Évaluation de méthodes et optimisation de protocoles de GWAS à locus unique et multi-locus pour l'identification de régions génomiques contrôlant des caractères d'intérêt chez le pois (Pisum sativum L.)

Mamadou Sene

▶ **To cite this version:**

Mamadou Sene. Évaluation de méthodes et optimisation de protocoles de GWAS à locus unique et multi-locus pour l'identification de régions génomiques contrôlant des caractères d'intérêt chez le pois (Pisum sativum L.). Sciences du Vivant [q-bio]. 2022. dumas-03977199

HAL Id: dumas-03977199

https://dumas.ccsd.cnrs.fr/dumas-03977199v1

Submitted on 7 Feb 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

L'Institut Agro Rennes-Angers

☐ **Site d'Angers** ☒ **Site de Rennes**

Année universitaire : 2021-2022 **Master** : Biologie, Agrosciences **Parcours** : Amélioration, Production, Valorisation du Végétal (**APVV**) **Option** : Génétique, Génomique et Amélioration des Plantes (**GGAP**)	**Mémoire de fin d'études** ☐d'ingénieur de l'Institut Agro Rennes-Angers (Institut national d'enseignement supérieur pour l'agriculture, l'alimentation et l'environnement) ☒de master de l'Institut Agro Rennes-Angers (Institut na+ tional d'enseignement supérieur pour l'agriculture, l'alimentation et l'environnement) ☐de l'Institut Agro Montpellier (étudiant arrivé en M2) ☐d'un autre établissement (étudiant arrivé en M2)

Evaluation de méthodes et optimisation de protocoles de GWAS à locus unique et multi-locus pour l'identification de régions génomiques contrôlant des caractères d'intérêt chez le pois (*Pisum sativum* L.)

Par : Mamadou SENE

Soutenu à Rennes le 24/06/2022

Devant le jury composé de :

Président : Mélanie JUBAULT

Maîtres de stage : Nadim TAYEH et Jonathan KREPLAK

Enseignant référent : Mélanie JUBAULT

Examinatrice : Maria MANZANARES-DAULEUX

Rapportrice : Marie-Laure NAYEL

Les analyses et les conclusions de ce travail d'étudiant n'engagent que la responsabilité de son auteur et non celle de l'Institut Agro Rennes-Angers et l'université de Rennes 1

Ce document est soumis aux conditions d'utilisation « Paternité-Pas d'Utilisation Commerciale-Pas de Modification 4.0 France » disponible en ligne http://creativecommons.org/licenses/by-nc-nd/4.0/deed.fr

Sommaire

Remerciements

Je tiens à adresser de grands remerciements à mes deux encadrants, Nadim TAYEH et Jonathan KREPLAK, de m'avoir donné la confiance et l'opportunité de travailler dans ce projet qui fut pour moi une expérience professionnelle et personnelle remarquable. Merci d'avoir suivi de près ce stage et d'avoir mis à ma disposition tous les moyens nécessaires pour sa réussite et son bon déroulement. Et enfin, merci pour votre écoute, votre patience, votre disponibilité, les réunions de mise au point, vos suggestions et les efforts fournis pour corriger ce mémoire.

Je remercie chaleureusement la responsable du pôle GEASPI, Judith BURSTIN, pour l'accueil qu'elle m'a réservé et pour ses conseils.

Un grand merci à mes collègues de bureau, Baptiste IMBERT (Doctorant en bio-informatique) et Codé DIOL (CDD en bioinformatique), pour votre accueil, votre sympathie, nos échanges intéressants et de m'avoir souvent débloqué lors de mes analyses.

Merci à Grégoire AUBERT pour votre aide dans ce travail et les corrections et suggestions apportées à ce travail.

Merci à Sandie BARBOT pour sa sympathie et à mon collègue stagiaire Virgilio FREITAS pour les discussions intéressantes autour de nos différentes thématiques de stage.

Merci à tout le personnel de l'équipe ECP, notamment Anthony KLEIN, ainsi que tous les autres pour leur accueil, leur sympathie et leur aide.

Merci à Mélanie JUBAULT, ma tutrice universitaire et responsable du parcours GGAP pour les conseils et le suivi de mon stage. A travers elle, je remercie l'ensemble du corps professoral du Master APVV.

Sigles et abréviations

ACM : et l'analyse des correspondances multiples
ACP : analyse en composantes principales
AFDM : Analyse Factorielle des Données Mixtes
BLINK (*Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway*)
***CMLM** : compressed mixed linear model*
***ECMLM** : enriched compressed mixed linear model*
***EM** : expectation and maximization*
***EMMA** : efficient mixed-model association*
***EMMAX** : efficient mixed-model association expedited*
FarmCPU (*Fixed and random model Circulating Probability Unification*)
***FaST-LMM** : factored spectrally transformed linear mixed model*
***FASTmrEMMA** : fast multi-locus random-SNP-effect EMMA.*
***FASTmrMLM** : a fast mrMLM multilocus mixed linear model.*
***FDR** : false discovery rate*
FEM:Fixed Effect Model
GAPIT: *Genomic Association and Prediction Integrated Tool*
GBC : *genotyping by capture* (Génotypage par capture)
***GEMMA** : genome-wide efficient mixed model analysis*
***GLM** : general linear model*
***GWAS** : genome-wide association studies*
***ISIS EM-BLASSO** : Iterative modified-Sure Independence Screening EM-Bayesian LASSO*
***LD** : linkage disequilibrium*
***MLM** : mixed linear model*
***MLMM** : multi-locus mixed model*
mrMLM :*multi-locus random-SNP-effect MLM*)
***mrMLM** : multi-locus random-SNP-effect mixed linear model*
***P3D** : population parameters previously determined*
***ACP** :* Analyse de composantes principales
PC : *Principal Component*
pKWmEB : *integration of Kruskal-Wallis test with empirical Bayes*)
pLARmEB : *polygenic-background-control-based least angle regression plus empirical Bayes*)
***QTL**: Quantitative trait locus*
***QTN** : quantitative trait nucleotide*
***REML** : restricted maximum likelihood*
***SNP**: Single nucleotide polymorphism*
***SUPER** : settlement of MLM under progressively exclusive relationship*
***UPGMA** : unweighted pair group method with arithmetic average*

Liste des figures

Liste des tableaux

Liste des annexes

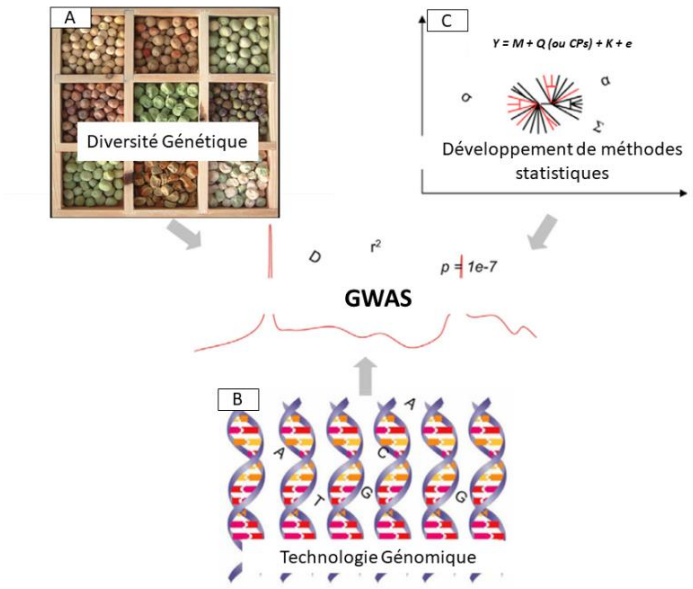


Figure 1 : Schématisation de l'étude d'association à l'échelle du génome (Source : modifié de Zhu et al., 2008)
La réalisation d'une étude de génétique d'association nécessite trois composantes : (**A**) une population de large diversité génétique appelée panel, phénotypée pour un ou des caractères d'intérêt ; (**B**) des données de génotypage avec des marqueurs moléculaires pour l'ensemble des individus de la population et (**C**) une méthode statistique permettant de relier les variations phénotypiques aux variations génotypiques afin d'identifier les loci causaux.

1. INTRODUCTION GENERALE

Les légumineuses sont l'un des piliers pour une agriculture agroécologique, en partie par leur symbiose avec les bactéries du sol fixatrices d'azote, ce qui réduit le besoin d'appliquer des engrais azotés, et donc permet d'atténuer les émissions de gaz à effet de serre (Nemecek et al., 2008 ; Crews et al., 2004). Elles sont aussi de précieuses sources de protéines végétales pour l'alimentation humaine et animale. Pour développer la culture des légumineuses, et en particulier des légumineuses à graines, il faudrait qu'elles puissent offrir une production régulière en étant plus robustes face aux aléas climatiques et plus résistantes face aux bioagresseurs. Pendant 8 ans (2012-2020), le projet public-privé PeaMUST (*Pea MultiStress Tolerance*), piloté par l'institut national de recherche pour l'agriculture, l'alimentation et l'environnement (INRAE) et financé par le programme « Investissements d'avenir », a permis de caractériser une large collection de diversité génétique de pois. De nombreux traits ont été évalués et mesurés à la fois dans des contextes pédoclimatiques différents et aussi en conditions contrôlées. Ces ressources phénotypiques ont été complétées avec des ressources génomiques issues du séquençage et de l'annotation du génome du pois ainsi que du génotypage haute densité des accessions de la collection de diversité. C'est dans ce contexte de recherche que s'insère mon stage, qui s'est déroulé pendant six mois, au niveau de l'équipe Espèces Cibles Protéagineux (ECP) de l'unité mixte de recherche (UMR) Agroécologie à l'INRAE de Dijon, qui était impliquée dans ce projet. Son objectif général était d'optimiser la détection des déterminants génétiques contrôlant différents traits d'intérêt chez le pois.

1.1. Etat de l'art

1.1.1. Génétique d'association chez les plantes

Chez les plantes, on peut retrouver des traits qualitatifs dont la variation est discontinue (couleurs des feuilles ou des fleurs, réponses à certaines maladies…). Ces traits sont généralement contrôlés par un faible nombre de gènes (traits mono- ou oligo-géniques) et faiblement influencés par l'environnement. D'autres traits sont des traits quantitatifs, notamment importants en agronomie (rendement, poids des grains, hauteur, etc.), souvent complexes, contrôlés par de nombreux gènes (polygéniques) et influencés par l'environnement (Galais, 2018 ; Samouelian et al., 2009). Relier le génotype aux traits permet d'analyser l'architecture génétique de ces derniers, de localiser les régions génomiques impliquées dans les variations phénotypiques, de comprendre les mécanismes biologiques sous-jacents et *in fine*, d'aider à la sélection des allèles d'intérêt (Mackay et al., 2009 ; Mackay, 2001). L'étude d'association pangénomique ou GWAS (*Genome Wide Association Study*) est une approche couramment utilisée pour disséquer les bases génétiques des traits complexes en s'appuyant sur la diversité génétique naturelle existante et sur des recombinaisons historiques (Korte & Farlow, 2013 ; Risch & Merikangas, 1996 ; Figure 1). Au-delà de l'espèce modèle *Arabidopsis thaliana*, la GWAS a permis de révéler l'architecture génétique de traits importants chez diverses espèces d'intérêt économique (Kaler et al., 2020 ; Gupta et al., 2019 ; Liu and Yan, 2019 ; Sukumaran and Yu, 2014 ; Ersoz et al., 2007).

Dans la pratique, afin d'identifier des marqueurs génétiques associés aux traits phénotypiques, la GWAS utilise des modèles statistiques comme illustré dans la Figure 1.

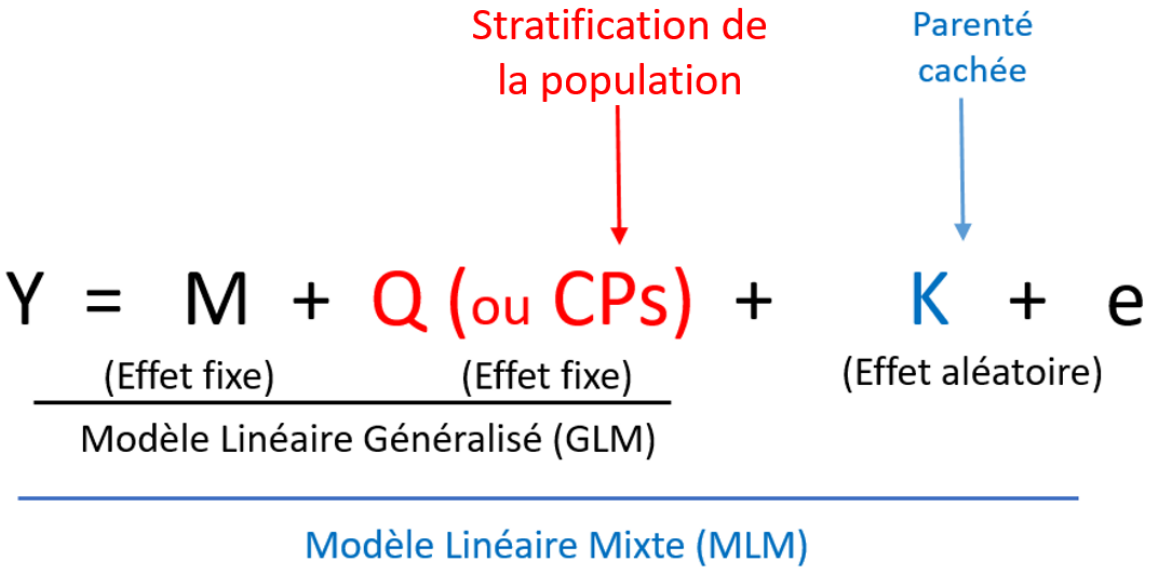


Figure 2 : Modèle linéaire généralisé (GLM) et modèle linéaire mixte (MLM) : deux modèles statistiques communément utilisés en GWAS
Y, Phénotype ou trait d'étude ; **M**, représente les marqueurs génétiques testés ; **Q** ou **CP**, représentent l'effet de la structure ; **K** (*kinship*), représente la parenté génétique entre les individus et **e**, l'effet résiduel. Pour contrôler la structure de la population, les modèles GLM et MLM prennent en compte l'effet des SNP et de la stratification de la population qu'ils considèrent comme fixes, en plus de l'effet résiduel qui est considéré comme aléatoire. A la différence de la GLM, le MLM ajoute un effet de parenté génétique entre individus (K), qu'il considère comme aléatoire, afin de mieux contrôler la covariance.

1.1.1.1. Modèles statistiques utilisés en GWAS

- ✓ **Modèles à locus unique**

Au niveau le plus simple, la GWAS peut être accomplie en effectuant un simple modèle naïf (test de t ou une analyse de variance) évaluant chaque marqueur séparément des autres pour sa contribution au phénotype (Bush and Moore, 2012). Les méthodes du modèle naïf ayant un fort taux de faux positifs, des méthodes basées sur un Modèle Linéaire Généralisé (GLM) ont été proposées (Jombart et al. 2010 ; Purcell et al., 2007 ; Price et al., 2006 ; Pritchard et al., 2000 ; Devlin and Roeder, 1999). Dans ces méthodes GLM, un seul marqueur génétique est aussi évalué, mais la stratification de la population est prise en compte comme cofacteur à effet fixe dans le modèle, grâce à une matrice (Q) représentant la proportion d'individus appartenant à un sous-groupe (Figure 2). L'utilisation du modèle linéaire mixte (MLM) (Yu et al., 2006) qui incorpore simultanément dans un modèle la stratification de la population (Q) comme effet fixe et la parenté génétique entre les individus (K) comme effet aléatoire (Figure 2) a ensuite permis d'éliminer encore plus efficacement les faux positifs. Depuis le premier modèle MLM proposé par Yu et al. (2006), de nombreuses méthodes statistiques avancées basées sur MLM ont été développées pour résoudre certaines limitations du modèle MLM telles que les besoins informatiques importants et la puissance statistique. EMMA (Kang et al., 2008), GEMMA (Zhou et Stephens, 2012), EMMAX et 3PD (Kang et al., 2010) ont été proposées pour minimiser la charge de calcul présentée dans les fonctions de probabilité du MLM. Elles considèrent l'effet des nucléotides de traits quantitatifs (QTN) comme un effet fixe. CMLM (Zhang et al., 2010), ECMLM (Li et al., 2014), FaST-LMM (Lippert et al., 2011), Fast-LMM select (Listgarten et al., 2012) et SUPER (Wang et al., 2014) ont été proposées pour contrôler les grandes matrices de génotypage en regroupant les individus en groupes. Ainsi, la matrice de parenté K est dérivée des individus regroupés.

Globalement, les méthodes MLM à locus unique présentent des limites pour résoudre les effets potentiels causés par plusieurs tests, et les effets pléiotropes (Wen et al., 2018 ; Buzdugan et al., 2016). Aussi, les interactions entre les variantes génétiques disponibles dans tout le génome ne sont pas explorées en profondeur lorsqu'un seul marqueur est testé à la fois.

- ✓ **Modèle multilocus**

Les méthodes multilocus sont une alternative pour résoudre les limites des méthodes à locus unique et permettent de prendre en compte la structure polygénique des traits complexes (Segura et al., 2012). De façon générale, ces méthodes multilocus impliquent des algorithmes en deux étapes, consistant en un balayage à un seul locus de l'ensemble du génome pour détecter tous les QTN possibles, puis en un test de tous les marqueurs associés à l'aide d'un modèle multilocus pour détecter les vrais QTN (Berhe et al., 2021 ; Liu et al., 2016 ; Wang et al., 2016 ; Segura et al., 2012). L'approche multilocus a été d'abord mise en œuvre dans le modèle MLMM (Segura et al., 2012), ensuite, d'autres méthodes s'appuyant sur MLMM et permettant d'améliorer l'efficacité de calcul et la puissance statistique ont été développées : FarmCPU (Liu et al., 2016) et BLINK (Huang et al., 2019). Egalement dans l'optique d'améliorer la puissance statistique de la GWAS et la prédiction de l'effet des QTN, six autres méthodes multilocus basées sur une approche bayésienne et considérant l'effet des marqueurs comme aléatoire dans le modèle ont été plus récemment développées : mrMLM ; pLARmEB (Zhang et al., 2017) ; ISIS EM-BLASSO (Tamba et al., 2017) ; FASTmrMLM (Tamba and Zhang, 2018) ; FASTmrEMMA (Wen et al., 2018) et pKWmEB (Wen et al., 2018).

Tableau 1 : Etudes comparatives de plusieurs méthodes de GWAS à locus unique et multilocus sur des plantes modèles et cultivées

Espèce	Données phénotypique et génotypiques				Modèle GWAS			Résultats		Source
	Taille panel	Caractère étudié	Nombre de traits mesurés	Nombre de SNP	Type de modèle	Nombre de méthodes	Méthodes	Meilleure methode	Approche recommandée	
Arabidopsis	199	Résistance aux maladies, développement, teneur en ion, période floraison	107	250000	UL	3	naïf, GLM et MLM	FarmCPU	ML	Lui et al., 2016
					ML	1	FarmCPU			
	188	Période de floraison	6	216130	UL	2	rMLM, EMMA	mrMLM	ML	Wang et al., 2016
					ML	1	mrMLM			
	199	Période de floraison	4	216130	UL	4	EMMA, CMLM, ECMLM, SUPER,	FASTmrEMMA	ML	Wen et al., 2018
					ML	2	E_Bayes, FASTmrEMMA			
Maïs	144	Capacité de régénération des cals embryonnaires	5	43427	UL	1	NA	ISIS EM BLASSO	ML	Ma et al., 2018
					ML	4	*package mrMLM			
	230	Propriétés collantes de l'amidon	7	145232	UL	1	GEMMA	FASTmrEMMA	Intégrée	Xu et al., 2018
					ML	3	FASTmrEMMA, FarmCPU			
Coton	160	Qualité de la fibre	6	72792	UL	1	MLM, EMMAX	NA	Intégrée	Su et al., 2018
					ML	6	package mrMLM			
	169	qualité de la fibre	5	53 848	UL	3	GLM, MLM, CMLM	ML	Intégrée	Li et al., 2018
					ML	3	mrMLM, FASTmrEMMA, ISIS EM-BLASSO			
Soja	200	Résistance aux nématodes	1	33194	UL	NA	GLM, CMLM et ECMLM	ECMLM, FarmCPU	Intégrée	Zhao et al., 2017
					ML	6	FarmCPU			
	219	Réponse à la photosynthèse		292035	UL	NA	NA	FASTmrEMMA	ML	Lu et al., 2018
					ML	6	Package mrMLM			
	368	Hauteur de la plante et nombre de nœuds sur la tige principale	6	62423	UL	1	MLM, CMLM	mrMLM	ML	Chang et al., 2018
					ML	1	mrMLM			
	346	Divers traits avec des héritabilités variables	6	42509	UL	6	ANOVA, GLM, MLM, CMLM, ECMLM, SUPER	FarmCPU	ML	Kaler et al., 2020
					ML	2	MLMM, FarmCPU			
Blé	182	Niveau d'acides Aminés libres	20	14646	UL	1	NA	pKWmEB	Intégrée	Peng et al., 2018
					ML	6	Package mrMLM			
Riz	478	Tolérance au sel	5	165529	NA	6	NA	ISI EM BLOSSO	Intégrée	Cui et al., 2018
					ML		mrMLM			
	529	Composantes de rendement	5	607201	UL	1	MLMclassique	ML	ML	Zhong et al., 2021
					ML	5	Package mrMLM			
	478	Tolérance au sel	5	165529	NA	NA	NA	ISI EM BLOSSO	Intégrée	Cui et al., 2018
					ML	6	Package mrMLM			
	191	Agronomique et teneur en ion	29	3 200 000	UL	2	GLM et MLM	FarmCPU	Intégrée	Liu et al., 2020
					ML	2	MLMM et FarmCPU			

*Package mrMLM : contient les six méthodes multilocus suivantes : mrMLM, ISIS EM-BLASSO, pLARmEB, FASTmrMLM, FASTmrEMMA et pKWmEB ; UL : méthodes GWAS à locus unique ; ML : méthodes GWAS avec mutlilocus ; NA : données numériques non trouvées dans nos recherches, ou approche non étudiée dans l'étude correspondante.

1.1.1.2. Impact de la méthode de GWAS sur la détection de locus significatifs

Différentes études comparatives récentes ont été menées pour évaluer la capacité de ces différentes méthodes de GWAS à détecter des associations marqueurs-traits chez différentes espèces végétales (Tableau 1). Dans ces études, il a été globalement constaté que, les méthodes multi-locus étaient plus efficaces et plus puissantes que les méthodes à locus unique pour détecter des résultats d'association hautement significatifs pour les traits d'intérêt. Néanmoins, il n'est probablement pas possible d'identifier une seule meilleure méthode GWAS pour toutes les situations étant donné la complexité biologique inhérente aux populations d'étude et à la nature du trait considéré. Chaque méthode de GWAS fournit un outil pour découvrir les associations qui peuvent être manquées par d'autres méthodes en fonction de l'architecture génétique du trait étudié et de la structure de la population d'étude (Cortes et al., 2021). Des études ont montré que, des marqueurs significatifs détectés par une méthode et validés expérimentalement comme biologiquement pertinents ne sont pas du tout détectés par d'autres méthodes (Kaler et al., 2020 ; Liu et al., 2020 ; Li et al., 2018 ; Yang et al., 2018 ; Liu et al., 2016b). On s'attend à ce que ces différences de résultats se produisent en raison de différences dans les détails des méthodes statistiques utilisées. Dans de nombreux cas, les forces connues de diverses méthodes GWAS dans les traits avec des architectures génétiques diverses et des populations avec des structures différentes expliquent les divergences (Cortes et al., 2021). Cependant, il a été prouvé que l'intégration de méthodes à locus unique et multilocus améliore la puissance et la validité de la GWAS pour des traits complexes (Berhe et al., 2021 ; Liu et al., 2020 ; Abed and Belzile, 2019 ; Zhang et al., 2019 ; Cui et al., 2018 ; Peng et al., 2018 ; Li et al., 2018 ; Su et al., 2018 ; Xu et al., 2018 ; Zhao et al., 2017). De plus, on peut potentiellement supposer que les QTL détectés simultanément avec plusieurs approches (à locus unique et multilocus) soient très susceptibles de capturer de véritables associations qui seront utiles pour la recherche et l'amélioration des plantes (Abed and Belzile, 2019 ; Zhang et al., 2019).

1.1.2. La recherche en génétique au service de l'amélioration génétique du pois

1.1.2.1. Importance économique du pois

Le pois (*Pisum sativum* L.) est la deuxième légumineuse à graine la plus importante au monde après le haricot commun (*Phaseolus vulgaris* L.) (faostat, 2022). Il est cultivé dans plus de 100 pays répartis dans les cinq continents. En 2020, plus de 24 millions d'hectares de pois ont été cultivés dans le monde, avec une production de plus de 34 millions de tonnes pour l'alimentation humaine et de plus de 8 millions de tonnes pour l'alimentation animale, dont des rendements en grains moyens respectifs de 2036.4 kg/ha et 591.2 kg/ha (faosat, 2022). La valeur économique du pois provient principalement de l'importance des services écosystémiques qu'il fournit. Le pois constitue une source précieuse de protéines alimentaires, de nutriments minéraux, d'amidon complexe, de vitamines et de fibres avec des bienfaits démontrés pour la santé (Foschia et al., 2017 ; Smýkal et al., 2012 ; Burstin et al., 2011). Sa symbiose avec les bactéries du sol fixatrices d'azote réduit le besoin d'appliquer des engrais azotés, et donc permet d'atténuer les émissions de gaz à effet de serre (Nemecek et al., 2008 ; Crews et al., 2004).

1.1.2.2. Diversité phénotypique et génétique du pois

Dans le genre *Pisum* on distingue particulièrement, l'espèce cultivée *P. sativum ssp. sativum* et l'espèce sauvage *P. fulvum*. Au sein de l'espèce *P. sativum*, nous avons des sous espèces sauvages à l'exemple de *eleatus* (Kreplak et al., 2019). Les accessions de pois diffèrent considérablement en termes de morphologie, de potentiel de rendement, de durée de floraison et de taille des grains (Rana et al., 2017 ; Ouafi et al., 2016, Warkentin et al., 2015) et de réponse aux stress. Ces stress sont liés à des facteurs climatiques comme le gel (Beiji et al., 2019), à des maladies comme la pourriture racinaire ou l'oïdium (Sulima et Zhukov, 2022), ou à des ravageurs comme les pucerons (Ollivier et al., 2022) et la bruche (Aznar-Fernández et al., 2020). Des études de diversité sur des collections de divers accessions de pois (variétés locales ; cultivars de pois potagers, de grande culture ou fourragers ; ainsi que des pois sauvages), utilisant différents types de marqueurs moléculaires ont montré une importante diversité génétique chez le pois (Siol et al., 2017 ; Burstin et al., 2015 ; Jing et al., 2012 ; Tar'an et al., 2005). Cette diversité existante peut être utilisée pour améliorer l'espèce cultivée à travers des croisements dirigés, y compris des introgressions de parents sauvages, vu que les barrières reproductives ne sont pas strictes parmi les espèces et sous-espèces de *Pisum* (Gali et al., 2019 ; Kreplak et al., 2019).

1.1.2.3. Génomique du pois et apport des nouveaux outils moléculaires dans les études génétiques

Le pois est autogame et son génome haploïde est constitué de sept chromosomes appariés (2 n = 2 x = 14). Sa taille haploïde est estimée à 4,45 Gb (Pandey et al., 2021 ; Kreplak et al., 2019). Le premier génome de référence du pois a été assemblé en 2019. Cette première version du génome est basée sur le cultivar "Cameor" et s'étend sur 3,92 Gb (88 % de la taille estimée du génome du pois), avec 3,23 Gb (82,5 %) de séquences attribuées aux sept pseudo-molécules et 14.266 scaffolds représentant 685 Mo (Kreplak et al., 2019).

L'émergence de technologies de séquençage de nouvelle génération (NGS) à haut débit à un prix abordable a fait de l'utilisation des marqueurs moléculaires et en particulier des SNP (*single nucleotide polymorphism*) à l'échelle du génome une solution idéale pour les études de diversité génétique et l'estimation du déséquilibre de liaison (DL) dans de nombreuses cultures, y compris le pois (Gali et al., 2019 ; Elshire et al., 2011). Chez le pois, des marqueurs SNP ont été développés et utilisés dans des études de diversité génétique (Siol et al., 2017 ; Burstin et al., 2015 ; Diapari et al., 2015) et de cartographie génétique (Sindhu et al., 2014 ; Tayeh et al. , 2015). Ces marqueurs SNP à l'échelle du génome ont été utilisés pour développer des puces à SNP, pour le génotypage à haut débit de génotypes de pois et la cartographie des populations (Tayeh et al., 2015 ; Sindhu et al., 2014). L'assemblage de la première version du génome du pois (Kreplak et al. 2019), et l'utilisation de la technologie de génotypage par capture (GBC) ont permis d'améliorer la précision de la GWAS chez le pois (Ollivier et al., 2021). La méthode de génotypage par capture d'exome, permet le séquençage ciblé de régions du génome codant pour des protéines ("exomes"), afin de réduire les coûts tout en enrichissant la découverte de variants intéressants (Ng et al., 2009).

1.1.3. Etudes GWAS chez le pois

Chez le pois, la GWAS a été d'ores et déjà utilisée pour révéler les régions génomiques contrôlant plusieurs caractères, notamment la résistance à *Aphanomyces euteiches*, agent de la pourriture racinaire, (Desgroux et al., 2018 ; Desgroux et al., 2016), la résistance aux pucerons (Ollivier et al., 2022), la tolérance au gel (Beji et al., 2020), l'architecture racinaire (Desgroux et al., 2018), la résistance à la verse, le rendement et ses composantes et la qualité des graines (Gali et al., 2019) et les concentrations en macroéléments des graines (Cartelier, 2021 ; Dissanayaka et al., 2020). Dans ces études GWAS sur le pois, les méthodes MLM ; Fast-LMM et MLMM ont été particulièrement utilisées (Ollivier et al., 2022 ; Beji et al., 2020 ; Dissanayaka et al., 2020 ; Desgroux et al., 2018 ; Desgroux et al., 2018).

1.2. Contexte et stratégies développées au cours du stage

Les études GWAS réalisées sur le pois ne sont pas assez nombreuses, mais augmentent depuis la publication de la première version du génome du pois. La majorité de ces études a été menée avec un nombre relativement faible de SNP (< 30.000), ce qui peut constituer une insuffisance pour révéler pleinement les bases génétiques des traits quantitatifs. De plus, à ma connaissance, parmi ces études, aucune n'a utilisé les méthodes multi-locus à effet SNP aléatoire basées sur une approche Bayésienne (mrMLM) et certaines autres nouvelles méthodes multi-locus (MLMLM). Également, une minorité de ces études a comparé ces différents types de méthodes multi-locus avec les méthodes à locus unique (ULMLM). En plus, avec l'évolution de la version d'assemblage du génome de pois et la diversité des méthodes statistiques nouvellement disponibles, il devient important de se saisir de l'ensemble des données pour les traiter ensemble et fournir des connaissances complètes et précises aux chercheurs et aux semenciers.

Dans ce contexte, dans les travaux de mon stage, à l'aide d'un panel de 240 individus génotypés avec un ensemble de plus de 1.900.000 SNP issus de GBC, j'ai réussi à atteindre trois des quatre objectifs fixés au départ. Dans un premier temps, pour permettre d'explorer au mieux les bases génétiques des traits clés du pois, j'ai évalué et comparé les performances de neuf méthodes de GWAS réparties au sein de trois approches identifiées dans la littérature (ULMLM, MLMLM et mrMLM), avant de comparer ensuite ces trois approches entre elles. Dans un second temps, pour évaluer l'impact de la qualité de l'assemblage du génome sur les résultats de GWAS, j'ai comparé les résultats de GWAS entre la première version du génome du pois déjà publiée et une nouvelle version améliorée disponible en interne, mais pas encore publiée. Enfin, pour recommander des outils nécessaires pour l'introgression des QTL dans les programmes de sélection, j'ai fait une synthèse de tous mes résultats et je les ai comparés avec les connaissances déjà disponibles. Le quatrième objectif, qui consistait à faire de la GWAS multitrait, n'a pas été réalisé dans les délais pour être intégré dans le manuscrit, mais sera poursuivi dans la période de stage restante.

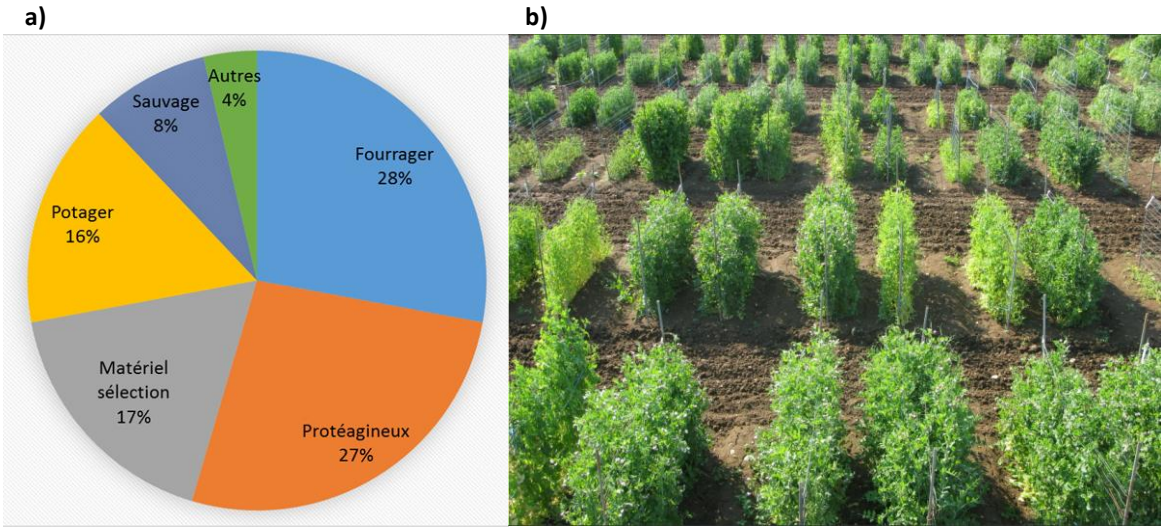


Figure 3 : Aperçu de la diversité au sein de la collection AMS (source : Anthony Klein, communication réunion finale PeaMUST WP3).
a) Part de chacun des différents constituants dans le panel de 240 génotypes, **b**) photo prise au champ qui illustre une diversité du panel en termes de vigueur, de précocité de floraison, de hauteur et d'intensité de coloration du couvert.

2. MATERIEL ET METHODES

2.1. Matériel végétal

Le matériel végétal utilisé est un panel de 240 génotypes de pois, appelé collection AMS (Architecture et Multi-Stress), construit dans le cadre du projet PeaMUST (Burstin et al., 2021). Les 240 génotypes sont d'origines et d'usages divers (Figure 3 ; Ollivier et al., 2022) et ont été sélectionnés pour représenter la diversité phénotypique des espèces et sous espèces de pois (cultivées et sauvages) pour l'architecture aérienne et racinaire ainsi que les principales réponses aux stress biotiques (aphanomyces et ascochytose) et abiotiques (essentiellement gel). La part des sauvages était faible dans le panel (20 dont trois *P. fulvum*). Des données de diversité moléculaire obtenues préalablement (Siol et al., 2017) ont été également utilisées pour maximiser la diversité.

2.2. Données phénotypiques disponibles

Les données phénotypiques analysées ont été acquises dans le cadre du projet PeaMUST (Burstin et al., 2021), avec des expérimentations multisites (en chambre climatique et dans cinq lieux en France : Bretenière, Le Rheu, Mauguio, Mons et Orsonville) et multi-années (2015, 2016, 2017 et 2018). La matrice complète contient 41 différents traits, dont certains ont été mesurés/observés dans différents sites et/ou sur différentes années, faisant au total 63 variables (Annexe 1). Les traits sont en lien avec l'architecture, la phénologie, la résistance aux stress, le rendement et la qualité des grains. Le nom des variables ainsi que la description brève des méthodes de phénotypage, des sites et des années d'expérimentation sont résumés dans Annexe 1.

2.3. Analyse statistique des données phénotypiques

Des statistiques simples par variable incluant le nombre d'observations réalisées, la valeur minimale, la valeur maximale, la moyenne et le coefficient de variation (cv) ont été calculées à partir de la matrice d'origine en utilisant le logiciel *R* (RStudio Team, 2022). Puis, l'héritabilité au sens étroit, définie comme le rapport de la variance génétique additive à la variance phénotypique (h^2), a été calculée pour chaque variable à l'aide de la version 3 du package *GAPIT (Genomic Association and Prediction Integrated Tool ;* Wang et Zhang, 2021), en se basant sur les informations fournies par les marqueurs moléculaires.

Les corrélations entre les variables ont été établies à l'aide d'une Analyse Factorielle des Données Mixtes (AFDM), combinant à la fois l'analyse en composantes principales (ACP) et l'analyse des correspondances multiples (ACM), compte tenu de l'existence à la fois de variables quantitatives et qualitatives (Pagès, 2004). L'AFMD a été réalisée avec les package *R* : *FactoMineR* (Lê et al., 2008) et *factoextra* (Kassambara et Mundt, 2020).

2.4. Données de génotypage

2.4.1. Données génotypiques basées sur la première version du génome du pois (V1)

Les données de génotypage de la collection AMS générées dans le cadre du projet PeaMUST obtenues avec la technique GBS (Ng et al., 2009) et en utilisant la séquence du génome V1 du cultivar « cameor » de pois (Kreplak et al. 2019) comme référence ont été utilisées lors de ce stage. La matrice de génotypage brute comportait 919144 SNP. J'ai appliqué un filtre de contrôle qualité avec une MAF (fréquence des allèles mineurs) > 5 %, à l'aide des logiciels *TASSEL* (Bradbury et al., 2007) pour les méthodes multi-locus à effet SNP aléatoires (mrMLM) et *GAPIT* (Wang et Zhang, 2021) pour les méthodes MLM à locus unique (ULMLM) et multilocus à effet SNP fixes ou partiellement fixes (MLMLM).

2.4.2. Données génotypiques basées sur la deuxième version du génome du pois (V2)

Les séquences contextes (+/- 250 pb) des différents SNP identifiés pour la collection AMS sur la V1 de l'assemblage du génome de référence de pois ont été alignées sur une nouvelle version améliorée et en cours de valorisation du génome du cultivar «Cameor» (V2). Cette version présente une meilleure continuité des séquences et une meilleure affectation des séquences génomiques aux pseudo-molécules ou chromosomes (J. Kreplak, communication personnelle). L'annotation de cette version n'étant pas disponible à ce stade, un transfert de celle de la première version a été réalisé grâce au logiciel *gmap* (Wu et Watanabe, 2005). L'ensemble des opérations de transfert des modèles de gènes de la V1 à la V2 a été réalisé par les bioinformaticiens de l'équipe et un fichier *gff* incluant les positions des gènes a été mis à ma disposition pour le stage.

2.5. Analyse d'association à l'échelle du génome (GWAS)

2.5.1. Structuration de la population

2.5.1.1. Nécessité de stratification et niveau de stratification en fonction des traits

Pour l'ensemble des 63 variables étudiées, la nécessité de prendre en compte ou pas la stratification de la population (Q) ainsi que le niveau de stratification optimal ont été décidés sur la base des valeurs du critère d'information bayésien (BIC), dérivé du critère d'information d'Akaike (Akaike, 1973). Les valeurs de BIC pour chaque variable ont été calculées avec le logiciel *GAPIT* en utilisant un modèle MLM et en variant le nombre de groupes de stratification optimal de 0 à 5 pour chaque variable.

2.5.1.2. Modèle de structuration

Pour évaluer la nécessité d'inclure dans le modèle GWAS à la fois la stratification (Q) et la matrice de parenté (K) comme des cofacteurs, j'ai comparé les valeurs de BIC de trois variantes de modèles : naïf (absence de Q et de K), MLM (avec Q seulement) et MLM (avec Q+K). Ensuite pour comparer l'efficacité de l'approche de calcul de la matrice de parenté (K), j'ai comparé trois modèles qui utilisent des approches différentes pour la calculer : MLM (Yu et al., 2006), CMLM (Zhang et al., 2010) et Fasta-LMM-Select (Listgarten et al., 2012) (Annexe 2). MLM utilise la méthode de VanRaden (2008) qui calcule les fréquences alléliques et l'identité par état pour estimer l'identité par descendance et donc les coefficients de parenté (Speed et Balding, 2015). CMLM calcule la matrice de parenté entre des groupes d'individus qu'il construit en utilisant des algorithmes de regroupement. Ainsi, la parenté entre les groupes est calculée simplement comme la moyenne de la parenté entre les individus et la parenté entre paires de groupes remplace la parenté entre paires d'individus pour l'effet aléatoire d'un MLM. Enfin, Fasta-LMM-Select calcule aussi la matrice de parenté par groupe d'individus. Il sélectionne un sous-ensemble de SNP associés au trait d'intérêt, de sorte que les matrices de parenté calculée soie spécifique à chaque trait tout en excluant les marqueurs qui sont en DL (2cM) avec les marqueurs représentant chaque groupe.

Également, dans une logique de comparaison globale des modèles à locus unique, les valeurs de BIC ont été comparées pour 11 variables prises pour test.

Pour la stratification (Q) dans les différentes méthodes, le nombre d'axes optimal de composantes principales (ACP) préalablement déterminé pour chaque variable a été utilisé. Ainsi, les matrices de stratification (nombre d'axe ACP) ont été générées avec le logiciel *GAPIT* selon le nombre d'axes défini.

2.5.2. Analyses GWAS proprement dites

Pour les analyses GWAS avec les méthodes ULMLM et MLMLM, le package *LegGWAS* en cours de développement au niveau du laboratoire a été utilisé. *LegGWAS* est un package *R* qui sert à traiter des données issues des méthodes de GWAS et est compatible avec *GAPIT*. Afin de minimiser au maximum le taux de faux positifs dans la détection des associations significatives entre marqueurs et phénotypes, j'ai appliqué la correction de bonferroni au seuil de 5 %, correspondant à une p-value = $6{,}94\ 10^{-8}$.

Concernant les méthodes mrMLM, j'ai utilisé le package *mrMLM* : *Multi-Locus Random-SNP-Effect Mixed Linear Model Tools for GWAS* (Wang et al., 2016). J'ai fixé un seuil de significativité de LOD = 3 (p-value = 0,0002), comme recommandé par Zhang et al. (2019).

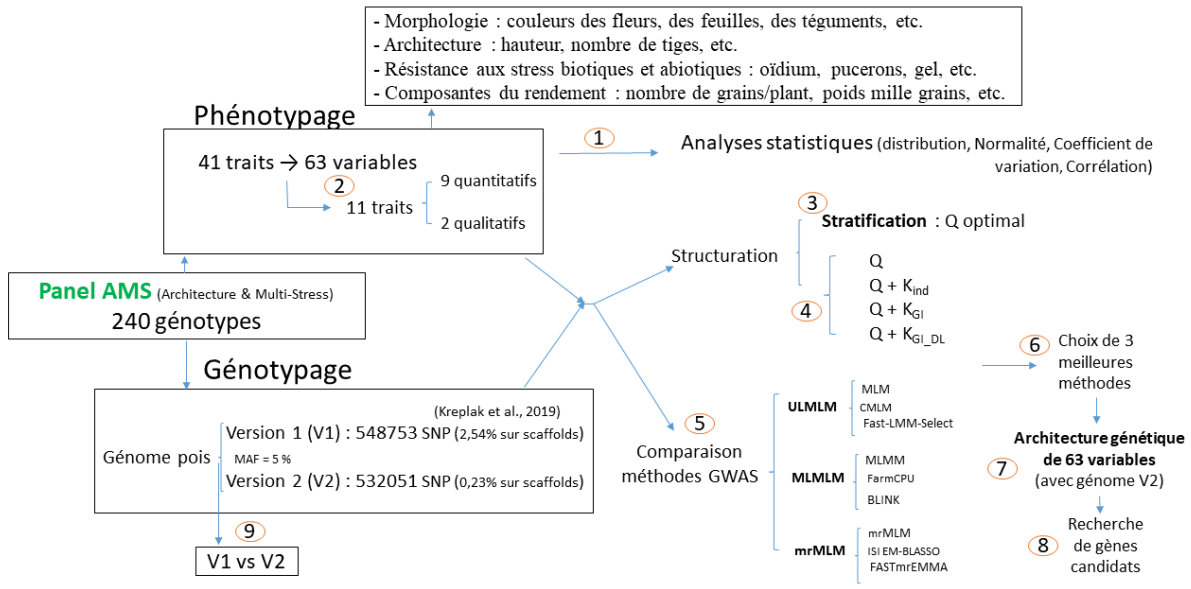
Les figures de Manhattan et de qqplot pour l'ensemble des méthodes ont été générées avec le package *qqman* (Turne, 2018).

2.5.3. Comparaison des résultats de GWAS issus de chacune des deux versions du génome du pois

Pour comparer les résultats de GWAS obtenus avec les SNP placés sur chacune des deux versions du génome du pois, j'ai utilisé les trois meilleures méthodes identifiées dans l'étape de comparaison des méthodes. Ainsi, j'ai analysé et comparé les sorties manhattan et qqplot de cinq des 11 variables test.

2.5.4. Comparaison des résultats avec les connaissances disponibles

Une approche avec *a priori* a été utilisée pour identifier des gènes connus correspondant à certains des pics de SNP identifiés. Pour cela, les déterminants génétiques connus pour des traits analysés lors du stage ont été recherchés dans la bibliographie puis placés sur le génome par alignement de séquences (BLAST). La colocalisation ou non de ces gènes avec les pics de GWAS identifiés a été ensuite discutée. J'ai également comparé les résultats obtenus avec une étude de GWAS (Ollivier et al. 2022) réalisée avec le même panel et les mêmes données de génotypage que j'ai utilisé.



Kreplak et al., 2019. *Nat Genet* : 1546-1718
K_{ind} : calcul de K par individus 2 à 2 ; K_{GI} = calcul de K par Groupe d'Individus ; K_{GI_DL} = calcul de K par GI et prise en compte du DL

<u>Figure 4</u> : Résumé schématique de la démarche adaptée pendant le stage

Dans un premier temps **(1)** j'ai fait l'analyse statistique des 63 variables de la collection, notamment, l'étude de la distribution et de la normalité des variables, leur coefficient de variation (CV) et une étude de corrélation entre elles. Ensuite **(2)**, en se basant sur les corrélations entre les variables, j'ai choisi 11 variables dont 9 quantitatives et 2 qualitatives, pour les tests des modèles de structuration et la comparaison des méthodes de GWAS. Pour faire les tests des modèles de structuration et la comparaison des méthodes GWAS, j'ai utilisé la version 1 (V1) du génome qui était déjà disponible. Pour les tests des modèles de structuration, j'ai d'abord étudié la stratification **(3)**, c'est-à-dire le nombre d'axes de PC (ou de sous-groupes) optimal pour chacune des 63 variables. Ensuite, j'ai testé quatre différents modèles de structuration **(4)** : un qui prend en compte que la stratification (Q, correspond au modèle GLM) et trois autres qui prennent à la fois la stratification et la matrice de parenté entre les individus, mais qu'ils calculent de façons différentes ($Q+K_{ind}$, correspond au modèle MLM de base, calcul la matrice de parenté entre paire d'individus en utilisant la méthode de VanRaden ; $Q+K_{GI}$, correspond au modèle CMLM, calcul la matrice de parenté par groupes d'individus préalablement formés ; $Q+K_{GI\_DL}$, correspond à Fast-LMM select, calcule la matrice de parenté aussi entre groupe d'individus de façon spécifiques à chaque trait et considère aussi le déséquilibre de liaison entre les marqueurs pour éviter la redondance). Après, **(5)** j'ai comparé 9 méthodes GWAS réparties entre trois différentes approches : MLM à locus unique (ULMLM), MLM multilocus avec effet marqueurs fixe ou partiellement fixe (MLMLM) et MLM multilocus à effet marqueurs aléatoire intégrant des méthodes de calcul bayésiennes (mrMLM). A l'issue de la comparaison des méthodes, **(6)** j'ai choisi les 3 meilleures méthodes. Ces trois meilleures méthodes ont été appliquées sur la deuxième version du génome V2, **(7)** pour étudier l'architecture génétique des 63 variables de l'étude. Ainsi, quelques marqueurs liés aux traits (identifiés au moins par deux des trois méthodes) ont été utilisés (**8**) pour rechercher les gènes candidats liés, en ajoutant un intervalle de 10kb de part et d'autre de la position du SNP. J'ai également **(9)** comparé, les résultats de GWAS des deux versions du génome pour évaluer la qualité de l'assemblage qui a été améliorée chez la V2.

3. RESULTATS

Dans l'objectif d'explorer les bases génétiques des traits clés du pois à travers un panel de 240 individus et de comparer des méthodes de GWAS, j'ai d'abord synthétisé et caractérisé les informations disponibles pour 63 variables étudiées. Ensuite, pour déterminer le meilleur modèle de structuration pour le panel, j'ai conduit une analyse comparative de modèles de structuration. Après, pour choisir les meilleures méthodes GWAS permettant d'explorer au mieux les bases génétiques, j'ai fait un choix de neuf méthodes sur l'ensemble des méthodes identifiées dans la bibliographie, avant de comparer ces neuf méthodes entre elles. Aussi, pour évaluer l'impact de la qualité de l'assemblage du génome sur les résultats de GWAS, avec les trois meilleures méthodes identifiées parmi les neuf, j'ai comparé les résultats de GWAS entre la première version du génome du pois et une nouvelle version améliorée disponible en interne, mais pas encore publiée. Enfin, pour recommander des outils nécessaires pour l'introgression des QTL (quantitative trait locus) dans les programmes de sélection, j'ai fait une synthèse de tous mes résultats et je les ai comparé avec les connaissances déjà disponibles (Figure 4).

3.1. Caractérisation phénotypique de la collection de référence de pois et corrélations entre les différents traits mesurés

3.1.1. Variabilité phénotypique et héritabilité des variables étudiées

J'ai travaillé sur 41 différents traits, dont certains ont été observés ou mesurés dans différents sites et/ou sur différentes années chez la collection de référence de pois dite collection AMS, faisant au total 63 variables dont 51 variables quantitatives et 12 variables qualitatives (Annexe 1). Pour l'ensemble des variables étudiées, le pourcentage de données manquantes a été assez faible. Le pourcentage le plus important rencontré concernait la variable forme des folioles (Form_foliol) avec 28,3% de données manquantes (Tableau 2).

Les 12 variables qualitatives sont sous forme de données binaires ou de trois à six classes (Annexe 3). L'observation de l'histogramme de distribution et du qqplot de normalité (Annexe 3) de chacune des 51 variables quantitatives montrent qu'elles sont distribuées en plusieurs classes, avec un comportement normal ou assez proche de la normalité.

Les héritabilités au sens étroit des 63 variables ont varié entre 0,005 % et 100%. Les variables teneur en cuivre (Tx_Cu_Gr_B15) et teneur en zinc des grains (Tx_Zn_Gr_B15) et pourcentage du volume du grain mangé par la bruche (Pourc_Vol_Mange_Bruch_B16) avaient des héritabilités presque nulles (0,005%) (Tableau 2). Les valeurs d'héritabilité les plus importantes (h^2 = 100%) ont été obtenues pour les variables qualitatives de coloration du hile (Col_Hil), de texture de la graine (Text_Gr), de coloration des fleurs (Col_fleur) et de moucheture des grains (Gr_mouch).

Tableau 2 : Paramètres statistiques caractérisant les 63 variables phénotypiques mesurées ou observées sur la collection AMS de pois

Trait	Nbr_Obs	% Phénotypage	Min	Max	Moy	CV(%)	h^2(%)
Biom_Aer_Matur_MG15*	236	98,33	2,64	28,02	11,17	33,07	67,59
Col_Feuil_SortiHv_MS17*	227	94,58	1,00	3,50	1,73	29,57	32,74
Col_Feuil_SortiHv_ORS17*	201	83,75	1,00	3,50	1,85	42,14	64,65
Col_fleur*	226	94,17	1,00	2,00	1,38	35,28	100,00
Col_Hil*	223	92,92	1,00	2,00	1,67	28,29	100,00
Col_teg_Gr*	208	86,67	1,00	6,00	1,81	52,48	93,13
Deb_Flor_B15	239	99,58	123,00	164,00	142,26	5,53	85,02
Deb_Flor_B16	217	90,42	143,00	171,00	156,47	2,61	62,69
Deb_Flor_semAut_MSS17	219	91,25	161,00	213,50	183,01	7,08	69,21
Deb_Rempl_Gr_B15	240	100,00	137,00	176,00	155,27	3,81	82,79
Deb_Rempl_Gr_B16	216	90,00	153,00	180,50	169,39	2,56	62,25
Dega_gel_SortiHv_MS17	230	95,83	0,00	5,00	1,39	86,94	28,80
Dega_gel_SortiHv_MS18	223	92,92	0,50	5,00	2,96	54,18	50,69
Dega_gel_SortiHv_ORS17	216	90,00	0,00	5,00	2,25	74,50	35,62
Fecond_Puc_ArPo28	240	100,00	31,50	256,50	129,50	22,95	49,63
Fecond_Puc_LSR1	240	100,00	0,00	287,52	57,29	113,53	35,53
Fin_Flor_B15	240	100,00	140,00	181,00	162,43	3,04	74,61
Fin_Flor_B16	216	90,00	164,00	182,00	173,09	1,99	53,53
Fin_Flor_SemAut_MS17	214	89,17	195,50	229,00	216,52	3,55	51,37
Form_feuil*	240	100,00	1,00	2,00	1,25	34,63	64,24
Form_foliol*	172	71,67	1,00	2,00	1,20	33,35	30,49
Gr_marbr*	236	98,33	1,00	2,00	1,85	19,50	49,04
Gr_mouch*	233	97,08	1,00	2,00	1,79	22,82	100,00
Haut_Fin_Flor_B15	240	100,00	10,00	155,00	93,60	34,15	60,11
Haut_Fin_Flor_MG15	239	99,58	2,47	107,74	59,89	29,30	67,67
Haut_Fin_Flor_R16	224	93,33	20,00	200,00	97,34	37,59	56,38
Haut_Fin_Flor_R17	225	93,75	30,00	160,00	94,80	29,55	57,78
Haut_TigPrinc_SortiHv_MS17	223	92,92	39,50	151,25	89,51	34,58	69,36
Haut_TigPrinc_SortiHv_ORS17	201	83,75	1,00	5,00	2,53	37,44	54,28
Matur_B15	240	100,00	170,00	210,00	186,60	2,55	75,00
Nbr_Gr_1Plt_B15	237	98,75	2,15	116,98	53,61	40,05	39,69
Nbr_Gr_1Plt_MG15	202	84,17	1,31	90,75	17,25	63,61	43,87
Nbr_Gr_Bruch_B15	236	98,33	0,00	859,00	149,64	86,87	19,52
Nbr_Jr_Flor_B16	216	90,00	7,50	26,50	16,67	19,97	44,97
Nbr_Jr_Flor_SemAut_MS17	214	89,17	13,50	55,50	33,79	23,70	21,93
Nbr_Ramif_Aer_1Plt_MG15	236	98,33	-0,03	4,77	0,61	107,68	33,07
Nbr_Ramif_Basal_1Plt_B15	236	98,33	0,92	4,05	1,66	29,82	27,38
Nbr_Ramif_Basal_1Plt_MG15	236	98,33	0,88	7,90	2,79	37,45	22,05
Nbr_Tig_DebFlor_ORS17	213	88,75	1,00	4,75	1,89	51,93	17,01
NDVI_Flor_MG15	240	100,00	0,14	0,64	0,43	18,18	61,38
PMG_B15	239	99,58	20,93	365,14	156,77	37,14	85,36
PMG_B16	210	87,50	12,50	304,48	132,41	36,15	62,07
PMG_MG15	202	84,17	4,43	262,05	122,18	32,80	73,90
Poids_Gr_1Plt_B15	238	99,17	0,05	18,68	8,12	44,69	51,70
Poids_Gr_1Plt_MG15	202	84,17	-0,07	9,52	2,12	68,70	67,09
Pourc_Gr_0Bruch_B15	236	98,33	59,45	100,00	94,41	4,62	8,20
Pourc_Gr_Bruch_B16	232	96,67	0,00	86,05	12,86	101,32	62,95
Pourc_Vol_Mange_B16	232	96,67	0,00	35,96	2,71	117,69	3,78
Pourc_Vol_Mange_Bruch_B16	226	94,17	7,76	73,34	19,38	30,30	0,00
Resist_Oid_MG15	237	98,75	0,04	5,05	3,80	21,77	65,47
Resist_Puc_MG15	237	98,75	2,12	5,72	4,27	17,59	25,53
Tail_Feuil_SortiHiv_MS17*	229	95,42	1,00	3,00	1,83	39,02	36,53
Tail_Feuil_SortiHiv_ORS17*	201	83,75	1,00	3,00	2,06	21,21	6,70
Text_Gr	234	97,50	1,00	2,00	1,07	24,25	100,00
Tx_Cu_Gr_B15	224	93,33	3,69	19,79	7,42	24,43	0,00
Tx_Fe_Gr_B15	224	93,33	19,92	153,75	40,63	31,04	14,28
Tx_Mn_Gr_B15	224	93,33	4,70	14,90	8,17	20,39	12,11
Tx_Prot_Gr_B15	227	94,58	16,27	30,88	22,28	9,85	48,37
Tx_Zn_Gr_B15	224	93,33	37,85	417,97	63,34	43,81	0,00
Volum_Gr_0Bruch_B16	232	96,67	20,51	155,22	89,99	31,62	40,89
Volum_Gr_B16	232	96,67	20,51	151,26	91,15	31,00	38,90
Volum_Gr_Bruch_B16	226	94,17	20,32	164,50	96,20	26,15	22,00
Volum_Gr_Bruch_Theor_B16	226	94,17	47,16	192,14	116,63	21,82	24,92

Nbr_Obs : Nombre d'individus de la collection pour lesquels des données phénotypiques sont disponibles pour la variable en question ; **% Phénotypage** : le rapport en pourcentage entre le nombre d'individus phénotypés pour une variable et le nombre total d'individus du panel ; **Min** : valeur minimale observée ou mesurée ; **Max** : valeur maximale mesurée ou observée ; **Moy**: moyenne des individus du panel pour la variable ; **CV** : Coefficient de variation de la variable ; **h^2 (%)** : héritabilité au sens étroit. ***** : indique les variables qualitatives.

3.1.2. Corrélations entre les différentes variables étudiées

La représentation de l'ensemble des variables sur les deux principaux axes de l'analyse factorielle de données mixtes (FAMD) (Figure 7) et le corrélogramme des variables quantitatives (Figure 6) ont permis d'apprécier les corrélations existantes entre les 12 variables qualitatives et les 51 variables quantitatives. Globalement, il a été observé plusieurs corrélations aussi bien entre les variables quantitatives (Figure 6) et entre les variables quantitatives et qualitatives (Figure 7).

3.1.3. Choix de variables pour l'étude comparative des méthodes GWAS

Pour comparer les approches et les méthodes de GWAS, j'ai sélectionné des variables test à partir des 63 variables de l'étude, dont des variables quantitatives et qualitatives. Dans un premier temps, les 51 variables quantitatives ont été réparties en groupes, selon leurs relations (Figure 5) puis une analyse plus approfondie à l'intérieur des groupes de corrélation (n=4) a permis de les subdiviser en sous-groupes en se basant sur le type de trait concerné. Ainsi, j'ai constitué neuf sous-groupes de variables quantitatives. Pour chaque sous-groupe, la variable la plus représentative (meilleure contribution dans la constitution des axes, Figure 5) a été retenue pour tester les variantes de méthodes GWAS. Dans un deuxième temps, et grâce à des critères similaires à ceux pour le choix des variables quantitatives, les deux variables qualitatives, couleur des feuilles à la sortie de l'hiver (Col_Feuil_SortiHiv_ORS17) et couleur des fleurs (Col_Fleur), ont été retenues à leur tour pour les tests de méthodes de GWAS (Figure 7). Les 11 variables choisies (quantitatives et qualitatives) sont listées dans le Tableau 3.

3.1.4. Analyses comparatives de modèles de structuration

3.1.4.1. Stratification de la population en fonction des variables étudiées

Pour l'optimisation de la prise en compte de la stratification dans les analyses GWAS sur le panel AMS, le calcul du BIC pour chacune des variables a été effectué comme précisé dans la section Matériel et Méthodes. Les résultats obtenus ont montré que, pour l'ensemble des 63 variables étudiées, le nombre de composantes principales (PC) idéal à prendre comme cofacteur pour un meilleur ajustement du modèle ne dépassait pas deux. La grande majorité des variables (63,49%) présentait des BIC optimaux en l'absence de stratification (PC = 0), 31,75% avec un seul axe de PC et seulement 4,76% pour deux axes de PC (Tableau 4).

3.1.4.2. Modèle optimal de structuration de la population dans le modèle GWAS

Le modèle de structuration peut varier d'une population à une autre et selon le trait étudié. Pour évaluer la nécessité pour notre panel d'étude, d'inclure dans le modèle GWAS à la fois la stratification (Q) et la matrice de parenté (K) comme des cofacteurs, avec les variables test choisies préalablement, j'ai comparé les valeurs de BIC de trois variantes de modèles : naïf (absence de Q et de K), GLM (MLM avec Q seulement) et MLM (avec Q+K). Ensuite pour comparer l'efficacité de l'approche de calcul de la matrice de parenté (K), j'ai comparé trois modèles qui utilisent des approches différentes pour le faire (MLM, CMLM et Fasta-LMM-Select).

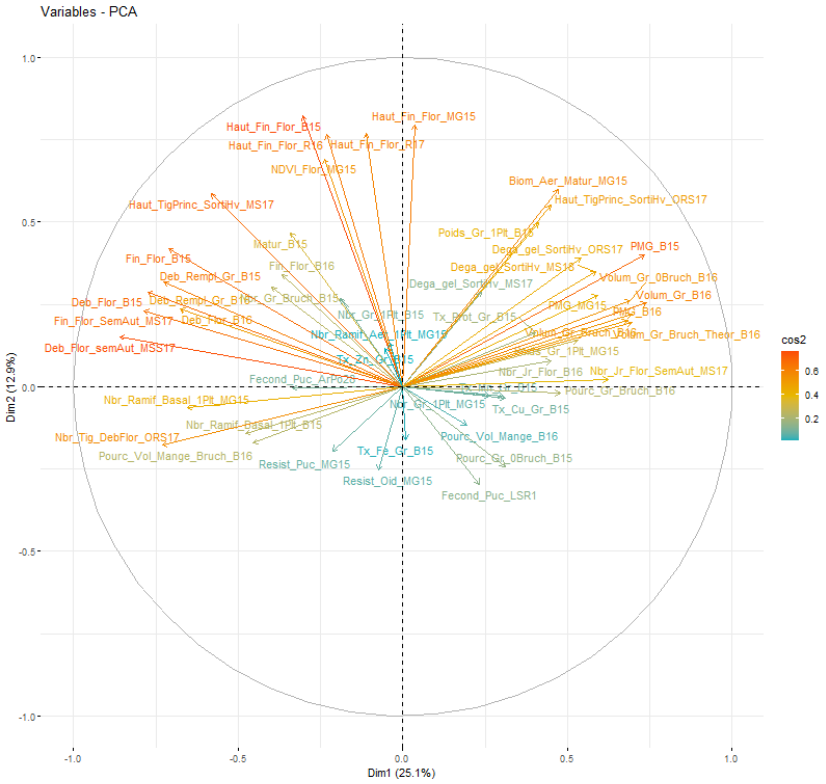


Figure 5 : Cercle de corrélation des variables quantitatives

Cercle de corrélation des 51 variables quantitatives avec indication du niveau de contribution de chaque variable à la construction des axes. Les deux axes présentés expliquent 38 % de la variabilité observée. Le cos^2 permet d'évaluer le niveau de contribution de chaque variable de trait à la construction des axes de l'ACP, plus le cos^2 est élevé, plus la variable en question participe à la construction des axes (explique plus la variabilité observée). Selon la grille de couleur, plus la variable tend vers le rouge, plus sa contribution à la construction des axes est importante et plus elle tend vers le bleu, moins sa contribution est importante (faible participation à la variabilité observée). On peut distinguer quatre groupes de corrélation, dans le sens des aiguilles d'une montre en commençant par la variable Biom_Aer_Matur_MG15. Un premier groupe avec 17 variables : dégât de gel (3), PMG et Poids Grains (5), Volume Grains (1), Volume Grains bruchés (3), durée floraison (2), Biomasse Aérienne (1), Taux de Protéine (1), hauteur tige principale (1). Un deuxième groupe avec 8 variables : en majorité des variables de Pourcentage Volume de Grain/Bruche (3), teneur en micro éléments (3), Fécondité de Pucerons (1), Nombre de Grains par plant (1). Un troisième groupe qui contient 7 variables : Nombre de Ramifications (3), Résistance à l'oïdium et aux pucerons (2), Fécondité des pucerons (1) et le dégât des bruches dans les graines (1). Un quatrième groupe comporte 19 variables avec en majorité des variables de floraison et de Hauteur : Hauteur (4), durée de floraison (6), début de remplissage des grains et maturité (3), Nombre de ramifications (1), Attaque de bruches (Nbr_Gr, 1), l'indice de végétation à floraison (1), Taux de Zinc dans les grains matures (1), Nombre de grains par plante (1).

Les résultats obtenus avec les 11 variables tests (Tableau 5) ont montré que, pour les quatre variables test dont le BIC était optimal sans stratification (hauteur à la fin floraison, pourcentage de grains bruchés, résistance à l'oïdium et teneur en manganèse dans les grains) le niveau d'ajustement du modèle naïf est égal à celui des modèles GLM. Les modèles MLM et Fast-LMM-Select, même s'ils intègrent la matrice de parenté entre individus (K), ont un ajustement égal à ceux des modèles naïfs et GLM. Par contre, le modèle CMLM qui intègre aussi la parenté génétique améliore sensiblement l'ajustement pour ces variables ne nécessitant pas une stratification.

Pour les variables nécessitant une stratification avec un ou deux axes de PC, le modèle GLM montre toujours un meilleur ajustement par rapport au modèle naïf (Tableau 5). Par contre, il est moins performant que les modèles MLM, CMLM et Fast-LMM-select. Les modèles MLM et Fast-LMM-Select ont le même ajustement. Le modèle CMLM montre un meilleur ajustement par rapport à MLM et Fast-LMM-Select pour toutes les variables nécessitant une stratification, sauf pour le volume du grain bruché (Volum_Grain_Bruch_B16) avec laquelle, ils présentent un ajustement égal.

Globalement, il a été observé que, les modèles MLM et Fast-LMM-Select même s'ils ont des approches différentes pour calculer la matrice de parenté (K), ils présentent des ajustements similaires, indépendamment de la prise en compte ou non de la stratification et aussi du niveau de stratification le cas échéant. Ces modèles n'apportent pas de différences par rapport au modèle naïf pour les variables ne nécessitant pas une stratification. L'inclusion de la matrice de parenté (K) dans le modèle CMLM améliore l'ajustement du modèle avec les variables nécessitant ou pas une stratification (Q). Ainsi, dans la suite pour la comparaison des méthodes GWAS, la structuration se fera avec le modèle Q + K.

3.1.5. Choix des méthodes GWAS à tester

Après recherches bibliographiques, j'ai identifié 30 méthodes GWAS simple trait. Ces méthodes peuvent être réparties sous trois modèles : modèle naïf (simple analyse de variance (ANOVA)), Modèle Linéaire Généralisé (GLM) et Modèle Linéaire Mixte à locus unique (MLM). Elles peuvent être également classées en fonction de l'approche utilisée : méthodes MLM à locus unique (ULMLM), méthodes MLM multi-locus à effet SNP fixes ou partiellement fixes (MLMLM) et méthodes multi-locus à effet SNP aléatoires et basées sur une approche Bayésienne (mrMLM).

Pour des raisons de temps, il ne m'était pas possible de tester l'ensemble des méthodes, j'ai choisi neuf d'entre elles, de façon à couvrir les trois approches simple trait, tout en prenant en compte leur niveau de citation dans la littérature, leur mode de calcul, leur puissance statistique et les connaissances issues d'études comparatives de méthodes GWAS chez des espèces végétales partageant des similarités avec le modèle pois (Wen et al., 2018 ; Xu et al., 2018 ; Lu et al., 2018) et/ou les résultats de tests préliminaires que j'ai réalisés. Ainsi, j'ai choisi :

- ✓ trois méthodes ULMLM : MLM classique (Yu et al., 2006), CMLM (Zhang et al., 2010) et FaST-LMM-Select (Listgarten et al., 2012) ;
- ✓ trois méthodes MLMLM : MLMM (Segura et al., 2012), FarmCPU (Liu et al., 2016) et BLINK (Huang et al., 2019) ;

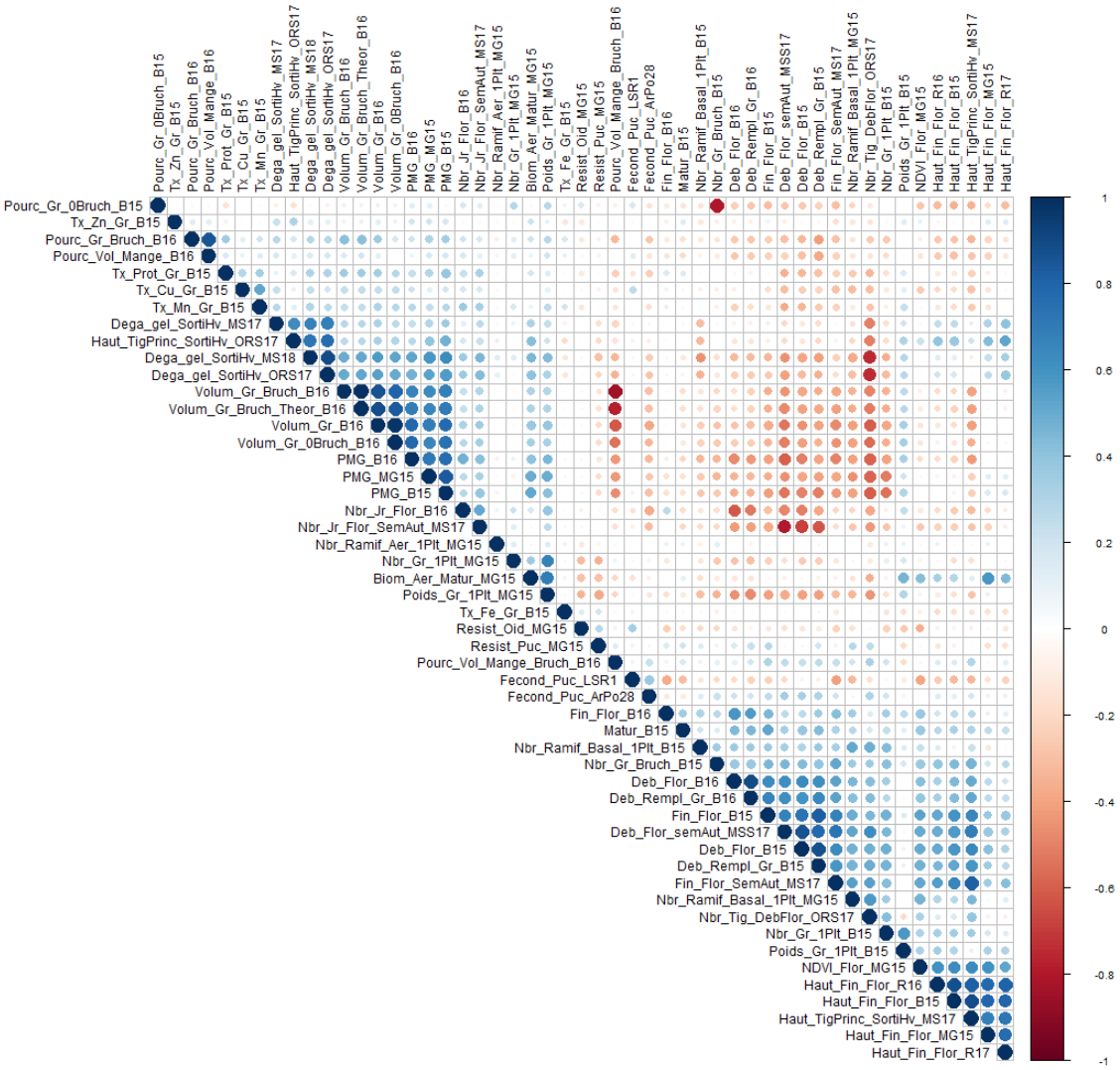


Figure 6 : Corrélogramme des 51 variables quantitatives étudiées

Corrélogramme construit à partir du package *corrplot* du logiciel R. pour 51 variables quantitatives étudiées sur la collection AMS et dont les données ont été mises à disposition pour le stage. Les corrélations non significatives au seuil de 1% sont colorées en blanc. Les coefficients de corrélation significatifs sont colorés en fonction de leur valeur de corrélation et du sens de la corrélation (positive ou négative). Plus le coefficient de corrélation est significatif et positif, plus il tend vers le bleu ; plus il est négatif et significatif, plus il tend vers le rouge.

- ✓ trois méthodes mrMLM : mrMLM (Wang et al., 2016) ; ISIS EM-BLASSO (Tamba et al., 2017) et FASTmrEMMA (Wen et al., 2018) (Annexe 2).

Les neuf méthodes choisies ont été testées par la suite sur les 11 variables choisies (Tableau 3), pour comparer les méthodes GWAS.

3.1.6. Analyse comparative des méthodes et approches GWAS simple trait

Les analyses comparatives des méthodes et approches GWAS ont été effectuées sur la première version du génome du pois en attendant la mise à disposition de la version améliorée de l'assemblage du génome. Après application d'un filtre de MAF à 5%, la matrice de génotypage comprenait 548753 marqueurs SNP dont 13925 SNP sur des scaffolds (2,54%). Les comparaisons des méthodes et des approches ont été basées sur l'examen des figures de Manhattan et des qqplot et sur l'identification des SNP ou des locus communs entre les méthodes.

Tableau 3 : Variables choisies pour la comparaison des approches et des méthodes de GWAS

Trait	Type de variable	Sous-groupe d'appartenance	CV(%)	h^2(%)	Nbr_PCs Optimal
Deb_Flor_semAut_MSS17	quantitative	date de floraison, de maturité et de remplissage	7,08	69,21	1
Dega_gel_SortiHv_ORS17	quantitative	dégâts dus au gel	74,50	35,62	2
Haut_Fin_Flor_B15	quantitative	Hauteur	34,15	60,11	0
Nbr_Tig_DebFlor_ORS17	quantitative	nombre de tiges et de ramifications	51,93	17,01	1
PMG_B15	quantitative	poids mille grains et poids Grains	37,14	85,36	1
Pourc_Gr_Bruch_B16	quantitative	pourcentage et volume de grains par rapport aux attaques de bruche	101,32	62,95	0
Resist_Oid_MG15	quantitative	résistance à l'oïdium et au puceron	21,77	65,47	0
Tx_Mn_Gr_B15	quantitative	taux de microéléments dans les grains	20,39	12,11	0
Volum_Gr_Bruch_B16	quantitative	volume Grains Bruché	26,15	22,00	1
Col_Feuil_SortiHv_ORS17	qualitative	Non binaire	42,14	64,65	1
Col_fleur	qualitative	binaire	35,28	100,00	1

Les variables ont été choisies en fonction de leur type (quantitatif ou qualitatif), à partir de sous-groupes de variables corrélées entre elles. Ces variables représentatives du nombre total de variables couvrent également une grande diversité d'héritabilités. L'ensemble des nombres de PC optimal (**Nbr_PCs Optimal)** trouvés avec les 63 variables a été représenté.

Tableau 4 : Tableau comparatif des modèles de structuration de la population selon la variable.

Trait	Nbr_PCs Optimal	Naïf	GLM	MLM	CMLM	Fast_LMM-Select
Col_Feuil_SortiHv_ORS	1	-242	-192	-162	-160	-162
Col_fleur	1	-165	-74,46	-2,40	62,76	-2,40
Deb_Flor_semAut_MSS17	1	-879	-800	-777	-773	-777
Dega_gel_SortiHv_ORS17	2	-426	-324	-313	-309	-313
Haut_Fin_Flor_B15	0	-1180	-1180	-1118	-1107	-1118
Nbr_Tig_DebFlor_ORS17	1	-306	-220	-215	-211	-215
PMG_B15	1	-1318	-1253	-1165	-1165	-1165
Pourc_Gr_Bruch_B16	0	-932	-932	-902	-901	-902
Resist_Oid_MG15	0	-299	-299	-258	-255	-258
Tx_Mn_Gr_B15	0	-440	-440	-430	-429	-430
Volum_Gr_Bruch_B16	1	-1057	-1010	-1007	-1007	-1007

Ce tableau présente les valeurs de BIC (Critère d'Information Bayésienne) obtenues des différents modèles à locus unique, permettant de comparer leur ajustement. Le modèle le plus ajusté est celui avec la valeur de BIC la plus élevée. Pour les modèles nécessitant une stratification avec Q, le nombre de PC optimal est mis en cofacteur. Le modèle naïf correspond à un MLM sans Q et K ; le modèle GLM correspond à un MLM avec Q, mais sans K.

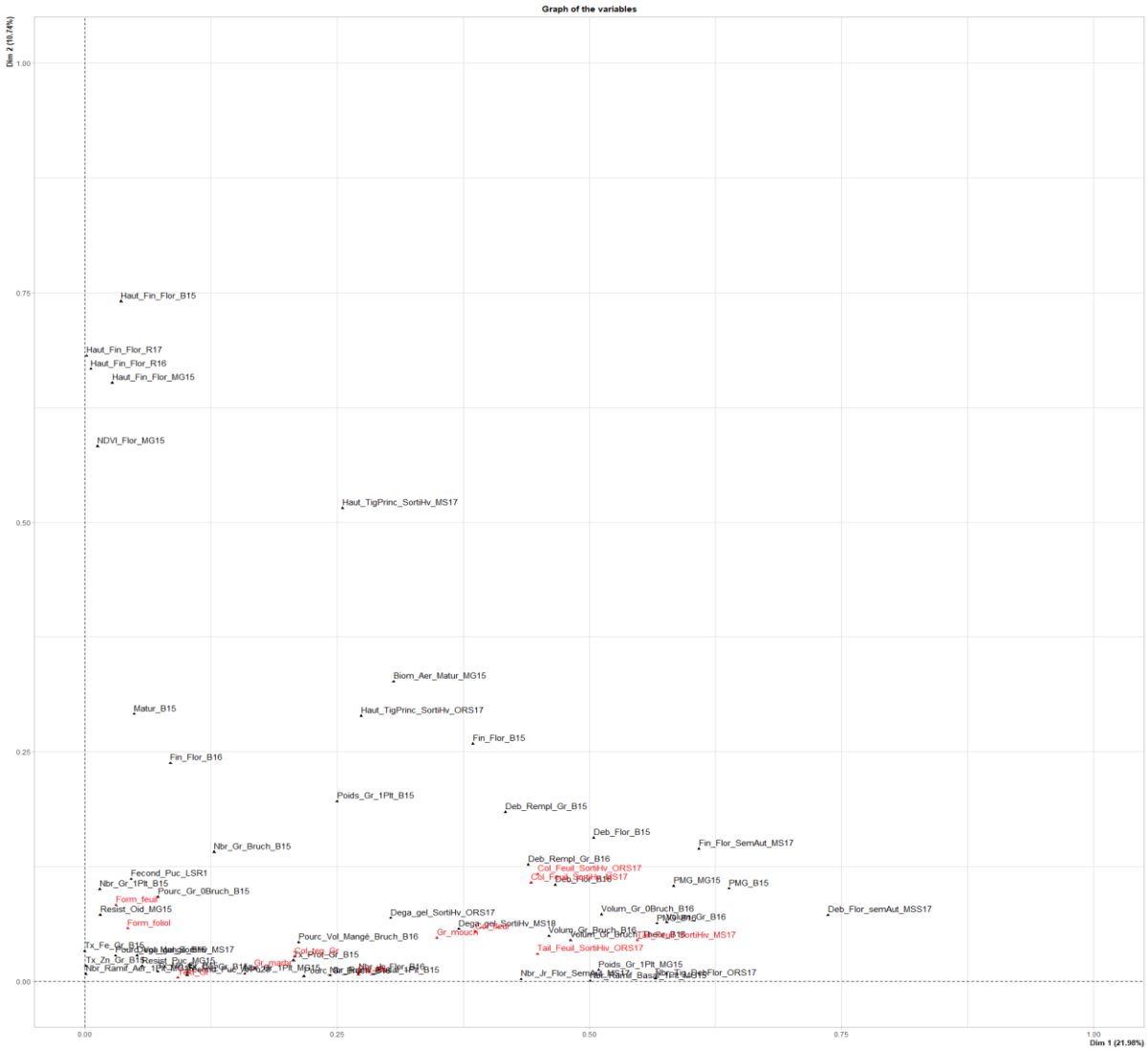


Figure 7 : Relation entre les variables quantitatives et qualitatives

Cette figure générée à l'issu de la FAMD (Analyse Factorielle de Données Mixtes) permet d'observer les relations entre les variables quantitatives et qualitatives, sur les deux axes principaux de composantes principales représentant 32,72 % de la variabilité observée. En rouge, nous avons les variables qualitatives et en noir, les variables quantitatives.

3.1.6.1. Comparaison des méthodes de l'approche GWAS à locus unique (ULMLM)

Les résultats de l'analyse comparative des méthodes GWAS à locus unique révèlent que, pour les variables qualitatives à l'exemple de la couleur des fleurs (Figure 8) et la couleur des feuilles à la sortie de l'hiver, les méthodes MLM et Fast-LMM-select sont égales en terme de nombre de SNP significatifs identifiés (240 SNP en commun) (Tableau 6 ; Figure 8.a1-b1-j) et de niveau d'ajustement du modèle (Tableau 6 ; Figure 8.a2-b2). Par contre, la méthode CMLM améliore l'ajustement du modèle (Tableau 4 ; Figure 8.z1.c2) et identifie moins de SNP significatifs (63 SNP) pour la couleur des fleurs, dont tous en commun avec MLM et Fast-LMM-Select (Figure 8.c1-j). CMLM parvient également à identifier des SNP significatifs pour la couleur des feuilles à la sortie de l'hiver, alors que MLM et Fast-LMM-Select n'en révèlent pas (Tableau 6).

Pour certaines variables quantitatives : dégâts dus au gel, poids de mille grains (Tableau 6, Figure 9.a1-b1-c1), teneur en manganèse des grains et volume du grain bruché, les méthodes de l'approche ULMLM ne parviennent pas à identifier des SNP significatifs. Pour les autres variables quantitatives pour lesquelles des SNP significatifs ont été identifiés avec les méthodes de l'approche ULMLM, les méthodes MLM et Fasta_LMM_Select sont égales en terme de nombre de SNP identifiés (Tableau 6 ; Figure 9.a1-c1) et d'ajustement du modèle (Tableau 5 ; Figure 9-10.a2-c2). Elles identifient plus de SNP significatifs (Tableau 6) et en même temps présentent un ajustement moins bon que CMLM (Tableau 5, Figure 8-9.b2). Par exemple, pour la résistance à l'oïdium, MLM et Fast-LMM-Select ont identifié 23 SNP identiques, dont 18 en commun avec CMLM.

3.1.6.2. Comparaison des méthodes de l'approche GWAS multilocus à effet SNP fixe ou partiellement fixe (MLMLM)

Les résultats de l'analyse comparative des méthodes de l'approche MLMLM ont montré que, pour la couleur des fleurs, MLMM, FarmCPU et BLINK ont identifié respectivement 8, 17 et 22 SNP significatifs, dont un seul identifié simultanément par les 3 méthodes et trois communs entre MLMM et FarmCPU (Tableau 6 ; Figure 8.k). Pour la couleur des feuilles à la sortie de l'hiver, MLMM, FarmCPU et BLINK ont identifié respectivement, 1, 5 et 4 SNP significatifs, dont 4 en commun entre FarmCPU et BLINK (Tableau 6). En terme d'ajustement du modèle GWAS pour les variables qualitatives, l'observation des qqplot pour la couleur des fleurs (Figure 8. d2-e2-f2) montre que la méthode BLINK est moins ajustée que MLMM, et que cette dernière est moins ajustée que FarmCPU. L'observation des figures qqplot d'autres variables qualitatives non présentées dans ce rapport (couleur du hile, couleur du tégument et couleur des feuilles à la sortie d'hiver) confirme la comparaison BLINK/FarmCPU. Même si dans certains cas, BLINK a un meilleur ajustement que MLMM, son ajustement reste inférieur à celui de FarmCPU dans la majorité des cas.

Tableau 5 : Résultats des calculs de nombre optimal de PC pour les 63 variables étudiées

Trait	0PCs	1PCs	2PCs	3PCs	4PCs	5PCs	Nbr_PCs Optimal
Biom_Aer_Matur_MG15	-592,5	-594,1	-596,7	-599,4	-600,3	-602,1	0PCs
Col_Feuil_SortiHv_MS17	-121,0	-120,1	-122,2	-124,8	-127,5	-128,2	1PCs
Col_Feuil_SortiHv_ORS17	-162,3	-162,0	-163,3	-165,8	-167,3	-168,5	1PCs
Col_fleur	2,9	3,0	5,2	2,5	-0,1	-2,4	1Pcs
Col_Hil	-28,2	-28,8	-29,6	-31,6	-34,3	-35,8	0PCs
Col_teg_Gr	-259,1	-261,7	-264,3	-266,9	-269,5	-271,9	0PCs
Deb_Flor_B15	-765,8	-766,9	-769,6	-772,3	-774,9	-777,6	0PCs
Deb_Flor_B16	-573,9	-575,1	-577,0	-579,7	-582,4	-584,5	0PCs
Deb_Flor_semAut_MSS17	-778,9	-777,2	-778,9	-781,0	-783,7	-786,3	1PCs
Deb_Rempl_Gr_B15	-708,3	-709,9	-712,6	-715,3	-717,9	-720,3	0PCs
Deb_Rempl_Gr_B16	-586,2	-587,6	-589,8	-592,5	-595,1	-597,1	0PCs
Dega_gel_SortiHv_MS17	-330,9	-333,0	-329,3	-330,1	-332,7	-335,4	2Pcs
Dega_gel_SortiHv_MS18	-310,4	-307,8	-305,3	-307,7	-309,2	-310,8	2Pcs
Dega_gel_SortiHv_ORS17	-321,5	-319,7	-313,5	-316,1	-317,6	-318,8	2Pcs
Fecond_Puc_ArPo28	-1136,4	-1138,6	-1141,2	-1143,5	-1145,9	-1148,6	0PCs
Fecond_Puc_LSR1	-1316,7	-1319,0	-1320,6	-1321,9	-1319,8	-1322,5	0PCs
Fin_Flor_B15	-669,4	-671,0	-673,7	-676,2	-678,6	-681,2	0PCs
Fin_Flor_B16	-547,1	-549,6	-552,1	-551,8	-554,5	-556,0	0PCs
Fin_Flor_SemAut_MS17	-650,6	-648,1	-649,1	-651,7	-652,8	-655,5	1PCs
Form_feuil	-62,1	-64,5	-63,8	-64,6	-67,1	-66,7	0PCs
Form_foliol	-81,7	-84,2	-85,3	-87,1	-88,1	-90,0	0PCs
Gr_marbr	-56,3	-58,2	-59,0	-61,1	-63,9	-66,7	0PCs
Gr_mouch	-28,1	-28,9	-31,5	-34,1	-36,2	-38,9	0PCs
Haut_Fin_Flor_B15	-1118,1	-1120,3	-1122,7	-1125,1	-1125,3	-1127,4	0PCs
Haut_Fin_Flor_MG15	-974,2	-977,0	-979,7	-982,4	-982,8	-984,4	0PCs
Haut_Fin_Flor_R16	-1085,1	-1087,5	-1089,9	-1092,5	-1092,3	-1094,4	0PCs
Haut_Fin_Flor_R17	-1035,2	-1037,9	-1040,2	-1042,9	-1043,2	-1045,7	0PCs
Haut_TigPrinc_SortiHv_MS17	-990,4	-990,6	-990,5	-993,1	-993,7	-996,1	0PCs
Haut_TigPrinc_SortiHv_ORS17	-231,4	-232,5	-233,6	-236,3	-237,7	-238,0	0PCs
Matur_B15	-674,3	-676,8	-679,3	-681,1	-683,2	-685,9	0PCs
Nbr_Gr_1Plt_B15	-1046,6	-1049,0	-1051,6	-1052,5	-1055,2	-1057,8	0PCs
Nbr_Gr_1Plt_MG15	-768,9	-771,4	-773,9	-776,6	-779,3	-781,2	0PCs
Nbr_Gr_Bruch_B15	-1469,6	-1470,4	-1473,0	-1473,8	-1475,0	-1476,2	0PCs
Nbr_Jr_Flor_B16	-542,1	-543,5	-545,7	-544,8	-547,6	-550,4	0PCs
Nbr_Jr_Flor_SemAut_MS17	-724,7	-724,4	-726,9	-726,6	-728,8	-731,5	1PCs
Nbr_Ramif_Aer_1Plt_MG15	-228,8	-231,5	-232,6	-235,3	-237,9	-240,7	0PCs
Nbr_Ramif_Basal_1Plt_B15	-154,5	-156,1	-158,8	-161,5	-164,2	-161,5	0PCs
Nbr_Ramif_Basal_1Plt_MG15	-300,4	-298,3	-299,9	-302,5	-305,3	-304,3	1PCs
Nbr_Tig_DebFlor_ORS17	-223,1	-215,4	-216,0	-218,7	-221,5	-223,3	1PCs
NDVI_Flor_MG15	319,5	316,9	314,2	313,7	313,0	310,4	0PCs
PMG_B15	-1165,9	-1164,5	-1166,0	-1168,1	-1165,2	-1167,8	1PCs
PMG_B16	-1018,4	-1016,3	-1016,6	-1018,6	-1017,8	-1020,1	1PCs
PMG_MG15	-955,5	-955,3	-957,2	-959,6	-958,4	-961,1	1PCs
Poids_Gr_1Plt_B15	-597,48	-599,20	-601,34	-601,30	-602,93	-605,58	0PCs
Poids_Gr_1Plt_MG15	-331,7	-333,1	-335,4	-338,0	-340,4	-342,2	0PCs
Pourc_Gr_0Bruch_B15	-684,8	-686,3	-688,6	-691,2	-690,7	-692,7	0PCs
Pourc_Gr_Bruch_B16	-605,1	-606,9	-608,4	-609,1	-611,8	-613,1	0PCs
Pourc_Vol_Mange_B16	-901,8	-903,7	-906,2	-908,5	-911,1	-913,2	0PCs
Pourc_Vol_Mangé_Bruch_B16	-717,5	-714,2	-716,7	-719,3	-721,2	-723,9	1PCs
Resist_Oid_MG15	-258,0	-260,6	-263,1	-263,8	-266,5	-268,8	0PCs
Resist_Puc_MG15	-265,5	-267,9	-270,6	-273,4	-276,2	-275,3	0PCs
Tail_Feuil_SortiHiv_MS17	-185,7	-184,2	-185,3	-188,1	-189,6	-191,8	1PCs
Tail_Feuil_SortiHiv_ORS17	-84,7	-79,8	-82,5	-85,0	-84,2	-87,7	1PCs
Text_Gr	73,3	70,9	68,1	66,0	63,3	62,4	1PCs
Tx_Cu_Gr_B15	-449,4	-447,0	-448,7	-451,1	-453,7	-455,8	1PCs
Tx_Fe_Gr_B15	-892,1	-894,6	-897,0	-896,4	-899,3	-897,0	0PCs
Tx_Mn_Gr_B15	-429,9	-430,1	-430,6	-431,7	-434,0	-436,0	0PCs
Tx_Prot_Gr_B15	-479,9	-481,7	-484,2	-485,7	-488,2	-490,9	0PCs
Tx_Zn_Gr_B15	-1066,7	-1069,4	-1072,1	-1074,9	-1077,6	-1080,3	0PCs
Volum_Gr_0Bruch_B16	-1033,1	-1031,0	-1032,3	-1034,4	-1035,2	-1037,2	1PCs
Volum_Gr_B16	-1023,3	-1020,2	-1021,1	-1022,8	-1022,9	-1025,4	1PCs
Volum_Gr_Bruch_B16	-1009,0	-1007,2	-1009,9	-1012,7	-1014,2	-1016,5	1PCs
Volum_Gr_Bruch_Theor_B16	-1006,6	-1004,7	-1007,5	-1010,2	-1011,8	-1013,9	1PCs

Ce tableau présente les valeurs du critère d'information bayésien (BIC) pour les 5 premiers axes de composantes principales (**PC**). Le nombre de PC optimal (**Nbr_PCs Optimal**) pour chaque variable est celui avec la valeur de BIC la plus élevée. Pour calculer les valeurs de BIC pour chaque variable, le modèle MLM a été utilisé en variant le nombre de groupes de stratification optimal de 0 à 5. Les calculs ont été faits avec GAPIT.

Pour l'ensemble des variables quantitatives testées, FarmCPU a identifié presque deux fois plus de SNP significatifs, suivi de BLINK, sauf pour le poids de mille grains où BLINK a identifié plus de SNP (Tableau 6 ; Figure 9.d1-e1-f1-k). Aucune des trois méthodes de l'approche multilocus n'a identifié de SNP significatifs pour le volume de grains bruchés. Pour la teneur en manganèse dans les grains, FarmCPU a révélé trois SNP significatifs sur les chromosomes 3, 4 et 5, alors que MLMM et BLINK n'en ont pas trouvé (Tableau 6). Concernant l'ajustement du modèle pour les variables quantitatives, l'observation des figures de qqplot a révélé que, BLINK et FarmCPU étaient meilleures que MLMM pour le poids de mille grains (Figure 9.d2-e2-f2), les dégâts dus au gel, le nombre de tiges en début de floraison et la teneur en manganèse des grains. Pour la durée de début floraison en semis d'automne et la résistance à l'oïdium (Figure 10. d2-e2-f2), BLINK et MLMM ont eu un meilleur ajustement que FarmCPU. Pour la hauteur en fin floraison, le pourcentage de grains bruchés et le volume de grains bruchés toutes les trois méthodes ont eu des ajustements assez similaires.

3.1.6.3. Comparaison des méthodes de l'approche GWAS multilocus à effet SNP aléatoire (mrMLM)

Les résultats de l'analyse comparative de méthodes de GWAS avec l'approche mrMLM (Tableau 6) ont montré que, pour les deux variables qualitatives étudiées, les méthodes mrMLM et FASTmrEMMA ont permis d'identifier plus de SNP significatifs que ISIS EM-BLASSO. Pour la couleur des fleurs, mrMLM, FASTmrEMMA et ISIS EM-BLASSO ont identifié respectivement 22, 45 et 8 SNP significatifs, dont aucun en commun entre les trois méthodes. mrMLM et FASTmrEMMA ont identifié 2 SNP en commun. mrMLM et ISIS EM-BLASSO ont identifié un seul SNP en commun, FASTmrEMMA et ISIS EM-BLASSO n'ont aucun SNP en commun (Figure 8.l). Pour la couleur des feuilles à la sortie de l'hiver, mrMLM a identifié 15 SNP significatifs, alors que FASTmrEMMA et ISIS-EM BLASSO ont identifié 6 et 5 SNP significatifs, respectivement (Tableau 6). L'analyse des résultats de la part de la variance phénotypique expliquée par les marqueurs significatifs identifiés (r^2) a montré que, les marqueurs identifiés par la méthode mrMLM expliquent plus largement la variance phénotypique pour la couleur des feuilles à la sortie de l'hiver. Pour la couleur des fleurs, les marqueurs identifiés par ISIS EM-BLASSO expliquent plus la variance phénotypique observée (Tableau 6). La méthode mrMLM a permis d'identifier un SNP expliquant plus de 50% de la variance liée à la couleur des fleurs. Pour l'ajustement du modèle de ces variables qualitatives, et vu que ISIS EM-BLASSO ne produit pas de qqplot, l'analyse des qqplot a porté uniquement sur les sorties des méthodes FASTmrEMMA et mrMLM. mrMLM a montré un meilleur ajustement aussi bien pour la couleur des fleurs (Figure 8.g2-h2) et pour la couleur des feuilles à la sortie de l'hiver que la méthode FASTmrEMMA.

Pour la grande majorité (7/9) des variables quantitatives testées, la méthode mrMLM a permis d'identifier plus de SNP significatifs (Tableau 6, Figure 9-10.g1-h1-i), suivie par la méthode ISIS EM-BLASSO. Par exemple, pour le poids de mille grains, mrMLM, FASTmrEMMA et ISIS EM-BLASSO ont identifié respectivement 29, 10 et 8 SNP significatifs, dont un seul en commun entre ces trois méthodes (Figure 9.l). Concernant la part de la variance phénotypique expliquée par les SNP identifiés (r^2), mrMLM a généré les meilleures valeurs de r^2 cumulées pour la grande majorité (8/9) des variables quantitatives testées (Tableau 6). Pour l'ajustement du modèle, l'analyse des qqplot a permis d'observer que mrMLM et FASTmrEMMA sont assez similaires et que leurs ajustements ne sont pas assez bons en général.

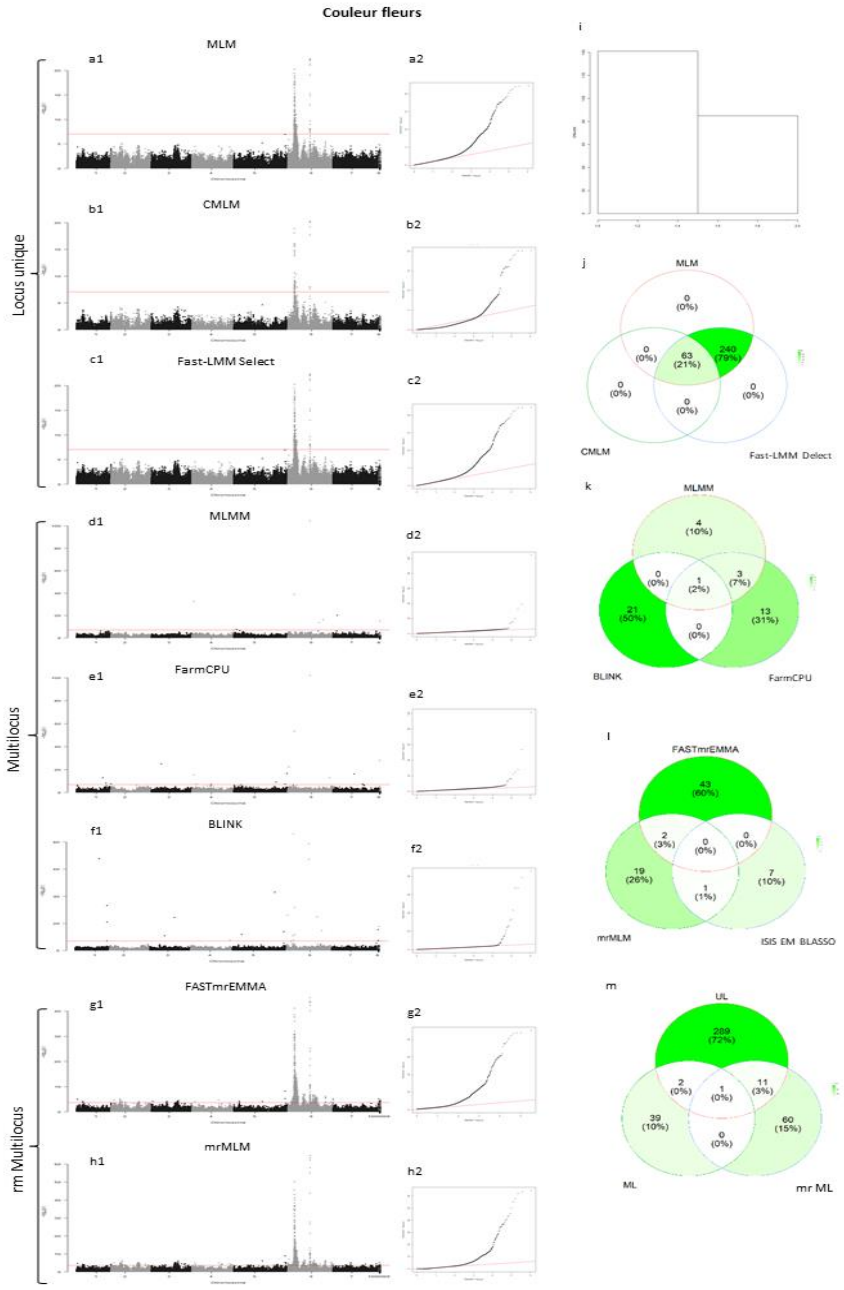


Figure 8 : Comparaison des méthodes et approches GWAS pour la couleur des feuilles

a1, b1, c1 correspondent respectivement aux Manhattan plot des méthodes de l'approche GWAS à locus unique (ULMLM) : MLM, Fast-LMM-Select et CMLM ; **d1, e1, f1**, correspondent respectivement aux Manhattan plot des méthodes de l'approche GWAS multilocus à effet SNP fixe ou partiellement fixe (MLMLM) : MLMM, FarmCPU et BLINK ; **g1, h1** correspondent respectivement aux Manhattan plot des méthodes de l'approche GWAS multilocus à effet SNP aléatoire (mrMLM) : FASTAmrEMMA et mrMLM ; **a2, b2, c2** correspondent respectivement aux qqplot des méthodes de l'approche ULMLM : MLM, Fast-LMM-Select et CMLM ; ; **d2, e2, f2** correspondent respectivement aux qqplot des méthodes de l'approche MLMLM : MLMM, FarmCPU et BLINK ; **g2, h2** correspondent respectivement aux qqplot des méthodes de l'approche mrMLM : FASTAmrEMMA et mrMLM ; **i** : histogramme de distribution de la variable couleur des feuilles, qui se présente sous forme binaire (Blanc =1 et Violet=2) ; **j** : diagramme de Venn pour les trois méthodes de l'approche ULMLM ; **k** : diagramme de Venn pour les trois méthodes de l'approche MLMLM ; **l** : diagramme de Venn pour les deux méthodes de l'approche mrMLM ; **m** : diagramme de Venn comparatif des trois approches UMLM (UL), MLMLM (ML) et mrMLM.

3.1.6.4. Comparaison des approches GWAS à locus unique et multilocus

Pour comparer les trois approches de GWAS simple trait étudiées (ULMLM, MLMLM et mrMLM), j'ai regroupé les résultats des méthodes relatives à chacune d'entre elles. Ainsi, les résultats obtenus ont montré que, pour la couleur des fleurs (Figure 8 ; Tableau 6), l'approche ULMLM a identifié deux pics de SNP significatifs sur le chromosome 6, tandisque MLMLM a permis d'identifier plusieurs SNP sur les chromosomes 1, 3, 4, 5, 6, 7 et sur les scaffolds, avec le SNP sur le chromosome 6 ayant la p-value la plus significative. L'approche mrMLM a identifié des SNP significatifs sur tous les sept chromosomes, avec deux importants pics de SNP sur le chromosome 6. Un seul SNP significatif est identifié comme commun aux trois approches. Les approches ULMLM et mrMLM ont 12 SNP en commun et ULMLM et MLMLM ont 3 SNP en commun (Figure 8. m). Pour la couleur des feuilles à la sortie de l'hiver (Tableau 6), l'approche ULMLM a identifié un seul SNP significatif sur le chromosome 5, particulièrement par la méthode CMLM. MLMLM a identifié des SNP significatifs sur les chromosomes 1 et 5 en plus d'un pic sur le chromosome 5. L'approche mrMLM a identifié des SNP significatifs sur l'ensemble des 7 chromosomes et sur les scaffolds également. Un seul SNP commun est identifié entre les approches ULMLM et mrMLM, par contre aucun SNP commun n'est identifié entre les trois approches.

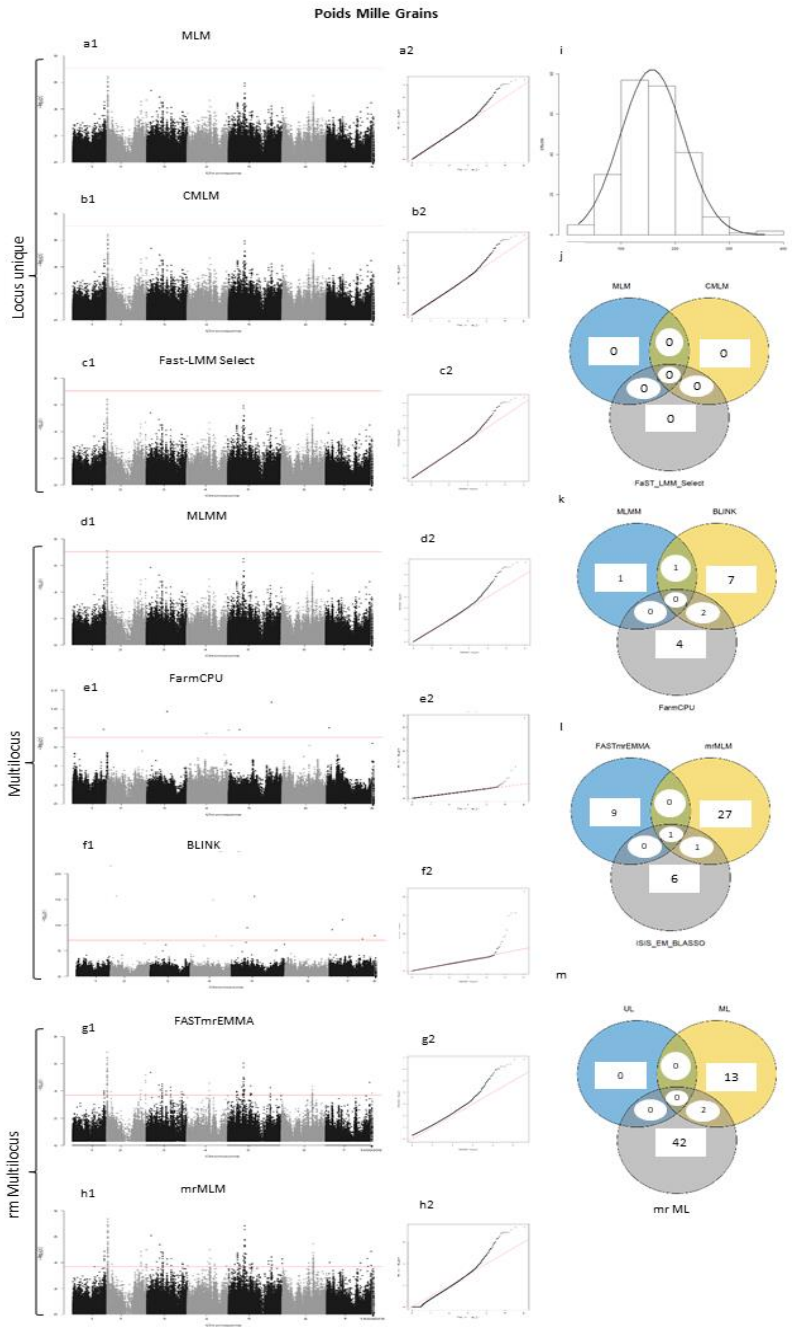


Figure 9 : Comparaison des méthodes et approches GWAS pour le poids mille grains

a1, b1, c1 correspondent respectivement aux Manhattan plot des méthodes de l'approche GWAS à locus unique (ULMLM) : MLM, Fast-LMM-Select et CMLM ; **d1, e1, f1**, correspondent respectivement aux Manhattan plot des méthodes de l'approche GWAS multilocus à effet SNP fixe ou partiellement fixe (MLMLM) : MLMM, FarmCPU et BLINK ; **g1, h1** correspondent respectivement aux Manhattan plot des méthodes de l'approche GWAS multilocus à effet SNP aléatoire (mrMLM) : FASTAmrEMMA et mrMLM ; **a2, b2, c2** correspondent respectivement aux qqplot des méthodes de l'approche ULMLM : MLM, Fast-LMM-Select et CMLM ; ; **d2, e2, f2** correspondent respectivement aux qqplot des méthodes de l'approche MLMLM : MLMM, FarmCPU et BLINK ; **g2, h2** correspondent respectivement aux qqplot des méthodes de l'approche mrMLM : FASTAmrEMMA et mrMLM ; **i** : histogramme de distribution de la variable poids mille grains ; **j** : diagramme de Venn pour les trois méthodes de l'approche ULMLM ; **k** : diagramme de Venn pour les trois méthodes de l'approche MLMLM ; **l** : diagramme de Venn pour les deux méthodes de l'approche mrMLM ; **m** : diagramme de Venn comparatif des trois approches UMLM (UL), MLMLM (ML) et mrMLM (mrML).

Concernant les variables quantitatives, pour la date de début floraison (Tableau 6), l'approche ULMLM a identifié un seul pic de SNP sur le chromosome 5. MLMLM a identifié des SNP significatifs sur les chromosomes 1, 2, 3, 5 et 7 et mrMLM sur les chromosomes 3, 5, 6, 7 et sur les scaffolds. Les approches ULMLM et mrMLM ont identifié deux SNP en commun et ULMLM et MLMLM ont un seul SNP en commun. Aucun SNP, ni de pics n'est identifié en commun entre les trois approches. Pour la résistance à l'oïdium (Tableau 6 ; Figure 10.m), l'approche ULMLM n'a identifié de pics ou SNP significatifs que sur le chromosome 1, alors que MLMLM en a identifié sur les chromosomes 1, 5 et 6 et mrMLM sur les chromosomes 1, 2, 3, 4, 5, 6 et sur les scaffolds. Aucun pic ou SNP significatif commun entre les trois approches n'est identifié. Les approches ULMLM et MLMLM ont un seul SNP significatif en commun. Pour la hauteur des plantes (Tableau 6), ULMLM a identifié deux pics de SNP significatifs, sur les chromosomes 1 et 5. MLMLM a identifié des SNP significatifs sur les chromosomes 2, 3 et 6 et sur les scaffolds en plus de ceux sur les chromosomes 1 et 5. mrMLM a identifié des SNP significatifs sur tous les sept chromosomes. Un seul SNP significatif est identifié en commun entre les trois approches, ULMLM et MLMLM ont deux SNP significatifs en commun et ULMLM et mrMLM ont quatre SNP significatifs en commun. Pour le volume de grains bruchés, seule l'approche mrMLM a identifié des SNP significatifs. Ces SNP sont répartis dans les sept chromosomes et sur les scaffolds. Pour les dégâts dus au gel, le poids de mille grains et la teneur en manganèse des grains (Tableau 6, Figure 9), l'approche ULMLM n'a pas identifié de pics de SNP significatifs, alors que MLMLM et mrMLM en ont identifié dans différentes régions du génome (Tableau 6). Pour le Nombre de tige en début floraison (Tableau 6), l'approche ULMLM a identifié 2 SNP significatifs sur le chromosome 2, MLMLM a identifié 2 sur les chromosomes 1 et 2, mrMLM a identifié 11 SNP répartis sur tous les chromosomes, sauf le chromosome 4. Pour le pourcentage de grains bruchés (Tableau 6), l'approche ULMLM a identifié un seul pic de SNP sur le chromosome 3, MLMLM a identifié 5 SNP sur les chromosomes 1 et 3 et 7, mrMLM a identifié 19 SNP répartis sur tous les chromosomes, sauf le chromosome 1. En termes d'ajustement de modèles, l'analyse des qqplot a permis de voir que, aussi bien pour les variables quantitatives que qualitatives, les méthodes de l'approche MLMLM ont donné de meilleurs ajustements. Les approches ULMLM et mrMLM ont des ajustements assez similaires, ou des fois même, les méthodes de l'approche ULMLM ont des meilleurs ajustements (Figure 8-9-10).

3.1.7. Impact de la qualité d'assemblage du génome du pois sur l'identification des bases génétiques des traits

Afin d'évaluer l'impact de la qualité de l'assemblage dans l'identification des bases génétiques des traits, les trois meilleures méthodes GWAS identifiées précédemment (CMLM, FarmCPU et mrMLM) ont été appliquées sur cinq traits quantitatifs (dégâts dus au gel, hauteur de la plante, poids de mille grains, résistance à l'oïdium et teneur en manganèse dans les grains) et un trait qualitatif (couleur des fleurs). Au total, 891463 SNP ont pu être placés avec confiance sur la V2 via un alignement des séquences contextes. Après filtrage, avec une MAF de 5%, 532051 SNP ont été retenus dont seulement 1208 SNP sur des scaffolds (0,23%).

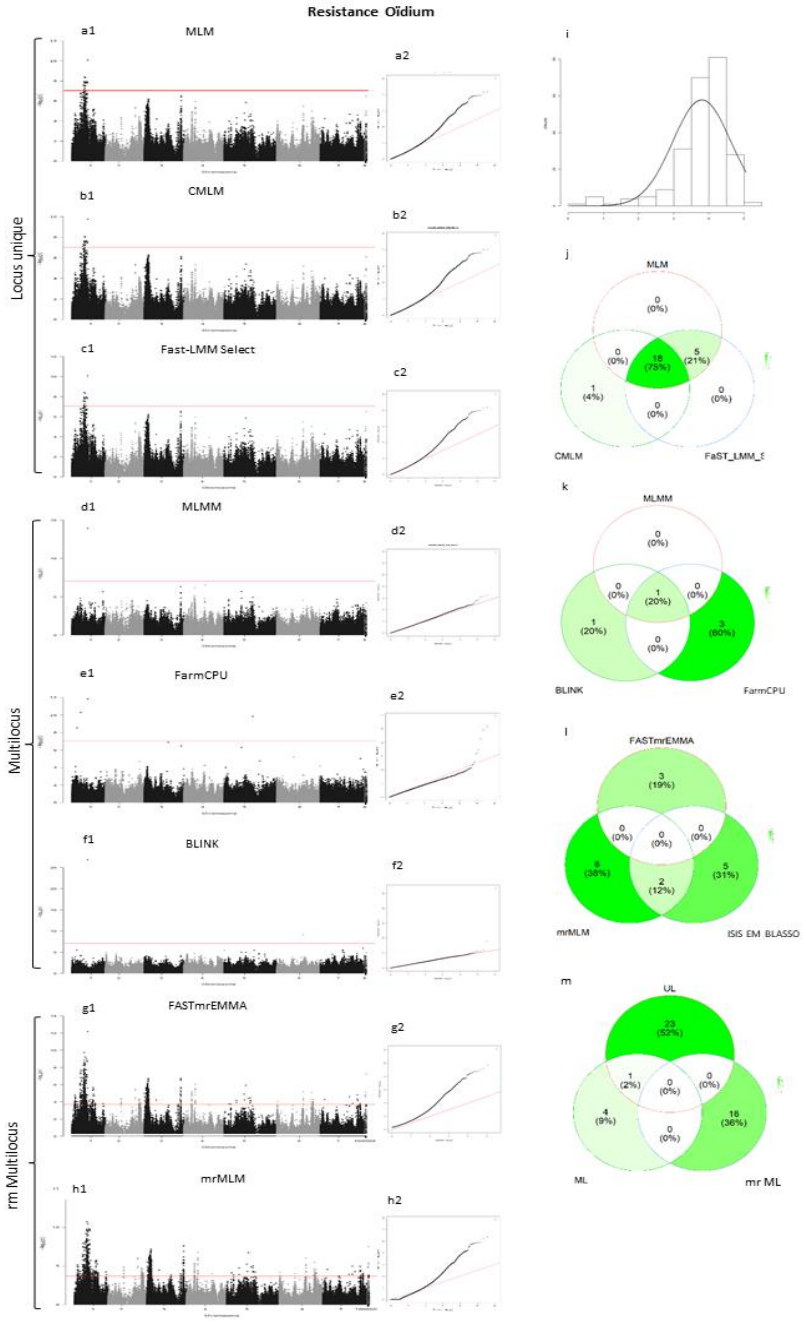


Figure 10 : Comparaison des méthodes et approches GWAS pour la résistance à l'oïdium
a1, b1, c1 correspondent respectivement aux Manhattan plot des méthodes de l'approche GWAS à locus unique (ULMLM) : MLM, Fast-LMM-Select et CMLM ; **d1, e1, f1**, correspondent respectivement aux Manhattan plot des méthodes de l'approche GWAS multilocus à effet SNP fixe ou partiellement fixe (MLMLM) : MLMM, FarmCPU et BLINK ; **g1, h1** correspondent respectivement aux Manhattan plot des méthodes de l'approche GWAS multilocus à effet SNP aléatoire (mrMLM) : FASTAmrEMMA et mrMLM ; **a2, b2, c2** correspondent respectivement aux qqplot des méthodes de l'approche ULMLM : MLM, Fast-LMM-Select et CMLM ; ; **d2, e2, f2** correspondent respectivement aux qqplot des méthodes de l'approche MLMLM : MLMM, FarmCPU et BLINK ; **g2, h2** correspondent respectivement aux qqplot des méthodes de l'approche mrMLM : FASTAmrEMMA et mrMLM ; **i** : histogramme de distribution de la variable résistance à l'oïdium ; **j** : diagramme de Venn pour les trois méthodes de l'approche ULMLM ; **k** : diagramme de Venn pour les trois méthodes de l'approche MLMLM ; **l** : diagramme de Venn pour les deux méthodes de l'approche mrMLM ; **m** : diagramme de Venn comparatif des trois approches UMLM (UL), MLMLM (ML) et mrMLM (mrML).

Les résultats obtenus (Tableau 7, Figure 11) ont montré que, pour la couleur des fleurs, en passant de la première version (V1) à la deuxième version (V2), le nombre de SNP identifiés par la méthode CMLM diminue de 63 à 26. Aussi, le deuxième pic de SNP significatifs sur le chromosome 6 de la V1, disparait avec la V2 (Figure 11.a1-a2). Avec la méthode FarmCPU, j'ai observé une disparition des SNP significatifs sur les chromosomes 3 et chromosomes 7 et sur les scaffolds, mais le nombre de SNP sur le chromosome 6 reste le même (Figure 11.c1-c2).

Tableau 6 : Résumé des résultats GWAS obtenus avec 3 approches et 9 méthodes différentes de GWAS à locus unique et multilocus

Trait	Approche	Méthode GWAS	Chr1	Chr2	Chr3	Chr4	Chr5	Chr6	Chr7	Scaffold	Total SNP	Max r^2	Cum r^2	Max effet
Col_Feuil_SortiHv_ORS17	ULMLM	MLM	0	0	0	0	0	0	0	0	0	-	-	-
		CMLM	0	0	0	0	5	0	0	0	5	-	-	-
		FaST-LMM-Select	0	0	0	0	0	0	0	0	0	-	-	-
	MLMLM	MLMM	0	0	0	0	1	0	0	0	1	-	-	-
		FarmCPU	1	0	0	0	4	0	0	0	5	-	-	-
		BLINK	1	0	0	0	3	0	0	0	4	-	-	-
	mrMLM	mrMLM	4	1	2	1	3	1	1	2	15	51.65	90.56	0.66
		FASTmrEMMA	1	0	0	1	1	2	0	1	6	8.17	22.47	0.45
		ISIS EM-BLASSO	0	0	0	0	3	1	1	0	5	4.16	12.71	0.27
Col_fleur	ULMLM	MLM	0	0	0	0	0	303	0	0	303	-	-	-
		CMLM	0	0	0	0	0	63	0	0	63	-	-	-
		FaST-LMM-Select	0	0	0	0	0	303	0	0	303	-	-	-
	MLMLM	MLMM	0	0	0	1	1	4	1	1	8	-	-	-
		FarmCPU	3	0	1	3	3	4	2	1	17	-	-	-
		BLINK	3	0	2	1	3	10	0	3	22	-	-	-
	mrMLM	mrMLM	0	3	1	2	1	13	2	0	22	14.05	46.64	0.20
		FASTmrEMMA	0	4	0	0	0	41	0	0	45	22.35	49.86	0.00
		ISIS EM-BLASSO	1	0	0	0	3	2	2	0	8	31.48	38.37	0.30
Deb_Flor_semAut_MSS17	ULMLM	MLM	0	0	0	0	39	0	0	0	39	-	-	-
		CMLM	0	0	0	0	18	0	0	0	18	-	-	-
		FaST-LMM-Select	0	0	0	0	39	0	0	0	39	-	-	-
	MLMLM	MLMM	0	0	0	0	1	0	0	0	1	-	-	-
		FarmCPU	1	1	1	0	1	0	2	0	6	-	-	-
		BLINK	0	0	1	0	2	0	0	0	3	-	-	-
	mrMLM	mrMLM	0	0	1	0	6	2	3	1	13	7.69	52.76	3.85
		FASTmrEMMA	0	0	0	0	1	0	0	0	1	21.80	21.80	13.19
		ISIS EM-BLASSO	0	0	0	0	4	1	0	0	5	16.39	24.78	5.57
Dega_gel_SortiHv_ORS17	ULMLM	MLM	0	0	0	0	0	0	0	0	0	-	-	-
		CMLM	0	0	0	0	0	0	0	0	0	-	-	-
		FaST-LMM-Select	0	0	0	0	0	0	0	0	0	-	-	-
	MLMLM	MLMM	1	0	0	0	0	1	0	0	2	-	-	-
		FarmCPU	2	0	1	0	1	0	0	0	4	-	-	-
		BLINK	1	1	0	0	0	0	0	0	2	-	-	-
	mrMLM	mrMLM	1	1	3	3	0	2	1	1	12	23.98	66.50	0.89
		FASTmrEMMA	0	0	1	0	2	0	1	0	4	7.89	9.59	1.06
		ISIS EM-BLASSO	2	1	0	1	2	0	0	1	7	2.08	10.03	0.32
Haut_Fin_Flor_B15	ULMLM	MLM	3	0	0	0	116	0	0	0	119	-	-	-
		CMLM	2	0	0	0	95	0	0	0	97	-	-	-
		FaST-LMM-Select	3	0	0	0	106	0	0	0	109	-	-	-
	MLMLM	MLMM	1	0	0	0	1	0	0	0	2	-	-	-
		FarmCPU	0	1	0	0	1	1	0	1	4	-	-	-
		BLINK	0	0	1	0	1	0	0	0	2	-	-	-
	mrMLM	mrMLM	3	1	0	1	4	2	1	0	12	37.26	80.50	19.53
		FASTmrEMMA	1	0	1	1	2	1	0	0	6	55.68	72.57	50.78
		ISIS EM-BLASSO	0	1	0	0	4	4	1	0	10	28.51	59.28	4.59
Nbr_Tig_DebFlor_ORS17	ULMLM	MLM	0	2	0	0	0	0	0	0	2	-	-	-
		CMLM	0	0	0	0	0	0	0	0	0	-	-	-
		FaST-LMM-Select	0	2	0	0	0	0	0	0	2	-	-	-
	MLMLM	MLMM	0	1	0	0	0	0	0	0	1	-	-	-
		FarmCPU	1	1	0	0	0	0	0	0	2	-	-	-
		BLINK	1	0	0	0	0	0	0	0	1	-	-	-
	mrMLM	mrMLM	1	1	0	0	0	0	0	2	4	29.45	63.70	0.45
		FASTmrEMMA	0	1	1	0	1	0	0	0	3	4.96	11.37	0.79
		ISIS EM-BLASSO	0	2	1	0	2	2	1	2	10	8.68	34.20	0.37
PMG_B15	ULMLM	MLM	0	0	0	0	0	0	0	0	0	-	-	-
		CMLM	0	0	0	0	0	0	0	0	0	-	-	-
		FaST-LMM-Select	0	0	0	0	0	0	0	0	0	-	-	-
	MLMLM	MLMM	0	2	0	0	0	0	0	0	2	-	-	-
		FarmCPU	0	1	1	1	2	0	1	0	6	-	-	-
		BLINK	0	2	0	2	2	0	3	1	10	-	-	-
	mrMLM	mrMLM	2	5	3	5	8	2	4	0	29	7.21	42.32	17.84
		FASTmrEMMA	1	3	1	1	3	1	0	0	10	5.09	25.77	43.34
		ISIS EM-BLASSO	0	3	2	0	2	0	0	1	8	4.07	14.47	7.46

Trait	Approche	Méthode GWAS	Chr1	Chr2	Chr3	Chr4	Chr5	Chr6	Chr7	Scaffold	Total SNP	Max r^2	Cum r^2	Max effet
Pourc_Gr_Bruch_B16	ULMLM	MLM	0	0	31	0	0	0	0	0	31	-	-	-
		CMLM	0	0	29	0	0	0	0	0	29	-	-	-
		FaST-LMM-Select	0	0	31	0	0	0	0	0	31	-	-	-
	MLMLM	MLMM	0	0	1	0	0	0	0	0	1	-	-	-
		FarmCPU	1	0	2	0	0	0	1	0	4	-	-	-
		BLINK	0	0	0	0	0	0	1	0	1	-	-	-
	mrMLM	mrMLM	0	3	1	1	2	1	5	0	13	13.48	62.91	3.28
		FASTmrEMMA	0	0	0	1	0	0	0	0	1	0.00	0.00	0.00
		ISIS EM-BLASSO	0	0	3	0	0	0	2	0	5	16.48	35.72	2.68
Resist_Oid_MG15	ULMLM	MLM	23	0	0	0	0	0	0	0	23	-	-	-
		CMLM	19	0	0	0	0	0	0	0	19	-	-	-
		FaST-LMM-Select	23	0	0	0	0	0	0	0	23	-	-	-
	MLMLM	MLMM	1	0	0	0	0	0	0	0	1	-	-	-
		FarmCPU	3	0	0	0	1	0	0	0	4	-	-	-
		BLINK	1	0	0	0	0	1	0	0	2	-	-	-
	mrMLM	mrMLM	3	0	1	2	1	1	0	0	8	24.01	67.21	0.66
		FASTmrEMMA	0	1	1	0	0	0	0	1	3	6.34	10.72	0.47
		ISIS EM-BLASSO	0	0	3	1	0	3	0	0	7	4.49	22.30	0.40
Tx_Mn_Gr_B15	ULMLM	MLM	0	0	0	0	0	0	0	0	0	-	-	-
		CMLM	0	0	0	0	0	0	0	0	0	-	-	-
		FaST-LMM-Select	0	0	0	0	0	0	0	0	0	-	-	-
	MLMLM	MLMM	0	0	0	0	0	0	0	0	0	-	-	-
		FarmCPU	0	0	1	1	1	0	0	0	3	-	-	-
		BLINK	0	0	0	0	0	0	0	0	0	-	-	-
	mrMLM	mrMLM	0	0	0	3	0	0	0	0	3	16.70	32.21	0.75
		FASTmrEMMA	0	0	0	0	1	0	0	0	1	10.73	10.73	1.31
		ISIS EM-BLASSO	0	0	2	1	2	2	0	0	7	11.31	38.27	1.16
Volum_Gr_Bruch_B16	ULMLM	MLM	0	0	0	0	0	0	0	0	0	-	-	-
		CMLM	0	0	0	0	0	0	0	0	0	-	-	-
		FaST-LMM-Select	0	0	0	0	0	0	0	0	0	-	-	-
	MLMLM	MLMM	0	0	0	0	0	0	0	0	0	-	-	-
		FarmCPU	0	0	0	0	0	0	0	0	0	-	-	-
		BLINK	0	0	0	0	0	0	0	0	0	-	-	-
	mrMLM	mrMLM	4	1	2	0	3	1	1	1	13	15.72	78.14	23.27
		FASTmrEMMA	0	1	1	0	1	0	0	0	3	4.96	11.37	0.79
		ISIS EM-BLASSO	0	0	2	2	3	1	1	0	9	7.43	28.96	8.15

ULMLM : approche GWAS à locus unique ; **MLMLM** : approche GWAS multilocus à effet SNP fixe ou partiellement fixe ; **mrMLM** : approche GWAS multilocus à effet SNP aléatoire ; **Chr1, Chr2, Chr3, Chr4, Chr5, Chr6, Chr7** correspondent aux sept chromosomes qui constituent le génome du pois – les colonnes contiennent les nombres de SNP significatifs par chromosome ; **Scaffold** : les SNP non assignés à des chromosomes ; **Total SNP** : nombre total de SNP significatifs identifiés ; **Max_r^2** : la part de la variance phénotypique expliquée par le marqueur le plus significatif identifié par la méthode ; **Cum_r^2** : la part de la variance phénotypique expliquée par l'ensemble des SNP identifiés par la méthode ; **Max_effet** : l'effet du marqueur le plus significatif identifié par la méthode.

Pour les dégâts dus au gel, avec la méthode CMLM, aucun changement n'est observé entre la V1 et la V2 (Tableau 7). Avec la méthode FarmCPU, j'ai noté une augmentation du nombre de SNP significatifs en passant de la V1 à la V2. J'ai également noté, la disparition du SNP significatif identifié au niveau des scaffolds de la V1 par la méthode mrMLM.
Pour la hauteur des plantes, avec la méthode CMLM (Figure 11.b1-b2), j'ai observé sur la V2, une disparition des deux SNP significatifs identifiés sur le chromosome 2 de la V1, concentrant ainsi, l'ensemble des SNP significatifs sur le chromosome 5 pour la V2.
Pour la résistance à l'oïdium (Tableau 7), j'ai noté une diminution du nombre de SNP significatifs sur le chromosome 1 en passant de la V1 à la V2, mais par contre des SNP supplémentaires sont identifiés sur les chromosomes 3, 4, 5 et 7.

3.1.8. Résultats de GWAS sur la résistance aux pucerons

Pour comparer mes résultats avec les connaissances disponibles dans la bibliographie, j'ai choisi de travailler sur la résistance aux pucerons. Avec la meilleure méthode identifiée sur chacune des trois approches étudiées (CMLM, FarmCPU et mrMLM), j'ai analysé sur les deux versions du génome les données des deux variables de fécondité des pucerons (Fecond_Puc_ArPo28 et Fecond_Puc_LSR1) générées par Ollivier et al., 2022 sur le panel AMS.

Sur la V1 pour le clone ArPo28 (biotype du pois), j'ai identifié 16 SNP significatifs (Annexe 3) dont quatre sur le chromosome 2 (chr2), trois sur le chr3, un sur le chr4, trois sur le chr5, un sur le chr6, trois sur le chr7 et un sur les scafolds). En termes de méthodes, sur ces 16 SNP significatifs, trois ont été identifiés par FarmCPU (1 sur ch2, un sur chr4 et un sur chr5) et 13 par mrMLM (3 sur chr2, 3 sur chr3, 2 sur chr5, un sur chr6, 3 sur chr7 et un sur les scaffolds). CMLM n'a pas identifié de SNP significatifs pour ce trait. Aucun SNP en commun n'a été identifié par les trois méthodes. Le SNP Pisum_contig067893_11297 identifié à la position 423Mb du chr2 et le SNP Pisum_contig180166_5326 identifié à la position 212Mb du chr3 avaient les meilleures valeurs de r^2 avec respectivement 10,5% et 11,5%. Ces deux SNP ont été identifiés par la méthode mrMLM. Toujours sur la V1, avec le clone LSR1 (biotype luzerne), j'ai identifié aussi 16 SNP significatifs, dont un sur le chr1, un sur le chr2, un sur le chr3, un sur le chr5, un sur le chr6 et 11 sur le chr7. En termes de méthodes, CMLM a identifié trois SNP significatifs tous sur le chr7, FarmCPU a identifié quatre SNP significatifs dont un sur le chr1 et trois sur le chr7, mrMLM a identifié 9 SNP significatifs dont un sur le chr2, un sur le chr3, un sur le chr5, un sur le chr6 et 5 sur le chr7. Les méthodes CMLM et FarmCPU ont identifié en commun un SNP significatif à la position 148Mb du chr7 (Pisum_contig115828_65327). Le SNP Pisum_contig115828_22928 identifié à la position 148Mb du chr7 avait la valeur de r^2 la plus élevée (14,65%).

Pour le clone ArPo28 en passant de la V1 à la V2, les SNP significatifs ont été identifiés dans les mêmes chromosomes. Il a été noté une diminution du nombre de SNP significatif passant de 16 SNP à 12 SNP, avec particulièrement la disparition du SNP identifié dans les scaffolds de la V1 et la réduction du nombre de SNP identifiés sur le chr3. Contrairement au clone ArP28, pour le clone LSR1, le nombre de SNP significatifs a augmenté de 16 SNP à 17 SNP en passant de la V1 à la V2.

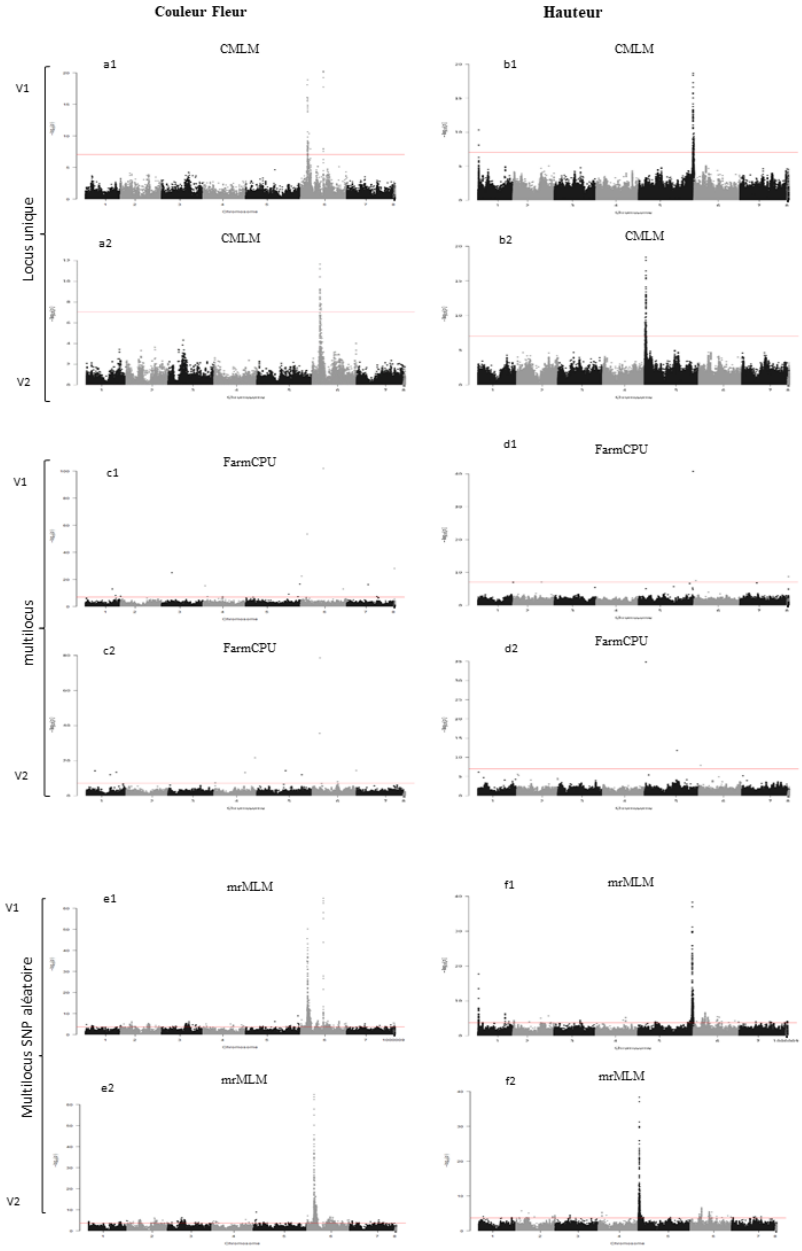


Figure 11: Résumé comparative des résultats de GWAS des deux versions du génome du pois pour la couleur des fleurs et la hauteur des plantes

a1, a2 respectivement les figures Manhattan de la première (V1) et de la deuxième version (V2) du génome du pois, pour la couleur des fleurs, obtenues avec la méthode GWAS à locus unique CMLM ; **b1, b2** respectivement les figures Manhattan de la V1 et de la V2, pour la hauteur des plantes, obtenus avec CMLM ; **c1, c2** respectivement les figures Manhattan de la V1 et de la V2, pour la couleur des fleurs, obtenues avec la méthode GWAS multilocus à effet SNP fixe FarmCPU ; **d1, d2** respectivement les figures Manhattan de la V1 et de la V2, pour la hauteur des plantes, obtenus avec FarmCPU ; **e1, e2** respectivement les figures Manhattan de la V1 et de la V2, pour la couleur des fleurs, obtenues avec la méthode GWAS multilocus à effet SNP aléatoire mrMLM ; **f1, f2** respectivement les figures Manhattan de la V1 et de la V2, pour la hauteur des plantes, obtenus avec mrMLM. Pour l'ensemble des traits, on peut observer une inversion des pics de SNP significatifs sur le chromosome 5 en passant de la V1 à la V2, cela s'explique par l'inversion qui a été réalisé lors de l'assemblage de la V2.

3.1.9. Robustesse régions génomiques détectées et lien avec les gènes causaux

Afin de vérifier la robustesse des pics détectés par les différentes méthodes, j'ai comparé les régions génomiques sous-jacentes aux pics d'association avec la position de gènes connus pour contrôler les traits étudiés. J'ai pu retrouver plusieurs fois ces gènes à proximité des pics d'association détectés.

Ainsi, plusieurs pics d'association avec une p-value très significative pour les variables hauteur de la plante colocalisent aux alentours de 13Mb sur le chromosome 5 (v2 du génome) à proximité du gène *Le*. Ce gène, qui code pour une gibberellin 3 beta-hydroxylase a été caractérisé pour ses mutations (identifiées par Mendel) qui affectent la longueur des entre-noeuds (Martin et al, 1997). De même, des pics d'association liés à la couleur des fleurs et à la moucheture des grains colocalisent sur le chromosome 6 autour de 82 Mb à proximité du gène *A* étudié aussi par Mendel et identifié comme un facteur de transcription qui gouverne la couleur blanche de la fleur (Hellens et al.,2010).

Comme autre exemple, un pic d'association pour la résistance à l'Oidium a été détecté sur le chromosome 1 autour de 281Mb à proximité du gène *Mlo1* dont les mutations sont connues pour conférer une résistance à cette maladie (Sulima et Zhukov, 2022).

Tableau 7 : Comparaison des résultats de GWAS obtenus en fonction de la version d'assemblage du génome du pois

Trait	Méthode GWAS	Génome	Chr1	chr2	Chr3	Chr4	Chr5	Chr6	Chr7	Scaffold	Total SNP	Max r^2	Cumul r^2	Max effet
Col_fleur	CMLM	V1	0	0	0	0	0	63	0	0	63	-	-	-
		V2	0	0	0	0	0	26	0	0	26	-	-	-
	FarmCPU	V1	3	0	1	3	3	4	2	1	17	-	-	-
		V2	3	0	0	3	2	4	0	0	11	-	-	-
	mrMLM	V1	0	3	1	2	1	13	2	0	22	14.05	46.64	0.2
		V2	0	2	0	1	2	10	2	0	17	11.07	29.03	0.18
Dega_gel_SortiHv_ORS17	CMLM	V1	0	0	0	0	0	0	0	0	0	-	-	-
		V2	0	0	0	0	0	0	0	0	0	-	-	-
	FarmCPU	V1	2	0	1	0	1	0	0	0	4	-	-	-
		V2	3	1	2	0	2	0	0	0	8	-	-	-
	mrMLM	V1	1	1	3	3	0	2	1	1	12	23.98	66.5	0.89
		V2	1	0	2	3	2	1	0	0	9	25.37	56.13	0.91
Haut_Fin_Flor_B15	CMLM	V1	2	0	0	0	95	0	0	0	97	-	-	-
		V2	0	0	0	0	86	0	0	0	86	-	-	-
	FarmCPU	V1	0	1	0	0	1	1	0	1	4	-	-	-
		V2	0	0	0	0	2	1	0	0	3	-	-	-
	mrMLM	V1	3	1	0	1	4	2	1	0	12	37.26	80.5	19.53
		V2	2	1	0	0	4	2	1	0	10	42.34	81.41	21.37
PMG_B15	CMLM	V1	0	0	0	0	0	0	0	0	0	-	-	-
		V2	0	0	0	0	0	0	0	0	0	-	-	-
	FarmCPU	V1	0	1	1	1	2	0	1	0	6	-	-	-
		V2	1	2	1	1	2	0	1	0	8	-	-	-
	mrMLM	V1	2	5	3	5	8	2	4	0	29	7.21	42.32	17.84
		V2	4	5	5	1	9	2	4	0	30	8.07	45.19	18.87
Resist_Oid_MG15	CMLM	V1	19	0	0	0	0	0	0	0	19	-	-	-
		V2	18	0	0	0	0	0	0	0	18	-	-	-
	FarmCPU	V1	3	0	0	0	1	0	0	0	4	-	-	-
		V2	2	0	1	1	0	0	0	0	4	-	-	-
	mrMLM	V1	3	0	1	2	1	1	0	0	8	24.01	67.21	0.66
		V2	2	0	2	2	3	0	1	0	10	18.41	70.26	0.94
Tx_Mn_Gr_B15	CMLM	V1	0	0	0	0	0	0	0	0	0	-	-	-
		V2	0	0	0	0	0	0	0	0	0	-	-	-
	FarmCPU	V1	0	0	1	1	1	0	0	0	3	-	-	-
		V2	0	0	1	0	1	0	1	0	3	-	-	-
	mrMLM	V1	0	0	0	3	0	0	0	0	3	16.7	32.21	0.75
		V2	2	0	1	1	1	0	0	0	5	15.43	44.92	1.49

Chr1, Chr2, Chr3, Chr4, Chr5, Chr6, Chr7 correspondent aux sept chromosomes que contient le génome du pois ;**Scaffold** : les SNP non assignés à des chromosomes ; **Total SNP** : nombre totale de SNP significatifs identifiés par la méthode ; **Max_r^2** : la part de la variance phénotypique expliquée par le marqueur le plus significatif identifié par la méthode ; **Cum_r^2** : la part de la variance phénotypique expliquée par l'ensemble des SNP identifiés par la méthode ; **Max_effet** : l'effet du marqueur le plus significatif identifié par la méthode.

4. DISCUSSION

La stratification d'un panel de GWAS peut dépendre du trait étudié

Pour un même panel d'individus et les mêmes données génotypiques, il peut exister des différences entre les traits observés ou mesurés, sur la nécessité de prendre en compte ou pas et le niveau de stratification (Q) comme covariable dans le modèle GWAS (Wang et Zhang, 2021). A ma connaissance, pour la plupart des études précédentes de cartographie d'association chez le pois (Ollivier et al., 2022 ; Cartelier, 2021 ; Beji et al., 2020 ; Dissanayaka et al., 2020 ; Gali et al., 2019 ; Desgroux et al., 2018 ; Desgroux et al., 2016), le nombre de groupes de structuration a été utilisé pour représenter la stratification (Q) dans le modèle, pour l'ensemble des traits étudiés, sans examiner au préalable, le niveau de stratification le mieux adapté à chaque trait. Les résultats que j'ai obtenu ont montré que pour l'ensemble des 63 variables étudiées le niveau de stratification optimal n'était pas le même. La plupart des variables (63,49%) n'ont même pas nécessité d'inclure la matrice Q comme covariable dans le modèle. Aussi, le nombre d'axes de PC optimal ne dépassait pas deux (3 groupes). Ollivier et al.(2022) ont identifiés trois grands groupes génétiques dans le panel AMS, ce qui peut correspondre aux groupes révélés par les deux axes de PC. Mes résultats sont cohérence, vu que pour chacun des traits répétés sur plusieurs sites et/ou sur plusieurs années, les variables correspondantes avaient toujours le même niveau optimal de stratification.

La méthode de calcul de la matrice de parenté peut améliorer l'ajustement du modèle de GWAS

Des études ont montré que, une confusion survient en raison de la structure (stratification et matrice de parenté) de la population, en particulier si elle est corrélée avec le trait étudié (Lui et al., 2016 ; Larsson et al., 2013 ; Mir et al., 2012 ; Jaiswal et al., 2012 ; Zhang et al., 2010 ; Thornsberry et al., 2001). Dans la présente étude, en plus de la prise en compte du niveau de stratification optimal pour chaque variable, la sélection de modèles a permis d'aborder ce problème de la structure de la population liée aux traits. Les résultats ont montré que pour l'ensemble des variables testées, nécessitant ou pas une stratification, l'inclusion de la matrice de parenté (K) avec l'approche de calcul intégrée dans la méthode CMLM (Zhang et al., 2010) améliore l'ajustement du modèle GWAS, ce qui permet d'espérer un meilleur niveau de confiance dans nos résultats. Ce résultat est en phase avec des études précédentes sur le pois (Beiji et al., 2020), le Soja (Kaler et al., 2020 ; Zhao et al., 2017) et d'autres espèces (Liu et al., 2020 ; Li et al., 2018 ; Lui et al., 2016 ; Li et al., 2014 ; Yu et al., 2006) qui ont suggéré que, les modèles incorporant à la fois la stratification et la parenté fonctionnent mieux que lorsqu'ils les incluent séparément.

Par ailleurs, les différences de résultats pour les méthodes MLM (Yu et al., 2006), CMLM (Zhang et al., 2010) et Fast_LMM_Select (Listgarten et al., 2012) qui utilisent toutes la matrice de parenté (K), peut résider dans la différence de leur approche de calcul de cette matrice. Cependant, on peut supposer que pour le panel AMS, le calcul de la matrice de parenté par groupe est plus adéquat que le calcul par paire d'individus (Beiji et al., 2019). Également, les différences entre Fast-LMM-Select et CMLM utilisant toutes deux une matrice K de rang réduit, peut-être dû au fait que, l'utilisation dans Fast-LMM-Select d'un intervalle arbitraire de 2 cM comme seuil d'exclusion pour un LD entre les pseudo QTN et le marqueur de test est moins efficace chez le pois que la méthode implémentée dans CMLM qui utilise un algorithme de regroupement pour diviser les individus en groupes basés sur des génotypes similaires et ensuite, utilise un résumé de la parenté au sein et entre les groupes comme matrice de parenté réduite lors de la résolution du modèle MLM (Zhang et al., 2010).

Les méthodes des approches multilocus ont permis d'identifier des loci non détéctés par les méthodes de l'approche à locus unique

Dans notre étude comparative de méthodes et approche GWAS, même si aucune méthode ou approche n'a été meilleure sur toutes les variables analysées, certaines méthodes ou approches se sont très bien distinguées. Ainsi, nos résultats ont montré que, au sein de l'approche GWAS à locus unique (ULMLM), par rapport aux méthodes MLM et Fast-LMM-Select, la méthode CMLM avait un meilleur ajustement de modèle et est parvenue à trouver des loci pour certaines variables avec lesquelles ces dernières n'ont rien identifié. Cette amélioration des résultats de GWAS par CMLM par rapport aux autres méthodes ULMLM a été suggérée dans des études comparatives de méthodes GWAS (Kaler et al., 2020 ; Zhang et al., 2010).

Concernant les méthodes de l'approche multilocus à effet marqueurs fixe ou partiellement fixe (MLMLM), toutes les méthodes avaient des niveaux d'ajustement assez bons. FarmCPU a permis d'identifier plus de loci et sur des régions et/ou des traits avec lesquels, les méthodes BLINK et MLMM n'ont pas réussi (exemple : teneur en manganèse dans les graines), suggérant ainsi de meilleures performances dans l'identification des loci liés aux traits. Ces résultats sont en phase avec ceux de plusieurs études comparatives de méthodes GWAS multilocus (Kaler et al., 2020 ; Lui et al., 2016 ; Zhao et al., 2017 ; Liu et al., 2020).

Au sein des méthodes multilocus à effet marqueurs aléatoires (mrMLM), la méthode mrMLM s'est distinguée, en termes de nombre de loci identifiés, du nombre de régions génomiques où des loci ont été identifiés et de la part de variance phénotypique expliquée par les marqueurs identifiés. Des résultats similaires ont été rapportés par Chang et al. (2018) et Wang et al. (2016).

Avec les résultats comparatifs des trois approches GWAS, nous avons globalement observé que, pour les traits simples (qualitatif ou quantitatif simple), les méthodes de l'approche ULMLM ont permis de détecter plus de loci significatifs avec des pics de SNP assez soutenus, alors que pour certains traits plus complexes (poids de mille grains, teneur de manganèse dans les grains, dégâts dus au gel, etc.), ces méthodes ULMLM ont été incapables d'identifier des loci significatifs. A côté, les méthodes multilocus MLMLM, plus particulièrement FarmCPU, n'ont pas identifié de pics de SNP, mais des SNP individuels plus significatifs et sur plusieurs chromosomes, même pour les traits complexes avec lesquels les méthodes à locus unique n'ont pas réussi à révéler des associations marqueurs-trait. En général, en GWAS, la notion de QTL fait référence au signal identifié par des méthodes à locus unique. De tels QTL, contient majoritairement de nombreux SNP associés au trait. Alors que dans les méthodes GWAS multilocus, lorsque tous les marqueurs potentiellement associés sont identifiés dans la première étape, ils seront soumis à un modèle multilocus pour une analyse plus approfondie et les vrais QTN sont confirmés par le test du rapport de vraisemblance (Berhe et al., 2021 ; Kaler., 2020 ; Wang et al., 2016 ; Liu et al., 2016a ; Segura et al., 2012). L'approche mrMLM, plus particulièrement la méthode mrMLM, ont réussi à identifier des loci significatifs dans tous les traits même les plus complexes (volume grains bruchés), pour lesquels les méthodes des approches ULMLM et MLMLM n'ont pas réussi à identifier des associations marqueurs-trait significatifs.

Concernant la puissance statistique, les analyses des qqplot ont montré que les méthodes de l'approche MLMLM avaient de meilleurs ajustements par rapport aux méthodes des approches ULMLM et mrMLM. Dans la littérature, plusieurs études ont confirmé que les méthodes de l'approche MLMLM étaient plus puissantes que les méthodes de ULMLM (Kaler et al., 2020 Huang et al., 2019 ; Liu et al., 2016 ; Segura et al., 2012). Par ailleurs, ce meilleur comportement des qqplot des méthodes de l'approche MLMLM par rapport aux méthodes de l'approche mrMLM ne traduit pas forcément leur meilleure puissance, car dans plusieurs études comparatives de méthodes MLMLM et mrMLM, les méthodes MLMLM avaient un meilleur comportement de qqplot, mais des calculs de puissances ont montré que les méthodes mrMLM étaient plus puissantes (Tamba and Zhang, 2018 ; Wen et al., 2018 ; Tamba et al., 2017 ; Zhang et al., 2017). Cette aberration sur le comportement des qqplot par rapport à la puissance statistique calculée des méthodes mrMLM, se retrouve aussi au niveau des Manhattan plot, dont certains SNP dépassant le seuil de significativité ne sont pas retenus comme significatifs dans les résultats finaux. Il est à supposer que ceci est dû à la particularité des calculs bayésiens qui utilisent les méthodes mrMLM, par rapport aux méthodes connues dans les autres approches.

Également, les résultats comparatifs de méthodes des trois approches ont montré que, en dehors des SNP identifiés en commun, pour plusieurs traits aussi bien simples que complexes, des SNP différents, dans des régions génomiques différentes ont été identifiés par les trois approches, indiquant ainsi une complémentarité de ces trois approches dans l'identification des associations marqueurs-traits. Globalement, dans plusieurs études antérieures, même si, il a été constaté que les modèles multilocus étaient plus efficaces et puissants que les modèles à locus unique pour détecter des résultats d'association hautement significatifs pour les traits d'intérêt (Berhe et al., 2021 ; Liu et al., 2020 ; Abed and Belzile, 2019 ; Zhang et al., 2019 ; Cui et al., 2018 ; Peng et al., 2018 ; Li et al., 2018 ; Su et al., 2018 ; Xu et al., 2018 ; Zhao et al., 2017), et que des SNP significatifs détectés par des méthodes à locus unique et validés expérimentalement comme biologiquement pertinents peuvent ne pas être du tout détectés par des méthodes multilocus (Kaler et al., 2020 ; Liu et al., 2020 ; Li et al., 2018 ; Yang et al., 2018 ; Liu et al., 2016b). Ces différences de résultats seraient dues aux différences dans les détails des méthodes statistiques utilisées (Cortes et al., 2021).

La nouvelle version du génome du pois améliore les résultats de la GWAS

Une deuxième version non-publiée et non-annotée du génome du pois a été assemblée dans le laboratoire où j'ai réalisé mon stage. Par rapport à la première version du génome (V1) (Kreplak et al., 2019) obtenue à partir de short-reads, cette deuxième version utilise la technologie de long-reads ONT d'Oxford Nanopore. Son N50 est cinq cent quatre-vingts fois plus grand que la V1. Cette meilleure continuité permet un meilleur ordonnancement des gènes sur les chromosomes.

Après un filtre avec une MAF à 5%, il a été constaté que le nombre de SNP était légèrement inférieur sur la V2 (532051 SNP) par rapport à la V1 (548.753 SNP), mais le taux de SNP non placés dans les sept chromosomes (Scaffold) était largement supérieur sur la V1 (2,54%) par rapport à la V2 (0,23%). La comparaison des résultats de GWAS réalisées sur les deux génomes avec différents traits simples et complexes a permis de voir que, globalement le nombre de SNP identifiés diminue en passant de la V1 à la V2. Néanmoins, avec l'analyse des figures de Manhattan, on observe une diminution des bruits de fond avec la V2 par rapport à la V1. Par exemple, pour la hauteur des plantes, les deux pics de SNP identifiés sur le chromosome 5 avec la V1, se sont unifiés en un seul sur la V2. Il a été observé aussi, une diminution des SNP significatifs localisés sur les Scaffolds en passant de la V1 à la V2, ce qui suggère que la V2 permet de replacer sur le génome les loci liés aux traits. Par ailleurs, il a été également observé avec l'approche mrMLM que, les marqueurs qui participaient plus à la variabilité phénotypique observé (plus grand r^2) ont été identifiés sur la V2 par rapport à la V1.

Notre étude GWAS a permis de confirmer des hypothèses sur les bases génétiques des traits d'intérêt chez le pois

Avec les trois meilleures méthodes (CMLM, FarmCPU et mrMLM), j'ai repris sur la V1, l'analyse des données de résistance à deux clones de pucerons (ArPo28 et LSR1) de Ollivier et al. (2022) utilisant également le même panel et les mêmes données de génotypage. Ainsi, comme résultats, pour le clone ArPo28 (biotype du pois), j'ai identifié 16 SNP significatifs dont quatre sur le chromosome 2 (chr2), trois sur le chr3, un sur le chr4, trois sur le chr5, un sur le chr6, trois sur le chr7 (entre les positions 141 et 204 Mb et un sur les scafolds). La méthode mrMLM a permis l'identification de 13 des 16 SNP significatifs. Avec le clone LSR1 (biotype de la luzerne), j'ai identifié aussi 16 SNP significatifs dont un sur chacun des sept chromosomes sauf le chromosome 4 et 11 sur le chr7 (entre 6,5 Mb et 410 Mb). Dans leurs analyses utilisant les méthodes Fasta-LMM et MLMM, Ollivier et al. (2022) ont identifié des SNP significatifs uniquement sur le chr7, dont ceux liés à la résistance à ArPo28 entre les positions 129-157 Mb et ceux liés à la résistance à LSR1 entre les positions 128-161 Mb. Dans une autre étude, un QTL lié aux scores de fécondité des pucerons chez le pois sauvage (*P. fulvum)* (Barilli et al. 2020) était synthénique avec le chr6 du pois cultivé (*P. sativum),* alors que Ollivier et al. (2022) n'ont détecté aucun SNP lié à la résistance aux pucerons sur le chr6. Par conséquent, ce QTL peut être l'une des sources du niveau de résistance plus élevé de *P. fulvum* par rapport à *P. sativum* (Ollivier et al., 2022). Vu que dans mes analyses, en plus des SNP identifiés sur le chr7 couvrant les intervalles des SNP identifiés par Ollivier et al. 2022, avec la méthode mrMLM, j'ai aussi identifié des SNP significatifs sur le chromosome 6, dont un pour chacun des deux clones : le clone LSR1 (à 238 Mb) et un autre pour le clone ArPo28 (à 346 Mb). Cela suggère que, les méthodes utilisées lors de ce stage et les protocoles permettraient d'améliorer les résultats d'Ollivier et al. (2022), de les relier avec ceux de Barilli et al.(2020) et d'offrir de nouvelles cibles pour la recherche et le développement.

Également Ollivier et al. (2022) ont observé un lien entre la hauteur de la plante et la résistance au puceron, vu que les génotypes les plus grands avaient une meilleure résistance, ce qui suppose des régions participant à la résistance au puceron dans le chr5 (contenant le gène *Le*), mais ils n'ont pas identifié de SNP lié à la résistance aux pucerons sur le chr5 pour confirmer leur hypothèse. Dans mes résultats, d'abord avec l'analyse des données phénotypiques, j'ai observé une corrélation entre les variables de la hauteur et les variables de fécondité des deux clones de pucerons. Aussi avec la GWAS, j'ai identifié trois SNP significatifs sur le chromosome 5 avec le clone ArPo28 (entre 108 et 424 Mb) et un SNP significatif pour le clone LSR1 à la position (169 Mb). Ainsi, ces résultats me permettent de confirmer l'hypothèse d'une potentielle liaison entre la hauteur de la plante et la résistance aux pucerons.

5. CONCLUSION ET PERSPECTIVES

L'identification des déterminants génétiques contrôlant les traits en lien avec la phénologie, l'architecture, le rendement et la réponse aux stress chez le pois est une étape clé pour aller plus loin en recherche et en développement et pour accompagner l'innovation variétale. Ce travail a montré que l'optimisation de la GWAS en améliorant la prise en compte de la structuration du panel d'individus et des liens de parenté et en adaptant le choix des modèles statistiques influe nettement sur la performance globale. Il a également montré qu'un bon positionnement des marqueurs sur le génome (version d'assemblage 2 versus version d'assemblage 1 du génome) améliore la détection des loci associés aux traits. Même si cela n'a pas été présenté dans ce rapport, un nombre important de SNP significatifs est commun entre les différents traits analysés. Ceci est en phase avec les corrélations significatives observées entre certaines variables et ouvre la voie à révéler des régions génomiques porteuses de gène(s) contrôlant plusieurs traits à la fois. Ceci montre la nécessité d'explorer les analyses GWAS multi-trait et d'exploiter davantage les ressources génomiques disponibles pour mettre en évidence des gènes candidats sous-jacents aux loci significatifs.

Références bibliographiques

Abecasis, G.R., Cardon, L.R., Cookson, W.O.C., 2000. A General Test of Association for Quantitative Traits in Nuclear Families. The American Journal of Human Genetics 66, 279–292. https://doi.org/10.1086/302698

Abed, A., Belzile, F., 2019. Comparing Single-SNP, Multi-SNP, and Haplotype-Based Approaches in Association Studies for Major Traits in Barley. The Plant Genome 12, 190036. https://doi.org/10.3835/plantgenome2019.05.0036

Alboukadel, K., Mundt, F., 2020. factoextra: Extract and Visualize the Results of Multivariate Data Analyses. https://CRAN.R-project.org/package=factoextra

Aulchenko, Y.S., de Koning, D.-J., Haley, C., 2007. Genomewide Rapid Association Using Mixed Model and Regression: A Fast and Simple Method For Genomewide Pedigree-Based Quantitative Trait Loci Association Analysis. Genetics 177, 577–585. https://doi.org/10.1534/genetics.107.075614

Aznar-Fernández, T., Barilli, E., Cobos, M.J. et al., 2020. Identification of quantitative trait loci (QTL) controlling resistance to pea weevil (Bruchus pisorum) in a high-density integrated DArTseq SNP-based genetic map of pea. Sci Rep 10, 33 (2020). https://doi.org/10.1038/s41598-019-56987-7

Beji, S., Fontaine, V., Devaux, R., Thomas, M., Negro, S. S., Bahrman, N., Siol, M., Aubert, G., Burstin, J., Hilbert, J.-L., Delbreil, B., & Lejeune-Hénaut, I., 2020. Genome-wide association study identifies favorable SNP alleles and candidate genes for frost tolerance in pea. *BMC Genomics*, *21*(1), 536. https://doi.org/10.1186/s12864-020-06928-w

Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) 57, 289–300.

Berhe, M., Dossa, K., You, J., Mboup, P.A., Diallo, I.N., Diouf, D., Zhang, X., Wang, L., 2021. Genome-wide association study and its applications in the non-model crop Sesamum indicum. BMC Plant Biology 21, 283. https://doi.org/10.1186/s12870-021-03046-x

Bush, W.S., Moore, J.H., 2012. Chapter 11: Genome-Wide Association Studies. PLOS Computational Biology 8, e1002822. https://doi.org/10.1371/journal.pcbi.1002822

Buzdugan, L., Kalisch, M., Navarro, A., Schunk, D., Fehr, E., Bühlmann, P., 2016. Assessing statistical significance in multivariable genome wide association analysis. Bioinformatics 32, 1990–2000.

Cardon, L.R., Abecasis, G.R., 2003. Using haplotype blocks to map human complex trait loci. Trends Genet 19, 135–140.

Cartelier, K., 2021. *Déterminisme génétique de la plasticité de la composition protéique des graines de légumineuses vis-à-vis de l'environnement : Rôle du métabolisme du soufre* [These de doctorat, Bourgogne Franche-Comté]. http://www.theses.fr/2021UBFCK012

Chang, F., Guo, C., Sun, F., Zhang, J., Wang, Z., Kong, J., He, Q., Sharmin, R.A., Zhao, T., 2018. Genome-Wide Association Studies for Dynamic Plant Height and Number of Nodes on the Main Stem in Summer Sowing Soybeans. Frontiers in Plant Science 9.

Chen, H., Hao, Z., Zhao, Y., Yang, R., 2020. A fast-linear mixed model for genome-wide haplotype association analysis: application to agronomic traits in maize. BMC Genomics 21, 151. https://doi.org/10.1186/s12864-020-6552-x

Cortes, L.T., Zhang, Z., Yu, J., 2021. Status and prospects of genome-wide association studies in plants. The Plant Genome 14, e20077. https://doi.org/10.1002/tpg2.20077

Crews F. T., Collins M. A., Dlugos C., Littleton J., Wilkins L., Neafsey E. J., et al., 2004. Alcohol-induced neurodegeneration: when, where and why? Alcohol Clin. Exp. Res. 28 350–364.

Cui, Y., Zhang, F., Zhou, Y., 2018. The Application of Multi-Locus GWAS for the Detection of Salt-Tolerance Loci in Rice. Frontiers in Plant Science 9.

Desgroux, A., Baudais, V. N., Aubert, V., Le Roy, G., de Larambergue, H., Miteul, H., Aubert, G., Boutet, G., Duc, G., Baranger, A., Burstin, J., Manzanares-Dauleux, M., Pilet-Nayel, M.-L., & Bourion, V., 2018. Comparative Genome-Wide-Association Mapping Identifies Common Loci Controlling Root System Architecture and Resistance to Aphanomyces euteiches in Pea. *Frontiers in Plant Science*, *8*.

Desgroux, A., L'Anthoëne, V., Roux-Duparque, M., Rivière, J.-P., Aubert, G., Tayeh, N., Moussart, A., Mangin, P., Vetel, P., Piriou, C., McGee, R. J., Coyne, C. J., Burstin, J., Baranger, A., Manzanares-Dauleux, M., Bourion, V., & Pilet-Nayel, M.-L., 2016. Genome-wide association mapping of partial resistance to Aphanomyces euteiches in pea. *BMC Genomics*, *17*, 124. https://doi.org/10.1186/s12864-016-2429-4

Devlin, B., Roeder, K., 1999. Genomic Control for Association Studies. Biometrics 55, 997–1004.

Dissanayaka, D. N., Gali, K. K., Jha, A. B., Lachagari, V. B. R., & Warkentin, T. D., 2020. Genome-wide association study to identify single nucleotide polymorphisms associated with Fe, Zn, and Se concentration in field pea. *Crop Science*, *60*(4), 2070-2084. https://doi.org/10.1002/csc2.20161

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least Angle Regression" (with discussions). The Annals of Statistics 32.

Ersoz, E.S., Yu, J., Buckler, E.S., 2007. Applications of Linkage Disequilibrium and Association Mapping in Crop Plants, in: Varshney, R.K., Tuberosa, R. (Eds.), Genomics-Assisted Crop Improvement: Vol. 1: Genomics Approaches and Platforms. Springer Netherlands, Dordrecht, pp. 97–119. https://doi.org/10.1007/978-1-4020-6295-7_5

Faostat (2022) : https://www.fao.org/faostat/fr/#data/QCL

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S.N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E.S., Daly, M.J., Altshuler, D., 2002. The structure of haplotype blocks in the human genome. Science 296, 2225–2229. https://doi.org/10.1126/science.1069424

Galais, A., 2018. Histoire de la génétique et de l'amélioration des plantes. Éditions Quæ, Inra, Versailles, 175 p.

Gali, K., Sackville, A., Tafesse, E., Lachagari, R., Mcphee, K., Hybl, M., Mikić, A., Smykal, P., Mcgee, R., Burstin, J., Domoney, C., Ellis, N., Tar'an, B., & Warkentin, T., 2019. Genome-Wide Association Mapping for Agronomic and Seed Quality Traits of Field Pea (Pisum sativum L.). *Frontiers in Plant Science*, *10*. https://doi.org/10.3389/fpls.2019.01538

Gawenda, I., Thorwarth, P., Günther, T., Ordon, F., Schmid, K.J., 2015. Genome-wide association studies in elite varieties of German winter barley using single-marker and haplotype-based methods. Plant Breeding 134, 28–39. https://doi.org/10.1111/pbr.12237

Gilmour, A.R., Thompson, R., Cullis, B.R., 1995. Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. Biometrics 51, 1440–1450. https://doi.org/10.2307/2533274

Goddard, M.E., Wray, N.R., Verbyla, K., Visscher, P.M., 2009. Estimating Effects and Making Predictions from Genome-Wide Marker Data. Statistical Science 24, 517–529. https://doi.org/10.1214/09-STS306

Gupta, P.K., Kulwal, P.L., Jaiswal, V., 2019. Chapter Two - Association mapping in plants in the post-GWAS genomics era, in: Kumar, D. (Ed.), Advances in Genetics. Academic Press, pp. 75–154. https://doi.org/10.1016/bs.adgen.2018.12.001

Gupta, P.K., Kulwal, P.L., Jaiswal, V., 2014. Association mapping in crop plants: opportunities and challenges. Adv Genet 85, 109–147. https://doi.org/10.1016/B978-0-12-800271-1.00002-0

Hellens RP, Moreau C, Lin-Wang K, Schwinn KE, Thomson SJ, Fiers MWEJ, et al. (2010) Identification of Mendel's White Flower Character. PLoS ONE 5(10): e13230. https://doi.org/10.1371/journal.pone.0013230

Hickey, L.T., N. Hafeez, A., Robinson, H., Jackson, S.A., Leal-Bertioli, S.C.M., Tester, M., Gao, C., Godwin, I.D., Hayes, B.J., Wulff, B.B.H., 2019. Breeding crops to feed 10 billion. Nat Biotechnol 37, 744–754. https://doi.org/10.1038/s41587-019-0152-9

Holm, S., 1979. A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics 6, 65–70.

Hu, S., Sanchez, D.L., Wang, C., Lipka, A.E., Yin, Y., Gardner, C.A.C., Lübberstedt, T., 2017. Brassinosteroid and gibberellin control of seedling traits in maize (Zea mays L.). Plant Science 263, 132–141. https://doi.org/10.1016/j.plantsci.2017.07.011

Huang, M., Liu, X., Zhou, Y., Summers, R.M., Zhang, Z., 2019. BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. GigaScience 8, giy154. https://doi.org/10.1093/gigascience/giy154

Huang, X., Han, B., 2014. Natural variations and genome-wide association studies in crop plants. Annu Rev Plant Biol 65, 531–551. https://doi.org/10.1146/annurev-arplant-050213-035715

Jaiswal, V., Gahlaut, V., Meher, P.K., Mir, R.R., Jaiswal, J.P., Rao, A.R., Balyan, H.S., Gupta, P.K., 2016. Genome Wide Single Locus Single Trait, Multi-Locus and Multi-Trait Association Mapping for Some Important Agronomic Traits in Common Wheat (T. aestivum L.). PLOS ONE 11, e0159343. https://doi.org/10.1371/journal.pone.0159343

Jaiswal V, Mir RR, Mohan A, Balyan HS, Gupta PK. Association mapping for pre-harvest sprouting tolerance in common wheat (Triticum aestivum L.). Euphytica. 2012; 188:89–102

Kaler, A.S., Gillman, J.D., Beissinger, T., Purcell, L.C., 2020. Comparing Different Statistical Models and Multiple Testing Corrections for Association Mapping in Soybean and Maize. Frontiers in Plant Science 10, 1794. https://doi.org/10.3389/fpls.2019.01794

Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S., Freimer, N.B., Sabatti, C., Eskin, E., 2010. Variance component model to account for sample structure in genome-wide association studies. Nat Genet 42, 348–354. https://doi.org/10.1038/ng.548

Kang, H.M., Zaitlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., Daly, M.J., Eskin, E., 2008. Efficient Control of Population Structure in Model Organism Association Mapping. Genetics 178, 1709–1723. https://doi.org/10.1534/genetics.107.080101

Khan, M.A., Tong, F., Wang, W., He, J., Zhao, T., Gai, J., 2018. Analysis of QTL–allele system conferring drought tolerance at seedling stage in a nested association mapping population of soybean [Glycine max (L.) Merr.] using a novel GWAS procedure. Planta 248, 947–962. https://doi.org/10.1007/s00425-018-2952-4

Klasen, J.R., Barbez, E., Meier, L., Meinshausen, N., Bühlmann, P., Koornneef, M., Busch, W., Schneeberger, K., 2016. A multi-marker association method for genome-wide association studies without the need for population structure correction. Nat Commun 7, 13299. https://doi.org/10.1038/ncomms13299

Korte, A., & Farlow, A., 2013. The advantages and limitations of trait analysis with GWAS : A review. *Plant Methods*, *9*(1), 29. https://doi.org/10.1186/1746-4811-9-29

Kumar, J., Saripalli, G., Gahlaut, V., Goel, N., Meher, P.K., Mishra, K.K., Mishra, P.C., Sehgal, D., Vikram, P., Sansaloni, C., Singh, S., Sharma, P.K., Gupta, P.K., 2018. Genetics of Fe, Zn, β-carotene, GPC and yield traits in bread wheat (Triticum aestivum L.) using multi-locus and multi-traits GWAS. Euphytica 214, 219. https://doi.org/10.1007/s10681-018-2284-2

Lê, S., Josse. J., Husson, F., 2008. "FactoMineR: A Package for Multivariate Analysis." Journal of Statistical Software, 25(1), 1–18. doi: 10.18637/jss.v025.i01

Li, C., Fu, Y., Sun, R., Wang, Y., Wang, Q., 2018. Single-Locus and Multi-Locus Genome-Wide Association Studies in the Genetic Dissection of Fiber Quality Traits in Upland Cotton (Gossypium hirsutum L.). Frontiers in Plant Science 9.

Li, M., Liu, X., Bradbury, P., Yu, J., Zhang, Y.-M., Todhunter, R.J., Buckler, E.S., Zhang, Z., 2014. Enrichment of statistical power for genome-wide association studies. BMC Biology 12, 73. https://doi.org/10.1186/s12915-014-0073-5

Lipka, A.E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P.J., Gore, M.A., Buckler, E.S., Zhang, Z., 2012. GAPIT: genome association and prediction integrated tool. Bioinformatics 28, 2397–2399. https://doi.org/10.1093/bioinformatics/bts444

Lippert, C., Listgarten, J., Liu, Y., Kadie, C.M., Davidson, R.I., Heckerman, D., 2011. FaST linear mixed models for genome-wide association studies. Nat Methods 8, 833–835. https://doi.org/10.1038/nmeth.1681

Listgarten, J., Lippert, C., Kadie, C.M., Davidson, R.I., Eskin, E., Heckerman, D., 2012. Improved linear mixed models for genome-wide association studies. Nat Methods 9, 525–526. https://doi.org/10.1038/nmeth.2037

Liu, H.-J., Yan, J., 2019. Crop genome-wide association study: a harvest of biological relevance. The Plant Journal 97, 8–18. https://doi.org/10.1111/tpj.14139

Liu, N., Zhang, K., Zhao, H., 2008. Haplotype-association analysis. Adv Genet 60, 335–405. https://doi.org/10.1016/S0065-2660(07)00414-2

Liu, R., Gong, J., Xiao, X., Zhang, Z., Li, J., Liu, A., Lu, Q., Shang, H., Shi, Y., Ge, Q., Iqbal, M.S., Deng, X., Li, S., Pan, J., Duan, L., Zhang, Q., Jiang, X., Zou, X., Hafeez, A., Chen, Q., Geng, H., Gong, W., Yuan, Y., 2018. GWAS Analysis and QTL Identification of Fiber Quality Traits and Yield Components in Upland Cotton Using Enriched High-Density SNP Markers. Frontiers in Plant Science 9.

Liu, S., Zhong, H., Meng, X., Sun, T., Li, Y., Pinson, S.R.M., Chang, S.K.C., Peng, Z., 2020. Genome-wide association studies of ionomic and agronomic traits in USDA mini core collection of rice and comparative analyses of different mapping methods. BMC Plant Biology 20, 441. https://doi.org/10.1186/s12870-020-02603-0

Liu, X., Huang, M., Fan, B., Buckler, E.S., Zhang, Z., 2016. Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. PLOS Genetics 12, e1005767. https://doi.org/10.1371/journal.pgen.1005767

Loh, P.-R., Tucker, G., Bulik-Sullivan, B.K., Vilhjálmsson, B.J., Finucane, H.K., Salem, R.M., Chasman, D.I., Ridker, P.M., Neale, B.M., Berger, B., Patterson, N., Price, A.L., 2015. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet 47, 284–290. https://doi.org/10.1038/ng.3190

Loiselle, B.A., Sork, V.L., Nason, J., Graham, C., 1995. Spatial genetic structure of a tropical understory shrub, PSYCHOTRIA OFFICINALIS (RuBIACEAE). American Journal of Botany 82, 1420–1425. https://doi.org/10.1002/j.1537-2197.1995.tb12679.x

Lü, H., Yang, Y., Li, H., Liu, Q., Zhang, J., Yin, J., Chu, S., Zhang, X., Yu, K., Lv, L., Chen, X., Zhang, D., 2018. Genome-Wide Association Studies of Photosynthetic Traits Related to Phosphorus Efficiency in Soybean. Frontiers in Plant Science 9.

Ma, L., Liu, M., Yan, Y., Qing, C., Zhang, X., Zhang, Y., Long, Y., Wang, L., Pan, L., Zou, C., Li, Z., Wang, Y., Peng, H., Pan, G., Jiang, Z., Shen, Y., 2018. Genetic Dissection of Maize Embryonic Callus Regenerative Capacity Using Multi-Locus Genome-Wide Association Studies. Frontiers in Plant Science 9.

Mackay, T.F.C., 2001. The Genetic Architecture of Quantitative Traits. Annu. Rev. Genet. 35, 303–339. https://doi.org/10.1146/annurev.genet.35.102401.090633

Mackay, T.F.C., Stone, E.A., Ayroles, J.F., 2009. The genetics of quantitative traits: challenges and prospects. Nat Rev Genet 10, 565–577. https://doi.org/10.1038/nrg2612

Martin, D.N., Proebsting, W.M. & Hedden, P. (1997) Mendel's dwarfing gene: cDNAs from the Le alleles and function of the expressed proteins. Proceedings of the National Academy of Sciences USA, 94, 8907– 8911.

Michael, T.P., Jackson, S., 2013. The First 50 Plant Genomes. The Plant Genome 6, plantgenome2013.03.0001in. https://doi.org/10.3835/plantgenome2013.03.0001in

Miller, K.S., 1981. On the Inverse of the Sum of Matrices. Mathematics Magazine 54, 67–72. https://doi.org/10.1080/0025570X.1981.11976898

Mir RR, Kumar N, Jaiswal V, Girdharwal N, Prasad M, Balyan HS, et al. Genetic dissection of grain weight in bread wheat through quantitative trait locus interval and association mapping. Mol Breed. 2012; 29: 963–972.

Misra, G., Badoni, S., Domingo, C.J., Cuevas, R.P.O., Llorente, C., Mbanjo, E.G.N., Sreenivasulu, N., 2018. Deciphering the Genetic Architecture of Cooked Rice Texture. Frontiers in Plant Science 9.

Naveed, S.A., Zhang, F., Zhang, J., Zheng, T.-Q., Meng, L.-J., Pang, Y.-L., Xu, J.-L., Li, Z.-K., 2018. Identification of QTN and candidate genes for Salinity Tolerance at the Germination and Seedling Stages in Rice by Genome-Wide Association Analyses. Sci Rep 8, 6505. https://doi.org/10.1038/s41598-018-24946-3

Nemecek, J.-S. von Richthofen, G. Dubois, P. Casta, R. Charles, H. Pahl., 2008. Environmental impacts of introducing grain legumes into European crop rotations Eur. J. Agron., 28 (2008), pp. 380-393, 10.1016/j.eja.2007.11.004

Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., ... & Shendure, J., 2009. Targeted capture and massively parallel sequencing of 12 human exomes. Nature, 461(7261), 272-276.

N'Diaye, A., Haile, J.K., Cory, A.T., Clarke, F.R., Clarke, J.M., Knox, R.E., Pozniak, C.J., 2017. Single Marker and Haplotype-Based Association Analysis of Semolina and Pasta Colour in Elite Durum Wheat Breeding Lines Using a High-Density Consensus Map. PLOS ONE 12, e0170941. https://doi.org/10.1371/journal.pone.0170941

Ollivier, R., Glory, I., Cloteau, R., Le Gallic, J.-F., Denis, G., Morlière, S., Miteul, H., Rivière, J.-P., Lesné, A., Klein, A., Aubert, G., Kreplak, J., Burstin, J., Pilet-Nayel, M.-L., Simon, J.-C., & Sugio, A., 2022. A major-effect genetic locus, ApRVII, controlling resistance against both adapted and non-adapted aphid biotypes in pea. *Theoretical and Applied Genetics*. https://doi.org/10.1007/s00122-022-04050-x

Peng, Y., Liu, H., Chen, J., Shi, T., Zhang, C., Sun, D., He, Z., Hao, Y., Chen, W., 2018. Genome-Wide Association Studies of Free Amino Acid Levels by Six Multi-Locus Models in Bread Wheat. Frontiers in Plant Science 9.

Pook, T., Schlather, M., de los Campos, G., Mayer, M., Schoen, C.C., Simianer, H., 2019. HaploBlocker: Creation of Subgroup-Specific Haplotype Blocks and Libraries. Genetics 212, 1045–1061. https://doi.org/10.1534/genetics.119.302283

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., Reich, D., 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38, 904–909. https://doi.org/10.1038/ng1847

Pritchard, J.K., Stephens, M., Rosenberg, N.A., Donnelly, P., 2000. Association Mapping in Structured Populations. The American Journal of Human Genetics 67, 170–181. https://doi.org/10.1086/302959

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., Sham, P.C., 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet 81, 559–575.

Reich, D., Price, A.L., Patterson, N., 2008. Principal component analysis of genetic data. Nat Genet 40, 491–492. https://doi.org/10.1038/ng0508-491

Ren, W.-L., Wen, Y.-J., Dunwell, J.M., Zhang, Y.-M., 2018. pKWmEB: integration of Kruskal–Wallis test with empirical Bayes under polygenic background control for multi-locus genome-wide association study. Heredity 120, 208–218. https://doi.org/10.1038/s41437-017-0007-4

Risch, N., Merikangas, K., 1996. The Future of Genetic Studies of Complex Human Diseases. Science 273, 1516–1517. https://doi.org/10.1126/science.273.5281.1516

RStudio Team., 2022. RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA. URL http://www.rstudio.com/.

Samouelian, F., Gaudin, V., Boccara, M., 2009. Génétique moléculaire des plantes Ed. 1. Editions Quae.

Sanchez, D.L., Liu, S., Ibrahim, R., Blanco, M., Lübberstedt, T., 2018. Genome-wide association studies of doubled haploid exotic introgression lines for root system architecture traits in maize (Zea mays L.). Plant Science 268, 30–38. https://doi.org/10.1016/j.plantsci.2017.12.004

Segura, V., Vilhjálmsson, B.J., Platt, A., Korte, A., Seren, Ü., Long, Q., Nordborg, M., 2012. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet 44, 825–830. https://doi.org/10.1038/ng.2314

Speed, D., Balding, D.J., 2015. Relatedness in the post-genomic era: is it still useful? Nat Rev Genet 16, 33–44. https://doi.org/10.1038/nrg3821

Su, J., Ma, Q., Li, M., Hao, F., Wang, C., 2018. Multi-Locus Genome-Wide Association Studies of Fiber-Quality Related Traits in Chinese Early-Maturity Upland Cotton. Frontiers in Plant Science 9.

Sukumaran, S., Yu, J., 2014. Association Mapping of Genetic Resources: Achievements and Future Perspectives, in: Tuberosa, R., Graner, A., Frison, E. (Eds.), Genomics of Plant Genetic Resources: Volume 1. Managing, Sequencing and Mining Genetic Resources. Springer Netherlands, Dordrecht, pp. 207–235. https://doi.org/10.1007/978-94-007-7572-5_9

Sulima, A.S., and Zhukov, V.A. (2022). "War and Peas: Molecular Bases of Resistance to Powdery Mildew in Pea (Pisum sativum L.) and Other Legumes" Plants 11, no. 3: 339. https://doi.org/10.3390/plants11030339

Svishcheva, G.R., Axenovich, T.I., Belonogova, N.M., van Duijn, C.M., Aulchenko, Y.S., 2012. Rapid variance components–based method for whole-genome association analysis. Nat Genet 44, 1166–1170. https://doi.org/10.1038/ng.2410

Tamba, C.L., Ni, Y.-L., Zhang, Y.-M., 2017. Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. PLOS Computational Biology 13, e1005357. https://doi.org/10.1371/journal.pcbi.1005357

Tamba, C.L., Zhang, Y.-M., 2018. A fast mrMLM algorithm for multi-locus genome-wide association studies. https://doi.org/10.1101/341784

Tayeh, N., Aluome, C., Falque, M., Jacquin, F., Klein, A., Chauveau, A., et al., 2015. Development of two major resources for pea genomics: the GenoPea 13.2K SNP array and a high-density, high-resolution consensus genetic map. Plant J. 84, 1257–1273. doi: 10.1111/tpj.13070

Thornsberry, J.M., Goodman, M.M., Doebley, J., Kresovich, S., Nielsen, D., Buckler, E.S., 2001. Dwarf8 polymorphisms associate with variation in flowering time. Nat Genet 28, 286–289. https://doi.org/10.1038/90135

Thomas D. Wu, Colin K. Watanabe., 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences, *Bioinformatics*, Volume 21, Issue 9, , Pages 1859–1875, https://doi.org/10.1093/bioinformatics/bti310

Turner, S., 2018. "qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots." The Journal of Open Source Software. doi: 10.21105/joss.00731.

VanRaden, P.M., 2008. Efficient Methods to Compute Genomic Predictions. Journal of Dairy Science 91, 4414–4423. https://doi.org/10.3168/jds.2007-0980

Wang, Q., Tian, F., Pan, Y., Buckler, E.S., Zhang, Z., 2014. A SUPER Powerful Method for Genome Wide Association Study. PLOS ONE 9, e107684. https://doi.org/10.1371/journal.pone.0107684

Wang, S.-B., Feng, J.-Y., Ren, W.-L., Huang, B., Zhou, L., Wen, Y.-J., Zhang, J., Dunwell, J.M., Xu, S., Zhang, Y.-M., 2016. Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. Sci Rep 6, 19444. https://doi.org/10.1038/srep19444

Wen, Y.-J., Zhang, H., Ni, Y.-L., Huang, B., Zhang, J., Feng, J.-Y., Wang, S.-B., Dunwell, J.M., Zhang, Y.-M., Wu, R., 2018. Methodological implementation of mixed linear models in multi-locus genome-wide association studies. Briefings in Bioinformatics 19, 700–712. https://doi.org/10.1093/bib/bbw145

Xu, S., 2013. Mapping Quantitative Trait Loci by Controlling Polygenic Background Effects. Genetics 195, 1209–1222. https://doi.org/10.1534/genetics.113.157032

Xu, Y., Yang, T., Zhou, Y., Yin, S., Li, P., Liu, J., Xu, S., Yang, Z., Xu, C., 2018. Genome-Wide Association Mapping of Starch Pasting Properties in Maize Using Single-Locus and Multi-Locus Models. Frontiers in Plant Science 9.

Yang, J., Yeh, C.-T. "Eddy," Ramamurthy, R.K., Qi, X., Fernando, R.L., Dekkers, J.C.M., Garrick, D.J., Nettleton, D., Schnable, P.S., 2018. Empirical Comparisons of Different Statistical Models To Identify and Validate Kernel Row Number-Associated Variants from Structured Multi-parent Mapping Populations of Maize. G3 Genes|Genomes|Genetics 8, 3567–3575. https://doi.org/10.1534/g3.118.200636

Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., Kresovich, S., Buckler, E.S., 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38, 203–208. https://doi.org/10.1038/ng1702

Zeng, Z.B., 1994. Precision mapping of quantitative trait loci. Genetics 136, 1457–1468. https://doi.org/10.1093/genetics/136.4.1457

Zhang, J., Feng, J.-Y., Ni, Y.-L., Wen, Y.-J., Niu, Y., Tamba, C.L., Yue, C., Song, Q., Zhang, Y.-M., 2017. pLARmEB: integration of least angle regression with empirical Bayes for multilocus genome-wide association studies. Heredity 118, 517–524. https://doi.org/10.1038/hdy.2017.8

Zhang, Y., Liu, P., Zhang, X., Zheng, Q., Chen, M., Ge, F., Li, Z., Sun, W., Guan, Z., Liang, T., Zheng, Y., Tan, X., Zou, C., Peng, H., Pan, G., Shen, Y., 2018. Multi-Locus Genome-Wide Association Study Reveals the Genetic Architecture of Stalk Lodging Resistance-Related Traits in Maize. Frontiers in Plant Science 9.

Zhang, Y.-M., Jia, Z., Dunwell, J.M., 2019. Editorial: The Applications of New Multi-Locus GWAS Methodologies in the Genetic Dissection of Complex Traits. Frontiers in Plant Science 10.

Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R.J., Tiwari, H.K., Gore, M.A., Bradbury, P.J., Yu, J., Arnett, D.K., Ordovas, J.M., Buckler, E.S., 2010. Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42, 355–360. https://doi.org/10.1038/ng.546

Zhao, K., Aranzana, M.J., Kim, S., Lister, C., Shindo, C., Tang, C., Toomajian, C., Zheng, H., Dean, C., Marjoram, P., Nordborg, M., 2007. An Arabidopsis Example of Association Mapping in Structured Samples. PLOS Genetics 3, e4. https://doi.org/10.1371/journal.pgen.0030004

Zhao, X., Teng, W., Li, Y., Liu, D., Cao, G., Li, D., Qiu, L., Zheng, H., Han, Y., Li, W., 2017. Loci and candidate genes conferring resistance to soybean cyst nematode HG type 2.5.7. BMC Genomics 18, 462. https://doi.org/10.1186/s12864-017-3843-y

Zhou, X., Stephens, M., 2012. Genome-wide efficient mixed-model analysis for association studies. Nat Genet 44, 821–824. https://doi.org/10.1038/ng.2310

Zhu, C., Gore, M., Buckler, E.S., Yu, J., 2008. Status and Prospects of Association Mapping in Plants. The Plant Genome 1. https://doi.org/10.3835/plantgenome2008.02.0089

Annexes

Annexe 1 : Variables étudiées, méthodes de Phénotypage, sites et années de phénotypage

Type de Trait	Nom du trait	Nom codé	Méthode d'observation/mesure	Site	Année
Biomasse aérienne	Biomasse_aerienne_maturite_Maug_15	Biom_Aer_Matur_MG15	cumul sur 10 plantes	INRAE Mauguio	2015
coloration	Couleur_feuilles_sortie_hiver_Mons_17	Col_Feuil_SortiHv_MS17		INRAE Mons	2016/2022
coloration	Couleur_feuilles_sortie_hiver_ORS_17	Col_Feuil_SortiHv_ORS17	évaluation en sortie hiver, le 25/02/2017	Agri Obtentions, Orsonville	2016/2023
coloration	Couleur_fleurs	Col_fleur	Blanc=1, Violet=2	INRAE Bretenière	2015
coloration	Couleur_hile	Col_Hil	Foncé=2, Clair=2	INRAE Bretenière	2015
coloration	Couleur_tegument_grains	Col_teg_Gr	Vert=1, Beige=2, Gris=3, Marron=4	INRAE Bretenière	2015
floraison	Date_debut_floraison_Bret_15	Deb_Flor_B15		INRAE Bretenière	2015
floraison	Date_debut_floraison_Bret_16	Deb_Flor_B16		INRAE Bretenière	2016
floraison	Date_debut_floraison_semis_automne_Mons_17	Deb_Flor_semAut_MS17	Nombre de jours depuis le semis à l'automne	INRAE Mons	2016/2017
durée remplissage grain	Date_debut_remplissage_grains_Bret_15	Deb_Rempl_Gr_B15		INRAE Bretenière	2015
durée remplissage grain	Date_debut_remplissage_grains_Bret_16	Deb_Rempl_Gr_B16		INRAE Bretenière	2016
floraison	Date_fin_floraison_Bret_15	Fin_Flor_B15		INRAE Bretenière	2015
floraison	Date_fin_floraison_Bret_16	Fin_Flor_B16		INRAE Bretenière	2016
floraison	Date_fin_floraison_semis_automne_Mons_17	Fin_Flor_SemAut_MS17	Nombre de jours depuis le semis à l'automne	INRAE Mons	2016/2018
maturité	Date_maturite_Bret_15	Matur_B15	enregistrée comme date de récolte	INRAE Bretenière	2015
dégât gel	Degats_gel_sortie_hiver_Mons_17	Dega_gel_SortiHv_MS17		INRAE Mons	2016/2017
dégât gel	Degats_gel_sortie_hiver_Mons_18	Dega_gel_SortiHv_MS18		INRAE Mons	2017/2018
dégât gel	Degats_gel_sortie_hiver_ORS_17	Dega_gel_SortiHv_ORS17	évaluation en sortie hiver, le 25/02/2017	Agri Obtentions, Orsonville	2016/2017
résistance puceron	Fecondite_puceron_clone_ArPo28	Fecond_Puc_ArPo28	collection entière testée en même temps avec 1 plante par accession - 13 répétitions	chambre climatique	
résistance puceron	Fecondite_puceron_clone_LSR1	Fecond_Puc_LSR1	collection entière testée en même temps avec 1 plante par accession - 12 répétitions	chambre climatique	
forme feuille	Forme_feuilles	Form_feuil	Feuillu=1, afila=2	INRAE Bretenière	2015
forme foliole	Forme_folioles	Form_foliol	Ovale=1, Elliptique=2	INRAE Bretenière	2015
marbrure grains	Grains_marbres	Gr_marbr	Présence=1, Absence=2	INRAE Bretenière	2015
moucheture grain	Grains_mouchetes	Gr_mouch	Présence=1, Absence=2	INRAE Bretenière	2015
Hauteur	Hauteur_fin_floraison_Bret_15	Haut_Fin_Flor_B15		INRAE Bretenière	2015
Hauteur	Hauteur_fin_floraison_Maug_15	Haut_Fin_Flor_MG15	Sur 1 plante représentive de la ligne	INRAE Mauguio	2015
Hauteur	Hauteur_fin_floraison_Rheu_16	Haut_Fin_Flor_R16		INRAE Le Rheu	2016
Hauteur	Hauteur_fin_floraison_Rheu_17	Haut_Fin_Flor_R17		INRAE Le Rheu	2017
Hauteur	Hauteur_tige_princip_sortie_hiver_Mons_17	Haut_TigPrinc_SortiHv_MS17		INRAE Mons	2016/2020
Hauteur	Hauteur_tige_princip_sortie_hiver_ORS_17	Haut_TigPrinc_SortiHv_ORS17	évaluation en sortie hiver, le 25/02/2017	Agri Obtentions, Orsonville	2016/2021

Type de Trait	Nom du trait	Nom codé	Méthode d'observation/mesure	Site	Année
Dégât bruche	Nb_grains_bruches_Bret_15	Nbr_Gr_1Plt_B15		INRAE Bretenière	2015
Nombre grains/plant	Nb_grains_par_plante_Bret_15	Nbr_Gr_Bruch_B15	(number_of_healthy_seeds+	INRAE Bretenière	2015
Nombre grains/plant	Nb_grains_par_plante_Maug_15	Nbr_Gr_1Plt_MG15	cumul sur 10 plantes (moyenne)	INRAE Mauguio	2015
floraison	Nb_jours_floraison_Bret_16	Nbr_Jr_Flor_B16		INRAE Bretenière	2016
floraison	Nb_jours_floraison_semis_automne_Mons_17	Nbr_Jr_Flor_SemAut_MS17		INRAE Mons	2016/2019
ramification aérienne	Nb_ramifications_aeriennes_par_plante_Maug_15	Nbr_Ramif_Aer_1Plt_MG15	moyenne sur 10 plantes, à maturité; ramifications aériennes	INRAE Mauguio	2015
ramification basale/Pl	Nb_ramifications_basales_par_plante_Bret_15	Nbr_Ramif_Basal_1Plt_B15	number_of_stems divisé par number_of_plants	INRAE Bretenière	2015
ramification basale/Pl	Nb_ramifications_basales_par_plante_Maug_15	Nbr_Ramif_Basal_1Plt_MG15	moyenne sur 10 plantes, à maturité	INRAE Mauguio	2015
nombre de tige	Nb_tiges_debut_floraison_ORS_17	Nbr_Tig_DebFlor_ORS17	évaluation début floraison vers le 20/04/2017	Agri Obtentions, Orsonville	2016/2026
physiologique	NDVI_floraison_Maug_15	NDVI_Flor_MG15	moyenne parcelle, à floraison (21 mai)	INRAE Mauguio	2015
poids grains/Plant	Poids_grains_par_plante_Bret_15			INRAE Bretenière	2015
poids grains/Plant	Poids_grains_par_plante_Maug_15	Poids_Gr_1Plt_MG15	cumul sur 10 plantes (moyenne)	INRAE Mauguio	2015
poids mille grains	Poids_mille_grains_Bret_15	PMG_B15	sur lot de grains sains non bruchés	INRAE Bretenière	2015
poids mille grains	Poids_mille_grains_Bret_16	PMG_B16		INRAE Bretenière	2016
poids mille grains	Poids_mille_grains_Maug_15	PMG_MG15		INRAE Mauguio	2015
Dégât bruche	Pourcent_grains_bruches_Bret_16	Pourc_Gr_Bruch_B16	Tomographie, pourcentage des semences bruchées du lot	INRAE Bretenière	2016
résistance bruche	Pourcent_grains_non_bruches_Bret_15	Pourc_Gr_0Bruch_B15		INRAE Bretenière	2015
sensibilité bruche	Pourcent_volume_mange_vs_total_grains_Bret_16	Pourc_Vol_Mange_B16	Tomographie, pourcentage des dégâts par rapport au volume total des semences du lot (%)	INRAE Bretenière	2016
sensibilité bruche	Pourcent_volume_mange_vs_total_grains_bruches_Bret_16	Pourc_Vol_Mangé_Bruch_B16	Tomographie, pourcentage moyen du dégât par semence par rapport au volume total de la semence bruchée	INRAE Bretenière	2016
résistance oïdium	Resistance_oidium_Maug_15	Resist_Oid_MG15	seulement sur 2 rep/3 - 1 valeur par microparcelle	INRAE Mauguio	2015
résistance puceron	Resistance_pucerons_Maug_15	Resist_Puc_MG15	seulement sur 2 rep/3 - 1 valeur par microparcelle	INRAE Mauguio	2015
taille feuilles	Taille_feuilles_sortie_hiver_Mons_17	Tail_Feuil_SortiHiv_MS17		Pépinière, INRAE Mons	2016/2024
taille feuilles	Taille_feuilles_sortie_hiver_ORS_17	Tail_Feuil_SortiHiv_ORS17	évaluation en sortie hiver, le 25/02/2017	Agri Obtentions, Orsonville	2016/2025
taux microéléments	Teneur_cuivre_grains_Bret_15	Tx_Cu_Gr_B15	à partir de farines de graines de pois (10-15 mg), Cu 324.754	INRAE Bretenière	2015
taux microéléments	Teneur_fer_grains_Bret_15	Tx_Fe_Gr_B15	à partir de farines de graines de pois (10-15 mg), Fe 259.940	INRAE Bretenière	2015
taux microéléments	Teneur_manganese_grains_Bret_15	Tx_Mn_Gr_B15	à partir de farines de graines de pois (10-15 mg), Mn 403.076	INRAE Bretenière	2015
taux protéine	Teneur_proteines_grains_Bret_15	Tx_Prot_Gr_B15	NIRS	INRAE Bretenière	2015
taux microéléments	Teneur_zinc_grains_Bret_15	Tx_Zn_Gr_B15	à partir de farines de graines de pois (10-15 mg), Zn 213.857	INRAE Bretenière	2015

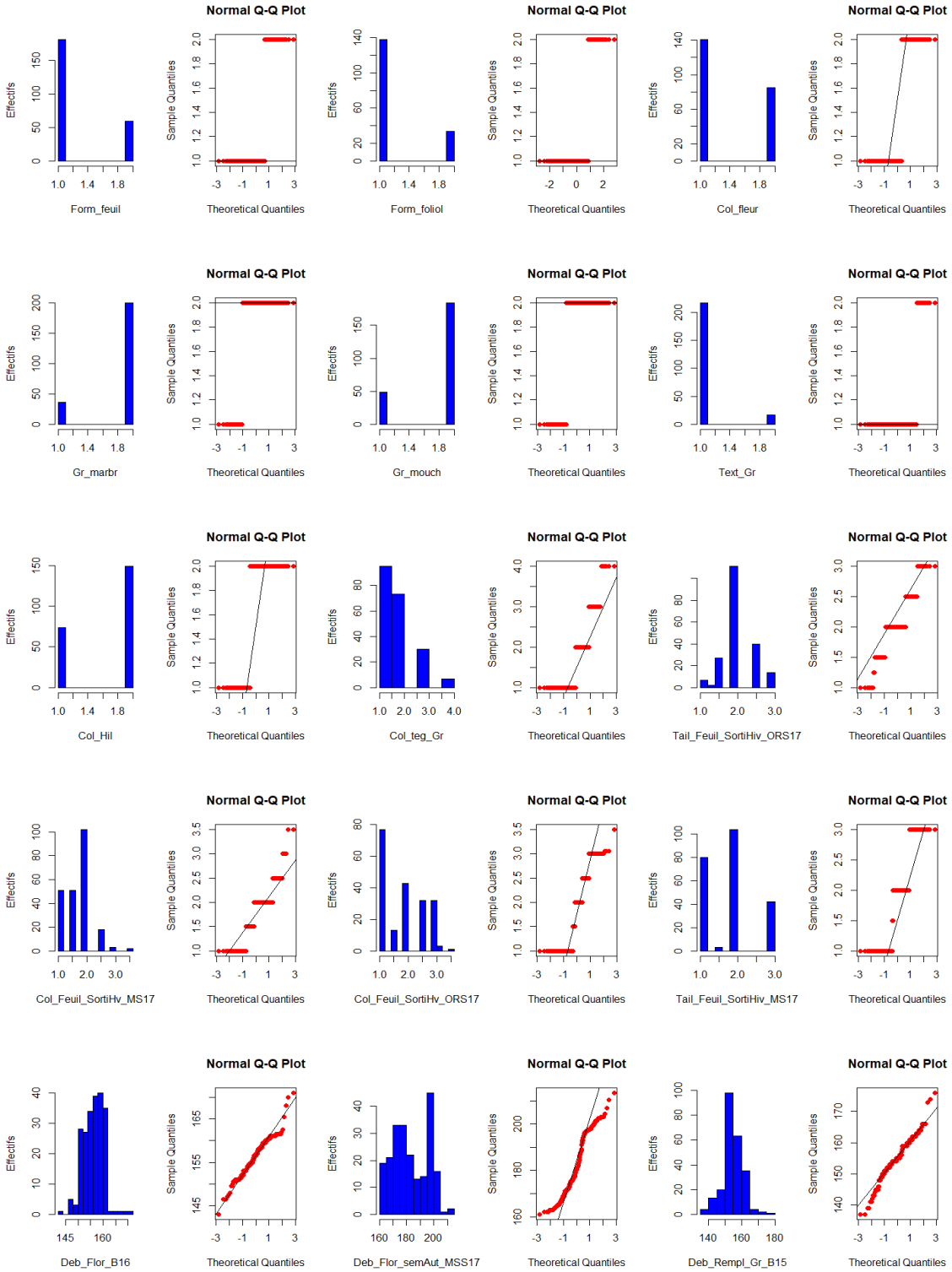
Type de Trait	Nom du trait	Nom codé	Méthode d'observation/mesure	Site	Année
volume grains	Volume_grains_Bret_16	Volum_Gr_B16	Tomographie, moyenne des volumes de semences du lot (mm3)	INRAE Bretenière	2015
sensibilité bruche	Volume_grains_bruches_Bret_16	Volum_Gr_Bruch_B16	Tomographie, moyenne des volumes de semences bruchées du lot (mm^3)	INRAE Bretenière	2016
sensibilité bruche	Volume_grains_bruches_theorique_Bret_16	Volum_Gr_Bruch_Theor_B16	Tomographie, moyenne des volumes de semences bruchées (volume théorique comme si elles n'étaient pas bruchées) du lot	INRAE Bretenière	2016
résistance bruche	Volume_grains_non_bruches_Bret_16	Volum_Gr_0Bruch_B16	Tomographie, moyenne des volumes de semences non bruchées du lot (mm3)	INRAE Bretenière	2016

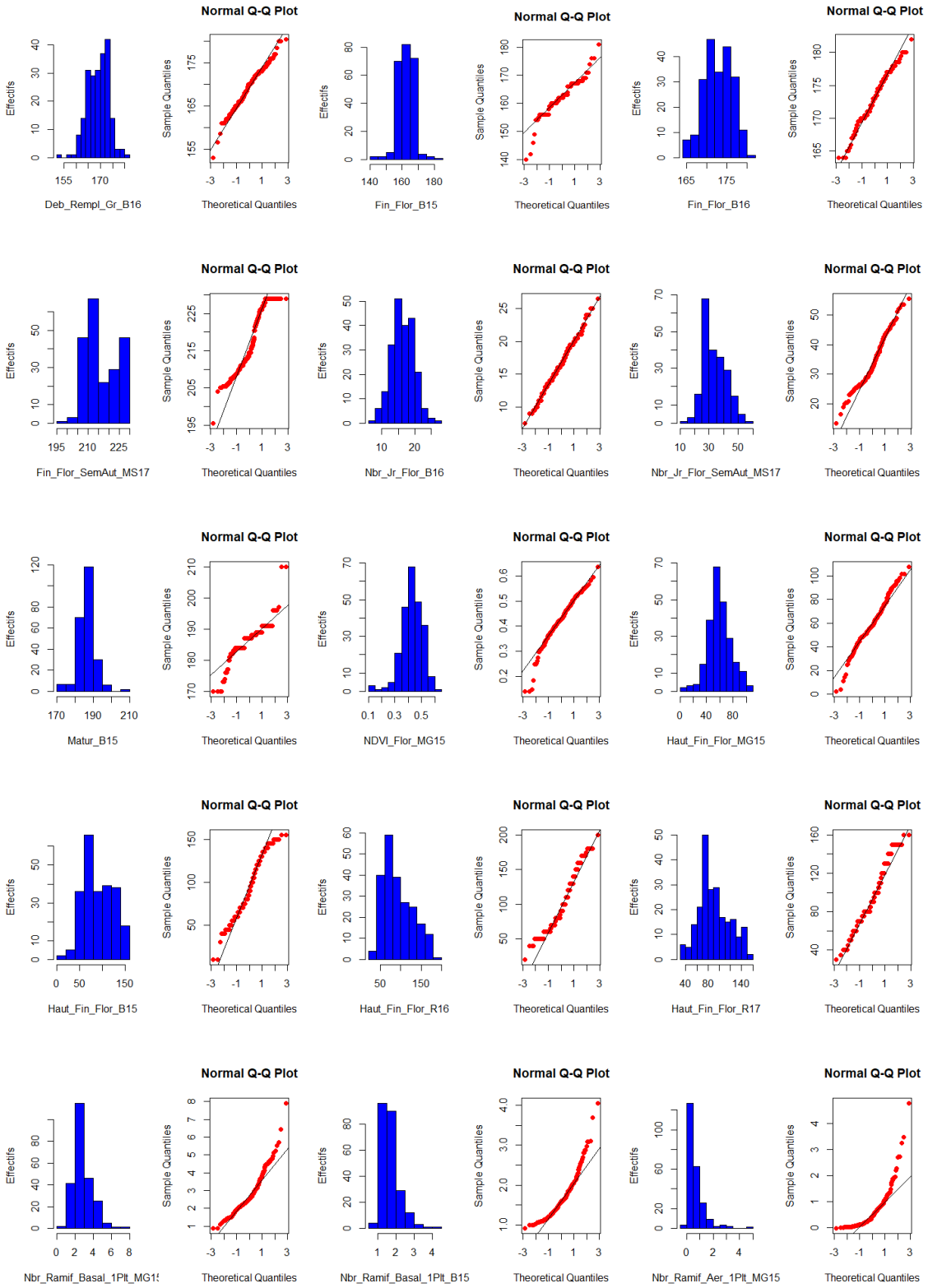
Annexe 2 : Caractéristique des neuf méthodes GWAS testés

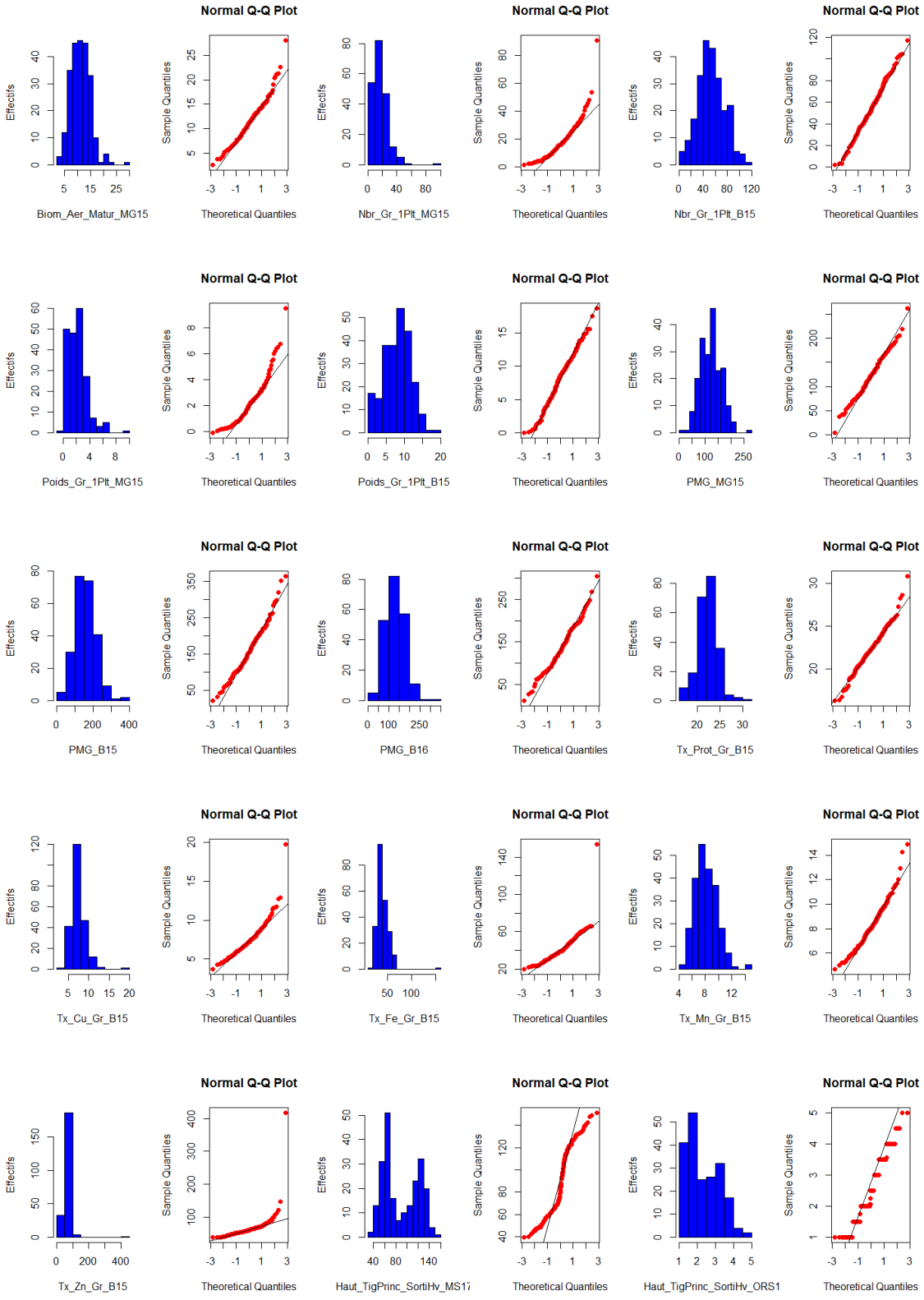
Approche	Méthode GWAS	Principe	Référence
locus unique (ULMLM)	MLM	Par rapport au GLM, le modèle MLM élimine efficacement les faux positifs en incorporant simultanément dans le modèle la stratification de la population (Q) comme effet fixe et la parenté génétique entre individus (K) comme effet aléatoire (Yu et al., 2006 ; Figure 2)	Yu et al., 2006
	CMLM	Améliore la puissance du MLM en utilisant une matrice de parenté de rang réduit. Elles utilisent un algorithme de regroupement pour diviser les individus en groupes basés sur des génotypes similaires. Un résumé de la parenté au sein et entre les groupes est ensuite utilisé comme matrice de parenté réduite lors de la résolution du MLM .	Zhang et al., 2010
	Fast-LMM-Select	utilise une approche simple de réduction du nombre de SNP pour le calcul de la matrice de parenté afin d'améliorer l'efficacité de calcul du MLM. Elle sélectionne un sous-ensemble de SNP associés au trait d'intérêt, de sorte que les matrices de parenté calculées soient spécifiques à chaque trait. Elle utilise un intervalle arbitraire de 2 cM comme seuil d'exclusion pour un LD entre les pseudo QTN (Quantitative Trait Nucleotide) et le marqueur de test.	Listgarten et al., 2012
Multilocus à effet SNP fixe ou partiellement fixe (MLMM)	MLMM	impliquent des algorithmes en deux étapes, consistant en un balayage à un seul locus de l'ensemble du génome pour détecter tous les SNP associés au trait, puis en testant tous les SNP associés à l'aide d'un modèle GWAS multi-locus avec comme cofacteur les SNP significatifs, pour détecter les vrais SNP	Segura et al., 2012
	FarmCPU	Également à deux étapes, d'abord un modèle à effets fixes (FEM) et ensuite un modèle à effets aléatoires (REM), qui sont utilisés de manière itérative, avec le maximum de vraisemblance restreint (REML) comme critère d'optimisation. FEM contient des marqueurs de test, un à la fois, et plusieurs marqueurs associés en tant que covariables pour contrôler les faux positifs. Elle utilise une matrice de parenté réduite pour améliorer la puissance statistique.	Lui et al., 2016
	BLINK	Elle constitue une modification de FarmCPU et permet d'améliorer la puissance en assouplissant l'exigence dans FarmCPU que les QTN soient uniformément répartis dans des groupes de liaisons à travers le génome. Elle améliore également la vitesse par le remplacement du modèle à effets aléatoires et l'optimisation associée par un modèle à effets fixes utilisant l'optimisation des critères d'information bayésiens (BIC)	Huang et al., 2019
multilocus à effet SNP aléatoire (mrMLM)	mrMLM	Méthode multilocus à deux étapes, tout d'abord, elle utilise un modèle MLM pour scanner tous les marqueurs du génome et quelques SNP potentiellement associés au trait sont choisis (p-value < 0,01). Ensuite, les effets des marqueurs sont estimés par Espérance de Maximisation Empirique de Bayes (EMEB).	Wang et al., 2016
	FASTmrEMMA	Également à deux étapes. D'abord, la méthode EMMA est utilisée pour sélection les marqueurs potentiellement liés au trait (p-value < 0,005). Ensuite, les effets des marqueurs sont estimés EMEB.	Wen et al., 2018
	ISIS EM-BLASSO	Méthode à deux étapes, utilisent un même seuil de sélection que mrMLM (p-value < 0,01) dans la première étape. Ensuite, les effets de marqueurs sont estimés par Espérance de Maximisation LASSO-Bayésienne.	Tamba et al., 2017

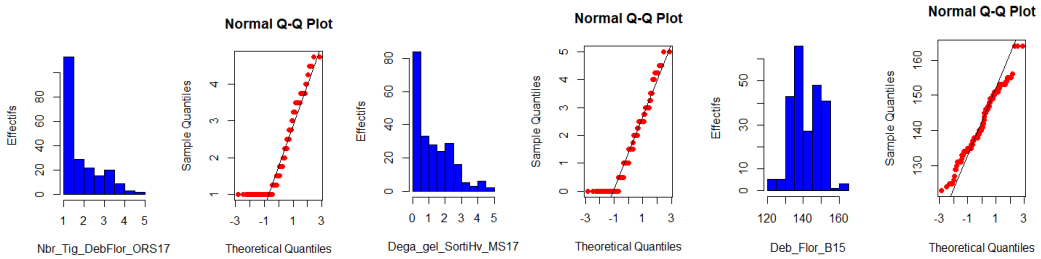
<u>Annexe 3</u>: Distribution et évaluation de la normalité des données pour chaque trait

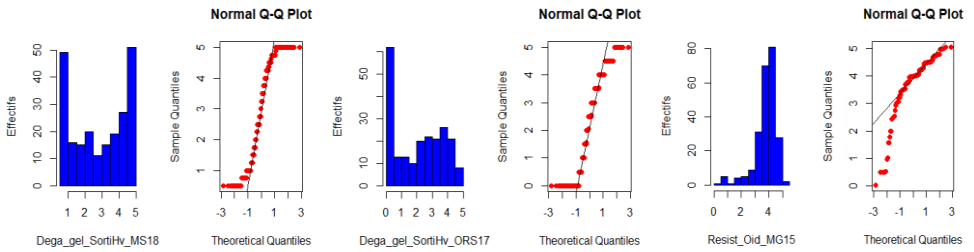
Histogramme de distribution (à gauche en bleu) et une courbe de normalité (qqnorm, à droite en rouge) pour chacun des 63 variables étudiées :

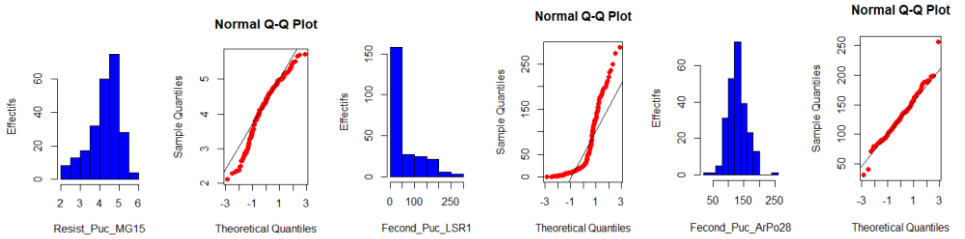


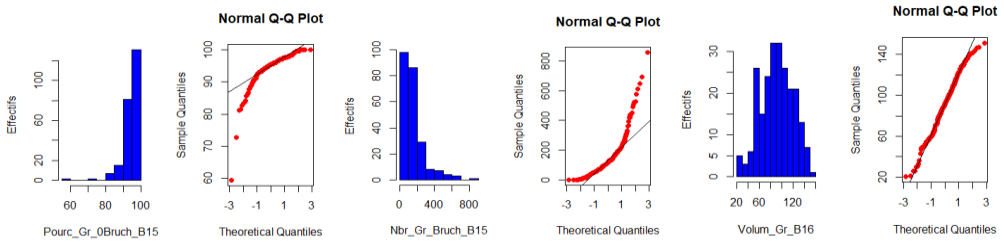

Normal Q-Q Plot
Effectifs
Deb_Rempl_Gr_B16
Sample Quantiles
Theoretical Quantiles
Fin_Flor_B15
Fin_Flor_B16
Fin_Flor_SemAut_MS17
Nbr_Jr_Flor_B16
Nbr_Jr_Flor_SemAut_MS17
Matur_B15
NDVI_Flor_MG15
Haut_Fin_Flor_MG15
Haut_Fin_Flor_B15
Haut_Fin_Flor_R16
Haut_Fin_Flor_R17
Nbr_Ramif_Basal_1Plt_MG15
Nbr_Ramif_Basal_1Plt_B15
Nbr_Ramif_Aer_1Plt_MG15

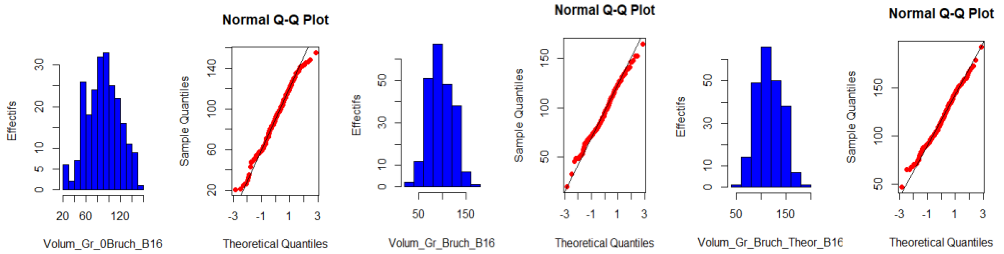

Normal Q-Q Plot
Effectifs
Sample Quantiles
Theoretical Quantiles
Biom_Aer_Matur_MG15
Nbr_Gr_1Plt_MG15
Nbr_Gr_1Plt_B15
Poids_Gr_1Plt_MG15
Poids_Gr_1Plt_B15
PMG_MG15
PMG_B15
PMG_B16
Tx_Prot_Gr_B15
Tx_Cu_Gr_B15
Tx_Fe_Gr_B15
Tx_Mn_Gr_B15
Tx_Zn_Gr_B15
Haut_TigPrinc_SortiHv_MS1
Haut_TigPrinc_SortiHv_ORS1

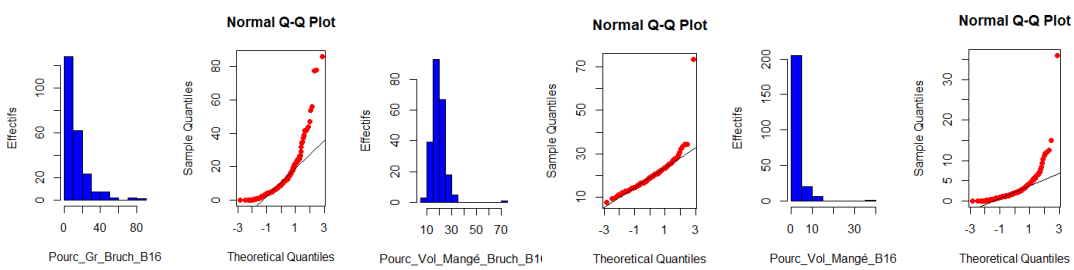

Normal Q-Q Plot
Effectifs
Sample Quantiles
Theoretical Quantiles
Nbr_Tig_DebFlor_ORS17
Dega_gel_SortiHv_MS17
Deb_Flor_B15


Normal Q-Q Plot
Effectifs
Sample Quantiles
Theoretical Quantiles
Dega_gel_SortiHv_MS18
Dega_gel_SortiHv_ORS17
Resist_Oid_MG15


Normal Q-Q Plot
Effectifs
Sample Quantiles
Theoretical Quantiles
Resist_Puc_MG15
Fecond_Puc_LSR1
Fecond_Puc_ArPo28


Normal Q-Q Plot
Effectifs
Sample Quantiles
Theoretical Quantiles
Pourc_Gr_0Bruch_B15
Nbr_Gr_Bruch_B15
Volum_Gr_B16


Normal Q-Q Plot
Effectifs
Sample Quantiles
Theoretical Quantiles
Volum_Gr_0Bruch_B16
Volum_Gr_Bruch_B16
Volum_Gr_Bruch_Theor_B16


Normal Q-Q Plot
Effectifs
Sample Quantiles
Theoretical Quantiles
Pourc_Gr_Bruch_B16
Pourc_Vol_Mangé_Bruch_B1
Pourc_Vol_Mangé_B16

Résumé

Diplôme et Mention : Master Biologie, Agrosciences

Parcours : Amélioration, Production et Valorisation du Végétal

Option : Génétique Génomique et Amélioration des Plantes

Responsable d'option : Mélanie JUBAULT

Auteur(s) : Mamadou SENE

Date de naissance* : 30/04/1992

Nb pages : 28 Annexe(s) : 7

Année de soutenance : 2022

Organisme d'accueil : INRAE Bourgogne-Franche-Comté

Adresse : 17 Rue Sully, 21000 Dijon

Maîtres de stage : Nadim TAYEH et Jonathan KREPLAK

Titre français : Evaluation de méthodes et optimisation de protocoles de GWAS à locus unique et multi-locus pour l'identification de régions génomiques contrôlant des caractères d'intérêt chez le pois (*Pisum sativum* L.)

Titre anglais : Evaluation of methods and optimization of single and multi-locus GWAS protocols for the identification of genomic regions controlling traits of interest in pea (*Pisum sativum* L.)

Résumé

Relier les génotypes aux traits permet de révéler l'architecture génétique de ces derniers et de localiser les régions génomiques les contrôlant ouvrant ainsi la voie pour comprendre les mécanismes biologiques sous-jacents et pour utiliser les connaissances en amélioration variétale. L'étude d'association pangénomique ou GWAS est une approche couramment utilisée pour disséquer les bases génétiques des traits complexes. Cette présente étude visait à optimiser la conduite de GWAS chez le pois qui est une légumineuse cultivée pour ses graines riches en protéines. Pour ce faire, des études GWAS avec différentes méthodes à locus uniques et multilocus ont été réalisées en exploitant les données disponibles pour un panel de 240 accessions de pois. Les accessions ont été génotypées avec un ensemble de plus de 1.900.000 SNP issus de génotypage par capture d'exome et ont été phénotypées dans des contextes pédoclimatiques différents pour 41 traits. Les résultats obtenus sur deux versions du génome ont suggéré une meilleure façon d'optimiser la structuration des panels dans les modèles de GWAS et ont mis en évidence trois meilleures méthodes de GWAS à locus unique et multilocus : CMLM, FarmCPU et mrMLM. Ces méthodes se distinguent par leur capacité à identifier des associations marqueurs-traits et par la puissance statistique sous-jacente. Croiser les données obtenues avec les connaissances dans la littérature ont permis de démontrer la robustesse des résultats et le potentiel des optimisations pour déceler des nouvelles associations.

Abstract

Linking genotypes to phenotypes makes it possible to reveal the genetic architecture of plants traits and to locate the genomic regions controlling them, thus opening the way to understanding the underlying biological mechanisms and to using knowledge in varietal improvement. The genome-wide association study or GWAS is a commonly used approach to dissect the genetic bases of complex traits. This present study aimed at optimizing the conduct of GWAS in pea, which is a legume grown for its protein-rich seeds. To do this, GWAS studies with different single-locus and multi-locus methods were carried out by exploiting the data available for a panel of 240 pea accessions. A set of over 1,900,000 SNPs obtained through exome capture genotyping is available and the accessions were phenotyped in different pedoclimatic contexts for 41 traits. The results obtained on two genome versions suggested a better way to optimize panel structuring in GWAS models and highlighted three best single-locus and multi-locus GWAS methods: CMLM, FarmCPU and mrMLM. These methods are distinguished by their ability to identify marker-trait associations and by the underlying statistical power. Cross-checking the data obtained in this work with knowledge available in the literature made it possible to demonstrate the robustness of the results and the potential for optimizations to detect new associations in pea.

Mots-clés : GWAS, pois (Pisum sativum L.), méthodologie, traits d'intérêt

Keywords: GWAS, pea (*Pisum sativum* L.), methods, traits