 189 markers used to constitute the 617 pairs of SNP, 43 SNPs (3.6%) were used more than once to constitute several pairs of SNPs.



2.

Pair number	SNP A	SNP B	M
1	1A	1B	10
2	1A	2B	20
3	1A	3B	15
4	2A	1B	3
5	2A	2B	7
6	2A	3B	11

3.

LD	1	2	3	5	6
1	-	min(1,0.45)	min(1,0.1)	min(0.45,0.45)	min(0.45,0.1)
2	-	-	min(1,0.3) = 0.3	min(0.45,0.5)	min(0.45,0.3) = 0.3
3	-	-	-	min(0.45,0.3)	min(0.45,0.5) = 0.45
5	-			-	min(1,0.3)
6	-	-	-	-	-

Figure 1: Sorting of the pairs of SNPs belonging to homoelogous genes

1. The two genes and their SNPs. In this example, the gene A contains two SNPs, 1A and 2A. The linkage disequilibrium between them is equal to $r_{A,12} = 0.45$. The gene B contains three SNPS, called 1B, 2B and 3B. The linkage disequilibrium is equal to :

- $r_{B,12} = 0.45$ between 1B and 2B
- $r_{B,13} = 0.1$ between 1B and 3B
- $r_{B,23} = 0.3$ between 2B and 3B

2. The six pairs of SNPs which can be formed with these SNPs. For each pair, the criterion M (the number of individuals carrying the rarest genotype) is shown in the third column. The pair number 2 is the best represented, with $M = 20$, and will be selected. The pair number 4 is discarded as its M is inferior to 5.

3. In this table, the linkage disequilibrium (LD) between pairs of SNPs is calculated. This enables to eventually select other pairs, in addition to the one with the largest M (pair n°2). Linkage disequilibrium between pairs of SNPs is the minimal linkage disequilibrium between the SNPs constituting the two pairs. For example, the linkage disequilibrium between the pairs 2 and 3 is equal to $r_{AB,23} = \min(r_{A,11}, r_{B,23}) = \min(1,0.3) = 0.3$. Indeed, the pairs 2 and 3 are both formed with the SNP 1A. Thus the linkage on the gene A is equal to the linkage disequilibrium of 1A with itself, i.e. the maximal possible value: $r_{A,11} = 1$. Then, the pair 2 contains 2B and the pair 3, 3B. The linkage disequilibrium on the gene B $r_{B,23} = 0.3$.

   The table shows that two pairs of SNPs have a linkage disequilibrium inferior to 0.4 ($r_{AB} < 0.4$) with the already selected pair (pair n°2): pairs n°3 and n°6. As the pair n°3 has a M criterion superior to that of the pair n°6 (15 against 11), it is preferentially selected. Then, as the linkage disequilibrium between pairs the pairs n°3 and n°6 is superior to 0.4 ($r_{AB,36} = 0.45$), the pair n°6 is not selected.

   Thus, finally, two pairs of SNPs are selected fort this pair of homoelogous genes: the pairs n°2 and n°3. They both contain the same SNP on the gene A: 1A. But as the SNPs selected on the B genomes (2B and 3B) are not in strong linkage disequilibrium, the two pairs are different. In addition, in subsequent analyses of the regulatory effects, this redundancy is taken into account.

III.2.2.5.2.1.1.2 Formation of non-homoelogous pairs

Ten thousands pairs of random SNPs were formed, from non homoelogous genes, with one gene belonging to the A genome and the other B genome. As before, only pairs where $M > 5$ were kept. It was checked that each pair of SNPs was unique. These 10 000 pairs were created by drawing randomly among the 8 776 and 9 437 polymorphic genes of the A genome and B genome, respectively. However, when looking at the 10 000 randomly created pairs, it can be observed they were formed only from 2 252 genes of the A genome and 2 206 genes of the B genome B. Moreover, only 5 658 and 6 047 SNPs of the A and B genome were used, whereas there were 38 981 SNPs in the A genome and 45 729 SNPs in the B genome, available to create pairs. This drastic reduction in the number of genes and SNPs may be due to the fact that only pairs of SNPs where each genotype was represented by at least 5 individuals were kept for analysis. This may have led to the elimination of SNPs with too much missing data, or with too extreme allelic frequencies.

III.2.2.5.2.1.2 Estimation model

Some methods for eQTL mapping use a linear model, and the observed variation in expression is attributed to the genotype of the considered marker and to a component of residual variation (Michaelson et al., 2009). Our model was as well linear, but two SNPs were included: one belonging to the gene whose expression is under consideration (*cis* effect), and one belonging to a gene of the other genome (*trans* effect).Their interaction effect is the epistasis effect. As we did not aim to estimate dominance effects, we considered that homozygous individuals gave us a good estimation of the additive effects and additive x additive epistasis effects, relevant for a highly selfing species. Indeed, it was assumed that in a self-pollinated species such as tetraploid wheat, dominance had only a marginal role. The model was haploid (one allele per locus) as only homozygous genotypes were used

Within our mixed-model framework, we partitioned the total genetic variation into a fixed part, corresponding to the two additive effects and the epistasis effect, and a random part associated with the genetic background. This model is derived from the unweighted regression model "UWR" (Cheverud and Routman, 1995), and allows to estimate the physiological genetic effects, independently of allelic frequencies (presented in section II.2.2.2 and II.4.2.2). Thus, weights were added such that each genotype had the same weight (¼), although the genotype frequencies were not equal to ¼. It was implemented with the function lmekin from the R package coxme, version 2.2-3 (Therneau, 2012), using the maximum likelihood method. This function allows to use a kinship matrix. It performs as well a test to determinate the significance of each genetic effect.

$Y_{ijk} = \mu + x_j\beta_A + x_k\beta_B + x_jx_k\alpha\alpha + b_i + \epsilon_{ijk}$ **(1)**

With:

- Y_{ijk} the expression measure of the gene A or B for the individual i carrying the jk genotype. When the model was applied on homoelogous pairs, the used expression level measure was RPDS. When it was applied on non-homoelogous pairs, the expression level measurement was RPKM if the gene, whose expression was considered, had no homoelog expressed in our dataset. In contrast, if it had one homoelog expressed in our dataset, the expression level was measured in RPDS. This enabled to better discriminate the expression levels between the two homoelogs.
- μ an intercept term
- β_A and β_B the fixed additive effect of the SNP A and the SNP B, respectively; x_j and x_k the coefficients corresponding to the allele of the locus A and B, respectively (x_j = ½ if the allele is A and -½ if the allele is a; x_j = ½ if the allele is B and -½ if the allele is b); $\alpha\alpha$ the fixed epistatic effect.
- b_i the random effect of the genetic background of the individual i; and ε_{ijk} a residual term. The random term b_i is assumed to be normally and independently distributed with zero mean and a proper variance $b_i \sim N(0, \sigma_b^2 K)$, with K being in a kinship matrix and σ_b^2 a variance term.

Estimated regulatory effects are functional effects, depending only on the physiology of the genes, and not on the sample properties. The graphical interpretation of the used model (UWR model) is shown in **Figure 2**. It is based on an unweighted regression.

No population structure was detected by multidimensional scaling, but there was some genetic covariance between individuals. This genetic covariance was included in the kinship matrix K. The kinship coefficients between two individuals were estimated as the identity by state, on the basis of the 84 710 SNPs. The matrix K was not positive definite, but only the last three eigenvalues of 175 were negative, which indicates a minor departure from positive-definite. This problem is common in genetic covariance matrices (Legarra and Fernando, 2009; Meyer and Kirkpatrick, 2010) because kinship coefficients are not estimated using a joint distribution. Instead, each kinship coefficient is computed for two individuals, marginalizing over the rest. The K matrix was corrected with a statistical method called "bending" (Jorjani et al., 2003), which replaces negative eigenvalues by a small positive value (here, 10^{-4}). After replacing the eigenvalues, the diagonal elements were different from 1. Thus, it was normalized by dividing each kinship coefficient $k_{ii'}$ by $\sqrt{k_{ii}\, k_{i'i'}}$ (where i and i' represent two individuals, and $k_{ii'}$ the kinship coefficient between them).

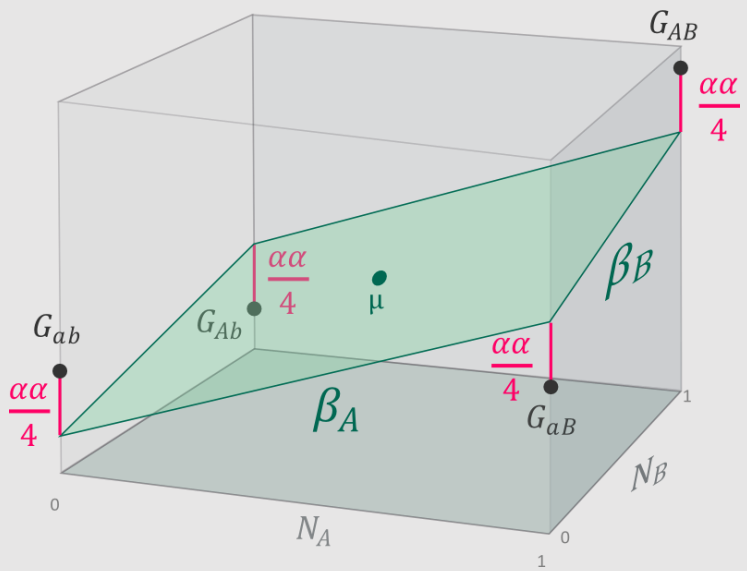


Figure 2 : Graphical illustration of the UWR model in the haploid and bilocus case (Cheverud and Routman, 1995)

G_{jk} = genotypic value of genotype jk = expected phenotype in the absence of environmental effect. Genotypic values are estimated as the average phenotype of individual carrying the jk $G_{jk} = \overline{Y_{jk}}$

N_A = number of A allele at locus A ; N_B = number of B allele at locus B ;

Genetic effects are estimated from an unweighted regression. A plan, passing at the nearest of each genotypic values, is plotted. The slopes of the two lines defining the plan correspond to the additive effects of SNP A (β_A) and SNP B (β_B). The difference between the plan and the genotypic values corresponds to epistasis ($\alpha\alpha$).

This model allows to easily implement kinship matrix and to estimate epistatic interactions. The measure of expression level we used, RPKM or RPDS, is equivalent to a RNA concentration. It has an unbiased mean, but its variance is biased in a gene's length-dependent manner (Oshlack and Wakefield, 2009). However, this did not affect our model, as it focuses on marginal means. It was confirmed that residuals were normally distributed and centred at 0. Overall, residuals did not vary with individuals, except for 10 of them that were a slightly underrestimated.

The effect of a SNP on the expression level of a gene was regarded as *cis* if the SNP belonged to this gene. Conversely, the effect of SNP was considered as *trans* if it altered the expression of a gene to which it did not belong. As only pairs of genes belonging to different genomes were used in the model **(1)**, the identified *trans* effects are regulations by SNPs on expression levels of a gene belonging to the other genome; thus, it was certain that those SNPs were not in linkage disequilibrium with genes they regulated, and therefore that the regulatory effect was not a *cis* effect wrongly categorized as *trans*. The different types of effects are shown in **Figure 3**.

If Y_{ijk} represents the expression measure of the gene A (Y_{ijk}^{A}), then β_A is a *cis* effect. Indeed it is the effect of a SNP on the expression of the gene A to which it belongs. Conversely, β_B is then a *trans* effect,

since it is the effect of a SNP on the expression of a gene to which it does not belong. These effects will be noted cis$A$ and trans$B$→A. The epistasis effect will be noted epi$A$. In contrast, if Y_{ijk} represents the expression level of the gene B (Y_{ijk}^{B}), then β_A is a *trans* effect of a SNP belonging to the A genome on the expression of a gene of the B genome (denoted trans$A$→B) and β_B is a *cis* effect (denoted cis$B$). The epistasis effect is noted epi$B$. *Cis* effects were defined as positive, as they were equal to the up-regulating allele (alleles A and B) effect minus the down-regulating allele (alleles a and b) effect. *Trans* effect were oriented according to the *cis* effect. For example, trans$A$→B (trans$B$→A) effects were calculated as the value of the effect of A (B) minus the value of the effect of a (b) on the expression of the gene B (gene A). This means that, if a *trans* effect is positive, then the causal marker increases both the expression of its own gene and the expression of the other gene. It can therefore be assumed that the product of the gene to which the causal SNP belongs has an activator regulatory effect. In the contrary, a negative *trans* effect may be the sign of inhibitory regulatory action.

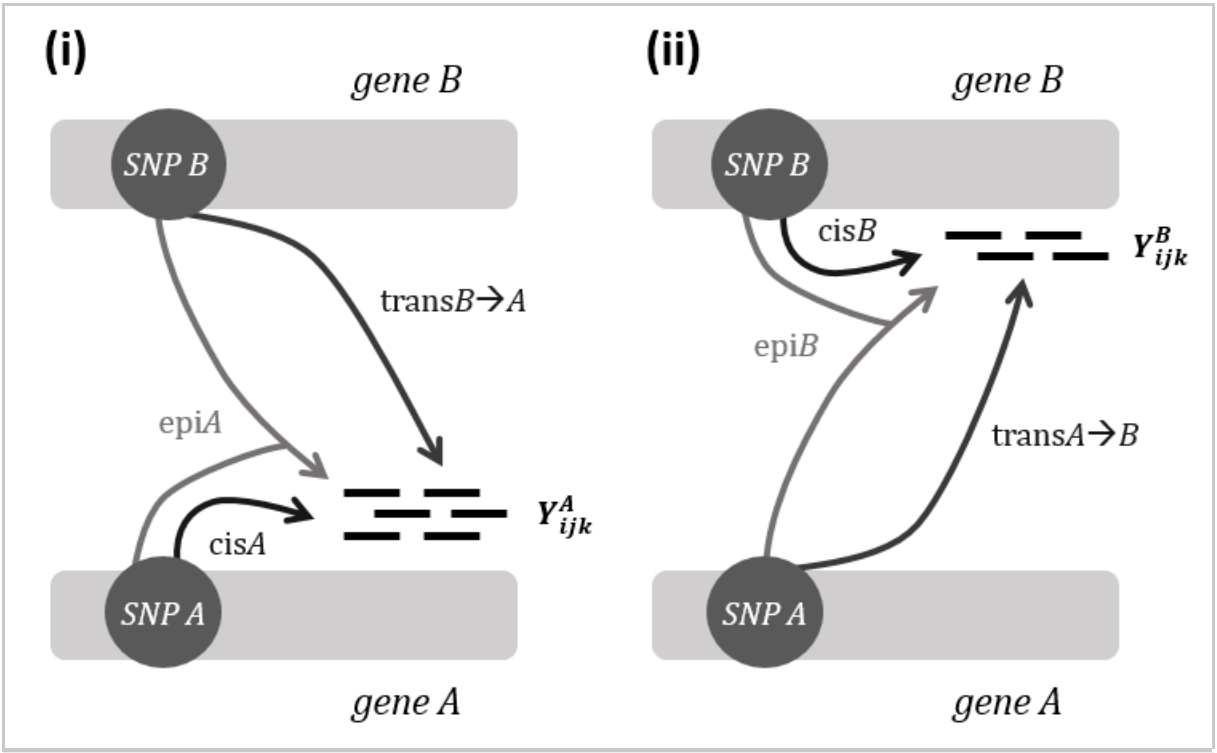


Figure 3: Regulatory effects detected on a pair of genes belonging to the two genomes
The model **(1)** considers two SNPs (SNP A and SNP B) belonging to two genes (genes A and B) of the two genomes of *Triticum turgidum*, A and B.
In (i), the model is applied to expression level of the gene A, $\boldsymbol{Y_{ijk}^{A}}$; in (ii), it is applied to expression level of the gene B, $\boldsymbol{Y_{ijk}^{B}}$.
Cis effects are the effects of a SNP on the expression level of the gene to which it belongs. They are called cis$A$ (cis$B$) if they regulate the expression level of a gene of the A (B) genome.
Trans effects are effects of a SNP on the expression level of another gene. They are denoted trans$B$→A (trans$A$→B) if a SNP belonging to the B (A) genome alters the expression of a gene belonging to the A (B) genome.
Epistasis effects are interaction effects between SNPs on the expression level of one of the two genes. They are called epi$A$ (epi$B$) if they alter expression level of a gene belonging to A (B) genome.

III.2.2.5.2.2 Analysis of eQTL

Once *cis*, *trans* and epistasis effects were estimated and tested, p-values of the significance test were adjusted using Holm-Bonferroni method. Then, we could define two types of elements (**Figure 4**):

- Regulatory effects designed all the β_A, β_B and $\alpha\alpha$ we estimated with the model **(1)**. There were 20 000 (2x10 000) *cis*, 20 000 *trans* and 20 000 epistasis effects estimated on non homoelogous pairs, and 1 234 (2x617) on homoelogous pairs. Several effects could have been estimated from the same SNP, but as the second SNP added in the model was always different, they were considered as different.
- However, to get rid of this redundancy, we examined as well eQTL. A SNP was considered as signing an eQTL if it had a statistically significant *cis* or *trans* effect on the expression level in at least one pair of genes. A pair of SNP was as well considered as an epistatic eQTL if it was involved in significant epistatic effects. Note that the SNP does not necessarily cause directly the expression change, but can be simply in linkage disequilibrium with the causal polymorphism. "*Cis* eQTL", "*trans* eQTL" and "epistasis eQTL" referred to SNPs or pairs of SNPs which had a statistical (but not necessarily biological) significant *cis*, *trans* or epistasis effects on gene expression.

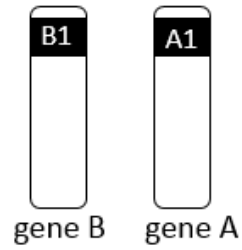
Our goal was to determine whether there were differences in terms of regulation:

- between the homoelogous and non-homoelogous pairs
- between the A and B genomes.

This was examined in two ways. First, in terms of proportion of *cis*, *trans* and epistatic significant effects/eQTL/regulatory genes/regulated genes. In this case, Pearson's chi-squared tests were used. Secondly, in terms of amplitude of the regulatory effects. In this case, the mean of the absolute values of all the effects were compared. Wilcoxon tests (for paired samples) or permutation tests (for unpaired samples) were used. Holm correction was used to adjust p-values.

Estimated values of the genetic effects were standardized. Indeed, regulatory effects were strong when they were estimated from a strong expression level, and weak if estimated from a weak expression level. We divided them by a factor s, so that regulatory effects estimated on high expression levels could be comparable to those estimated from low expression levels.

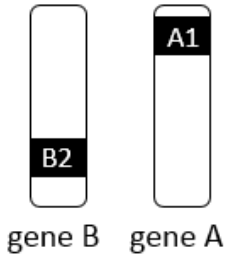
$$s = \sqrt{\sum_{jk} (\overline{Y_{jk}} - \mu)^2}$$



Pair of SNPs n°1, constituted by SNPs B1 and A1

Three effects are estimated for gene A expression (Y^A) :

- Effect of A1 on Y^A (cis$A$ effect) : **significant**
- Effect of B1 on Y^A(trans$B$→A effect) : **significant**
- Effect of A1:B1 interaction on Y^A (epi$A$ effect) : non significant



Pair of SNPs n°2, constituted by SNPs B2 and A1

Six effects are estimated (Y^A) :

- Effect of A1 on Y^A (cis$A$ effect) : **significant**
- Effect of B2 on Y^A(trans$B$→A effect) : non significant
- Effect of A1:B2 interaction on Y^A (epi$A$ effect) : non significant

	Number of effects detected	Number of eQTL detected
cis$A$	2	1 (effect of A1)
trans$B$→A	2	1 (effect of B1)
epi$A$	2	0

Figure 4: Distinction between regulatory effects and eQTL

Consider a pair of genes, gene A and gene B. There is one SNP on the gene A (called A1), and two on the gene B (called B1 and B2). Two pairs of SNPs are constituted: A1-B1 (n°1) and A1-B2 (n°2). For this example, only the expression level of the gene A ($\boldsymbol{Y^A}$) is considered, thus only cis$A$, trans$B$→A and epi$A$ effects are accounted.

As there are two pairs of SNPs, two cis$A$, two trans$B$→A and two epi$A$ effects were estimated. In other words, there are as many effects of each type as pairs of SNPs.

The number of eQTL is the number of SNPs which have a significant effect. In this example, only one SNP has a significant cis$A$ effect on $\boldsymbol{Y^A}$: A1. Thus, there is only one significant cis$A$ eQTL. Then, only one SNP has a trans$B$→A effects (B2), and thus, there is only one trans$B$→A eQTL. Finally, no pairs of SNPs had a significant epistatic effect, so no epistatic eQTL was detected.

With $\overline{Y_{jk}}$ the marginalized mean of the individuals carrying the *jk* genotype; μ the intercept. Let notice that $E(\overline{Y_{jk}}) = \mu + x_j\beta_A + x_k\beta_B + x_jx_k\alpha\alpha$. Here, μ is equal to $\mu = \frac{\sum_{jk}\overline{Y_{jk}}}{4}$. Once standardized, the genetic effects were comprised in intervals equal to [0,1] for *cis* effects, [-1,1] for *trans* effects and [-2,2] for epistasis effects.

Standardization does not influence the detection of eQTL (i.e, tests stating if regulatory effects are significant or not), and consequently it does not have an impact on results of chi-squared tests comparing the proportion of eQTL on each genome, or between homoelogous and non-homoelogous pairs of SNPs. However, this standardization adjusts the magnitude of the estimated regulatory effects, and therefore influences results of the tests comparing magnitudes of regulatory effects between genomes, or between homoelogous and non-homoelogous pairs of SNPs (Wilcoxon tests and permutation tests).

III.2.2.5.3 Relation between the expression level of two homoelogous genes

In a third step, we aimed to determine whether the expression of two homoelogous genes or two non-homoelogous genes could be related at the individual level, and if the genotype of SNPs could create interactions in these relations. The model was:

$$Y_{ijk}^{A} = \mu + \gamma Y_{ijk}^{B} + x_j\beta_A + x_k\beta_B + x_jx_k\alpha\alpha + x_jY_{ijk}^{B}\beta_{A:Y_{ijk}^{B}} + x_kY_{ijk}^{B}\beta_{B:Y_{ijk}^{B}} + x_jx_kY_{ijk}^{B}\alpha\alpha_{AB:Y_{ijk}^{B}} + b_i + \epsilon_{ijk} \quad \textbf{(2)}$$

It was as well implemented with the function lmekin in R. The expression level of the gene B, Y_{ijk}^{B}, was included as a quantitative variable which allows to predict the expression level of the gene A, Y_{ijk}^{A}. If the effect of Y_{ijk}^{B} is significant, this means that the expression level of the gene of the B genome, Y_{ijk}^{B}, is related to the expression level of another gene on the A genome, Y_{ijk}^{A}, independently to any SNP in the model. The genotype of the locus A, the locus B and the epistasis between then were added as factors. The effects noted $x_jY_{ijk}^{B}\beta_{A:Y_{ijk}^{B}}$, $x_kY_{ijk}^{B}\beta_{B:Y_{ijk}^{B}}$ and $x_jx_kY_{ijk}^{B}\alpha\alpha_{AB:Y_{ijk}^{B}}$ represent the interactions between the factors and the expression level of the gene B. If they are significant, this means that SNPs or their cross-allelic interaction (epistasis) can interact with Y_{ijk}^{B} to explain Y_{ijk}^{A}. These effects may be the sign of regulatory effects.

We wanted to test several hypotheses. First, we assumed that similar gene sequences tended to be subject to the same *trans* regulations, from across the genome. Therefore, we assumed that their expression levels were correlated. Our hypothesis is as follows: the expression levels of two homoelogous genes (similar sequences) have more chance to be correlated than the expression levels of two non-homoelogous genes (dissimilar sequences).

Second, we assumed that this model would enable us to identify, within pairs of homoelogs, regulatory effects which were not detected with the model **(1)**. Indeed if a SNP has a *cis* statistical effect on the expression level of a gene, but that the expression is also subject to strong *trans* effects from the entire genome, the *cis* effect may not be detectable by the model **(1)**, because it may be hidden by the *trans* regulations from the wide genome. These regulations are genetic effects due to the genetic background. Model **(2)** allows the detection of *cis* effects in another way, not by examining the difference in expression levels between alleles, but looking at the expression ratios between the two genes. If there is a *cis* effect which regulates the expression of one of the two homoelogs, then it will modify the intercept of the $Y_{ijk}^{A} = f(Y_{ijk}^{B})$ regression, in an allele-specific way. In addition, interactions between a SNP and the genetic background can be observed if the slope of the $Y_{ijk}^{A} = f(Y_{ijk}^{B})$ regression is changed depending on the allele at the SNP. Therefore, a regulatory effect can be detected if a SNP interacts with the $Y_{ijk}^{A} = f(Y_{ijk}^{B})$ regression in an allele-specific or a genotype-specific manner. The **Figure 5** illustrates graphically this model.

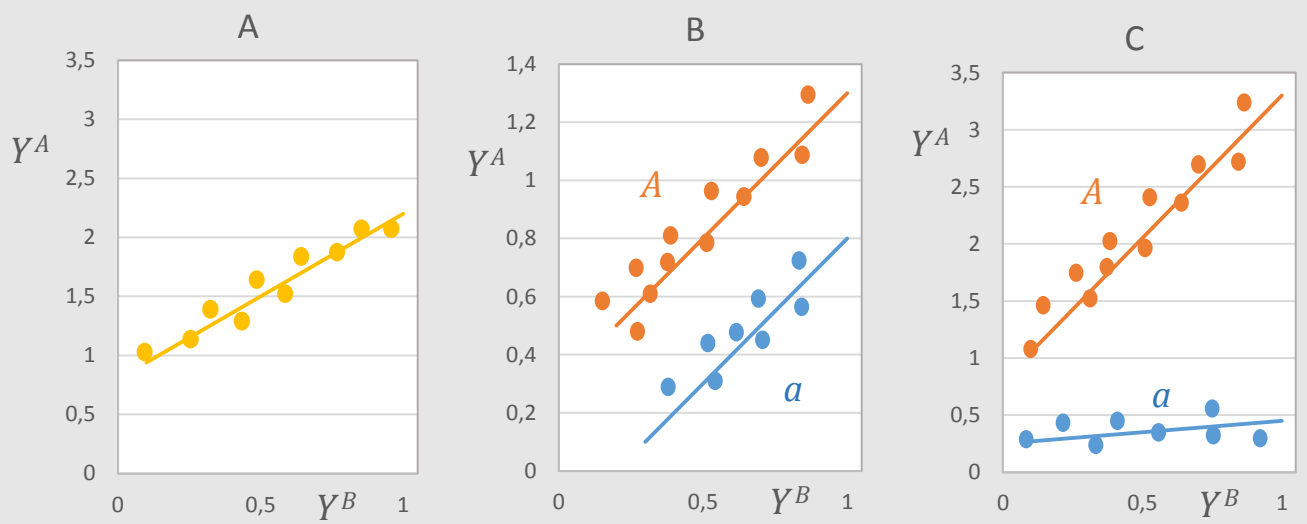


Figure 5: Graphical illustration of the model (2).
x-axis **:** Y^B = expression level of gene B
y-axis : Y^A = expression level of gene A
Each point represent one individual. Lines represent significant $Y_{ijk}^{A} = f(Y_{ijk}^{B})$ regression

A. Significant effect of Y^B on Y^A. The SNPs included in the model have no significant effect. The fact that there is a straight line and not a single point, indicates that the genetic background has an effect on the expression levels of both genes. Thus the regression is due to *trans* effects from the whole genome.
B. Significant effect of Y^B on Y^A. The SNP A has a main significant effect on Y^A, as individuals carrying the A allele have a stronger Y^A (expression level of the A gene) than those carrying the a allele. The slopes of the two regressions remain identical, which indicates no interaction between the SNP B and the genetic background.
C. Significant effect of Y^B on Y^A. The SNP A has a main significant effect on Y^A, as individuals carrying the A allele have a stronger Y^A (expression level of the A gene) than those carrying the a allele. In addition, the slopes of the two regressions are different, which indicates interactions between the SNP B and the genetic background.

This model was applied to the 617 pairs of homoelogs genes (to identify significant effects of Y_{ijk}^{B}, $x_j Y_{ijk}^{B} \beta_{A:Y_{ijk}^{B}}$, $x_k Y_{ijk}^{B} \beta_{B:Y_{ijk}^{B}}$ and $x_j x_k Y_{ijk}^{B} \alpha\alpha_{AB:Y_{ijk}^{B}}$), and to the 10 000 pairs of non-homoelogous genes (to identify significant effects of Y_{ijk}^{B}).

III.2.2.5.4 Role of regulations in expression bias between homoelogs

We aimed to determine the role of eQTL in expression bias between homoelogs. Homoelogous pairs of genes were classified in three categories: 1) no significant expression bias, 2) significant expression bias in favour of the A genome, and 3) significant expression bias in favour of the B genome, based on the p-values of the Wilcoxon test, corrected with the Holm method.

Using a Pearson's chi-squared test, we aimed to compare these three categories of homoelogous pairs, in terms of proportion of:

- *cis*, *trans* and epistatic eQTL detected by the model **(1)**
- SNPs interacting with the $Y_{ijk}^{A} = f(Y_{ijk}^{B})$ regression, detected by the model **(2)**

Intensity of effects were also compared, using permutation tests (10 000 permutations).

III.2.2.5.5 Gene ontologies

Gene ontologies (Ashburner et al., 2000) are available on the Ensembl database (Hubbard et al., 2002). Chi squared tests were performed to compare the proportion of gene functions between two samples of genes, such as:

- *Cis* regulated genes *versus* non *cis* regulated genes
- *Trans* regulated genes *versus* non *trans* regulated genes
- *Trans* regulatory genes *versus* non *trans* regulatory genes
- Epistasis regulated genes *versus* non epistasis regulated genes

P-values were corrected using Holm correction.

III.2.3 Results

III.2.3.1 Comparison of the expression level and fractionation process between the two genomes

Among the 62 166 expressed genes identified in our dataset, 30 130 (48%) belonged to the A genome and 32 036 (52%) the B genome. These numbers were compared to the number of genes (annotated and showing homology to related grass) that the two genomes of *Triticum aestivum* contain (Mayer et al., 2014a): 40 253 genes (47%) for the A genome and 44 523 (53%) for the B genome. The difference between proportions of A and B genes in our data and in genomic data was slight but significant (χ^2 =

13.92, df = 1, p-value = 0.00019), which means there was a transcription bias in favour of the A genome, compared to the gene composition of the bread wheat genome.

The average expression level of the 30 130 genes of A genome - expressed as RPKM - was compared to that of the 32 081 genes of the B genome, using a permutation test. The difference of expression level (11.9 RPKM for the A genome and 11.4 for the B genome) was not significant (p-value = 0.17). Thus, no average expression bias was found with this method. Secondly, the average expression level (expressed in RPDS) was compared between homoelogous genes. A Wilcoxon test indicated that the average expression levels of genes from the A genome and the B genome were not significantly different (p-value = 0.19; average RPDS: 1.06 for the A genome and 1.05 for the B genome). Finally, among the 13 605 pairs of homoelogous genes available for our test, 8 894 (68.1%) had a significantly expression bias. For 4 579 (51.5%) pairs of genes, the level of expression of the gene belonging to the A genome was significantly stronger than that of the gene of the B genome. In contrast, in 4 315 pairs (48.5%), the expression level of the genes from the B genome was significantly stronger. A chi-squared test indicated that the distribution was not uniform (χ^2 = 7.84, df = 1, p-value = 0.0051): there was significantly more homoelogous pairs where the expression level of the gene copy of the A genome was significantly higher than that of the copy belonging to the genome B, even if the gap was small.

III.2.3.2 Regulatory effects

The main findings on regulatory effects are summarized in **Table 1** and **Table 2**. Table 1 shows the proportions of significant effects, eQTL, regulated genes and regulatory genes detected on non-homoelogous and homoelogous pairs of genes. Table 2 lists the intensity of regulatory effects.

III.2.3.2.1 Among non homoelogous pairs of genes

The 10 000 non-homoelogous pairs of SNPs were constituted with 5 658 and 6 047 SNPs of the A and B genome, respectively. Among them, there were 1 278 *cis* eQTL (22.59%), spread on 406 of the 2252 genes from the A genome, and 1 992 *cis* eQTL (32.94%), spread on 501 of the 2206 genes from the B genome (Table 1, first column). The chi-squared test indicated there were significantly more *cis* eQTL detected on the genome B than on the genome A (χ^2 = 155.1533, df = 1, adjusted p-value < 2.2e-16). There were also significantly more genes subjected to *cis* regulations on the B genome than on the A genome (501 genes from the B genome, again 406 from the A genome: χ^2 = 16.93, df = 1, adjusted p-value = 0.00027). In addition, the average value of 10 000 cis$B$ effects (estimated on the 10 000 pairs), were significantly stronger than cis$A$ effects (average value equal 0.76 against 0.73; adjusted Wilcoxon signed-rank test p-value < 2.2e-16 (Table 2, first column). Remember that these effects were standardized.

Table 1: Number and proportion of significant eQTL detected on non-homoelogous and homoelogous pairs of genes

		Detected on non homoelogous pairs	Detected on homoelogous pairs	Chi squared test p-values after Holm correction
eQTL	cis$A$	1278/5658 (22.59%)[b]	72/596 (12.08%)[b]	**1,66E-08**
	cis$B$	1992/6047 (32.94%)[a]	106/593 (17.88%)[a]	**3,57E-13**
	transAB	11/5658 (0.19%)[c]	10/596 (1.68%)[c,d]	**7,12E-08**
	transBA	9/6047 (0.15%)[c]	17/593 (2.87%)[c]	**9,17E-22**
	epi$A$	4/10000 (0.04%)[c]	1/617 (0.16%)[d]	1,00E+00
	epi$B$	10/10000 (0.10%)[c]	1/617 (0.16%)[d]	1,00E+00

Different letters indicate significant differences (p-value ≤ 0.05) after Holm correction, within the same type of pairs (homoelogous or non homoelogous).
Values in the last column on the right are corrected chi-squared tests p-values, comparing the proportion of elements detected on homoelogous and non-homoelogous pairs of genes. In bold: significant differences (p-value ≤ 0.05) after Holm correction.
Each individual SNP constitutes a potential *cis* or *trans* eQTL. Each pairs of SNP constitutes a potential epistatic eQTL.
5 648 = number of SNPs on the genome A, constituting non homoelogous pairs
6 047 = number of SNPs on the genome B, constituting non homoelogous pairs
10 000 = number of non homoelogous pairs
596 = number of SNPs on the genome A, constituting homoelogous pairs
593 = number of SNPs on the genome B, constituting homoelogous pairs
617 = number of homoelogous pairs

Table 2: Intensity of regulatory effects detected on non-homoelogous and homoelogous pairs of genes

	Detected on non homoelogous pairs	Detected on homoelogous pairs	Permutation test p-values	Permutation test p-values after Holm correction
cis$A$	0.73[b]	0.64[b]	0,00E+00	**0,00E+00**
cis$B$	0.76[a]	0.68[a]	0,00E+00	**0,00E+00**
transAB	0.31[f]	0.38[c]	0,00E+00	**0,00E+00**
transBA	0.35[e]	0.44[c]	0,00E+00	**0,00E+00**
epi$A$	0.67[c]	0.75[a]	2,00E-04	**2,00E-04**
epi$B$	0.60[d]	0.74[a]	0,00E+00	**0,00E+00**

The intensity of regulatory effect is equal to the mean of the absolute standardized values.
Different letters indicate significant differences (p-value ≤ 0.05) after Holm correction, within the same type of pairs (homoelogous or non homoelogous).
Values in the last two columns on the right are corrected and non-corrected permutation tests p-values (10 000 permutations), comparing the intensity of regulatory effects detected on homoelogous and non-homoelogous pairs of genes. In bold: significant differences (p-value ≤ 0.05) after Holm correction.

There were only 9 trans$A$→B (0.15%) and 11 trans$B$→A (0.19%) eQTL, spread on 9 and 10 genes, respectively. The mean of all the trans$B$→A effects (0.35) was significantly stronger than that of the trans$A$→B effects (0.31) (adjusted Wilcoxon signed-rank test p-value < 2.2e-16). There were 14 significant epistatic eQTL (0.14%) altering the expression of 4 genes belonging to the A genome (epi$A$) and 10 genes from the B genome (epi$B$). In addition, epi$A$ effects were significantly stronger than epi$B$ effects (0.67 against 0.60; adjusted Wilcoxon signed-rank test p-value < 2.2e-16).

Globally, there were significantly more *cis* eQTL, than *trans* (χ^2= 3661.86, df = 1, adjusted p-value < 2.2e-1) and epistatic ones (χ^2= 3677.262, df = 1, adjusted p-value < 2.2e-16). Furthermore, *cis* effects were significantly stronger than *trans* (average absolute values: 0.74 against 0.33; adjusted Wilcoxon test p-value < 2.2e-16) and epistasis ones (average absolute values: 0.74 against 0.64; adjusted Wilcoxon test p-value < 2.2e-16). However, these effects are statistical effects, not necessarily biological ones: SNPs that have a statistical effect on expression are not necessarily causal polymorphisms, but may be in linkage disequilibrium with causal polymorphisms. It might be easier to get SNPs close to a *cis*-acting causal polymorphism than to a *trans*-acting one, and therefore to detect *cis* effects than *trans* ones.

The polymorphism levels of the SNPs from the B genome were slightly stronger than that of the A genome: the average expected Nei's heterozygosity was equal to 0.42 for the B genome against 0.41 for the A genome. This slight difference was significant (p-value=0.010).

III.2.3.2.2 Among homoelogous pairs of genes

There were 596 and 593 SNPs of the A and B genome, respectively, constituting the 617 pairs of SNPs belonging to homoelogous genes. Among them 76 on 596 (12.32%) had a significant cis$A$ effect, and 109 on 593 (17.67%) a significant cis$B$ effect (Table 1, second column). These eQTL were spread on 70 genes of the A genome and 105 genes of the B genome. There were significantly more *cis* eQTL on the B genome than on the A genome (χ^2= 7.39, df = 1, adjusted p-value = 0.033), and significantly more cis-regulated genes on the B genome than on the A genome (105 genes of the B genome against 76 on the A genome: χ^2= 7.79, df = 1, adjusted p-value = 0.026). Moreover, the average value of cis$B$ effects (0.68) was significantly higher than that of cis$A$ effects (0.64) (adjusted Wilcoxon signed-rank test p-value = 0.020) (Table 2, second column).

There were respectively 10 (1.7%) and 17 (2.9%) trans$A$→B and trans$B$→A eQTL, with one eQTL per gene. This difference was not significant (χ^2= 1.36, df = 1, adjusted p-value = 0.48). However the trans$B$→A effects were on average significantly stronger than the trans$A$→B effects (average absolute values: 0.44 against 0.38; adjusted Wilcoxon signed-rank test p-value = 0.0017). No pair of homoelogs had both significant trans$A$→B and trans$B$→A effects.

There were only one significant epistatic eQTL, altering the expression of the two genes of the same pair of homoelogs. Values of $\mathrm{epi}A$ effects were not significantly different from values of $\mathrm{epi}B$ effects (adjusted Wilcoxon signed-rank test p-value = 1).

There were significantly more *cis* eQTL, than *trans* (χ^2= 117.97, df = 1, adjusted p-value < 2.2e-16) and epistatic ones (χ^2= 181.7445, df = 1, adjusted p-value < 2.2e-16). *Trans* eQTL were as well significantly more numerous than epistasis ones (χ^2= 20.11, df = 1, adjusted p-value = 2.20e-05). Furthermore, *cis* effects were significantly stronger than *trans* ones (average absolute values: 0.66 against 0.41; adjusted Wilcoxon test p-value < 2.2e-16). Epistasis effects were as well stronger than *trans* ones (average absolute values: 0.75 against 0.41; adjusted Wilcoxon test p-value < 2.2e-16).

The polymorphism level of the SNPs from the B genome were slightly higher than that of the SNPs from the A genome (He = 0.41 for B, He = 0.40 for A). However, the difference was not significant (p-value = 0.061).

III.2.3.2.3 Comparison of the estimates of the genetic effects between homoelogous and non-homoelogous pairs of genes

We wanted to see if there was a larger or smaller proportion of eQTL detected on homoelogous pairs, compared to non-homoelogous pairs (Table 1, third column). In addition, intensities of effects detected on the two types of pairs were compared (Table 2, third column).

It is expected that there is the same proportion of significant *cis* effects among the two types of pairs. Indeed, pairs of genes differ in their nature (homeologous pairs vs aleatory pairs), but not the genes which constitute these pairs. Indeed, among the genes constituting non-homoelogous pairs, some of them had homoelogs in the rest of the genome and were selected to constitute homeologous pairs. *Cis* effects are the property of one single gene, because it is basically a gene which self-regulates. In contrast, *trans* effects are the property of a pair of genes (one which regulates, and the other one which is regulated). It is the same for interaction effects (epistasis). Therefore, if *trans* and/or epistatic effects vary (in intensity or in proportion of in frequency of significance), depending on the type of pairs of genes, it is likely due to biology, at least partially. However, if *cis* effects do, this is probably an artefact.

There was a significantly greater proportion of significant *cis* eQTL among the non-homoelogous pairs than among pairs of homoelogous genes (χ^2= 88.9, df = 1, p-value < 2.2e-16). Moreover, values of *cis* effects estimated on non-homoelogous pairs were stronger than those estimated on homoelogous pairs (averages values: 0.74 against 0.66; adjusted permutation test p-value = 0). As explained above, these differences in intensity and proportion of significant *cis* effects/eQTL/regulatory genes are

probably due to the model, not to an intrinsic biological property of genes having no homoelogous copy.

In contrast, significant *trans* eQTL were far more numerous among the homoelogous pairs than among the non-homoelogous ones (χ^2= 125.34, df = 1, adjusted p-value < 2.2e-16). In addition, *trans* effects were significantly stronger in homoelogous pairs (average absolute values: 0.41 for homoelogs versus 0.33 for non-homoelogs; adjusted permutation test p-value = 0).

Finally, there was not a greater proportion of significant epistatic eQTL among homoelogs than among non-homoelogs (χ^2 = 0.3715, df = 1, p-value = 1). However, epistasis effects estimated on homoelogous pairs were significantly stronger (average absolute values: 0.75 for homoelogs versus 0.64 for non-homoelogs; adjusted permutation test p-value = 0). The number of epistatic eQTL is surprisingly low.

III.2.3.2.4 Effect of the kinship matrix

The model **(1)** was run with and without the kinship matrix, on all the pairs of SNPs. This allowed us to obtain for all the pairs the deviance (d) between the two models: the "full model" (with the K matrix) and the "constrained model" (without the K matrix). The distribution of the deviances are shown in **Figure 6**.

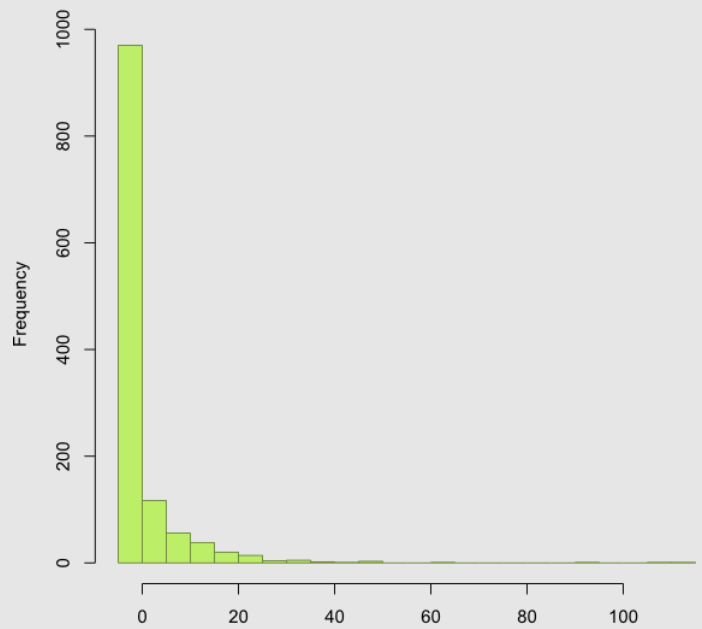


Figure 6: Distribution of the deviances d for the 617 homoelogous pairs of SNPs.

The distribution of the deviances of the 10 000 non-homoelogous pairs of SNPs was similar (data not shown).

The more positive the deviance d is, the higher the likelihood of the full model (with K) is high relative to the one of the constrained model (without K) model. In other words, the stronger d is, the more

suitable the use of the matrix K is. Therefore, it can be observed that for most pairs of SNPs where $d < 0$, the use of the kinship matrix K is not necessary. However, the kinship matrix is useful for approximately 20% of them, where $d > 0$.

III.2.3.3 Relation between the expression of two homoelogous genes

In this section, we examined if the variation in the expression level of a gene of the B genome among individuals,Y_{ijk}^{B}, could affect the expression level of another gene on the A genome, Y_{ijk}^{A}.

Indeed, homoelogous genes having similar sequences, it was assumed that they were subject to the same overall *trans* regulations, and that therefore, their expression levels were frequently correlated. This was observed. Indeed, among the 548 pairs of homoelogous genes, 427 (77.9%), expression levels of the two genes were significantly (p-value=0.05) related. All the correlations were positive. Moreover, 109 pairs (19.8%) did not have a significant Y_{ijk}^{B} effect on Y_{ijk}^{A}, and 12 pairs (2.2%) were ambiguous, which means that the effect of Y_{ijk}^{B} on Y_{ijk}^{A} was significant if some SNPs were included in the model, but that if other SNPs, belonging to the same gene, were considered, this significant effect disappeared. An example is given in **Figure 7**.

In contrast, among the 10 000 random pairs of non-homoelogous genes, only 631 (6.31%) showed a significant Y_{ijk}^{B} effect on Y_{ijk}^{A}. Thus, pairs of homoelogous genes were significantly more likely to have two significantly related levels of expressions than randomly drawn pairs of genes (χ^2= 2944.25, df = 1, adjusted p-value < 2.2e-16).

We also assumed that if one of the two homoelogous genes was specifically subjected to a *cis* or *trans* regulation, the involved SNP would interact with the $Y_{ijk}^{A} = f(Y_{ijk}^{B})$ regression, by modifying the intercept and/or the slope of the regression in a allele or genotype-specific way. Consequently, we also verified if SNPs could interact significantly – solely or in epistatic interaction - with Y_{ijk}^{B} to explain Y_{ijk}^{A} in homoelogous pairs. Results are summarized in **Table 3**.

Among the 596 SNPs spread over the 548 genes of the A genome, 13 induced an interaction on the $Y_{ijk}^{A} = f(Y_{ijk}^{B})$ regression between homoelogous genes. Similarly, among the 593 SNPs in the B genome SNP, 27 had an interaction effect on this regression. Each of these interacting SNPs belonged to a different gene. Proportions of these interactions in the two genomes were compared. However, the chi-squared test was not significant any more after Holm correction (χ^2 = 4.44, df = 1, adjusted chi-squared test p-value = 0.070), so it cannot be stated that there were significantly more SNPs of the B genome which induced interactions with the $Y_{ijk}^{A} = f(Y_{ijk}^{B})$ regression, than SNPs of A genome. Finally, epistasis of 6 pairs of SNPs induced an interaction effect on the $Y_{ijk}^{A} = f(Y_{ijk}^{B})$ regression. Among these 6 pairs, 3 had already a simply interaction effect.

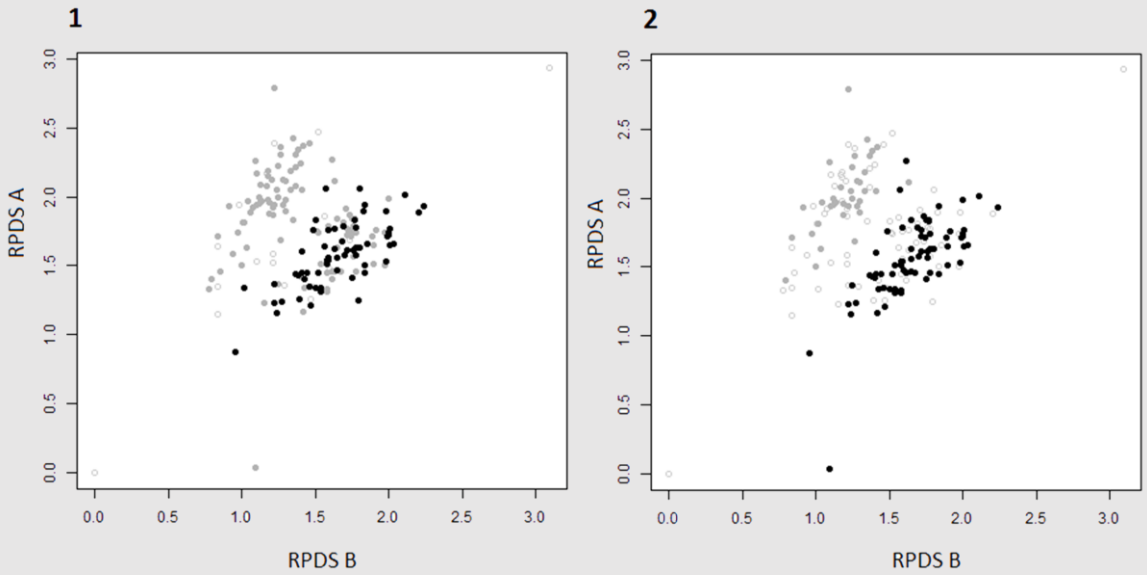


Figure 7: An example of ambiguous correlation between the expression levels of two genes
Pair of homoelogs: Traes_6AS_4F0EAD407 and Traes_6BS_FF3A794F8. Each point represent one individual.
x-axis: RPDS B: expression level of the gene belonging to the genome B (Traes_6BS_FF3A794F8), expressed in RPDS
y-axis: RPDS A, the expression level of the gene belonging to the genome A (Traes_6AS_4F0EAD407), expressed in RPDS
1: the relation between the expression levels is not significant (p-value = 1), when a specific SNP of the B genome is considered. Grey points: allele b of this SNP; black points: allele B; empty circles: missing data
2: the relation between the expression levels becomes significant (p-value = 0.033) when another SNP belonging to the gene Traes_6BS_FF3A794F8 is included in the model. Grey points: allele b of this SNP; black points: allele B; empty circles: missing data

Table 3: Results of the model (2). $Y_{ijk}^{A} = f(Y_{ijk}^{B})$ regression and proportion of significant interaction effects detected in homoelogous pairs.

	Detected on non homoelogous pairs	Detected on homoelogous pairs	Chi squared test p-values
Pairs of genes with significant or ambiguous Y_{ijk}^{B} effects	631/10000 (6.31%)	439/548 (80.11%)	0,00E+00
Interacting SNPs A	-	13/596 (2.22%)[a,b]	-
Interacting SNPs B	-	27/593 (4.55%)[a]	-
Cross-allelic interacting pairs of SNPs (epistasis)	-	6/617 (0.97%)[b]	-

Different letters indicate significant differences (p-value ≤ 0.05) between proportions, after Holm correction, within pairs of homoelogous genes.
The p-value in the last column is the result of the chi-squared test comparing the significant proportion of expression level correlations, detected on homoelogous and non-homoelogous pairs of genes.
Y_{ijk}^{A}= expression level of the gene A; Y_{ijk}^{B}= expression level of the gene B
Pairs of genes with significant or ambiguous Y_{ijk}^{B} effects = pairs of genes where the effect of Y_{ijk}^{B} was significant for at least one pair of SNPs
Interacting SNPs = SNPs involved in at least one significant simple interaction effect ($\beta_{A:Y_{ijk}^{B}}$ or $\beta_{B:Y_{ijk}^{B}}$)
Cross-allelic interacting pairs of SNPs (epistasis) = pairs of SNPs involved in a least one significant cross-allelic interaction effect ($\alpha\alpha_{AB:Y_{ijk}^{B}}$)

Usually, in pairs of homoelogous genes where there were significant interaction effects ($\beta_{A:Y_{ijk}^{B}}$, $\beta_{B:Y_{ijk}^{B}}$ and $\alpha\alpha_{AB:Y_{ijk}^{B}}$), Y_{ijk}^{B} had a significant effect on the expression level of Y_{ijk}^{A}, i.e. expression levels of the two genes were related. It was the case for 39 (95%) of the 41 pairs with a significant interaction effect. Furthermore, nearly half of the SNPs involved in such interactions were also statistically associated to a *cis* effect. For example, of the 13 SNPs in the A genome and the 27 SNPs in the B genome which interacted significantly with the $Y_{ijk}^{A} = f(Y_{ijk}^{B})$ regression, 6 (46.2%) and 16 (59.3%) had a significant *cis* effect, respectively. As for the 6 cross-allelic interaction effects (epistasis), they all involved at least one SNP which had a significant *cis* effect.

Graphics representing $Y_{ijk}^{A} = f(Y_{ijk}^{B})$ were plotted for pairs of SNPs where significant interactions were detected. Several graphic types have been identified and some are shown in **Figure 8**:

- When the interacting SNP was also statistically associated to a significant *cis* effect, two distinct clouds were visible (Figure 8.1). In some cases, these clouds were imperfect, i.e. one cloud was mostly composed of genotypes carrying A (or B) alleles, and the other one of a (or b) alleles, but there were some outliers. This may be due to the fact that the SNP is in strong – but not maximal – linkage disequilibrium with the causal polymorphism inducing *cis* effect. This type of graph (two allele-specific clouds associated with a significant *cis* effect) involved 46.2% (6 of 13), 59.3% (16 of 27) and 66.7% (4 of 6) of the significant interactions between Y_{ijk}^{B} and a SNP A, a SNP B and an epistatic effect, respectively.
- Sometimes clouds were not allele-specific, but genotype-specific (Figure 8.2). There could be four clouds, with one genotype per cloud; or two clouds, one containing one genotype, and the other one, three genotypes. This type of graphics involved 7.7% (1 in 13), 7.4% (2 of 27) and 33.3% (2 of 6) significant interactions between Y_{ijk}^{B} and a SNP A, a SNP B and an epistatic effect, respectively.
- In rare cases (2 out of 27 cases of significant interaction between Y_{ijk}^{B} and a SNP B), allele-specific clouds were formed, but there were two clouds for a SNP (Figure 8.3). This could be a sign of interaction with a third gene.
- In some cases, there was only one cloud, but two allele-specific sub-clouds could be distinguished from each other (Figure 8.4). The interaction may be due to a weak *cis* eQTL which remained undetected because of a lack of power. This concerned 38.5% (5 of 13) and 7.4% (2 of 27) of significant interactions between Y_{ijk}^{B} and a SNP A, and between Y_{ijk}^{B} and a SNP B, respectively.

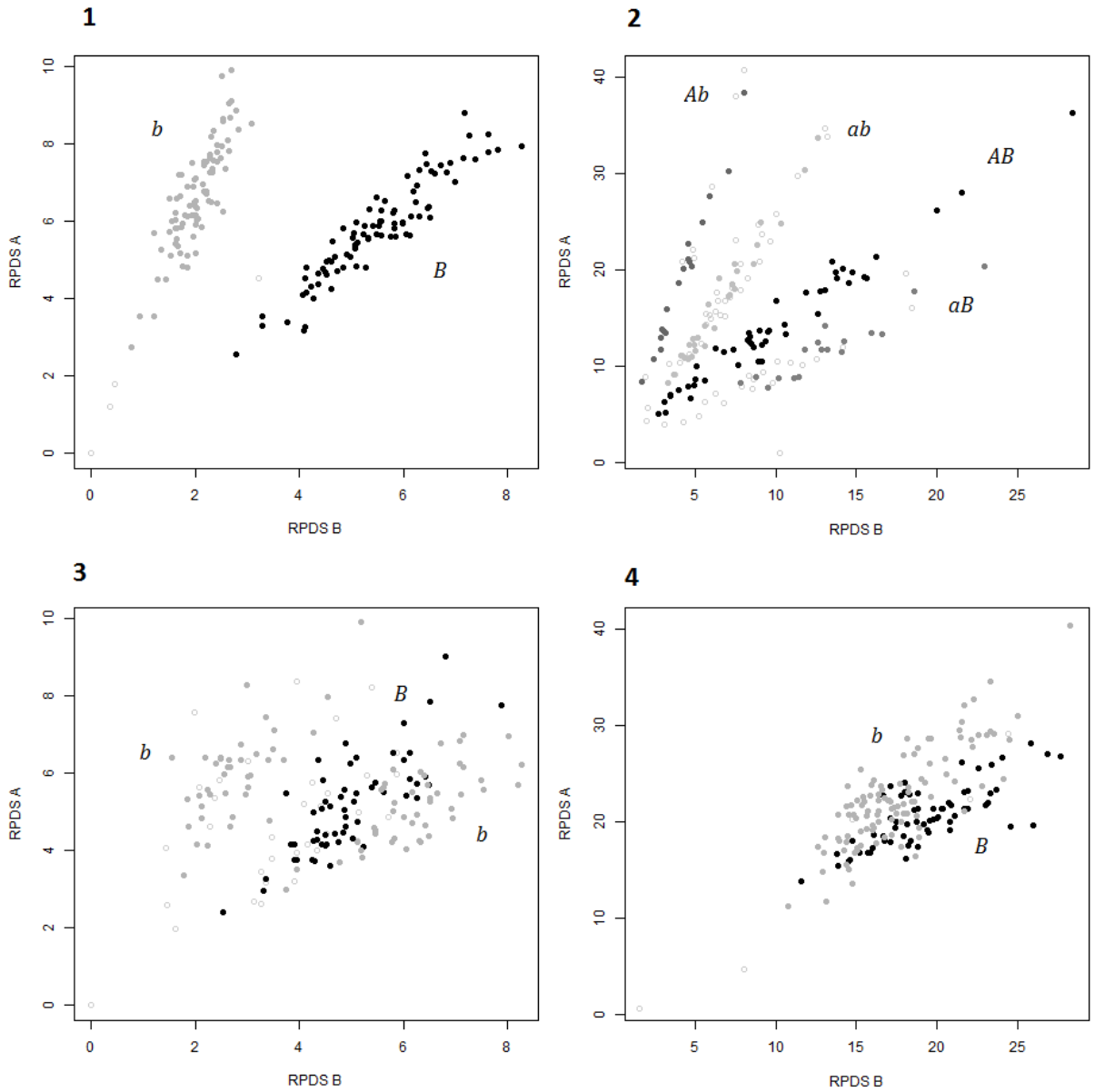


Figure 8: Interactions between genotypes and the expression level of the gene B to explain the expression level of the gene A

x-axis: RPDS B: expression level of the gene belonging to the genome B, expressed in RPDS

y-axis: RPDS A, the expression level of the gene belonging to the genome A, expressed in RPDS

Each point represent one individual. Empty circles: missing data

1: Pair of homoelogs: Traes_7AL_CBE99239D and Traes_7BL_8849C9100. Grey points: allele b; black points: allele B. The interaction effect between the genotype of the SNP B and the expression level of the gene B was significant (model **(2)**, adjusted p-value = 0.0017). In addition the SNP B was also associated to a significant cis effect (model **(1)**, adjusted p-value = 0)

2: Pair of homoelogs: Traes_6AS_236AF32FD and Traes_6BS_804DA0429. From lightest to darkest gray: genotypes ab, aB, Ab and AB. All interaction effects (SNP A interaction, SNP B interaction, and cross-allelic interaction) were significant (model **(2)**, adjusted p-values = 0, 0 and 1.35e-13, respectively). In addition the SNP B was also associated to a significant cis effect (model **(1)**, adjusted p-value = 3.37e-08)

3: Pair of homoelogs: Traes_7AL_16DB61F77 and Traes_7BL_B97BBAFAD. Grey points: allele b; black points: allele B. The interaction effect between the genotype of the SNP B and the expression level of the gene B was significant (model **(2)**, adjusted p-value = 1.29e-08). No significant regulation effect was detected on this pair by the model **(1)**.

4: Pair of homoelogs: Traes_2AL_1CC09E982 and Traes_2BL_89749BD6C. Grey points: allele b; black points: allele B. The interaction effect between the genotype of the SNP B and the expression level of the gene B was significant (model **(2)**, adjusted p-value = 0.0040). No significant regulation effect was detected on this pair by the model **(1)**.

- Finally, in some cases, the graph did not show any particular form, and therefore was not really informative. This type of graphics involved 7.7% (1 of 13) and 18.5% (5 of 27) of the significant interactions between Y_{ijk}^{B} and a SNP A, and between Y_{ijk}^{B} and a SNP B, respectively.

For most of the cases where there were allele or genotype-specific clouds, a change in the slope of the regression was observed, which indicates interactions with the genetic background. This was the case for 5 of the 6 allele-specific regressions involving an allele of the SNP A, 13 of the 16 allele-specific regressions involving an allele of the SNP B, and for the 2 genotype-specific regressions. Therefore, 20 pairs of homoelogous genes showed interactions with the genetic background.

In conclusion, interaction effects seemed to occur when there was a relation between the expression levels of the two homoelogous genes. This relation is probably due to similarities in their sequence, and thus to *trans* regulations they are subjected to. Interaction effects seem mainly due to *cis* regulations that modify the intercept of regression line, and to SNP x genetic background interactions that modify the slope of the regression. **Figures 9** and **10** provide details on how *cis* regulations can lead to this type of interaction, in particularly when allele-specific or genotype-specific clouds can be distinguished on the graphic $Y_{ijk}^{A} = f(Y_{ijk}^{B})$.

Figure 9 shows the case of the pair of homoelogs Traes_7AL_CBE99239D (gene A) and Traes_7BL_8849C9100 (gene B). Two distinct elongated clouds can be observed, one with individuals carrying the allele b, the other with those carrying the allele B. These clouds were modelled as regression lines. In this homoelogous pair, the expression of the B gene is subjected to *cis* and *trans* regulations (from the entire genome), while the gene A is only subjected to the latter. The points ❶ and ❷ represent the allele b subjected to two different genetic backgrounds, and therefore to different *trans* regulations (inhibitory in ❶, enhancer in ❷). The points ❸ and ❹ represent the allele B subjected to the same *trans* regulations. The change in slope as a function of the allele at the SNP B indicates that there are also interactions between the SNP B and the genetic background. Indeed, between points 1 and 3 (similar genetic backgrounds with inhibitory *trans* effects, but different alleles at the SNP B), the additive *cis* effect of the SNP B is equal to $\beta_B = 4 - 1 = 3$. Conversely, between points 2 and 4 (equivalent genetic background effect with enhancer *trans* effects), the additive *cis* effect of the SNP B is equal to $\beta_B{}' = 8 - 3 = 5$.

Figure 10 shows the pair of homoelogs Traes_6AS_236AF32FD (gene A) and Traes_6BS_804DA0429 (gene B). Four genotype-specific regressions (elongated clouds) can be distinguished. These four regressions can be explained by a cis$B$ effect and a cis$A$ effect, and by *trans* regulations from the entire genome, affecting both the expression of the gene A and the gene B. Note however that in this pair of homoelogs, the cis$A$ effect was not considered as significant by the model **(1)** (adjusted p-value= 1).

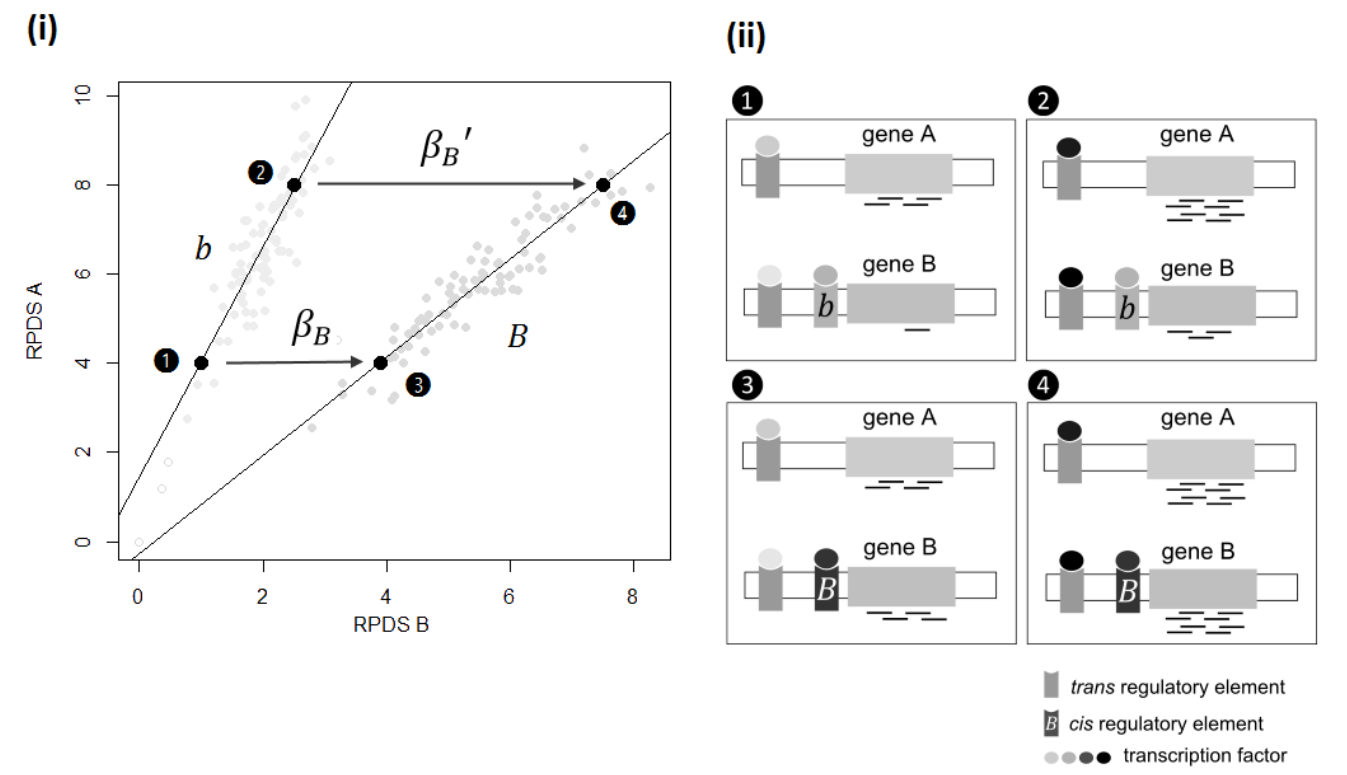


Figure 9: *Cis* regulations and interactions in the $Y^A_{ijk} = f(Y^B_{ijk})$ regression in the pair of homoelogs Traes_7AL_CBE99239D (gene A) and Traes_7BL_8849C9100 (gene B).

(i) $Y^A_{ijk} = f(Y^B_{ijk})$ regression
x-axis: RPDS B: expression level of the gene belonging to the genome B, expressed in RPDS
y-axis: RPDS A, the expression level of the gene belonging to the genome A, expressed in RPDS
Each grey point represent one individual. Black points represent situations schematized in (ii). Empty circles: missing data. Each regression line corresponds to a B allele (b or B), involved in a *cis* regulation. Moving along a line corresponds to changes in *trans* regulations, shared by the two homoelogous genes.
$\boldsymbol{\beta_B}$ = cis$B$ effect in an inhibitory genetic background
$\boldsymbol{\beta_B}'$ = cis$B$ effect in an enhancer genetic background

(ii) Schematization of four situations
❶ Allele b and inhibitory *trans* regulations
❷ Allele b and enhancer *trans* regulations
❸ Allele B and inhibitory *trans* regulations
❹ Allele B and enhancer *trans* regulations

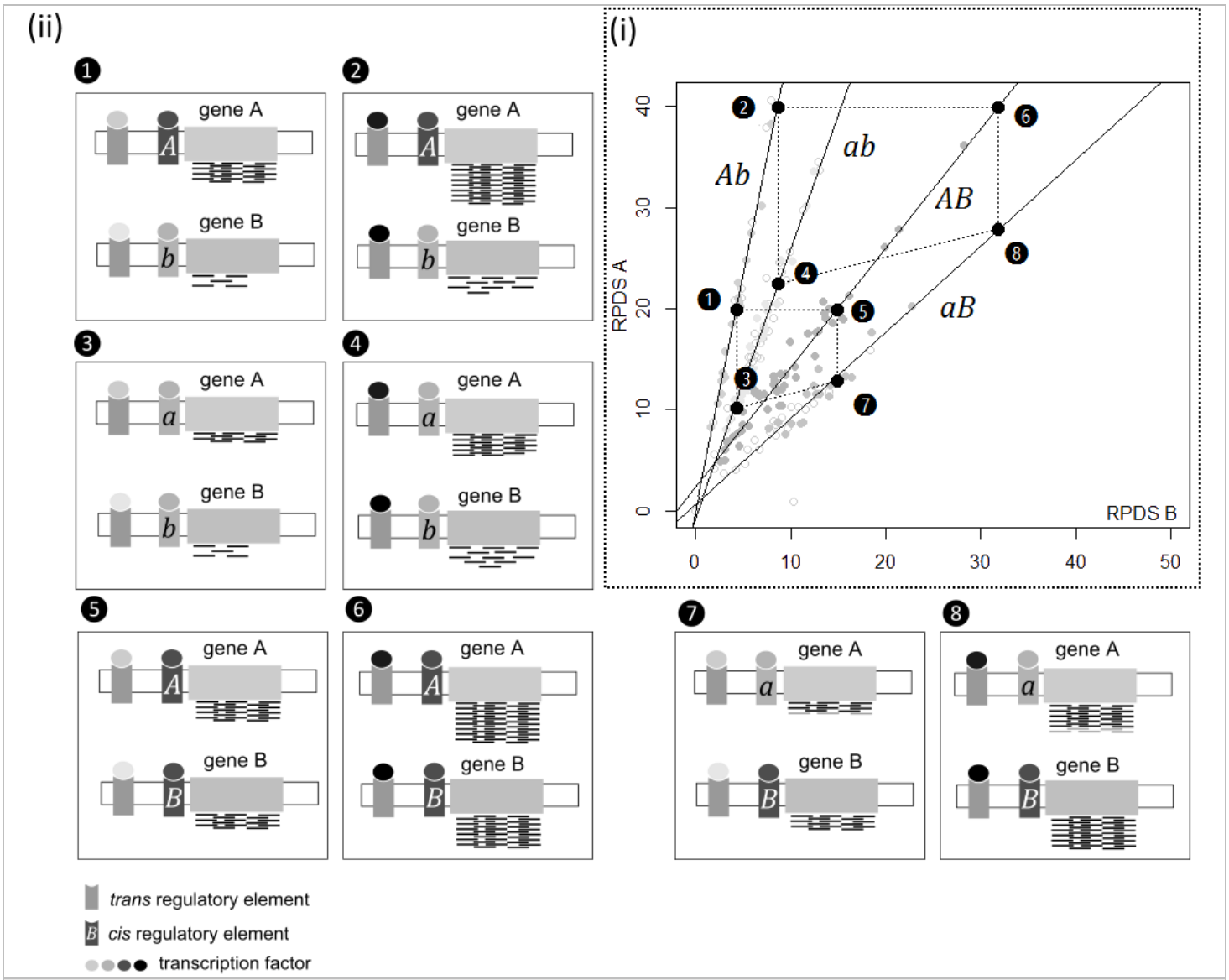


Figure 10: *Cis* regulations and interactions in the $Y^{A}_{ijk} = f(Y^{B}_{ijk})$ regression in the pair of homoelogs Traes_6AS_236AF32FD (gene A) and Traes_6BS_804DA0429 (gene B).

(i) $Y^{A}_{ijk} = f(Y^{B}_{ijk})$ regression
x-axis: RPDS B: expression level of the gene belonging to the genome B, expressed in RPDS
y-axis: RPDS A, the expression level of the gene belonging to the genome A, expressed in RPDS
Each regression line corresponds to a genotype (ab, Ab, aB or AB). Moving along a line corresponds to changes in *trans* regulations, shared by the two homoelogous genes. Black points represent situations schematized in (ii).

(ii) Schematization of eight situations
❶ Genotype Ab and inhibitory *trans* regulations
❷ Genotype Ab and enhancer *trans* regulations
❸ Genotype ab and inhibitory *trans* regulations
❹ Genotype ab and enhancer *trans* regulations
❺ Genotype AB and inhibitory *trans* regulations
❻ Genotype AB and enhancer *trans* regulations
❼ Genotype aB and inhibitory *trans* regulations
❽ Genotype aB and enhancer *trans* regulations

In ❼ and ❽, Y^{A}_{ijk} is higher than expected, as ❼ and ❽ are not aligned with ❸ and ❹ on the graph of the $Y^{A}_{ijk} = f(Y^{B}_{ijk})$ regression in (i). The proportion of mRNA additionally transcribed appears in gray in schemas ❼ and ❽.

In addition, even in the model **(2)**, the effect of the single SNP A on the expression level of gene A was not significant (p-value-adjusted = 1). This could be due to *trans* effects from the entire genome, which blurred the signal. Moreover, it seems that there is also an interaction effect between the two considered SNPs, since the expression level of the gene A is stronger than expected, when the genotype is aB. Finally, as before, the change in the slop of the regression indicates interactions between the bilocus genotypes and the genetic background.

III.2.3.4 Regulation among homoelogous pairs of genes

Among the 548 pairs of polymorphic homoelogous genes, 414 (75.5%) had an expression level bias, i.e., one gene was significantly more expressed than its homoelog among the 175 individuals, according to the Wilcoxon signed-rank test. There were 218 (52.7%) pairs where this bias was in favour of the A genome and 196 (47.3%) where it was in favour of the B genome. This distribution did not significantly deviate from the expected one of 50/50 (χ^2= 1.17, df = 1, adjusted p-value = 0.28).

Table 4: number and proportion of genes subjected to regulatory and interaction effects, for each category of expression bias (expression bias in favour of the A genome, expression bias in favour of the B genome and no expression bias)

	Expression bias in favour of the A genome	Expression bias in favour of the B genome	No expression bias
cis$A$	26/218 (11.93%)[a]	28/196 (14.29%)[a]	17/134 (12.69%)[a]
cis$B$	66/218 (30.28%)[a]	16/196 (8.16%)[c]	24/134 (17.91%)[b]
trans$A$→B	5/218 (2.29%)[a]	3/196 (1.53%)[a]	2/134 (1.49%)[a]
trans$B$→A	15/218 (6.88%)[a]	0/196 (0%)[b]	2/134 (1.49%)[a],[b]
epi$A$	1/218 (0.46%)[a]	0/196 (0%)[a]	0/134 (0%)[a]
epi$B$	1/218 (0.46%)[a]	0/196 (0%)[a]	0/134 (0%)[a]
Interaction effect with a SNP A on $Y_{ijk}^{A} = f(Y_{ijk}^{B})$	3/218 (1.38%)[a]	6/196 (3.06%)[a]	4/134 (2.99%)[a]
Interaction effect with a SNP B on $Y_{ijk}^{A} = f(Y_{ijk}^{B})$	16/218 (7.34%)[a]	2/196 (1.02%)[b]	9/134 (6.72%)[a]
Cross-allelic interaction effects (epistasis) on $Y_{ijk}^{A} = f(Y_{ijk}^{B})$	5/218 (2.29%)[a]	1/196 (0.51%)[a]	0/134 (0%)[a]

Different letters indicate significant differences (p-value ≤ 0.05) after Holm correction, within the same type of regulatory or interaction effect.
cis$A$, cis$B$, trans$A$→B, trans$B$→A, epi$A$ and epi$B$ effects are estimated with the model **(1)**.
The three last rows refer to interaction effects on $\boldsymbol{Y_{ijk}^{A} = f(Y_{ijk}^{B})}$ regression, which were estimated with the model **(2)**.

Distribution of genes subjected to regulatory effects among these pairs was examined and is shown in **Table 4**. Our results demonstrated that there were significantly more genes subjected to significant cis$B$ effects among pairs of homoelogs where the bias was in favour of the A genome, compared to pairs where the bias was in favour of the B genome (χ^2= 30.40, df = 1, adjusted p-value = 1.06e-07) and compared to pairs where there was no significant expression bias (χ^2= 6.033, df = 1, adjusted p-value

= 0.025). Genes subjected to trans$B \rightarrow A$ effects were also significantly more numerous when the expression bias was in favour of the A gene, than when it was in favour of B one (χ^2= 12.09, df = 1, adjusted p-value = 0.0015).

It was equally found there were significantly more SNPs of the genome B which significantly interact with the to $Y_{ijk}^{A} = f(Y_{ijk}^{B})$ regression, in pairs where the expression bias is in favour of the A gene, compared to pairs with an expression bias in favour of the B genome (χ^2= 8.44, df = 1, p-value = 0.010). We also examined whether certain types of effects were stronger when the expression bias was in favour of the genome A, of the B genome, or when there was no expression bias. No results were significant.

III.2.4. Discussion

III.2.4.1. Originality of this work

To our knowledge, this work is the first eQTL study on a polyploid species and the first aiming to detect epistasis on transcriptomic data of a plant species. In addition, the plant material was a sample from a broad genetic base population of durum wheat, whereas to date, most eQTL studies on plant species were realized on biparental populations, using standard interval mapping or composite interval mapping methods of detection of QTL (Drost et al., 2010; Moscou et al., 2011; Potokina et al., 2008; Schadt et al., 2003; Shi et al., 2007; Wang et al., 2010; West et al., 2007). However, few used more diversified materials (Gan et al., 2011; Zhang et al., 2011). The latter used models similar to ours, except it was unilocus: a linear regression of expression level on the genotype of the considered SNP, with eventually corrections for confounding sources of variation (e.g., kinship).

Our study allowed to examine the specificity of each of the two genomes of *Triticum turgidum*. It also enabled to establish whether certain types of expression regulations were more common in pairs of homoelogous genes, compared to pairs of non-homoelogous genes. In this study, more than 62 000 genes were identified, including approximatively 18 000 polymorphic ones. On these genes, approximately 85,000 SNPs were identified. Approximatively 12 000 SNPs, spread over 5000 genes, were selected for the detection of eQTL.

III.2.4.2. Technical discussion

III.2.4.2.1 RNA-seq

RNAseq was used in this study. It is a very useful and accurate method for eQTL detection (Majewski and Pastinen, 2011; Wang et al., 2009), but which may be subject to technical artefacts. For example, some may occur during germination of plants: variable germination dates can induce expression variations due to environmental factors (temperature, humidity, etc.). Human technical errors can as

well cause variations between cDNA libraries, especially since RNA is easily degradable. In addition, there may be variations by lane during sequencing itself. However, normalizing the expression level by the total number of reads for each individual should help to overcome these environmentally induced variations in expression levels. Finally, errors may occur during the bioinformatic analysis: two genes may be confused or, conversely, two alleles of a single gene may be perceived as being two different genes. Another disadvantage of RNAseq, in addition to the risk of errors, is that it may be difficult to distinguish weakly expressed genes from unexpressed genes. Moreover, it is impossible to determine if a value of 0 reads corresponds to missing data or to the absence expression. However, repeatability of the expression data was examined and was globally consistent (data not shown).

In addition we used a new measurement of the expression level in homoelogs: Reads Per Divergent Sites (RPDS), instead of the RPKM (Reads per kilobase per million mapped reads). This measure takes into account only reads on divergent sites, enabling better discrimination between the two homoelogous genes.

III.2.4.2.2. Interest of both models

Two estimation models were used to determine if a SNP (or a combination of SNPs via epistatic effect) could have a significant effect on the expression level of a gene. Both models were derived from the UWR model (Cheverud and Routman, 1995), which is one of the linear models inspired by quantitative genetics model of Fisher (Fisher, 1918) (see the presentation of these models in Part II).

The model **(1)** is directly inspired by the main idea of the classical linear models of quantitative genetics: the effect of a SNP is measured as the mean difference between its two alleles. It is derived from a functional model, the UWR model (described in II.2.2.2). The advantage of this model is that it estimates effects which do not depend on allele frequencies, and therefore which do not depend on the properties of the sample. In contrast, statistical models estimate genetic effects depending on allelic frequencies. In this study, allelic frequencies were very frequently unbalanced; therefore, if a statistical model had been used, estimated effects would have been very different from functional effects, and very dependent on allele frequencies at each SNP. With a functional model, estimated effects are supposed to reflect only the properties of the genes, and are comparable with each other.

Model **(2)** includes the fact that a regulatory effect can also change the slope of the $Y_{ijk}^{A} = f(Y_{ijk}^{B})$ regression, notably if the expression levels of both genes are correlated, which is frequently the case for homoelogous genes. For example, in the pairs of homoelogous genes Traes_6AS_236AF32FD (gene A) and Traes_6BS_804DA0429 (gene B), shown in Figure 9, the alleles of the SNP B change the intercept and the slope of the $Y_{ijk}^{A} = f(Y_{ijk}^{B})$ regression, which is indicative of the existence of a slight *trans* regulation and of interaction between the SNP B and the genetic background. However, the model **(1)**

could not detect the cis effect because it was blurred by the $Y_{ijk}^{A} = f(Y_{ijk}^{B})$ regression (generated by *trans* effects altering expression levels of both genes) and by the *cis* effect caused by the SNP B (cis$B$ effect). Conversely, the model **(2)** can lack of power and miss some regulatory effects detected by the model **(1)**, as shown in **Figure 11**.

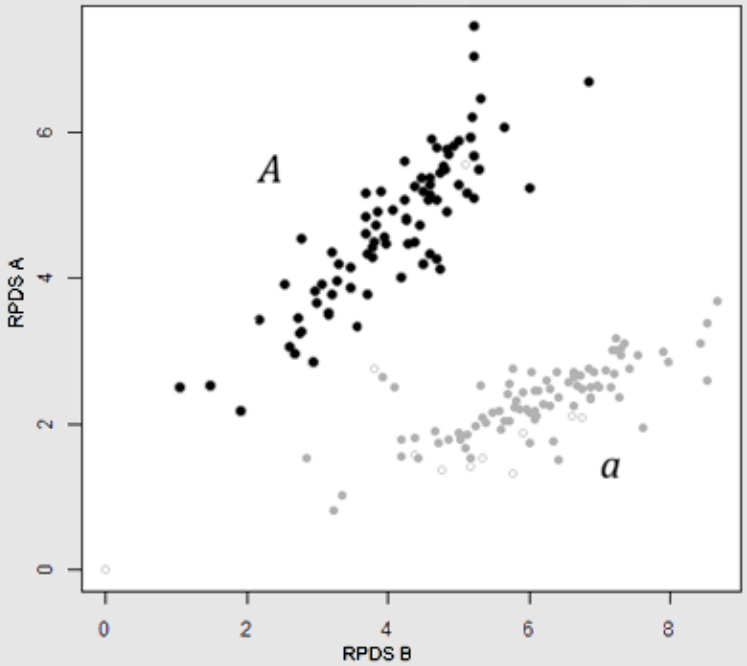


Figure 11: $Y_{ijk}^{A} = f(Y_{ijk}^{B})$ regression in the pair of homoelogs Traes_3AS_BB185F2EB (gene A) and Traes_3B_77C6EE076 (gene B).
x-axis: RPDS B: expression level of the gene belonging to the genome B, expressed in RPDS
y-axis: RPDS A, the expression level of the gene belonging to the genome A, expressed in RPDS
Each point represent one individual. Empty circles: missing data
Grey points: allele a; black points: allele A.
The model **(1)** detected a significant *cis* effect (adjusted p-value = 0). However, despite the existence of two distinct clouds, and hence of two allele-specific $Y_{ijk}^{A} = f(Y_{ijk}^{B})$ regressions, the model **(2)** did not detect any effect of the SNP A on Y_{ijk}^{A} (neither simple effect, nor interaction effect with Y_{ijk}^{B}).

III.2.4.3. Interpretation of results

III.2.4.3.1. Transcription bias

Our data seemed to show that there was a slight, but significant transcription bias in favour of the A genome, when we compared the proportion of genes of the A genome expressed in our data (48%), to the proportion of genes found in the A genome of bread wheat (Mayer et al., 2014b) (47%). Globally, it does not seem that one genome was more expressed than the other. However, there were slightly more homoelogous pairs where the expression bias was if favour of the gene A (51.5%), than pairs where the opposite was observed (48.5%). However, further evidence is needed to show that

expression of the genome A is globally favoured over that of the B genome. As far, no global expression bias was found in wheat between A and B genomes (Mayer et al., 2014b; Pfeifer et al., 2014).

III.2.4.3.2. Characteristics of eQTL

III.2.4.3.2.1. Importance of *cis* and *trans* eQTL

More *cis* eQTL than *trans* eQTL were detected in both homoelogous and non-homoelogous pairs. . This is not surprising, since all possible pairs of genes have not been studied, and many *trans* effects were thus missed. Our study did not allow us to quantify the total number of *cis* eQTL compared to the total number of possible *trans* eQTL, because it would have needed to examine all possible pairs of genes (for 5000 genes on each genome, the total number pairs would be $\binom{5000}{2} = 12\ 497\ 500$). It was not what we did. For instance, trans$A$→A and trans$B$→B effects were excluded from the study.

However, we found that *cis* effects were stronger in intensity than *trans* effects, which is commonly found in eQTL studies on plants (Hammond et al., 2011; Holloway et al., 2011; Keurentjes et al., 2007; Schadt et al., 2003; Swanson-Wagner et al., 2009; Wang et al., 2006). However, we did not find that *trans* eQTL had more extreme allelic frequencies than *cis* eQTL (He = 0.44 for SNPs involved in *cis* eQTL; He = 0.45 for SNPs involved in *trans* eQTL; p-value = 0.24), as it was found previously (Cubillos et al., 2012). *Trans* effects are thus involved in the genetic variation of the population, and therefore may play a role in its ability to adapt. However, one must be cautious while concluding, because the expression levels correspond to a very early stage.

III.2.4.3.2.2. Regulations among homoelogous pairs of genes

III.2.4.3.2.2.1. Correlation of expression levels between homoelogs

We found that many homoelogous (77.9%) had their expression levels correlated in our study. Significant correlations of expression levels between homoelogs were always positive, which indicates that the total expression level of the two homoelogs is not controlled to be maintained below a certain level. Our results suggest that homoelogous genes are globally subjected to the same *trans* regulations. This was expected because homoelogs are relatively similar, so could have been able to maintain similar transcription factor binding sites. The divergence between sequences and expressions of homoelogs increase over evolutionary time. However, it has been shown that homoelogous sequences remain relatively similar, compared to gene copies derived from other types of duplication (De Smet and Van de Peer, 2012). This can be understood in terms of the gene balance hypothesis (Birchler and Veitia, 2007), which stipulates that a change in the copy number of certain genes (particularly regulatory genes), either through duplication or deletion, affects gene dosage and can lead to a decrease in fitness. Thus, the duplication of a whole gene network, as in polyploidization, is better

tolerated than the duplication of only one single element, because it does not completely disrupt stoichiometry, i.e., the appropriate ratio of gene products involved in biological reactions.

III.2.4.3.2.2.2. More trans *effects among homoelogous pairs of genes*

Another important result is that there are more *trans* effects between homoelogous genes than between random pairs of genes. This means that a gene is more likely to be able to regulate the expression of another gene - either directly or indirectly - if it is its homoelog. This can perhaps be explained by the fact that homoelogs are more likely to belong to the same gene network and to share the same systems of regulation, inherited from their common ancestor sequence. Homoelogs may thus influence the transcription of each other.

As expression levels of homoelogous genes are positively correlated, it can be assumed that they are more frequently involved in positive feedback loops rather than negative feedback loops. Negative feedback is associated with homeostasis and positive feedback is a necessary prerequisite for multistationarity (Plahte et al., 1995; Thomas and Richelle, 1988). Multistationarity is of great importance in biological systems, allowing cells and organisms to move from one discrete state to another (Angeli et al., 2004). For example, it enables cell differentiation (Thomas and Kaufman, 2001).

If networks based on positive feedback are duplicated in an organism, it is likely that the two copies of one gene influence the expression of each other. For example, if one of the two copies has an allele which greatly increases its expression (*cis* effect), it will also indirectly increase the expression of its homoelogs (*trans* effect). However, note that following an event of polyploidization, gene networks are gradually reorganized (De Smet and Van de Peer, 2012), and the reciprocal influence that could have some homoelogs on each other may have been be lost.

Another difference between homoelogous pairs and non-homoelogous pairs, is that more significant *cis* effects have been detected on non-homoelogous pairs. Moreover, the effects were significantly stronger when detected on these pairs. Among the genes constituting non-homoelogous pairs, some of them had homoelogs in the rest of the genome. Therefore these differences in intensity and proportion of significant *cis* effects were probably due to the model, not to an intrinsic biological property of genes having no homoelogous copy. The increase in the estimations of *cis* effects when a homoelogous gene is not added to the model, can be explained by an artificial correlation induced by the standardization. For example, non-standardized $\mathrm{cis}A$ and $\mathrm{trans}B \rightarrow A$ effects are positively correlated (correlation coefficient ≈ 0.60, adjusted p-value < 2.2e-16). But standardisation created negative correlation between these effects (correlation coefficient ≈ -0.65, adjusted p-value < 2.2e-16). Indeed, adding a second genetic effect will *de facto* decrease the estimation of the first one. This is due to the fact that the standardization factor, s, increases when an effect is added. An example is

given in **Table 5**. In the first case, there is only one cis$A$ effect, as in non-homoelogous pairs. When a trans$B$→A effect is added (as in homoelogous pairs), the value of the standardized cis$A$ effect decreases.

Table 5: effect of the standardization on the correlation between effects.

		case 1	case 2
expression level of gene A (Y_{ijk}^{A})	AB	1	2
	Ab	1	1
	aB	0	1
	ab	0	0
Estimation of effects	mu	0.5	1
	cis$A$	1	1
	trans$B$→A	0	1
	epi$A$	0	0
Standardisation factor	s	1	1.41
Standardized effects	standardized cis$A$	1	0.71
	standardized trans$B$→A	0	0.71
	Standardized epi$A$	0	0.00

In case 1, only the SNP A has an effect on the expression on Y_{ijk}^{A} (cis$A$ effect). This is representative of non-homoelogous pairs of genes. In the case 2, a trans$B$→A effect is added, as in homoelogous pairs of genes where *trans* effects are stronger. In this case, the value of standardized cis$A$ decreases.

In conclusion, *cis* effects may be over-estimated, and false positives may be detected when the genotype of the homoelog is not added to the model.

III.2.4.3.2.3. Low detection of epistasis

A surprising result is that very few epistatic eQTL were detected. We expected to find more of them, since tetraploid wheat is a self-pollinating species. Indeed, some authors have argued that inbreeding allows selection to act more directly on gene combinations, facilitating the evolution of co-adapted gene complexes (Allard et al., 1993; Fenster et al., 1997; Nordborg et al., 2002; Pollak and Sabran, 1999)**.** Thus, selfing may permit the development and maintenance of combinations of alleles of highest selective value, and thus of positive or sign epistasis. In addition, genetic composition of predominantly selfing populations strongly supports the hypothesis of the evolution of co-adapted gene complexes. Although epistasis is not always found in selfing species, it seems to be more commonly detected in autogamous plants than in outcrossing species (Holland, J. B., 2001). Outbreeding depression, the phenomenon which results from the breakdown by recombination of co-adapted gene pools, was observed in several autogamous species (Edmands and Timmerman, 2003; Frankham et al., 2011; Lynch, 1991). This was particularly studied in barley (Volis et al., 2011). In addition, disomic polyploidy could be a way to compensate the loss of allelic interactions between

homologous loci by allelic interactions between homoelogous loci in selfing plants (Mac Key, 1970). A consequence of this would be that epistasis may be less important in polysomic tetraploid such as alfalfa than in disomic autogamous plants, such as tetraploid wheat, where homoelogous chromosomes are distinguished during meiosis. Studies on the selfing tetraploid grass *Avena barbata,* showed that polyploidization allowed fixing heterosis (Allard et al., 1993; García et al., 1991). Indeed *Avena barbata* is an autotetraploid disomic plant, derived from two ecotypes of single a biological species, the diploid *Avena hirtula.* Allelic interactions between homologous loci (dominance) were converted into interactions between homoelogous loci (epistasis), which protects the favorable allelic complex from breakup by segregation. Such co-adapted gene pools, involving homoelogs, were thus expected in a selfing population, as it could play an important role in the compensation of the loss of heterozygosity by epistatic interactions. Epistatic QTL were found in durum wheat for numerous phenotypic traits such as yield-associated characters (Maccaferri et al., 2008; Patil et al., 2013; Sharma and Sain, 2004), disease resistance (Bnejdi and Gazzah, 2008; Haile et al., 2012), morphological characters (Maccaferri et al., 2008; Sharma et al., 2003), or grain quality (Conti et al., 2011), which indicates epistasis can be observed in this species. But expression level does not seem to be a character subjected to epistasis. However, it is possible that a transcription level subjected to *cis* and *trans* effects (two additive effects) may correspond to an epistatic trait at the phenotypic level. In other words, epistatic effects may not be detectable at the elementary level, within a pair of homoelogs, but may appear on complex and integrative character, measurable at the phenotypic level.

III.2.4.3.2.4. Asymmetric roles of the A and B genomes

In both models, there have always been more significant eQTL detected on the B genome than on the A genome. This is true for both homoelogous non-homoelogous pairs of genes, and this concerns both SNPs involved in *cis* or *trans* significant effects and SNPs altering the $Y_{ijk}^{A} = f(Y_{ijk}^{B})$ regression. In addition, regulatory *cis* and *trans* effects were as well stronger when coming from the B genome.

The polymorphism levels (He) of SNPs of the genome B, used to detect regulatory effects, were slightly higher than that of the SNPs of the A genome (He = 0.42 for B and 0.41 for A, in non-homoelogous pairs of genes; He = 0.41 for B and 0.40 for A, in homoelogous pairs of genes). This difference is not strong enough to explain a better detection of eQTL on the B genome than on the A genome A. Therefore, the difference in number of detected eQTL is not due to the statistical model.

There may be a lower density of genes producing regulatory proteins on the A genome, and a lower density of *cis* element.

More probably, it may also be due to the fact that cis$B$ and trans$B$→A effects have stronger amplitudes than cis$A$ and trans$A$→B effects. Therefore, the former could be more easily detected than

the latter. This difference in amplitude may be due to artefacts: for example, SNPs B may be physically closer to causal polymorphisms than SNPs A. The linkage disequilibrium may be also stronger on the B genome, making the effects SNPs B statistically more correlated to causal locus, than those of SNPs A. Finally, this may be due to real biological effects: the eQTL of the B genome may have stronger regulatory effects than the eQTL of the A genome.

The current *Triticum urartu* is a strongly selfing species while *Aegilops speltoides* is an auto-incompatible species. One could then reasonably speculates that the ancestor of *T. urartu* involved in the paternal genome of the first spontaneous *T. dicoccoides* hybrids was already a selfing species. Selfing species may be better in controlling the abundance and activity of repetitive genetic elements (Wright et al., 2008). Therefore, repetitive selfish elements may be less frequently inserted in regulatory sequences of selfing species. In addition, repetitive selfish element have been often described for their capacity of regulating genes while inserting in regulatory sequences. Consequently, *cis* effects due to repetitive elements in A genome may be less frequent than in B genome. If this could partly explain the larger amplitude of $\mathrm{cis}B$ effects, compared to $\mathrm{cis}A$ effects, this hypothesis cannot elucidate the larger amplitude of $\mathrm{trans}B \rightarrow A$ effects, compared to $\mathrm{trans}A \rightarrow B$ effects.

Another possible explanation could be linked to the genetic load. In outbreeding species, slight deleterious recessive mutations may accumulate in population, masked at the heterozygous state by a dominant and functional allele, and remain in the population as crypto polymorphism. When their frequency increases by chance due to genetic drift, or by the development of some consanguinity in the population, homozygous carriers appear and are counter selected. This selection-mutation-drift equilibrium is one of the major explanation of the strong inbreeding depression observed in outbreeding species having large population genetic size. When these deleterious mutations alter the efficiency of enzymes or the function of a gene, an increase in the quantity of the enzyme may permit the carrier individual to achieve the gene function with the same efficiency than the wild allele. Thus, a mutation in the *cis* regulating region of a deleterious allele may compensate the functional deficiency of the allele. As the two mutants (the *cis* regulating one and the functionally one) are closely linked together, they are transmitted as only one haplotype. Of course, the new up-regulating mutant may also have an effect on the regulation of the wild allele and disturb the function of this allele. In reaction, a down regulator mutant could be selected in tight linkage with the wild allele, in order to repress the impact of the up regulating mutant in a heterozygous situation. This race of arms leads to a conflict between the combinations of haplotypes. This conflict may thus result in a build-up of polymorphisms in regulating areas, with more and more contrasted regulation amplitudes when deleterious mutations accumulate. In selfing species, this conflict does not occur, since deleterious alleles associated with

their up-regulators are homozygous and are therefore eliminated by selection. This idea has to be clarified and tested through modelling (T. Lenormand and S. Glémin, personal communication).

Finally, more *cis* and *trans* eQTL of the B genome were detected in pairs of homoelogs where there was an expression bias in favour of the A genome, compared to homoelogous pairs where there was a bias in favour of the B genome, or no bias. This was a very pronounced trend, since the proportion of significant cis$B$ effects was almost 4 times higher among pairs of homoelogous where the bias was in favour of A, compared to pairs where it was in favour of B. In addition, the proportion of significant trans$B$→A effects were equal to 6.88% for the first, against 0% for the latter. However, we could not demonstrate that was due to a weaker selection when the bias was in favour of the A genome.

III.2.4.4. Perspectives

This study was performed on only one stage (sprout) and only for aboveground organs (Coleoptiles and primary leaves). It needs to be repeated in other organs, in other stages, and in more environmental conditions (stress/non-stress) so that these results could be confirmed. It would be also worthwhile to estimate the trans$A$→A and trans$B$→B effects, to compare their number and amplitude to inter-genome *trans* effects. Finally, it would also be interesting to calculate genetic variances and covariances with the G2A-D model (see section II.4.2.3) , to estimate the contributions (see section II.4.3) of additive and epistatic effects to genetic variance.

III.2.5. References

Adams, K.L., Cronn, R., Percifield, R., and Wendel, J.F. (2003). Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. Proc. Natl. Acad. Sci. 100, 4649–4654.

Adams, K.L., Percifield, R., and Wendel, J.F. (2004). Organ-Specific Silencing of Duplicated Genes in a Newly Synthesized Cotton Allotetraploid. Genetics 168, 2217–2226.

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

Allard, R.W., García, P., Sáenz-de-Miera, L.E., and Vega, M.P. de la (1993). Evolution of multilocus genetic structure in Avena hirtula and Avena barbata. Genetics 135, 1125–1139.

Angeli, D., Ferrell, J.E., and Sontag, E.D. (2004). Detection of multistability, bifurcations, and hysteresis in a large class of biological positive-feedback systems. Proc. Natl. Acad. Sci. U. S. A. 101, 1822–1827.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene Ontology: tool for the unification of biology. Nat. Genet. 25, 25–29.

Becker, J., Wendland, J.R., Haenisch, B., Nöthen, M.M., and Schumacher, J. (2012). A systematic eQTL study of cis–trans epistasis in 210 HapMap individuals. Eur. J. Hum. Genet. 20, 97–101.

Birchler, J.A., and Veitia, R.A. (2007). The Gene Balance Hypothesis: From Classical Genetics to Modern Genomics. Plant Cell Online 19, 395–402.

Bnejdi, F., and Gazzah, M.E. (2008). Inheritance of resistance to yellowberry in durum wheat. Euphytica 163, 225–230.

Brem, R.B., and Kruglyak, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proc. Natl. Acad. Sci. U. S. A. 102, 1572–1577.

Brem, R.B., Storey, J.D., Whittle, J., and Kruglyak, L. (2005). Genetic interactions between polymorphisms that affect gene expression in yeast. Nature 436, 701–703.

Brenchley, R., Spannagl, M., Pfeifer, M., Barker, G.L.A., D'Amore, R., Allen, A.M., McKenzie, N., Kramer, M., Kerhornou, A., Bolser, D., et al. (2012). Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature 491, 705–710.

Buggs, R.J.A., Chamala, S., Wu, W., Gao, L., May, G.D., Schnable, P.S., Soltis, D.E., Soltis, P.S., and Barbazuk, W.B. (2010). Characterization of duplicate gene evolution in the recent natural allopolyploid Tragopogon miscellus by next-generation sequencing and Sequenom iPLEX MassARRAY genotyping. Mol. Ecol. 19 Suppl 1, 132–146.

Chagué, V., Just, J., Mestiri, I., Balzergue, S., Tanguy, A.-M., Huneau, C., Huteau, V., Belcram, H., Coriton, O., Jahier, J., et al. (2010). Genome-wide gene expression changes in genetically stable synthetic and natural wheat allohexaploids. New Phytol. 187, 1181–1194.

Chang, P.L., Dilkes, B.P., McMahon, M., Comai, L., and Nuzhdin, S.V. (2010). Homoeolog-specific retention and use in allotetraploid Arabidopsis suecica depends on parent of origin and network partners. Genome Biol. 11, R125.

Charmet, G. (2011). Wheat domestication: Lessons for the future. C. R. Biol. 334, 212–220.

Chaudhary, B., Flagel, L., Stupar, R.M., Udall, J.A., Verma, N., Springer, N.M., and Wendel, J.F. (2009). Reciprocal Silencing, Transcriptional Bias and Functional Divergence of Homeologs in Polyploid Cotton (Gossypium). Genetics 182, 503–517.

Chen, Z.J. (2007). Genetic and Epigenetic Mechanisms for Gene Expression and Phenotypic Variation in Plant Polyploids. Annu. Rev. Plant Biol. 58, 377–406.

Chen, Z.J., and Ni, Z. (2006). Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. BioEssays 28, 240–252.

Chen, Z.J., and Pikaard, C.S. (1997). Epigenetic silencing of RNA polymerase I transcription: a role for DNA methylation and histone modification in nucleolar dominance. Genes Dev. 11, 2124–2136.

Chen, X., Hackett, C.A., Niks, R.E., Hedley, P.E., Booth, C., Druka, A., Marcel, T.C., Vels, A., Bayer, M., Milne, I., et al. (2010). An eQTL Analysis of Partial Resistance to Puccinia hordei in Barley. PLoS ONE 5, e8598.

Cheng, F., Wu, J., Fang, L., Sun, S., Liu, B., Lin, K., Bonnema, G., and Wang, X. (2012). Biased Gene Fractionation and Dominant Gene Expression among the Subgenomes of Brassica rapa. PLoS ONE 7, e36442.

Cheung, V.G., Spielman, R.S., Ewens, K.G., Weber, T.M., Morley, M., and Burdick, J.T. (2005). Mapping determinants of human gene expression by regional and genome-wide association. Nature 437, 1365–1369.

Cheverud, J.M., and Routman, E.J. (1995). Epistasis and its contribution to genetic variance components. Genetics 139, 1455–1461.

Comai, L. (2005). The advantages and disadvantages of being polyploid. Nat. Rev. Genet. 6, 836–846.

Comai, L., Tyagi, A.P., Winter, K., Holmes-Davis, R., Reynolds, S.H., Stevens, Y., and Byers, B. (2000). Phenotypic Instability and Rapid Gene Silencing in Newly Formed Arabidopsis Allotetraploids. Plant Cell Online 12, 1551–1567.

Conti, V., Roncallo, P.F., Beaufort, V., Cervigni, G.L., Miranda, R., Jensen, C.A., and Echenique, V.3. (2011). Mapping of main and epistatic effect QTLs associated to grain protein and gluten strength using a RIL population of durum wheat. J. Appl. Genet. 52, 287–298.

Cubillos, F.A., Coustham, V., and Loudet, O. (2012). Lessons from eQTL mapping studies: non-coding regions and their role behind natural phenotypic variation in plants. Curr. Opin. Plant Biol. 15, 192–198.

Cui, L., Wall, P.K., Leebens-Mack, J.H., Lindsay, B.G., Soltis, D.E., Doyle, J.J., Soltis, P.S., Carlson, J.E., Arumuganathan, K., Barakat, A., et al. (2006). Widespread genome duplications throughout the history of flowering plants. Genome Res. 16, 738–749.

Daud, H.M., and Gustafson, J.P. (1996). Molecular evidence for Triticum speltoides as a B-genome progenitor of wheat (Triticum aestivum). Genome 39, 543–548.

David, J., Holtz, Y., Ranwez, V., Santoni, S., Sarah, G., Ardisson, M., Poux, G., Choulet, F., Genthon, C., Roumet, P., et al. (in press). Genotyping by sequencing transcriptomes in an evolutionary pre-breeding durum wheat population. Mol. Breed.

DeCook, R., Lall, S., Nettleton, D., and Howell, S.H. (2006). Genetic Regulation of Gene Expression During Shoot Development in Arabidopsis. Genetics 172, 1155–1164.

Drost, D.R., Benedict, C.I., Berg, A., Novaes, E., Novaes, C.R.D.B., Yu, Q., Dervinis, C., Maia, J.M., Yap, J., Miles, B., et al. (2010). Diversification in the genetic architecture of gene expression and transcriptional networks in organ differentiation of Populus. Proc. Natl. Acad. Sci. 107, 8492–8497.

Dubcovsky, J., and Dvorak, J. (2007). Genome Plasticity a Key Factor in the Success of Polyploid Wheat Under Domestication. Science 316, 1862–1866.

Dvořák, J., McGuire, P.E., and Cassidy, B. (1988). Apparent sources of the A genomes of wheats inferred from polymorphism in abundance and restriction fragment length of repeated nucleotide sequences. Genome 30, 680–689.

Dvořák, J., Terlizzi, P. di, Zhang, H.-B., and Resta, P. (1993). The evolution of polyploid wheats: identification of the A genome donor species. Genome 36, 21–31.

Edmands, S., and Timmerman, C.C. (2003). Factores de Modelaje que Afectan la Severidad de la Depresión Exogámica. Conserv. Biol. 17, 883–892.

Escobar, J.S., Cenci, A., Bolognini, J., Haudry, A., Laurent, S., David, J., and Glémin, S. (2010). An integrative test of the dead-end hypothesis of selfing evolution in Triticeae (Poaceae). Evol. Int. J. Org. Evol. 64, 2855–2872.

Fawcett, J.A., Maere, S., and Peer, Y.V. de (2009). Plants with double genomes might have had a better chance to survive the Cretaceous–Tertiary extinction event. Proc. Natl. Acad. Sci. 106, 5737–5742.

Feldman, M., and Levy, A.A. (2009). Genome evolution in allopolyploid wheat—a revolutionary reprogramming followed by gradual changes. J. Genet. Genomics 36, 511–518.

Feldman, M., Liu, B., Segal, G., Abbo, S., Levy, A.A., and Vega, J.M. (1997). Rapid Elimination of Low-Copy DNA Sequences in Polyploid Wheat: A Possible Mechanism for Differentiation of Homoeologous Chromosomes. Genetics 147, 1381–1387.

Fenster, C.B., Galloway, L.F., and Chao, L. (1997). Epistasis and its consequences for the evolution of natural populations. Trends Ecol. Evol. 12, 282–286.

Fisher, R.A. (1918). The Correlation between Relatives on the Supposition of Mendelian Inheritance. Trans. R. Soc. Edinb. 52, 399–433.

Flagel, L.E., and Wendel, J.F. (2010). Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. New Phytol. 186, 184–193.

Frankham, R., Ballou, J.D., Eldridge, M.D.B., Lacy, R.C., Ralls, K., Dudash, M.R., and Fenster, C.B. (2011). Predicción de la Depresión por Exogámica. Conserv. Biol. 25, 465–475.

Freeling, M., Woodhouse, M.R., Subramaniam, S., Turco, G., Lisch, D., and Schnable, J.C. (2012). Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. Curr. Opin. Plant Biol. 15, 131–139.

Gaeta, R.T., Pires, J.C., Iniguez-Luy, F., Leon, E., and Osborn, T.C. (2007). Genomic Changes in Resynthesized Brassica napus and Their Effect on Gene Expression and Phenotype. Plant Cell Online 19, 3403–3417.

Gan, X., Stegle, O., Behr, J., Steffen, J.G., Drewe, P., Hildebrand, K.L., Lyngsoe, R., Schultheiss, S.J., Osborne, E.J., Sreedharan, V.T., et al. (2011). Multiple reference genomes and transcriptomes for Arabidopsis thaliana. Nature 477, 419–423.

García, P., Morris, M.I., Sáenz-de-Miera, L.E., Allard, R.W., Vega, M.P. de la, and Ladizinsky, G. (1991). Genetic diversity and adaptedness in tetraploid Avena barbata and its diploid ancestors Avena hirtula and Avena wiestii. Proc. Natl. Acad. Sci. 88, 1207–1211.

Garsmeur, O., Schnable, J.C., Almeida, A., Jourda, C., D'Hont, A., and Freeling, M. (2014). Two Evolutionarily Distinct Classes of Paleopolyploidy. Mol. Biol. Evol. 31, 448–454.

Gayral, P., Melo-Ferreira, J., Glémin, S., Bierne, N., Carneiro, M., Nabholz, B., Lourenco, J.M., Alves, P.C., Ballenghien, M., Faivre, N., et al. (2013). Reference-Free Population Genomics from Next-Generation Transcriptome Data and the Vertebrate–Invertebrate Gap. PLoS Genet 9, e1003457.

Gill, D.B.S., and Waines, J.G. (1978). Paternal regulation of seed development in wheat hybrids. Theor. Appl. Genet. 51, 265–270.

Gjuvsland, A.B., Hayes, B.J., Omholt, S.W., and Carlborg, O. (2007). Statistical Epistasis Is a Generic Feature of Gene Regulatory Networks. Genetics 175, 411–420.

Grover, C.E., Gallagher, J.P., Szadkowski, E.P., Yoo, M.J., Flagel, L.E., and Wendel, J.F. (2012a). Homoeolog expression bias and expression level dominance in allopolyploids. New Phytol. 196, 966–971.

Grover, C.E., Gallagher, J.P., Szadkowski, E.P., Yoo, M.J., Flagel, L.E., and Wendel, J.F. (2012b). Homoeolog expression bias and expression level dominance in allopolyploids. New Phytol. 196, 966–971.

Haile, J.K., Nachit, M.M., Hammer, K., Badebo, A., and Röder, M.S. (2012). QTL mapping of resistance to race Ug99 of Puccinia graminis f. sp. tritici in durum wheat (Triticum durum Desf.). Mol. Breed. 30, 1479–1493.

Hammond, J.P., Mayes, S., Bowen, H.C., Graham, N.S., Hayden, R.M., Love, C.G., Spracklen, W.P., Wang, J., Welham, S.J., White, P.J., et al. (2011). Regulatory Hotspots Are Associated with Plant Gene Expression under Varying Soil Phosphorus Supply in Brassica rapa. Plant Physiol. 156, 1230–1241.

He, P., Friebe, B.R., Gill, B.S., and Zhou, J.-M. (2003). Allopolyploidy alters gene expression in the highly stable hexaploid wheat. Plant Mol. Biol. 52, 401–414.

Hemani, G., Shakhbazov, K., Westra, H.-J., Esko, T., Henders, A.K., McRae, A.F., Yang, J., Gibson, G., Martin, N.G., Metspalu, A., et al. (2014). Detection and replication of epistasis influencing transcription in humans. Nature 508, 249–253.

Hirschhorn, J.N., and Daly, M.J. (2005). Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. 6, 95–108.

Holland, J. B. (2001). Epistasis and plant breeding. In Plant Breeding Reviews, pp. 27–92.

Holloway, B., Luck, S., Beatty, M., Rafalski, J.-A., and Li, B. (2011). Genome-wide expression quantitative trait loci (eQTL) analysis in maize. BMC Genomics 12, 336.

Houchins, K., O'Dell, M., Flavell, R.B., and Gustafson, J.P. (1997). Cytosine methylation and nucleolar dominance in cereal hybrids. Mol. Gen. Genet. MGG 255, 294–301.

Huang, S., Sirikhachornkit, A., Su, X., Faris, J., Gill, B., Haselkorn, R., and Gornicki, P. (2002). Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the Triticum/Aegilops complex and the evolutionary history of polyploid wheat. Proc. Natl. Acad. Sci. 99, 8133–8138.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., et al. (2002). The Ensembl genome database project. Nucleic Acids Res. 30, 38–41.

Jiao, Y., Wickett, N.J., Ayyampalayam, S., Chanderbali, A.S., Landherr, L., Ralph, P.E., Tomsho, L.P., Hu, Y., Liang, H., Soltis, P.S., et al. (2011). Ancestral polyploidy in seed plants and angiosperms. Nature 473, 97–100.

Jorjani, H., Klei, L., and Emanuelson, U. (2003). A Simple Method for Weighted Bending of Genetic (Co)variance Matrices. J. Dairy Sci. 86, 677–679.

Kashkush, K., Feldman, M., and Levy, A.A. (2002). Gene Loss, Silencing and Activation in a Newly Synthesized Wheat Allotetraploid. Genetics 160, 1651–1659.

Keurentjes, J.J.B., Fu, J., Terpstra, I.R., Garcia, J.M., Ackerveken, G. van den, Snoek, L.B., Peeters, A.J.M., Vreugdenhil, D., Koornneef, M., and Jansen, R.C. (2007). Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. Proc. Natl. Acad. Sci. 104, 1708–1713.

Kilian, B., Özkan, H., Deusch, O., Effgen, S., Brandolini, A., Kohl, J., Martin, W., and Salamini, F. (2007). Independent Wheat B and G Genome Origins in Outcrossing Aegilops Progenitor Haplotypes. Mol. Biol. Evol. 24, 217–227.

Kliebenstein, D. (2009). Quantitative Genomics: Analyzing Intraspecific Variation Using Global Gene Expression Polymorphisms or eQTLs. Annu. Rev. Plant Biol. 60, 93–114.

Kotseruba, V., Gernand, D., Meister, A., and Houben, A. (2003). Uniparental loss of ribosomal DNA in the allotetraploid grass Zingeria trichopoda (2n = 8). Genome 46, 156–163.

Kovarik, A., Matyasek, R., Lim, K.Y., Skalická, K., Koukalová, B., Knapp, S., Chase, M., and Leitch, A.R. (2004). Concerted evolution of 18–5.8–26S rDNA repeats in Nicotiana allotetraploids. Biol. J. Linn. Soc. 82, 615–625.

Kovarik, A., Pires, J.C., Leitch, A.R., Lim, K.Y., Sherwood, A.M., Matyasek, R., Rocca, J., Soltis, D.E., and Soltis, P.S. (2005). Rapid Concerted Evolution of Nuclear Ribosomal DNA in Two Tragopogon Allopolyploids of Recent and Recurrent Origin. Genetics 169, 931–944.

Legarra, A., and Fernando, R.L. (2009). Linear models for joint association and linkage QTL mapping. Genet. Sel. Evol. 41, 43.

Lukens, L.N., Pires, J.C., Leon, E., Vogelzang, R., Oslach, L., and Osborn, T. (2006). Patterns of Sequence Loss and Cytosine Methylation within a Population of Newly Resynthesized Brassica napus Allopolyploids. Plant Physiol. 140, 336–348.

Lynch, M. (1991). The Genetic Interpretation of Inbreeding Depression and Outbreeding Depression. Evolution 45, 622.

Ma, X.-F., and Gustafson, J.P. (2006). Timing and rate of genome variation in triticale following allopolyploidization. Genome 49, 950–958.

Maccaferri, M., Sanguineti, M.C., Corneti, S., Ortega, J.L.A., Salem, M.B., Bort, J., DeAmbrogio, E., Moral, L.F.G. del, Demontis, A., El-Ahmed, A., et al. (2008). Quantitative Trait Loci for Grain Yield and Adaptation of Durum Wheat (Triticum durum Desf.) Across a Wide Range of Water Availability. Genetics 178, 489–511.

Mac Key, J. (1970). Significance of mating systems for chromosomes and gametes in polyploids. Hereditas 66, 165–176.

Madlung, A., Masuelli, R.W., Watson, B., Reynolds, S.H., Davison, J., and Comai, L. (2002). Remodeling of DNA Methylation and Phenotypic and Transcriptional Changes in Synthetic Arabidopsis Allotetraploids. Plant Physiol. 129, 733–746.

Madlung, A., Tyagi, A.P., Watson, B., Jiang, H., Kagochi, T., Doerge, R.W., Martienssen, R., and Comai, L. (2005). Genomic changes in synthetic Arabidopsis polyploids. Plant J. 41, 221–230.

Majewski, J., and Pastinen, T. (2011). The study of eQTL variations by RNA-seq: from SNPs to phenotypes. Trends Genet. 27, 72–79.

Mayer, K.F.X., Rogers, J., Doležel, J., Pozniak, C., Eversole, K., Feuillet, C., Gill, B., Friebe, B., Lukaszewski, A.J., Sourdille, P., et al. (2014a). A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science 345, 1251788.

Mayer, K.F.X., Rogers, J., Doležel, J., Pozniak, C., Eversole, K., Feuillet, C., Gill, B., Friebe, B., Lukaszewski, A.J., Sourdille, P., et al. (2014b). A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science 345, 1251788.

McClintock, B. (1984). The significance of responses of the genome to challenge. Science 226, 792–801.

Meyer, K., and Kirkpatrick, M. (2010). Better Estimates of Genetic Covariance Matrices by “Bending” Using Penalized Maximum Likelihood. Genetics 185, 1097–1110.

Michaelson, J.J., Loguercio, S., and Beyer, A. (2009). Detection and interpretation of expression quantitative trait loci (eQTL). Methods 48, 265–276.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods 5, 621–628.

Moscou, M.J., Lauter, N., Steffenson, B., and Wise, R.P. (2011). Quantitative and Qualitative Stem Rust Resistance Factors in Barley Are Associated with Transcriptional Suppression of Defense Regulons. PLoS Genet 7, e1002208.

Nei, M. (1978). Estimation of Average Heterozygosity and Genetic Distance from a Small Number of Individuals. Genetics 89, 583–590.

Nordborg, M., Borevitz, J.O., Bergelson, J., Berry, C.C., Chory, J., Hagenblad, J., Kreitman, M., Maloof, J.N., Noyes, T., Oefner, P.J., et al. (2002). The extent of linkage disequilibrium in Arabidopsis thaliana. Nat. Genet. 30, 190–193.

Osborn, T.C., Chris Pires, J., Birchler, J.A., Auger, D.L., Jeffery Chen, Z., Lee, H.-S., Comai, L., Madlung, A., Doerge, R.W., Colot, V., et al. (2003). Understanding mechanisms of novel gene expression in polyploids. Trends Genet. 19, 141–147.

Oshlack, A., and Wakefield, M.J. (2009). Transcript length bias in RNA-seq data confounds systems biology. Biol. Direct 4, 1–10.

Ozkan, H., Levy, A.A., and Feldman, M. (2001). Allopolyploidy-Induced Rapid Genome Evolution in the Wheat (Aegilops–Triticum) Group. Plant Cell Online 13, 1735–1747.

Özkan, H., Brandolini, A., Schäfer-Pregl, R., and Salamini, F. (2002). AFLP Analysis of a Collection of Tetraploid Wheats Indicates the Origin of Emmer and Hard Wheat Domestication in Southeast Turkey. Mol. Biol. Evol. 19, 1797–1801.

Patil, R.M., Tamhankar, S.A., Oak, M.D., Raut, A.L., Honrao, B.K., Rao, V.S., and Misra, S.C. (2013). Mapping of QTL for agronomic traits and kernel characters in durum wheat (Triticum durum Desf.). Euphytica 190, 117–129.

Van de Peer, Y., Fawcett, J.A., Proost, S., Sterck, L., and Vandepoele, K. (2009). The flowering world: a tale of duplications. Trends Plant Sci. 14, 680–688.

Petersen, G., Seberg, O., Yde, M., and Berthelsen, K. (2006). Phylogenetic relationships of Triticum and Aegilops and evidence for the origin of the A, B, and D genomes of common wheat (Triticum aestivum). Mol. Phylogenet. Evol. 39, 70–82.

Pfeifer, M., Kugler, K.G., Sandve, S.R., Zhan, B., Rudi, H., Hvidsten, T.R., Mayer, K.F.X., and Olsen, O.-A. (2014). Genome interplay in the grain transcriptome of hexaploid bread wheat. Science 345, 1250091.

Pires, J.C., Zhao, J., Schranz, M.E., Leon, E.J., Quijada, P.A., Lukens, L.N., and Osborn, T.C. (2004). Flowering time divergence and genomic rearrangements in resynthesized Brassica polyploids (Brassicaceae). Biol. J. Linn. Soc. 82, 675–688.

Plahte, E., Mestl, T., and Omholt, S.W. (1995). Feedback loops, stability and multistationarity in dynamical systems. J. Biol. Syst. 03, 409–413.

Pollak, E., and Sabran, M. (1999). On the theory of partially inbreeding finite populations. VI. The survival probability of a two-locus allele combination when there is partial selfing. Math. Biosci. 159, 97–112.

Potokina, E., Druka, A., Luo, Z., Wise, R., Waugh, R., and Kearsey, M. (2008). Gene expression quantitative trait locus analysis of 16 000 barley genes reveals a complex pattern of genome-wide transcriptional regulation. Plant J. 53, 90–101.

Qi, B., Huang, W., Zhu, B., Zhong, X., Guo, J., Zhao, N., Xu, C., Zhang, H., Pang, J., Han, F., et al. (2012). Global transgenerational gene expression dynamics in two newly synthesized allohexaploid wheat (Triticum aestivum) lines. BMC Biol. 10, 3.

Salamini, F., Özkan, H., Brandolini, A., Schäfer-Pregl, R., and Martin, W. (2002). Genetics and geography of wild cereal domestication in the near east. Nat. Rev. Genet. 3, 429–441.

Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G., et al. (2003). Genetics of gene expression surveyed in maize, mouse and man. Nature 422, 297–302.

Schnable, J.C., Springer, N.M., and Freeling, M. (2011). Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc. Natl. Acad. Sci. 108, 4069–4074.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., et al. (2009). The B73 Maize Genome: Complexity, Diversity, and Dynamics. Science 326, 1112–1115.

Sears, E.R. (1976). Genetic Control of Chromosome Pairing in Wheat. Annu. Rev. Genet. 10, 31–51.

Shaked, H., Kashkush, K., Ozkan, H., Feldman, M., and Levy, A.A. (2001). Sequence Elimination and Cytosine Methylation Are Rapid and Reproducible Responses of the Genome to Wide Hybridization and Allopolyploidy in Wheat. Plant Cell Online 13, 1749–1759.

Sharma, S.N., and Sain, R.S. (2004). Genetics of grains per spike in durum wheat under normal and late planting conditions. Euphytica 139, 1–7.

Sharma, S.N., Sain, R.S., and Sharma, R.K. (2003). Genetics of spike length in durum wheat. Euphytica 130, 155–161.

Shi, C., Uzarowska, A., Ouzunova, M., Landbeck, M., Wenzel, G., and Lübberstedt, T. (2007). Identification of candidate genes associated with cell wall digestibility and eQTL (expression quantitative trait loci) analysis in a Flint × Flint maize recombinant inbred line population. BMC Genomics 8, 22.

Shi, X., Ng, D.W.-K., Zhang, C., Comai, L., Ye, W., and Jeffrey Chen, Z. (2012). Cis- and trans-regulatory divergence between progenitor species determines gene-expression novelty in Arabidopsis allopolyploids. Nat. Commun. 3, 950.

De Smet, R., and Van de Peer, Y. (2012). Redundancy and rewiring of genetic networks following genome-wide duplication events. Curr. Opin. Plant Biol. 15, 168–176.

Soltis, D.E., Bell, C.D., Kim, S., and Soltis, P.S. (2008). Origin and Early Evolution of Angiosperms. Ann. N. Y. Acad. Sci. 1133, 3–25.

Song, K., Lu, P., Tang, K., and Osborn, T.C. (1995). Rapid genome change in synthetic polyploids of Brassica and its implications for polyploid evolution. Proc. Natl. Acad. Sci. 92, 7719–7723.

Storey, J.D., Akey, J.M., and Kruglyak, L. (2005). Multiple Locus Linkage Analysis of Genomewide Expression in Yeast. PLoS Biol 3, e267.

Stupar, R.M., and Springer, N.M. (2006). Cis-transcriptional Variation in Maize Inbred Lines B73 and Mo17 Leads to Additive Expression Patterns in the F1 Hybrid. Genetics 173, 2199–2210.

Sun, W., and Hu, Y. (2013). eQTL Mapping Using RNA-seq Data. Stat. Biosci. 5.

Swanson-Wagner, R.A., DeCook, R., Jia, Y., Bancroft, T., Ji, T., Zhao, X., Nettleton, D., and Schnable, P.S. (2009). Paternal Dominance of Trans-eQTL Influences Gene Expression Patterns in Maize Hybrids. Science 326, 1118–1120.

Tang, H., Woodhouse, M.R., Cheng, F., Schnable, J.C., Pedersen, B.S., Conant, G., Wang, X., Freeling, M., and Pires, J.C. (2012). Altered patterns of fractionation and exon deletions in Brassica rapa support a two-step model of paleohexaploidy. Genetics 190, 1563–1574.

Therneau, T. (2012). coxme: Mixed Effects Cox Models.

Thomas, R., and Kaufman, M. (2001). Multistationarity, the basis of cell differentiation and memory. II. Logical analysis of regulatory networks in terms of feedback circuits. Chaos Interdiscip. J. Nonlinear Sci. 11, 180–195.

Thomas, R., and Richelle, J. (1988). Positive feedback loops and multistationarity. Discrete Appl. Math. 19, 381–396.

Thomas, B.C., Pedersen, B., and Freeling, M. (2006). Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. Genome Res. 16, 934–946.

Udall, J.A., and Wendel, J.F. (2006). Polyploidy and Crop Improvement. Crop Sci. 46, S – 3.

Volis, S., Shulgina, I., Zaretsky, M., and Koren, O. (2011). Epistasis in natural populations of a predominantly selfing plant. Heredity 106, 300–309.

Wang, J., Tian, L., Madlung, A., Lee, H.-S., Chen, M., Lee, J.J., Watson, B., Kagochi, T., Comai, L., and Chen, Z.J. (2004). Stochastic and Epigenetic Changes of Gene Expression in Arabidopsis Polyploids. Genetics 167, 1961–1973.

Wang, J., Tian, L., Lee, H.-S., Wei, N.E., Jiang, H., Watson, B., Madlung, A., Osborn, T.C., Doerge, R.W., Comai, L., et al. (2006). Genomewide Nonadditive Gene Regulation in Arabidopsis Allotetraploids. Genetics 172, 507–517.

Wang, J., Yu, H., Xie, W., Xing, Y., Yu, S., Xu, C., Li, X., Xiao, J., and Zhang, Q. (2010). A global analysis of QTLs for expression variations in rice shoots at the early seedling stage. Plant J. 63, 1063–1074.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63.

Weiss-Schneeweiss, H., Schneeweiss, G.M., Stuessy, T.F., Mabuchi, T., Park, J.-M., Jang, C.-G., and Sun, B.-Y. (2007). Chromosomal stasis in diploids contrasts with genome restructuring in auto- and allopolyploid taxa of Hepatica (Ranunculaceae). New Phytol. 174, 669–682.

Wendel, J.F., Schnabel, A., and Seelanan, T. (1995). Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (Gossypium). Proc. Natl. Acad. Sci. 92, 280–284.

West, M.A.L., Kim, K., Kliebenstein, D.J., Leeuwen, H. van, Michelmore, R.W., Doerge, R.W., and Clair, D.A.S. (2007). Global eQTL Mapping Reveals the Complex Genetic Architecture of Transcript-Level Variation in Arabidopsis. Genetics 175, 1441–1450.

Wittkopp, P.J., Haerum, B.K., and Clark, A.G. (2004). Evolutionary changes in cis and trans gene regulation. Nature 430, 85–88.

Wittkopp, P.J., Haerum, B.K., and Clark, A.G. (2008). Regulatory changes underlying expression differences within and between Drosophila species. Nat. Genet. 40, 346–350.

Woodhouse, M.R., Schnable, J.C., Pedersen, B.S., Lyons, E., Lisch, D., Subramaniam, S., and Freeling, M. (2010). Following Tetraploidy in Maize, a Short Deletion Mechanism Removed Genes Preferentially from One of the Two Homeologs. PLoS Biol 8, e1000409.

Woodhouse, M.R., Cheng, F., Pires, J.C., Lisch, D., Freeling, M., and Wang, X. (2014). Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. Proc. Natl. Acad. Sci. 111, 5283–5288.

Wright, S. (1950). Genetical structure of populations. Nature 166, 247–249.

Wright, S.I., Ness, R.W., Foxe, J.P., and Barrett, S.C.H. (2008). Genomic Consequences of Outcrossing and Selfing in Plants. Int. J. Plant Sci. 169, 105–118.

Xu, Y., Zhong, L., Wu, X., Fang, X., and Wang, J. (2009). Rapid alterations of gene expression and cytosine methylation in newly synthesized Brassica napus allopolyploids. Planta 229, 471–483.

Zhang, X., Cal, A.J., and Borevitz, J.O. (2011). Genetic architecture of regulatory variation in Arabidopsis thaliana. Genome Res. 21, 725–733.

III.3. Limites du modèle utilisé et perspectives

Les modèles utilisés dans la publication *Asymmetry of A and B genomes of tetraploid wheat* Triticum turgidum *in cis and trans regulations of the expression* (III.2), connaissent des limites. Voici nos suggestions pour l'améliorer.

III.3.1 Indépendance entre les effets génétiques

Le modèle d'estimation des effets sur lequel nos deux modèles sont basés, est le modèle UWR (Cheverud and Routman, 1995), présenté dans le paragraphe II.4.2.2. Ce modèle a été choisi car il permet d'estimer des effets indépendamment des fréquences alléliques (effets fonctionnels). Néanmoins, les effets génétiques ($x_j\beta_A$, $x_k\beta_B$, $x_jx_k\alpha\alpha$) du modèle (1) du III.2.2.5 ne sont pas indépendants, car en appliquant les calculs du II.4.3 à ce modèle, on trouve que la matrice de variance-covariance des variables explicatives n'est pas diagonale :

$$\begin{pmatrix} V(x_j) & \mathrm{cov}(x_j,x_k) & \mathrm{cov}(x_j,x_jx_k) \\ \mathrm{cov}(x_j,x_k) & V(x_k) & \mathrm{cov}(x_j,x_jx_k) \\ \mathrm{cov}(x_j,x_jx_k) & \mathrm{cov}(x_j,x_jx_k) & V(x_jx_k) \end{pmatrix}$$

$$= \begin{pmatrix} 1/4 & \frac{p_{AB}-p_{Ab}-p_{aB}+p_{ab}}{4} & \frac{p_B-p_b}{8} \\ \frac{p_{AB}-p_{Ab}-p_{aB}+p_{ab}}{4} & 1/4 & \frac{p_A-p_a}{8} \\ \frac{p_B-p_b}{8} & \frac{p_A-p_a}{8} & 1/16 \end{pmatrix}$$

De plus, notons qu'en présence d'épistasie, les estimateurs (des moindres carrés) pour ce modèle sont biaisés. En effet, supposons qu'il existe trois SNPs, notés X, W et Z, dont on a estimé l'effet sur le niveau d'expression d'un gène. On suppose que ces SNPs sont en équilibre de liaison. Leurs allèles sont respectivement notés X/x, W/w et Z/z. Supposons que X appartienne au gène dont l'expression est considérée. Deux modèles sont alors possibles :

$Y_{ijk} = \mu + x_j\beta_X + x_k\beta_W + x_jx_k\alpha\alpha_{XW} + b_i + \epsilon_{ijk}$ (1.i)

$Y_{ijm} = \mu + x_j\beta_X + x_m\beta_Z + x_jx_m\alpha\alpha_{XZ} + b_i + \epsilon_{ijm}$ (1.ii)

Avec $x_{j(i)} = \begin{cases} \frac{1}{2} \text{ si } j = X \\ -\frac{1}{2} \text{ si } j = x \end{cases}$, $x_{k(i)} = \begin{cases} \frac{1}{2} \text{ si } k = W \\ -\frac{1}{2} \text{ si } k = w \end{cases}$ et $x_{n(i)} = \begin{cases} \frac{1}{2} \text{ si } m = Z \\ -\frac{1}{2} \text{ si } m = z \end{cases}$

L'effet additif β_X est donc estimé dans deux contextes différents. On appele $\beta_{X,1}$ l'estimation par le modèle (1.i) et $\beta_{X,2}$ l'estimation par le modèle (1.ii). Ces deux modèles sont des cas particuliers du modèle (1).

Si on définit le modèle UWR avec trois locus (SNPs X, W et Z), alors les valeurs génotypiques trilocus sont définies ainsi :

$$G_{jkm} = E(\bar{Y}_{jkm})$$

Dans ce modèle trilocus, les valeurs génotypiques bilocus sont définies comme des moyennes ajustées, calculées à partir des valeurs génotypiques trilocus (de même, dans le modèle bilocus, les valeurs génotypiques unilocus sont des moyennes ajustées calculées à partir des valeurs génotypiques bilocus). Ces valeurs génotypiques sont égales à :

$$G_{jk} = \frac{G_{jkZ} + G_{jkz}}{2}$$

$$G_{jm} = \frac{G_{jWm} + G_{jwm}}{2}$$

Le paramètre β_{X} est par conséquent égal à

$$\beta_{\mathrm{X}} = \frac{(G_{XW} + G_{Xw}) - (G_{xW} + G_{xw})}{2} = \frac{(G_{XZ} + G_{Xz}) - (G_{xZ} + G_{xz})}{2}$$

Par ailleurs,

$$\alpha\alpha_{\mathrm{XW}} = G_{XW} + G_{xw} - G_{Xw} - G_{xW}$$

Et

$$\alpha\alpha_{\mathrm{XZ}} = G_{XZ} + G_{xz} - G_{Xz} - G_{xZ}$$

Enfin, notons que l'épistasie de deuxième ordre est égale à :

$$aaa = (G_{XWZ} - G_{XWz}) + (G_{xwZ} - G_{xwz}) - (G_{XwZ} - G_{Xwz}) - (G_{xWZ} - G_{xWz})$$

En utilisant un modèle bilocus (comme dans l'étude), et en ignorant donc le troisième SNP, on n'estime pas les valeurs génotypiques trilocus, qui permettraient une estimation correcte des valeurs génotypiques bilocus pour le modèle UWR. A la place, les valeurs génotypiques bilocus sont estimées en calculant la moyenne du phénotype des individus porteurs du génotype bilocus considéré. Nous noterons Γ_{jk} et Γ_{jn} ces valeurs génotypiques bilocus incorrectes pour le modèle UWR. Par exemple :

$$\Gamma_{XW} = E(\bar{Y}_{XW}) \neq \frac{G_{XWZ}+G_{XWz}}{2} = G_{XW} \text{ (sauf si } p_Z = 1/2)$$

Par conséquent, en supposant un équilibre de liaison :

$$\Gamma_{XW} = G_{XWZ} \times p_Z + G_{XWz} \times (1 - p_Z) = \frac{G_{XWZ} + G_{XWz}}{2} + \left(p_Z - \frac{1}{2}\right)(G_{XWZ} - G_{XWz})$$

$$= G_{XW} + \left(p_Z - \frac{1}{2}\right)(G_{XWZ} - G_{XWz})$$

De la même manière :

$$\Gamma_{Xw} = G_{Xw} + \left(p_Z - \frac{1}{2}\right)(G_{XwZ} - G_{Xwz})$$

Et ainsi de suite pour Γ_{xW} et Γ_{xw}.

Dans le modèle (1.i), l'estimation de $\beta_{\mathrm{X},1}$est biaisé :

$$\beta_{\mathrm{X},1} = \frac{(\Gamma_{XW} + \Gamma_{Xw}) - (\Gamma_{xW} + \Gamma_{xw})}{2} = \beta_{\mathrm{X}} + \left(p_Z - \frac{1}{2}\right)\alpha\alpha_{\mathrm{XZ}}$$

Par conséquent, s'il n'y a pas d'épistasie entre le SNP X et le SNP Z, l'estimation par le modèle (1) sera correcte. De la même façon :

$$\beta_{\mathrm{X},2} = \beta_{\mathrm{X}} + \left(p_W - \frac{1}{2}\right)\alpha\alpha_{\mathrm{XW}}$$

L'effet épistatique entre le SNP X et le SNP W est estimé par le modèle 1 comme :

$$\begin{aligned}\alpha\alpha_{\mathrm{XW},1} &= \Gamma_{XW} + \Gamma_{xw} - \Gamma_{Xw} - \Gamma_{xW} \\ &= (G_{XW} + G_{xw} - G_{Xw} - G_{xW}) \\ &+ \left(p_Z - \frac{1}{2}\right)[(G_{XWZ} - G_{XWz}) + (G_{xwZ} - G_{xwz}) - (G_{XwZ} - G_{Xwz}) \\ &- (G_{xWZ} - G_{xWz})] = \alpha\alpha_{\mathrm{XW}} + \left(p_Z - \frac{1}{2}\right)aaa\end{aligned}$$

De la même façon :

$$\alpha\alpha_{\mathrm{XZ},2} = \alpha\alpha_{\mathrm{XZ}} + \left(p_W - \frac{1}{2}\right)aaa$$

En conclusion, le modèle (1) est sans biais en l'absence d'épistasie. Dans notre travail, malgré le fait que peu d'épistasie significative ait été détectée, les corrélations entre estimateurs (cf. matrice de variance-covariance des variables explicatives) ont potentiellement eu un impact, comme le montre la **Figure 15**, qui représente l'estimation de l'effet d'un SNP sur une paire non-homéologues, en fonction de l'effet de ce même SNP estimé sur une paire homéologue.

Notons que si l'on trace le graphique de la moyenne des effets estimés d'un SNP sur des paires non homéologues, en fonction de la moyenne des effets estimés de ce SNP sur des paires homéologues, le graphique ressemble beaucoup plus à une régression linéaire, comme le montre la **Figure 16**. Par exemple, si un SNP X a été utilisé dans deux paires de SNPs homéologues, il y a deux estimations de son effet sur ce type de paires : $\beta_{\mathrm{X},1}$ et $\beta_{\mathrm{X},2}$. S'il a ensuite été employé trois fois dans des paires non-homéologues, il y a trois estimations de son effet *cis* au sein de paires non homéologues : $\beta_{\mathrm{X},3}$**,** $\beta_{\mathrm{X},4}$ et $\beta_{\mathrm{X},5}$. La figure 16 représente donc un point par SNP, avec $E(\beta_{\mathrm{X},3}, \beta_{\mathrm{X},4}, \beta_{\mathrm{X},5}) = f(E(\beta_{\mathrm{X},1}, \beta_{\mathrm{X},2}))$.

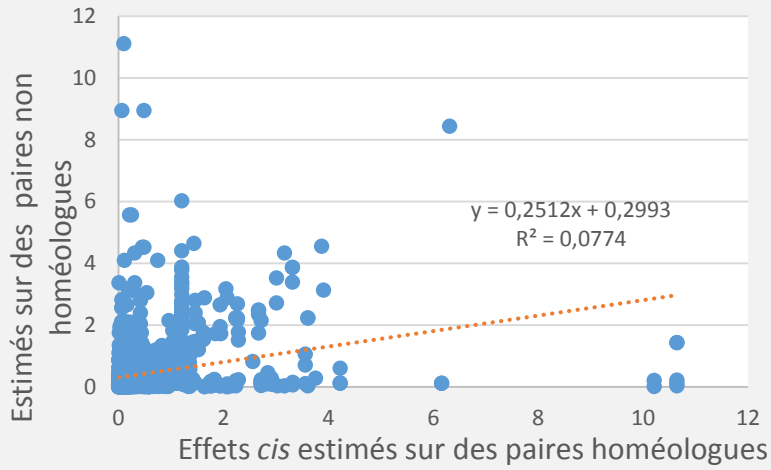


Figure 15 : Effets *cis* estimés sur les paires non homéologues, en fonction des effets *cis* estimés sur des paires homéologues (valeurs non standardisées)

En abscisse : les effets *cis* estimés au sein de paires d'homéologues. Il s'agit par exemple de la valeur de $\boldsymbol{\beta}_{\mathrm{X},1}$ estimé sur ce modèle :

$\boldsymbol{Y}_{ijk} = \mu + \boldsymbol{x}_j\boldsymbol{\beta}_{\mathrm{X},1} + \boldsymbol{x}_k\boldsymbol{\beta}_{\mathrm{W}} + \boldsymbol{x}_j\boldsymbol{x}_k\boldsymbol{\alpha}\boldsymbol{\alpha}_{\mathrm{XW}} + \boldsymbol{b}_i + \boldsymbol{\epsilon}_{ijk}$ **(i)**

Où X et W sont deux SNPs appartenant à des gènes homéologues. Dans cet exemple, $\boldsymbol{Y}_{ijk}$ est le niveau d'expression du gène auquel X appartient.

En ordonnées : les mêmes effets *cis* estimés au sein de paires de gènes non homéologues. Il s'agit par exemple de la valeur de $\boldsymbol{\beta}_{\mathrm{X},2}$ estimé sur ce modèle :

$\boldsymbol{Y}_{imk} = \mu + \boldsymbol{x}_j\boldsymbol{\beta}_{\mathrm{X},2} + \boldsymbol{x}_m\boldsymbol{\beta}_{\mathrm{Z}} + \boldsymbol{x}_j\boldsymbol{x}_m\boldsymbol{\alpha}\boldsymbol{\alpha}_{\mathrm{XZ}} + \boldsymbol{b}_i + \boldsymbol{\epsilon}_{imk}$ **(ii)**

Où X et Z sont deux SNPs appartenant à des gènes non homéologues. Dans cet exemple, $\boldsymbol{Y}_{imk}$ est à nouveau le niveau d'expression du gène auquel X appartient.

Cette figure représente donc $\boldsymbol{\beta}_{\mathrm{X},2} = \boldsymbol{f}(\boldsymbol{\beta}_{\mathrm{X},1})$

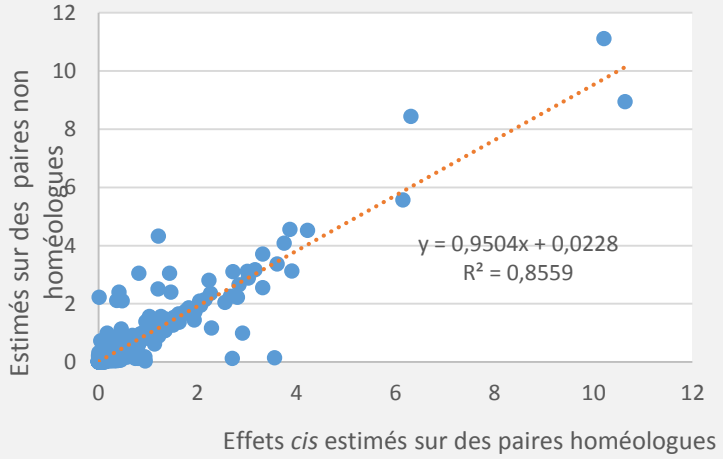


Figure 16 : Valeur moyennée des effets *cis* estimés sur les paires non homéologues, en fonction de la valeur moyennée des effets cis estimés sur des paires homéologues (valeurs non standardisées)

Un point = un SNP

En abscisse : moyenne des effets *cis* d'un SNP, estimés au sein de paires d'homéologues.

En ordonnées : moyenne des effets *cis* du même SNP, estimés au sein de paires de gènes non homéologues.

La figure 16 indique que quand les valeurs des effets *cis* sont moyennées sur les deux types de paires (homéologues ou non homéologues), alors les moyennes des valeurs estimées sur les deux types de paires sont la plupart du temps similaires. Cela pourrait indiquer qu'il existe un biais dans nos estimations, mais le biais s'annule quand les estimations des effets sont moyennées.

Une solution serait d'utiliser un modèle complet, comprenant tous les SNPs, et d'effectuer une ridge régression (Hoerl and Kennard, 1970). Le problème est que nous avons beaucoup de données manquantes qui rendent cette méthode difficile.

Álvarez-Castro et Carlborg ont proposé une méthode pour permettre de « convertir » les effets statistiques en effets fonctionnels, et vice-versa (Álvarez-Castro and Carlborg, 2007). Une solution pour pallier ce problème de non-orthogonalité serait éventuellement d'utiliser un autre modèle, comme le modèle centré G2A-D, présenté en II.4.2.3, qui est orthogonal en l'absence de déséquilibre de liaison. Pour s'assurer de l'orthogonalité du modèle, il ne faudrait sélectionner des SNPs qui ne soient pas en déséquilibre de liaison, et appliquer un tri plus strict que celui présenté dans l'article ($r < 0.4$). Ensuite, il serait possible de convertir les effets estimés afin d'obtenir des effets fonctionnels indépendants des fréquences, à l'aide de la méthode d'Álvarez-Castro et Carlborg.

Pour un cas avec trois locus, comme celui présenté ci-dessus avec les SNPs X, W et Z, et les modèles (1.i) et (1.ii), le processus de conversion serait le suivant : six effets ont été estimés avec deux modèles bilocus basé sur le modèle G2A-D: μ^{G2A-D}, $\beta_{\mathrm{X}}^{G2A-D}$, $\beta_{\mathrm{W}}^{G2A-D}$, $\beta_{\mathrm{Z}}^{G2A-D}$, $\alpha\alpha_{\mathrm{XW}}^{G2A-D}$ et $\alpha\alpha_{\mathrm{XZ}}^{G2A-D}$. Ces modèles bilocus sont semblables au modèle (1), sauf qu'ils ne sont pas basés sur le même modèle (G2A-D au lieu de UWR). Comme les modèles sont orthogonaux, alors les estimations des effets estimés deux fois (μ^{G2A-D} et $\beta_{\mathrm{X}}^{G2A-D}$) sont les mêmes pour les deux modèles.

Soit $\boldsymbol{S}_{G2A-D}^{\mathrm{L}}$ une matrice de transition unilocus haploïde du modèle G2A-D associé au locus L. Elle est égale à :

$\boldsymbol{S}_{G2A-D}^{\mathrm{L}} = \begin{pmatrix} 1 & 1-p_L \\ 1 & -p_L \end{pmatrix}$, avec p_L la fréquence d'un des deux allèles à ce locus.

Soit $\boldsymbol{S}_{UWR}$ une matrice de transition unilocus haploïde du modèle UWR :

$$\boldsymbol{S}_{UWR} = \begin{pmatrix} 1 & 1/2 \\ 1 & -1/2 \end{pmatrix}$$

Soit $\boldsymbol{E}_{G2A-D}$ le vecteur des effets génétiques estimés par le modèle centré :

$$\boldsymbol{E}_{G2A-D} = \begin{pmatrix} \mu^{G2A-D} \\ \beta_{\mathrm{X}}^{G2A-D} \\ \beta_{\mathrm{W}}^{G2A-D} \\ \beta_{\mathrm{Z}}^{G2A-D} \\ \alpha\alpha_{\mathrm{XW}}^{G2A-D} \\ \alpha\alpha_{\mathrm{XZ}}^{G2A-D} \\ \alpha\alpha_{\mathrm{WZ}}^{G2A-D} \\ aaa^{G2A-D} \end{pmatrix}$$

Les effets $\alpha\alpha_{\mathrm{WZ}}^{G2A-D}$ et aaa^{G2A-D} (épistasie de deuxième ordre) n'ont pas pu être estimés avec les deux modèles bilocus. On va donc les supposer nuls : $\alpha\alpha_{\mathrm{WZ}}^{G2A-D} = 0$ et $aaa^{G2A-D} = 0$

La matrice de design à trois locus du modèle centré est égal à : $\boldsymbol{S}_{G2A-D}^{\mathrm{XWZ}} = \boldsymbol{S}_{G2A-D}^{\mathrm{Z}} \otimes \boldsymbol{S}_{G2A-D}^{\mathrm{W}} \otimes \boldsymbol{S}_{G2A-D}^{\mathrm{X}}$ puisque le déséquilibre de liaison est nul. Celle du modèle UWR est égal à $\boldsymbol{S}_{UWR}^{\mathrm{XWZ}} = \boldsymbol{S}_{UWR} \otimes \boldsymbol{S}_{UWR} \otimes \boldsymbol{S}_{UWR}$.

Dans les faits, il faudra calculer ces matrices de design pour tous les SNPs considérés. Pour N paires de SNPs :

$$\boldsymbol{S}_{G2A-D}^{\mathrm{total}} = \boldsymbol{S}_{G2A-D}^{\mathrm{A1}} \otimes \ldots \otimes \boldsymbol{S}_{G2A-D}^{\mathrm{A}N} \otimes \boldsymbol{S}_{G2A-D}^{\mathrm{B1}} \ldots \otimes \boldsymbol{S}_{G2A-D}^{\mathrm{B}N}$$

$$\boldsymbol{S}_{UWR}^{\mathrm{total}} = \boldsymbol{S}_{UWR} \otimes \ldots \otimes \boldsymbol{S}_{UWR}$$

De même, il faudra construire un vecteur des effets génétiques estimés avec le modèle centré : $\boldsymbol{E}_{G2A-D}^{\mathrm{total}}$, à partir des estimations des modèles bilocus. Les effets épistatiques qui n'auront pas été estimés seront fixés arbitrairement à 0 (ce qui risque d'induire un biais). Ces matrices et vecteurs seront à très grandes dimensions. Concrètement, les très grandes dimensions des matrices et vecteurs vont poser problème. La mise en œuvre pratique de cette méthode proposée par Álvarez-Castro et Carlborg mérite donc encore d'être approfondie.

Les effets génétiques fonctionnels pourront alors être estimés comme:

$$\boldsymbol{E}_{UWR}^{\mathrm{total}} = \boldsymbol{S}_{UWR}^{\mathrm{total}\,-1} \otimes \boldsymbol{S}_{G2A-D}^{\mathrm{total}} \otimes \boldsymbol{E}_{G2A-D}^{\mathrm{total}}$$

L'utilisation du modèle G2A permettra par ailleurs d'estimer les variances et covariances génétiques (ces dernières seront nulles en l'absence de déséquilibre de liaison). Elle se calcule selon la méthode présentée dans le paragraphe II.4.3.

III.3.2 Standardisation des effets

Comme expliqué dans l'article, la standardisation des effets induit une corrélation artificielle entre les effets génétiques, notamment entre les effets additifs. C'est très probablement cet artefact qui fait apparaître les effets *cis* estimés sur des paires non-homéologues comme plus importants, par rapport à ceux estimés sur des paires homéologues.

A l'état « brute » (sans standardisation), les valeurs absolues des effets *cis* et des effets *trans* sont corrélées positivement (le coefficient de corrélation est de l'ordre de 0,60). Cela est normal car si un gène a un niveau d'expression fort, alors les effets régulateurs de son niveau d'expression seront forts (en valeur absolue). A l'inverse, si le niveau d'expression est faible, alors les effets estimés seront faibles. Cette corrélation entre le niveau d'expression et les effets estimés, est la raison pour laquelle nous avons souhaité standardiser les effets.

$$s = \sqrt{\sum_{jk} (\overline{Y_{jk}} - \hat{\mu})^2}$$

Le coefficient de standardisation utilisé dépend du niveau d'expression, mais aussi du génotype.

De plus, mes analyses ont par ailleurs montré que la moyenne des valeurs des mêmes effets *cis* estimés, d'une part dans une paire homéologue, d'autre part au sein de paires non-homéologue, ne variait pas de manière substantielle à l'état brut (voir Figure 16). En revanche, une fois standardisés, la moyenne des mêmes effets *cis* pouvaient considérablement changer de valeur (**Figure 17**), puisque le facteur s utilisé pour standardiser les effets, dépendait également des effets *trans* et des effets épistatiques : il était propre à chaque paire de SNPs testée.

Il serait préférable de standardiser par un facteur ne dépendant que du niveau d'expression du gène considéré, indépendamment du génotype. Ainsi, la standardisation n'induirait pas une corrélation artificielle entre effets standardisés, tout en permettant une comparaison entre les effets estimés sur des forts niveaux d'expression, et ceux estimés sur de faibles niveaux d'expression.

Notons néanmoins, que même si ce facteur de standardisation n'est pas des plus pertinents, il n'affecte pas les résultats les plus importants de notre étude, à savoir que l'on détecte plus d'effets *trans* au sein des paires homéologues qu'au sein des paires non homéologues, et que le génome B présente plus d'eQTL *cis* et *trans* que le génome A.

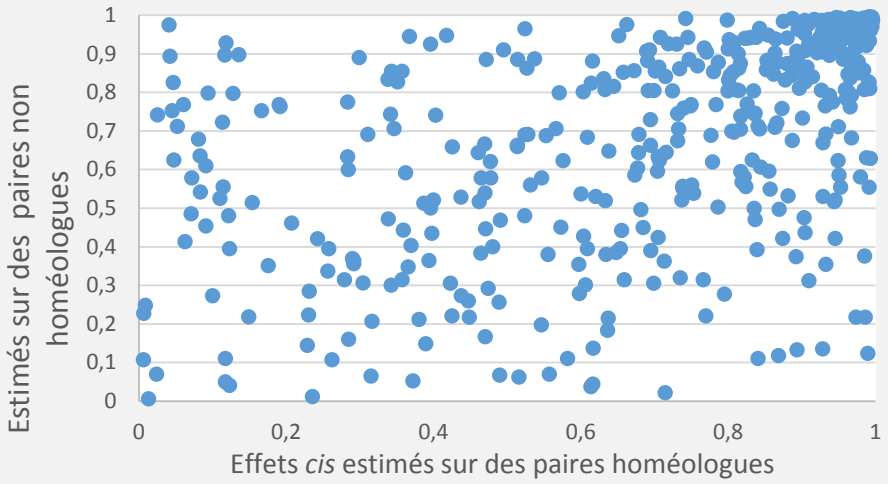


Figure 17 : Valeur moyennée des effets *cis* estimés sur les paires non homéologues, en fonction de la valeur moyennée des effets cis estimés sur des paires homéologues (valeurs standardisées)

Un point = un SNP

En abscisse : moyenne des effets *cis* d'un SNP, estimés au sein de paires d'homéologues.

En ordonnées : moyenne des effets *cis* du même SNP, estimés au sein de paires de gènes non homéologues.

III.3.3 Résidus

La distribution des résidus étaient normale et centrée en 0. Néanmoins, nous avons pu constater que pour 10 individus (sur 175), les estimations des valeurs phénotypiques étaient systématiquement sous-estimées par le modèle. Ces dix individus sont heureusement en faible proportion, et n'affectent probablement pas outre mesure les estimations.

Si les estimations sont sous-estimées, cela signifie que ces 10 individus ont, en moyenne, des niveaux d'expression plus forts, que les autres. Prenons par exemple, deux individus ayant le même génotype kj. L'un (le numéro 1) provient du groupe des 10 individus, l'autre (le numéro 2) provient du reste de l'échantillon. Alors :

$$r_1 = Y_{1jk} - \widehat{Y_{jk}}$$

$$r_2 = Y_{2jk} - \widehat{Y_{jk}}$$

Donc si $r_1 > r_2$, alors $Y_{1jk} > Y_{2jk}$.

Nos données confirment que le RPKM moyen des 10 individus surestimés étaient légèrement plus élevé (18,3) que celui des autres individus (17,8). Les 10 individus avaient été génotypés lors d'une première série, où il y avait 48 génotypes par lane. Leur profondeur de séquençage était donc réduite,

et leur niveau d'expression global, en nombre total de reads, était plus faible que celui des autres individus.

Le RPKM (Reads Per Kilobase per Million mapped reads), utilisé dans cette étude, permet pourtant de normaliser l'expression. Chaque niveau d'expression d'un gène correspond à une proportion du niveau d'expression globale. Le biais dû à la profondeur de l'expression devrait donc être normalement corrigé. Néanmoins, s'il y a un biais d'expression en fonction de la longueur des gènes, et plus précisément, si les gènes les plus courts ont un nombre de reads proportionnellement plus grand (c'est-à-dire, ont tendance à être plus exprimés), le RPKM moyen des individus génotypés avec une faible profondeur sera plus fort.

Supposons qu'il y ait deux individus : l'individu n°1 et l'individu n°2. L'individu n°1 a été génotypé avec une plus faible profondeur, et comprend donc un nombre plus réduit de reads. Soient deux gènes : le gène A et le gène B. Le gène A est moins long que le gène B. Supposons qu'on puisse exprimer le nombre de reads associé au gène B, en fonction du nombre de reads associés au gène A :

$$r_{B,j} = r_{A,j}\frac{L_B}{L_A} + x$$

Avec $r_{g,j}$ le nombre de reads associés au gène g pour l'individu j ($r_{A,1} < r_{A,2}$ et $r_{B,1} < r_{B,2}$) et L_g la longueur du gène g ($L_B > L_A$). Le nombre x est un nombre réel, qui indique s'il existe un biais de transcription en fonction de la longueur du gène. Si $x = 0$, alors les deux gènes ont un niveau d'expression proportionnel à leur longueur. Si $x > 0$, alors le gène long (B) a proportionnellement une expression plus forte, par rapport au gène court (A). A l'inverse, si $x < 0$, alors le gène court (A) a proportionnellement un niveau d'expression plus fort que le gène long (B). On suppose, pour simplifier, que cette relation est vraie pour les individus de l'exemple.

Le RPKM est égal à :

$$RPKM_{g,j} = \frac{r_{g,j} \times 10^9}{L_g \times R_j}$$

R_j est la somme totale des reads d'un individu j sur les G gènes exprimés dans le jeu de données.

$$R_j = \sum_{g=1}^{G} r_{g,j}$$

Ici, cette somme est égale à : $R_j = r_{A,j} + r_{B,j}$. Dans notre exemple $R_1 < R_2$ car l'individu 1 a été génotypé avec une plus faible profondeur.

Calculons les RPKM moyens pour les deux individus :

$$E(RPKM_1) = \frac{1}{2}\left[\frac{r_{A,1} \times 10^9}{L_A \times R_1} + \frac{r_{B,1} \times 10^9}{L_B \times R_1}\right] = \frac{10^9}{2}\left[\frac{r_{A,1}}{L_A \times R_1} + \frac{r_{B,1}}{L_B \times R_1}\right]$$

$$= \frac{10^9}{2}\left[\frac{r_{A,1}}{L_A \times R_1} + \frac{1}{L_B \times R_1} r_{A,1} \frac{L_B}{L_A} + \frac{x}{L_B \times R_1}\right] = \frac{10^9}{2 \times R_1}\left[\frac{r_{A,1}}{L_A} + \frac{r_{A,1}}{L_A} + \frac{x}{L_B}\right]$$

$$= \frac{10^9}{2 \times R_1}\left[2\frac{r_{A,1}}{L_A} + \frac{x}{L_B}\right]$$

$$E(RPKM_2) = \frac{10^9}{2 \times R_2}\left[2\frac{r_{A,2}}{L_A} + \frac{x}{L_B}\right]$$

On trouve que le RPKM moyen de l'individu 1 sera supérieur à celui de l'individu 2 ($E(RPKM_1) > E(RPKM_2)$) si :

$$2\frac{r_{A,1}}{L_A \times R_1} + \frac{x}{L_B \times R_1} > 2\frac{r_{A,2}}{L_A \times R_2} + \frac{x}{L_B \times R_2}$$

$$\leftrightarrow x > 2\frac{L_B}{L_A} \times \frac{(r_{A,2} \times R_1 - r_{A,1} \times R_2)}{R_2 - R_1}$$

Or $R_1 = r_{A,1} + r_{B,1} = r_{A,1} + r_{A,1}\frac{L_B}{L_A} + x$ et $R_2 = r_{A,2} + r_{B,1} = r_{A,2} + r_{A,2}\frac{L_B}{L_A} + x$. Donc $R_2 - R_1 = \frac{L_A + L_B}{L_A}(r_{A,2} - r_{A,1})$.

Par ailleurs, $r_{A,2} \times R_1 = r_{A,2} \times r_{A,1}\frac{L_A + L_B}{L_A} + r_{A,2} \times x$ et$r_{A,1} \times R_2 = r_{A,2} \times r_{A,1}\frac{L_A + L_B}{L_A} + r_{A,1} \times x$. Par conséquent, $r_{A,2} \times R_1 - r_{A,1} \times R_2 = (r_{A,2} - r_{A,1}) \times x$

Donc $E(RPKM_1) > E(RPKM_2)$ si $x > 2\frac{L_B}{L_A} \times \frac{L_A}{L_A + L_B} \times \frac{(r_{A,2} - r_{A,1}) \times x}{r_{A,2} - r_{A,1}}$

$$\leftrightarrow x > 2\frac{L_B}{L_A + L_B} \times x$$

$$\leftrightarrow x(L_A + L_B) - 2L_B x > 0$$

$$\leftrightarrow x \times [L_A - L_B] > 0$$

Comme $L_B > L_A$, alors $L_A - L_B < 0$. Donc :

- $E(RPKM_1) < E(RPKM_2)$ si $x > 0$
- $E(RPKM_1) = E(RPKM_2)$ si $x = 0$
- $E(RPKM_1) > E(RPKM_2)$ si $x < 0$

Ce calcul simplifié démontre que si les gènes longs ont une expression proportionnellement plus faible que les gènes courts ($x < 0$), alors des génotypes séquencés avec une plus faible profondeur, vont avoir un RPKM moyen plus fort. Cela explique pourquoi les 10 individus séquencés avec une plus faible

profondeur ont été surestimés. Les effets du modèle, ont été estimés à partir d'individus dont la majorité avait des RPKM plus faibles. Ces effets ne seront donc pas adaptés aux 10 individus dont le niveau d'expression est plus fort : ils seront alors sous-estimés.

Une manière de pallier ce problème est de ne plus utiliser le RPKM, mais une mesure corrigée, telle que le TPM (Transcripts per million), proposé par Wagner et al. (Wagner et al., 2012). Cette mesure tient compte de la longueur des reads et est insensible à la profondeur du séquençage, comme nous allons le montrer.

Soit l_j la longueur moyenne des reads d'un individu j. Pour chaque individu, on définit une valeur T_j :

$$T_j = \sum_{g=1}^{G} \frac{r_{g,j} \times l_j}{L_g}$$

Avec G le nombre total de gènes transcripts dans l'ensemble du jeu de données.

Le TPM de chaque individu pour un gène g donné est égal à :

$$TPM_j = \frac{r_{g,j} \times l_j \times 10^6}{L_g \times T_j}$$

Le TPM total de chaque individu est :

$$TPM_{tot} = \sum_{g=1}^{G} \frac{r_{g,j} \times l_j \times 10^6}{L_g \times T_j} = \frac{10^6}{T_j} \sum_{g=1}^{G} \frac{r_{g,j} \times l_j}{L_g} = \frac{10^6}{T_j} T_j = 10^6$$

Par conséquent, le TPM moyen de chaque individu (égal à TPM_{tot}/G) sera identique, indépendamment de la profondeur du séquençage.

Une autre manière de régler ce problème serait d'inclure la date ou la modalité de séquençage comme effet fixe dans le modèle.

III.3.2 Corrélation entre niveaux d'expression

Pour une paire de SNPs, le même modèle de détection des effets a été appliqué à deux niveaux d'expression : ceux des deux gènes auxquels appartiennent les SNPs.

Par exemple, si un SNP X appartient à un gène Ξ, et si un SNP L appartient à un gène Λ, alors il existe deux modèles :

$$Y_{ijk}^{\Xi} = \mu_{\Xi} + x_j \beta_{X,\Xi} + x_k \beta_{W,\Xi} + x_j x_k \alpha\alpha_{\Xi} + b_i + \epsilon_{ijk,\Xi}$$

$$Y_{ijk}^{\Lambda} = \mu_{\Lambda} + x_j \beta_{X,\Lambda} + x_k \beta_{W,\Lambda} + x_j x_k \alpha\alpha_{\Lambda} + b_i + \epsilon_{ijk,\Lambda}$$

Les effets $\beta_{X,\Xi}$, $\beta_{W,\Xi}$ et $\alpha\alpha_{\Xi}$ sont estimés indépendamment des effets $\beta_{X,\Lambda}$, $\beta_{W,\Lambda}$ et $\alpha\alpha_{\Lambda}$.

Or, en particulier si les gènes Ξ et Λ sont homéologues, les niveaux d'expression Y_{ijk}^{Ξ} et Y_{ijk}^{Λ} ont de fortes chances d'être corrélés.

Effectuer ainsi des régressions indépendantes ne permet pas de tenir compte de cette corrélation, qui induit une non-indépendance entre les résidus $\epsilon_{ijk,\Xi}$ et $\epsilon_{ijk,\Lambda}$. Il serait donc plus approprié d'utiliser un modèle multivarié. Ce modèle peut s'écrire de manière matricielle :

$$\boldsymbol{Y} = \boldsymbol{ZE} + \boldsymbol{\epsilon}$$

Avec $\boldsymbol{Y}$ la matrice des niveaux d'expression pour chacun des n individus, de dimension nx2 pour n individus ; $\boldsymbol{Z}$ est une matrice de design comprenant les coefficients du modèle UWR, de dimension n x4 ; $\boldsymbol{E}$ est la matrice des effets génétiques, de dimension 4x2 ; et $\boldsymbol{\epsilon}$ la matrice des résidus, de dimensions nx2.

$$\begin{pmatrix} Y_1^{\Xi} & Y_1^{\Lambda} \\ \dots & \dots \\ Y_i^{\Xi} & Y_i^{\Lambda} \\ \dots & \dots \\ Y_n^{\Xi} & Y_n^{\Lambda} \end{pmatrix} = \begin{pmatrix} 1 & x_{j(1)} & x_{k(1)} & x_{j(1)}x_{k(1)} \\ \dots & \dots & \dots & \dots \\ 1 & x_{j(i)} & x_{k(i)} & x_{j(i)}x_{k(i)} \\ \dots & \dots & \dots & \dots \\ 1 & x_{j(n)} & x_{k(n)} & x_{j(n)}x_{k(n)} \end{pmatrix} \begin{pmatrix} \mu_{\Xi} & \mu_{\Lambda} \\ \beta_{X,\Xi} & \beta_{X,\Lambda} \\ \beta_{W,\Xi} & \beta_{W,\Lambda} \\ \alpha\alpha_{\Xi} & \alpha\alpha_{\Lambda} \end{pmatrix} + \begin{pmatrix} \epsilon_1^{\Xi} & \epsilon_1^{\Lambda} \\ \dots & \dots \\ \epsilon_i^{\Xi} & \epsilon_i^{\Lambda} \\ \dots & \dots \\ \epsilon_n^{\Xi} & \epsilon_n^{\Lambda} \end{pmatrix}$$

Les erreurs pour chaque couple de valeurs d'expression d'un individu i suivent une loi normale multidimensionnelle. Au minimum, on pourrait définir :

$$\begin{pmatrix} \epsilon_i^{\Xi} \\ \epsilon_i^{\Lambda} \end{pmatrix} \sim_{iid} N\left(O, \sigma_E^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

PARTIE IV

ROLE DE L'EPISTASIE DANS L'EVOLUTION

IV. Role de l'epistasie dans l'evolution

IV.1. Prédire l'évolution à court terme avec l'épistasie statistique

Les modèles statistiques permettent d'estimer la part de variance génétique provenant des effets additifs et celles provenant de l'épistasie (voir paragraphe II.4.3.). Ces variances génétiques peuvent ensuite être utilisées pour calculer la covariance parent-enfant, qui est une mesure de la ressemblance entre parents et descendants, mais qui permet surtout de prédire la réponse à la sélection, après une génération de sélection. La réponse à la sélection est le changement de moyenne phénotypique entre la génération parentale et les descendants issus des individus sélectionnés.

IV.1.1. Calcul de la covariance parent-enfant dans le cas haploïde bi-locus

La covariance parent-enfant est la covariance entre le phénotype des parents et celui de leurs descendants. C'est une mesure de la ressemblance entre parents et descendants. Elle est utile pour prédire le progrès génétique, comme nous le montrerons ultérieurement (cf IV.1.2).

Gianola et al. 2013 ont remarqué que quand il y avait du déséquilibre de liaison entre gènes, il était trompeur de considérer la variance additive d'un locus comme représentatif de la contribution de ce locus à la variance génétique totale. En effet, dans ce cas-là, des covariances apparaissent entre les effets (cf. II.4.3). Ils développèrent la notion de « contribution ». Dans un modèle sans épistasie, la contribution d'un locus L, parmi h locus, est égale à la variance additive de ce locus L, plus toutes les covariances additives auxquelles il participe :

$$C_L = V_L + \sum_{\substack{H=1 \\ H \neq L}}^{h} \mathrm{cov}_{LH}$$

Nous avons étendu ce concept au cas où il y a de l'épistasie. Nous avons considéré que la contribution de l'épistasie était équivalente à la contribution additive d'un locus.

Considérons un cas bilocus haploïde : les contributions que l'on peut estimer sont celles des effets additifs de chacun des locus et celle de l'épistasie. Soient C_A, C_B et C_{aa} les contributions du locus A, du locus B et de l'épistasie, à la variance génétique. Elles sont égales à :

$$C_A = V_A + \mathrm{cov}_{AB} + \mathrm{cov}_{Aaa}$$

$$C_B = V_B + \mathrm{cov}_{AB} + \mathrm{cov}_{Baa}$$

$C_{aa} = V_{aa} + \mathrm{cov}_{Aaa} + \mathrm{cov}_{Baa}$

Avec V_A la variance additive au locus A, V_B la variance additive au locus B, V_{aa} la variance épistatique, cov_{AB} la covariance entre effets additifs, cov_{Aaa} la covariance entre l'épistasie et l'effet additif au locus A et cov_{Baa} la covariance entre l'épistasie et l'effet additif au locus B.

On remarque que la variance génétique V_G est égale à : $V_G = C_A + C_B + C_{aa}$

Nous allons maintenant définir les coefficients d'apparentement entre un parent et son descendant. Considérons quatre évènements :

- α : Le parent et le descendant partagent le même allèle au locus A. La probabilité de α est noté $P(\alpha)$
- $\overline{\alpha}$: Le parent et le descendant ne partagent le même allèle au locus A. La probabilité de $\overline{\alpha}$ est note $P(\overline{\alpha})$. $P(\overline{\alpha}) = 1 - P(\alpha)$
- β : Le parent et le descendant partagent le même allèle au locus B. La probabilité de β est noté $P(\beta)$
- $\overline{\beta}$: Le parent et le descendant ne partagent pas le même allèle au locus B. La probabilité de β est noté $P(\overline{\beta})$. $P(\overline{\beta}) = 1 - P(\beta)$

Il existe donc 4 cas dans un modèle bilocus haploïde. Rappelons que les fréquences génotypiques sont égales aux fréquences gamétiques. La probabilité de chacun de ces quatre cas est :

$\Phi_{AB} = P(\alpha \cap \beta) = (1-r)/2$; $\Phi_A = P(\alpha \cap \overline{\beta}) = r/2$; $\Phi_B = P(\overline{\alpha} \cap \beta) = r/2$; $\Phi_\emptyset = P(\overline{\alpha} \cap \overline{\beta}) = (1-r)/2$

Où r est le taux de recombinaison.

Ces quatre probabilités correspondent aux types de descendants qui peuvent être obtenus pour un parent : un descendant haploïde ayant hérité de l'allèle parental aux deux locus A et B (Φ_{AB}), un descendant ayant hérité d'un seul allèle seulement (au locus A (Φ_A) ou au locus B (Φ_B)) ou un descendant n'ayant hérité d'aucun allèle ($\Phi_\emptyset$).

Pour calculer la covariance parent-enfant, il faut d'abord estimer sa valeur pour chacun des quatre types de descendants :

- Si le descendant a hérité des deux allèles, alors la covariance parent-enfant est égale à la variance génétique totale, V_G.
- Si le descendant n'a hérité que de l'allèle au locus A, alors elle est égale à la contribution du locus A, C_A.

- Si le descendant n'a hérité que de l'allèle au locus B, alors elle est égale à la contribution du locus B, C_B.
- Si le descendant n'a hérité d'aucun allèle, alors elle est nulle.

On trouve donc :

$$\begin{aligned}\mathrm{cov}_{PO} &= \Phi_A \mathrm{cov}_{PO,A} + \Phi_B \mathrm{cov}_{PO,B} + \Phi_{AB} \mathrm{cov}_{PO,AB} + \Phi_\emptyset \mathrm{cov}_{PO,\emptyset} \\ &= \Phi_A C_A + \Phi_B C_B + \Phi_{AB}(C_{aa} + C_A + C_B)\end{aligned}$$

$$\mathrm{cov}_{PO} = \frac{1}{2}V_A + \frac{1}{2}V_B + \frac{1-r}{2}V_{aa} + \mathrm{cov}_{AB} + \frac{2-r}{2}\mathrm{cov}_{Aaa} + \frac{2-r}{2}\mathrm{cov}_{Baa}$$

On peut remarquer que le taux de recombinaison peut agir à plusieurs niveau dans la covariance parents-enfants :

- De manière directe, puisqu'il intervient devant certains termes liés à l'épistasie. Ce taux de recombinaison peut augmenter ou diminuer la covariance parents-enfants, puisque cov_{Aaa} et cov_{Baa} peuvent être négatifs. Plus exactement, si $\mathrm{cov}_{Aaa} + \mathrm{cov}_{Baa} + V_{aa} > 0$, alors r diminuera la covariance parents-enfants. En revanche, si $\mathrm{cov}_{Aaa} + \mathrm{cov}_{Baa} + V_{aa} < 0$, alors r augmentera la covariance parents-enfants.
- Le taux de recombinaison a également un effet plus indirect. En effet, il agit sur le déséquilibre de liaison, qu'il a tendance à réduire au cours des générations. Dans une population de taille infinie, panmictique et non soumise à la sélection, on peut écrire que $D_g = (1-r)\, D_{g-1}$ où D_g est le déséquilibre de liaison à la génération g (Hill and Robertson, 1968). Or le déséquilibre de liaison a un effet sur les covariances et variances génétiques (cf partie II.4.3.). Plus précisément, il augmente la covariance entre effets additifs (cov_{AB}). Il peut augmenter ou diminuer les covariances entre effets additifs et épistasie (cov_{Aaa} et cov_{Baa}) et la variance épistatique en fonction des fréquences alléliques et des valeurs des effets génétiques.

En l'absence d'épistasie et de déséquilibre de liaison, on tombe sur un résultat connu :

$$\mathrm{cov}_{PO} = \frac{1}{2}V_A + \frac{1}{2}V_B$$

IV.1.2 Prédiction du progrès génétique à l'aide de la covariance parent-enfant

IV.1.2.1. Le phénotype moyen des descendants en fonction du phénotype moyen des parents

En génétique quantitative, le phénotype moyen des descendants est classiquement modélisé comme étant une fonction dépendant du phénotype des parents (Falconer and Mackay, 1996) :

$$P_{O,j} = bP_{P,j} + c \textbf{ (1)}$$

Où $P_{O,j}$ est le phénotype moyen des descendants d'une famille j et $P_{P,j}$ est celui des parents de la famille j. Le coefficient b peut être calculé comme étant (Falconer and Mackay, 1996):

$$b = 2\frac{\mathrm{cov}(P_O, P_P)}{V_{P,P}}$$

Avec $V_{P,P}$ la variance phénotypique des parents et $\mathrm{cov}(P_O, P_P)$, la covariance parents-enfants. Le facteur 2 est ajouté à la formule car la régression se fait non pas en fonction des phénotypes des parents, mais en fonction de la moyenne des parents. Par conséquent, la variance phénotypique de ceux-ci est divisée par deux.

Soient μ_P la moyenne phénotypique de la population parentale avant sélection, μ_S la moyenne phénotypique de la population parentale sélectionnée, μ_O la moyenne phénotypique des descendants de la population en l'absence de sélection, et μ_{OS} la moyenne phénotypique des descendants de la population sélectionnée. Remarquez que $\mu_O = b\mu_P + c$ et $\mu_{OS} = b\mu_S + c$.

Par convention, on fixe μ_P à 0. L'ordonnée à l'origine de l'équation **(1)** est donc égale à μ_O. On définit S, le différentiel de sélection. Il s'agit de la différence de moyennes entre les individus sélectionnés et l'ensemble de la population (sélection artificielle ou naturelle).

$$S = \mu_S - \mu_P$$

On définit R, la réponse à la sélection. C'est la différence phénotypique entre les descendants des individus sélectionnés, et les descendants des individus non sélectionnés :

$$R = \mu_{OS} - \mu_O = b(\mu_S - \mu_P) = bS$$

La **figure 18** illustre la fonction parent-enfant dans plusieurs cas (trait additif, trait épistatique, trait additif et épistatique).

IV.1.2.2. Prédiction de la réponse à la sélection en l'absence d'épistasie

En l'absence d'épistasie (Figure 18, première colonne), la covariance parents-enfants est égale à la moitié de la variance additive, qui est elle-même égale à :

$$\mathrm{V}_{add} = V_A + V_B + 2\mathrm{cov}_{AB}$$

Par conséquent, b correspond à l'héritabilité au sens strict : $b = h^2$, avec $h^2 = \frac{\mathrm{V}_{add}}{V_{P,P}}$.

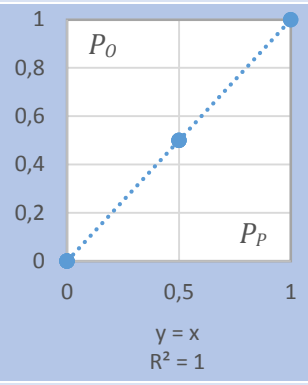
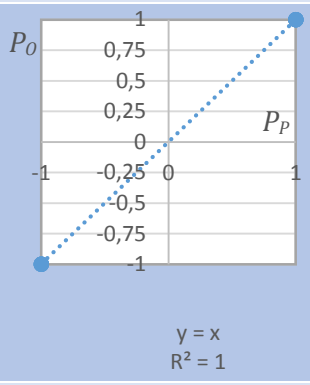
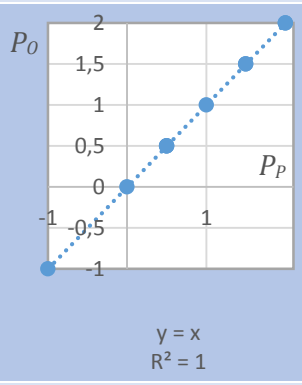
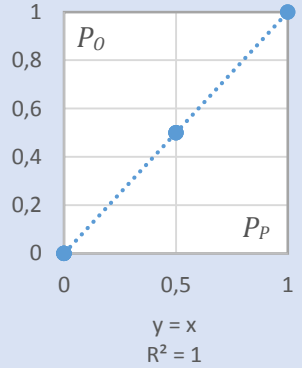
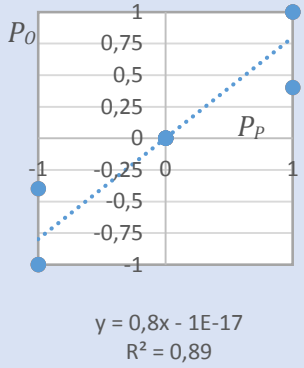
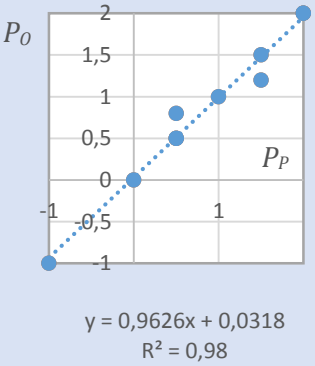
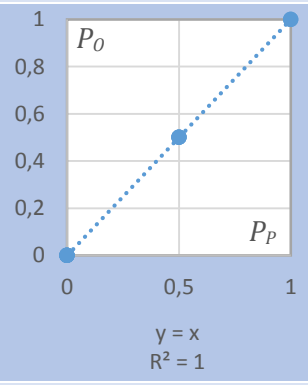
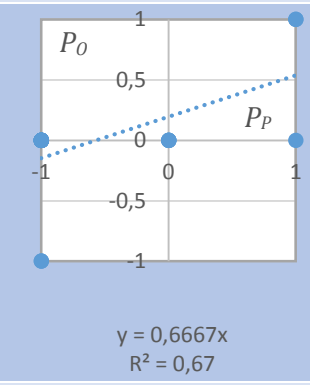
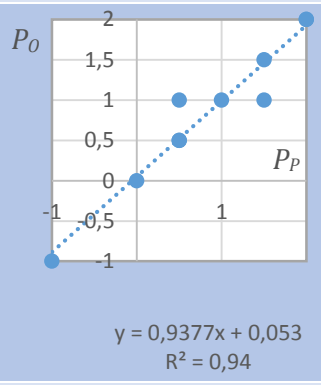
	Additivité	Epistasie	Additivité + épistasie
Valeurs génotypiques	$G_{AB} = 1$ $G_{Ab} = 1$ $G_{aB} = 0$ $G_{ab} = 0$	$G_{AB} = 1$ $G_{Ab} = -1$ $G_{aB} = -1$ $G_{ab} = 1$	$G_{AB} = 2$ $G_{Ab} = 0$ $G_{aB} = -1$ $G_{ab} = 1$
Effets génétiques (estimés avec le modèle UWR)	$\mu = 0$ $a_A = 1$ $a_B = 0$ $aa = 0$	$\mu = 0$ $a_A = 0$ $a_B = 0$ $aa = 1$	$\mu = 0$ $a_A = 1$ $a_B = 0$ $aa = 1$
$r = 0$			
$r = 0,3$			
$r = 0,5$			

Figure 18 : Phénotype des descendants (P_O) fonction du phénotype des parents (P_P)
r = taux de recombinaison
Variance environnementale $V_E = 0$ dans cet exemple
Déséquilibre de liaison nul

En moyenne, les parents transmettront chacun de leur allèle à la fréquence ½. Chaque allèle apportera un effet indépendant sur le trait étudié. La fonction $P_{O,j} = f(P_{P,j})$ est donc une fonction linéaire qui permet de permettre de prédire le progrès génétique en présence de sélection, à l'aide de l'équation du sélectionneur (Lush, 1937) :

$$R = h^2 S$$

Par ailleurs, en l'absence de sélection, on suppose que les fréquences alléliques ne sont pas modifiées (on suppose l'absence de dérive ou de migration). Or, dans un cas purement additif, seuls les allèles sont porteurs d'un effet génétique. Autrement dit, en l'absence de sélection et d'épistasie, la moyenne phénotypique d'une population ne va pas changer d'une génération à l'autre. Par conséquent $\mu_O = \mu_P = 0$. L'ordonnée à l'origine de la fonction linéaire défini par l'équation **(1)** est nulle.

De plus, $R = \mu_{OS} - \mu_O = \mu_{OS} - \mu_P$. Par conséquent, ici R ne représente pas seulement la différence de valeur phénotypique entre les descendants d'une population sélectionnée et ceux d'une population non sélectionnée. R est aussi le changement de moyenne phénotypique entre la génération parentale et les descendants issus des individus sélectionnés.

Une illustration de la sélection dans le cas additif est donnée en **figure 19**.

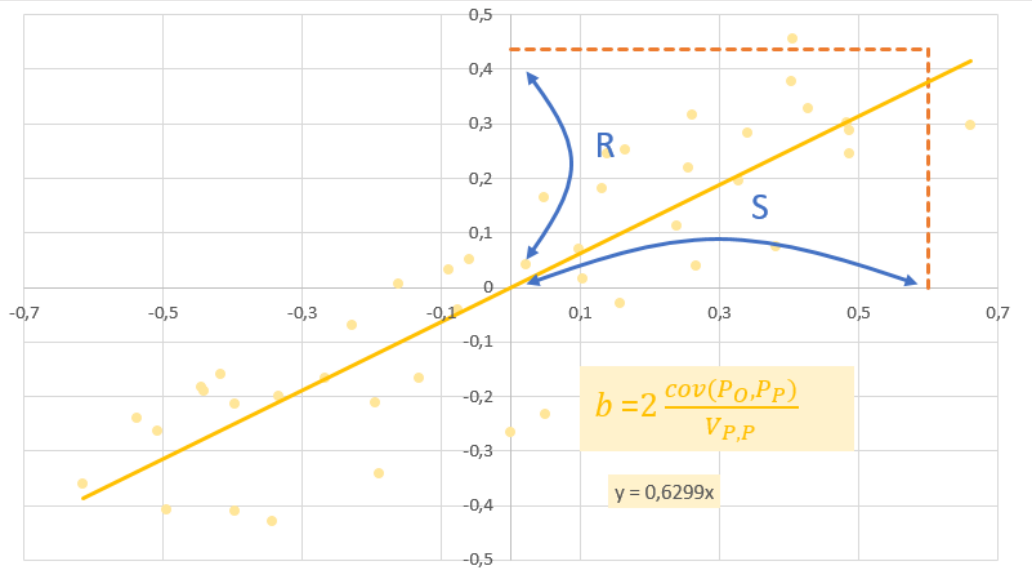


Figure 19 : Réponse à la sélection pour un caractère additif (équation du sélectionneur)
En abscisse : phénotype des parents (P_P)
En ordonné: phénotype des descendants (P_O)
S = différentiel de sélection
R = réponse à la sélection

IV.1.2.3. Prédiction de la réponse à la sélection en l'absence d'épistasie

En présence d'épistasie, la moyenne phénotypique d'une population dépend des fréquences des génotypes. Elle ne dépend donc pas seulement des fréquences alléliques, mais également du

déséquilibre de liaison. Par conséquent, en présence de déséquilibre de liaison, la moyenne d'une population change d'une génération à l'autre, même en l'absence de sélection. En effet, ce déséquilibre de liaison sera réduit par la recombinaison. Ainsi, l'ordonnée à l'origine,μ_O, ne sera pas nulle s'il y a du déséquilibre de liaison.

Par ailleurs, en présence d'épistasie, $P_{O,j} = f(P_{P,j})$ n'est plus une fonction, puisque pour un $P_{P,j}$ donné, plusieurs $P_{O,j}$ sont possibles. En effet, prenons un cas où le trait n'est qu'épistatique (Figure 18, deuxième colonne). Par exemple, $G_{AB} = 1$, $G_{Ab} = -1$, $G_{aB} = -1$ et $G_{ab} = 1$. Soit r le taux de recombinaison. Si un individu AB se reproduit avec un individu également AB, alors ils n'auront que des descendants AB. Par conséquent, dans ce cas, et $P_{P,j} = 1$ et $P_{O,j} = 1$. En revanche, si un individu AB se reproduit avec un individu ab, une proportion $1 - r$ des descendants aura un phénotype moyen égal à 1 ($G_{AB} = 1$et $G_{ab} = 1$) et une proportion r aura un phénotype moyen égal à -1 ($G_{Ab} = -1$ et $G_{aB} = -1$). Par conséquent, le phénotype moyen des descendants sera égal à $1 - 2r$. Donc, $P_{P,j} = 1$ et $P_{O,j} = 1 - 2r$. Dans cet exemple, deux descendances moyennes sont donc possibles pour des parents dont le phénotype moyen est égal à $P_{P,j} = 1$.

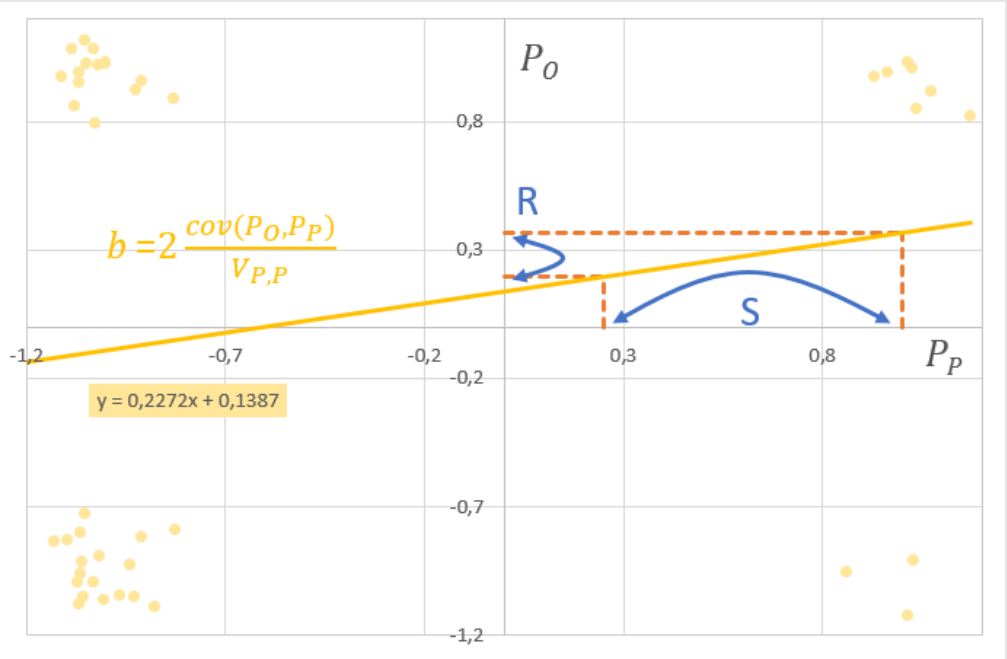


Figure 20 : Réponse à la sélection pour un caractère épistatique
En abscisse : phénotype des parents (P_P)
En ordonné: phénotype des descendants (P_O)
S = différentiel de sélection
R = réponse à la sélection

Néanmoins, même si $P_{O,j} = f(P_{P,j})$ n'est pas une fonction en présence d'épistasie, il est possible de déterminer une courbe de tendance sous forme de fonction linéaire. Cette fonction permet de prédire le phénotype moyen des descendants. La pente est égale, comme précédemment, à $b = 2\frac{cov(P_O,P_P)}{V_{P,P}}$.

Il est donc possible, même en présence d'épistasie, de prédire la réponse à la sélection :

$$R = 2\frac{cov(P_O, P_P)}{V_{P,P}}S$$

La réponse à la sélection pour un caractère purement épistatique est montrée en **figure 20**. Elle illustre que pour un caractère épistatique, la prédiction de la réponse à la sélection peut se faire de la même manière que pour un caractère additif. Il manque néanmoins une connaissance sur l'ordonnée à l'origine.

IV.2 Prédire l'évolution à long terme avec l'épistasie fonctionnelle

Si le calcul de variances génétiques est essentiel pour prédire la réponse à la sélection, les modèles fonctionnels peuvent aussi aider à la compréhension de l'évolution d'un trait soumis à épistasie. Contrairement aux variances génétique, qui changent en fonction des fréquences alléliques et sont donc prédictives de l'évolution à court terme, les effets génétiques fonctionnels peuvent permettre de comprendre l'évolution sur une échelle de temps beaucoup plus longue (Weinreich et al., 2005).

IV.2.2 Théorie des paysages adaptatifs

Pour mieux comprendre comment des effets fonctionnels peuvent avoir un effet sur l'évolution d'une population, il faut faire appel à la théorie des paysages adaptatifs, développée par Wright (Wright, 1932). Il existe deux types de paysages adaptatifs, qui sont représentés en **figure 21** :

- le paysage des valeurs sélectives génotypiques (Figure 21.A). Chaque haploïde possible contenant L locus peut être représenté dans un espace discret à L+1 dimensions. Les L premières dimensions représentent un locus. La valeur sélective est représentée dans la dimension L+1. Les points adjacents sont des génotypes haploïdes n'ayant qu'un allèle de différence.
- le paysage de la valeur sélective moyenne de la population (Figure 21.B-C). Chaque composition génétique possible d'une population, pour L locus peut être représentée dans un espace continu à L+1 dimensions. Les L premières dimensions représentent la fréquence allélique à un locus. La valeur sélective est représentée dans la dimension L+1.

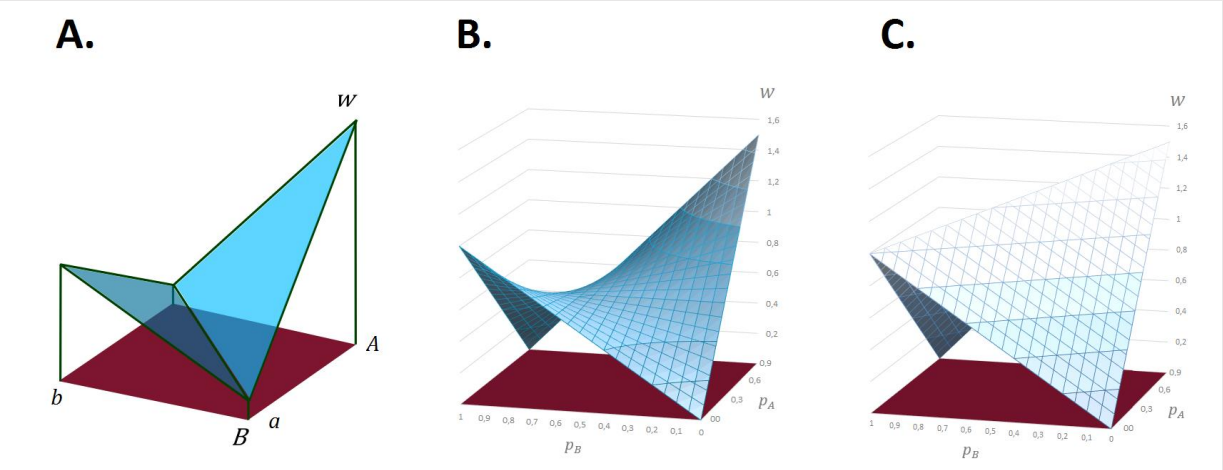


Figure 21 : Paysages adaptatifs quand la valeur sélective est soumise à de l'épistasie de signe réciproque

D' : déséquilibre de liaison normalisé

- **A.** Paysage des valeurs sélectives génotypiques.
  En abscisse : absence/présence de l'allèle B (absence de l'allèle B = présence de l'allèle b)
  En ordonnée : absence/présence de l'allèle A (absence de l'allèle A = présence de l'allèle a)
  En cote : la valeur sélective du génotype
- **B.** Paysage de la valeur sélective moyenne de la population, pour un déséquilibre de liaison nul (D' = 0)
  En abscisse : fréquence de l'allèle B (p_B)
  En ordonnée : fréquence de l'allèle A (p_A)
  En cote : la valeur sélective de la population
- **C.** Paysage de la valeur sélective moyenne de la population, pour un déséquilibre de liaison maximal (D' = 1 ; $D' = D/D_{max}$ avec $D_{max} = \min(p_A p_b, p_a p_B)$)

Les deux types de paysages adaptatifs se ressemblent, le paysage de la valeur sélective moyenne de la population étant une version continue du paysage des valeurs sélectives génotypiques. Il peut donc être perçu comme simplement remplissant les interstices du paysage des valeurs sélectives génotypiques. Cependant, les choses se compliquent quand la valeur sélective est soumise à de l'épistasie. En effet, dans ce cas ce ne sont pas les fréquences alléliques seules qui déterminent la moyenne de la valeur sélective d'une population, mais les fréquences génotypiques, et donc le déséquilibre de liaison. Il existe ainsi plusieurs paysages de la valeur sélective moyenne de la population en présence d'épistasie (Figure 21.B-C) (Weinreich et al., 2005). Par conséquent, il est plus difficile de prédire l'évolution de la valeur sélective moyenne de la population en présence d'épistasie.

Une population ne peut se déplacer sur le paysage adaptatif qu'en montant en altitude, c'est-à-dire qu'en augmentant sa valeur sélective. Cependant, l'épistasie de signe réciproque peut induire des vallées et des pics dans le paysage adaptatif (Poelwijk et al., 2011), contraignant la capacité des populations à atteindre (c'est-à-dire, fixer) le génotype avec la plus haute valeur sélective. Un exemple est par exemple donné en Figure 21.B, où l'on voit qu'il y a deux pics : l'un quand p_A et p_B sont égales

à 1 (pic optimal), correspondant à la fixation du génotype AB ; l'autre quand p_A et p_B sont nulles (pic sous-optimal) correspondant à la fixation du génotype ab. On parle alors de paysage adaptatif rugueux. Le problème avec ce type de paysage, est qu'une population risque d'escalader un pic sous-optimal et d'y restée coincée.

Néanmoins, il est possible de traverser ces vallées adaptatives. D'abord, dans de petites populations isolées, la dérive peut devenir plus forte que la sélection, et permettre à la population de passer par des régions de faible valeur sélective. Cette théorie a été avancée dès 1932 par Wright (*Shifting balance theory*) (Wright, 1932). L'importance de ce phénomène reste controversé (Coyne et al., 1997, 2000), mais des preuves empiriques démontrent qu'il existe (Chouteau and Angers, 2012; Mallet, 2010; Wade and Goodnight, 1991). Cependant, même dans des populations de grande taille, cette traversée est possible (Weinreich and Chao, 2005; Weissman et al., 2009). En effet, quand des mutations individuellement délétères (mais bénéfiques quand elles sont associées) apparaissent par mutation, il est peu probable qu'elles soient immédiatement éliminées à moins d'être létales : elles persistent quelques générations. Pendant ce temps, une mutation à l'autre locus peut apparaître, générant alors le génotype le plus optimal.

La recombinaison affecte ce processus. D'un côté, elle pourrait l'accélérer, en générant la combinaison allélique optimale. L'efficacité de ce processus va dépendre du polymorphisme. S'il est très faible, il n'aura quasiment aucun effet (Weinreich and Chao, 2005). D'un autre côté, elle peut aussi détruire ce génotype optimal et réduire son temps de fixation. Cela peut être perçu graphiquement, notamment sur les figures 21.B-C. L'apparition d'un génome optimal, et sa sélection, va générer un déséquilibre de liaison fortement positif. Ce déséquilibre de liaison peut créer un « pont » dans le paysage adaptatif, comme montré en figure 21.C. Cependant, la recombinaison va détruire ce pont, en réduisant le déséquilibre de liaison (figure 21.B). L'effet de la recombinaison sur la traversée des vallées adaptatives a fait l'objet de nombreuses études. Il a été démontré dans plusieurs situations (pics de même hauteur, présence/absence de mutation, présence/absence de dérive) que la recombinaison avait tendance à ralentir, voire empêcher, la traversées des vallées (Jain, 2010; Stephan, 1996; de Visser et al., 2009). Cependant, dans certains cas, un faible taux de recombinaison pourrait accélérer la traversée, en permettant de créer plus de génotypes optimaux qu'il n'en détruit (Weinreich and Chao, 2005). Plusieurs auteurs ont également cherché à déterminer le taux de recombinaison critique, au-delà duquel la traversée de la vallée n'est pas possible et la population reste coincée sur le pic sous-optimal, dans diverses situations (Bodmer and Felsenstein, 1967; Bürger, 1989; Feldman, 1971; Higgs, 1998; Park and Krug, 2011; Rutschman, 1994). Ce taux de recombinaison critique dépend des valeurs sélectives de chacun des génotypes et éventuellement du taux de mutation.

Nous avons voulu comprendre l'impact de l'épistasie de signe réciproque dans une population de pre-breeding.

IV.2.3 Epistasie et sélection par troncation

Dans les populations de pre-breeding ou de management dynamique, une légère sélection anthropique est parfois effectuée. Il s'agit d'éliminer un petit pourcentage d'individus ayant des traits trop défavorables (trop grande taille, sensibilité à des maladies). La sélection est très faible pour ne pas réduire la diversité génétique en créant un goulot d'étranglement. Nous avons considéré que nous pouvions modéliser cette sélection comme une sélection par troncation : chaque année, une fraction p de la population est gardée. Cette fraction p ($p \in$]0,1]) a été considérée comme à peu près constante. Dans la pratique, p est élevée (>0.90), mais, pour notre étude, nous avons modélisé des valeurs de p balayant tout l'intervalle]0,1].

Comme cette fraction p est constante durant le temps d'une simulation, les valeurs sélectives de chaque génotype vont varier au cours des générations. Soit w_{ij}^n la valeur sélective relative du génotype ij à la génération n. Elle est égale à :

$$w_{ij}^n = \frac{p_{ij}^{n\,\prime}}{p_{ij}^n}$$

Où p_{ij}^n est la fréquence du génotype ij à la génération n avant la sélection, et $p_{ij}^{n\,\prime}$ la fréquence du génotype ij à la génération n après la sélection.

Par convention, on suppose que le sélectionneur choisit en priorité les individus qui ont le plus grand phénotype. Les valeurs génotypiques ne varient pas, contrairement aux valeurs sélectives relatives. Cependant les valeurs sélectives relatives sont toujours ordonnées dans le même ordre que les valeurs génotypiques. Par exemple, si $G_{AB} > G_{Ab} > G_{aB} > G_{ab}$, alors $w_{AB}^n \geq w_{Ab}^n \geq w_{aB}^n \geq w_{ab}^n$. En effet, si AB est le meilleur génotype, il va être sélectionné en priorité. Si $p_{AB}^n > p$, alors le sélectionneur ne va choisir que des génotypes AB pour renouveler la population. A l'inverse, si $p_{AB}^n < p$, le sélectionneur va ensuite choisir le deuxième meilleur génotype, et ainsi de suite, jusqu'à obtenir une fraction p d'individus. La **figure 22** détaille ce processus de sélection et la variabilité des valeurs sélectives qu'il engendre, dans le cas de trois génotypes, dans le cas où il n'y a pas d'effet environnementaux.

En effet, supposons trois génotypes, nommés 1,2 et 3 et dont les valeurs sélectives suivent l'inégalité $G_1 > G_2 > G_3$. Trois cas de figure se présentent :

- Soit $p_1^n > p$ (Figure 22.A). Dans ce cas, seul le génotype 1 est sélectionné. Sa valeur sélective w_1^n est égale à $\frac{p_1^{n\,\prime}}{p_1^n} = \frac{1}{p_1^n} \geq 1$. La valeur sélective des autres génotypes (non sélectionnés) est

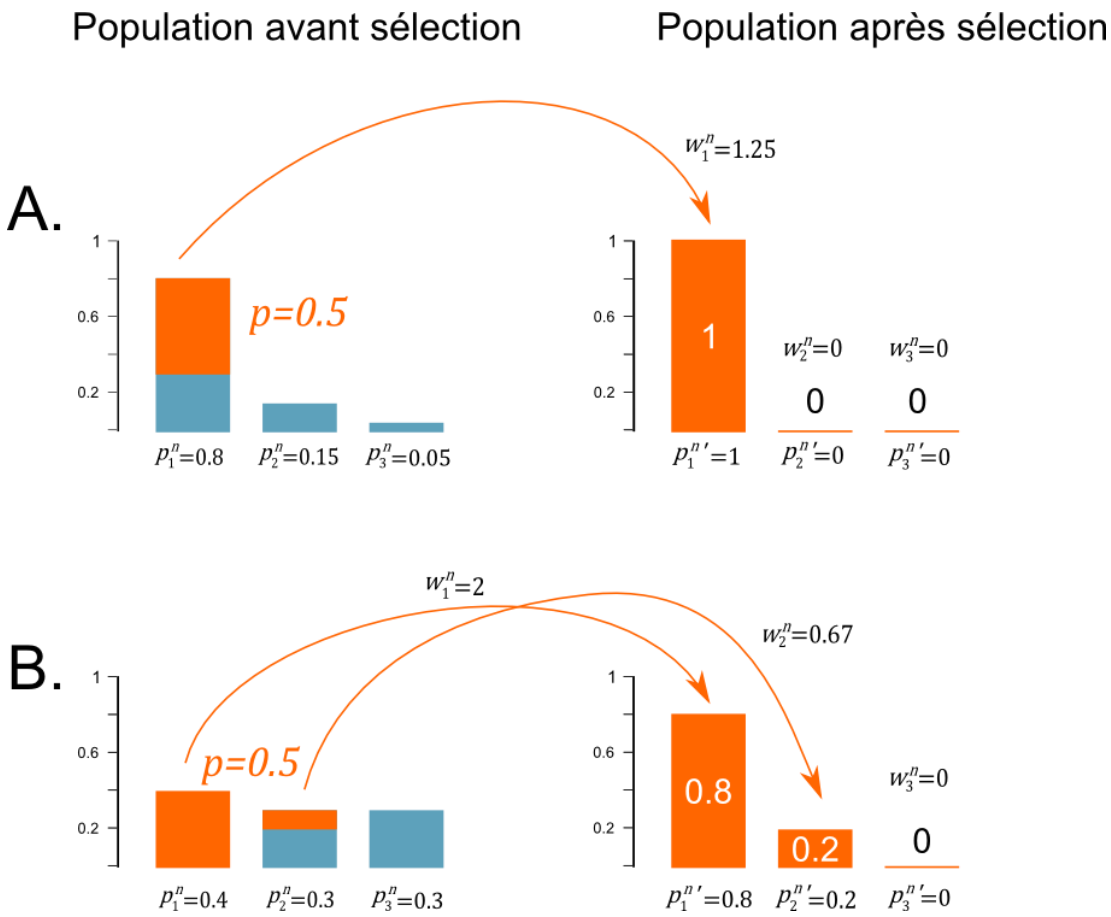


Figure XX3: Sélection par troncation et variabilité des valeurs sélectives
Soient trois génotypes : génotypes 1,2 et 3.
Le génotype 1 est celui qui a la meilleure valeur génotypique, suivi du génotype 2 et enfin du génotype 3 : $G_1 > G_2 > G_3$.
La fraction p est fixée dans cet exemple à 0,5.
n = génération

- Cas où $p_1^n > p$. Seul le génotype 1 est sélectionné. Sa valeur sélective w_1^n est égale à $\frac{p_1^{n\prime}}{p_1^n} = \frac{1}{0.8} = 1.25$. La valeur sélective des autres génotypes (non sélectionnés) est nulle.
- Cas où $p_1^n < p$. Tous les individus porteurs du génotype 1 sont sélectionnés, puis une partie des individus portant le génotype 2, afin qu'une fraction p de la population soit sélectionnée. Le génotype 3 n'est pas sélectionné.
  $w_1^n = \frac{0.8}{0.4} = 2$; $w_2^n = \frac{0.2}{0.3} = 2/3$; $w_3^n = 0$

nulle. Donc $w_1^n > w_2^n = w_3^n$

- Soit $p_1^n < p$ et $p_1^n + p_2^n > p$ (Figure 22.B). Dans ce cas, tous les individus du génotype 1 sont sélectionnés. Par conséquent $p_1^{n\prime} = \frac{p_1^n}{p}$. Donc w_1^n est égale à $\frac{p_1^{n\prime}}{p_1^n} = \frac{1}{p}$. La fréquence du génotype 2 après sélection est égale à $\frac{p-p_1^n}{p}$, et par conséquent $w_2^n = \frac{p-p_1^n}{p \cdot p_2^n}$.

  Or $p_1^n + p_2^n > p$ donc :

  $p_2^n > p - p_1^n$

  $\leftrightarrow 1 > \dfrac{p - p_1^n}{p_2^n}$

  $\leftrightarrow \dfrac{1}{p} > \dfrac{p - p_1^n}{p \cdot p_2^n}$

  $\leftrightarrow w_1^n > w_2^n$

  Par ailleurs, le génotype 3 n'a pas été sélectionné, donc $w_3^n = 0$.

  Par conséquent $w_1^n > w_2^n > w_3^n$

- Enfin, la dernière situation possible est : $p_1^n < p$ et $p_1^n + p_2^n < p$. Dans ce cas, tous les individus du génotype 1 et tous les individus du génotype 2 sont sélectionnés. Par conséquent $p_1^{n\prime} = \frac{p_1^n}{p}$ et $p_2^{n\prime} = \frac{p_2^n}{p}$. Donc w_1^n est égale à $\frac{p_1^{n\prime}}{p_1^n} = \frac{1}{p}$ et w_2^n à $\frac{p_2^{n\prime}}{p_2^n} = \frac{1}{p}$. La fréquence du génotype 3 après sélection est égale à $\frac{p-p_1^n-p_2^n}{p}$, et par conséquent $w_3^n = \frac{p-p_1^n-p_2^n}{p \cdot p_3^n}$.

  Or $p_1^n + p_2^n + p_3^n > p$ donc :

  $p_3^n > p - p_1^n - p_2^n$

  $\leftrightarrow 1 > \dfrac{p - p_1^n - p_2^n}{p_3^n}$

  $\leftrightarrow \dfrac{1}{p} > \dfrac{p - p_1^n}{p \cdot p_3^n}$

  Donc $w_1^n = w_2^n > w_3^n$

Le raisonnement peut s'étendre à un nombre plus grand de génotypes. Sachant que les valeurs sélectives relatives suivent le même ordre que les valeurs génotypiques, on peut déterminer le génotype qui sera fixé, dans une population de taille infinie (sans dérive). Dans un paysage adaptatif plat (en l'absence d'épistasie de signe réciproque) le génotype qui sera fixé sera simplement celui ayant la plus forte valeur génotypique. En présence d'épistasie de signe réciproque, les choses se compliquent, puisque le paysage de la valeur sélective moyenne de la population devient rugueux 2.2), et la population peut rester coincée sur un pic sous-optimal.

Nous avons décidé d'étudier en détails, à l'aide de calculs numériques et analytiques, l'évolution d'une population de pre-breeding sélectionnée pour un trait soumis à de l'épistasie de signe réciproque. Cette épistasie de signe réciproque est à l'origine de phénomènes comme les incompatibilités de Dobzhansky-Müller ou la dépression hybride, qui a ont observés par exemple chez le blé tendre (Hermsen, 1963a, 1963b; Heyne et al., 1943; Iehisa et al., 2014; Rousselle et al., 2011)..

Le génotype optimal était AB tandis que ab constituait un pic sous-optimal. Puisque le trait phénotypique subissait une sélection de troncation telle que celle décrite ci-dessus, les valeurs sélectives génotypiques étaient également soumises à de l'épistasie de signe réciproque, mais étaient variables à cause du processus de sélection. Le modèle utilisé était bi-locus et haploïde, car seul l'épistasie additive x additive nous intéressait. La population initiale ne contenait pas le génotype le plus favorable (AB), mais comprenait au moins chacun des allèles (A, a, B et b). Elle n'était donc pas située initialement sur le pic sous-optimal, mais à proximité dans la vallée adaptative. Le taux de mutation était nul, et donc la formation du génotype AB n'était due qu'à la recombinaison.

Nos résultats nous ont montré qu'un fort taux de recombinaison empêchait la fixation du génotype AB. Cependant, un minimum de recombinaison est nécessaire pour créer le génotype AB. Par ailleurs, de faibles taux de recombinaison peuvent accélérer sa fixation dans la population. Si un trait semble fortement soumis à de l'épistasie de signe réciproque, nous recommandons pour les sélectionneurs de favoriser dans les premières générations la recombinaison (par exemple, en permettant un fort taux d'allogamie à l'aide d'un gène de stérilité mâle), puis d'effectuer une phase de sélection sans recombinaison une fois que le déséquilibre de liaison a été fortement réduit.

Les résultats détaillés sont présentés dans l'article qui suit : *Reciprocal sign epistasis and truncation selection: when is recombination favorable in a pre-breeding program with a selfing species?* Les fichiers supplémentaires de l'article sont présentés en **Annexe VII.6**.

Cet article a été soumis à Heredity. Il n'a pas été accepté pour publication, en dehors du scope de la revue. Les commentaires des reviewers sont donnés en **Annexe VII.7**. Ce papier sera resoumis après légères modification au plus tôt.

IV.3 Reciprocal sign epistasis and truncation selection: when is recombination favorable in a pre-breeding program with a selfing species?

Authors: Constance Vagne[a], Jacques David[a], Muriel Tavaud[a], Bénédicte Fontez[b]

[a]Montpellier SupAgro,

UMR AGAP,

F-34060 Montpellier, France

constance.vagne@supagro.inra.fr

muriel.tavaud@supagro.inra.fr

jacques.david@supagro.inra.fr

[b]Montpellier SupAgro,

UMR MISTEA,

F-34060 Montpellier, France

benedicte.fontez@supagro.inra.fr

Corresponding author:

Constance Vagne

Montpellier SupAgro,

UMR AGAP, Bâtiment 21, bureau 217

F-34060 Montpellier, France

+33609223119

Running title: Reciprocal sign epistasis and recombination

Word count: 6745

Since the dawn of agriculture, humans have applied artificial selection on traits of interest, regardless of their genetic architecture. Yet, still today, most models used to study and streamline this process overlook genetic interactions. In this study, we determined the conditions in which a target genotype

can be fixed when truncation selection is applied on an epistatic trait. Previous studies have shown that reciprocal sign epistasis with two fitness peaks of unequal height involves multiple equilibrium states, i.e. below one critical parameter value, such as a critical recombination rate, one genotype may be fixed, and above it, another one may be fixed. Using a haploid bi-locus model, we identified which genotype would be fixed, and how quickly, in an infinite population selected for a phenotypic trait subject to reciprocal sign epistasis with unequal peak heights, depending on two criteria: the recombination rate and percentage of selected individuals. The critical parameter values at which bistability sets in, were also calculated. These results were complemented by stochastic simulations in finite populations. Our results confirmed that, in the case of fitness under reciprocal sign epistasis, high recombination rates induce blockage at the local optimum or attainment of an equilibrium state between the two peaks. However, if linkage disequilibrium is negative in the initial population, recombination is necessary to create the most favorable genotype. Therefore, in this case, reciprocal sign epistasis favors non-null recombination rates, particularly if selection is intense. Based on our results, recommendations are put forward for the monitoring of pre-breeding populations.

Keywords: reciprocal sign epistasis, recombination, truncation selection, pre-breeding, bistability

IV.3.1 Introduction

The near ubiquity of recombination is one of the greatest enigmas in evolutionary biology (Otto, 2009). Asexual reproduction does not involve recombination, only sexual reproduction does, while selfing reduces its efficiency and also heterozygosity. Indeed, the effective recombination rate (Nordborg, 2000), which measures the recombination rate between heteroallelic chromosomal regions, is the relevant evolution parameter. It corresponds to the probability of producing one recombinant per generation.

It was predicted that negative epistasis promotes recombination (Barton, 1995; Charlesworth, 1990; Kondrashov, 1988), but experimental evidence indicates that negative epistasis is not widespread enough to be the main explanation for the evolution of recombination (Rice, 2002). In contrast, there is substantial experimental evidence of sign epistasis in nature (Colegrave and Buckling, 2005; Lalić and Elena, 2012; Poon and Chao, 2005; Schenk et al., 2013; Weinreich et al., 2005, 2006). Sign epistasis is when the effect of an allele can have a different sign depending on the genetic background and may involve 'compensatory mutations' since two unfavorable mutations may be beneficial when they are associated. Reciprocal sign epistasis is a special case of sign epistasis where the genotype values of AB and ab are superior to those of aB and Ab in a bilocus haploid model. Note that fitness is often regarded as a phenotype, and many studies are focused on fitness epistasis.

Sign epistasis could disadvantage recombination (Otto and Feldman, 1997). Indeed, de Visser (2009) found that sex (and then recombination) had a general disadvantage, i.e. breaking up favorable allelic combinations and thus limiting the fixation of the fittest genotypes. Recombination was only favourable for certain values of some parameters (fitness, mutation rate and population size). This suggests that, under reciprocal sign epistasis, fixation of the best genotype depends on key parameters. We addressed this field of study in the context of pre-breeding populations, which can be monitored via these key parameters. The term 'pre-breeding' refers to the transfer of genes from related wild ancestors or from ancient varieties to breeding material (FAO, 1996). Pre-breeding populations are created to make effective use of genetic resources and consist of a broad genetic base

evolving under slight recurrent selection for traits of agronomic interest. The evolution of these populations can be complex because accessions constituting it diverged over a period of a few hundred to a few thousand years, and may have developed particular allelic combinations adapted to specific environmental conditions. Breaking up these combinations may have different evolutionary consequences. In autogamous plants, breeders can use sterility genes to increase the outcrossing rate and thus improve the efficiency of recombination between genotypes (Enjalbert et al., 2011; Henry et al., 1991; Porcher et al., 2004; Suneson, 1956). When monitoring a pre-breeding population of a predominantly selfing species, should a breeder promote high recombination rates, notably by increasing the outcrossing rate via a male sterility gene, in order to improve a phenotypic trait, or should it be kept at a low or medium level?

Truncation selection is usually applied in pre-breeding populations. Indeed, the breeder maintains the population at a roughly constant demographic size, while eliminating a small percentage of poor individuals. Strict truncation selection never occurs in nature (Crow and Kimura, 1979). However, quasi-truncation selection may occur, notably when resources are scarce. In this case, only a limited number of individuals, i.e. the most efficient ones, would survive and reproduce. Under truncation selection with a constant demographic size, a proportion p of individuals is selected for reproduction, regardless of the genotype frequencies in the population. The probability that a genotype will be selected thus depends on the genotype frequencies in the population. Therefore genotype fitness values vary over time in this case. However, as the population size remains roughly constant, the average fitness of the population actually does not change, as fitness is the ratio of the number of progeny to that of the parents. Note that the opposite situation generally occurs in natural populations, which are not limited by resources or by the size of a habitat, and whose demographic size can increase—genotype fitness values are constant, but the mean fitness of the population evolves to become increasingly strong.

Some authors have already investigated the role of recombination in cases where the genotype fitness values are constant and subjected to reciprocal sign epistasis. Eshel and Feldman (1970) noted that

sexual reproduction can prevent a population from fixing the fittest genotype, contrary to asexual reproduction. Several authors have calculated the critical value of the recombination rate, r_c, when the fitness value of AB is superior to that of ab ($W_{AB} > W_{ab}$). A population whose recombination rate is above this critical value will fix the ab genotype, whereas AB will become fixed if r is inferior to r_c. More precisely, Park and Krug (2011), Rutschman (1994) and Weinreich *et al.* (2005) expressed a critical recombination rate r_c as a function of the genotype fitness values for different population models (population of infinite or finite size, null or positive mutation rate, etc.). In addition, Higgs (1998), Jain (2010), Kimura (1985), Stephan (1996) and Weinreich and Chao (2005) calculated the expected time needed to fix AB in a population where there were initially only ab individuals. They all found that recombination increases this fixation time because it breaks up favorable allelic combinations.

Here we propose a method to determine the effect of recombination in the presence of reciprocal sign epistasis in the case of a population subjected to selection by truncation with varying genotype fitness values. What is the effect of recombination in this kind of population? When fitness values are not constant, does recombination improve the population mean of a selected trait subjected to reciprocal sign epistasis and allow fixation of the most effective genotype? If there is a critical value of the recombination rate, what is it?

We aimed at determining which genotype becomes fixed and how quickly, according to the recombination rate r, the percentage of individuals selected p and the initial genotype frequencies. We determined the evolutionary pathways that a population could take. This step was used to estimate the different critical values of the studied parameters (percentage of selected individuals p and recombination rate r), below and above which a genotype is fixed. The critical values are expressed in terms of genotype frequencies, r and p. The equations used to model the evolution of the population or to estimate critical values are easy to implement. Finally, based on these results, we put forward recommendations for the monitoring of pre-breeding populations in terms of outcrossing rate and selection.

IV.3.2 Materials and Methods

A three-step process was applied to gain insight into the evolution of a population selected for a trait subject to reciprocal sign epistasis. First we identified the different possible population configurations. We established the conditions for moving from one configuration to another and determined the equilibrium states. The critical values that determine whether the population will follow a specific evolutionary path were also calculated. Secondly, times to fixation were estimated by numerical calculations in random mating populations of infinite size. Finally, simulations of finite populations enabled us to investigate the role of genetic drift. Mutation was neglected in all the cases. We applied the simplifying assumption that all genotypes have the same fertility and thus that each parent generates the same number of offspring. Selection was performed on the parents.

IV.3.2.1 Model

We derived equations in a haploid, bi-locus and biallelic model tailored for describing selfing diploid species. Consider two loci: locus A with alleles A and a having frequencies p_A and 1-p_A, respectively; and locus B with alleles B and b with frequencies p_B and 1-p_B, respectively. The genotypes and their frequencies are noted in **Table 1**. A haploid model was chosen because it is a simpler framework for deriving equations.

Table 1: Genotypes and their frequencies in a haploid population, in the case of bi-locus and biallelic model

	a	***A***
b	ab	Ab
	$p_{ab} = (1-p_A)(1-p_B) + D$	$p_{Ab} = p_A \cdot (1-p_B) - D$
B	aB	AB
	$p_{aB} = (1-p_A) \cdot p_B - D$	$p_{AB} = p_A \cdot p_B + D$

D: linkage disequilibrium coefficient; p_A = frequency of allele A; p_B = frequency of allele B; p_{AB}= frequency of genotype AB; p_{ab}= frequency of genotype ab; p_{aB} = frequency of genotype aB; p_{Ab}= frequency of genotype aB

p_{ij}^n denotes the genotype frequency of genotype ij at generation n before selection, $p_{ij}^{n\prime}$ represents the genotype frequency of genotype ij at generation n after selection, and also $p_H^n = p_{aB}^n + p_{Ab}^n$. Finally, D^n and $D^{n\prime}$ represent the linkage disequilibrium values at generation n before and after selection, respectively. Genotype values are the phenotypic values that a genotype would produce in

the absence of environmental effects. The genotype values of genotypes AB, ab, aB and Ab were respectively set as: $G_{AB} = 1 + x$, $G_{ab} = 1$ and $G_{aB} = G_{Ab} = 1 - y$, with x and $y > 0$, such that $G_{AB} > G_{ab} > G_{aB} = G_{Ab}$.

We used analytical and numerical approaches to analyze the evolution of a pre-breeding population subject to truncation selection. The modeling included two main steps in the reproduction of the population:

- **Selection of parents**: directional selection by truncation was chosen. A percentage p ($p \in]0,1]$) of individuals with the highest phenotypic values was selected as parents for the following generation. As breeders preferably select individuals with the strongest genotype, they first select individuals having the genotype AB. Then, if more individuals are needed to create the parental population, individuals carrying the ab genotype are selected, and finally individuals with the Ab or aB genotype which have the lowest phenotypes. The relative fitness values are thus subject to reciprocal sign epistasis: $w_{AB} > w_{ab} > w_{aB} = w_{Ab}$ ($w_{ij}^{n} = p_{ij}^{n}{}'/p_{ij}^{n}$).
- **Mating and recombination**: mating consisted of combining two gametes drawn at random from a pair of parents, thus creating a diploid individual. This diploid individual underwent meiosis, and the resulting gametes were considered as the haploid offspring.

IV.3.2.2 Step 1: Identification of the different population configurations

An analytical approach was used to deterministically establish the evolutionary path that the population would follow, depending on r, p and genotype frequencies. Analytical calculations were done to determine if a genotype would become fixed, and which one, but they did not provide any estimate of the time to fixation. Genotype frequencies at generation $n + 1$ depend on the genotype frequencies at the previous generation n, the proportion of individuals selected as parents at each generation (p) and the recombination rate (r). Genotype frequencies are first modified by selection and then changed by recombination. Mating actually does not change the allelic frequencies, but instead reduces the linkage disequilibrium via recombination: $D^{n+1} = (1 - r)D^{n}$ (Lewontin and Kojima, 1960).

At any generation, the population is found in one the possible configurations. The selection and mating processes are shown in **Figure 1**, whereas the selection process is illustrated in **Figure 2**. We hypothesized that all alleles were present in the population at the initial state.

The different configurations and the expected genotype frequencies are:

- **Configuration 1**: the proportion of selected individuals p exceeds the proportion of the genotypes with the highest genotype value, AB and ab: $p_{AB}^n + p_{ab}^n < p$. In this case, the breeder selects all genotypes at generation n, including aB and Ab.

  Genotype frequencies after selection are equal to: $p_{AB}^n{}' = \frac{p_{AB}^n}{p}$, $p_{ab}^n{}' = \frac{p_{ab}^n}{p}$ and $p_H^{n\prime} = \frac{p - p_{AB}^n - p_{ab}^n}{p}$. Note that $p_{ij}^{n\prime} = \frac{p_H^{n\prime}}{p_H^n}\, p_{ij}^n$ with ij = Ab or aB. The linkage disequilibrium after selection in generation n includes frequencies of aB and Ab: $D^{n\prime} = p_{AB}^n{}' \times p_{ab}^n{}' - p_{aB}^n{}' \times p_{Ab}^n{}' = \frac{p_{AB}^n \times p_{ab}^n}{p^2} - p_{aB}^n \times p_{Ab}^n \left(\frac{p - p_{AB}^n - p_{ab}^n}{p(p_{aB}^n + p_{Ab}^n)}\right)^2$

  Therefore, the genotype frequencies at the generation n + 1 are equal to: $p_{AB}^{n+1} = \frac{p_{AB}^n}{p} - rD^{n\prime}$, $p_{ab}^{n+1} = \frac{p_{ab}^n}{p} - rD^{n\prime}$, and $p_{ij}^{n+1} = \frac{p - p_{AB}^n - p_{ab}^n}{p} \times \frac{p_{ij}^n}{p_{aB}^n + p_{Ab}^n} + rD^{n\prime}$ (ij = Ab or aB).

- **Configuration 2:** the proportion of selected individuals p exceeds the proportion of genotype AB, but not the proportion of AB and ab: $p_{AB}^n < p < p_{AB}^n + p_{ab}^n$. So the breeder selects only genotypes AB and ab at generation n.

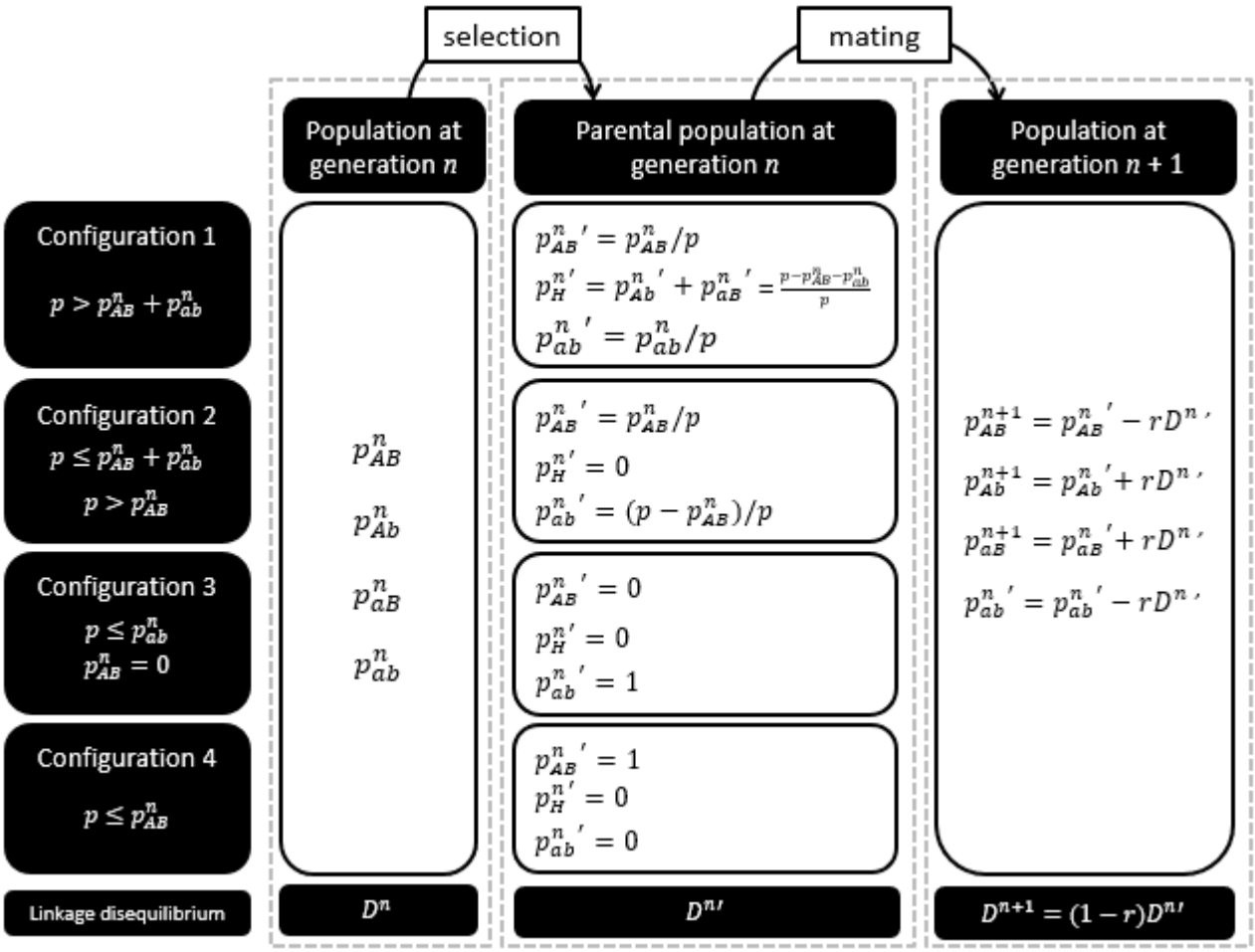


Figure 1: Processes of selection and mating in the different configurations

p = percentage of selected individuals

r = recombination rate

p_{ij}^n = frequency of genotype *ij* at generation *n* before selection

$p_{ij}^{n\prime}$= frequency of genotype *ij* at generation *n* after selection

$D^n = p_{AB}^n \times p_{ab}^n - p_{aB}^n \times p_{Ab}^n$= linkage disequilibrium at generation *n*, before selection

$D^{n\prime} = p_{AB}^{n\,\prime} \times p_{ab}^{n\,\prime} - p_{aB}^{n\,\prime} \times p_{Ab}^{n\,\prime}$ = linkage disequilibrium at generation *n* after selection.

$p_H^{n\prime} = p_{aB}^{n\,\prime} + p_{Ab}^{n\,\prime}$; $p_H^n = p_{aB}^n + p_{Ab}^n$. Nota that $p_{ij}^{n\,\prime} = \frac{p_H^{n\prime}}{p_H^n}$ with *ij* = *Ab* or *aB*.

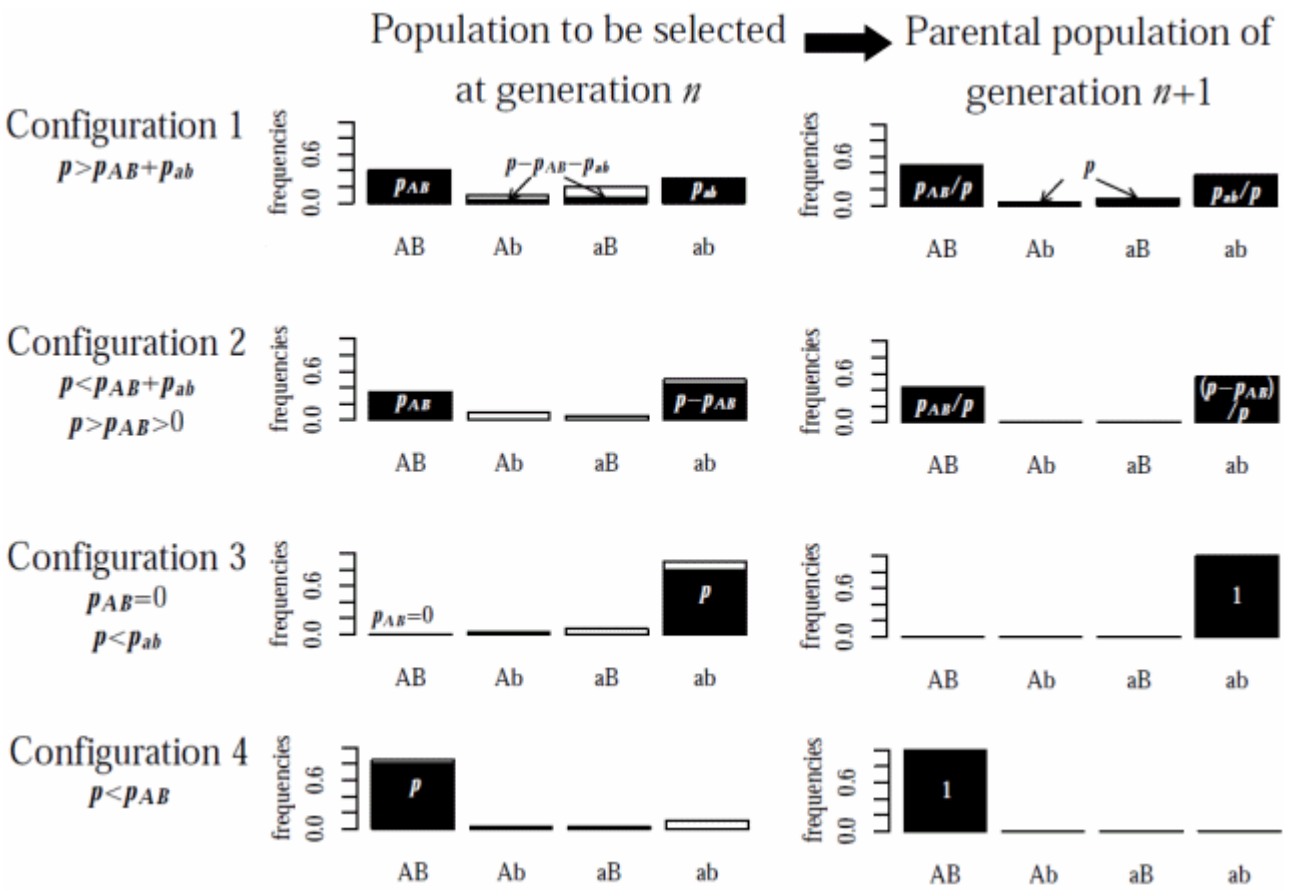


Figure 2: Selection in the different populational configurations

p = percentage of selected individuals

p_{ij}: frequency of genotype ij

Black blocs: fraction p of the population which is selected to be progenitors of the population at the next population

Genotype frequencies after selection are equal to: $p_{AB}^{n}{}' = \frac{p_{AB}^{n}}{p}$, $p_{ab}^{n}{}' = \frac{p-p_{AB}^{n}}{p}$ and $p_{H}^{n}{}' = 0$. The linkage disequilibrium after selection is equal to $D^{n\prime} = \frac{p_{AB}^{n}(p-p_{AB}^{n})}{p^2}$

Therefore at generation $n + 1$: $p_{AB}^{n+1} = \frac{p_{AB}^{n}}{p} - rD^{n\prime}$, $p_{ab}^{n+1} = \frac{p - p_{AB}^{n}}{p} - rD^{n\prime}$ and $p_{aB}^{n+1} = p_{Ab}^{n+1} = rD^{n\prime}$

- **Configuration 3**: There is no AB genotype in the population and the ab genotype is present at higher frequency than p: $p \leq p_{ab}^{n}$ and $p_{AB}^{n} = 0$. The breeder selects only ab individuals to create the new population at generation n+1, therefore $p_{ab}^{n}{}' = p_{ab}^{n+1} = 1$. Genotype ab becomes fixed at generation $n + 1$. This configuration must be avoided because it leads to the fixation of ab, a moderately efficient genotype.
- **Configuration 4**: the AB genotype is present at higher frequency than p: $p \leq p_{AB}^{n}$. The breeder selects only AB individuals to create the new population at generation n+1, therefore $p_{AB}^{n}{}' = p_{AB}^{n+1} = 1$. Genotype AB becomes fixed at generation $n + 1$. This is the configuration sought by the breeder if the goal is to fix AB, the best genotype.

Impossible configurations and initial states

We assume that all alleles are present in the population at the initial state. Hence there are some impossible initial states:

- $p_{ij}^{0} = 1$ for any *ij* genotype.
- $p_{Ab}^{0} + p_{AB}^{0} = 1$, $p_{ab}^{0} + p_{aB}^{0} = 1$, $p_{ab}^{0} + p_{Ab}^{0} = 1$or $p_{aB}^{0} + p_{AB}^{0} = 1$

As genotypes Ab and aB have the lowest fitness, they will be counterselected in the presence of ab or AB genotypes. Moreover, in the absence of ab and AB genotypes in the population, neither Ab nor aB will be selected because their fitness values are equal. So a population consisting only of genotypes Ab and aB is without selection. Thus $p_{Ab}^{n} \neq 1$ and $p_{aB}^{n} \neq 1$ for any n. Only genotypes ab or AB can be fixed.

IV.3.2.3 Step 2: Numerical calculations of times to fixation in populations of infinite size

The above equations were implemented in the R 3.0.2 software program (R Core Team, 2013), with the Rmpfr package, using 5000 b precision. This numerical approach was used to confirm which genotype becomes fixed and to calculate the time to fixation T, in the case of specific initial states, still in populations of infinite size. Numerical calculations of genotype frequencies were performed from the starting generation, as defined in **Table 2,** to the 10 000th generation. This number was chosen to be high enough to ensure numerical convergence to a stable state, regardless of whether there is fixation or not.

Two initial states were modeled (Table 2). The first initial state mimics an admixture situation, as may be encountered in the pre-breeding. The starting population is at maximal negative linkage disequilibrium. The second initial state corresponds to a situation where all genotypes are present, except the most effective (AB), which also induces negative linkage disequilibrium.

Table 2: Initial states in the numerical calculations

Genotype	Genotypic value	Genotypic frequencies in the initial state 1	Genotypic frequencies in the initial state 2
ab	1	$p_{ab}^0 = 0$	p_{ab}^0
aB	$1 - y$	p_{aB}^0	$p_{aB}^0 = p_{Ab}^0 = \frac{1 - p_{ab}^0}{2}$
Ab	$1 - y$	$p_{Ab}^0 = 1 - p_{aB}^0$	$p_{Ab}^0 = p_{aB}^0 = \frac{1 - p_{ab}^0}{2}$
AB	$1 + x$	$p_{AB}^0 = 0$	$p_{AB}^0 = 0$
Sign of linkage disequilibrium		< 0	< 0

p_{AB}^0 = initial frequency of genotype AB; p_{ab}^0 = initial frequency of genotype ab; p_{aB}^0 = initial frequency of genotype aB; p_{Ab}^0 = initial frequency of genotype Ab; $x, y > 0$.

IV.3.2.4 Step 3: Simulation of the evolution of a finite population

Drift was taken into account by setting a finite population size (N = 1000 and N = 5000). Using the software R program, stochastic simulations were carried out to determine which genotype becomes fixed in a finite population, especially near critical parameters (r_c, p_c) below and above which one genotype or another may be fixed. These critical parameters had been calculated beforehand with analytical and numerical calculations regarding populations of infinite size (steps 1 and 2). The

populations evolved for 100 generations. One hundred repetitions were performed. Some specific situations were studied and are shown in **Table 3**.

Table 3: Initial parameter values in the simulations for the two initial states

	Initial state 1	Initial state 2	
Initial genotypic frequencies	$p_{ab}^0 = 0$ $p_{aB}^0 = 0.6$ $p_{Ab}^0 = 0.4$ $p_{AB}^0 = 0$	$p_{ab}^0 = 0.8$ $p_{aB}^0 = 0.1$ $p_{Ab}^0 = 0.1$ $p_{AB}^0 = 0$	$p_{ab}^0 = 0.1$ $p_{aB}^0 = 0.45$ $p_{Ab}^0 = 0.45$ $p_{AB}^0 = 0$
Recombination rate *r*	p fixed to 0.9 r from 0.18 to 0.22	p fixed to 0.9 r from 0.09 to 0.11	p fixed to 0.9 r from 0.09 to 0.16
Percentage of individuals selected *p*		r fixed to 0.1 p from $0.9p_c$ to $1.1p_c$ (with $p_c = 0.8$)	r fixed to 0.01 p from $0.9p_c$ to $1.1p_c$ (with $p_c = 0.1$)

r = recombination rate ; p = percentage of selected individuals; r_c and p_c = critical parameters; p_{ab}^0=initial frequency of genotype ab; p_{aB}^0=initial frequency of genotype aB; p_{Ab}^0 = initial frequency of genotype Ab ; p_{AB}^0= initial frequency of genotype AB

IV.3.2.5 Rationale of the model

IV.3.2.5.1 Haploidy

As long as the outcrossing rate remains low, heterozygosity is weak enough to not substantially affect the evolution of the population. So in this case the main evolution factors were correctly modeled. However, the evolution of a population with high recombination rates, generating significant heterozygosity, may be influenced by other factors (notably dominance) which are not taken into account in this model.

IV.3.2.5.2 Recombination and outcrossing

In this study, the outcrossing rate was equal to one (random mating), and several recombination rates ($r \in [0,0.5]$) were tested. But in this case testing the recombination rate is equivalent to testing the outcrossing rate as well, since the effective recombination rate r_e is the determining factor in the evolution of a population. According to Nordborg (2000), r_e is equal to the recombination rate r,

multiplied by the outcrossing rate 1-S (with S being the selfing rate): r_e = r·(1-S). Thus, a haploid random-mating population in which the recombination rate between loci A and B is equal to 0.1 evolves in the same way as a population in which A and B are independent (r = 0.5), but in which the outcrossing rate is equal to 0.2. Consequently, in this model, the outcrossing rate is low if r is small.

IV.3.2.5.3 Initial states in numerical calculations of times to fixation

In all the initial states, the AB genotype was not present. Indeed, the implicit pre-breeding assumption is that the best allelic combination must be created by assembling alleles that are present in different lineages: thus, *a priori*, a genotype such as AB does not exist in the initial pre-breeding population.

In the first initial state, the first generation of pre-breeding population corresponds to a population where wild and elite individuals coexist but have not yet reproduced together. Genotypes aB and Ab are the only genotypes present in the initial population, and therefore represent the elite and wild genotypes. Recombinant genotypes (AB and ab here) would be more successful than expected in a genetic architecture with no sign epistasis.

In the second initial state, there is greater genotype variety than in the first one (three genotypes are present versus two in the first initial state). This is also relatively representative of a pre-breeding population. Indeed, in a pre-breeding population, genotypes present in the initial population are likely to have a diversity of phenotypes and therefore do not all have the same fitness (as genotypes aB, Ab and ab present in this initial state).

IV.3.3 Results

IV.3.3.1 Evolutionary paths and final equilibria

When $p \leq p_{AB}^{n}$ (configuration 4), the genotype with the highest genotype value, AB, is fixed within a single generation regardless of the value of the recombination rate r, because only AB individuals are chosen to create the new generation. Similarly, in configuration 3 ($p_{AB}^{n} = 0$ and $p \leq p_{ab}^{n}$), ab, the suboptimal genotype, becomes fixed. If neither configuration 3 or 4 is true, the population is in one of

the two other possible configurations: configuration 1 ($p > p_{AB}^n$ and $p > p_{AB}^n + p_{ab}^n$) or configuration 2 ($p > p_{AB}^n$ and $p \leq p_{AB}^n + p_{ab}^n$). Conditions to move from one configuration to another or to fix a genotype are described below and summarized in **Figure 3**. Details of the calculations can be found in **Supplementary File 1**.

The first configuration is generally instable. From it, a population can take:

- **path 1 (staying in configuration 1) if:**
  - $p_{AB}^n - p_{ab}^n = 0$
    - $r \geq r_c \geq 2(1\text{-}p)$ ($r_c = 2(1\text{-}p)$ when $p_{aB}^n = p_{Ab}^n$)

If the population fulfills these conditions at generation n, then it will fulfill them at generation n+1 because r and p (so r_c) are constant, and because p_{AB}^n and p_{ab}^n will not increase. This means that if the population satisfies these conditions at one generation, it will remain trapped within configuration 1 forever. This situation is an equilibrium where polymorphism is maintained at both loci.

- **path 2 (going to configuration 4 at generation n+1 and fixing the genotype AB at generation n+2) if:**
  - $rD^{n\prime} \leq \frac{p_{AB}^n - p^2}{p}$

This path is unlikely because it requires specific conditions: p_{AB}^n must be quite strong because $\frac{p_{AB}^n - p^2}{p}$ must be equal or superior to $rD^{n\prime}$, but at the same time $p_{AB}^n + p_{ab}^n < p$ as the population is in configuration 1 at generation n.

- **path 3 (going to configuration 3 at generation n+1 and fixing the genotype ab at generation n+2) if:**

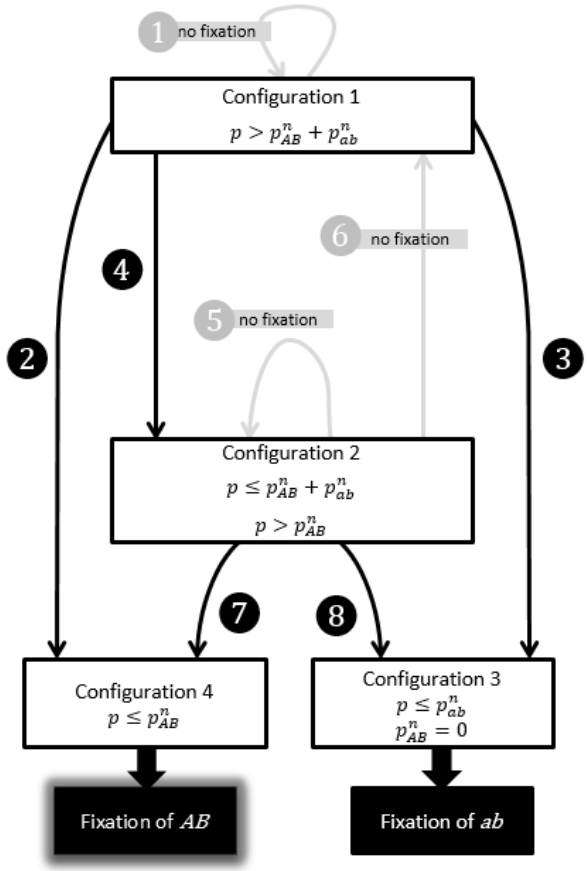


Figure 3: Evolutionary paths possibly taken by the population
p_{ij}^n : frequency of genotype ij at generation n before selection

Configuration 1: $p > p_{AB}^n$ and $p > p_{AB}^n + p_{ab}^n$. All genotypes contribute to the next generation because individuals with genotypes AB and ab are not frequent enough in the population.

Configuration 2: $p > p_{AB}^n$ and $p \leq p_{AB}^n + p_{ab}^n$. Only individuals with genotypes AB and ab contribute to the next generation.

Configuration 3: $p \leq p_{ab}^n$ and $p_{AB}^n = 0$. Only ab genotypes are selected to create the new population at generation n+1.

Configuration 4: $p \leq p_{AB}^n$. Only AB genotypes are selected to create the new population at generation n+1.

1,2…, 8: number of the evolutionary paths. Conditions for the population to take them are detailed in the text.

Light grey arrow: path leading to an equilibrium state where polymorphism is maintained at the two loci (no fixation)

- $p_{AB}^n - p_{ab}^n < 0$
    - $rD^{n\prime} = \frac{p_{AB}^n}{p}$
    - $p_{ab}^n - p_{AB}^n \geq p^2$

Like path 2, path 3 is unlikely because it requires very specific conditions. In particular, $rD^{n\prime}$ must be exactly equal to $\frac{p_{AB}^n}{p}$. However, this path may exist for example when $r = 0$, $p_{AB}^n = 0$ and $p_{ab}^n \geq p^2$.

- **path 4 (going to configuration 2) if:**
    - $p_{AB}^n - p_{ab}^n \neq 0$ with conditions that differ from those leading to path 2 or 3

  or

    - $p_{AB}^n - p_{ab}^n = 0$
        - $r < r_c$ (if $p_{aB}^n = p_{Ab}^n$ then $r_c = 2(1\text{-}p)$, else $r_c \geq 2(1\text{-}p)$)

          In this case, the population will then necessarily follow path (see below), i.e. it will eventually fix AB.

Configuration 2 is crucial because it is the easiest way to achieve fixation compared to configuration 1, where direct fixations (paths 2 and 3) are unlikely. If the population is in configuration 2, it can go through:

- **path 5 (staying in configuration 2) if:**
    - $p_{AB}^n \leq \frac{p}{2}$
        - $r = \frac{(1-p)p}{p-p_{AB}^n}$

If the population fulfills these conditions at generation n, then $p_{AB}^{n+1} = p_{AB}^n$. So r will again fulfill them at generation n+1. Therefore, the population will remain trapped within configuration 2. No genotype will be fixed.

- **path 6 (going back to configuration 1 and staying there) if:**
    - $p_{AB}^n = \frac{p}{2}$

- $r \geq r_c \geq 2(1\text{-}p)$ ($r_c = 2(1\text{-}p)$ when $p_{aB}^n = p_{Ab}^n$)

If $p_{AB}^n = \frac{p}{2}$, $p_{AB}^{n+1} = p_{ab}^{n+1} < p/2$. So at generation n+1, the population goes back to configuration 1 and meets the conditions of path 1. The population remains trapped within configuration 1.

- **path 7 (going to configuration 4 and fixing AB) if :**
  - $p_{AB}^n > \frac{p}{2}$

  If r is strong ($r \geq \frac{(1-p)\cdot p}{p - p_{AB}^n}$), the population transiently goes back to configuration 1.

  or

  - $p_{AB}^n \leq \frac{p}{2}$
    - $r < \frac{(1-p)p}{p-p_{AB}^n}$
- **path 8 (going to configuration 3 and fixing ab) if :**
  - $p_{AB}^n < \frac{p}{2}$
    - $r > \frac{(1-p)p}{p-p_{AB}^n}$

  If r is strong, the population transiently goes back to configuration 1.

It is often possible, knowing the initial state of the population, to predict the path it will take after several generations. This is the case for paths 1, 4, 5, 6, 7 and 8. If the initial population meets the criteria of one of these paths, it will necessarily follow it after a given number of generations. Conversely, for paths 2 and 3, if the conditions presented above are met at generation n, fixation of AB or ab will occur at generation n + 2.

IV.3.3.2 Fixation according to the recombination rate and strength of selection

These predictions from the analytical model were supplemented with numerical calculations. Changes in genotype frequencies were estimated every generation from the starting generation defined in Table 2 to the 10 000th generation using the equation formulated in step one for an infinite sized population. In addition to the fixed genotype, the time to fixation T was numerically estimated. Graphs plotting the logarithm of the time to fixation T and the fixed genotype, according to the recombination

rate r and the percentage p of selected individuals are shown in **Figures 4** and **5**. These numerical calculations were complemented by simulations in populations of finite size. Tables summarizing the simulation results are shown in **Supplementary File 2**.

IV.3.3.2.1 Initial state 1: optimal and suboptimal genotypes are absent from the initial population

This initial state corresponds to the situation where recombinant genotypes (here AB and ab) have a higher genotype value than the initial ones, Ab and aB, which correspond to the two sub-populations present in the starting pre-breeding population. This initial state 1 matches the above-mentioned configuration 1 because $p_{AB}^0 + p_{ab}^0 < p$.

IV.3.3.2.1.1 Numerical calculations in populations of infinite size

The initial state 1 starts in configuration 1, so two areas can be identified within the time to fixation landscape (Figure 4.A). If $r < r_c$, the population will take path 4 and then path 7: AB will be fixed. On the contrary, if $r \geq r_c$, the population will follow path 1 and no genotype will be fixed. Note that when the population is in the AB fixation area, the closer it gets to the other area (by increasing p and/or r), the longer the time to fixation T. Note that, as $r_c \geq 2(1\text{-}p)$, when $p < 0.75$, then $r_c > 0.5$: thus in this condition, r is always inferior to r_c and AB always becomes fixed.

Our numerical calculations suggest that, regardless of the frequencies of aB and Ab (even if $p_{aB}^0 \neq p_{Ab}^0$), the assumption $r < 2(1\text{-}p)$ is sufficient to ensure that AB is always fixed, as $2(1\text{-}p)$ is a lower bound for r_c. Furthermore, the fixation time landscapes obtained when $p_{aB}^0 \neq p_{Ab}^0$ (shown in **Supplementary File 3**) are very similar to the one achieved with $p_{aB}^0 = p_{Ab}^0 = 0.5$.

Note also that there is a "wall" of infinite height at $r = 0$. This makes sense because, in the complete absence of recombination, genotypes ab and AB cannot be created. And as genotypes aB and Ab have the same fitness, there is no selection and the genotype frequencies remain constant.

Figure 4.B details $\log(T)$ for some p and r in the case where $p_{aB}^0 = p_{Ab}^0 = 0.5$. We find that T increases with p, but its relationship with r is more complex. For low p (strong selection), a high recombination

rate allows quicker fixation of AB. However, for high p (weak selection), an intermediate recombination rate leads to a quicker fixation.

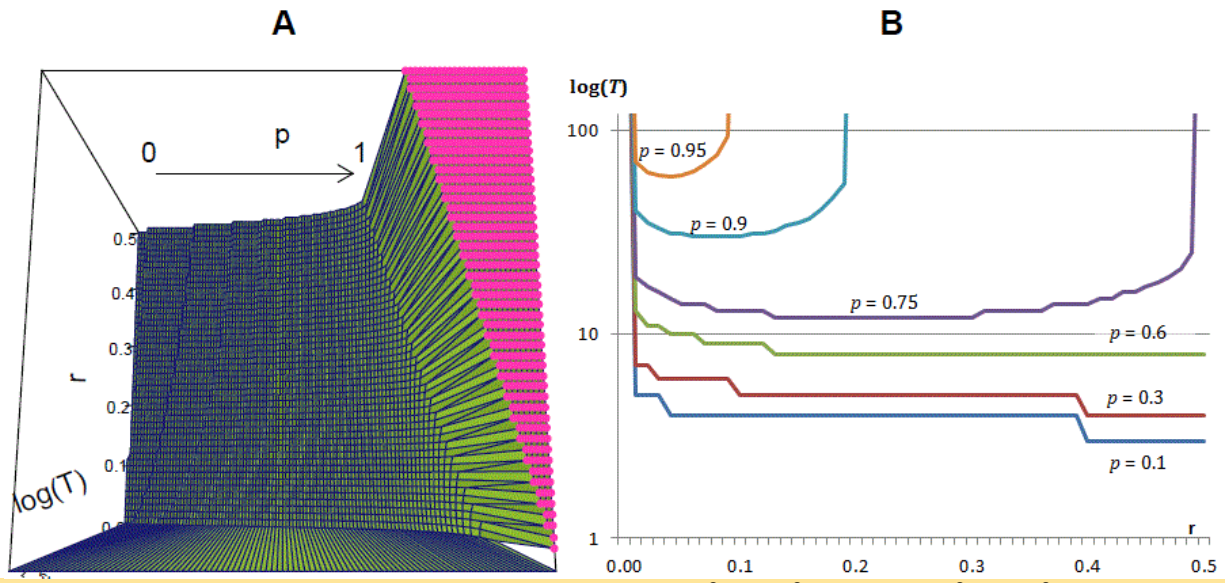


Figure 4: Logarithm of the time to fixation in initial state 1 ($p^0_{ab} = p^0_{AB} = 0$) for $p^0_{aB} = p^0_{Ab} = 0.5$

A. Depending on r and p
   Dark blue grid area: fixation of AB (paths 4 and 7); pink points: no fixation (path 1)
B. Focus on some particular p

r = recombination rate; p = percentage of selected individuals; T = time to fixation; p^n_{ij} = frequency of genotype ij at generation n before selection

IV.3.3.2.1.2 Simulations in finite populations

As detailed above, theoretical predictions in infinite populations suggest that if $r < r_c$, AB should be fixed and if $r \geq r_c$, the polymorphism should be maintained. Simulations were carried out for $p^0_{aB} = 0.4$ and $p^0_{Ab} = 0.6$ to examine which genotypes were fixed around the critical recombination rate. The chosen percentage of selected individual p was equal to 0.9. Therefore, $r_c > 0.2$.

We found that the simulations were consistent with our theoretical predictions. Indeed, when $r < 0.2$, AB becomes more commonly fixed than ab. However, ab also becomes fixed in a high number of simulations, particularly if N is small, i.e. if genetic drift is strong. For example, when $r = 0.18$, ab becomes fixed in 28% of the replicates when $N = 1000$. When genetic drift is weaker ($N = 5000$), genetic polymorphism is frequently maintained (no fixation). This could imply that the time to fixation is longer under these conditions.

For $r \geq 0.2$, AB is fixed in roughly half of the repetitions. For example, when $N = 1000$ and $r = 0.2$, 47% of the repetitions are concluded by the fixation of AB. When $N = 1000$, ab is fixed in about half of the repetitions: e.g. it is fixed in 57% of the repetitions when $r = 0.22$. But when $N = 5000$, its fixation is less common (e.g. in only 35% of the repetitions for $r = 0.21$).

In conclusion, to increase the probability of fixing AB, in the initial state 1, the recombination rate r should be low but non-null, and selection should be strong (low p). A large population size could further promote AB fixation. But, if the breeder's goal is to maintain polymorphism, then the recombination rate should be maintained at its critical value, in addition to reducing the effect of drift.

IV.3.3.2.2 Initial state 2: the optimal genotype is absent from the initial population

In this initial state, $p_{AB}^0 = 0$ and $p_{aB}^0 = p_{Ab}^0$. This corresponds to the situation where all genotypes are present, except the one (AB) with the highest genotype value and fitness. It must be created by recombination.

IV.3.3.2.2.1 Numerical calculations in populations of infinite size

Several areas can be observed in Figure 5. On the left of graphs A and B, there is an area of low p ($\leq p_{ab}^0$), which corresponds to configuration 3 ($p \leq p_{ab}^0$ and $p_{AB}^0 = 0$) and leads to the fixation of ab.

On the right of the two graphs in Figure 5, where $p > p_{ab}^0$, initial state 2 begins in configuration 1 ($p_{ab}^0 + p_{AB}^0 = p_{ab}^0 < p$). First, the line where $r = 0$ corresponds to path 3. Indeed, p_{AB}^n will remain null for any generation n since genotype AB cannot be created. Thus, $rD^n > \frac{p_{AB}^n}{p} = 0$. In addition, as $p_{ab}^n > p_{AB}^n$, frequency p_{ab} will increase, so at some point p_{ab} will exceed threshold p^2 and ab will be fixed two generations later. The rest of the region, where $r > 0$ and $p > p_{ab}^0$ is divided into two areas:

- An area of relatively low p and low r, where AB is fixed.
- An area of strong p and strong r, where ab is fixed.

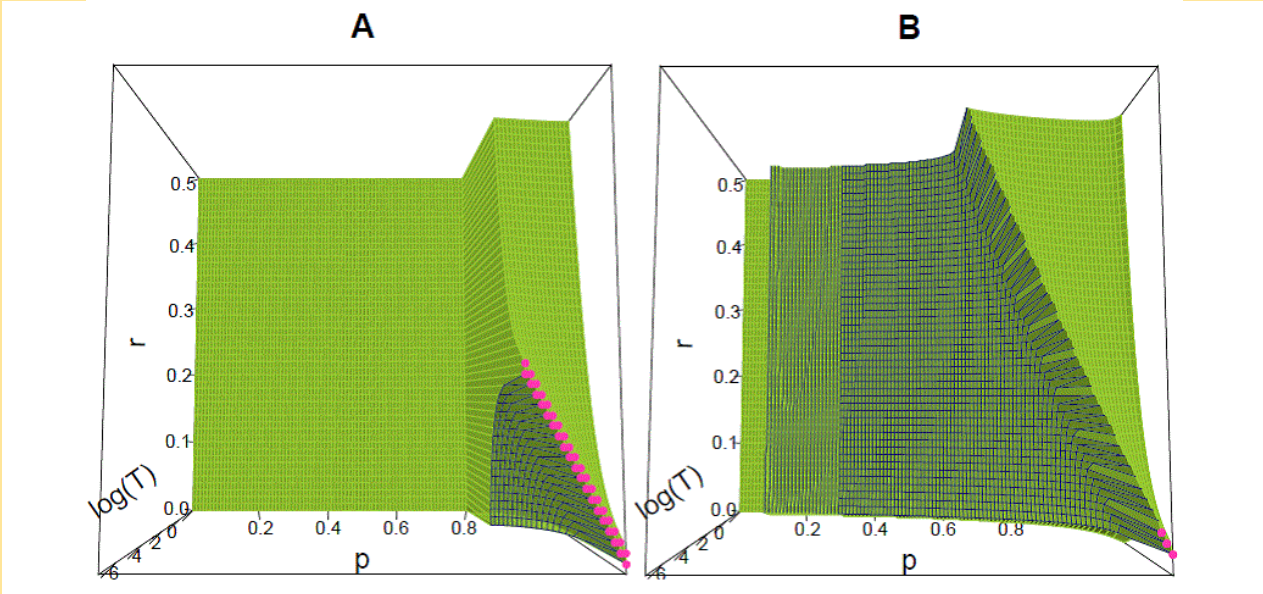


Figure 5: Logarithm of the time to fixation and fixed genotype in initial state 2 ($p_{AB}^0 = 0$ and $p_{aB}^0 = p_{Ab}^0$) for different initial genotypic frequencies p_{ab}^0

A: $p_{ab}^0 = 0.8$
B: $p_{ab}^0 = 0.1$

Dark blue grid area: fixation of AB; pink points: no fixation (infinite time to fixation); rest of the graph: fixation of ab.

r = recombination rate; p = percentage of selected individuals; T = time to fixation ; p_{ij}^n= frequency of genotype ij at generation n before selection

As $p_{ab}^0 > p_{AB}^0 = 0$ and $r > 0$, the population will always be in configuration 2 (path 4). Then it will take either path 7 (fixation of AB) or path 8 (fixation of ab). The two areas that correspond to these two paths are separated by a wall of infinite height (polymorphic situation and no fixation), which matches path 5 (blocked within configuration 2). The time to fixation T increases when the population approaches this wall. Note that in Figure 5.B this wall is visible only for very high p values due to the lack of calculation accuracy.

Let us call n_0 the generation at which the population enters into configuration 2. The conditions for which the population uses each path 7, 8 or 5 depends on the critical rate $r_c = \frac{(1-p)p}{p-p_{AB}^{n_0}}$ (see evolution paths). It is hard to predict n_0 and $p_{AB}^{n_0}$, and thus the critical value r_c. However, a lower bound for r_c is 1 - p. Moreover, this lower bound can be used as an approximation of r_c for high values of p_{ab}^0.

The area of AB fixation is at a lower height on the $\log(T)$ axis than that of ab fixation. This means AB is fixed faster than ab. In the area of AB fixation, $T(r,p)$ meets a minimum only when p is not too high. For example, in the space of the studied parameters (i.e. r = 0, 0.01,... 0.5 and p = 0.01, 0.02,... 0.99) we found that for $p_{ab}^0 = 0.1$, $T(r,p)_{p\geq 0.97}$ is simply an increasing function. But elsewhere, a minimum value is obtained along the r-axis. For example, a minimum of $T(r, p = 0.45)$ = 6 generations are observed when $r \in [0.19, 0.47]$ and a minimum of $T(r, p = 0.8)$ = 19 generations when $r \in [0.06, 0.14]$. The r values decrease as p increases, where $T(r,p)$ is minimal. In the same way, the function $T(r,p)$ reaches a minimum along the p-axis. For example, for $p_{ab}^0 = 0.1$ and r = 0.2, the minimum is obtained when $p \in [0.14, 0.21]$, with $T(r = 0.2, p)$ = 4 generations.

In the area of ab fixation, strong recombination reduces the time to fixation. Along the p-axis, the function $T(r,p)$ reaches a minimum, which increases when r is low.

IV.3.3.2.2.2 Simulations in finite populations

Let $p_{c,1}$ and $p_{c,2}$ denote the two critical values for p:

- $p_{c,1} = p_{ab}^0$
- $p_{c,2}$ is the value of p at which r equals r_c. If $r_c \approx 1 - p$ (as for high values of p_{ab}^0), thus $p_{c,2} \approx 1 - r$

Predictions in infinite populations are:

- If $p \leq p_{c,1}$, ab becomes fixed
- If $p > p_{c,1}$
    - $r < r_c$ (equivalent to $p < p_{c,2}$), AB becomes fixed
    - $r > r_c$ (equivalent to $p > p_{c,2}$), ab becomes fixed
    - $r = r_c$ (equivalent to $p = p_{c,2}$), there is no fixation

Two sets of simulations were run: one with $p_{ab}^0 = 0.8$ and $p_{aB}^0 = p_{Ab}^0 = 0.1$, the other one with $p_{ab}^0 = 0.1$ and $p_{aB}^0 = p_{Ab}^0 = 0.45$.

In the first set ($p_{ab}^0 = 0.8$ and $p_{AB}^0 = 0$), ab was fixed in 98% of the replicates, irrespective of the N, r and p values.

The results of the second set of simulations ($p_{ab}^0 = 0.1$ and $p_{AB}^0 = 0$) when $p > p_{c,1}$ showed that the critical recombination rate r_c, which governs the orientation toward path 7 (fixation of AB) or 8 (fixation of ab), is close to 0.13. Indeed, when $r \in [0.090, 0.12]$ AB becomes fixed in most replicates. At $r = 0.13$, AB becomes fixed in only 54% and 37% of the replicates, respectively when $N = 1000$ and $N = 5000$. Next, for $0.13 < r < 0.16$, the fixation rate of AB decreases (from 23 to 2% of the replicates when $N = 1000$). Note however that for these recombination rates ($r = 0.13$, 0.14 and 0.15), ab is not fixed very frequently (respectively in 1, 18 and 42% of the repetitions when $N = 1000$, and 0, 0, and 4% of the repetitions when $N = 5000$). Actually, in a certain number of repetitions (from 28 to 59% when $N = 1000$ and from 66% to 96% for $N = 5000$), no genotype becomes fixed after 100 generations of evolution at these rates. However, when $r = 0.16$, ab becomes fixed more often: in 70% of them for $N = 1000$ and in 34 % of them for $N = 5000$.

The results of the second set of simulations ($p_{ab}^0 = 0.1$) around $p_{c,1}$, when $r = 0.01 < r_c$, showed that AB is rarely fixed, even when $p > p_{c,1}$. For example, when $p = 0.11$, AB becomes fixed in only 19% of the repetitions (ab: 81%) when $N = 1000$. Thus, even when the recombination rate is small and largely inferior to r_c, the suboptimal genotype ab is very likely to become fixed in finite populations if $p \approx p_{c,1}$, even if $p > p_{c,1}$.

Numerical calculations in populations of infinite size showed that the frequency p_{ab} can increase significantly in earlier generations. In the area of AB fixation, the maximum reached by the frequency p_{ab} over generations is close to 1 in a large part of the parameter space, especially when p_{ab}^0 is large. Nevertheless, p_{ab} always remains inferior to 1 and the genotype AB can finally be fixed in populations of infinite size, as there is no drift. But this significant increase in p_{ab} leads to the fixation of ab in finite populations.

The first set of simulation shows that when p_{ab}^0 is high, ab will be fixed in finite populations. To ensure that AB gets a chance to be fixed, p_{ab}^0 must be quite inferior to p, while p must nevertheless remain below $p_{c,2}$. Hence, our results show that the fixation of AB in a finite population is possible only if $p_{ab}^0 < < p < p_{c,2}$. Thus, we found that the fixation area of AB is much less extensive in finite populations.

In conclusion, in the initial state 2, the most efficient genotype, AB, is in very strong concurrence with ab, which has the advantage of being already present in the initial population. In a finite population, the risk of losing genotype AB due to genetic drift is strong, especially if the frequency of ab in the population is high. Thus, a pre-breeder should absolutely seek to reduce drift in the population by choosing a large population size and by exercising weak selection. The recombination rate should be low, but not null.

IV.3.4 Discussion

Here we have presented an analysis of a haploid two-locus model focused on a population subject to truncation selection on a phenotypic trait displaying reciprocal sign epistasis, with unequal fitness values for the two best genotypes. The originality of our work can be summed up in two main features. First, the genotype fitness values varied over time because they depended on genotype frequencies, mimicking what happens in pre-breeding populations. Critical values for truncation selection and recombination rate were thus dependent, complex and not expressed in terms of differences between fitness values, but instead relative to the key parameters (recombination rate r and proportion of individuals selected p) and to the genotype frequencies. Moreover, contrary to most previous studies dealing with the dynamic aspect of the evolution of populations under reciprocal sign epistasis (e.g. Higgs 1998, Jain 2010, Stephan 1996 and Weinreich et al. 2005), in our study the initial states did not correspond to a population dominated by the suboptimal genotype ($p_{ab}^0 = 1$), they corresponded to a pre-breeding population with two or three genotypes.

Our results confirmed that, in the case of fitness displaying reciprocal sign epistasis, high recombination rates have a negative effect on the fixation of the optimal genotype (de Visser et al.,

2009). Like Park and Krug (2011), Rutschman, (1994) and Weinreich et al. (2005), we have shown that for given parameter values, there may be a critical recombination rate above which polymorphism is maintained (as in initial state 1) or the suboptimal genotype becomes fixed (as in initial state 2). However, we have also shown that when the initial linkage disequilibrium is negative, recombination is necessary to create the most favorable genotype AB. In this situation, recombination can even lead to faster fixation. Indeed, there is an optimal recombination rate r for which the time to fixation is minimal. Our results are thus consistent with those of de Visser et al., (2009) who showed that recombination is beneficial for fixing the optimal genotype only in the presence of some particular parameters in the case of reciprocal sign epistasis. De Visser et al. (2009) used the 'fitness landscapes' concept in their work, whereby fitness landscapes allow visualization of the relationship between genotypes and fitness values. Height represents fitness and the points at the surface represent the genotype composition of the population. As de Visser et al. (2009) worked on a multilocus mutational model with adaptive landscapes consisting of multiple fitness peaks and valleys, our results cannot be fully compared to theirs. However, some of their conclusions were confirmed in our study. First, de Visser et al. (2009) highlighted the role of the adaptive landscape topography on the evolution of populations. In our study, the adaptive landscape always involved two fitness peaks (corresponding to the fitness of genotypes AB and ab) and, depending on the initial genotype frequencies, the initial population may be positioned in a fitness valley or on a peak, and this obviously influences the evolution of the population. We were able to show that the initial genotype frequencies had a major influence on the evolutionary direction that a population can take. Secondly, like de Visser et al. (2009), we were able to demonstrate the existence of an optimal recombination rate, which reduces the time to fixation under certain conditions. Finally, simulations in finite populations confirmed that the population size also has an effect on the fixation of optimal and suboptimal genotypes, leading to the loss of the one that is present at the lowest frequency.

The aim of pre-breeding populations is to mix substantial genetic diversity and to gradually increase the average phenotype of the population. In pre-breeding, it is possible for some selfing species to

control the outcrossing rate using a male sterility gene. If the breeder's goal is to create and fix the best genotype, the rule to be applied on the outcrossing rate may be complex. Though a minimum of recombination is necessary to create the optimal genotype, an excessively efficient recombination rate can also break up favorable allelic combinations that occur in autogamous plants (Allard et al., 1993; Fenster et al., 1997; Nordborg et al., 2002; Pollak and Sabran, 1999). The effect of the outcrossing rate will actually depend on two factors, i.e. the tightness of the linkage between loci (crossing-over rate r) and the selection strength (p). Generally in pre-breeding populations, selection is too weak to avoid excessive loss of alleles (p is around 0.80 or 0.90). In this case, the recombination rate should not be excessively effective. When linkage disequilibrium is negative, recombination is favorable for the fixation of AB in the first generations because it allows the creation of the most favorable genotype, AB. However, when the linkage disequilibrium becomes positive via the selection of AB and ab, recombination becomes unfavorable because it destroys these favorable allelic combinations.

The initial state 2 (cf. Table 3, where genotypes aB, Ab and ab are present initially) is relatively representative of an initial pre-breeding population with several phenotypes. The initial state 2 thus provides information on the potential of a real pre-breeding population if selection is too strong—the most favorable genotype AB cannot become fixed because others with a relatively favorable phenotype will be selected before it is created at relatively high frequency by recombination. The simulations showed that genotype AB has a high risk of being lost because of genetic drift. Weak selection is commonly applied on pre-breeding populations to avoid losing too many alleles, and this seems to be a wise practice. In addition, it might be advisable to not apply selection in the early generations so as to correctly monitor the pre-breeding populations. The first pre-breeding generations would therefore just be dedicated to recombination in order to reduce linkage disequilibrium. Selection should be performed at later generations, once recombination has been able to combine alleles that were initially present in different lineages within new genotypes.

In order to determine a monitoring strategy based on selection and the outcrossing rate for a pre-breeding population, the genetic architecture of the phenotypic trait that the breeder wishes to improve could be analyzed, notably to assess the extent and type of epistasis. Next-generation sequencing technologies (Metzker, 2010) could be valuable tools for this purpose. So far, very few studies have focused on measuring epistasis in pre-breeding or dynamic management populations. For example, Goldringer et al. (1997) estimated the additive and epistatic variances of doubled-haploid lines derived from a panmictic wheat population subjected to recurrent selection and found high epistatic variance regarding earliness and powdery mildew resistance. Epistatic QTLs for grain yield (Blanc et al., 2006) and days to flowering (Buckler et al., 2009) were found in maize populations with a broad genetic base, such as nested association mapping (NAM) populations, where multiple lines are crossed to a particular inbred line. In rice, the analysis of an F2 population showed epistasis strongly influenced yield (Huang et al., 2014).

It would be of interest to extend the present study by testing different mutation rates. Here mutation was overlooked (μ = 0), but Park and Krug (2011), using a haploid two-locus model with reversible mutations, demonstrated that the expression of r_c depends on μ. The role of the mutation rate in the evolution of populations subjected to reciprocal sign epistasis when fitness values are not constant has yet to be studied. It would also be interesting to create a multilocus model with various genetic effects that would better reflect the reality and help to draw up more specific guidelines on the management of pre-breeding populations.

IV.3.5 Conflict of interest

The authors declare no conflicts of interest.

IV.3.6 Data Archiving

This article does not report new empirical data or software.

IV.3.7 Acknowledgements

We thank L. Gay, M-H. Müller, V. Ranwez and T. Flutre for valuable discussion and comments that substantially improved the manuscript.

IV.3.8 References

Allard, RW, García, P, Sáenz-de-Miera, LE, de la Vega, MP (1993). Evolution of multilocus genetic structure in Avena hirtula and Avena barbata. Genetics *135*, 1125–1139.

Barton, NH (1995). A general model for the evolution of recombination. Genet. Res. *65*, 123–144.

Blanc, G, Charcosset, A, Mangin, B, Gallais, A, Moreau, L (2006). Connected populations for detecting quantitative trait loci and testing for epistasis: an application in maize. Theor. Appl. Genet. *113*, 206–224.

Buckler, ES, Holland, JB, Bradbury, PJ, Acharya, CB, Brown, PJ, Browne, et al. (2009). The Genetic Architecture of Maize Flowering Time. Science *325*, 714–718.

Charlesworth, B (1990). Mutation-selection balance and the evolutionary advantage of sex and recombination. Genet. Res. *55*, 199–221.

Colegrave, N, Buckling, A (2005). Microbial experiments on adaptive landscapes. BioEssays *27*, 1167–1173.

Crow, JF, Kimura, M (1979). Efficiency of truncation selection. Proc. Natl. Acad. Sci. *76*, 396–399.

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

Eshel, I, Feldman, M (1970). On the evolutionary effect of recombination. Theor. Popul. Biol. *1*, 88–100.

FAO (1996). Report of the State of the World's Plant Genetic Resources for Food and Agriculture (Rome).

Fenster, CB, Galloway, LF, Chao, L (1997). Epistasis and its consequences for the evolution of natural populations. Trends Ecol. Evol. *12*, 282–286.

Goldringer, I, Brabant, P, Gallais, A (1997). Estimation of additive and epistatic genetic variances for agronomic traits in a population of doubled-haploid lines of wheat. Heredity *79*, 60–71.

Henry, JP, Pontis, C, David, J,Gouyon, PH (1991). An Experiment on Dynamic Conservation of Genetic Resources With Metapopulations. In Species Conservation: A Population-Biological Approach, P.A. Seitz, and P.V. Loeschcke, eds. (Birkhäuser Basel), pp. 185–198.

Higgs, PG (1998). Compensatory neutral mutations and the evolution of RNA. Genetica *102-103*, 91–101.

Huang, A, Xu, S,Cai, X (2014). Whole-Genome Quantitative Trait Locus Mapping Reveals Major Role of Epistasis on Yield of Rice. PLoS ONE *9*, e87330.

Jain, K (2010). Time to fixation in the presence of recombination. Theor. Popul. Biol. *77*, 23–31.

Kimura, M (1985). The role of compensatory neutral mutations in molecular evolution. J. Genet. *64*, 7–19.

Kondrashov, AS (1988). Deleterious mutations and the evolution of sexual reproduction. Nature *336*, 435–440.

Lalić, J, Elena, SF (2012). Magnitude and sign epistasis among deleterious mutations in a positive-sense plant RNA virus. Heredity *109*, 71–77.

Lewontin, RC, Kojima, K (1960). The evolutionary dynamics of complex polymorphisms. Evolution *14*, 458–472.

Metzker, ML (2010). Sequencing technologies — the next generation. Nat. Rev. Genet. *11*, 31–46.

Nordborg, M (2000). Linkage Disequilibrium, Gene Trees and Selfing: An Ancestral Recombination Graph With Partial Self-Fertilization. Genetics *154*, 923–929.

Nordborg, M, Borevitz, JO, Bergelson, J, Berry, CC, Chory, J, Hagenblad, et al. (2002). The extent of linkage disequilibrium in Arabidopsis thaliana. Nat. Genet. *30*, 190–193.

Otto, SP (2009). The evolutionary enigma of sex. Am. Nat. *174 Suppl 1*, S1–S14.

Otto, SP, Feldman, MW (1997). Deleterious Mutations, Variable Epistatic Interactions, and the Evolution of Recombination. Theor. Popul. Biol. *51*, 134–147.

Park, S-C, Krug, J (2011). Bistability in two-locus models with selection, mutation, and recombination. J. Math. Biol. *62*, 763–788.

Pollak, E, Sabran, M (1999). On the theory of partially inbreeding finite populations. VI. The survival probability of a two-locus allele combination when there is partial selfing. Math. Biosci. *159*, 97–112.

Poon, A, Chao, L (2005). The Rate of Compensatory Mutation in the DNA Bacteriophage ϕX174. Genetics *170*, 989–999.

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

R Core Team (2013). R: A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing).

Rice, WR (2002). Experimental tests of the adaptive significance of sexual recombination. Nat. Rev. Genet. *3*, 241–251.

Rutschman, DH (1994). Dynamics of the Two-Locus Haploid Model. Theor. Popul. Biol. *45*, 167–176.

Schenk, M.F., Szendro, I.G., Salverda, M.L.M., Krug, J., de Visser, JAGM (2013). Patterns of epistasis between beneficial mutations in an antibiotic resistance gene. Mol. Biol. Evol. *8*, 419–426.

Suneson, CA (1956). An Evolutionary Plant Breeding Method1. Agron. J. *48*, 188.

De Visser, JAGM, Park, S, Krug, J (2009). Exploring the Effect of Sex on Empirical Fitness Landscapes. Am. Nat. *174*, S15–S30.

Weinreich, DM, Chao, L (2005). Rapid Evolutionary Escape by Large Populations from Local Fitness Peaks Is Likely in Nature. Evolution *59*, 1175–1182.

Weinreich, DM, Watson, RA, Chao, L (2005). Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. Evol. Int. J. Org. Evol. *59*, 1165–1174.

Weinreich, DM, Delaney, NF, DePristo, MA, Hartl, DL (2006). Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins. Science *312*, 111–114.

PARTIE V

DISCUSSION GENERALE ET PERSPECTIVES

V. DISCUSSION GENERALE ET PERSPECTIVES

V.1. Difficulté de la détection de l'épistasie

L'épistasie est difficile à détecter. Premièrement, sa détection nécessite un plus grand nombre de données que celle des les effets simples. Ainsi pour détecter l'épistasie entre deux locus A et B, il faut avoir à disposition un certain nombre d'individus porteurs de chaque génotype (AB, aB, Ab et ab). Cela se complique encore plus pour l'épistasie de deuxième ordre ou plus (entre plus de deux locus).

Plusieurs modèles de génétique quantitative classique ont été présentés dans le chapitre II. Chacun de ces modèles présentent des inconvénients et des avantages. Les modèles statistiques sont utiles pour estimer les variances et covariances génétiques, mais les effets qu'ils estiment dépendent de l'échantillon étudié. Il est alors difficile de comparer les effets génétiques d'un échantillon à l'autre, ou de se faire une idée du fonctionnement du gène à partir de ces effets statistiques. Les modèles fonctionnels sont censés pallier ce problème, mais l'un (la formulation fonctionnel du NOIA) dépend tout de même des fréquences alléliques et n'est orthogonal que dans certains cas, et l'autre (le modèle UWR) n'est jamais orthogonal. Or la non-orthogonalité induit une dépendance entre les effets, ce qui pose problème lors de l'estimation d'effets, puisque les estimations varient entre un modèle réduit et un modèle complet.

La plupart du temps, les modèles utilisés en étude d'association pangénomique sont relativement basiques. Il s'agit de la régression logistique (pour les traits binaires) ou de la régression linéaire (pour les traits continus) du phénotype par rapport au génotype du locus étudié, avec éventuellement ajout d'autres variables comme le sexe, l'âge, l'appartenance à un sous-groupe, etc. (Sikorska et al., 2013; Soller et al., 1976). Ce sont des modèles unilocus, qui sont appliqué à chaque marqueur individuellement. Ils sont généralement implémentés dans des logiciels comme PLINK (Purcell et al., 2007), QTL Cartographer (Basten et al., 2004), Mach2qtl (Li et al., 2009b, 2010) ou ProbABEL (Aulchenko et al., 2010), ou dans des packages R comme SNPassoc (González et al., 2007) ou GenABEL (Aulchenko et al., 2007). Seuls deux parmi eux permettent d'estimer l'épistasie : PLINK et SNPassoc. Le modèle implémenté dans QTL Cartographer, logiciel très populaire, est le modèle F2 (Cockerham, 1954). PLINK, utilise lui le modèle F∞ (Hayman, 1957), gérant différemment la dominance. Dans le cas haploïde, ces deux modèles sont similaires au nôtre.

Outre ces modèles basés sur la régression linéaire, SNP par SNP, d'autres approches ont été proposées, notamment pour les données d'association pangénomique. L'épistasie a connu un regain d'intérêt ces dernières années, quand le problème de l'héritabilité manquante a fait surface dans les études

d'association pangénomique (Manolio et al., 2009; Zuk et al., 2014). Une méthode développée tout récemment, propose de prendre une région génomique comme une unité de base (Zhang et al., 2014). L'idée n'est plus de tester toutes les interactions entre tous les SNPs deux à deux, mais de tester simultanément toutes les interactions possibles entre paires de SNPs au sein de deux régions génomiques. Des simulations ont démontré que l'erreur de type I était à un niveau correct (par rapport aux méthodes classiques de régression linéaire classique, qui ont une erreur de type I élevée), et que la capacité de détection étaient bien plus élevée qu'avec les analyses d'interaction classiques. Cette méthode, plus puissante, pourrait éventuellement être utilisée sur nos données d'expression chez le blé dur, sur lesquelles nous n'avons pas pu détecter de l'épistasie de premier ordre. Nous pourrions considérer qu'un gène est une région génomique, et tester simultanément l'interaction entre toutes les paires de SNPs possibles entre deux gènes. Dans notre cas, nous avons bien utilisé parfois plusieurs SNPs présents à chaque locus, mais en les sélectionnant pour être non liés par le déséquilibre de liaison. A la réflexion et à la lumière des travaux de Zhang et al., 2014, nous avons certainement manqué une occasion d'augmenter la puissance de détection des interactions. Nous reprendrons cet aspect pour la publication définitive des résultats.

V.2. Epistasie et pléiotropie sur le niveau d'expression de paires de gènes chez le blé dur

V.2.1. Détection d'épistasie SNP x fond génétique

Malgré le fait que de l'épistasie de premier ordre (SNP x SNP) soit probablement présente dans la population de pre-breeding de blé dur cultivée au laboratoire, nous n'en avons presque pas trouvé sur le niveau d'expression de paires de gènes (**Partie III**). Une première hypothèse concernant l'absence de détection d'effets épistatiques SNP x SNP est que le niveau d'expression d'un gène est peu dépendant de l'interaction avec un autre gène (bien que des effets *trans*, additifs, puissent exister) et n'est donc pas un caractère soumis à l'épistasie de premier ordre. L'organe sur lequel nous avons travaillé (germe) n'était peut-être pas un organe où l'expression est soumise à de l'épistasie de premier ordre. Néanmoins, il est plus probable qu'il y eût de l'épistasie SNP x SNP, mais que nous n'avons pas réussi à la détecter. Cette hypothèse est plus vraisemblable pour plusieurs raisons.

Premièrement, nous avons mis en évidence l'existence d'épistasie entre le fond génétique et des SNPs. En effet, nous avons trouvé, pour des paires d'homéologues dont les niveaux d'expression étaient corrélés, que les régressions $Y^A = f(Y^B)$ (où Y^A et Y^B sont les niveaux d'expression des gènes du génome A et du génome B, respectivement) étaient parfois allèle-spécifiques, se manifestant par des changements de pente de la régression en fonction de l'allèle. Un illustration de régression allèle-

spécifique (allèle du SNP B, B et b) est représenté **Figure 23**. Dans cet exemple, le niveau d'expression du gène B, Y^B, ne dépend que du fond génétique. Le déplacement le long de chacune des droites de régression correspond à un changement de fond génétique pour chacun des allèles du SNP appartenant au génome B. La distance entre les deux droites de régression peut être interprétée comme l'effet trans$B$→A. Dans cet exemple, comme les pentes des deux régressions sont différentes, cette distance varie, ce qui suggère une interaction avec d'autres régulateurs de l'expression du gène répartis ailleurs dans le génome. Nous pourrions donc considérer cette interaction comme le signe d'une somme d'effets d'interactions entre le locus B et l'ensemble de ces régulateurs, plutôt situés en *trans* ailleurs dans le génome. Sur la figure 23, trois fonds génétiques sont présentés : le 1, avec des effets du fond génétique inhibiteurs fixant Y^B à 0,2 (ronds verts) ; le 2, avec des effets du fond génétique intermédiaires, fixant Y^B à 0,6 (triangles bleus) ; et enfin, le 3, avec des effets du fond génétique activateurs, fixant Y^B à 1 (ronds jaunes). Si les deux droites de régression ne différaient que par leur ordonnée à l'origine, elles seraient parallèles, et cela ne serait le signe que d'un effet trans$B$→A simple. Dans notre étude, environ 4,6% (20 sur 439) des paires d'homéologues ayant des niveaux d'expression corrélés présentaient un tel profil. L'épistasie SNP x fond génétique n'est donc pas un phénomène complètement marginal. Si l'épistasie SNP x fond génétique existe, il est probable que l'épistasie SNP x SNP existe également mais qu'elle ait des effets difficilement détectables. D'ailleurs, 6 paires de gènes homéologues (1,4% des paires de gènes homéologues ayant des niveaux d'expression corrélés) présentaient des interactions SNP x SNP x fond génétique.

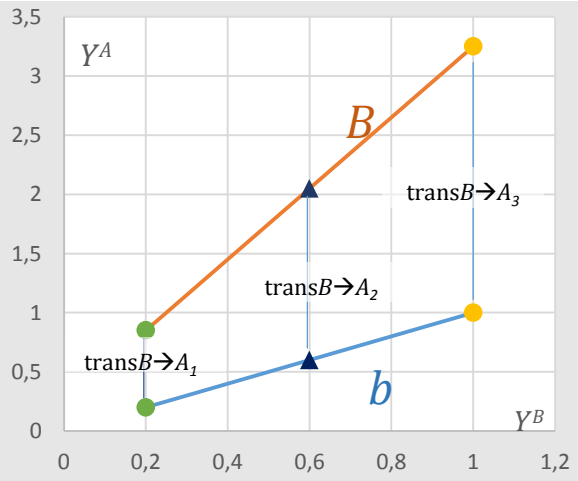


Figure 23 : Exemple de régressions $Y^A = f(Y^B)$ allèle-spécifique avant changement de pente.
Y^A et Y^B = niveaux d'expression des gènes du génome A et du génome B (gènes homéologues)

Deuxièmement, la majorité des études ayant cherché à mettre en évidence de l'épistasie de premier ordre (locus x locus) pour le niveau d'expression ont réussi à en détecter, parfois de manière importante. Cela a été le cas chez la levure (Brem and Kruglyak, 2005; Brem et al., 2005; Huang et al., 2012b; Storey et al., 2005), chez l'humain (Becker et al., 2012; Hemani et al., 2014) et chez *Plasmodium falciparum*, le parasite provoquant la malaria (Huang et al., 2012b). Une étude sur la levure, cependant, a trouvé peu d'effets épistatique (seulement 49 interactions parmi 5 182 eQTL détectés) (Zou et al., 2009). Notons également qu'avec la même méthode, Huang et al. 2012 ont trouvé beaucoup plus d'épistasie chez *Plasmodium falciparum* que chez la levure, ce qui indique une variabilité entre les espèces. Certains auteurs ont détecté l'épistasie en utilisant une méthode très différente de la nôtre (Brem and Kruglyak, 2005; Zou et al., 2009) : pour chaque gène, ils ont calculé l'effet de l'épistasie sur son niveau d'expression comme étant la différence entre le niveau d'expression des parents et celui des lignées descendantes. Mais certains ont utilisé des modèles fonctionnels, classiquement utilisés en génétique d'association, étendu à deux gènes (Huang et al., 2012b; Storey et al., 2005). Ces modèles sont relativement proches – voire identique, dans le cas haploïde – à celui que nous avons utilisé, le modèle UWR (Cheverud and Routman, 1995). Il est donc a priori surprenant que notre modèle n'ait pas réussi à détecter plus d'épistasie de premier ordre (SNP x SNP). Il est possible que notre échantillon fût trop petit. Par ailleurs, dans notre cas, le RNA seq qui génère un grand nombre de données manquantes a réduit clairement la puissance de notre dispositif sur le blé dur. Les individus viennent d'être génotypés par la puce Axiom de 480K mis au point sur le blé tendre dans le cadre du programme BreedWheat (Paux, 2013). Reprendre nos analyses avec ces nouvelles données nous permettra peut-être de détecter plus de situations significatives. Enfin contrairement aux études classiques d'eQTL, nous n'avons pas cherché à détecter tous les eQTL *cis*, *trans* ou épistatiques existants, mais de déterminer certains motifs dans la régulation de l'expression. Nous n'avons donc pas cherché à tester toutes les paires, et nous n'avons par exemple pas étudié les paires de gènes appartenant au même génome, dans lesquelles plus d'épistasie serait peut-être détectée. Cela est notre prochain objectif.

V.2.2. Détection de pléiotropie

Sans avoir recherché de manière exhaustive tous les effets régulateurs, nous avons pu mettre en évidence un certain nombre d'effets régulateurs *cis* et *trans*. Les effets *trans* sont plus souvent détectés entre paires de gènes homéologues qu'entre paires de gènes non-homéologues. Par ailleurs, plus d'eQTL, *cis* et *trans*, ont été trouvés sur le génome B que sur le génome A, révélant un rôle asymétrique des deux génomes sur la régulation de l'expression.

Les effets *trans* peuvent être interprétés en termes de pléiotropie (le fait qu'un gène puisse avoir un effet sur plusieurs caractères). D'abord, au niveau transcriptionnel, un certain nombre de SNPs avait à la fois un effet sur l'expression du gène auquel ils appartenaient (effet *cis*) et sur l'expression d'un

autre gène (effet *trans*) ; certains pouvaient aussi avoir un effet *trans* sur l'expression de plusieurs gènes. Cela signifie que ces SNPs avaient un effet pléiotrope, puisqu'ils agissaient sur plusieurs caractères : les niveaux d'expression de deux ou plusieurs gènes. Plus généralement, ces effets suggèrent que des gènes ont une influence sur l'expression d'un autre gène. Si, au niveau phénotypique, ces deux gènes ne sont pas directement impliqués dans le même caractère, alors le gène régulateur sera pléiotrope.

V.2.3. Les réseaux de gènes, la prochaine étape ?

Nous avons donc réussi à mettre en évidence de l'épistasie (locus x fond génétique) et la pléiotropie. Or ces phénomènes seraient des propriétés inhérentes aux réseaux de gènes, et seraient donc quasi-omniprésentes (Tyler et al., 2009). Par ailleurs, les effets trans pourraient indiquer la présence de réseaux de gènes, puisqu'ils pourraient correspondre à un lien entre deux nœuds d'un réseau (= gènes). Une étude simulant des réseaux à trois gènes a montré que ceux-ci pouvaient générer de la variance épistatique, en particulier si ces réseaux constituent des boucles de rétroaction positive (Gjuvsland et al., 2007). La prochaine étape de notre travail serait donc d'inférer ces réseaux de gènes.

Les réseaux de gènes ont une structure bien particulière : ce sont des « réseaux invariant d'échelle » (McKinney and Pajewski, 2012; Tyler et al., 2009). Autrement dit, un faible nombre de gènes interagisse avec beaucoup d'autres gènes. On appelle ces gènes des « concentrateurs » (hub). Ces « concentrateurs » n'ont pas nécessairement un fort effet additif sur le phénotype, mais sont d'une importance capitale à cause de l'effet modulateur qu'ils ont sur un nombre important de gènes impliqués dans le caractère.

Les réseaux invariant d'échelle se retrouvent dans d'autres systèmes biologiques, comme les interactions protéine-protéine (Wuchty, 2002), les réseaux métaboliques (Jeong et al., 2000) ou les interactions au niveau écologique (Montoya and Solé, 2002). Dans ces réseaux, le chemin le plus court entre deux nœuds (= gènes pour un réseau de gènes) est généralement plus court que dans les autres types de réseaux (Tyler et al., 2009). Cela permet une transmission rapide de l'information. Si ces réseaux sont particulièrement vulnérables à la suppression ou à la dégradation d'un concentrateur (chez la levure, il a été démontré que la délétion d'un concentrateur conduisait plus fréquemment à la létalité que la délétion d'un gène non concentrateur (Jeong et al., 2001)), ils sont particulièrement résistants à la suppression des autres nœuds. Ces réseaux sont donc particulièrement robustes à des mutations aléatoires car ils n'ont que quelques points faibles (les concentrateurs).

Notre travail sur le blé dur a permis de trouver des résultats intéressants : les effets *trans* sont plus nombreux entre gènes homéologues, ces derniers ont fréquemment des niveaux d'expression corrélés (ce qui revient très probablement à dire que les gènes homéologues ont tendance à appartenir

fréquemment aux mêmes réseaux de gènes co-regulés), et le génome B contient plus de eQTL. Un travail plus exhaustif sur les réseaux de gènes, permettant de connaître plus en détails leur structure, serait intéressant sur cette espèce polyploïde. Il serait notamment pertinent de comprendre quelle place ont les génomes ou les homéologues dans la structure ces réseaux. Ce type de travail pourrait confirmer (ou infirmer) que les gènes homéologues ont plus de chances d'appartenir au même réseau de gènes que des gènes aléatoires. Il pourrait permettre de déterminer si les homéologues ont tendance à avoir le même rôle au sein des réseaux, et notamment s'ils sont généralement reliés à peu près au même nombre de gènes. Enfin, on pourrait déterminer si par exemple, les gènes concentrateurs, appartiennent plus souvent au génome B ou au génome A.

Pour établir un réseau de gènes, il faudrait d'abord faire une recherche exhaustive de tous les eQTL existants, et notamment de tous les eQTL *trans*. Or, il est techniquement impossible de tester toutes les paires de SNP deux à deux : non seulement, cela demanderait un temps de calculs trop important, mais en plus, la puissance des multiples tests statistiques serait sérieusement réduite par leur nombre extrêmement importants. La méthode présentée ci-dessus en V.1 (Zhang et al., 2014), permettant de tester simultanément toutes les interactions possibles entre paires de SNPs au sein de deux gènes ou de deux régions génomiques, peut être une solution pour gagner en temps de calcul et en puissance. Il existe d'autres méthodes pour réduire la dimensionnalité des données. Un type d'approche est « l'extraction de caractéristiques » (feature extraction), qui consiste à projeter les données de hautes dimensions dans un espace à plus faible dimensions. L'utilisation de composantes principales a ainsi été suggérée (Li et al., 2009a). Mais l'une des méthodes les plus courantes pour traiter les données de séquençage haut débit est la « méthode de regroupement » (collapsing method), dont le principe est de regrouper plusieurs SNPs en une variable (Bansal et al., 2010; Li and Leal, 2008; Luo et al., 2011; Wu et al., 2011). Le désavantage de cette méthode est qu'elle peut engendrer une importante perte d'informations (Zhang et al., 2014). Un autre type d'approche est la « sélection de caractéristiques » (feature selection), dont l'objectif est de sélectionner les variables les plus pertinentes (He and Lin, 2011; Moore, 2007). Contrairement aux approches d' « extraction », il n'y a pas de transformation des données. Enfin, de nouveaux algorithmes de calcul sont aussi développés pour réduire le temps de calculs (Prabhu and Pe'er, 2012).

Une fois que les eQTL sont détectés, il faut pouvoir identifier le gène sous-jacent de l'eQTL. En effet, si un gène est en fort déséquilibre de liaison avec un gène agissant en *trans* sur l'expression d'un autre gène, alors le premier sera statistiquement aussi détecté comme un *trans*-régulateur, bien qu'il n'ait pas forcément d'influence biologique sur l'expression d'autres gènes. Autrement dit, si plusieurs gènes en fort déséquilibre de liaison ont un effet trans sur le même gène, il se peut qu'il s'agisse d'un artefact dû au déséquilibre de liaison, et qu'un seul de ces gènes soit réellement régulateur. Il est alors probable

que le gène régulateur soit celui qui contient le SNP ayant l'effet le plus significatif sur le niveau d'expression du gène régulé.

A priori, les réseaux de gènes pourront être construits simplement en reliant tous les gènes régulateurs en *trans* aux gènes régulés. Des « hotspots » de régulations en *trans*, pléiotropiques, ont pu être mis en évidence dans plusieurs études sur les eQTL (Drost et al., 2010; Schadt et al., 2003; West et al., 2007). Ces hotspots correspondent aux gènes concentrateurs des réseaux, mentionnés ci-dessus. Les gènes ayant un effet interactif (épistasie) avec un autre gène devront également figurer dans les réseaux.

Dans les faits, cela pourrait être plus compliqué, puisqu'il faudra distinguer les effets indirects des effets directs : biologiquement, un premier gène peut modifier l'expression d'un deuxième gène, qui aura lui-même un effet sur un troisième gène. Or, le premier gène pourra être détecté comme un régulateur du troisième gène, alors qu'il ne le régule pas directement.

Par ailleurs, l'approche eQTL ne permet pas de couvrir l'ensemble des relations entre gènes. Premièrement, les effets *trans* sont des effets additifs. Mais un gène peut réguler un autre, non pas en ayant un effet +/- sur l'expression d'un autre gène, mais via un effet se modélisant par une équation différentielle. Ce type de relation entre niveaux d'expression génétique pourrait se détecter via la mise en évidence de corrélation entre niveaux d'expression. Néanmoins, la mise en évidence de corrélation ne permet pas de définir de relations causales : un gène A et un gène B peuvent avoir un niveau d'expression corrélé parce que (i) le gène A régule le gène B, (ii) le gène B régule le gène A, (iii) les deux gènes sont régulés par le même facteur. Notons que nous avons à notre disposition le niveau d'expression de gènes non polymorphes, qui représentaient 71% des gènes transcrits dans notre jeu de données. Ces gènes peuvent être des régulateurs, mais en absence de polymorphisme ils ne pourront pas être détectés par la méthode de détection d'eQTL ; ils peuvent également être régulés. Ils peuvent donc appartenir à des réseaux de gènes. La façon dont les données d'expression de ces gènes non polymorphes pourraient être exploitées reste à approfondir. Un modèle unilocus peut permettre de déterminer si des gènes polymorphes ont un effet sur ces gènes non polymorphes. En revanche, il sera sans doute plus difficile de tester si ces gènes non polymorphes ont un effet causal sur l'expression d'autres gènes.

Plusieurs auteurs ont proposé des méthodes pour inférer des réseaux de gènes à partir d'étude d'eQTL (Huang et al., 2009). Par ailleurs, certains proposent de combiner les données d'expression et les données de génotypage pour reconstruire de manière optimale les relations de causes à effets entre données d'expression et maladie (Schadt et al., 2005; Zhu et al., 2007). Ces méthodes peuvent peut-être également être utilisées pour l'inférence de réseaux de gènes.

V.3. Rôle de l'épistasie fonctionnelle dans l'adaptation et application au pre-breeding

L'épistasie peut tenir un rôle important dans l'évolution, notamment quand elle induit une incompatibilité ou une dépression hybride (épistasie de signe réciproque). L'article *Reciprocal sign epistasis and truncation selection: when is recombination favourable in a pre-breeding program with a selfing species?* (**Partie IV.2**) a montré qu'en présence de sélection par troncation, l'épistasie de signe réciproque pouvait induire la fixation d'un génotype sous-optimal, notamment si le taux de recombinaison était élevé.

Des exemples de dépression ou d'incompatibilité hybride ont pu être mis en évidence chez le blé tendre. Cette dépression/incompatibilité peut se présenter sous forme de nécroses (Hermsen, 1963a, 1963b; Heyne et al., 1943; Iehisa et al., 2014; Tsunewaki, 1960). Deux gènes, interagissant pour provoquer ces nécroses, ont été identifiés : Ne1 et Ne2, situés respectivement sur les chromosomes 5BL et 2BS (Nishikawa et al.; Tsunewaki, 1960). Dans d'autres cas, les hybrides ressemblent à des « touffes d'herbe naines ». Ce nanisme est en réalité provoqué par une nécrose inactivant le méristème apical (Canvin and Mcvetty, 1976). Par ailleurs, une légère dépression hybride a pu être observée lors de croisements entre populations de gestion dynamique de blé tendre, ayant divergé depuis 12 ans, pour des caractères liés au rendement (Rousselle et al., 2011). Il est très vraisemblable de tels phénomènes existent aussi chez le blé dur.

Par conséquent, il est probable que lors de la création d'une population de pre-breeding de blé dur, issue de croisement de variétés récentes avec des individus ancestraux ou sauvages, de telles incompatibilités ou dépressions hybrides apparaissent. Il est donc quasiment certain que le paysage adaptatif d'une population de pre-breeding d'une espèce autogame (plus sujette à l'accumulation d'incompatibilité de Dobzhansky-Müller) comporte des pics adaptatifs.

Le modèle présenté dans l'article est basique, puisque le déterminisme génétique est bilocus. Il ne représente donc pas fidèlement la réalité. Néanmoins, il peut tout de même permettre de comprendre certains mécanismes. Ainsi, si comme dans l'article, nous sommes en présence d'un cas où la combinaison génotypique optimale n'est pas présente dans la population, et qu'il existe une vallée adaptative entre la combinaison optimale et sous-optimale, il est recommandé de ne pas sélectionner dès les premières générations, qui devraient être consacrées au cassage du déséquilibre de liaison via la gestion de l'allofécondation par un gène de stérilité mâle. Quand le déséquilibre de liaison entre locus distants commence à être réduit, une phase de sélection, faible pour éviter des pertes via la dérive, pourra commencer.

Dans la réalité, les caractères phénotypes sont souvent contrôlés par de multiples locus. Un modèle plus complexe, comprenant des effets additifs et des effets épistatiques sur plusieurs locus, permettrait peut-être d'appréhender de manière plus précise le comportement d'une population de pre-breeding d'une espèce autogame. Grâce à ce modèle plus réaliste, il serait possible de déterminer l'effet du régime de reproduction (donc de la recombinaison) sur l'évolution d'une telle population en fonction de l'importance des effets épistatiques par rapport aux effets additifs. Ces travaux de modélisation pourraient être complétés par des expérimentations aux champs. Des populations d'admixture (blé dur et amidonnier) pourraient être cultivées dans un même environnement et soumise à une légère sélection anthropique pour un caractère d'intérêt (par exemple, on pourrait sélectionner pour un idéotype cultivé de petite taille, mais présentant des caractères favorables hérités de l'amidonnier, comme une plus grande robustesse, une meilleure adaptation à un mode de culture à faibles intrants, ou une meilleure résistance à une maladie). Dans les différentes populations d'admixture, un taux d'allofécondation différent pourrait être fixé. Un suivi évolutif, sur le plan phénotypique et sur le plan génotypique, pourrait être mis en place, afin de suivre l'évolution des fréquences alléliques et du phénotype. Au bout de quelques générations, il sera peut-être possible de déterminer si l'une des populations s'est mieux adaptée, en fonction du taux d'allofécondation.

V.4. Conclusion

L'épistasie a longtemps été négligée, et pourtant, elle joue un rôle extrêmement important d'un point de vue évolutif, puisqu'elle permet notamment d'expliquer un phénomène comme la spéciation. Elle est également indispensable à la compréhension de l'architecture génétique des traits phénotypiques. En effet, les composés biologiques (gènes, protéines, métabolites) interagissent les uns avec les autres via des réseaux complexes. Afin d'élucider comment des gènes peuvent expliquer un phénotype, il est donc nécessaire de comprendre comment ces composés interagissent, et d'étudier les effets des gènes dans leur contexte. Enfin, l'impact d'autres phénomènes non mendéliens, comme l'épigénétique ou les interactions génotype x environnement, restent encore à explorer.

PARTIE VI

REFERENCES

VI. References

Adams, K.L., Cronn, R., Percifield, R., and Wendel, J.F. (2003). Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. Proc. Natl. Acad. Sci. *100*, 4649–4654.

Adams, K.L., Percifield, R., and Wendel, J.F. (2004). Organ-Specific Silencing of Duplicated Genes in a Newly Synthesized Cotton Allotetraploid. Genetics *168*, 2217–2226.

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

Allard, R.W., García, P., Sáenz-de-Miera, L.E., and Vega, M.P. de la (1993). Evolution of multilocus genetic structure in Avena hirtula and Avena barbata. Genetics *135*, 1125–1139.

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

Anderson, V.L., and Kempthorne, O. (1954). A Model for the Study of Quantitative Inheritance. Genetics *39*, 883–898.

Angeli, D., Ferrell, J.E., and Sontag, E.D. (2004). Detection of multistability, bifurcations, and hysteresis in a large class of biological positive-feedback systems. Proc. Natl. Acad. Sci. U. S. A. *101*, 1822–1827.

Anholt, R.R.H., Dilda, C.L., Chang, S., Fanara, J.-J., Kulkarni, N.H., Ganguly, I., Rollmann, S.M., Kamdar, K.P., and Mackay, T.F.C. (2003). The genetic architecture of odor-guided behavior in Drosophila: epistasis and the transcriptome. Nat. Genet. *35*, 180–184.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene Ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.

Aulchenko, Y.S., Ripke, S., Isaacs, A., and Duijn, C.M. van (2007). GenABEL: an R library for genome-wide association analysis. Bioinformatics *23*, 1294–1296.

Aulchenko, Y.S., Struchalin, M.V., and Duijn, C.M. van (2010). ProbABEL package for genome-wide association analysis of imputed data. BMC Bioinformatics *11*, 134.

Bading, H., Ginty, D.D., and Greenberg, M.E. (1993). Regulation of gene expression in hippocampal neurons by distinct calcium signaling pathways. Science *260*, 181–186.

Bansal, V., Libiger, O., Torkamani, A., and Schork, N.J. (2010). Statistical analysis strategies for association studies involving rare variants. Nat. Rev. Genet. *11*, 773–785.

Barton, N. h., and Turelli, M. (2004). Effects of Genetic Drift on Variance Components Under a General Model of Epistasis. Evolution *58*, 2111–2132.

Basten, C., Weir, B., and Zeng, Z.B. (2004). QTL cartographer. Dep. Stat. N. C. State Univ. Raleigh NC.

Bateson, W. (1909). Heredity and variation in modern lights. Darwin Mod. Sci. 85–101.

Bateson, W., and Mendel, G. (1909). Mendel's Principles of Heredity (Courier Dover Publications).

Becker, J., Wendland, J.R., Haenisch, B., Nöthen, M.M., and Schumacher, J. (2012). A systematic eQTL study of cis–trans epistasis in 210 HapMap individuals. Eur. J. Hum. Genet. *20*, 97–101.

Belkhir, K., Borsa, P., Chikhi, L., Raufaste, N., and Bonhomme, F. (1996). GENETIX 4.05, logiciel sous Windows TM pour la génétique des populations. Lab. Génome Popul. Interact. CNRS UMR *5000*, 1996–2004.

Bikard, D., Patel, D., Metté, C.L., Giorgi, V., Camilleri, C., Bennett, M.J., and Loudet, O. (2009). Divergent Evolution of Duplicate Genes Leads to Genetic Incompatibilities Within A. thaliana. Science *323*, 623–626.

Birchler, J.A., and Veitia, R.A. (2007). The Gene Balance Hypothesis: From Classical Genetics to Modern Genomics. Plant Cell Online *19*, 395–402.

Bloom, J.S., Ehrenreich, I.M., Loo, W.T., Lite, T.-L.V., and Kruglyak, L. (2013). Finding the sources of missing heritability in a yeast cross. Nature *494*, 234–237.

Bnejdi, F., and Gazzah, M.E. (2008). Inheritance of resistance to yellowberry in durum wheat. Euphytica *163*, 225–230.

Bodmer, W.F., and Felsenstein, J. (1967). Linkage and Selection: Theoretical Analysis of the Deterministic Two Locus Random Mating Model. Genetics *57*, 237–265.

Bomblies, K., and Weigel, D. (2007). Hybrid necrosis: autoimmunity as a potential gene-flow barrier in plant species. Nat. Rev. Genet. *8*, 382–393.

Bomblies, K., Lempe, J., Epple, P., Warthmann, N., Lanz, C., Dangl, J.L., and Weigel, D. (2007). Autoimmune Response as a Mechanism for a Dobzhansky-Muller-Type Incompatibility Syndrome in Plants. PLoS Biol *5*, e236.

Boulc'h, V.L., David, J.L., Brabant, P., and Vallavieille-Pope, C. de (1994). Dynamic conservation of variability: responses of wheat populations to different selective forces including powdery mildew. Genet. Sel. Evol. *26*, 1–20.

Brem, R.B., and Kruglyak, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proc. Natl. Acad. Sci. U. S. A. *102*, 1572–1577.

Brem, R.B., Storey, J.D., Whittle, J., and Kruglyak, L. (2005). Genetic interactions between polymorphisms that affect gene expression in yeast. Nature *436*, 701–703.

Brenchley, R., Spannagl, M., Pfeifer, M., Barker, G.L.A., D'Amore, R., Allen, A.M., McKenzie, N., Kramer, M., Kerhornou, A., Bolser, D., et al. (2012). Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature *491*, 705–710.

Brideau, N.J., Flores, H.A., Wang, J., Maheshwari, S., Wang, X., and Barbash, D.A. (2006). Two Dobzhansky-Muller Genes Interact to Cause Hybrid Lethality in Drosophila. Science *314*, 1292–1295.

Briggs, W.H., and Goldman, I.L. (2006). Genetic Variation and Selection Response in Model Breeding Populations of Brassica rapa Following a Diversity Bottleneck. Genetics *172*, 457–465.

Bryant, E.H., McCommas, S.A., and Combs, L.M. (1986). The Effect of an Experimental Bottleneck upon Quantitative Genetic Variation in the Housefly. Genetics *114*, 1191–1211.

Buckler, E.S., Thornsberry, J.M., and Kresovich, S. (2001). Molecular Diversity, Structure and Domestication of Grasses. Genet. Res. *77*, 213–218.

Buggs, R.J.A., Chamala, S., Wu, W., Gao, L., May, G.D., Schnable, P.S., Soltis, D.E., Soltis, P.S., and Barbazuk, W.B. (2010). Characterization of duplicate gene evolution in the recent natural allopolyploid Tragopogon miscellus by next-generation sequencing and Sequenom iPLEX MassARRAY genotyping. Mol. Ecol. *19 Suppl 1*, 132–146.

Bürger, R. (1989). Linkage and the maintenance of heritable variation by mutation-selection balance. Genetics *121*, 175–184.

Caballero, A., and Hill, W.G. (1992). Effective size of nonrandom mating populations. Genetics *130*, 909–916.

Cabot, E.L., Davis, A.W., Johnson, N.A., and Wu, C.I. (1994). Genetics of reproductive isolation in the Drosophila simulans clade: complex epistasis underlying hybrid male sterility. Genetics *137*, 175–189.

Canvin, D.T., and Mcvetty, P.B.E. (1976). Hybrid grass-clump dwarfness in wheat: Physiology and genetics. Euphytica *25*, 471–483.

Cao, G., Zhu, J., He, C., Gao, Y., Yan, J., and Wu, P. (2001). Impact of epistasis and QTL×environment interaction on the developmental behavior of plant height in rice (Oryza sativa L.). Theor. Appl. Genet. *103*, 153–160.

Carlborg, Ö., and Haley, C.S. (2004). Epistasis: too often neglected in complex trait studies? Nat. Rev. Genet. *5*, 618–625.

Carlborg, Ö., Kerje, S., Schütz, K., Jacobsson, L., Jensen, P., and Andersson, L. (2003). A Global Search Reveals Epistatic Interaction Between QTL for Early Growth in the Chicken. Genome Res. *13*, 413–421.

Carrillo, J.M., Rousset, M., Qualset, C.O., and Kasarda, D.D. (1990). Use of recombinant inbred lines of wheat for study of associations of high-molecular-weight glutenin subunit alleles to quantitative traits. Theor. Appl. Genet. *79*, 321–330.

Cavanagh, C., Morell, M., Mackay, I., and Powell, W. (2008). From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. Curr. Opin. Plant Biol. *11*, 215–221.

Chagué, V., Just, J., Mestiri, I., Balzergue, S., Tanguy, A.-M., Huneau, C., Huteau, V., Belcram, H., Coriton, O., Jahier, J., et al. (2010). Genome-wide gene expression changes in genetically stable synthetic and natural wheat allohexaploids. New Phytol. *187*, 1181–1194.

Chang, P.L., Dilkes, B.P., McMahon, M., Comai, L., and Nuzhdin, S.V. (2010). Homoeolog-specific retention and use in allotetraploid Arabidopsis suecica depends on parent of origin and network partners. Genome Biol. *11*, R125.

Charlesworth, D., and Pannell, J.R. (2001). Mating systems and population genetic structure in the light of coalescent theory. In Integrating Ecology and Evolution in a Spatial Context: 14th Special Symposium of the British Ecological Society, (Cambridge University Press),.

Charmet, G. (2011). Wheat domestication: Lessons for the future. C. R. Biol. *334*, 212–220.

Chaudhary, B., Flagel, L., Stupar, R.M., Udall, J.A., Verma, N., Springer, N.M., and Wendel, J.F. (2009). Reciprocal Silencing, Transcriptional Bias and Functional Divergence of Homeologs in Polyploid Cotton (Gossypium). Genetics *182*, 503–517.

Chen, Z.J. (2007). Genetic and Epigenetic Mechanisms for Gene Expression and Phenotypic Variation in Plant Polyploids. Annu. Rev. Plant Biol. *58*, 377–406.

Chen, Z.J., and Ni, Z. (2006). Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. BioEssays *28*, 240–252.

Chen, Z.J., and Pikaard, C.S. (1997). Epigenetic silencing of RNA polymerase I transcription: a role for DNA methylation and histone modification in nucleolar dominance. Genes Dev. *11*, 2124–2136.

Chen, X., Hackett, C.A., Niks, R.E., Hedley, P.E., Booth, C., Druka, A., Marcel, T.C., Vels, A., Bayer, M., Milne, I., et al. (2010). An eQTL Analysis of Partial Resistance to Puccinia hordei in Barley. PLoS ONE *5*, e8598.

Cheng, F., Wu, J., Fang, L., Sun, S., Liu, B., Lin, K., Bonnema, G., and Wang, X. (2012). Biased Gene Fractionation and Dominant Gene Expression among the Subgenomes of Brassica rapa. PLoS ONE *7*, e36442.

Chesler, E.J., Lu, L., Shou, S., Qu, Y., Gu, J., Wang, J., Hsu, H.C., Mountz, J.D., Baldwin, N.E., Langston, M.A., et al. (2005). Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. Nat. Genet. *37*, 233–242.

Cheung, V.G., Spielman, R.S., Ewens, K.G., Weber, T.M., Morley, M., and Burdick, J.T. (2005). Mapping determinants of human gene expression by regional and genome-wide association. Nature *437*, 1365–1369.

Cheverud, J.M. (2000). Detecting epistasis among quantitative trait loci. In Epistasis and the Evolutionary Process, (Oxford University Press), pp. 58–81.

Cheverud, J.M., and Routman, E.J. (1995). Epistasis and its contribution to genetic variance components. Genetics *139*, 1455–1461.

Cheverud, J.M., and Routman, E.J. (1996). Epistasis as a Source of Increased Additive Genetic Variance at Population Bottlenecks. Evolution *50*, 1042.

Cheverud, J.M., Vaughn, T.T., Pletscher, L.S., King-Ellison, K., Bailiff, J., Adams, E., Erickson, C., and Bonislawski, A. (1999). Epistasis and the Evolution of Additive Genetic Variance in Populations That Pass Through a Bottleneck. Evolution *53*, 1009.

Chouteau, M., and Angers, B. (2012). Wright's Shifting Balance Theory and the Diversification of Aposematic Signals. PLoS ONE *7*, e34028.

Cockerham, C.C. (1954). An Extension of the Concept of Partitioning Hereditary Variance for Analysis of Covariances among Relatives When Epistasis Is Present. Genetics *39*, 859–882.

Colegrave, N., and Buckling, A. (2005). Microbial experiments on adaptive landscapes. BioEssays News Rev. Mol. Cell. Dev. Biol. *27*, 1167–1173.

Comai, L. (2005). The advantages and disadvantages of being polyploid. Nat. Rev. Genet. *6*, 836–846.

Comai, L., Tyagi, A.P., Winter, K., Holmes-Davis, R., Reynolds, S.H., Stevens, Y., and Byers, B. (2000). Phenotypic Instability and Rapid Gene Silencing in Newly Formed Arabidopsis Allotetraploids. Plant Cell Online *12*, 1551–1567.

Conti, V., Roncallo, P.F., Beaufort, V., Cervigni, G.L., Miranda, R., Jensen, C.A., and Echenique, V.3. (2011). Mapping of main and epistatic effect QTLs associated to grain protein and gluten strength using a RIL population of durum wheat. J. Appl. Genet. *52*, 287–298.

Cordell, H.J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum. Mol. Genet. *11*, 2463–2468.

Costa e Silva, J., Potts, B.M., and Tilyard, P. (2012). Epistasis causes outbreeding depression in eucalypt hybrids. Tree Genet. Genomes *8*, 249–265.

Cowles, C.R., Hirschhorn, J.N., Altshuler, D., and Lander, E.S. (2002). Detection of regulatory variation in mouse genes. Nat. Genet. *32*, 432–437.

Coyne, J.A., Barton, N.H., and Turelli, M. (1997). Perspective: A Critique of Sewall Wright's Shifting Balance Theory of Evolution. Evolution *51*, 643.

Coyne, J.A., Barton, N.H., and Turelli, M. (2000). Is Wright's shifting balance process important in evolution? Evolution *54*, 306–317.

Crow, J.F. (2010). On epistasis: why it is unimportant in polygenic directional selection. Philos. Trans. R. Soc. B Biol. Sci. *365*, 1241–1244.

Crow, J.F., and Kimura, M. (1970). An introduction to population genetics theory. (Minneapolis).

Cubillos, F.A., Coustham, V., and Loudet, O. (2012). Lessons from eQTL mapping studies: non-coding regions and their role behind natural phenotypic variation in plants. Curr. Opin. Plant Biol. *15*, 192–198.

Cui, L., Wall, P.K., Leebens-Mack, J.H., Lindsay, B.G., Soltis, D.E., Doyle, J.J., Soltis, P.S., Carlson, J.E., Arumuganathan, K., Barakat, A., et al. (2006). Widespread genome duplications throughout the history of flowering plants. Genome Res. *16*, 738–749.

Darwin, C. (1859). On the Origin of Species. (London).

Daud, H.M., and Gustafson, J.P. (1996). Molecular evidence for Triticum speltoides as a B-genome progenitor of wheat (Triticum aestivum). Genome *39*, 543–548.

Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M., and Blaxter, M.L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. *12*, 499–510.

David, J., Tavaud-Pirra, M., Roumet, P., Muller, M.-H., Santoni, S., Sarah, G., Holtz, Y., Ranwez, V., Ardisson, M., Poux, G., et al. (in press). Broadening the genetic basis of durum wheat.

David, J., Holtz, Y., Ranwez, V., Santoni, S., Sarah, G., Ardisson, M., Poux, G., Choulet, F., Genthon, C., Roumet, P., et al. (in press). Genotyping by sequencing transcriptomes in an evolutionary pre-breeding durum wheat population. Mol. Breed.

David, J.L., Zivy, M., Cardin, M.L., and Brabant, P. (1997). Protein evolution in dynamically managed populations of wheat: adaptive responses to macro-environmental conditions. Theor. Appl. Genet. *95*, 932–941.

DeCook, R., Lall, S., Nettleton, D., and Howell, S.H. (2006). Genetic Regulation of Gene Expression During Shoot Development in Arabidopsis. Genetics *172*, 1155–1164.

Dettman, J.R., Sirjusingh, C., Kohn, L.M., and Anderson, J.B. (2007). Incipient speciation by divergent adaptation and antagonistic epistasis in yeast. Nature *447*, 585–588.

Dobzhansky, T. (1934). Studies on Hybrid Sterility. I. Spermatogenesis in pure and hybrid Drosophila pseudoobscura. Z. Für Zellforsch. Mikrosk. Anat. *21*.

Dolgin, E.S., Charlesworth, B., Baird, S.E., and Cutter, A.D. (2007). Inbreeding and Outbreeding Depression in Caenorhabditis Nematodes. Evolution *61*, 1339–1352.

Drost, D.R., Benedict, C.I., Berg, A., Novaes, E., Novaes, C.R.D.B., Yu, Q., Dervinis, C., Maia, J.M., Yap, J., Miles, B., et al. (2010). Diversification in the genetic architecture of gene expression and transcriptional networks in organ differentiation of Populus. Proc. Natl. Acad. Sci. *107*, 8492–8497.

Dubcovsky, J., and Dvorak, J. (2007). Genome Plasticity a Key Factor in the Success of Polyploid Wheat Under Domestication. Science *316*, 1862–1866.

Dvořák, J., McGuire, P.E., and Cassidy, B. (1988). Apparent sources of the A genomes of wheats inferred from polymorphism in abundance and restriction fragment length of repeated nucleotide sequences. Genome *30*, 680–689.

Dvořák, J., Terlizzi, P. di, Zhang, H.-B., and Resta, P. (1993). The evolution of polyploid wheats: identification of the A genome donor species. Genome *36*, 21–31.

Edmands, S., and Timmerman, C.C. (2003). Factores de Modelaje que Afectan la Severidad de la Depresión Exogámica. Conserv. Biol. *17*, 883–892.

Engelmann, F., and Engels, J.M.M. (2002). Technologies and strategies for ex situ conservation. In Managing Plant Genetic Diversity, V.R. Rao, A.H.D. Brown, and M. Jackson, eds. (CABI), pp. 89–103.

Enjalbert, J., and David, J.L. (2000). Inferring Recent Outcrossing Rates Using Multilocus Individual Heterozygosity: Application to Evolving Wheat Populations. Genetics *156*, 1973–1982.

Enjalbert, J., Goldringer, I., David, J., and Brabant, P. (1998). The relevance of outcrossing for the dynamic management of genetic resources in predominantly selfing Triticum aestivum L. (bread wheat). Genet. Sel. Evol. *30*, 1–15.

Escobar, J.S., Cenci, A., Bolognini, J., Haudry, A., Laurent, S., David, J., and Glémin, S. (2010). An integrative test of the dead-end hypothesis of selfing evolution in Triticeae (Poaceae). Evol. Int. J. Org. Evol. *64*, 2855–2872.

Falconer, D.S., and Mackay, T.F.C. (1996). Introduction to Quantitative Genetics (Essex, England: Benjamin Cummings).

FAO (1996). The state of the world's plant genetic resources for food and agriculture (Rome: FAO).

Fawcett, J.A., Maere, S., and Peer, Y.V. de (2009). Plants with double genomes might have had a better chance to survive the Cretaceous–Tertiary extinction event. Proc. Natl. Acad. Sci. *106*, 5737–5742.

Fedorowicz, G.M., Fry, J.D., Anholt, R.R.H., and Mackay, T.F.C. (1998). Epistatic Interactions Between smell-impaired Loci in Drosophila melanogaster. Genetics *148*, 1885–1891.

Feldman, M.W. (1971). Equilibrium studies of two locus haploid populations with recombination. Theor. Popul. Biol. *2*, 299–318.

Feldman, M., and Levy, A.A. (2009). Genome evolution in allopolyploid wheat—a revolutionary reprogramming followed by gradual changes. J. Genet. Genomics *36*, 511–518.

Feldman, M., Liu, B., Segal, G., Abbo, S., Levy, A.A., and Vega, J.M. (1997). Rapid Elimination of Low-Copy DNA Sequences in Polyploid Wheat: A Possible Mechanism for Differentiation of Homoeologous Chromosomes. Genetics *147*, 1381–1387.

Fenster, C.B., and Galloway, L.F. (2000). Population Differentiation in an Annual Legume: Genetic Architecture. Evolution *54*, 1157–1172.

Fenster, C.B., Galloway, L.F., and Chao, L. (1997). Epistasis and its consequences for the evolution of natural populations. Trends Ecol. Evol. *12*, 282–286.

Feuillet, C., Langridge, P., and Waugh, R. (2008). Cereal breeding takes a walk on the wild side. Trends Genet. *24*, 24–32.

Fisher, R.A. (1918). The Correlation between Relatives on the Supposition of Mendelian Inheritance. Trans. R. Soc. Edinb. *52*, 399–433.

Fishman, L., and Willis, J.H. (2001). Evidence for Dobzhansky-Muller Incompatibilites Contributing to the Sterility of Hybrids Between Mimulus Guttatus and M. Nasutus. Evolution *55*, 1932–1942.

Flagel, L.E., and Wendel, J.F. (2010). Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. New Phytol. *186*, 184–193.

Flint, J., and Mackay, T.F.C. (2009). Genetic architecture of quantitative traits in mice, flies, and humans. Genome Res. *19*, 723–733.

FranceAgriMer (2013). Marché du blé dur. Monde, Europe, France (FranceAgriMer).

Frankel, O.H. (1995). The Conservation of Plant Biodiversity (Cambridge University Press).

Frankham, R., Ballou, J.D., Eldridge, M.D.B., Lacy, R.C., Ralls, K., Dudash, M.R., and Fenster, C.B. (2011). Predicción de la Depresión por Exogámica. Conserv. Biol. *25*, 465–475.

Freeling, M., Woodhouse, M.R., Subramaniam, S., Turco, G., Lisch, D., and Schnable, J.C. (2012). Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. Curr. Opin. Plant Biol. *15*, 131–139.

Gaeta, R.T., Pires, J.C., Iniguez-Luy, F., Leon, E., and Osborn, T.C. (2007). Genomic Changes in Resynthesized Brassica napus and Their Effect on Gene Expression and Phenotype. Plant Cell Online *19*, 3403–3417.

Gallais, A. (1990). Théorie de la sélection en amélioration des plantes (Masson).

Gan, X., Stegle, O., Behr, J., Steffen, J.G., Drewe, P., Hildebrand, K.L., Lyngsoe, R., Schultheiss, S.J., Osborne, E.J., Sreedharan, V.T., et al. (2011). Multiple reference genomes and transcriptomes for Arabidopsis thaliana. Nature *477*, 419–423.

García, P., Morris, M.I., Sáenz-de-Miera, L.E., Allard, R.W., Vega, M.P. de la, and Ladizinsky, G. (1991). Genetic diversity and adaptedness in tetraploid Avena barbata and its diploid ancestors Avena hirtula and Avena wiestii. Proc. Natl. Acad. Sci. *88*, 1207–1211.

Garsmeur, O., Schnable, J.C., Almeida, A., Jourda, C., D'Hont, A., and Freeling, M. (2014). Two Evolutionarily Distinct Classes of Paleopolyploidy. Mol. Biol. Evol. *31*, 448–454.

Gayral, P., Melo-Ferreira, J., Glémin, S., Bierne, N., Carneiro, M., Nabholz, B., Lourenco, J.M., Alves, P.C., Ballenghien, M., Faivre, N., et al. (2013). Reference-Free Population Genomics from Next-Generation Transcriptome Data and the Vertebrate–Invertebrate Gap. PLoS Genet *9*, e1003457.

Geigy, C.A., Heid, S., Steffen, F., Danielson, K., Jaggy, A., and Gaillard, C. (2007). Does a pleiotropic gene explain deafness and blue irises in white cats? Vet. J. *173*, 548–553.

Gianola, D., Hospital, F., and Verrier, E. (2013). Contribution of an additive locus to genetic variance when inheritance is multi-factorial with implications on interpretation of GWAS. TAG Theor. Appl. Genet. Theor. Angew. Genet. *126*, 1457–1472.

Gill, D.B.S., and Waines, J.G. (1978). Paternal regulation of seed development in wheat hybrids. Theor. Appl. Genet. *51*, 265–270.

Gjuvsland, A.B., Hayes, B.J., Omholt, S.W., and Carlborg, O. (2007). Statistical Epistasis Is a Generic Feature of Gene Regulatory Networks. Genetics *175*, 411–420.

Goldringer, I., and Bataillon, T. (2004). On the Distribution of Temporal Variations in Allele Frequency Consequences for the Estimation of Effective Population Size and the Detection of Loci Undergoing Selection. Genetics *168*, 563–568.

Goldringer, I., Brabant, P., and Gallais, A. (1997). Estimation of additive and epistatic genetic variances for agronomic traits in a population of doubled-haploid lines of wheat. Heredity *79*, 60–71.

Goldringer, I., Paillard, S., Enjalbert, J., David, J.L., and Brabant, P. (1998). Divergent evolution of wheat populations conducted under recurrent selection and dynamic management. Agronomie *18*, 413–425.

Goldringer, I., Enjalbert, J., David, J.L., Paillard, S., Pham, J.-L., and Brabant, P. (2001). Dynamic Management of Genetic Resources: a 13-year Experiment on Wheat. In Broadening the Genetic Base of Crop Production, (CABI), pp. 245–260.

González, J.R., Armengol, L., Solé, X., Guinó, E., Mercader, J.M., Estivill, X., and Moreno, V. (2007). SNPassoc: an R package to perform whole genome association studies. Bioinforma. Oxf. Engl. *23*, 644–645.

Goodnight, C.J. (1988). Epistasis and the Effect of Founder Events on the Additive Genetic Variance. Evolution *42*, 441.

Grover, C.E., Gallagher, J.P., Szadkowski, E.P., Yoo, M.J., Flagel, L.E., and Wendel, J.F. (2012). Homoeolog expression bias and expression level dominance in allopolyploids. New Phytol. *196*, 966–971.

Grundt, H.H., Kjølner, S., Borgen, L., Rieseberg, L.H., and Brochmann, C. (2006). High biological species diversity in the arctic flora. Proc. Natl. Acad. Sci. U. S. A. *103*, 972–975.

Haerty, W., and Singh, R.S. (2006). Gene Regulation Divergence Is a Major Contributor to the Evolution of Dobzhansky–Muller Incompatibilities between Species of Drosophila. Mol. Biol. Evol. *23*, 1707–1714.

Haile, J.K., Nachit, M.M., Hammer, K., Badebo, A., and Röder, M.S. (2012). QTL mapping of resistance to race Ug99 of Puccinia graminis f. sp. tritici in durum wheat (Triticum durum Desf.). Mol. Breed. *30*, 1479–1493.

Hammond, J.P., Mayes, S., Bowen, H.C., Graham, N.S., Hayden, R.M., Love, C.G., Spracklen, W.P., Wang, J., Welham, S.J., White, P.J., et al. (2011). Regulatory Hotspots Are Associated with Plant Gene Expression under Varying Soil Phosphorus Supply in Brassica rapa. Plant Physiol. *156*, 1230–1241.

Hamrick, J.L., and Godt, M.J.W. (1996). Effects of Life History Traits on Genetic Diversity in Plant Species. Philos. Trans. R. Soc. Lond. B. Biol. Sci. *351*, 1291–1298.

Hansen, T.F., and Wagner, G.P. (2001). Modeling Genetic Architecture: A Multilinear Theory of Gene Interaction. Theor. Popul. Biol. *59*, 61–86.

Hard, J.J., Bradshaw, W.E., and Holzapfel, C.M. (1992). Epistasis and the genetic divergence of photoperiodism between populations of the pitcher-plant mosquito, Wyeomyia smithii. Genetics *131*, 389–396.

Haudry, A., Cenci, A., Ravel, C., Bataillon, T., Brunel, D., Poncet, C., Hochu, I., Poirier, S., Santoni, S., Glémin, S., et al. (2007). Grinding up Wheat: A Massive Loss of Nucleotide Diversity Since Domestication. Mol. Biol. Evol. *24*, 1506–1517.

Hayman, B.I. (1957). Interaction, Heterosis and Diallel Crosses. Genetics *42*, 336–355.

He, Q., and Lin, D.-Y. (2011). A variable selection method for genome-wide association studies. Bioinformatics *27*, 1–8.

He, P., Friebe, B.R., Gill, B.S., and Zhou, J.-M. (2003). Allopolyploidy alters gene expression in the highly stable hexaploid wheat. Plant Mol. Biol. *52*, 401–414.

Heerwaarden, B. van, Willi, Y., Kristensen, T.N., and Hoffmann, A.A. (2008). Population Bottlenecks Increase Additive Genetic Variance But Do Not Break a Selection Limit in Rain Forest Drosophila. Genetics *179*, 2135–2146.

Helgason, A., Yngvadóttir, B., Hrafnkelsson, B., Gulcher, J., and Stefánsson, K. (2005). An Icelandic example of the impact of population structure on association studies. Nat. Genet. *37*, 90–95.

Hemani, G., Shakhbazov, K., Westra, H.-J., Esko, T., Henders, A.K., McRae, A.F., Yang, J., Gibson, G., Martin, N.G., Metspalu, A., et al. (2014). Detection and replication of epistasis influencing transcription in humans. Nature *508*, 249–253.

Henry, J.P., Pontis, C., David, J., and Gouyon, P.H. (1991). An Experiment on Dynamic Conservation of Genetic Resources With Metapopulations. In Species Conservation: A Population-Biological Approach, P.A. Seitz, and P.V. Loeschcke, eds. (Birkhäuser Basel), pp. 185–198.

Hermsen, J.G.T. (1963a). Hybrid necrosis as a problem for the wheat breeder. Euphytica *12*, 1–16.

Hermsen, J.G.T. (1963b). The genetic basis of hybrid necrosis in wheat. Genetica *33*, 245–287.

Heyne, E.G., Wiebe, G.A., and Painter, R.H. (1943). Complementary genes in wheat causing death of F1 plants. J. Hered. *34*, 243–245.

Higgs, P.G. (1998). Compensatory neutral mutations and the evolution of RNA. Genetica *102-103*, 91–101.

Hill, W.G., and Robertson, A. (1968). Linkage disequilibrium in finite populations. Theor. Appl. Genet. *38*, 226–231.

Hill, W.G., Goddard, M.E., and Visscher, P.M. (2008). Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits. PLoS Genet *4*, e1000008.

Hirschhorn, J.N., and Daly, M.J. (2005). Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. *6*, 95–108.

Hoerl, A.E., and Kennard, R.W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics *12*, 55–67.

Hoisington, D., Khairallah, M., Reeves, T., Ribaut, J.-M., Skovmand, B., Taba, S., and Warburton, M. (1999). Plant genetic resources: What can they contribute toward increased crop productivity? Proc. Natl. Acad. Sci. *96*, 5937–5943.

Holland, J. B. (2001). Epistasis and plant breeding. In Plant Breeding Reviews, pp. 27–92.

Holloway, B., Luck, S., Beatty, M., Rafalski, J.-A., and Li, B. (2011). Genome-wide expression quantitative trait loci (eQTL) analysis in maize. BMC Genomics *12*, 336.

Houchins, K., O'Dell, M., Flavell, R.B., and Gustafson, J.P. (1997). Cytosine methylation and nucleolar dominance in cereal hybrids. Mol. Gen. Genet. MGG *255*, 294–301.

Hu, Z.-L., Li, P., Zhou, M.-Q., Zhang, Z.-H., Wang, L.-X., Zhu, L.-H., and Zhu, Y.-G. (2004). Mapping of quantitative trait loci (QTLs) for rice protein and fat content using doubled haploid lines. Euphytica *135*, 47–54.

Huang, S., Sirikhachornkit, A., Su, X., Faris, J., Gill, B., Haselkorn, R., and Gornicki, P. (2002). Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the Triticum/Aegilops complex and the evolutionary history of polyploid wheat. Proc. Natl. Acad. Sci. *99*, 8133–8138.

Huang, W., Richards, S., Carbone, M.A., Zhu, D., Anholt, R.R.H., Ayroles, J.F., Duncan, L., Jordan, K.W., Lawrence, F., Magwire, M.M., et al. (2012a). Epistasis dominates the genetic architecture of Drosophila quantitative traits. Proc. Natl. Acad. Sci. *109*, 15553–15559.

Huang, Y., Zheng, J., and Przytycka, T.M. (2009). Discovery of regulatory mechanisms from gene expression variation by eQTL analysis. In Biological Data Mining, (CRC Press), pp. 205–228.

Huang, Y., Siwo, G., Wuchty, S., Ferdig, M.T., and Przytycka, T.M. (2012b). Symmetric Epistasis Estimation (SEE) and its application to dissecting interaction map of Plasmodium falciparum. Mol. Biosyst. *8*, 1544–1552.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., et al. (2002). The Ensembl genome database project. Nucleic Acids Res. *30*, 38–41.

Iehisa, J.C.M., Ohno, R., Kimura, T., Enoki, H., Nishimura, S., Okamoto, Y., Nasuda, S., and Takumi, S. (2014). A High-Density Genetic Map with Array-Based Markers Facilitates Structural and Quantitative Trait Locus Analyses of the Common Wheat Genome. DNA Res. dsu020.

Imielinski, M., and Belta, C. (2008). Exploiting the pathway structure of metabolism to reveal high-order epistasis. BMC Syst. Biol. *2*, 40.

International Grains Council (2012). World Grains Statistics (London, UK).

Jain, K. (2010). Time to fixation in the presence of recombination. Theor. Popul. Biol. *77*, 23–31.

Jain, S.K., and Suneson, C.A. (1963). Male Sterility for Increased Outbreeding in Populations of Barley. Nature *199*, 407–408.

Jain, S.K., and Suneson, C.A. (1966). Increased Recombination and Selection in Barley Populations Carrying a Male Sterility Factor. I. Quantitative Variability. Genetics *54*, 1215–1224.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., and Barabási, A.-L. (2000). The large-scale organization of metabolic networks. Nature *407*, 651–654.

Jeong, H., Mason, S.P., Barabási, A.-L., and Oltvai, Z.N. (2001). Lethality and centrality in protein networks. Nature *411*, 41–42.

Jiao, Y., Wickett, N.J., Ayyampalayam, S., Chanderbali, A.S., Landherr, L., Ralph, P.E., Tomsho, L.P., Hu, Y., Liang, H., Soltis, P.S., et al. (2011). Ancestral polyploidy in seed plants and angiosperms. Nature *473*, 97–100.

Jorjani, H., Klei, L., and Emanuelson, U. (2003). A Simple Method for Weighted Bending of Genetic (Co)variance Matrices. J. Dairy Sci. *86*, 677–679.

Kao, C.-H., and Zeng, Z.-B. (2002). Modeling Epistasis of Quantitative Trait Loci Using Cockerham's Model. Genetics *160*, 1243–1261.

Kashkush, K., Feldman, M., and Levy, A.A. (2002). Gene Loss, Silencing and Activation in a Newly Synthesized Wheat Allotetraploid. Genetics *160*, 1651–1659.

Kempthorne, O. (1954). The Correlation between Relatives in a Random Mating Population. Proc. R. Soc. Lond. Ser. B - Biol. Sci. *143*, 103–113.

Keurentjes, J.J.B., Fu, J., Terpstra, I.R., Garcia, J.M., Ackerveken, G. van den, Snoek, L.B., Peeters, A.J.M., Vreugdenhil, D., Koornneef, M., and Jansen, R.C. (2007). Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. Proc. Natl. Acad. Sci. *104*, 1708–1713.

Kihara, H. (1944). Discovery of the DD-analyzer, one of the ancestors of vulgare wheats. Agric Hort *19*, 889–890.

Kilian, B., Özkan, H., Deusch, O., Effgen, S., Brandolini, A., Kohl, J., Martin, W., and Salamini, F. (2007). Independent Wheat B and G Genome Origins in Outcrossing Aegilops Progenitor Haplotypes. Mol. Biol. Evol. *24*, 217–227.

Kilian, B., Özkan, H., Pozzi, C., and Salamini, F. (2009). Domestication of the Triticeae in the Fertile Crescent. In Genetics and Genomics of the Triticeae, G.J. Muehlbauer, and C. Feuillet, eds. (Springer US), pp. 81–119.

Kliebenstein, D. (2009). Quantitative Genomics: Analyzing Intraspecific Variation Using Global Gene Expression Polymorphisms or eQTLs. Annu. Rev. Plant Biol. *60*, 93–114.

Kolster, P., Eeuwijk, F.A. van, and Gelder, W.M.J. van (1991). Additive and epistatic effects of allelic variation at the high molecular weight glutenin subunit loci in determining the bread-making quality of breeding lines of wheat. Euphytica *55*, 277–285.

Kotseruba, V., Gernand, D., Meister, A., and Houben, A. (2003). Uniparental loss of ribosomal DNA in the allotetraploid grass Zingeria trichopoda (2n = 8). Genome *46*, 156–163.

Kovarik, A., Matyasek, R., Lim, K.Y., Skalická, K., Koukalová, B., Knapp, S., Chase, M., and Leitch, A.R. (2004). Concerted evolution of 18–5.8–26S rDNA repeats in Nicotiana allotetraploids. Biol. J. Linn. Soc. *82*, 615–625.

Kovarik, A., Pires, J.C., Leitch, A.R., Lim, K.Y., Sherwood, A.M., Matyasek, R., Rocca, J., Soltis, D.E., and Soltis, P.S. (2005). Rapid Concerted Evolution of Nuclear Ribosomal DNA in Two Tragopogon Allopolyploids of Recent and Recurrent Origin. Genetics *169*, 931–944.

Kubo, T., Yamagata, Y., Eguchi, M., and Yoshimura, A. (2008). A novel epistatic interaction at two loci causing hybrid male sterility in an inter-subspecific cross of rice (Oryza sativa L.). Genes Genet. Syst. *83*, 443–453.

Kubo, T., Yoshimura, A., and Kurata, N. (2011). Hybrid Male Sterility in Rice Is Due to Epistatic Interactions with a Pollen Killer Locus. Genetics *189*, 1083–1092.

Lair, K.P., Bradshaw, W.E., and Holzapfel, C.M. (1997). Evolutionary Divergence of the Genetic Architecture Underlying Photoperiodism in the Pitcher-Plant Mosquito, Wyeomyia smithii. Genetics *147*, 1873–1883.

Lalić, J., and Elena, S.F. (2012). Magnitude and sign epistasis among deleterious mutations in a positive-sense plant RNA virus. Heredity *109*, 71–77.

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature *467*, 832–838.

Lavigne, C., Reboud, X., Lefranc, M., Porcher, E., Roux, F., Olivieri, I., Godelle, B., Lavigne, C., Reboud, X., Lefranc, M., et al. (2001). Evolution of genetic diversity in metapopulations: Arabidopsis thaliana as an experimental model, Evolution of genetic diversity in metapopulations: Arabidopsis thaliana as an experimental model. Genet. Sel. Evol. Genet. Sel. Evol. *33, 33*, S399–S423, S399–S423.

Legarra, A., and Fernando, R.L. (2009). Linear models for joint association and linkage QTL mapping. Genet. Sel. Evol. *41*, 43.

Levin, D.A. (1971). The Origin of Reproductive Isolating Mechanisms in Flowering Plants. Taxon *20*, 91.

Li, B., and Leal, S.M. (2008). Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. Am. J. Hum. Genet. *83*, 311–321.

Li, J., Tang, R., Biernacka, J.M., and Andrade, M. de (2009a). Identification of gene-gene interaction using principal components. BMC Proc. *3*, S78.

Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009b). Genotype Imputation. Annu. Rev. Genomics Hum. Genet. *10*, 387–406.

Li, Y., Willer, C.J., Ding, J., Scheet, P., and Abecasis, G.R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol. *34*, 816–834.

Li, Z., Pinson, S.R., Park, W.D., Paterson, A.H., and Stansel, J.W. (1997a). Epistasis for three grain yield components in rice (Oryza sativa L.). Genetics *145*, 453–465.

Li, Z., Pinson, S.R.M., Paterson, A.H., Park, W.D., and Stansel, J.W. (1997b). Genetics of Hybrid Sterility and Hybrid Breakdown in an Intersubspecific Rice (Oryza sativa L.) Population. Genetics *145*, 1139–1148.

Liu, A., and Burke, J.M. (2006). Patterns of Nucleotide Diversity in Wild and Cultivated Sunflower. Genetics *173*, 321–330.

Lopez-Fanjul, C., Fernandez, A., and Toro, M.A. (1999). The role of epistasis in the increase in the additive genetic variance after population bottlenecks. Genet. Res. *73*, 45–59.

Luikart, G., Allendorf, F.W., Cornuet, J.-M., and Sherwin, W.B. (1998). Distortion of allele frequency distributions provides a test for recent population bottlenecks. J. Hered. *89*, 238–247.

Lukens, L.N., Pires, J.C., Leon, E., Vogelzang, R., Oslach, L., and Osborn, T. (2006). Patterns of Sequence Loss and Cytosine Methylation within a Population of Newly Resynthesized Brassica napus Allopolyploids. Plant Physiol. *140*, 336–348.

Luo, L., Boerwinkle, E., and Xiong, M. (2011). Association studies for next-generation sequencing. Genome Res. *21*, 1099–1108.

Lush, J.L. (1937). Animal Breeding Plans (Ames, Iowa: Iowa State Press).

Lynch, M. (1991). The Genetic Interpretation of Inbreeding Depression and Outbreeding Depression. Evolution *45*, 622.

Ma, X.-F., and Gustafson, J.P. (2006). Timing and rate of genome variation in triticale following allopolyploidization. Genome *49*, 950–958.

Maccaferri, M., Sanguineti, M.C., Corneti, S., Ortega, J.L.A., Salem, M.B., Bort, J., DeAmbrogio, E., Moral, L.F.G. del, Demontis, A., El-Ahmed, A., et al. (2008a). Quantitative Trait Loci for Grain Yield and Adaptation of Durum Wheat (Triticum durum Desf.) Across a Wide Range of Water Availability. Genetics *178*, 489–511.

Macgregor, S., Cornes, B.K., Martin, N.G., and Visscher, P.M. (2006). Bias, precision and heritability of self-reported and clinically measured height in Australian twins. Hum. Genet. *120*, 571–580.

Mackay, I., and Powell, W. (2007). Methods for linkage disequilibrium mapping in crops. Trends Plant Sci. *12*, 57–63.

Mac Key, J. (1970). Significance of mating systems for chromosomes and gametes in polyploids. Hereditas *66*, 165–176.

Madlung, A., Masuelli, R.W., Watson, B., Reynolds, S.H., Davison, J., and Comai, L. (2002). Remodeling of DNA Methylation and Phenotypic and Transcriptional Changes in Synthetic Arabidopsis Allotetraploids. Plant Physiol. *129*, 733–746.

Madlung, A., Tyagi, A.P., Watson, B., Jiang, H., Kagochi, T., Doerge, R.W., Martienssen, R., and Comai, L. (2005). Genomic changes in synthetic Arabidopsis polyploids. Plant J. *41*, 221–230.

Majewski, J., and Pastinen, T. (2011). The study of eQTL variations by RNA-seq: from SNPs to phenotypes. Trends Genet. *27*, 72–79.

Mäki-Tanila, A., and Hill, W.G. (2014). Influence of Gene Interaction on Complex Trait Variation with Multilocus Models. Genetics *198*, 355–367.

Mallet, J. (2010). Shift happens! Shifting balance and the evolution of diversity in warning colour and mimicry. Ecol. Entomol. *35*, 90–104.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.

Marais, G., and Charlesworth, B. (2003). Genome Evolution: Recombination Speeds Up Adaptive Evolution. Curr. Biol. *13*, R68–R70.

Martin, C.H., and Wainwright, P.C. (2013). Multiple Fitness Peaks on the Adaptive Landscape Drive Adaptive Radiation in the Wild. Science *339*, 208–211.

Martin, N.H., and Willis, J.H. (2007). Ecological Divergence Associated with Mating System Causes Nearly Complete Reproductive Isolation Between Sympatric Mimulus Species. Evolution *61*, 68–82.

Mayer, K.F.X., Rogers, J., Doležel, J., Pozniak, C., Eversole, K., Feuillet, C., Gill, B., Friebe, B., Lukaszewski, A.J., Sourdille, P., et al. (2014a). A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science *345*, 1251788.

Mayer, K.F.X., Rogers, J., Doležel, J., Pozniak, C., Eversole, K., Feuillet, C., Gill, B., Friebe, B., Lukaszewski, A.J., Sourdille, P., et al. (2014b). A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science *345*, 1251788.

McClintock, B. (1984). The significance of responses of the genome to challenge. Science *226*, 792–801.

Mcfadden, E.S., and Sears, E.R. (1946). The Origin of Triticum Spelta and Its Free-Threshing Hexaploid Relatives. J. Hered. *37*, 107–116.

McKinney, B.A., and Pajewski, N.M. (2012). Six Degrees of Epistasis: Statistical Network Models for GWAS. Front. Genet. *2*.

Meyer, K., and Kirkpatrick, M. (2010). Better Estimates of Genetic Covariance Matrices by “Bending” Using Penalized Maximum Likelihood. Genetics *185*, 1097–1110.

Michaelson, J.J., Loguercio, S., and Beyer, A. (2009). Detection and interpretation of expression quantitative trait loci (eQTL). Methods *48*, 265–276.

Miedaner, T., Würschum, T., Maurer, H.P., Korzun, V., Ebmeyer, E., and Reif, J.C. (2011). Association mapping for Fusarium head blight resistance in European soft winter wheat. Mol. Breed. *28*, 647–655.

Montoya, J., and Solé, R. (2002). Small World Patterns in Food Webs. J. Theor. Biol. *214*, 405–412.

Moore, J.H. (2007). Genome-wide analysis of epistasis using multifactor dimensionality reduction: Feature selection and construction. In Knowledge Discovery and Data Mining: Challenges and Realities, (Idea Group Inc (IGI)),.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods *5*, 621–628.

Moscou, M.J., Lauter, N., Steffenson, B., and Wise, R.P. (2011). Quantitative and Qualitative Stem Rust Resistance Factors in Barley Are Associated with Transcriptional Suppression of Defense Regulons. PLoS Genet *7*, e1002208.

Muller, H. (1939). Reversibility in evolution considered from the standpoint of genetics. Biol. Rev. *14*, 261–280.

Muller, H. (1940). Bearing of the Drosophila work on systematics. In The New Systematics, (Oxford), pp. 185–268.

Naciri-Graven, Y., and Goudet, J. (2003). The Additive Genetic Variance After Bottlenecks Is Affected by the Number of Loci Involved in Epistatic Interactions. Evolution *57*, 706–716.

Nahum, J.R., Godfrey-Smith, P., Harding, B.N., Marcus, J.H., Carlson-Stevermer, J., and Kerr, B. (2014). How the tortoise beats the hare: Slow and steady adaptation in structured populations suggests a rugged fitness landscape in bacteria. bioRxiv 005793.

Nei, M. (1978). Estimation of Average Heterozygosity and Genetic Distance from a Small Number of Individuals. Genetics *89*, 583–590.

Nei, M., and Tajima, F. (1981). Genetic Drift and Estimation of Effective Population Size. Genetics *98*, 625–640.

Nesbitt, M., and Samuel, D. (1998). Wheat domestication: archaeobotanical evidence. Science.

Newman, D.L., Abney, M., McPeek, M.S., Ober, C., and Cox, N.J. (2001). The Importance of Genealogy in Determining Genetic Associations with Complex Traits. Am. J. Hum. Genet. *69*, 1146–1148.

Nishikawa, K., Mori, T., Takami, N., and Furuta, Y. Mapping of progressive necrosis gene Ne1 and Ne2 of common wheat by the telocentric method. Jpn. J Breed *4:*, 277–281.

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

Nordborg, M., Borevitz, J.O., Bergelson, J., Berry, C.C., Chory, J., Hagenblad, J., Kreitman, M., Maloof, J.N., Noyes, T., Oefner, P.J., et al. (2002). The extent of linkage disequilibrium in Arabidopsis thaliana. Nat. Genet. *30*, 190–193.

Orr, H.A. (1995). The population genetics of speciation: the evolution of hybrid incompatibilities. Genetics *139*, 1805–1813.

Orr, H.A. (2005). The genetic theory of adaptation: a brief history. Nat. Rev. Genet. *6*, 119–127.

Orr, H.A., and Irving, S. (2001). Complex Epistasis and the Genetic Basis of Hybrid Sterility in the Drosophila pseudoobscura Bogota-USA Hybridization. Genetics *158*, 1089–1100.

Osborn, T.C., Chris Pires, J., Birchler, J.A., Auger, D.L., Jeffery Chen, Z., Lee, H.-S., Comai, L., Madlung, A., Doerge, R.W., Colot, V., et al. (2003). Understanding mechanisms of novel gene expression in polyploids. Trends Genet. *19*, 141–147.

Oshlack, A., and Wakefield, M.J. (2009). Transcript length bias in RNA-seq data confounds systems biology. Biol. Direct *4*, 1–10.

Otto, S.P. (2007). The Evolutionary Consequences of Polyploidy. Cell *131*, 452–462.

Otto, S.P., and Feldman, M.W. (1997). Deleterious Mutations, Variable Epistatic Interactions, and the Evolution of Recombination. Theor. Popul. Biol. *51*, 134–147.

Ozkan, H., Levy, A.A., and Feldman, M. (2001). Allopolyploidy-Induced Rapid Genome Evolution in the Wheat (Aegilops–Triticum) Group. Plant Cell Online *13*, 1735–1747.

Özkan, H., Brandolini, A., Schäfer-Pregl, R., and Salamini, F. (2002). AFLP Analysis of a Collection of Tetraploid Wheats Indicates the Origin of Emmer and Hard Wheat Domestication in Southeast Turkey. Mol. Biol. Evol. *19*, 1797–1801.

Paillard, S., Goldringer, I., Enjalbert, J., Doussinault, G., Vallavieille-Pope, C. de, and Brabant, P. (2000a). Evolution of resistance against powdery mildew in winter wheat populations conducted under dynamic management. I – Is specific seedling resistance selected? Theor. Appl. Genet. *101*, 449–456.

Paillard, S., Goldringer, I., Enjalbert, J., Trottet, M., David, J., Vallavieille-Pope, C. de, and Brabant, P. (2000b). Evolution of resistance against powdery mildew in winter wheat populations conducted under dynamic management. II. Adult plant resistance. Theor. Appl. Genet. *101*, 457–462.

Palmer, A.C., and Kishony, R. (2013). Understanding, predicting and manipulating the genotypic evolution of antibiotic resistance. Nat. Rev. Genet. *14*, 243–248.

Palopoli, M.F., and Wu, C.I. (1994). Genetics of hybrid male sterility between drosophila sibling species: a complex web of epistasis is revealed in interspecific studies. Genetics *138*, 329–341.

Park, S.-C., and Krug, J. (2011). Bistability in two-locus models with selection, mutation, and recombination. J. Math. Biol. *62*, 763–788.

Patil, R.M., Tamhankar, S.A., Oak, M.D., Raut, A.L., Honrao, B.K., Rao, V.S., and Misra, S.C. (2013). Mapping of QTL for agronomic traits and kernel characters in durum wheat (Triticum durum Desf.). Euphytica *190*, 117–129.

Paux, E. (2013). High-Throughput SNP Discovery and Genotyping in Hexaploid Wheat. (Plant and Animal Genome),.

Van de Peer, Y., Fawcett, J.A., Proost, S., Sterck, L., and Vandepoele, K. (2009). The flowering world: a tale of duplications. Trends Plant Sci. *14*, 680–688.

Petersen, G., Seberg, O., Yde, M., and Berthelsen, K. (2006). Phylogenetic relationships of Triticum and Aegilops and evidence for the origin of the A, B, and D genomes of common wheat (Triticum aestivum). Mol. Phylogenet. Evol. *39*, 70–82.

Pettersson, M., Besnier, F., Siegel, P.B., and Carlborg, Ö. (2011). Replication and Explorations of High-Order Epistasis Using a Large Advanced Intercross Line Pedigree. PLoS Genet *7*, e1002180.

Pfeifer, M., Kugler, K.G., Sandve, S.R., Zhan, B., Rudi, H., Hvidsten, T.R., Mayer, K.F.X., and Olsen, O.-A. (2014). Genome interplay in the grain transcriptome of hexaploid bread wheat. Science *345*, 1250091.

Phillips, S.L., and Wolfe, M.S. (2005). Evolutionary plant breeding for low input systems. J. Agric. Sci. *143*, 245–254.

Pires, J.C., Zhao, J., Schranz, M.E., Leon, E.J., Quijada, P.A., Lukens, L.N., and Osborn, T.C. (2004). Flowering time divergence and genomic rearrangements in resynthesized Brassica polyploids (Brassicaceae). Biol. J. Linn. Soc. *82*, 675–688.

Plahte, E., Mestl, T., and Omholt, S.W. (1995). Feedback loops, stability and multistationarity in dynamical systems. J. Biol. Syst. *03*, 409–413.

Poelwijk, F.J., Kiviet, D.J., Weinreich, D.M., and Tans, S.J. (2007). Empirical fitness landscapes reveal accessible evolutionary paths. Nature *445*, 383–386.

Poelwijk, F.J., Tănase-Nicola, S., Kiviet, D.J., and Tans, S.J. (2011). Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes. J. Theor. Biol. *272*, 141–144.

Pollak, E. (1987). On the Theory of Partially Inbreeding Finite Populations. I. Partial Selfing. Genetics *117*, 353–360.

Pollak, E., and Sabran, M. (1999). On the theory of partially inbreeding finite populations. VI. The survival probability of a two-locus allele combination when there is partial selfing. Math. Biosci. *159*, 97–112.

Poon, A., and Chao, L. (2005). The Rate of Compensatory Mutation in the DNA Bacteriophage ϕX174. Genetics *170*, 989–999.

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

Porcher, E., Giraud, T., and Lavigne, C. (2006). Genetic differentiation of neutral markers and quantitative traits in predominantly selfing metapopulations: confronting theory and experiments with Arabidopsis thaliana. Genet. Res. *87*, 1–12.

Potokina, E., Druka, A., Luo, Z., Wise, R., Waugh, R., and Kearsey, M. (2008). Gene expression quantitative trait locus analysis of 16 000 barley genes reveals a complex pattern of genome-wide transcriptional regulation. Plant J. *53*, 90–101.

Prabhu, S., and Pe'er, I. (2012). Ultrafast genome-wide scan for SNP–SNP interactions in common complex disease. Genome Res. *22*, 2230–2240.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet. *81*, 559–575.

Qi, B., Huang, W., Zhu, B., Zhong, X., Guo, J., Zhao, N., Xu, C., Zhang, H., Pang, J., Han, F., et al. (2012). Global transgenerational gene expression dynamics in two newly synthesized allohexaploid wheat (Triticum aestivum) lines. BMC Biol. *10*, 3.

Rafalski, J.A. (2010). Association genetics in crop improvement. Curr. Opin. Plant Biol. *13*, 174–180.

Ramsey, J., and Schemske, D.W. (2002). Neopolyploidy in Flowering Plants. Annu. Rev. Ecol. Syst. *33*, 589–639.

Reif, J.C., Gowda, M., Maurer, H.P., Longin, C.F.H., Korzun, V., Ebmeyer, E., Bothe, R., Pietsch, C., and Würschum, T. (2011). Association mapping for quality traits in soft winter wheat. Theor. Appl. Genet. *122*, 961–970.

Rhoné, B., Remoué, C., Galic, N., Goldringer, I., and Bonnin, I. (2008). Insight into the genetic bases of climatic adaptation in experimentally evolving wheat populations. Mol. Ecol. *17*, 930–943.

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

Rice, W.R. (2002). Experimental tests of the adaptive significance of sexual recombination. Nat. Rev. Genet. *3*, 241–251.

Roff, D.A., and Emerson, K. (2006). Epistasis and Dominance: Evidence for Differential Effects in Life-History Versus Morphological Traits. Evolution *60*, 1981–1990.

Roussel, V., Koenig, J., Beckert, M., and Balfourier, F. (2004). Molecular diversity in French bread wheat accessions related to temporal trends and breeding programmes. Theor. Appl. Genet. *108*, 920–930.

Roussel, V., Leisova, L., Exbrayat, F., Stehno, Z., and Balfourier, F. (2005). SSR allelic diversity changes in 480 European bread wheat varieties released from 1840 to 2000. Theor. Appl. Genet. *111*, 162–170.

Rousselle, Y., Thomas, M., Galic, N., Bonnin, I., and Goldringer, I. (2011). Inbreeding depression and low between-population heterosis in recently diverged experimental populations of a selfing species. Heredity *106*, 289–299.

Rousset, M., Carrillo, J.M., Qualset, C.O., and Kasarda, D.D. (1992). Use of recombinant inbred lines of wheat for study of associations of high-molecular-weight glutenin subunit alleles to quantitative traits. Theor. Appl. Genet. *83*, 403–412.

Rutschman, D.H. (1994). Dynamics of the Two-Locus Haploid Model. Theor. Popul. Biol. *45*, 167–176.

Salamini, F., Özkan, H., Brandolini, A., Schäfer-Pregl, R., and Martin, W. (2002). Genetics and geography of wild cereal domestication in the near east. Nat. Rev. Genet. *3*, 429–441.

Santra, M., Santra, D.K., Rao, V.S., Taware, S.P., and Tamhankar, S.A. (2005). Inheritance of β-carotene concentration in durum wheat (Triticum turgidum L. ssp. durum). Euphytica *144*, 215–221.

Sayar, R., Khemira, H., and Kharrat, M. (2007). Inheritance of deeper root length and grain yield in half-diallel durum wheat (Triticum durum) crosses. Ann. Appl. Biol. *151*, 213–220.

Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G., et al. (2003). Genetics of gene expression surveyed in maize, mouse and man. Nature *422*, 297–302.

Schadt, E.E., Lamb, J., Yang, X., Zhu, J., Edwards, S., GuhaThakurta, D., Sieberts, S.K., Monks, S., Reitman, M., Zhang, C., et al. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. Nat. Genet. *37*, 710–717.

Schenk, M.F., Szendro, I.G., Salverda, M.L.M., Krug, J., and Visser, J.A.G.M. de (2013). Patterns of Epistasis between Beneficial Mutations in an Antibiotic Resistance Gene. Mol. Biol. Evol. *30*, 1779–1787.

Schluter, D. (2000). The Ecology of Adaptive Radiation (Oxford University Press).

Schmidt-Küntzel, A., Nelson, G., David, V.A., Schäffer, A.A., Eizirik, E., Roelke, M.E., Kehler, J.S., Hannah, S.S., O'Brien, S.J., and Menotti-Raymond, M. (2009). A Domestic cat X Chromosome Linkage Map and the Sex-Linked orange Locus: Mapping of orange, Multiple Origins and Epistasis Over nonagouti. Genetics *181*, 1415–1425.

Schnable, J.C., Springer, N.M., and Freeling, M. (2011). Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc. Natl. Acad. Sci. *108*, 4069–4074.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., et al. (2009). The B73 Maize Genome: Complexity, Diversity, and Dynamics. Science *326*, 1112–1115.

Sears, E.R. (1976). Genetic Control of Chromosome Pairing in Wheat. Annu. Rev. Genet. *10*, 31–51.

Selander, R.K., and Kaufman, D.W. (1973). Self-Fertilization and Genetic Population Structure in a Colonizing Land Snail. Proc. Natl. Acad. Sci. U. S. A. *70*, 1186–1190.

Shaked, H., Kashkush, K., Ozkan, H., Feldman, M., and Levy, A.A. (2001). Sequence Elimination and Cytosine Methylation Are Rapid and Reproducible Responses of the Genome to Wide Hybridization and Allopolyploidy in Wheat. Plant Cell Online *13*, 1749–1759.

Sharma, S.N., and Sain, R.S. (2004). Genetics of grains per spike in durum wheat under normal and late planting conditions. Euphytica *139*, 1–7.

Sharma, S.N., Sharma, R.K., Bhatnagar, V.K., and Sain, R.S. (1998). Inheritance of spike area in three inter-varietal durum-wheat crosses under different environments. RACHIS ICARDA.

Sharma, S.N., Sain, R.S., and Sharma, R.K. (2003). Genetics of spike length in durum wheat. Euphytica *130*, 155–161.

Shi, C., Uzarowska, A., Ouzunova, M., Landbeck, M., Wenzel, G., and Lübberstedt, T. (2007). Identification of candidate genes associated with cell wall digestibility and eQTL (expression quantitative trait loci) analysis in a Flint × Flint maize recombinant inbred line population. BMC Genomics *8*, 22.

Shi, X., Ng, D.W.-K., Zhang, C., Comai, L., Ye, W., and Jeffrey Chen, Z. (2012). Cis- and trans-regulatory divergence between progenitor species determines gene-expression novelty in Arabidopsis allopolyploids. Nat. Commun. *3*, 950.

Shimomura, K., Low-Zeddies, S.S., King, D.P., Steeves, T.D.L., Whiteley, A., Kushla, J., Zemenides, P.D., Lin, A., Vitaterna, M.H., Churchill, G.A., et al. (2001). Genome-Wide Epistatic Interaction Analysis Reveals Complex Genetic Determinants of Circadian Behavior in Mice. Genome Res. *11*, 959–980.

Sikorska, K., Lesaffre, E., Groenen, P.F., and Eilers, P.H. (2013). GWAS on your notebook: fast semi-parallel linear and logistic regression for genome-wide association studies. BMC Bioinformatics *14*, 166.

Silventoinen, K., Sammalisto, S., Perola, M., Boomsma, D.I., Cornes, B.K., Davis, C., Dunkel, L., de Lange, M., Harris, J.R., Hjelmborg, J.V.2., et al. (2003). Heritability of Adult Body Height: A Comparative Study of Twin Cohorts in Eight Countries. Twin Res. Hum. Genet. *6*, 399–408.

Simmonds, N.W. (1993). Introgression and Incorporation. Strategies for the Use of Crop Genetic Resources. Biol. Rev. *68*, 539–562.

Simpson, G.G. (1984). Tempo and Mode in Evolution (New York: Columbia University Press).

De Smet, R., and Van de Peer, Y. (2012). Redundancy and rewiring of genetic networks following genome-wide duplication events. Curr. Opin. Plant Biol. *15*, 168–176.

Soller, M., Brody, T., and Genizi, A. (1976). On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. Theor. Appl. Genet. *47*, 35–39.

Soltis, D.E., Bell, C.D., Kim, S., and Soltis, P.S. (2008). Origin and Early Evolution of Angiosperms. Ann. N. Y. Acad. Sci. *1133*, 3–25.

Somers, D.J., Fedak, G., Clarke, J., and Cao, W. (2006). Mapping of FHB resistance QTLs in tetraploid wheat. Genome *49*, 1586–1593.

Song, K., Lu, P., Tang, K., and Osborn, T.C. (1995). Rapid genome change in synthetic polyploids of Brassica and its implications for polyploid evolution. Proc. Natl. Acad. Sci. *92*, 7719–7723.

Song, L., Guo, W., and Zhang, T. (2009). Interaction of novel Dobzhansky–Muller type genes for the induction of hybrid lethality between Gossypium hirsutum and G. barbadense cv. Coastland R4-4. Theor. Appl. Genet. *119*, 33–41.

Stephan, W. (1996). The Rate of Compensatory Evolution. Genetics *144*, 419–426.

Stich, B., Yu, J., Melchinger, A.E., Piepho, H.-P., Utz, H.F., Maurer, H.P., and Buckler, E.S. (2007). Power to Detect Higher-Order Epistatic Interactions in a Metabolic Pathway Using a New Mapping Strategy. Genetics *176*, 563–570.

Storey, J.D., Akey, J.M., and Kruglyak, L. (2005). Multiple Locus Linkage Analysis of Genomewide Expression in Yeast. PLoS Biol *3*, e267.

Stupar, R.M., and Springer, N.M. (2006). Cis-transcriptional Variation in Maize Inbred Lines B73 and Mo17 Leads to Additive Expression Patterns in the F1 Hybrid. Genetics *173*, 2199–2210.

Sun, W., and Hu, Y. (2013). eQTL Mapping Using RNA-seq Data. Stat. Biosci. *5*.

Suneson, C.A. (1956). An Evolutionary Plant Breeding Method. Agron. J. *48*, 188.

Suzuki, R., and Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinforma. Oxf. Engl. *22*, 1540–1542.

Svensson, E., and Calsbeek, R. (2012). The Adaptive Landscape in Evolutionary Biology (Oxford University Press).

Swanson-Wagner, R.A., DeCook, R., Jia, Y., Bancroft, T., Ji, T., Zhao, X., Nettleton, D., and Schnable, P.S. (2009). Paternal Dominance of Trans-eQTL Influences Gene Expression Patterns in Maize Hybrids. Science *326*, 1118–1120.

Szűcs, P., Skinner, J.S., Karsai, I., Cuesta-Marcos, A., Haggard, K.G., Corey, A.E., Chen, T.H.H., and Hayes, P.M. (2007). Validation of the VRN-H2/VRN-H1 epistatic model in barley reveals that intron length variation in VRN-H1 may account for a continuum of vernalization sensitivity. Mol. Genet. Genomics *277*, 249–261.

Tai, T.H., and Tanksley, S.D. (1990). A rapid and inexpensive method for isolation of total DNA from dehydrated plant tissue. Plant Mol. Biol. Report. *8*, 297–303.

Tang, H., Woodhouse, M.R., Cheng, F., Schnable, J.C., Pedersen, B.S., Conant, G., Wang, X., Freeling, M., and Pires, J.C. (2012). Altered patterns of fractionation and exon deletions in Brassica rapa support a two-step model of paleohexaploidy. Genetics *190*, 1563–1574.

Taylor, S.J., Arnold, M., and Martin, N.H. (2009). The Genetic Architecture of Reproductive Isolation in Louisiana Irises: Hybrid Fitness in Nature. Evolution *63*, 2581–2594.

Templeton, A.R. (1986). Coadaptation and outbreeding depression.

Teulat, B., This, D., Khairallah, M., Borries, C., Ragot, C., Sourdille, P., Leroy, P., Monneveux, P., and Charrier, A. (1998). Several QTLs involved in osmotic-adjustment trait variation in barley (Hordeum vulgare L.). Theor. Appl. Genet. *96*, 688–698.

Therneau, T. (2012). coxme: Mixed Effects Cox Models.

Thomas, R., and Kaufman, M. (2001). Multistationarity, the basis of cell differentiation and memory. II. Logical analysis of regulatory networks in terms of feedback circuits. Chaos Interdiscip. J. Nonlinear Sci. *11*, 180–195.

Thomas, R., and Richelle, J. (1988). Positive feedback loops and multistationarity. Discrete Appl. Math. *19*, 381–396.

Thomas, B.C., Pedersen, B., and Freeling, M. (2006). Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. Genome Res. *16*, 934–946.

Thuillet, A.-C., Bataillon, T., Poirier, S., Santoni, S., and David, J.L. (2005). Estimation of Long-Term Effective Population Sizes Through the History of Durum Wheat Using Microsatellite Data. Genetics *169*, 1589–1599.

Tranquilli, G., and Dubcovsky, J. (2000). Epistatic interaction between vernalization genes Vrn-Am1 and Vrn-Am2 in diploid wheat. J. Hered. *91*, 304–306.

Tsegaye, S. (1996). Estimation of outcrossing rate in landraces of tetraploid wheat (Triticum turgidum L.). Plant Breed. *115*, 195–197.

Tsunewaki, K. (1960). Monosomic and conventional analyses in common wheat. III. Lethality. Idengaku Zasshi Jap J Genet *35*, 71–75.

Tyler, A.L., Asselbergs, F.W., Williams, S.M., and Moore, J.H. (2009). Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. BioEssays *31*, 220–227.

Udall, J.A., and Wendel, J.F. (2006). Polyploidy and Crop Improvement. Crop Sci. *46*, S – 3.

De Visser, J.A.G.M., Park, S., and Krug, J. (2009). Exploring the Effect of Sex on Empirical Fitness Landscapes. Am. Nat. *174*, S15–S30.

Volis, S., Shulgina, I., Zaretsky, M., and Koren, O. (2011). Epistasis in natural populations of a predominantly selfing plant. Heredity *106*, 300–309.

Wade, M.J., and Goodnight, C.J. (1991). Wright's shifting balance theory: an experimental study. Science *253*, 1015–1018.

Wade, M.J., Winther, R.G., Agrawal, A.F., and Goodnight, C.J. (2001). Alternative definitions of epistasis: dependence and interaction. Trends Ecol. Evol. *16*, 498–504.

Wagner, G.P., Kin, K., and Lynch, V.J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci. *131*, 281–285.

Wang, J., and Caballero, A. (1999). Developments in predicting the effective size of subdivided populations. Heredity *82*, 212–226.

Wang, J., Tian, L., Madlung, A., Lee, H.-S., Chen, M., Lee, J.J., Watson, B., Kagochi, T., Comai, L., and Chen, Z.J. (2004). Stochastic and Epigenetic Changes of Gene Expression in Arabidopsis Polyploids. Genetics *167*, 1961–1973.

Wang, J., Tian, L., Lee, H.-S., Wei, N.E., Jiang, H., Watson, B., Madlung, A., Osborn, T.C., Doerge, R.W., Comai, L., et al. (2006). Genomewide Nonadditive Gene Regulation in Arabidopsis Allotetraploids. Genetics *172*, 507–517.

Wang, J., Yu, H., Xie, W., Xing, Y., Yu, S., Xu, C., Li, X., Xiao, J., and Zhang, Q. (2010). A global analysis of QTLs for expression variations in rice shoots at the early seedling stage. Plant J. *63*, 1063–1074.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. *10*, 57–63.

Waples, R.S. (1989). A generalized approach for estimating effective population size from temporal changes in allele frequency. Genetics *121*, 379–391.

Ward, J.H. (1963). Hierarchical Grouping to Optimize an Objective Function. J. Am. Stat. Assoc. *58*, 236–244.

Weinreich, D.M., and Chao, L. (2005). Rapid Evolutionary Escape by Large Populations from Local Fitness Peaks Is Likely in Nature. Evolution *59*, 1175–1182.

Weinreich, D.M., Watson, R.A., and Chao, L. (2005). Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. Evol. Int. J. Org. Evol. *59*, 1165–1174.

Weinreich, D.M., Delaney, N.F., DePristo, M.A., and Hartl, D.L. (2006). Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins. Science *312*, 111–114.

Weissman, D.B., Desai, M.M., Fisher, D.S., and Feldman, M.W. (2009). The rate at which asexual populations cross fitness valleys. Theor. Popul. Biol. *75*, 286–300.

Weiss-Schneeweiss, H., Schneeweiss, G.M., Stuessy, T.F., Mabuchi, T., Park, J.-M., Jang, C.-G., and Sun, B.-Y. (2007). Chromosomal stasis in diploids contrasts with genome restructuring in auto- and allopolyploid taxa of Hepatica (Ranunculaceae). New Phytol. *174*, 669–682.

Wendel, J.F., Schnabel, A., and Seelanan, T. (1995). Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (Gossypium). Proc. Natl. Acad. Sci. *92*, 280–284.

Wendt, T., Canela, M.B.F., Klein, D.E., and Rios, R.I. (2002). Selfing facilitates reproductive isolation among three sympatric species of Pitcairnia (Bromeliaceae). Plant Syst. Evol. *232*, 201–212.

West, M.A.L., Kim, K., Kliebenstein, D.J., Leeuwen, H. van, Michelmore, R.W., Doerge, R.W., and Clair, D.A.S. (2007). Global eQTL Mapping Reveals the Complex Genetic Architecture of Transcript-Level Variation in Arabidopsis. Genetics *175*, 1441–1450.

Whitlock, M.C., Phillips, P.C., and Wade, M.J. (1993). Gene Interaction Affects the Additive Genetic Variance in Subdivided Populations with Migration and Extinction. Evolution *47*, 1758.

Wittkopp, P.J., Haerum, B.K., and Clark, A.G. (2004). Evolutionary changes in cis and trans gene regulation. Nature *430*, 85–88.

Wittkopp, P.J., Haerum, B.K., and Clark, A.G. (2008). Regulatory changes underlying expression differences within and between Drosophila species. Nat. Genet. *40*, 346–350.

Woodhouse, M.R., Schnable, J.C., Pedersen, B.S., Lyons, E., Lisch, D., Subramaniam, S., and Freeling, M. (2010). Following Tetraploidy in Maize, a Short Deletion Mechanism Removed Genes Preferentially from One of the Two Homeologs. PLoS Biol *8*, e1000409.

Woodhouse, M.R., Cheng, F., Pires, J.C., Lisch, D., Freeling, M., and Wang, X. (2014). Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. Proc. Natl. Acad. Sci. *111*, 5283–5288.

Wooten, D.R., Livingston, D.P., Holland, J.B., Marshall, D.S., and Murphy, J.P. (2008). Quantitative Trait Loci and Epistasis for Crown Freezing Tolerance in the “Kanota” × “Ogle” Hexaploid Oat Mapping Population. Crop Sci. *48*, 149.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding, and selection in evolution.

Wright, S. (1950). Genetical structure of populations. Nature *166*, 247–249.

Wright, S.I., Bi, I.V., Schroeder, S.G., Yamasaki, M., Doebley, J.F., McMullen, M.D., and Gaut, B.S. (2005). The Effects of Artificial Selection on the Maize Genome. Science *308*, 1310–1314.

Wright, S.I., Ness, R.W., Foxe, J.P., and Barrett, S.C.H. (2008). Genomic Consequences of Outcrossing and Selfing in Plants. Int. J. Plant Sci. *169*, 105–118.

Wu, C.-I., and Beckenbach, A.T. (1983). Evidence for Extensive Genetic Differentiation Between the Sex-Ratio and the Standard Arrangement of Drosophila Pseudoobscura and D. Persimilis and Identification of Hybrid Sterility Factors. Genetics *105*, 71–86.

Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. Am. J. Hum. Genet. *89*, 82–93.

Wuchty, S. (2002). Interaction and domain networks of yeast. Proteomics *2*, 1715–1723.

Xu, S., and Jia, Z. (2007). Genomewide Analysis of Epistatic Effects for Quantitative Traits in Barley. Genetics *175*, 1955–1963.

Xu, Y., Zhong, L., Wu, X., Fang, X., and Wang, J. (2009). Rapid alterations of gene expression and cytosine methylation in newly synthesized Brassica napus allopolyploids. Planta *229*, 471–483.

Yang, R.-C. (2004). Epistasis of Quantitative Trait Loci Under Different Gene Action Models. Genetics *167*, 1493–1505.

Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. *38*, 203–208.

Yu, S.B., Li, J.X., Xu, C.G., Tan, Y.F., Gao, Y.J., Li, X.H., Zhang, Q., and Maroof, M.A.S. (1997). Importance of epistasis as the genetic basis of heterosis in an elite rice hybrid. Proc. Natl. Acad. Sci. *94*, 9226–9231.

Zeng, Z.-B., Wang, T., and Zou, W. (2005). Modeling Quantitative Trait Loci and Interpretation of Models. Genetics *169*, 1711–1725.

Zhang, F., Boerwinkle, E., and Xiong, M. (2014). Epistasis analysis for quantitative traits by functional regression model. Genome Res.

Zhang, K., Tian, J., Zhao, L., and Wang, S. (2008). Mapping QTLs with epistatic effects and QTL × environment interactions for plant height using a doubled haploid population in cultivated wheat. J. Genet. Genomics *35*, 119–127.

Zhang, X., Cal, A.J., and Borevitz, J.O. (2011). Genetic architecture of regulatory variation in Arabidopsis thaliana. Genome Res. *21*, 725–733.

Zhu, J., Wiener, M.C., Zhang, C., Fridman, A., Minch, E., Lum, P.Y., Sachs, J.R., and Schadt, E.E. (2007). Increasing the Power to Detect Causal Associations by Combining Genotypic and Expression Data in Segregating Populations. PLoS Comput Biol *3*, e69.

Zou, Y., Su, Z., Yang, J., Zeng, Y., and Gu, X. (2009). Uncovering genetic regulatory network divergence between duplicate genes using yeast eQTL landscape. J. Exp. Zoolog. B Mol. Dev. Evol. *312B*, 722–733.

Zuk, O., Hechter, E., Sunyaev, S.R., and Lander, E.S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. Proc. Natl. Acad. Sci. *109*, 1193–1198.

Zuk, O., Schaffner, S.F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M.J., Neale, B.M., Sunyaev, S.R., and Lander, E.S. (2014). Searching for missing heritability: Designing rare variant association studies. Proc. Natl. Acad. Sci. *111*, E455–E464.

PARTIE VII

ANNEXES

VII. ANNEXES

VII.1 Calculs : types d'épistasie et effets génétiques

Nous nous situons dans un modèle bilocus haploïde, inspiré du modèle UWR (Cheverud and Routman, 1995). Soient le locus A et le locus B. Quatre génotypes sont possible : AB, Ab, aB et ab. La valeur génotypique du génotype ij est égale à :

$$G_{ij} = \mu + x_i \cdot a_A + x_j \cdot a_B + x_j \cdot x_i \cdot aa$$

Avec $x_i = \begin{cases} 1/2 & \text{si } i = A \\ -1/2 & \text{si } i = a \end{cases}$ et $x_j = \begin{cases} 1/2 & \text{si } j = B \\ -1/2 & \text{si } j = b \end{cases}$

Nous posons que la valeur génotypiques du génotype AB est toujours la plus forte. Par ailleurs, les effets additifs sont également posés comme toujours positifs, donc $a_A \geq 0$ et $a_B \geq 0$. Par conséquent, dans tous les cas :

$$G_{AB} \geq G_{Ab}$$
$$\leftrightarrow 2a_B \geq -aa \quad \textbf{(i)}$$

De la même façon :

$$G_{AB} \geq G_{aB}$$
$$\leftrightarrow 2a_A \geq -aa \quad \textbf{(ii)}$$

Quand l'épistasie est positive, les deux inégalités **(i)** et **(ii)** sont toujours vraies. En revanche si $aa < 0$, alors aa doit toujours être supérieur à $-2a_A$ et à $-2a_B$: $aa \geq -2a_A, -2a_B$.

Si on pose que : $G_{Ab} \geq G_{ab}$ alors, $2a_A \geq aa$. De la même façon si $G_{aB} \geq G_{ab}$ alors, $2a_B \geq aa$

En résumé :

Inégalité sur les valeurs génotypiques	Cas	Inégalité sur les effets génétiques fonctionnels
$G_{AB} \geq G_{Ab}$	tous	$aa \geq -2a_A$
$G_{AB} \geq G_{aB}$	tous	$aa \geq -2a_B$
$G_{Ab} \geq G_{ab}$	- Additivité - Epistasie de magnitude	$2a_A \geq aa$
$G_{aB} \geq G_{ab}$	- Additivité - Epistasie de magnitude	$2a_B \geq aa$
$G_{Ab} \leq G_{ab}$	- Epistasie de signe simple sur le locus A (changement de signe de l'allèle A) - Epistasie de signe réciproque	$2a_A \leq aa$
$G_{aB} \leq G_{ab}$	- Epistasie de signe simple sur le locus B (changement de signe de l'allèle B) - Epistasie de signe réciproque	$2a_B \leq aa$

En connaissant l'ordre des valeurs génotypiques dans chacun des cas d'épistasie, il est possible d'en déduire les inégalités sur les effets génétiques :

Type d'épistasie	Conditions	Ordre des valeurs génotypiques
Pas d'épistasie	$aa = 0$	$G_{AB} \geq G_{Ab}, G_{aB} \geq G_{ab}$
Epistasis de magnitude, positive	$aa > 0$ $aa \leq 2a_A, 2a_B$	$G_{AB} \geq G_{Ab}, G_{aB} \geq G_{ab}$
Epistasis de magnitude, négative	$aa < 0$ $-aa \leq 2a_A, 2a_B$	$G_{AB} \geq G_{Ab}, G_{aB} \geq G_{ab}$
Epistasie de signe simple	$0 < 2a_A < aa \leq 2a_B$ ou $0 < 2a_B < aa \leq 2a_A$	$G_{AB} > G_{aB} > G_{ab} > G_{Ab}$ ou $G_{AB} > G_{Ab} > G_{aB} > G_{ab}$
Epistasie de signe réciproque	$0 < 2a_A, 2a_B < aa$	$G_{AB} \geq G_{ab} > G_{Ab}, G_{aB}$

VII.2 Broadening the genetic basis of durum wheat

Jacques L. David[1], Muriel Tavaud[1], Pierre Roumet[1], Marie-Hélène Muller[1], Sylvain Santoni[1], Gautier Sarah[1], Yan Holtz[1], Vincent Ranwez[1], Morgane Ardisson[1], Gérard Poux[1], Constance Vagne[1]

Abstract

A large reservoir of genetic diversity is available in the *Triticum turgidum* sub-species. This diversity can be used in breeding programs but the method by which it could be introgressed in the elite durum pool is challenging once the objective is to cover the whole genome. *T. turgidum* accessions are usually classified in sub-species or taxa while evidence are accumulating that a more complex genetic structure can be revealed using no *a priori* approaches. Wild and cultivated emmers are the most diverse compartment and some cultivated emmers are strongly differentiated from naked wheats while other emmers are closer to the current durum. Recent and successive bottlenecks (in the XXth century) explain a large part of the structuration of the modern durum pool and the loss of molecular diversity. The use of *T. turgidum* germplasm in classical genealogical breeding may lead to some disappointments if a quick return is researched while some advances in breeding elite lines can be achieved with more recombination and selection. Evolutionary Pre-Breeding appears as a valuable alternative. Building and managing composite cross populations for a long period yields in innovative genitors, with an enriched allelic diversity and reduced long range linkage disequilibrium.

Keywords

Triticum turgidum, durum, pre breeding, structuration, NGS, linkage disequilibrium

INTRODUCTION

Durum wheat belongs to the family of the AABB tetraploid wheats (*Triticum turgidum* spp.), where others cultivated forms such as emmer, polonicum and many others taxa co-exists (Bozzini, 1988, Nesbitt and Samuel, 1996). A lot of work has been carried out to describe and explain this current diversity (Özkan et al., 2002, 2011, Kilian et al 2009, Luo et al., 2007, Thuillet et al., 2005, Haudry et al 2007). All these domestic forms derive from a common wild ancestor, the wild emmer, *Triticum dicoccoides* (Kilian et al., 2009). Among the cultivated forms, emmer, *T. dicoccum* appears as the current remnant form of the first domesticated taxa developed by the first farmers 10 000 years ago in the Middle East (Zaharieva et al., 2010). It presents hulled grains and has lost the seed dispersal habit of its wild ancestor. Grains have also increased in size and their shape have dramatically evolved from a triangular to a more round section with a reduction in the grain length (Zaharieva et al., 2010). Dicoccum spread from its center of domestication and spread westward and eastward giving rise to number of other different hulled taxa. A more recent transition (7000 years ago) has lead to the fixation in new taxa of an innovating trait, the free threshing of the kernels (Feldman and Kislev, 2007). The free threshing wheat group, sometimes improperly designed as naked wheats is large and morphologically diversified and the current durum wheat is its most representative and farmed member (Bozzini 1988, Kilian et al., 2009).

Evolutionary factors such as spread and selection for the adaptation to new environments and varied farming practices, rise and fixation of new mutations, gene flows, seeds exchanges among and between farmers communities and lately modern breeding for adapting cultivars to the intensification and specialization of the crop production had a profound impact on cultivated plant growth, development and physiology (Alonso-Blanco et al., 2009) agronomic performance. Those evolutionary factors also deeply affect the level of genetic diversity between and among each group (Spilane and

[1] UMR AGAP, Montpellier Supagro, 2 place Viala, F-34060 Montpellier, France

Gepts, 2001 for a review, Buckler and Thornsberry 2002, Luo et al., 2007, Thuillet et al., 2005, Haudry et al., 2007, Laidò et al., 2013 for *T. turgidum*).

The complex genetic landscape of old and modern *T. turgidum* wheats resulting of these 120 centuries of evolution has been partly elucidated. The major split appears between hulled wheats that constitutes diversified genetic pool and free threshing wheats, that tare comparatively very poor in allelic diversity (Thuillet et al., 2005, Haudry et al., 2007). Diversity is usually seen as an essential resource for breeding (Tanksley and Mc Couch, 1997, Acosta-Gallegos et al 2007, Cooper, Spillane and Hodgkin, 2001, Spillane and Gepts, 2001), especially for new and unpredictable environments., Exploiting this "lost" or "neglected" diversity of hulled wheats for breeding elite cultivars is thus very attractive (FAO, 1996). This diversity is classically screened to seek for new resistance, find adaptation to harsh environments or to detect capacities to uptake more resources (water, light and nutrients) from the environment (Tanskley and McCouch, 1997).

But the exploitation of this diversity is very challenging since the gene pools spanning these 12000 years of evolution have been diverging for growth habits, adaptation to new cropping practices and to very different environments, from the harsh and competing condition in the wild to the highly controlled and fertile conditions of a modern field. In maize, it has been suggested that 4% of genes experienced a selective episode from the wild form to the crop (Wright et al 2005). In durum the transition from wild to domesticates also involved numerous QTLs (Peleg et al., 2011) and many physiological pathways have been differentially tuned (Papa, 2013). Mixing alleles selected in different conditions may therefore results in some physiological incompatibilities at the whole genome level, since many traits are constrained by contradictory trade-offs, e.g., the classical apparent and strong negative correlation between productivity and protein content (Bogard et al., 2010).

The knowledge of the genetic structure of diversity of the *T. turgidum* compartment taken as a whole, the creating of a pre-breeding germplasm gathering the diversity of the wild and the primary domesticated relatives, and the use of new technologies for its exploration and valorization is challenging. In this lecture note, we will first briefly expose recent work on the genetic structure and diversity of the *T. turgidum* gene pool without any prior on the different taxa. Then we will sum up some results obtained during a classical breeding program, lead in collaboration with French private companies, that included wild accessions, old landraces. Eventually, we will illustrate how the concept of evolutionary breeding (Suneson, 1956, 1969, Brown et al., 1990, Phillips and Wolfe, 2005, Wolfe et al., 2008) can be extended to the pre-breeding on durum wheat by creating and monitoring composite cross population to broaden the genetic basis of durum wheat. Finally we will detail how high-throughput sequencing technologies can be used to detect the allelic diversity introgressed in such composite populations. More generally, we are convinced that current breakthroughs in massive DNA sequencing and in massive genotyping relying on thousands of single nucleotide polymorphisms (SNPs) (Kilian and Graner, 2012) are preparing an avenue for the use of this, so far neglected diversity in pre breeding activities (Hajjar and Hodgkin, 1997).

SAMPLING DIVERSITY : LINKS BETWEEN GENETIC STRUCTURE OF *T. TURGIDUM* AND EROSION OF DIVERSITY

Many previous studies on the structure of *T. turgidum* genetic diversity relied on a priori assignation of the samples to the different taxa, based on discriminant morphological traits. This *a priori* classification was used to study the differentiation between taxa and to compare their levels of genetic diversity. But as the taxa discriminant traits may be based on very few major genes, they may not reflect the shared ancestry or the divergence within and between groups (i.e., common morphological

traits may have arisen through different history), and the classification may conceal a very different genetic structure. Moreover, recent or ancient crosses may have altered the initial genetic structure, by introgressing new traits in the different taxa. We applied here a clustering method identifying groups of genetically related individuals without any a priori on the origin of those individuals (DAPC, discriminant analysis of principal components, Jombart, Devillard, & Balloux 2010). We then projected the individual taxon information on the groups obtained by the classification procedure.

Our sample was made 492 individuals: 52 *T turgidum sp. dicoccoides (*DD), 52 *T. turgidum sp. dicoccum* (DC), 29 *T. turgidum sp. polonicum* (PO), 33 *T. turgidum sp. turgidum* (TU), 252 *T. turgidum sp. durum (*DR) covering traditional landraces and elite varieties mostly from the French catalog and 33 *T. turgidum sp. carthlicum* (CA) on which we firstly checked the ploidy level using flow cytometry. For these latter, we kept only 4X accessions since carthlicum accessions may count (2n=4X=28 chromosomes, 4X) or (2n=6X=42 chromosomes, 6X) (see Thuillet et al., 2005 for details). Fourteen microsatellites locus (table 1) were used to genotype the whole sample on a capillary sequencer. The ADEGENET R package was used for the discriminant analysis of the groups (Jombart, 2008).

Table 1: List and position of the 14 microsatellite locus used to genotype the 457 accessions.

Locus	Chromosome location
Xgpw7577	1B
Xgwm312	2A
Xgwm257	2B
Xgwm374	2B
Xgwm413	2B
Xgwm2	3A
Xgwm285	3B
Xgwm601	4A
Xgwm495	4B
Xgwm234	5B
Xgwm193	6B
Xgpw2103	7A
Xgwm297	7B
Xgwm537	7B

Nine groups were detected using the procedures defined by Jombart et al., (2010), and the distribution of the different a priori taxa among groups is plotted figure 1. Taxa have been sorted according to their relative level of Nei's diversity. This suggests an historical interpretation.

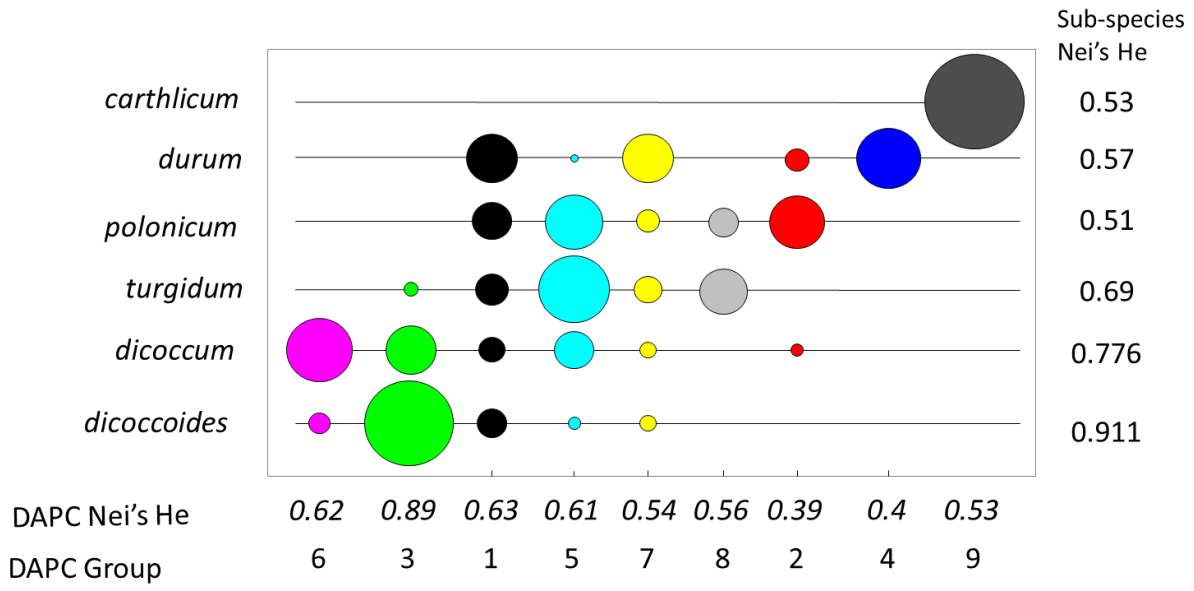


Figure 1. Distribution of the *a priori* taxon assignation of 492 *Triticum turgidum* spp. accessions within the DAPC groups obtained by a discriminant analysis of principal components obtained from 14 microsatellites locus (DAPC). The areas of the circle are proportional to the relative proportion within taxon. Nei's He are the Nei diversity index.

Dicoccoides is mostly present in DAPC group 3, very few accessions being attributed to other groups for this sub-species. It may be seen as the basal group of the turgidum species with the highest level of diversity (He=0.89). A significant fraction of cultivated emmer accessions also belongs to this DAPC group 3, they could be considered as the closest cultivated emmer to the wild emmer. The DAPC group 6 is mostly built on a portion of the sampled cultivated emmers and only a tiny portion of wild emmer accessions also belongs to this group. Note that some wild emmers in the DAPC group 6 could have been misclassified or be somewhat introgressed by domesticated emmers (Luo et al., 2007). No other sub-species contribute to this DAPC group 6 which appears then to be relatively disconnected of the rest of the cultivated turgidum sampled (graph not shown). This remarkable result suggests that this group did not participate to the emergence of the free-threshing wheat and remained isolated from the other cultivated sub-species. It has a relatively high level of diversity (He= 0.62). No obvious geographic localization could explain this structuration among cultivated emmers. Recent work support a polyphyletic origin of domesticated emmer from different sources of wild emmer (Civáň et al., 2013) and our results are somewhat congruent with this assumption.

The groups 1, 5, 7 and 8 have a complex composition which underlines the difficulty to resolve the taxonomy of *T. turgidum* in terms of history, using molecular markers. Namely, ssp. polonicum, turgidum and durum do not correspond to clear and distinguishable genetic entities. The group 1 appears relatively polymorphic (He=0.63) and spans all the sub species with the exception of carthlicum. This group may descent from the original domesticated genepool from which evolved the free threshing forms. Its complex structure is closed to that of the group 7 except that this latter has a reduced level of diversity.

The DAPC groups 2 and 4 are made almost exclusively with durum accessions and show a strong reduction of diversity (He=0.4). All *carthlicum* are grouped in a specific goup, DAPC group 9, with a medium level of diversity. These results are in strong agreement with the recent work of Laidò et al ; (2013). Our data does not permit to elucidate the origin of this group. *Triticum carthlicum* spikes

resemble those of *Triticum aestivum* L. rather than those of free-threshing tetraploid wheats (Haque et al., 2011). The existence of 6X accessions in co-existence with 4X accessions suggests that this sub-species has had a specific evolutionary pattern and it may result from recurrent intercrossing between 4X and 6X specific gene pools in Georgia, Armenia, Azerbaijan, northern Iraq, and Iran where it is still cultivated (Metakovski et al., 1989)

In brief, like other authors in recent works (Civáň et al., 2013, Laidò et al., 2013), we found that *T. turgidum spp* diversity should not be based only on keys determining their sub species status. Like in molecular phylogeny, morphological resemblance or difference may be or may be not linked to a common or divergent evolutionary history. More works should be dedicated to a fine analysis of the origin of the different emmers, their potential and respective implication in the formation of the naked wheats. The understanding of origin and evolution of carthlicum also deserves deeper and appropriate sampling. Indeed these wheats are a very valuable source of traits for durum breeding. They have resistance to drought, frost, and resistance to ergot infection.

IMPACT OF MODERN BREEDING ON DURUM WHEAT STRUCTURE AND GENETIC DIVERSITY

Focusing on durum wheat, a more precise and recent pattern appears. We split the sample before and after 1950. After 1950, varieties were distributed by decades according to their registration to the French catalog. Their distribution between the different DAPC groups and their relative Nei's heterozygosity are plotted on figure 2. Landraces clearly belong to the DAPC group 1, the somewhat undifferentiated group described before. In 1950, two main groups 7 and 4 appear and a minor group (group 2) as well. These groups clearly experienced a strong reduction of diversity, the group 4 being the less diverse. A temporal evolution is also observed from 1950 to the post 1990's varieties. If the Nei's heterozygosity was around 0.56 in landraces and in the 1950's varieties, it regularly decreased and is now as low as 0.4, less than half of that found in dicoccoides (group 3). Modern breeding for the transition to short stature but probably also more recent effort for developing varieties with high quality standards (e.g., selection on the gliadin profile) led to a strong and continuous reduction in genetic diversity. Selection in interaction with genetic drift, probably at the whole genome level (selective sweep like in bread wheat (Cavanagh et al., 2013)) is likely responsible of this dramatic reduction of genetic diversity in modern cultivars. This confirms previous results (Thuillet et al., 2005) and more recent work on durum (Laidò et al., 2013). This continuing erosion of genetic diversity is alarming.

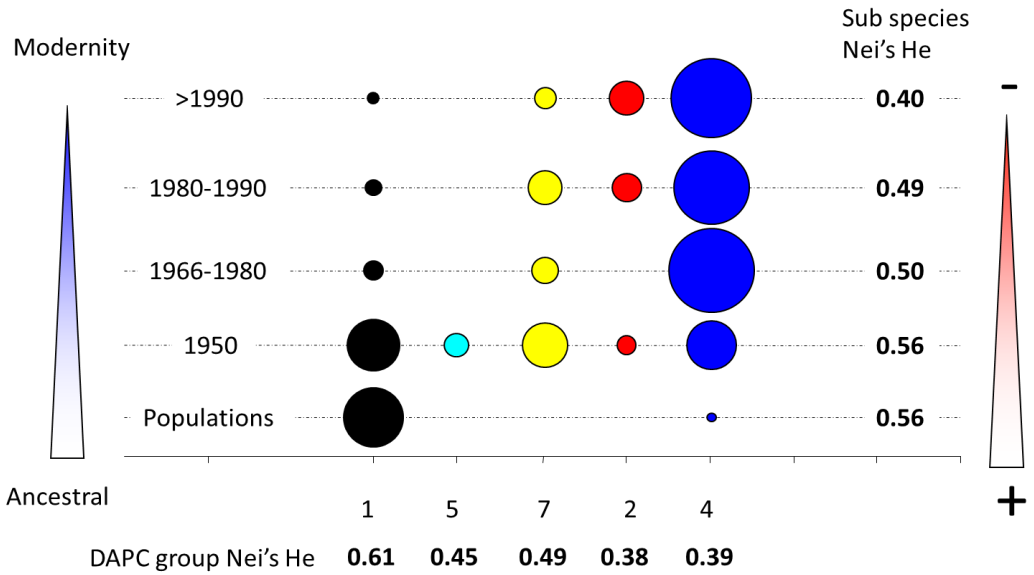


Figure 2: Distribution of the assignation of 252 *T. turgidum sp. durum* accessions within the DAPC groups obtained by a discriminant analysis of principal components obtained from 14 microsatellites locus (DAPC). These accessions are varieties sorted according to their period of release (see text for detail.). The areas of the circle are proportional to the relative proportion within time period. Nei's He are the Nei diversity index.

USE OF EXOTIC DIVERSITY IN PEDIGREE BREEDING

Diversity *per se* may not be an interest even if the pledge for new alleles is attracting. Introgressing dicoccoides alleles in elite germplasm is usually realized after an identification of promising parents, a cross and successive backcrosses to eliminate undesirable chromosome fragments from the donor. This method of backcross has demonstrated a real efficiency to transfer monogenic traits, mostly resistance. In a larger view, broadening the genetic basis of a crop necessitates another, less targeted approach. Observing that the loss of diversity in modern durum wheat is really strong and assuming that wild and exotic germplasm can carry a number of valuable alleles for many traits, not easy to evaluate, or even not easy to identify at their sub species level, methods of non-targeted introgressions, guided by the idea of a broad and non-targeted introgression of new diversity have been proposed as a new pre-breeding challenge. With the help of a guilde of French durum wheat breeders (GIE Blé dur) we investigated the interests of the use of such germplasm.

During several years, more than 200 crosses have been realized between a core collection of tetraploid accessions and a set of elite genitors provided by the private partners. The core collection has been built from a 600 accessions sample by maximizing allelic richness on a set of 30 microsatellite locus used in diversity study (Thuillet et al., 2005). F2 seeds were distributed in a multi-site network of public (INRA) and French private partners (DESPREZ, SERASEM-RAGT, EURODUR-LIMAGRAIN, SYNGENTA, BENOIST, GAE), and a classical pedigree breeding method has been carried out to start the fixation of valuable lines that were generally used as genitors for backcrossing on durum elite lines. Several hundreds of thousands of individual F2 plants were evaluated by the partners. High throughput phenotyping for quality traits (protein content, yellow colour, semolina yield) were applied in the F4 selected families. Multi-site evaluation for frost and rust resistance were carried out on F5 families. The positive qualities of this material lied mostly in disease resistance (leaf rust, head blight, Mosaic virus (Wheat Spindle Streak Mosaic Virus, Soilborne Cereal Mosaic Virus) and a large morphological

diversity. After the removal of the unfavorable undomesticated or primitive traits such as brittle rachis, hulled kernels and tall tillers, the main caveats of this material were defaults in the kernel size and color in crosses involving *T. dicoccoides* in their genealogy, a lack of productivity and possibly a inefficient remobilization of nitrogen from leaves to the kernel during the senescence period. *T. polonicum* appeared as a very good source of kernel color and some accessions had good roots implantation. These primary and empirical observations justify new studies about the impact of domestication and recent breeding on the *durum* plant physiology, its N economy during its whole lifespan from uptake to remobilization. Domesticated and elite favorable alleles at some key locus, yet to be identified, may be important to explain a good end-use quality and productivity. Due to this lack of productivity, most of private breeders finally stopped this program since the agronomical level of the advanced inbred lines was not sufficient to register elite varieties for the current fertilization and treatment practices. To keep up recombination and pre-breeding effort, INRA and Agri-Obtention went forward for more years with a policy of recombining several promising lines together as a priority instead of backcrossing recurrently on elite durum. The resulting most promising material increased in productivity, kept a good level of protein content and showed a high level of leaf rust resistance in untreated, and in treated as well, experiments (fig 3). Two lines are currently (in 2013) following the French registering procedure and we hope they will finally be registered in 2014-2015. Their advantage seems to root in their good level of leaf rust resistance even under treated conditions, that may itself come from the introgression of new major genes of resistance to this disease. In this case, this advantage might only last some years until new virulent strains of leaf rust become adapted to these new sources if the new lines finally succeed in being cultivated on a large area. This demonstrates that genetic advance for yield and quality relies on complex interactions between sanitary aspects, potential productivity in varying environments. These results clearly indicate that pursuing efforts in long term recombination and selection could permit a valorization of exotic germplasm for durum wheat breeding. Productivity and quality traits can be improved, either via better exploitation of the resources and adaptation to harsh environment but also by using new patterns of genetic resistance to main disease. In the present case, investigations are necessary to identify the genetic basis of this enhanced leaf rust resistance in order to assess their sustainability. If finally registered, these lines will constitute the demonstration that the erosion of genetic diversity in the modern elite pool of durum wheat can be stopped and that new diversity will be available for all breeders. Whole genome investigation will rapidly permit to estimate in which chromosomic regions new alleles are brought by these new lines.

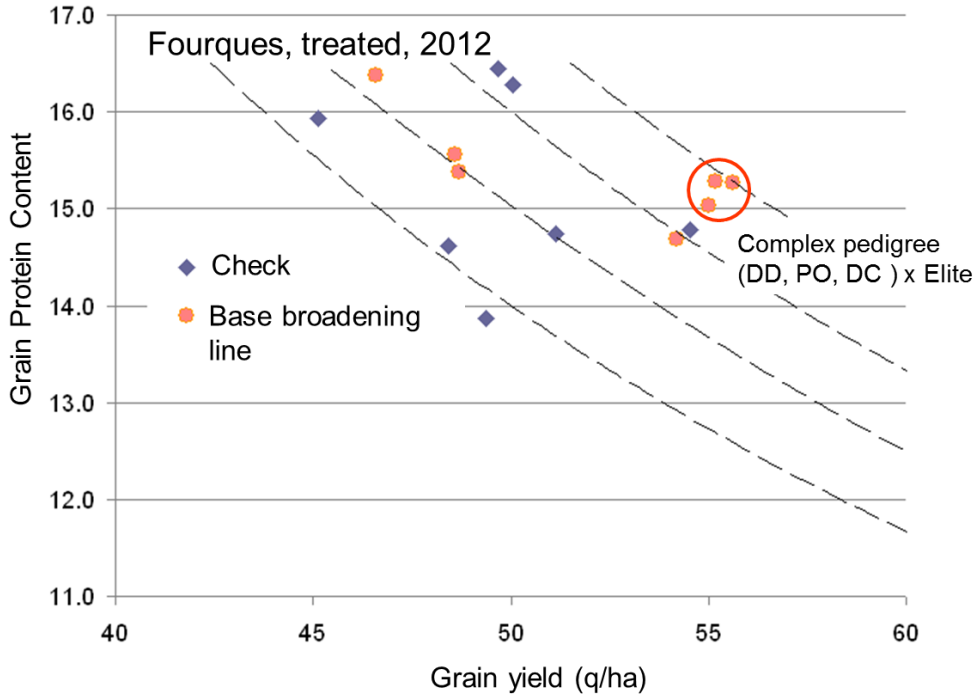


Figure 3. Comparison of agronomic performances of check lines (recent elite French durum cultivars) in blue versus lines derived from the pedigree base broadening program led by INRA Montpellier. Data are from Fourques in 2012 in a treated experiment. Dashed lines are the gradient of the grain yield x protein content product, i.e., the yield in protein/ha. The circled lines apply currently for a registration to the French durum wheat catalog.

EVOLUTIONARY PRE-BREEDING : PRESENTATION OF THE PRE-BREEDING POPULATION OF DURUM WHEAT

The term "pre-breeding" refers to the transfer of genes from related wild ancestors or from ancient varieties to breeding material (FAO, 1996). Pre-breeding activities span a very large set of methods, from interspecific crosses followed by recurrent back crosses to the management on the long term of composite cross populations. In this latter case, recombination and soft selection are used to introgress exotic material in an elite gene pool. Barley composite cross, started by Suneson (1956, 1969) and whose evolution was described by Allard and many others brought information about the very dynamic evolutionary processes at work in such long term monitored composite cross. More recent work on bread wheat confirmed that heterogeneous gene pools can adapt rapidly to different situations including climate gradients (Le Boulc'h et al., 1994, Goldringer et al., 2006, Rhoné et al., 2010), pathogen pressures (Paillard et al., 2000a, 2000b) maintaining their genetic diversity (Enjalbert et al., 1999). The lessons drawn from such experiments are that natural selection leads to adaptation to local condition (climate, disease) but also to competition between different architectural traits (Goldringer et al., 2001). Then, increasing the plant height in response to light competition may also drive to the fixation of bad alleles for productivity, as observed in all wheat populations in which semi-dwarf alleles disappeared (Le Boulc'h et al., 19994) and the harvest index evolved negatively.

Creating and managing composite cross (CC) on the long term could be a really interesting pre-breeding method but methodological work should be devoted to understand the interplay between recombination, natural and human selection and genetic drift. Empirical and efficient rules for managing and improve such composites would permit to avoid undesirable evolution and create

interesting new germplasm. Their role is to introgress massively interesting diversity at a whole genomic level. This needs that recombination and selection are finely tuned and like the CC cross of barley they have to be maintained for an indefinite number of generations and constitute an evolving reservoir of diversity (Henry et al., 1991). On durum wheat, our laboratory launched a pilot experiment ; we used a durum wheat population in which segregates a nuclear male sterility gene donated by a former French INRA scientist, François Kaan. The male fertile allele Ms is dominant on the male sterile allele. Plants can be either hermaphrodites (Ms/Ms or ms/ms) or male sterile (ms/ms). A collection of flowering *T.t.dicoccoides*, T.t. *dicoccum*, Tt *polonicum* accessions were crossed in 1997 on male sterile plants of this population. The resulting seeds were used to found a pre-breeding composite cross, the INRA Pre breeding durum wheat population (hereafter named IPBDWP).

Our aim was to combine recombination by promoting outcrossing and rapid fixation of favorable combinations by permitting selfing. The population is thus monitored under a mixed mating system thanks to the male sterility gene. This population is being reproduced as follows: every year, once the flowering starts and until harvest, the tallest tillers are eliminated to avoid a detrimental evolution of the IPBDWP due to competition of tall plants on short plants, male sterile spikes are identified by their wide glume opening at the blooming stage and marked by a red twist. These marked spikes are harvested and threshed in bulk separately from the selected fertile spikes. Hermaphrodite spikes are chosen visually at harvest for their shape, vigor and health status and then threshed in bulk. The new generation is composed then by 20% of seeds coming from the marked male ms/ms sterile spikes (outcrossing portion), 70% of the selected hermaphrodite spikes Ms/ms and Ms/Ms (selfing portion) and 10% coming from the best lines selected in the pedigree selection scheme presented above to bring new diversity and agronomic performance. The population introgressed a new diversity and is experiencing recombination at each generation, fixation of new combinations under the combined effect of anthropic selection for a return to agronomical conditions, natural selection for adaptation to the environment and of course random genetic drift. The restricted amount of outcrossing (20%) reduced also the selective pressure to adapt to allogamy which can be the major evolutionary force in such population of usually selfing crops (David et Pham, 1993). The project is now to verify the interest of such resources for breeding, either as a source of new alleles or gene combination or as a tool for deciphering the genetic basis of traits.

Recently interest in genome wide association studies (GWAS) pointed out the value of diversified panels to accurately detect chromosomal segments carrying valuable alleles for interesting agronomical traits (Maccaferri et al., 2010, 2012). As most of these panels are assembled from large and diverse collections, genetic structuration among accessions may lead to a high level of false positive associations. Even if several methods take into account this structuration, coping with it remains a challenge (Macafferri et al., 2005). The interest of evolving composite populations in the GWA approach is that the population can be seen as a reproductive unit and after several generations of partial outcrossing and effective recombination, a reduction of the genetic structuration and a consequent reduction of the statistical linkage between locus, especially those that are not closely physically linked is expected. Consequently the False Discovery Rate (FDR*), i.e.,* the ratio with spurious association between a polymorphism and a variation of a trait should decrease substantially in a composite cross compared to a panel made of lines from different geographical areas, from different periods or different breeding programs. The other interesting aspect of GWA in a diversified panels compared to biparental segregating population comes from a more robust estimation of allelic effects.

In the following, we investigated for the first time the genetic diversity content of a composite cross of durum IPBDWP and estimated the extent of linkage disequilibrium along chromosomal segments to determine whether such populations might be a good support for GWAS studies.

FIRST GENOMIC INVESTIGATIONS IN THE IPBDWP

Using New Generation Sequencing, information on thousands of candidate genes and candidate regions can be harnessed for thousands of individuals to sample genetic diversity within and between germplasm pools, to map Quantitative Trait Loci (QTLs), to identify individual genes and to determine their functional diversity (Kilian and Graner, 2012). Here we applied for the first time in durum wheat such an approach on our population.

Data production and SNP detection : In 2009, 500 spikes were randomly harvested in IPBDWP and entered a 2 year fixation process. Hundred and six (106) of these lines were used to investigate the level of genetic diversity available in this composite. Seeds were germinated in growth chamber in standardized conditions and young coleoptiles were used to extract RNAs. CDNAs libraries were produced and tagged for each of these 106 genotypes. These 106 libraries were pooled, either by 24 samples or by 48 samples and sequenced on a HiSeq 2000 to produce 100 pb read pairs. Finally 813,110,268 cleaned reads were used to produce a *de novo* assembly using a bio-informatic pipeline (publication in prep). To separate homeologs between their A and B copies, we used an algorithm based on unbalanced expression ratio between the two copies implemented in the Homeosplitter software (Ranwez et al., 2013). The good split of copies were verified when possible by mapping reads sequences on *T. urartu and Ae. speltoides* transcriptomes produced and assembled by the same protocols. Finally, only good quality SNPs with no heterozygous excess were used in this preliminary study to evaluate the level of diversity and the extent of the linkage disequilibrium in IPBDWP.

Nucleotide diversity (π) was estimated as proposed by Tajima (Tajima 1983). To estimate the decay of linkage disequilibrium, it was first necessary to obtain the position of the SNP on a reference genetic map. In this preliminary work, no segregation data were available for these SNPs in durum wheat and we used external and public data from bread wheat. Contigs containing SNPs were blasted against the sequence of the 9K SNP array defined and mapped on bread wheat polymorphism using several segregating populations (Cavanagh et al., 2013). To eliminate possible errors and to ensure appropriate genome localization, we kept only SNPs for which the genome localization was identified identically by mapping on bread wheat (Cavanagh et al., 2013) and properly assigned to a donor diploid species in our data (*T. urartu* for the A genome, *Ae. speltoides* for the S genome). Pairwise linkage disequilibrium was then computed and plotted against genetical distance between locus, distance estimated from the bread wheat data (Cavanagh et al., 2013)

DIVERSITY AND LINKAGE

Finally 13 911 SNPs on 5980 locus fulfilled the conditions to be kept for the study*, i.e.*, no excess of heterozygous individuals. The nucleotide diversity π was computed for these 13, 911 SNP and the obtained values vary between $\pi \sim 0.5$ to $0.9\ 10^{-3}$ per base pair. From previous evaluation on 21 genes (Haudry et al., 2007), estimations for the wild dicoccoides is $\pi \sim 2.5\ 10^{-3}$, $\pi \sim 1.3\ 10^{-3}$ for dicoccum and $\pi \sim 0.4\ 10^{-3}$ for all durum. It seems thus that if the population has effectively a good level of diversity compared to the whole durum sub-species it is still far from what it could have been if a large part of diversity from wild and cultivated emmer had been successfully introgressed in IPBDWP.

Out of the 5980 contigs, 553 blasted on the sequences of 9k bread wheat SNP arrays, giving a total of 1858 SNPs for studying the linkage disequilibrium decay. Discarding ambiguous SNPs, attributed to different genomes by the bread wheat mapping and by the proximity to one of the diploid ancestor, 1577 SNPs could be eventually used to evaluate the decay of the linkage disequilibrium in IPBDWP. Figure 4 illustrates this decay on the chromosome 1A, the other chromosomes showing very similar patterns. As expected, the disequilibrium values between pairs of SNPs located in the same contig have

the highest value, but their average value is far from the maximum value of 1, which would have mean complete linkage within genes and a low haplotype complexity. In this presentation paper, this apparent lack of linkage has not been fully investigated but it could mean that introgression of wild and exotic accessions has effectively enriched the haplotype diversity at very short genetic distance. Nevertheless, if some *de novo* assembled contigs are still chimeric between the A and B genome, low spurious r^2 values between some pairs of homeologous SNPs could decrease artificially the within contig linkage estimation. Between different contigs the decay of the linkage disequilibrium is decreasing very fast and is lower than the value found by Maccaferri et al 2005 using microsatellites. A threshold value for r^2 around 0.1 is found after 70 cM very close to the value (dashed line) found for SNPs located on different chromosomes. Naturally, deeper investigations are needed to ascertain these linkage disequilibrium patterns but anyway this preliminary data suggest that evolving composite cross such as IPBWDWP could have very good and interesting properties for detecting markers closely linked to causal polymorphisms. They could constitute then very good alternative to association panels.

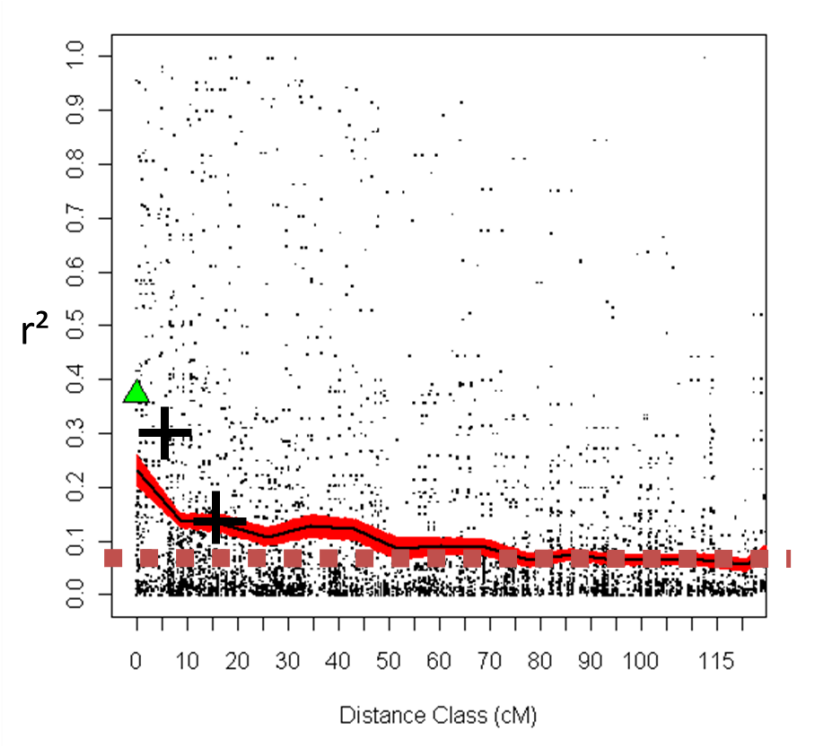


Figure 3. Evolution of linkage disequilibrium (r^2) between pair of SNPs located at different distances (cM) on the chromosome 1A in the INRA pre breeding Durum wheat population (IPBDWP). Mapping positions of the SNPs on the chromosome 1A were predicted by blasting contig sequence containing the SNPs on the sequence of the mapped markers of the 9K bread wheat micro array (Cavanagh et al., 2013). The green triangle is the average value of r^2 when the two SNP of a pair are located in the same contig. Black crosses are the values of r^2 estimated by Maccaferri et al., 2005 at similar distance classes in a durum wheat panel. In red, the within segmental average value of r^2 for 20 subsequent windows of equal genetic distance spanning the whole chromosome. The dotted brown line is the average value of pairwise r^2 when the two SNPs were assigned to different chromosomes.

Conclusions & Perspectives

The modern elite pool of durum wheat has experienced several severe reduction of genetic diversity, and there is evidence that this genetic erosion is still continuing. The recent cultivars share a lot of common alleles and their deviation from the historical genetic background of the species seems

accelerating at least until the end of the XXth century and for the French elite catalog. The same trend has been observed in the sample of durum recently investigated by Laido et al., (2013). If a lack of diversity in the elite pool is susceptible to impair future advances in the development of a sustainable durum wheat production, for which disease resistances, efficient nutrient uptake capacities, growth and flowering in harsher environmental conditions will be needed, the use of the genetic diversity of the whole *T. turgidum* species may be a key element of a germplasm development and integration strategy. But this will be a real challenge. If many studies demonstrated the worth of the genetic diversity hold in genebank in the whole *Turgidum subsp*, especially in the wild and cultivated emmers, the use of this valuable diversity is not easy and may not be really successful if one excepts the use of valuable alleles at some major genes, such as disease resistance. In a collaborative program between INRA and GIE Blé dur, classical breeding led to some results by using intensive back crossing after the initial cross but the selected lines were not sufficiently productive in the first place to be registered as elite varieties in the French catalog by the Private Breeders. Nevertheless, some success was obtained by persevering in recurrently crossing advanced lines with introgressed backgrounds. Productivity eventually increased and some lines might become registered varieties in a close future, probably thanks to their good level of resistance to brown rust. This tolerance to rust probably provided a yield advantage to these advanced lines in an experiment where rust attack was important. This success should be confirmed on the long range since a quick overcome of the allelic of resistance is likely in the case of their commercial development

As an alternative to this quick use of valuable germplasm, long term evolutionary pre-breeding programs may be of a great interest for creating new germplasm, integrating new alleles, promoting recombination and soft selection in populations of reasonable population effective size. In this paper, we reported the very first results on a composite cross population of a durum wheat population with a broaden genetic basis monitored for 12 years under a 20% outcrossing mating system. This current IPBDW population appears as an interesting resource for GWAS because of its reasonable level of genetic diversity, reduced long distance linkage disequilibrium and large phenotypic variation (data not shown). We are currently accumulating phenotypic data on a large number of traits (morphology, phenology, N status of leaves and grains) to verify if the sequencing effort yielded sufficient data to detect associations. RNAseq data obtained here will be directly be used as a genotyping method (Genotyping by sequencing, GBS) but a number of missing data arose since gene expression may greatly vary among individuals. The coverage of RNA seq for each individual, the standardization of the growing conditions before RNA extraction and the development of adapted bio informatics pipelines are key elements for the success of a RNA seq GBS approach. The detected SNPs here can also contribute to the assembly of a specific durum polymorphism database that can be used to develop a micro array chip within a durum wheat consortium.

Our 20% outcrossing mating regimes clearly reduced the long distance linkage disequilibrium in the population and also probably also reduced greatly the within population genetic structure that usually creates spurious association in GWA studies using panels assembled from different genetic sources. More methodological work is needed to set the most efficient value of the outcrossing rate in order to promote effective reduction of haplotype length, reduction of kinship structuration but also to promote a rapid fixation of valuable homozygous individuals in the population. If IPBWP appears as genetically diverse compared to a durum wheat panel, its nucleotide diversity is still much lower than the potential diversity available in the exotic parents of the composite. Selection for plant height, removal of plants showing genetic incompatibility and other unidentified selective pressures for adaptation to climate and local pathogens may explain a strong loss of diversity by linkage drag and selective sweeps around the domesticated alleles at locus determining minimum agronomical values. If the decay of linkage disequilibrium is rather steep in the population, low levels of linkage are still

present at 50 – 70 cM. This suggests that effective recombination, led by the 20% outcrossing level, was not sufficient to break rapidly and efficiently mix the elite haplotypes with the introgressed ones. Furthermore, if many genes, and not only some major genes major responsible for dramatic and apparent changes in morphology and shape, (e.g., brittle rachis), have been involved in domestication and further improvement of durum quality and agronomic performance, it is likely at the whole genome level that valuable alleles in the exotic germplasm have good chance to be regularly associated with unfavorable alleles. In this case, a more appropriate method to enrich the allelic diversity of such pre-breeding populations would have been first to promote 100% outcrossing and recombination during the first generations before starting any conscious massal selection for a return to a "durum" like morphology compatible with modern agronomical practices. Such a strategy should reduce the number, strength and extent of the selective sweeps.

In conclusion we claim that new composite populations should be created by controlled crosses of male sterile plants with wild and cultivated emmers, traditional durum, polonicum and turgidum landraces and carthlicum as well sampled to cover the whole diversity of the genetic groups described in this paper (figure 1) or discovered elsewhere. From our first experience in IPBDW, outcrossing rate should be increased to promote effective and rapid recombination to avoid strong selective sweeps. The interplay between outcrossing and selection practices should be theoretically investigated. Our selection practices in IPBDW were probably too strong to eliminate wild traits such as dispersal or asynchronous growth habits, tallness and hulled kernels. Accepting that these traits co-exist for longer period along with the domesticated phenotypes could be a key for a good introgression of larger levels of diversity in valuable pre-breeding composites. This claims for theoretical approaches to deliver methodological recipes to create, monitor and use of evolutionary pre-breeding populations.

If our population is evolving in only one environment (Montpellier; Southern France), such prebreeding composites can be used to create a network of connected populations evolving in contrasted environments. Diversifying selection on a similar genetic background may help to detect chromosomal regions involved in different adaptations patterns (Beaumont & Nichols, 1996, Enjalbert et al., 1999) and are very well adapted to an international collaboration. IPBDW is available for distribution, lines and associated molecular data will soon be released.

REFERENCES

Acosta-Gallegos, J. A., Kelly, J. D., & Gepts, P. (2007). Pre-breeding and genetic diversity in common bean (Phaseolus vulgaris). In *Proceedings of international plant breeding symposium, Ciudad de México. Crop Sci* (Vol. 47, No. S3, pp. S44-S59).

Alonso-Blanco, C., Aarts, M. G., Bentsink, L., Keurentjes, J. J., Reymond, M., Vreugdenhil, D., & Koornneef, M. (2009). What has natural variation taught us about plant development, physiology, and adaptation?. *The Plant Cell Online*, *21*(7), 1877-1896.

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.

Bogard, M., Allard, V., Brancourt-Hulmel, M., Heumez, E., Machet, J. M., Jeuffroy, M. H., ... & Le Gouis, J. (2010). Deviation from the grain protein concentration–grain yield negative relationship is highly correlated to post-anthesis N uptake in winter wheat. *Journal of experimental botany*, *61*(15), 4303-4312.

Bozzini, A. (1988). Origin, distribution, and production of durum wheat in the world. Pages 1-16 in: *Durum Wheat: Chemistry &Technology. G. Fabriani and C. Lintas, eds. Am. Assoc. Cereal Chem.: St. Paul, MN.*

Brown, A. H. D., Clegg, M. T., Kahler, A. L., & Weir, B. S. (1990). *Plant population genetics, breeding, and genetic resources*. Sinauer Associates Inc..

Buckler IV, E. S., & Thornsberry, J. M. (2002). Plant molecular diversity and applications to genomics. *Current opinion in plant biology*, *5*(2), 107-111.

Cavanagh, C. R., Chao, S., Wang, S., Huang, B. E., Stephen, S., Kiani, S., ... & Akhunov, E. (2013). Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *Proceedings of the National Academy of Sciences*, *110*(20), 8057-8062.

Civáň P, Ivaničová Z, Brown TA (2013). Reticulated Origin of Domesticated Emmer Wheat Supports a Dynamic Model for the Emergence of Agriculture in the Fertile Crescent. PLoS ONE 8(11): e81955. doi:10.1371/journal.pone.0081955

Cooper, H. D., Spillane, C., & Hodgkin, T. (Eds.). (2001). *Broadening the genetic base of crop production*. CABI.

David, J. L., & Pham, J. L. (1993). Rapid changes in pollen production in experimental outcrossing populations of wheat. Journal of evolutionary biology, 6(5), 659-676.

Enjalbert, J., Goldringer, I., Paillard, S., & Brabant, P. (1999). Molecular markers to study genetic drift and selection in wheat populations. *Journal of experimental botany*, *50*(332), 283-290.

FAO (1996). Report of the State of the World's Plant Genetic Resources for Food and Agriculture (Rome).

Feldman, M., and Kislev, M. E. 2007. Domestication of emmer wheat and evolution of free-threshing tetraploid wheat Special Issue: A Century of Wheat Research-From Wild Emmer Discovery to Genome Analysis. Israel Journal of Plant Sciences. 55:207-221.

Goldringer, I., Prouin, C., Rousset, M., Galic, N., & Bonnin, I. (2006). Rapid differentiation of experimental populations of wheat for heading time in response to local climatic conditions. *Annals of Botany*, *98*(4), 805-817.

Goldringert, I., Enjalbert, J., David, J., Paillard, S., Pham, J. L., & Brabant, P. (2001). 13 Dynamic Management of Genetic Resources: a 13-year Experiment on Wheat. *Broadening the Genetic Base of Crop Production*, 245.

Hajjar, R., Hodgkin, T., (2007). The use of wild relatives in crop improvement: a survey of developments over the last 20 years. Euphytica, 156: 1-13.

Haque, M. A., Takayama, A., Watanabe, N., & Kuboyama, T. (2011). Cytological and genetic mapping of the gene for four-awned phenotype in Triticum carthlicum Nevski. *Genetic Resources and Crop Evolution*, *58*(7), 1087-1093.

Haudry, A., Cenci, A., Ravel, C., Bataillon T., Brunel D., Poncet C., Hochu I., Poirier S., Santoni, , S., Glémin S. & David, J. (2007). Grinding up Wheat: A Massive Loss of Nucleotide Diversity since Domestication Molecular Biology and Evolution 2007 24(7):1506-1517.

Henry, J. P., Pontis, C., David, J., & Gouyon, P. H. (1991). An experiment on dynamic conservation of genetic resources with metapopulations. In *Species conservation: a population-biological approach* (pp. 185-198). Birkhäuser Basel.

Jombart, T., (2008). Adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics, 24:1403-1405.

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC genetics*, *11*(1), 94.

Kilian, B., & Graner, A. (2012). NGS technologies for analyzing germplasm diversity in genebanks. *Briefings in functional genomics*, *11*(1), 38-50.

Kilian, B., Özkan, H., Pozzi, C., & Salamini, F. (2009). Domestication of the Triticeae in the Fertile Crescent. In *Genetics and Genomics of the Triticeae* (pp. 81-119). Springer US.

Laidò, G., Mangini, G., Taranto, F., Gadaleta, A., Blanco. A., et al., (2013) Genetic Diversity and Population Structure of Tetraploid Wheats (*Triticum turgidum* L.) Estimated by SSR, DArT and Pedigree Data. PLoS ONE 8(6): e67280. doi:10.1371/journal.pone.0067280

Le Boulc'h, V., David, J. L., Brabant, P., & de Vallavieille-Pope, C. (1994). Dynamic conservation of variability: responses of wheat populations to different selective forces including powdery mildew. *Genetics Selection Evolution*, *26*(Supplement), 221s-240s.

Luo, M. C., Yang, Z. L., You, F. M., Kawahara, T., Waines, J. G., and Dvorak, J. (2007). The structure of wild and domesticated emmer wheat populations gene flow between them, and the site of emmer domestication. Theor. Appl. Genet. 114:947–959.

Maccaferri, M., Sanguineti, M. C., Demontis, A., El-Ahmed, A., del Moral, L. G., Maalouf, F., ... & Tuberosa, R. (2011). Association mapping in durum wheat grown across a broad range of water regimes. *Journal of experimental botany*, *62*(2), 409-438.

Maccaferri, M., Sanguineti, M. C., Mantovani, P., Demontis, A., Massi, A., Ammar, K., ... & Tuberosa, R. (2010). Association mapping of leaf rust response in durum wheat. *Molecular breeding*, *26*(2), 189-228.

Maccaferri, M., Sanguineti, M. C., Noli, E., & Tuberosa, R. (2005). Population structure and long-range linkage disequilibrium in a durum wheat elite collection. *Molecular Breeding*, *15*(3), 271-290.

Metakovsky, E. V., Kudryavtsev, A. M., Iakobashvili, Z. A., & Novoselskaya, A. Y. (1989). Analysis of phylogenetic relations of durum, carthlicum and common wheats by means of comparison of alleles of gliadin-coding loci. *Theoretical and applied genetics*, *77*(6), 881-887.

Nesbitt, M., and Samuel, D. 1996. From stable crop to extinction? The archaeology and history of the hulled wheats. Pages 41–100 In: Hulled Wheats. Proceedings of the first international workshop on hulled wheats, Castelvecchio, Pascoli, Tuscany. S. Padulosi, K. Hammer, J. Heller (Eds.). International Plant Genetic Resources Institute, Rome.

Özkan, H., Willcox, G., Graner, A., Salamini, F. and Kilian, B. (2011). Geographic distribution and domestication of wild emmer wheat (*Triticum dicoccoides*) Genet Resour Crop Evol 58:11–53.

Özkan, H., Brandolini, A., Schaefer-Pregl, R., and Salamini, F. 2002. AFLP analysis of a collection of tetraploid wheats indicates the origin of emmer and hard wheat domestication in southeast Turkey. Mol. Biol. Evol. 19:1797–1801.

Paillard, S., Goldringer, I., Enjalbert, J., Doussinault, G., de Vallavieille-Pope, C., & Brabant, P. (2000a). Evolution of resistance against powdery mildew in winter wheat populations conducted under dynamic management. I–Is specific seedling resistance selected?. *Theoretical and Applied Genetics*, *101*(3), 449-456.

Paillard, S., Goldringer, I., Enjalbert, J., Trottet, M., David, J., de Vallavieille-Pope, C., & Brabant, P. (2000b). Evolution of resistance against powdery mildew in winter wheat populations conducted under dynamic management. II. Adult plant resistance. *Theoretical and Applied Genetics*, *101*(3), 457-462.

Papa, R. (2013, January). Evolutionary Metabolomics of Durum Wheat Domestication. In *Plant and Animal Genome XXI Conference*. Plant and Animal Genome.

Peleg, Z., Fahima, T., Korol, A. B., Abbo, S., & Saranga, Y. (2011). Genetic analysis of wheat domestication and evolution under domestication. *Journal of experimental botany*, *62*(14), 5051-5061.

Phillips, S. L., & Wolfe, M. S. (2005). Evolutionary plant breeding for low input systems. *The Journal of Agricultural Science*, *143*(04), 245-254.

Rhoné, B., Vitalis, R., Goldringer, I., & Bonnin, I. (2010). Evolution of flowering time in experimental wheat populations: a comprehensive approach to detect genetic signatures of natural selection. *Evolution*, *64*(7), 2110-2125.

Spillane, C., & Gepts, P. (2001). Evolutionary and genetic perspectives on the dynamics of crop genepools. *Broadening the genetic base of crop production*, 25-70.

Suneson, C. A. (1956). An evolutionary plant breeding method. *Agronomy Journal*, *48*(4), 188-191.

Suneson, C. A. (1969). Evolutionary plant breeding. *Crop Science*, *9*(2), 119-121.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, *105*(2), 437-460.

Tanksley SD, McCouch SR, 1997.*Seed banks and molecular maps: unlocking genetic potential from the wild. Science 1997;277:1063-6.*

Thuillet, A-C., Bataillon, T., Poirier, S., Santoni, S., David, JL. (2005) Estimation of Long-Term Effective Population Sizes Through the History of Durum Wheat Using Microsatellite DataGenetics 169: 1589-1599. .

Wolfe, M. S., Baresel, J. P., Desclaux, D., Goldringer, I., Hoad, S., Kovacs, G., & Van Bueren, E. L. (2008). Developments in breeding cereals for organic agriculture. *Euphytica*, *163*(3), 323-346.

Wright, S. I., Bi, I. V., Schroeder, S. G., Yamasaki, M., Doebley, J. F., McMullen, M. D., & Gaut, B. S. (2005). The effects of artificial selection on the maize genome. *Science*, *308*(5726), 1310-1314.

Zaharieva M, Geleta Ayana N., Al Hakimi A., Misra S.C. and P. Monneveux, 2010 Cultivated emmer wheat (*Triticum dicoccon* Schrank), an old crop with promising future: a review. Genet Resour Crop Evol (2010) 57:937–962

VII.3. Amorces et sources des microsatellites utilisés dans la partie III.1.

Microsatellite name	Linkage group	Primers F	Primers R	Publication source
Xgpw3142	1A	GCCTTGCTCCTCTCCTACCT	GTATCACCCATCGCTGCC	Sourdille et al, 2009
Xgwm164	1A	ACATTTCTCCCCCATCGTC	TTGTAAACAAATCGCATGCG	Korzun V et al, 1997/ Roder MS et al, 1998
Xgpw7577	1B	CCCAGCAGCCTTCTCAATAC	CCAGCACCACAACAAGAGTG	Sourdille et al, 2009
Xgwm413	1B	TGCTTGTCTAGATTGCTTGGG	GATCGTCTCGTCCTTGGCA	Roder MS et al, 1998
Xgpw7101	2A	CCTGGTATGTATGGTAGAGCCC	ACCGCAGTGGCTGAAATC	Sourdille et al, 2009
Xgwm312	2A	ATCGCATGATGCACGTAGAG	ACATGCATGCCTACCTAATGG	Roder MS et al, 1998
Xgmw614a	2A	GATCACATGCATGCGTCATG	TTTTACCGTTCCGGCCTT	Roder MS et al, 1998
Xgwm374	2B	ATAGTGTGTTGCATGCTGTGTG	TCTAATTAGCGTTGGCTGCC	Roder MS et al, 1998
Xgmw257	2B	AGAGTGCATGGTGGGACG	CCAAGACGATGCTGAAGTCA	Roder MS et al, 1998 / Paillards S et al, 2003 / Somers DL et al, 2004
Xgwm2	3A	CTGCAAGCCTGTGATCAACT	CATTCTCAAATGATCGAACA	Roder MS et al, 1995 / Roder MS et al, 1998
Xgwm299	3B	ACTACTTAGGCCTCCCGCC	TGACCCACTTGCAATTCATC	Roder MS et al, 1998
Xgwm285	3B	ATGACCCTTCTGCCAAACAC	ATCGACCGGGATCTAGCC	Roder MS et al, 1998 / Somers DL et al, 2003 / Somers DL et al, 2004
Xgpw2239	4A	CAACCATATGCCCAGGAGAC	TGTTGCTGTCTGAAACAGGG	Sourdille et al, 2009
Xgwm601	4A	ATCGAGGACGACATGAAGGT	TTAAGTTGCTGCCAATGTTCC	Roder MS et al, 1998
Xgwm495	4B	GAGAGCCTCGCGAAATATAGG	TGCTTCTGGTGTTCCTTCG	Roder MS et al, 1998
Xwmc710	4B	GTAAGAAGGCAGCACGTATGAA	TAAGCATTCCCAATCACTCTCA	Somers DJ et al, 2004
Xgwm126	5A	CACACGCTCCACCATGAC	GTTGAGTTGATGCGGGAGG	Roder MS et al, 1998
Xgwm213	5B	TGCCTGGCTCGTTCTATCTC	CTAGCTTAGCACTGTCGCCC	Roder MS et al, 1998
Xgwm234	5B	GAGTCCTGATGTGAAGCTGTTG	CTCATTGGGGTGTGTACGTG	Korzun V et al, 1997/ Roder MS et al, 1998
Xgpw3101	6A	CAGGGAGATAGAGGGAAGGG	CTAAAATCGCCAGAAGAAGGG	Sourdille et al, 2009
Xgmw193	6B	CTTTGTGCACCTCTCTCTCC	AATTGTGTTGATGATTTGGGG	Korzun V et al, 1997/ Roder MS et al, 1998
Xgpw2103	7A(-7B)	CGTATGCAGCATGGCATC	GCTATGTTGTGTGGCATTGG	Sourdille et al, 2004
Xgwm297	7B	ATCGTCACGTATTTTGCAATG	TGCGTAAGTCTAGCATTTTCTG	Korzun V et al, 1997/ Roder MS et al, 1998
Xgwm537	7B	ACATAATGCTTCCTGTGCACC	GCCACTTTTGTGTCGTTCCT	Roder MS et al, 1998

VII.4. Détails des calculs du modèle d'évolution d'une population connaissant un épisode de migration

VII.4.1. Expression analytique de l'évolution des fréquences génotypiques au locus responsable de la stérilité mâle

Soit F, le locus du gène de stérilité mâle. Comme il est récessif, les plantes portant les génotypes *FF* et *Ff* sont hermaphrodites, tandis que les plantes portant le génotype *ff* sont des femelles. Les plantes hermaphrodites se reproduisent par autogamie, tandis que les femelles ne se reproduisent que par allofécondation

Soit $P_{jk}^{(g)}$ la fréquence du génotype jk dans la population, à la $g^{\text{ième}}$ génération, avec $j, k = \{F, f\}$. Notons $P_{j,p}^{(g)}$ la fréquence de l'allèle j au sein du nuage pollinique, avec $j = \{F, f\}$. Seuls les plantes hermaphrodites (génotypes *FF* et *Ff*) produisent du pollen. Par conséquent :

$$P_{F,p}^{(g)} = \frac{2P_{FF}^{(g)} + P_{Ff}^{(g)}}{2(P_{FF}^{(g)} + P_{Ff}^{(g)})} \text{ et } P_{f,p}^{(g)} = \frac{P_{Ff}^{(g)}}{2(P_{FF}^{(g)} + P_{Ff}^{(g)})}$$

Soit α la proportion de graines provenant de plantes mâles stériles, et donc d'allofécondation. Par conséquent $1 - \alpha$ est la proportion de graines provenant d'hermaphrodites, et donc provenant d'autogamie.

Notons $P_{jk,♀}^{(g)}$ la fréquence du génotype jk parmi les graines récoltées sur les plantes femelles. La fréquence du génotype jk des graines récoltées sur des hermaphrodites va être notée $P_{jk,♂}^{(g)}$. Enfin, soient $P_{j,o♀}^{(g)}$ et $P_{j,o♂}^{(g)}$ la fréquence des ovules portant l'allèle j, chez les femelles et chez les hermaphrodites, respectivement.

La composition génétique de chaque type de graines (issues de plantes femelles ou de plantes hermaphrodites), va être décrite ci-dessous.

VII.4.1.1 Composition génétique des graines récoltées sur les plantes femelles (allofécondation), pour le locus F

Le processus de fécondation des plantes femelles est montré en **figure 1.** Les plantes femelles ont un génotype *ff*. Par conséquent, elles ne produisent que des ovules avec l'allèle *f*. Ces ovules vont être fécondés par un grain de pollen. Ce grain de pollen va porter l'allèle f avec la probabilité $P_{f,p}^{(g)}$, et l'allèle *F* avec la probabilité $P_{F,p}^{(g)}$.

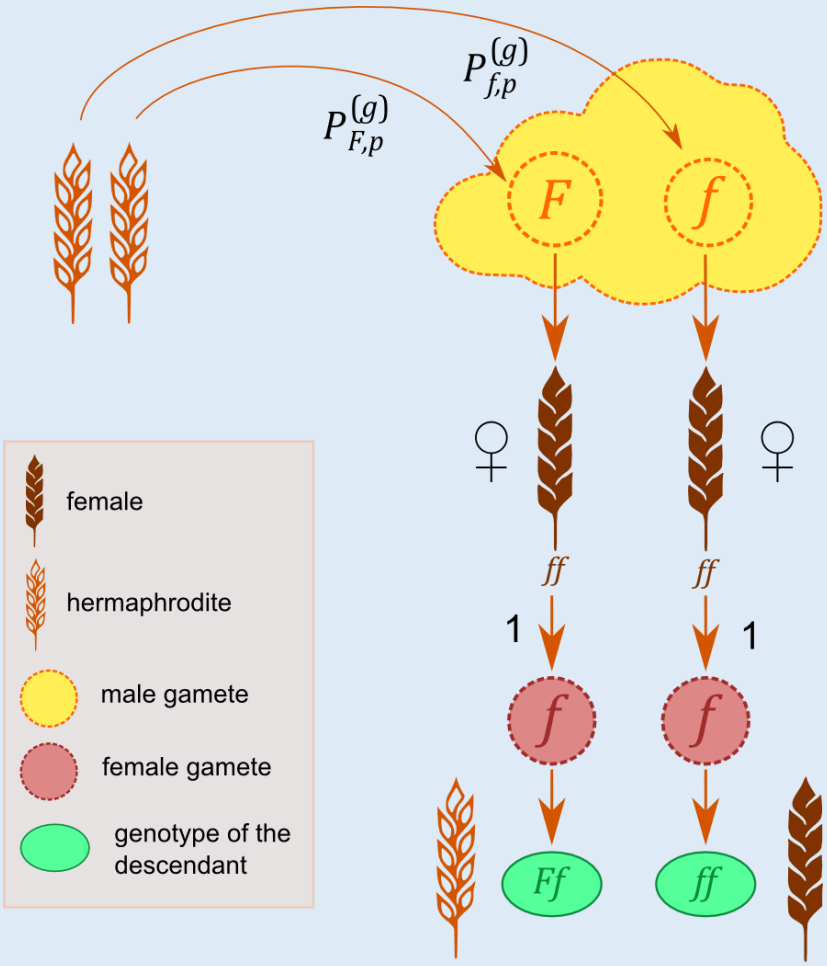


Figure 1: Fécondation des plantes femelles

Les génotypes des graines récoltées sur les plantes femelles dépendent de la composition génétique du nuage de pollen. En effet, les femelles portent toutes le génotype *ff* au locus F.

Par conséquent :

$$P_{ff,♀}^{(g+1)} = P_{f,o♀}^{(g)} \cdot P_{f,p}^{(g)} = P_{f,p}^{(g)} = \frac{P_{Ff}^{(g)}}{2(P_{FF}^{(g)} + P_{Ff}^{(g)})}$$

$$P_{Ff,♀}^{(g+1)} = P_{f,o♀}^{(g)} \cdot P_{F,p}^{(g)} = P_{F,p}^{(g)} = \frac{2P_{FF}^{(g)} + P_{Ff}^{(g)}}{2(P_{FF}^{(g)} + P_{Ff}^{(g)})}$$

$$P_{FF,♀}^{(g+1)} = 0 \text{ comme } P_{fFo♀}^{(g)} = 0$$

Ces équations peuvent être écrites de manière matricielle :

$$\begin{bmatrix} P_{ff,♀}^{(g+1)} \\ P_{Ff,♀}^{(g+1)} \\ P_{FF,♀}^{(g+1)} \end{bmatrix} = \frac{1}{P_{FF}^{(g)} + P_{Ff}^{(g)}} \begin{bmatrix} 0 & 1/2 & 0 \\ 0 & 1/2 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} P_{ff}^{(g)} \\ P_{Ff}^{(g)} \\ P_{FF}^{(g)} \end{bmatrix}$$

Appelons $\boldsymbol{O_F}$ la matrice de transition $\begin{bmatrix} 0 & 1/2 & 0 \\ 0 & 1/2 & 1 \\ 0 & 0 & 0 \end{bmatrix}$

VII.4.1.2 Composition génétique des graines récoltées sur les plantes femelles (autofécondation), pour le locus F

Les hermaphrodites portant le génotype *FF* peuvent uniquement produire des gamètes *F*. Par conséquent, leurs descendants par autofécondation ne peuvent que porter le génotype *FF*.

A l'inverse, la moitié des gamètes des hermaphrodites hétérozygotes (*Ff*) seront *f*, et l'autre moitié *F*. Par conséquent, un quart de leurs descendants seront *FF*, un quart *ff*, et la moitié *Ff*. Donc :

$$P_{ff,♂}^{(g+1)} = \frac{P_{Ff}^{(g)}}{4(P_{FF}^{(g)} + P_{Ff}^{(g)})}$$

$$P_{Ff,♂}^{(g+1)} = \frac{P_{Ff}^{(g)}}{2(P_{FF}^{(g)} + P_{Ff}^{(g)})}$$

$$P_{FF,♂}^{(g+1)} = (P_{FF}^{(g)} + \frac{P_{Ff}^{(g)}}{4})/(P_{FF}^{(g)} + P_{Ff}^{(g)})$$

Les fréquences génotypiques des graines récoltées sur les hermaphrodites doivent être diviser par $P_{FF}^{(g)} + P_{Ff}^{(g)}$ (la fréquence des hermaphrodites dans la population) car seuls-ci se reproduisent par autogamie.

Les formules peuvent s'écrire comme :

$$\begin{bmatrix} P_{ff,♂}^{(g+1)} \\ P_{Ff,♂}^{(g+1)} \\ P_{FF,♂}^{(g+1)} \end{bmatrix} = \frac{1}{P_{FF}^{(g)} + P_{Ff}^{(g)}} \begin{bmatrix} 0 & 1/4 & 0 \\ 0 & 1/2 & 0 \\ 0 & 1/4 & 1 \end{bmatrix} \begin{bmatrix} P_{ff}^{(g)} \\ P_{Ff}^{(g)} \\ P_{FF}^{(g)} \end{bmatrix}$$

Appelons $\boldsymbol{S_F}$ la matrice de transition $\begin{bmatrix} 0 & 1/4 & 0 \\ 0 & 1/2 & 0 \\ 0 & 1/4 & 1 \end{bmatrix}$

VII.4.1.3 Evolution des fréquences génotypique au locus F d'une génération à l'autre

Pour obtenir les fréquences populationnelles à la génération g+1, il faut écrire qu'une proportion α des graines a été récoltée sur les femelles, et une proportion 1-α sur les hermaphrodites. Par soucis de simplification, les migrants ne sont pas inclus dans le modèle.

$$\begin{bmatrix} P_{ff}^{(g+1)} \\ P_{Ff}^{(g+1)} \\ P_{FF}^{(g+1)} \end{bmatrix} = \alpha \begin{bmatrix} P_{ff,♀}^{(g+1)} \\ P_{Ff,♀}^{(g+1)} \\ P_{FF,♀}^{(g+1)} \end{bmatrix} + (1-\alpha) \begin{bmatrix} P_{ff,♂}^{(g+1)} \\ P_{Ff,♂}^{(g+1)} \\ P_{FF,♂}^{(g+1)} \end{bmatrix}$$

$$\begin{bmatrix} P_{ff}^{(g+1)} \\ P_{Ff}^{(g+1)} \\ P_{FF}^{(g+1)} \end{bmatrix} = \frac{\alpha}{P_{FF}^{(g)} + P_{Ff}^{(g)}} \begin{bmatrix} 0 & 1/2 & 0 \\ 0 & 1/2 & 1 \\ 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} P_{ff}^{(g)} \\ P_{Ff}^{(g)} \\ P_{FF}^{(g)} \end{bmatrix} + \frac{1-\alpha}{P_{FF}^{(g)} + P_{Ff}^{(g)}} \begin{bmatrix} 0 & 1/4 & 0 \\ 0 & 1/2 & 0 \\ 0 & 1/4 & 1 \end{bmatrix} \cdot \begin{bmatrix} P_{ff}^{(g)} \\ P_{Ff}^{(g)} \\ P_{FF}^{(g)} \end{bmatrix}$$

Donc :

$$\begin{bmatrix} P_{ff}^{(g+1)} \\ P_{Ff}^{(g+1)} \\ P_{FF}^{(g+1)} \end{bmatrix} = \frac{1}{P_{FF}^{(g)} + P_{Ff}^{(g)}} \begin{bmatrix} 0 & (2-\alpha)/2 & 0 \\ 0 & 1/2 & \alpha \\ 0 & (1-\alpha)/4 & 1-\alpha \end{bmatrix} \cdot \begin{bmatrix} P_{ff}^{(g)} \\ P_{Ff}^{(g)} \\ P_{FF}^{(g)} \end{bmatrix}$$

VII.4.1.4 Fréquences génotypiques au locus F à l'équilibre

Ces formules permettent de déterminer quelles seraient les fréquences génotypiques à l'équilibre. A l'équilibre, on trouve que :

$$\begin{bmatrix} P_{ff}^{(g+1)} \\ P_{Ff}^{(g+1)} \\ P_{FF}^{(g+1)} \end{bmatrix} = \begin{bmatrix} P_{ff}^{(g)} \\ P_{Ff}^{(g)} \\ P_{FF}^{(g)} \end{bmatrix}$$

Par conséquent, en posant les équations :

$$P_{ff}^{\infty} = \frac{(2-\alpha)P_{Ff}^{\infty}}{2(P_{FF}^{\infty} + P_{Ff}^{\infty})}$$

$$P_{Ff}^{\infty} = \frac{P_{Ff}^{\infty} + 2\alpha P_{FF}^{\infty}}{2(P_{FF}^{\infty} + P_{Ff}^{\infty})}$$

$$P_{FF}^{\infty} = \frac{(1-\alpha)P_{Ff}^{\infty} + 4(1-\alpha)P_{FF}^{\infty}}{4(P_{FF}^{(g)} + P_{Ff}^{(g)})}$$

On trouve que :

$$P_{ff}^{\infty} = \frac{\alpha}{2}$$

$$P_{Ff}^{\infty} = \frac{(2-\alpha)\alpha}{(\alpha+1)}$$

$$P_{FF}^{\infty} = \frac{(2-\alpha)(1-\alpha)}{2(\alpha+1)}$$

VII.4.2. Expression analytique de l'évolution des fréquences génotypiques à un locus A

Soit A un locus portant deux allèles : A et a. On suppose que ces allèles sont équitablement répartis chez les femelles et les hermaphrodites. Soient $P_A^{(g)}$ et $P_a^{(g)}$ les fréquences alléliques dans la population à la génération g. Soient $P_{AA}^{(g)}$, $P_{Aa}^{(g)}$ et $P_{aa}^{(g)}$ les fréquences des génotypes AA, Aa et aa dans la population à la génération g.

VII.4.2.1 Composition génétique des graines récoltées sur les plantes femelles (allofécondation), pour le locus A

Soient $P_{A,o}^{(g)}$ et $P_{a,o}^{(g)}$ les fréquences alléliques parmi les ovocytes des femelles à la génération g. Notons que $P_{A,o}^{(g)} = P_A^{(g)}$ et $P_{a,o}^{(g)} = P_a^{(g)}$.

$$P_{aa,♀}^{(g+1)} = {P_a^{(g)}}^2 = P_{a,o}^{(g)}\left(P_{aa}^{(g)} + \frac{P_{Aa}^{(g)}}{2}\right) = P_{a,o}^{(g)} \times P_{aa}^{(g)} + P_{a,o}^{(g)} \times \frac{P_{Aa}^{(g)}}{2}$$

$$P_{Aa,♀}^{(g+1)} = 2\,P_A^{(g)}\,P_a^{(g)} = P_{a,o}^{(g)}\left(P_{AA}^{(g)} + \frac{P_{Aa}^{(g)}}{2}\right) + P_{A,o}^{(g)}\left(P_{aa}^{(g)} + \frac{P_{Aa}^{(g)}}{2}\right)$$

$$= P_{aa}^{(g)} \times P_{A,o}^{(g)} + P_{AA}^{(g)} \times P_{a,o}^{(g)} + \frac{P_{Aa}^{(g)}}{2} \times (P_{a,o}^{(g)} + P_{A,o}^{(g)})$$

$$P_{AA,♀}^{(g+1)} = {P_A^{(g)}}^2 = P_{A,p}^{(g)}\left(P_{AA}^{(g)} + \frac{P_{A,o}^{(g)}}{2}\right) = P_{AA}^{(g)} \times P_{A,o}^{(g)} + P_{A,o}^{(g)} \times \frac{P_{Aa}^{(g)}}{2}$$

$$\begin{bmatrix} P_{aa,♀}^{(g+1)} \\ P_{Aa,♀}^{(g+1)} \\ P_{AA,♀}^{(g+1)} \end{bmatrix} = \begin{bmatrix} P_{a,o}^{(g)} & P_{a,o}^{(g)}/2 & 0 \\ P_{A,o}^{(g)} & 1/2 & P_{a,o}^{(g)} \\ 0 & P_{A,o}^{(g)}/2 & P_{A,o}^{(g)} \end{bmatrix} \begin{bmatrix} P_{aa}^{(g)} \\ P_{Aa}^{(g)} \\ P_{AA}^{(g)} \end{bmatrix}$$

Appelons $\boldsymbol{O_A}$ la matrice de transition $\begin{bmatrix} P_{a,o}^{(g)} & P_{a,o}^{(g)}/2 & 0 \\ P_{A,o}^{(g)} & 1/2 & P_{a,o}^{(g)} \\ 0 & P_{A,o}^{(g)}/2 & P_{A,o}^{(g)} \end{bmatrix}$

VII.4.2.2 Composition génétique des graines récoltées sur les plantes hermaphrodites (autofécondation), pour le locus A

Par autofécondation, les homozygotes donnent des individus génétiquement identiques à eux-mêmes. Les hétérozygotes, à l'inverse, donnent une moitié d'hétérozygotes, un quart de l'un des homozygotes, et un quart de l'autre homozygote. On trouve donc que :

$$\begin{bmatrix} P_{aa,♂}^{(g+1)} \\ P_{Aa,♂}^{(g+1)} \\ P_{AA,♂}^{(g+1)} \end{bmatrix} = \begin{bmatrix} 1 & 1/4 & 0 \\ 0 & 1/2 & 0 \\ 0 & 1/4 & 1 \end{bmatrix} \begin{bmatrix} P_{aa}^{(g)} \\ P_{Aa}^{(g)} \\ P_{AA}^{(g)} \end{bmatrix}$$

Appelons $\boldsymbol{S_A}$ la matrice de transition $\begin{bmatrix} 1 & 1/4 & 0 \\ 0 & 1/2 & 0 \\ 0 & 1/4 & 1 \end{bmatrix}$

VII.4.2.3 Evolution des fréquences génotypique au locus A d'une génération à l'autre

Comme précédent, pour trouver comment évoluent les fréquences à l'échelle de la population, il faut écrire qu'une proportion α provient d'allofécondation, et une proportion 1- α, d'autofécondation.

$$\begin{bmatrix} P_{aa}^{(g+1)} \\ P_{Aa}^{(g+1)} \\ P_{AA}^{(g+1)} \end{bmatrix} = \alpha \begin{bmatrix} P_{aa,♀}^{(g+1)} \\ P_{Aa,♀}^{(g+1)} \\ P_{AA,♀}^{(g+1)} \end{bmatrix} + (1-\alpha) \begin{bmatrix} P_{aa,♂}^{(g+1)} \\ P_{Aa,♂}^{(g+1)} \\ P_{AA,♂}^{(g+1)} \end{bmatrix}$$

$$\begin{bmatrix} P_{aa}^{(g+1)} \\ P_{Aa}^{(g+1)} \\ P_{AA}^{(g+1)} \end{bmatrix} = \left(\alpha \begin{bmatrix} P_{a,o}^{(g)} & P_{a,o}^{(g)}/2 & 0 \\ P_{A,o}^{(g)} & 1/2 & P_{a,o}^{(g)} \\ 0 & P_{A,o}^{(g)}/2 & P_{A,o}^{(g)} \end{bmatrix} + (1-\alpha) \begin{bmatrix} 1 & 1/4 & 0 \\ 0 & 1/2 & 0 \\ 0 & 1/4 & 1 \end{bmatrix} \right) \begin{bmatrix} P_{aa}^{(g)} \\ P_{Aa}^{(g)} \\ P_{AA}^{(g)} \end{bmatrix}$$

VII.4.3. Expression analytique de l'évolution des fréquences génotypiques pour les locus F et A

Dans cette partie, nous regarderons comment évoluent les fréquences des génotypes bilocus, incluant les deux locus A et F.

VII.4.3.1 Composition génétique des graines récoltées sur les plantes femelles (allofécondation), pour les locus F et A

Pour trouver la matrice de transition d'allofécondation pour les deux locus, $\boldsymbol{O}$, il suffit d'appliquer un produit de Kronecker :

$$\boldsymbol{O} = \frac{\boldsymbol{O}_F \otimes \boldsymbol{O}_A}{P_{FF}^{(g)} + P_{Ff}^{(g)}}$$

$$\boldsymbol{O} = \frac{1}{P_{FF}^{(g)} + P_{Ff}^{(g)}} \begin{bmatrix} 0 & 1/2 & 0 \\ 0 & 1/2 & 1 \\ 0 & 0 & 0 \end{bmatrix} \otimes \begin{bmatrix} P_{a,o}^{(g)} & P_{a,o}^{(g)}/2 & 0 \\ P_{A,o}^{(g)} & 1/2 & P_{a,o}^{(g)} \\ 0 & P_{A,o}^{(g)}/2 & P_{A,o}^{(g)} \end{bmatrix}$$

$$\begin{bmatrix} P_{ffaa,♀} \\ P_{ffaA,♀} \\ P_{ffAA,♀} \\ P_{fFaa,♀} \\ P_{fFaA,♀} \\ P_{FfAA,♀} \\ P_{FFaa,♀} \\ P_{FFaA,♀} \\ P_{FFAA,♀} \end{bmatrix}^{(g+1)} = \frac{1}{P_{FF}^{(g)} + P_{Ff}^{(g)}} \begin{bmatrix} 0 & 0 & 0 & P_{a,o}^{(g)}/2 & P_{a,o}^{(g)}/4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & P_{A,o}^{(g)}/2 & 1/4 & P_{a,o}^{(g)}/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & P_{A,p}^{(g)}/4 & P_{A,o}^{(g)}/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & P_{a,o}^{(g)}/2 & P_{a,o}^{(g)}/4 & 0 & P_{a,o}^{(g)} & P_{a,o}^{(g)}/2 & 0 \\ 0 & 0 & 0 & P_{A,o}^{(g)}/2 & 1/4 & P_{a,p}^{(g)}/2 & P_{a,o}^{(g)} & 1/2 & P_{a,o}^{(g)} \\ 0 & 0 & 0 & 0 & P_{A,o}^{(g)}/4 & P_{A,o}^{(g)}/2 & 0 & P_{A,o}^{(g)}/2 & P_{A,o}^{(g)} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} P_{ffaa} \\ P_{ffaA} \\ P_{ffAA} \\ P_{fFaa} \\ P_{fFaA} \\ P_{FfAA} \\ P_{FFaa} \\ P_{FFaA} \\ P_{FFAA} \end{bmatrix}^{(g)}$$

VII.4.3.2 Composition génétique des graines récoltées sur les plantes hermaphrodites (autofécondation), pour les locus F et A

Comme précédemment, pour trouver la matrice de transition d'autofécondation bilocus, un produit de Kronecker doit être appliqué entre les deux matrices unilocus :

$$\boldsymbol{S} = \frac{\boldsymbol{S}_F \otimes \boldsymbol{S}_A}{P_{FF}^{(g)} + P_{Ff}^{(g)}}$$

$$\boldsymbol{S} = \frac{1}{P_{FF}^{(g)} + P_{Ff}^{(g)}} \begin{bmatrix} 0 & 1/4 & 0 \\ 0 & 1/2 & 0 \\ 0 & 1/4 & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 1/4 & 0 \\ 0 & 1/2 & 0 \\ 0 & 1/4 & 1 \end{bmatrix}$$

$$\begin{bmatrix} P_{ffaa,♂} \\ P_{ffaA,♂} \\ P_{ffAA,♂} \\ P_{fFaa,♂} \\ P_{fFaA,♂} \\ P_{FfAA,♂} \\ P_{FFaa,♂} \\ P_{FFaA,♂} \\ P_{FFAA,♂} \end{bmatrix}^{(g+1)} = \frac{1}{P_{FF}^{(g)} + P_{Ff}^{(g)}} \begin{bmatrix} 0 & 0 & 0 & 1/4 & 1/16 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/16 & 1/4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/8 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 1/16 & 0 & 1 & 1/4 & 0 \\ 0 & 0 & 0 & 0 & 1/8 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 1/16 & 1/4 & 0 & 1/4 & 1 \end{bmatrix} \begin{bmatrix} P_{ffaa} \\ P_{ffaA} \\ P_{ffAA} \\ P_{fFaa} \\ P_{fFaA} \\ P_{FfAA} \\ P_{FFaa} \\ P_{FFaA} \\ P_{FFAA} \end{bmatrix}^{(g)}$$

VII.4.3.3 Evolution des fréquences génotypique aux locus A et F d'une génération à l'autre

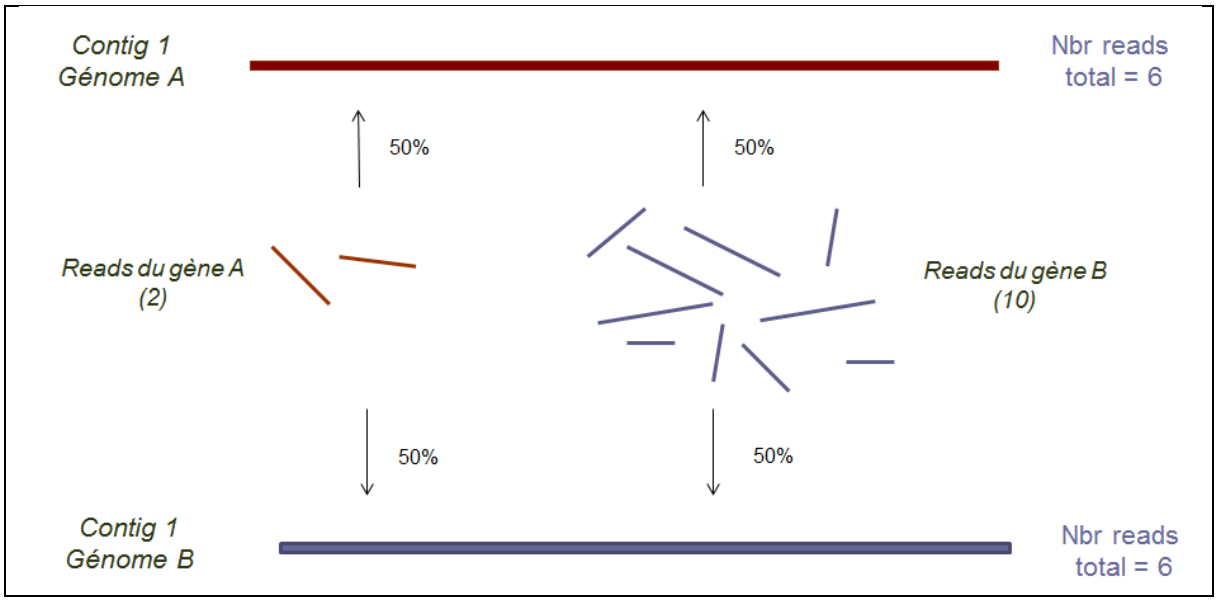
Pour former la génération suivante, on sèmera une proportion α de graines issues d'allofécondation avec les mâles stériles et une proportion (1 – α) de graine provenant d'autofécondation des mâles stériles. La matrice de transition pour passer d'une génération à l'autre devrait donc être égale à (1 – α)$\boldsymbol{O}$ + α$\boldsymbol{S}$.

$$\begin{bmatrix} P_{ffaa} \\ P_{ffaA} \\ P_{ffAA} \\ P_{fFaa} \\ P_{fFaA} \\ P_{FfAA} \\ P_{FFaa} \\ P_{FFaA} \\ P_{FFAA} \end{bmatrix}^{(g+1)} = \frac{\alpha\boldsymbol{O} + (1-\alpha)\boldsymbol{S}}{P_{FF}^{(g)} + P_{Ff}^{(g)}} \begin{bmatrix} P_{ffaa} \\ P_{ffaA} \\ P_{ffAA} \\ P_{fFaa} \\ P_{fFaA} \\ P_{FfAA} \\ P_{FFaa} \\ P_{FFaA} \\ P_{FFAA} \end{bmatrix}^{(g)}$$

VII.5. Principe du RPDS (Reads per Divergent Sites)

Lors du mapping sur des gènes homéologues, les reads se placent de manière aléatoire les zones non divergentes de la référence. Il n'est donc pas possible de discriminer les reads des zones non divergentes de deux homéologues.

Un exemple est donné dans la figure ci-dessous. Y sont figurées une zone non divergente entre deux gènes homéologues : en rouge, la zone du gène du génome A, en bleu, la zone du gène du génome B. Douze reads sont y représentés : deux provenant du gène du génome A, 10 du gène du génome B. Le gène B est donc plus exprimé que le gène A. Comme les séquences des deux gènes sont identiques dans cette zone, les reads se placent aléatoirement: 6 sur un gène et 6 sur l'autre.



En zone divergente, les reads se placeront sur le bon gène. Si deux gènes homéologues sont très similaires, et que leurs séquences sont identiques sur la plupart de leur longueur, alors l'éventuelle différence de niveau d'expression sera masquée. Il a donc été décidé que seuls les reads des zones divergentes seront pris en compte pour évaluer le niveau d'expression des gènes homéologues.

Pour cela, on a procédé ainsi :

1) Alignement des deux séquences homéologues de l'IWSCG (séquence de référence)
2) Détection des sites divergents
3) Comptage des reads sur les sites divergents <u>uniquement</u>
4) Calcul de la moyenne de ce comptage pour chaque contig : nombre de reads sur le nombre de nucléotides divergents
5) Ponderation par le nombre de reads total de l'individu, ce qui donne le RPDS (Reads per Divergent Sites)

VII.6. Fichiers supplémentaires de l'article *Reciprocal sign epistasis and truncation selection: when is recombination favorable in a pre-breeding program with a selfing species?*

VII.6.1. Supplementary File 1: Conditions to move from one configuration to another or to fix a genotype

VII.6.1.1. Proposition 1: Frequency differential between the two competitor genotypes

We tried to determine the value of the frequency differential between AB and ab in configurations 1 and 2.

Configuration 1: $p_{AB}^{n+1} - p_{ab}^{n+1} = \frac{p_{AB}^{n} - p_{ab}^{n}}{p} = \frac{p_{AB}^{0} - p_{ab}^{0}}{p^{n}}$ **(1)**

Indeed, $p_{AB}^{n+1} = \frac{p_{AB}^{n}}{p} - rD^{n}{}'$ and $p_{ab}^{n+1} = \frac{p_{ab}^{n}}{p} - rD^{n}{}'$ in configuration 1.

Configuration 2: $p_{AB}^{n+1} - p_{ab}^{n+1} = \frac{2p_{AB}^{n} - p}{p}$ **(2)**

Indeed, $p_{AB}^{n+1} = \frac{p_{AB}^{n}}{p} - rD^{n}{}'$ and $p_{ab}^{n+1} = \frac{p - p_{AB}^{n}}{p} - rD^{n}{}'$ in configuration 2.

VII.6.1.2. Proposition 2: Evolution of the frequency of the most efficient genotype *AB*

We sought to determine how p_{AB} evolved from one generation to another. Into both configurations 1 and 2, the difference $p_{AB}^{n+1} - p_{AB}^{n}$ is equal to:

$$p_{AB}^{n+1} - p_{AB}^{n} = \frac{p_{AB}^{n}}{p} - rD^{n}{}' - p_{AB}^{n} = \frac{p_{AB}^{n}(1-p)}{p} - rD^{n}{}' \quad \textbf{(3)}$$

The value of $D^{n}{}'$ changes depending on the configuration.

VII.6.1.3. Conditions for the evolution of the population

Propositions 1 and 2 are used to define conditions that determine the evolution of the population into the two first configurations:

Configuration 1

Given that $p_{AB}^{n} - p_{ab}^{n} \in [-1,1]$ and $\lim_{n\to\infty} \frac{1}{p^{n}} = \infty$. Therefore, we can deduce three conditions from equation **(1)**:

If $p_{AB}^{n} - p_{ab}^{n} > 0 \leftrightarrow p_{AB}^{n+1} - p_{ab}^{n+1} > 0$ **(4.a)**

If $p_{AB}^{n} - p_{ab}^{n} = 0 \leftrightarrow p_{AB}^{n+1} - p_{ab}^{n+1} = 0$ **(4.b)**

If $p_{AB}^{n} - p_{ab}^{n} < 0 \leftrightarrow p_{AB}^{n+1} - p_{ab}^{n+1} < 0$ **(4.c)**

In addition, from equation **(3)**, it can be deduced that the sign of $p_{AB}^{n+1} - p_{AB}^{n}$ will depend on these conditions:

If $rD^{n} p < p_{AB}^{n}(1-p) \leftrightarrow p_{AB}^{n+1} - p_{AB}^{n} > 0$ **(5.a)**

If $rD^{n} p = p_{AB}^{n}(1-p) \leftrightarrow p_{AB}^{n+1} - p_{AB}^{n} = 0$ **(5.b)**

If $rD^{n} p > p_{AB}^{n}(1-p) \leftrightarrow p_{AB}^{n+1} - p_{AB}^{n} < 0$ **(5.c)**

Configuration 2

Three conditions can be deduced from equation **(2)**:

If $p_{AB}^{n} > \frac{p}{2} \leftrightarrow p_{AB}^{n+1} - p_{ab}^{n+1} > 0$ **(6.a)**

If $p_{AB}^{n} = \frac{p}{2} \leftrightarrow p_{AB}^{n+1} - p_{ab}^{n+1} = 0$ **(6.b)**

If $p_{AB}^{n} < \frac{p}{2} \leftrightarrow p_{AB}^{n+1} - p_{ab}^{n+1} < 0$ **(6.c)**

Equation **(3)** can be written as $p_{AB}^{n+1} - p_{AB}^{n} = \frac{p_{AB}^{n}(1-p)}{p} - r\frac{p_{AB}^{n}(p - p_{AB}^{n})}{p^{2}}$ as $D^{n}{}' = \frac{p_{AB}^{n}(p - p_{AB}^{n})}{p^{2}}$ in configuration 2. We can deduce three conditions from it:

If $r < r_{c,2}(n) \leftrightarrow p_{AB}^{n+1} - p_{AB}^{n} > 0$ **(7.a)**

If $r = r_{c,2}(n) \leftrightarrow p_{AB}^{n+1} - p_{AB}^{n} = 0$ **(7.b)**

If $r > r_{c,2}(n) \leftrightarrow p_{AB}^{n+1} - p_{AB}^{n} < 0$ **(7.c)**

With $r_{c,2}(n) = \dfrac{(1-p)\cdot p}{p - p_{AB}^{n}}$

VII.6.1.4. Possible evolutionary paths from configuration 1

When the genotype AB is in higher frequency than ab: path 2 or 4

Equation **(4.a)** shows that if genotype AB is at higher frequency than ab at generation n, this will be the case at generation $n + 1$. Furthermore, the frequency difference between the two genotypes will continuously increase (see equation **(1)**). Thus:

$$p_{AB}^{n} - p_{ab}^{n} > 0 \rightarrow \lim_{n\to\infty} p_{AB}^{n} - p_{ab}^{n} = 1$$

So p_{AB}^{n} will increase and, after a certain number of generations, $p_{AB}^{n} + p_{ab}^{n} > p$. This will allow the population to leave configuration 1 to go to configuration 2, with genotype AB having an advantage over ab (path 4). Genotype AB can also go directly to configuration 4 without going through configuration 2 (path 2):

Conditions of path 2 (direct path to configuration 4 from configuration 1 – leading to the fixation of *AB*)

The population will be in configuration 4 at generation n+1 if p_{AB}^{n+1} is superior or equal to p. Then:

$$p_{AB}^{n+1} \geq p \leftrightarrow \frac{p_{AB}^{n}}{p} - rD^{n}{}' \geq p \leftrightarrow D^{n}{}' \leq \frac{p_{AB}^{n} - p^{2}}{p}$$

This may occur if:

- $D^n = 0$

- $p_{AB}^{n} - p^2 \geq 0$

or

- $D^n > 0$
- $r \leq \dfrac{p_{AB}^{n} - p^2}{pD^n}$

or

- $D^n < 0$
- $p_{AB}^{n} - p^2 \leq 0$
- $r \geq \dfrac{p_{AB}^{n} - p^2}{pD^n}$

When genotype AB is at the same frequency as ab: path 1 or 4

Equation **(4.b)** shows that if genotype AB is at the same frequency as ab at generation n, this will be the case at generation $n + 1$.

$$p_{AB}^{n} - p_{ab}^{n} = 0 \rightarrow \lim_{n\to\infty} p_{AB}^{n} - p_{ab}^{n} = 0$$

Genotype frequencies p_{AB} and p_{ab} will remain equal, but they could increase and reach a threshold. If this threshold is superior to $p/2$, the population will be able to attain configuration 2 (path 4). We determined the circumstances in which p_{AB}^{n} (= p_{ab}^{n}) will increase and take path 4.

In configuration 1, $D^n = \dfrac{{p_{AB}^{n}}^2}{p^2} - p_{aB}^{n} \cdot p_{Ab}^{n}\left(\dfrac{p - 2p_{AB}^{n}}{p \cdot (p_{aB}^{n} + p_{Ab}^{n})}\right)^2$

Note that $f(p_{aB}^{n}, p_{Ab}^{n}) = \dfrac{(p_{aB}^{n} + p_{Ab}^{n})^2}{p_{aB}^{n} \cdot p_{Ab}^{n}}$. The function $f(X,Y) = \dfrac{(X+Y)^2}{XY}$ accepts 4 as minimum on [0,1] x [0,1], obtained notably for $f(X,Y = X)$.

With this notation, $D^n = \dfrac{4p_{AB}^{n} - p}{fp} + \dfrac{{p_{AB}^{n}}^2(f-4)}{fp^2}$.

From equation **(5.a)**, it can be deduced that p_{AB} will increase if:

$$r(4p_{AB}^{n} - p + \frac{p_{AB}^{n}{}^{2}(f-4)}{p}) < fp_{AB}^{n}(1-p) \quad \textbf{(8)}$$

Note that $\frac{p_{AB}^{n}{}^{2}(f-4)}{p} \geq 0$, as $f \geq 4$.

Condition **(8)** is fulfilled if:

- $4p_{AB}^{n} - p + \frac{p_{AB}^{n}{}^{2}(f-4)}{p} > 0$
- $r < \frac{fp_{AB}^{n}(1-p)}{4p_{AB}^{n} - p + \frac{p_{AB}^{n}{}^{2}(f-4)}{p}} = r_{c,1}(n)$

or

- $4p_{AB}^{n} - p + \frac{p_{AB}^{n}{}^{2}(f-4)}{p} \leq 0$

$r_{c,1}(n)$ is a decreasing function of p_{AB}^{n}. The minimum value that can reach $r_{c,1}(n)$ is obtained for the maximum value of p_{AB}^{n} which is $p/2$ in configuration 1 when $p_{AB} = p_{ab}$. This defines a critical threshold value which is min $r_{c,1}(n) = f(1-p)/2$. If $p_{aB}^{n} = p_{Ab}^{n}$, this value equals $2(1-p)$. Note also that, due to recombination, frequencies of genotypes Ab and Ab are homogenized over time. Thus $\lim_{n\to\infty} p_{aB}^{n} - p_{Ab}^{n} = 0$ and $\lim_{n\to\infty} r_{c,1}(n) = 2(1-p)$. Therefore, considering the value of a threshold reached after many generations, $r_{c,1}(n)$ can be approximated by $2(1-p)$ in this case.

To conclude, when $p_{AB}^{n} - p_{ab}^{n} = 0$, the population will go into configuration 2 (path 4) if $r < f(1-p)/2$. Otherwise, it will stay in configuration 1 (path 1).

When genotype AB is at a lower frequency than ab: path 3 or 4

Equation **(4.c)** shows that if genotype AB is at a lower frequency than ab at generation n, this will be the case at generation n + 1. In addition, the frequency difference between the two genotypes will continuously increase (see equation **(1)**). Thus:

$$p_{AB}^{n} - p_{ab}^{n} < 0 \rightarrow \lim_{n\to\infty} p_{AB}^{n} - p_{ab}^{n} = -1$$

The genotype frequency p_{ab}^{n} will increase, so after a certain number of generations, $p_{AB}^{n} + p_{ab}^{n} > p$. This will allow the population to leave configuration 1 to go to configuration 2, with ab being advantaged (path 4). Genotype ab can also go directly to configuration 3 without going through configuration 2 (path 3)

Conditions of the path 3 (direct path to configuration 3 from configuration 1 – leading to the fixation of ab)

The population will take path 3 if:

- $p_{AB}^{n+1} = 0$ **(9)**
- $p_{ab}^{n+1} \geq p$ **(10)**

Expression **(9)** and expression **(3)** imply that $rD^{n}{}' = \frac{p_{AB}^{n}}{p}$. In addition, as $p_{ab}^{n+1} = \frac{p_{ab}^{n}}{p} - rD^{n}{}'$, expression **(10)** involves $rD^{n}{}' \leq \frac{p_{ab}^{n} - p^{2}}{p}$ **(11)**. We replace $rD^{n}{}'$ by $\frac{p_{AB}^{n}}{p}$ in **(10)** and we find that $p_{ab}^{n} - p_{AB}^{n} \geq p^{2}$.

Thus the population will take path 3 if $rD^{n}{}' = \frac{p_{AB}^{n}}{p}$ and if $p_{ab}^{n} - p_{AB}^{n} \geq p^{2}$. These two conditions are fulfilled for example if:

- $p_{AB}^{n} = 0$
- $r = 0$

- $p_{ab}^{n} \geq p^2$

VII.6.1.5. Possible evolutionary paths from configuration 2

The populations can follow different paths depending on the 6 conditions listed in section 4 (conditions 6.a, 6.b, 6.c, 7.a, 7.b and 7.c).

<table>
<tr><th>Conditions</th><th>7.a: $r < r_{c,2}(n)$</th><th>7.b: $r = r_{c,2}(n)$</th><th>7.c: $r > r_{c,2}(n)$</th></tr>
<tr><td>6.a: $p_{AB}^{n} > p/2$</td><td rowspan="3">The population goes to configuration 4 (Fixation of AB) (Path 7, paragraph i)</td><td colspan="2">The population goes to configuration 4 (Fixation of AB) with a return to configuration 1 (Path 7, paragraph ii)</td></tr>
<tr><td>6.b $p_{AB}^{n} = p/2$</td><td rowspan="2">The population stays in configuration 2 (Path 5, paragraph iii)</td><td>The population stays in configuration 1 (Path 6, paragraph iv)</td></tr>
<tr><td>6.c $p_{AB}^{n} < p/2$</td><td>The population goes to configuration 3 (Fixation of ab) (Path 8, paragraph v)</td></tr>
</table>

This will be demonstrated below:

(i) Configuration 4 and fixation of AB without a return to configuration 1: path 7

In this case, the population goes directly to configuration 4, leading to the fixation of AB. Condition **(7.a)** ($r < r_{c,2}(n)$) implies that p_{11} will continually increase until it is fixed (path 7). Indeed, the critical value $r_{c,2}(n) = \dfrac{(1-p)\cdot p}{p - p_{AB}^{n}}$ is an increasing function of p_{AB}. Thus if $p_{AB}^{n+1} > p_{AB}^{n}$, therefore $r_{c,2}(n) < r_{c,2}(n+1)$. If r is inferior to $r_{c,2}(n)$, it will be inferior to $r_{c,2}(n+1)$.

(ii) Configuration 4 and fixation of AB with a return to configuration 1: path 7

In this case, the population returns to configuration 1 before going to configuration 4. The conditions for this are:

$$p_{AB}^{n} > \frac{p}{2} \leftrightarrow p_{AB}^{n+1} - p_{ab}^{n+1} > 0 \quad \textbf{(6.a)}$$

$r \geq r_{c,2}(n) \leftrightarrow p_{AB}^{n+1} - p_{AB}^{n} \leq 0$ **(7.b)** and **(7.c)**

First note that simultaneously fulfilling condition **(6.a)** and condition **(7.b)** or **(7.c)** means that $p > 0.75$. Indeed, the conditions **(7.b)** and **(7.c)** ($r \geq r_{c,2}(n)$) are possible only if $r_{c,2}(n) \leq 1/2$. As $p_{AB}^{n} > p/2$ (condition **(6.a)**), $r_{c,2}(n) = \frac{(1-p)\cdot p}{p - p_{AB}^{n}} > 2(1-p)$. We find that the term $2(1 - p)$ is inferior to $1/2$ only if $p > 0.75$.

As $p_{AB}^{n} > p/2$ (condition **(6.a)**), $\frac{p}{2p_{AB}^{n}} < 1$, so $r \geq r_{c,2}(n) > r_{c,2}(n)\frac{p}{2p_{AB}^{n}}$

As $p_{AB}^{n+1} - p_{AB}^{n} = 0$ if $r = r_{c,2(n)}$, then $\frac{p_{AB}^{n}(1-p)}{p} - r_{c,2}(n)D^{n'} = 0$. Thus:

$$r > r_{c,2}(n)\frac{p}{2p_{AB}^{n}} = \frac{1-p}{2D^{n'}} \quad \textbf{(11)}$$

In addition, as $p_{AB}^{n+1} = \frac{p_{AB}^{n}}{p} - rD^{n'}$ and $p_{ab}^{n+1} = \frac{p - p_{AB}^{n}}{p} - rD^{n'}$, $p_{AB}^{n+1} + p_{ab}^{n+1} = 1 - 2rD^{n'}$ **(12)**. By equations **(11)** and **(12)**, we found that $p_{AB}^{n+1} + p_{ab}^{n+1} < p$.

This means that the population goes back into configuration 1 at generation n+1. But as $p_{AB}^{n+1} > p_{ab}^{n+1}$ (condition **(6.a)**), the population will come back to configuration 2 with $p_{AB} >> p_{ab}$. Thus this oscillation will just reinforce p_{11} at the expense of p_{00}.

Let n_2 denote the generation at which the population returns to configuration 2. The population will not come back a second time to configuration 1 if: $p_{AB}^{n_2+1} > p_{AB}^{n_2} \leftrightarrow p_{AB}^{n_2+1} - p_{AB}^{n_2} > 0$.

In addition, according to equation **(3)**:

$$p_{AB}^{n_2+1} - p_{AB}^{n_2} = \frac{p_{AB}^{n_2}}{p} - p_{AB}^{n_2} - r \cdot D^{n}' =$$

$$\frac{p_{AB}^{n_2}}{p} - p_{AB}^{n_2} - r \cdot p_{AB}^{n_2} \frac{(p - p_{AB}^{n_2})}{p^2} = \frac{p_{AB}^{n_2}}{p}(1 - p - r\frac{(p - p_{AB}^{n_2})}{p})$$

Consequently $p_{AB}^{n_2+1} > p_{AB}^{n_2}$ only if $1 - p > r\frac{(p - p_{AB}^{n_2})}{p}$, i.e. if $p_{AB}^{n_2}$ is close enough to p or if r is small enough. This condition is not guaranteed, but in the numerical calculations, 1 - p was always lower than $r\frac{(p - p_{AB}^{n_2})}{p}$: there was no more than one return to configuration 1. In any event, the return to configuration 1 only increases the frequency p_{AB} to the detriment of p_{ab}. So finally, p_{AB} will inevitably increase and genotype AB will finally be fixed (path 7).

(iii) The population stays in the configuration 2: path 5

The conditions are:

$p_{AB}^{n} \leq \frac{p}{2} \leftrightarrow p_{AB}^{n+1} - p_{ab}^{n+1} \leq 0$ **(6.b)** and **(6.c)**

$r = r_{c,2}(n) \leftrightarrow p_{AB}^{n+1} - p_{AB}^{n} = 0$ **(7.b)**

$r = r_{c,2}(n) = \frac{(1-p) \cdot p}{p - p_{AB}^{n}} = r_{c,2}(n+1)$ so p_{AB} does not evolve and stays equal to p_{AB}^{n}. In the same way p_{ab} is maintained at a constant value equal to $1 - p_{AB}^{n}(\frac{2}{p} - 1)$ from the generation n+1. Then $p_{ab} + p_{AB} = 1 - 2p_{AB}^{n}(\frac{1}{p} - 1)$ from generation n+1.

Note that $1 - 2p_{AB}^{n}(\frac{1}{p} - 1)$ is superior or equal to p because $p_{AB}^{n} \leq \frac{p}{2}$.

As $p_{AB} \leq \frac{p}{2}$ and $p_{ab} + p_{AB} \geq p$, the population stays in configuration 2 at any generation (path 5).

(iv) The population goes back to configuration 1 and stays there: path 6

The conditions are:

$$p_{AB}^{n} = \frac{p}{2} \leftrightarrow p_{AB}^{n+1} - p_{ab}^{n+1} = 0 \quad \textbf{(6.b)}$$

$$r > r_{c,2}(n) \leftrightarrow p_{AB}^{n+1} - p_{AB}^{n} < 0 \quad \textbf{(7.c)}$$

In these conditions, $p_{AB}^{n+1} = p_{ab}^{n+1} < p/2$. The population goes back to configuration 1 at generation n+1, with genotypes AB and ab at the same frequency. We find that:

$$p_{AB}^{n+1} = \frac{p_{AB}^{n}}{p} - rD^{n\prime} = \frac{p_{AB}^{n}}{p} - r\frac{p_{AB}^{n}}{p}\frac{p - p_{AB}^{n}}{p} = \frac{1}{2} - \frac{r}{4},$$

$$p_{ab}^{n+1} = \frac{p - p_{AB}^{n}}{p} - rD^{n\prime} = \frac{p - p_{AB}^{n}}{p} - r\frac{p_{AB}^{n}}{p}\frac{p - p_{AB}^{n}}{p} = \frac{1}{2} - \frac{r}{4}$$ and

$$p_{aB}^{n+1} = p_{Ab}^{n+1} = rD^{n\prime} = r\frac{p_{AB}^{n}}{p}\frac{p - p_{AB}^{n}}{p} = \frac{r}{4}$$

The critical value $r_{c,1}(n+1)$ is the value above which the population stays trapped in configuration 1 (see above). It is equal to:

$$r_{c,1}(n+1) = \frac{f \cdot p_{AB}^{n+1}(1-p) \cdot p}{4p_{AB}^{n+1} \cdot p - p^2 + p_{AB}^{n+1\,2}(f-4)}$$ with $f \geq 4$ (see above, at paragraph4.ii)

At generation n+1, $p_{aB}^{n+1} = p_{Ab}^{n+1}$, because only individuals ab and AB were selected at generation n, as the population was in configuration 2, so $f = 4$. Thus $r_{c,1}(n+1) = \dfrac{4p_{AB}^{n+1}(1-p)}{4p_{AB}^{n+1} - p}$ **(14)**

By replacing p_{AB}^{n+1} by $\frac{1}{2} - \frac{r}{4}$ in **(14)**, an inequation on r can be solved. We find that if $r \in [2(1 - p), 1]$, $r_{c,1}(n+1)$ is inferior to r. As $p_{AB}^{n} = \frac{p}{2}$, then $r_{c,2}(n) = \dfrac{(1-p)\cdot p}{p - p_{AB}^{n}} = 2(1-p)$. Thus, as $r > r_{c,2}(n) = 2(1-p)$ (condition **(7.c)**), $r > r_{c,1}(n+1)$.

Therefore, in this case, when the population falls into configuration 1, it is in conditions such that p_{AB} continues to decline, and stays equal to p_{ab}. The population remains trapped in configuration 1 (path 6).

(v) Configuration 3 and fixation of ab: path 8

The conditions are:

$$p_{AB}^{n} < \frac{p}{2} \leftrightarrow p_{AB}^{n+1} - p_{ab}^{n+1} < 0$$

$$r > r_{c,2}(n) \leftrightarrow p_{AB}^{n+1} - p_{AB}^{n} < 0$$

This involves a decrease in p_{AB}, which remains inferior to p_{ab}. We demonstrated that p_{00} increases:

First $p_{ab}^{n+1} - p_{ab}^{n} > 0$ if:

$r < p\dfrac{p\cdot(1-p_{ab}^{n}) - p_{AB}^{n}}{p_{AB}^{n}\cdot(p-p_{AB}^{n})} = r_{c,3}(n)$, by simply writing that

$$p_{ab}^{n+1} = \frac{p-p_{AB}^{n}}{p} - r\cdot D^{n}{}' = \frac{p-p_{AB}^{n}}{p} - r\cdot\frac{p_{AB}^{n}}{p}\cdot\frac{p-p_{AB}^{n}}{p}.$$

Note that $r_{c,3}(n) = r_{c,2}(n)\dfrac{p\cdot(1-p_{ab}^{n}) - p_{AB}^{n}}{p_{AB}^{n}\cdot(1-p)}$

- If $\dfrac{p\cdot(1-p_{ab}^{n}) - p_{AB}^{n}}{p_{AB}^{n}\cdot(1-p)} > 1 \leftrightarrow r_{c,2}(n) < r < r_{c,3}(n)$

  In this case, p_{AB} decreases and p_{ab} increases.

- If $r > r_{c,2}(n)$ and $r > r_{c,3}(n)$

  In this case, both p_{AB} and p_{ab} decrease. However, $r_{c,3}(n)$ is a decreasing function of p_{AB}^{n} and p_{ab}^{n}. This means that this critical value will increase with time. The recombination rate r will thus eventually become inferior to $r_{c,3}(n)$, and p_{ab} will finally increase.

In both cases, genotype ab will be fixed (path 8).

VII.6.2. Supplementary File 2: Results of simulations in finite populations

Table 1: results of simulations testing the recombination rate *r*

Initial state 1		r = 0.18			r = 0.19			r = 0.20			r = 0.21			r = 0.22		
	Fixation of	ab	AB	NF	ab	AB	NF	ab	AB	NF	ab	AB	NF	ab	AB	NF
	N = 1000	28%	56%	16%	42%	46%	12%	49%	47%	4%	42%	53%	5%	57%	43%	0%
	N = 5000	2%	66%	33%	10%	53%	37%	23%	49%	28%	35%	55%	10%	45%	50%	5%
Initial state 2 $p_{ab}^0 = 0.8$		r = 0.09			r = 0.095			r = 0.1			r = 0.105			r = 0.11		
	Fixation of	ab	AB	NF	ab	AB	NF	ab	AB	NF	ab	AB	NF	ab	AB	NF
	N = 1000	97%	1%	3%	97%	0%	3%	97%	0%	3%	97%	0%	3%	97%	0%	3%
	N = 5000	95%	0%	5%	94%	0%	6%	95%	0%	5%	98%	0%	2%	98%	0%	2%
Initial state 2 $p_{ab}^0 = 0.1$		r = 0.09			r = 0.095			r = 0.1			r = 0.105			r = 0.11		
	Fixation of	ab	AB	NF	ab	AB	NF	ab	AB	NF	ab	AB	NF	ab	AB	NF
	N = 1000	0%	98%	2%	1%	97%	2%	1%	94%	5%	0%	95%	5%	1%	85%	14%
	N = 5000	0%	100%	0%	0%	100%	0%	0%	100%	0%	0%	100%	0%	0%	100%	0%
		r = 0.12			r = 0.13			r = 0.14			r = 0.15			r = 0.16		
	Fixation of	ab	AB	NF	ab	AB	NF	ab	AB	NF	ab	AB	NF	ab	AB	NF
	N = 1000	2%	76%	22%	1%	54%	45%	18%	23%	59%	42%	12%	46%	70%	2%	28%
	N = 5000	0%	92%	8%	0%	37%	63%	0%	7%	93%	4%	0%	96%	34%	0%	66%

r = recombination rate; N = demographic population size; the percentage of selected individuals p is fixed at 0.9.

NF = no fixation; number of repetitions: 100; number of generations: 100

Initial state 1: $p_{ab}^0 = p_{AB}^0 = 0$. In the simulations $p_{aB}^0 = 0.6$ and $p_{Ab}^0 = 0.4$

Initial state 2: $p_{AB}^0 = 0$ and $p_{aB}^0 = p_{Ab}^0$

Table 2: Results of simulations testing the percentage of selected individuals p

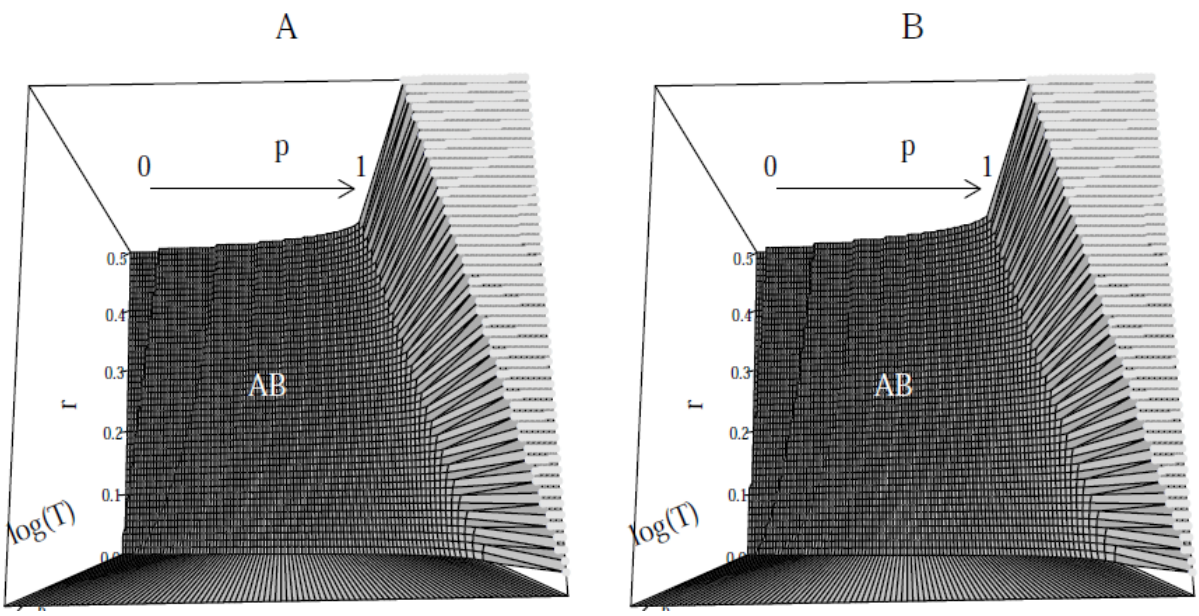
		$p = 0.72$			$p = 0.76$			$p = 0.8$			$p = 0.84$			$p = 0.88$		
Initial state 2 $p^0_{ab} = 0.8$	Expectation	Fixation of ab			Fixation of ab			No fixation			Fixation of AB			Fixation of AB		
	Fixation of	ab	AB	NF	ab	AB	NF	ab	AB	NF	ab	AB	NF	ab	AB	NF
	N = 1000	100%	0%	0%	100%	0%	0%	100%	0%	0%	100%	0%	0%	100%	0%	0%
	N = 5000	100%	0%	0%	100%	0%	0%	100%	0%	0%	100%	0%	0%	94%	0%	6%
		$p = 0.09$			$p = 0.095$			$p = 0.1$			$p = 0.105$			$p = 0.11$		
Initial state 2 $p^0_{ab} = 0.1$	Expectation	Fixation of ab			Fixation of ab			No fixation			Fixation of AB			Fixation of AB		
	Fixation of	ab	AB	NF	ab	AB	NF	ab	AB	NF	ab	AB	NF	ab	AB	NF
	N = 1000	100%	0%	0%	100%	0%	0%	100%	0%	0%	89%	11%	0%	81%	19%	0%
	N = 5000	100%	0%	0%	99%	1%	0%	100%	0%	0%	100%	0%	0%	84%	16%	0%

p = percentage of selected individuals; N = demographic population size; the recombination rate r is fixed at 0.1.

NF = no fixation; number of repetitions: 100; number of generations: 100

Initial state 2: $p^0_{AB} = 0$ and $p^0_{aB} = p^0_{Ab}$

VII.6.3. Supplementary File 3: Initial State 1, when *aB* is not at the same frequency as *Ab*



Logarithm of the time to fixation in initial state 1 ($p^0_{ab} = p^0_{AB} = 0$) for different initial genotypic frequencies

A: $p^0_{aB} = 0.3$ and $p^0_{Ab} = 0.7$

B: $p^0_{aB} = 0.1$ and $p^0_{Ab} = 0.9$

Black grid area: fixation of AB; light grey points: no fixation (infinite time to fixation); rest of the graph: fixation of ab.

r = recombination rate; p = percentage of selected individuals; T = time to fixation

VII.7. Commentaires des reviewers sur l'article *Reciprocal sign epistasis and truncation selection: when is recombination favorable in a pre-breeding program with a selfing species?*

Dear Ms Vagne,

Your paper has been reviewed. I regret to say that it is not considered suitable for publication in Heredity. The Subject Editor has outlined the reasons for this decision below.

I wish to thank you for the opportunity to review your work. I hope that the referees' comments benefit the future presentation of this work to a different journal, and that this decision will not deter you from submitting future work to Heredity.

Yours sincerely,

Professor Michael Bruford

Editor-in-Chief, Heredity

School of Biosciences

Cardiff University

Cathays Park, Cardiff

CF10 3AX, Wales, UK

Visit us at http://www.nature.com/hdy

Submit at: http://www.mts-hdy.nature.com

Subscribe at: http://www.nature.com/hdy/subscribe

_____

Editor (Comments to the Author):

All three reviewers found that the manuscript is hard to read. There are two many different cases, configurations and initials and readers can get lost easily. In addition, the work is purely theoretical and its application to quantitative genetics or population genetics in the genome era is not well addressed.

Referee #1 (Remarks to the Author):

1. Manuscript Summary

The authors explored the relationship between reciprocal sign epistasis, truncation selection, and recombination. This work generalizes previous studies by allowing the genotype fitness values to vary over time and by considering an extensive range for initial allelic frequencies.

2. Review Summary

I do not believe the paper should be published in its present form. The core of this work certainly has academic value, and the mathematics presented appear to be correct. However, the authors should revaluate the scope of this work. I would also suggest a more accessible presentation of the results.

3. Scope of the Manuscript

Overall, the authors offer a thorough treatment of infinite populations. Their model has novel features, and the authors strengthen the validity of their conclusions by showing consistency with finite simulations. This is the publishable portion of the manuscript.

Unfortunately, the manuscript goes way beyond the evidence and starts speculating about practical breeding programs. I do not see the relevant evidence in a model that fails to account for the heritability or the magnitude of the epistatic effect. It's very hard to imagine that these parameters would not influence a population's convergence to a final configuration. This work is best discussed for what it is: a theoretical study of an analytical model.

Furthermore, as the authors themselves clearly state, the design confounds outcrossing and recombination. Are you sure these results are strictly attributable to recombination? Couldn't selfing impact another population parameter? If so, this would distort your estimates of the impact of recombination. The paper could use a more formal treatment of this issue. Alternatively, you could recast the work as a study of the impact of outcrossing instead of recombination, which seems like a more accurate description of this experiment.

Finally, I'd like to see a clearer comparison with previous work. For example, I was left with the impression that these results were entirely consistent with previous work. Does this mean that varying genotype fitness values had no effect of the impact of recombination? That seems like an interesting conclusion, and I would encourage you to explore that and present the result explicitly.

Referee #2 (Remarks to the Author):

The title of the manuscript suggests that the main goal of the manuscript from Constance Vagne et al. is to examine when recombination is favorable in a pre-breeding program of selfing species considering reciprocal sign epistasis and truncation selection. The authors implemented therefore a theoretical framework based on a haploid bi-locus model and used this to identify the time of fixation of particular genotypes in an infinite population selected for a trait which is determined by reciprocal sign epistasis. In addition, stochastic simulations were performed to complement the results of the infinite populations.

I am not an expert in population genetics. From a non-expert view point the introduction is not convincing to show that the performed research is novel enough filling relevant gaps. This observation made when reading the introduction is supported by statements in the discussion of the type "our results confirmed" (e.g., page 28 line 562-565; page 329 line 572). Therefore, I doubt that the novelty of the research presented in manuscript warrants publication in Heredity, but I asked the editor to take into consideration that the heritability of my evaluation of the novelty is only moderate.

My expertise lies in the area of quantitative genetics in relation with prebreeding programs. Therefore, I evaluated the manuscript mainly based on the plausibility of the derived conclusions for prebreeding.

First of all, it is not clear to me what type of population (parents are only genetic resources or also elite lines, how many genetic resources would you use) and prebreeding design the authors had in mind. In addition, it is not clear what trait the authors consider as representative trait for their scenarios? Are the authors assuming trait with major QTL? The statement on page 4 line 76 - 78 suggests this at least:

"The term 'pre-breeding' refers to the transfer of genes from related wild ancestors or from ancient varieties to breeding material (FAO, 1996)". If the target is to transfer major genes into breeding material the assumptions of the investigated scenarios are not relevant, because backcross strategies are most efficient.

In contrast, if the targets of prebreeding are complex traits such as grain yield, I wonder how phenotypic values are entering in your developed scenarios and influence all the conclusions made in the discussion with regard to prebreeding? On page 30 line 608-610 you state: "When linkage disequilibrium is negative, recombination is favorable for the fixation of AB in the first generations because it allows the creation of the most favorable genotype". Creation of the AB genotype is important but in the full discussion you ignored that it is also relevant to identify the AB genotype in prebreeding. Is phenotyping in prebreeding conducted on individual plants or on families? Using individual plants makes no sense for complex traits such as grain yield, because the heritability of the evaluation at level of individual plants is extremely low. My guess is that phenotyping has then to be done using families. How is such a family structure considered in the existing framework? How relevant is epistasis in families used to evaluate the performance of heterozygous genotypes? I would think that epistasis contributes only little to the genetic variation for most common types of families (especially those which can easily be developed in selfing species with a dominant male sterility gene).

The authors state on page 31 that "In order to determine a monitoring strategy based on selection and the outcrossing rate for a pre-breeding population, the genetic architecture of the phenotypic trait that the breeder wishes to improve could be analyzed, notably to assess the extent and type of epistasis. Next-generation sequencing technologies (Metzker, 2010) could be valuable tools for this purpose". How would you implement this in the context of prebreeding? Based on which population? Epistasis is depended on the population under consideration. Is it really in range that we can disentangle the genetic architecture for agronomic traits up to the level of epistasis? To cite the conclusion of a recent paper on the genetic architecture of maize height based on > 7300!! inbreds (Peiffer et al. Genetics 2014): "In conclusion, the molecular mechanisms of natural PHT (plant height) variation remain difficult to disentangle and it seems unlikely that the rapid systematic elucidation of a causal model for the genetic architecture of this variation is possible". I would greatly appreciate if such findings are taken into account.

Summarizing, I would recommend rejecting the paper.

Referee #3 (Remarks to the Author):

The authors investigate the consequences of sign epistasis for the effect of recombination on the efficiency of selection in a simple haploid two-locus model. The study is motivated by previous work that addressed this question in the context of standard mutation-selection models, where fitness is assigned to genotypes in a time-independent fashion. The main outcome of these studies is the existence of a critical recombination rate beyond which selection becomes extremely inefficient, in the sense that the fixation of the most fit genotype may be completely suppressed in infinite populations. Here the authors consider this scenario for a truncation selection scheme, where a fixed fraction of individuals (those with the largest values of the desired trait) is transferred in each generation. This is an interesting and nontrivial variation of the standard problem which is of obvious relevance to breeding programmes. The analysis is sound and the presentation is generally clear. I recommend publication after the authors have addressed the remarks made above concerning the 'Results' section and Figure 2, and additionally have fixed the following minor issues:

line 349: The path number is missing in this sentence.

lines 368 and 376: What does it mean that "the population transiently goes back to configuration 1."? Is this a single excursion, after which AB or ab are fixed, and how many generations does it last?

line 511: The fixation rate of AB decreases from 23% to 2%, not the other way around.

line 520: I believe the recombination rate in this figure is 0.1, not 0.01.

Rôle des effets épistatiques dans l'évolution d'une population à régime de reproduction mixte, et dans la régulation de l'expression des gènes homéologues chez une espèce autogame allotétraploïde, le blé dur (*Triticum turgidum*)

Résumé : L'épistasie a longtemps été négligée, mais aujourd'hui elle connait un regain d'intérêt avec le développement du séquençage haut-débit et des études d'association pangénomique (GWAS). Elle est quasi-omniprésente, ayant été détectée sur de nombreux traits et chez de nombreuses espèces. Elle joue un rôle important d'un point de vue évolutif, puisqu'elle est à l'origine d'incompatibilités de Dobzhansky-Müller, expliquant l'isolement reproductif entre lignées apparentées. Par ailleurs, chez les polyploïdes à héritabilité disomique (essentiellement des allopolyploïdes), elle peut permettre de fixer l'hétérosis. L'objectif de cette thèse a été d'étudier l'épistasie dans ce contexte évolutif. Le matériel d'étude a été une population à base génétique large d'une espèce autogame allopolyploïde, le blé dur (*Triticum turgidum*), possédant deux génomes : le génome A et le génome B. Cette population a été constituée à partir de différents taxons : blé dur (*T. t. durum*), amidonnier domestique (*T. t. dicoccum*) et amidonnier sauvage (*T. t. dicoccoides*). Il est donc possible qu'elle présente des combinaisons de gènes coadaptées, et donc de l'épistasie générant des incompatibilités de Dobzhansky-Müller.

Une première partie présente différents modèles permettant d'estimer l'épistasie. Ces modèles ont des inconvénients et des avantages, qui dépendent de l'objectif que l'on s'est fixé. Dans une deuxième partie, nous avons utilisé l'un de ces modèles sur des données de RNA-seq (données d'expression et de génotypage) de la population de blé dur, afin de détecter des effets régulateurs de l'expression, et notamment de l'épistasie entre gènes homéologues. Cette analyse est une première pour le blé dur et probablement pour les espèces polyploïdes. Nous n'avons presque pas détecté d'épistasie de premier ordre, mais des interactions avec le fond génétique ont été observées. Nous avons trouvé plus d'eQTL sur le génome B que sur le génome A, ce qui est peut-être dû au fait que les ancêtres diploïdes du blé dur n'avaient pas le même régime de reproduction. Par ailleurs, nous avons montré que les niveaux d'expression des gènes homéologues étaient très fréquemment (80%) corrélés de manière positive, ce qui indique que les gènes homéologues appartiennent fréquemment aux mêmes complexes de gènes corégulés. Enfin, les régulations *trans* sont plus fréquentes au sein des paires de gènes d'homéologues, ce qui corrobore également l'idée que les gènes homéologues appartiennent aux mêmes réseaux de gènes. Dans une troisième partie, le rôle évolutif de l'épistasie a été examiné. D'abord, nous avons montré comment l'épistasie influençait la covariance parents-enfants, et donc l'évolution à court terme d'une population. Ensuite, nous avons étudié l'impact d'un type particulier d'épistasie (celui à l'origine des d'incompatibilités de Dobzhansky-Müller, présentant des combinaisons de gènes coadaptés) sur l'évolution d'une population soumise à de la sélection par troncation (comme dans les populations de pre-breeding), à l'aide d'un modèle haploïde bilocus, en fonction du taux de recombinaison et de l'intensité de la sélection. Nous avons montré que si le génotype optimal n'est pas présent dans la population initiale, un fort taux de recombinaison risque de mener à la fixation du génotype sous-optimal. Néanmoins, un peu de recombinaison est nécessaire pour créer ce génotype optimal et le fixer.

En perspectives, nous proposons d'adopter une méthode basée sur les réseaux de gènes pour approfondir les mécanismes de la régulation de l'expression chez le blé dur. Nous proposons également de complexifier le modèle permettant d'étudier l'effet de la recombinaison sur l'évolution d'un trait épistatique, sélectionné par troncation, et de compléter ces travaux de modélisation par des études expérimentales.

Mots-clés : épistasie, autogamie, polyploïdie, paysage adaptatif, pre-breeding, recombinaison, régime de reproduction, incompatibilité de Dobzhansky-Muller, dépression hybride, eQTL, GWAS

Role of epistatic effects in the evolution of a population with mixed mating system, and in the regulation of the expression of homoelogous genes in an allotetraploid selfing species, durum wheat (*Triticum turgidum*)

Abstract: Epistasis has long been neglected, but it is currently subject to a renewed interest with the development of high-throughput sequencing technologies and of genome-wide association studies (GWAS). It is virtually ubiquitous, having been detected on many traits and in many species. It plays an important evolutionary role, since it is one of the sources of Dobzhansky-Muller incompatibilities, explaining reproductive isolation between related lineages. Moreover, it can enable to fix heterosis in disomic polyploids (mostly allopolyploids). The objective of this thesis was to study epistasis in this evolutionary context. The study material was a broad genetic basis population of a self-pollinating allopolyploid species, durum wheat (*Triticum turgidum*), having two genomes: A and B genomes. This population was composed from different taxa: durum wheat (*T. t. durum*), domestic emmer (*T. t. dicoccum*) and wild emmer (*T. t. dicoccoides*). Therefore, there are probably combinations of coadapted genes in this population, and thus epistasis generating Dobzhansky-Muller incompatibilities.

The first part presents different models to estimate epistasis. These models have advantages and drawbacks, depending on the overall objective. In a second part, we used one of these models on RNA-seq data (expression and genotyping data) of the population of durum wheat to detect regulatory effects of expression, including epistasis between homoelogous genes. This analysis is unique in durum wheat, and probably in polyploid species. We did not detect first-order epistasis, but interactions with genetic background were observed. We found more eQTL on the B genome than on the A genome, which may be due to the fact that the diploid progenitors of durum wheat did not have the same mating system. Furthermore, we showed that the expression levels of homoelogous genes were frequently (80%) positively correlated, indicating that homoelogous genes often belong to the same complex of coregulated genes. Finally, *trans* regulations are more common among pairs of homoelogous genes, which also supports the idea that homoelogous genes belong to the same gene networks. In the third part, the evolving role of epistasis was examined. First, we showed how epistasis influences the parent-offspring covariance, and thus the short-term evolution of a population. Then we studied the impact of a particular type of epistasis (the one which can generate Dobzhansky-Muller incompatibilities, with combinations of co-adapted genes) on the evolution of a population subjected to truncation selection (as in pre-breeding populations), using a haploid bilocus model, based on the recombination rate and selection intensity. We have shown that if the optimal genotype is not present in the initial population, a high rate of recombination may lead to the fixation of the sub-optimal genotype. However, some recombination is needed to create this optimal genotype and fix it.

In perspective, we propose to adopt a network-based method to further the regulation mechanisms of expression in durum wheat. We also propose to make the model more complex, in order to study the effect of recombination on the evolution of an epistatic trait selected by truncation, and complete these modelling works by experimental studies.

Keywords: epistasis, selfing, polyploidy, adaptive landscape, pre-breeding- recombination, mating system, Dobzhansky-Muller incompatibilities, hybrid depression, eQTL, GWAS