Recherche par génétique d'association de gènes impliqués dans l'interaction pois x rhizobium

Valentin Delefortrie

AGROCAMPUS OUEST

☐ **CFR Angers** ☑ **CFR Rennes**

Année universitaire : 2017-2018

Spécialité : APVV
(Amélioration, Production, Valorisation du Végétal)

Spécialisation (et option éventuelle) :
GGAP (Génétique, Génomique et Amélioration des Plantes)

Mémoire de Fin d'Études

☐ d'Ingénieur de l'Institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage

☑ de Master de l'Institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage

☐ d'un autre établissement (étudiant arrivé en M2)

RECHERCHE PAR GENETIQUE D'ASSOCIATION DE GENES IMPLIQUES DANS L'INTERACTION POIS x RHIZOBIUM

Par : Valentin DELEFORTRIE

Rapport Confidentiel

Soutenu à Rennes le 28/06/ 2018

Devant le jury composé de :

Président : Mélanie Jubault

Maître de stage : Virginie Bourion

Enseignant référent : Mélanie Jubault

Autres membres du jury (Nom, Qualité) :

Examinateur : Maria Manzanares-Dauleux

Rapporteur : Marie-Laure Pilet-Nayel

Confidentialité

☐ Non ☑ Oui si oui : ☐ 1 an ☑ 5 ans ☐ 10 ans

Pendant toute la durée de confidentialité, aucune diffusion du mémoire n'est possible [1].

Date et signature du maître de stage [2] : 14/06/2018

A la fin de la période de confidentialité, sa diffusion est soumise aux règles ci-dessous (droits d'auteur et autorisation de diffusion par l'enseignant à renseigner).

Droits d'auteur

L'auteur[3] **Delefortrie Valentin**

autorise la diffusion de son travail (immédiatement ou à la fin de la période de confidentialité)

☑ Oui ☐ Non

Si oui, il autorise

☐ la diffusion papier du mémoire uniquement(4)

☐ la diffusion papier du mémoire et la diffusion électronique du résumé

☑ la diffusion papier et électronique du mémoire (joindre dans ce cas la fiche de conformité du mémoire numérique et le contrat de diffusion)

(Facultatif) ☐ accepte de placer son mémoire sous licence Creative commons CC-By-Nc-Nd (voir Guide du mémoire Chap 1.4 page 6)

Date et signature de l'auteur : 14/06/2018 Delefortrie

Autorisation de diffusion par le responsable de spécialisation ou son représentant

L'enseignant juge le mémoire de qualité suffisante pour être diffusé (immédiatement ou à la fin de la période de confidentialité)

☐ Oui ☐ Non

Si non, seul le titre du mémoire apparaîtra dans les bases de données.

Si oui, il autorise

☐ la diffusion papier du mémoire uniquement(4)

☐ la diffusion papier du mémoire et la diffusion électronique du résumé

☐ la diffusion papier et électronique du mémoire

Date et signature de l'enseignant :

(1) L'administration, les enseignants et les différents services de documentation d'AGROCAMPUS OUEST s'engagent à respecter cette confidentialité.
(2) Signature et cachet de l'organisme
(3) Auteur = étudiant qui réalise son mémoire de fin d'études
(4) La référence bibliographique (= Nom de l'auteur, titre du mémoire, année de soutenance, diplôme, spécialité et spécialisation/Option)) sera signalée dans les bases de données documentaires sans le résumé

Remerciements

Je tiens tout d'abord à remercier Virginie Bourion pour m'avoir formé et suivi durant ce stage.

Je tiens à remercier Coraline Bris pour m'avoir formé sur de nombreux logiciels et m'avoir appris de nombreuses choses sur les études de GWAS.

Je tiens à remercier Mathieu Siol pour sa pédagogie et pour son aide durant ce stage.

Je tiens à remercier Jonathan Kreplak pour sa pédagogie et son aide en bio-informatique.

Je tiens à remercier Adrien, Amandine, Aurélie, Aurélien, Fati, Luc, Mayla, Morgane, Mégane, Kevin, Nadim, Thibault pour leurs soutiens lors de ce stage et les nombreuses sorties !

Je tiens à remercier le reste de l'équipe GEAPSI pour la bonne ambiance lors de ce stage.

Glossaire

Accession : Lot identifié de semences d'une collection (banque) de semences.

Allèle : Variant d'un gène à un locus.

Déséquilibre de liaison : Associations préférentielles entre allèles à différents locus.

Inférence bayésienne : Méthode statistique qui déduit la probabilité d'un événement à partir d'autres événements déjà évalués.

Lixiviation : Passage dans la couverture pédologique et vers les nappes phréatiques de substances dissoutes (nitrates, phosphates, produit phytosanitaires…).

PCR : Méthode d'amplification de l'ADN basée sur trois grandes phases : la première correspondant à la dénaturation, la seconde à l'hybridation et la dernière à l'élongation. Ces phases sont spécifiques à des températures précises en fonction des nucléotides des séquences à amplifier.

Scarification : Traitement des semences par écorchage des téguments afin de faciliter la germination des graines. Chez le pois, la scarification est généralement faite à l'aide d'un scalpel.

Symbiose : Association préférentielle réciproque entre deux organismes.

Synténie : Groupe de gènes dont le voisinage et l'organisation sont conservés sur plusieurs génomes.

Abréviations

ADN : Acide DésoxyriboNucléique

ADNr : Acide DésoxyriboNucléique ribosomique

ACP : Analyse en Composante Principale

ANOVA : ANalysis Of Variance

ANR : Agence Nationale de la Recherche

BIC : Bayesian Information Criterion

DL : Déséquilibre de Liaison

FAO : Food and Agriculture Organization of the United Nations

FST : Indice de fixation

GRaSP : Genetics of Rhizobia Selection by Pea

GWAS : Genome-Wide Association Study

Ha : Hectare

INRA : Institut National de la Recherche Agronomique

LSTM : Laboratoire des Symbioses Tropicales et Méditerranéennes

MAF : Minor Allele Frequency

MLMM : Multivariate Linear Mixed Model

NGS : New Generation Sequencing

Pb : Paire de bases

PCR : Polymerase Chain Reaction

PeaMUST : Adaptation Multi-Stress et régulations biologiques pour l'amélioration du rendement et de la stabilité du POIS protéagineux

QTL : Quantitative Trait Locus

Rlv : *R. leguminosarum* sv *viciae*

SNP : Single Nucleotide Polymorphism

SSR : Simple Sequence Repeat

TILLING : Targeting Induced Local Lesions in Genomes

Liste des annexes

Liste des illustrations

Table des matières

Préface

Avec une estimation de la FAO qui porte la population mondiale à 10 milliards d'individus à l'horizon 2050, l'agriculture se trouve face à un défi de taille à relever : être toujours plus productive pour répondre aux besoins d'une population en constante augmentation tout en devenant plus écoresponsable. Les légumineuses pourraient participer à ce challenge. En effet, elles constituent une famille de plantes importante avec des taux protéiques nutritifs de l'ordre de 20 à 24% **(FAO, 2016)** et une capacité à fixer l'azote (grâce aux bactéries en symbiose avec elles) intéressante d'un point de vue environnemental et économique.

La nutrition azotée de la légumineuse se fait via 2 voies distinctes : une par l'alimentation azotée racinaire présente chez toutes les plantes et l'autre par la voie de fixation symbiotique bactérienne régit par des mécanismes de reconnaissances plante/rhizobium formant ainsi les nodosités dans lesquelles des bactéries fixent l'azote atmosphérique **(Bourion *et al*. ,2010)**.

Celle-ci se fait par le biais d'une symbiose plante/rhizobium : les rhizobia fixent l'azote atmosphérique qui sera alors assimilable grâce à différents mécanismes biochimiques propres aux légumineuses en échange de photosynthétats. La fixation symbiotique est sensible à différents stress environnementaux biotiques et abiotiques et reste très complexe à étudier **(Prudent *et al*., 2016)**.

Diverses études démontrent que la reconnaissance plante/bactérie varie selon le génotype de la plante hôte **(Laguerre *et al*., 2007; Depret and Laguerre, 2008)** ainsi que l'existence de spécificités entre certaines souches et certaines accessions de légumineuses **(Davis *et al*., 1988; Bourion *et al*., 2018)**.

L'objectif de ce stage est de participer à la recherche et à la découverte des gènes ou des locus de caractères quantitatifs (QTL) impliqués dans la nodulation chez le pois.
Nous avons recherché des gènes spécifiques ou QTL de choix par différents génotypes de pois entre différentes souches de *Rhizobium leguminosarum* sv. *viciae*.

Cette recherche est réalisée suivant une analyse de génétique d'association qui cherche à corréler la variation phénotypique d'un caractère avec des polymorphismes à un ou plusieurs locus. Ce type d'analyse a connu un fort développement dans les dernières décennies du fait des techniques de génotypage à haut débit couplées à un coût toujours plus réduit. La génétique d'association peut se pratiquer à l'échelle du génome entier, pour peu que l'on dispose de suffisamment de marqueurs le long du génome. Cette démarche doit permettre d'identifier les marqueurs en déséquilibre de liaison avec un polymorphisme responsable de la nodulation chez le pois.

Une fois ces marqueurs ou gènes identifiés, l'idée est d'introgresser les allèles d'intérêts (si leurs impacts bénéfiques est validé) dans des cultivars actuels aux caractéristiques agronomiques intéressantes. Les nouvelles variétés ainsi créées devraient, de ce fait, permettre une meilleure valorisation de l'azote atmosphérique et réduire significativement les pollutions azotées dans les agrosystèmes.

La comparaison de différentes méthodes de génétique d'association a été mise en œuvre lors de mon stage afin de déterminer le meilleur modèle d'analyse à utiliser.

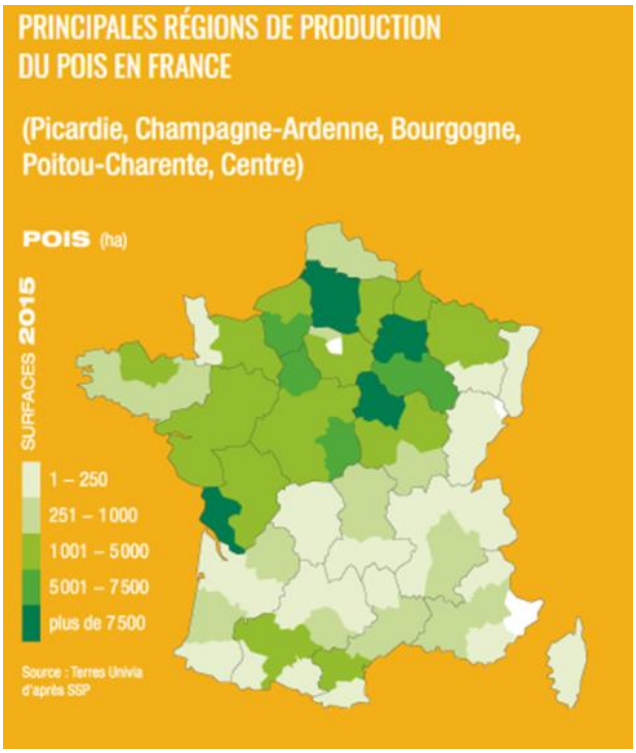


Figure 1: Répartition de la culture du pois en France **(Terres Univia, 2015)**

La culture du pois est majoritairement présente en Picardie, Champagne-Ardenne, Bourgogne, Poitou-Charentes et dans le Centre.

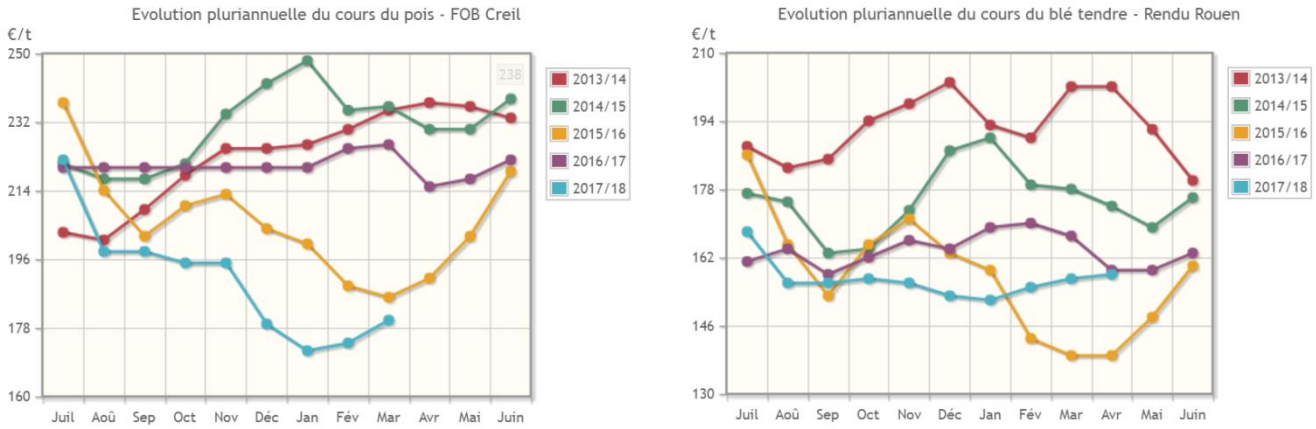


Figure 2 : Comparaison des cours du pois et du blé **(BASF France)**

Les cours du pois et du blé fluctuent très fortement selon les années avec un prix actuel de 180 €/Tonne pour le pois (mars 2018). Le prix du pois est normalement plus élevé que celui du blé.

Contexte général

1) L'importance de l'azote

Elément essentiel à la vie, principalement dans la biosynthèse des acides aminés et donc des protéines, l'azote se présente principalement sous forme de diazote gazeux et constitue 78.1% de l'atmosphère **(Ferguson *et al*., 2010)**. Il s'agit d'une molécule stable formée par triple liaison entre deux atomes d'azote et peu valorisable par les êtres vivants excepté certaines bactéries qui possèdent des nitrogénases dont le fonctionnement est très coûteux en énergie.

Avec l'essor de l'agriculture industrielle, due notamment à la révolution verte et l'utilisation de procédé d'Haber-Bosch pour la fabrication des engrais, de nouvelles variétés valorisant les apports azotés ont été développées. En moyenne 120MT d'engrais industriels sont utilisées annuellement dans le monde **(Galloway *et al*., 2008)**. Cette utilisation massive d'engrais industriels impacte grandement l'environnement en raison de lixiviations d'azote dans les nappes phréatiques d'une part, et d'émissions de gaz à effet de serre dans l'air d'autre part. En outre la fabrication de ces engrais de synthèses représente un fort coup énergétique.

2) Le pois cultivé

Le pois cultivé est une espèce diploïde (2n=14) appartenant à la famille des Fabaceae au genre *Pisum* et à l'espèce *P.sativum* **(Kaló *et al*., 2004)**. Il s'agit d'une espèce annuelle connue comme une des premières espèces cultivées **(Zohary *et al*., 2012).** Les espèces sauvages connues sont *Pisum fulvum* et *Pisum sativum* sbsp. *elatius* ; elles sont originaires du Moyen-Orient. La domestication du pois date d'environ 10 000 ans **(pour revue Smýkal *et al*., 2011)** ; sa culture s'est ensuite répandue dans toute l'Europe, en passant par la Turquie, la Grèce et le Caucase, et à l'est, en Inde et Chine, en passant par l'Iran et l'Afghanistan **(Zohary, 2012)**. Les pois cultivés appartiennent principalement à la sous-espèce *P.sativum* ; *P. sativum* subsp. *abyssinicum* est un pois dont la culture se limite au Yémen et à l'Ethiope. Le pois est la troisième légumineuse la plus cultivée au monde, après le soja et le haricot, avec 11 000 kT de production annuelle **(Terres Univia, 2017)**.

En France, le pois est principalement cultivé dans les régions du Nord et du Centre (Figure 1) avec 140 000 hectares semés sur l'année 2017 pour une production de 590 000 tonnes **(FranceAgriMer, 2017)**. Les rendements de cette espèce végétale peuvent avoisiner les 60 q/hectare mais sont très hétérogènes selon les années notamment à cause des stress biotiques et abiotiques **(Terres Univia, 2017)**. Malgré les primes versées aux agriculteurs pour promouvoir cette culture, le pois demeure peu attractif en raison des prix fluctuants et des rendements plutôt faibles (environ 45 à 60 q/ha) par rapport à ceux du blé (72 q/ha) (Figure 2). Le pois est principalement cultivé pour sa grande valeur protéique (20-24% de protéines pour un grain sec) **(Terres Univia, 2017)** et il est destiné à la fois à l'alimentation animale et humaine.

Les bactéries symbiotiques fixatrices d'azote spécifiques du pois sont des *Rhizobium leguminosarum* sv. *viciae* **(Bourion *et al*., 2018)** . Cette association, bénéfique pour la plante et les bactéries, permet de limiter l'utilisation d'azote dans les cultures qui succèdent à la culture de pois lors de la rotation culturale.

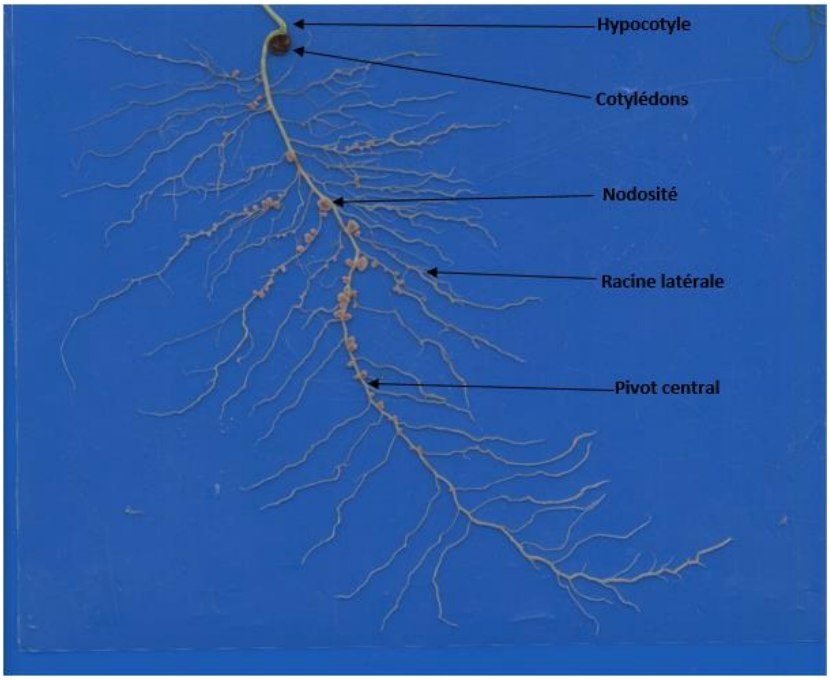


Figure 3 : Nodosités observées sur une racine de pois (×2)

Les nodosités se répartissent sur le pivot mais aussi sur les racines latérales avec une taille variant en fonction de l'âge des nodosités.

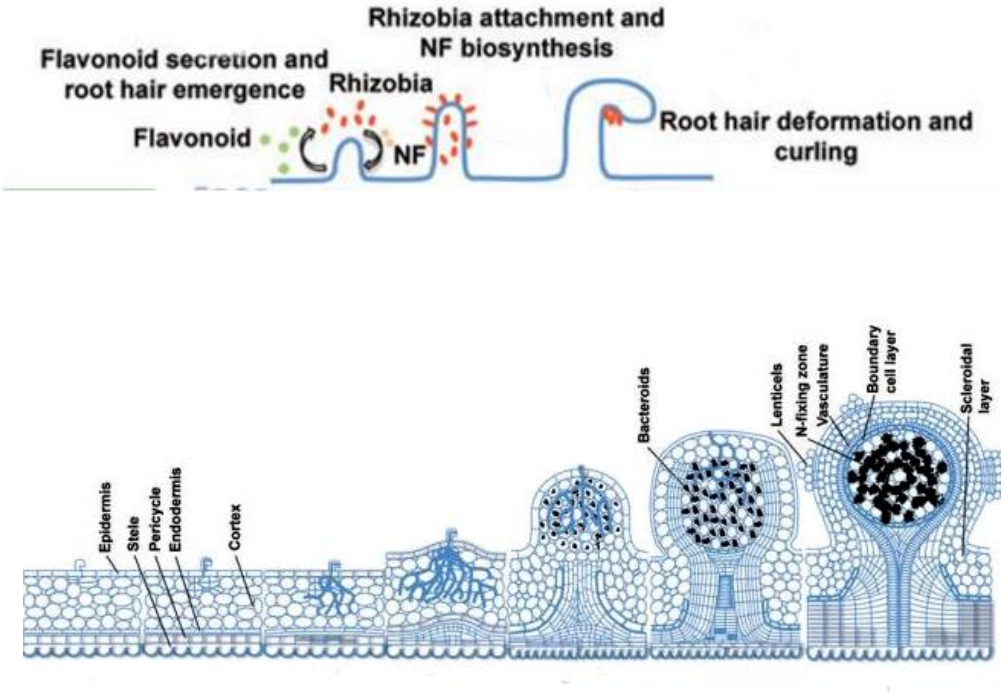


Figure 4 : interaction plantes-bactéries et formation des nodosités (d'après **Ferguson *et al*., 2010**)

La reconnaissance des deux partenaires se fait par émissions de composés biochimiques dans le milieu extra racinaire. La plante hôte émet des métabolites secondaires (flavonoïdes) et les bactéries émettent des facteurs Nod. Les cordons infectieux vont alors se former et les bactéroïdes se développeront dans le système racinaire de la plante pour assurer les échanges symbiotiques.

Le pois possède un génome de 4.45 Gb **(Dolezel *et al*., 1998; Dolezel and Greilhuber, 2010)**. Un consortium international a été lancé et a permis de séquencer récemment l'entièreté du génome **(International Conference on Legume Genetics and Genomics, 2017)**. En parallèle, différents projets menés à l'INRA par l'équipe de Dijon (ANR Genopea ; ANR GRaSP ; Projet Investissement d'avenir PeaMUST) ont permis le génotypage de large collection d'écotypes de pois permettant ainsi de mener des approches de génétique d'association.

3) Les rhizobia et la nodogénèse

Découvertes en 1888, les bactéries de genre rhizobium permettent la fixation de l'azote au sein d'organes spécifiques, les nodosités, qui se développent sur les racines de la plante hôte (Figure 3). Il existe une grande diversité génétique de rhizobia, comme cela est révélé par de l'ADNr 16S **(Young *et al*., 2004).** Aujourd'hui, cette diversité comporte 98 espèces en 13 genres **(http://www.rhizobia.co.nz/taxonomy/rhizobia)**. La majorité de ces bactéries appartient au groupe des Alphaprotéobactéries **(Willems, 2006)** et une infime partie à celui des bêtaprotéobactéries. Les études taxonomiques concernant les propriétés de fixations d'azote se concentrent plus particulièrement sur les gènes induisant la nodulation et la fixation d'azote par le pois. En effet, l'analyse des ADNr 16S ne distingue pas la variabilité de fixation symbiotique d'individus d'une même espèce **(Peix *et al*., 2015)**.

La formation de nodosités se fait par le biais d'échanges biochimiques entre les bactéries et le système racinaire de la plante (Figure 4). Il existe différents moyens d'attraction des bactéries par la plante mais le plus connu est celui d'émission de flavonoïdes (ou iso-flavonoïdes). Les flavonoïdes activent la protéine NodD, qui est un facteur de transcription des gènes bactériens *nodD*. Ces flavonoïdes sont des métabolites secondaires très polymorphes ; la reconnaissance entre les protéines NodD d'une espèce de rhizobium et leur flavonoïde activateur est un des premiers mécanismes de spécificité entre les deux partenaires **(Maj *et al*., 2010)**. En plus des gènes *nodD*, il existe deux autres classes de gènes *nod* : les gènes *nodABC* et le gène décoratif de l'hôte **(Broughton *et al*., 2000)**. Les gènes *nodABC* participent à la constitution du squelette de type chitine des facteurs Nod. Les facteurs Nod sont des lipochitine-oligosaccharides formés de trois à six (variation selon les espèces) résidus N-acetyl-glucosamines (chitine) avec une extrémité réductrice et une extrémité non réductrice **(Broughton *et al*., 2000)**. Les gènes décoratifs présents chez certaines espèces bactériennes permettent une variation des extrémités réductrices et non réductrices.

Il existe une diversité dans les facteurs Nod produits au sein d'une même espèce de bactérie. C'est par exemple le cas des souches de *R. leguminosarum* sv *viciae* (Rlv) qui nodulent le pois : il a été montré que des souches Rlv doivent posséder des gènes *nod* spécifiques pour permettre la nodulation sur certaines accessions de pois ; ainsi certains pois originaires d'Afghanistan chez lesquels le gène *PsSYM2* est sous une forme allélique particulière ne peuvent être nodulés que par des Rlv possédant le gène spécifique *nodX* **(Davis *et al*., 1988).**

Les facteurs Nod sont reconnus sur la plante hôte par des récepteurs de type kinase composés de lysine : les LysM-RLKs. Ces récepteurs sont formés d'un domaine intracellulaire et d'un domaine extracellulaire composé de plusieurs motifs de lysine. Ils sont codés par une famille multigénique (au moins 20 membres) **(Gough and Cullimore, 2011).** La majorité des gènes impliqués dans la nodulation a été identifiée chez les deux espèces modèles de légumineuses, *Lotus japonicus* (*Lj*) et *Medicago truncatula* (*Mt*).

Quelques gènes majeurs impactant la nodulation chez le pois ont été identifiés, comme les gènes *PsSYM10 (orthologues de MtNFP et LjNFR5) et PsK1* codant pour des LysM-RLKs et *PsSYm37* impactant la croissance du cordon d'infection **(Geurts *et al*., 2005; Zhukov *et al*., 2008)**. Un autre gène, *PsySYM19,* codant pour un récepteur sérine/thréonine kinase a été identifié comme impliqué dans la perception des facteurs Nod.

Enfin d'autres gènes connus impactent le trait de nodulation chez le pois **(Borisov *et al*., 2004)** et d'autres gènes impactent le trait de nodulation chez *Medicago truncatula*. Par exemple *MtDME* code une DNA déméthylase : ce sont des régulateurs épigénétiques qui modifient le développement des nodosités ; *MtEFD* code pour un facteur de transcription qui régule également le développement nodulaire chez *Medicago truncatula* **(Satgé *et al*., 2016)**.

4) La symbiose bactérienne

Suite à la reconnaissance entre les deux partenaires, la symbiose commence par la création d'un « cordon infectieux », lieu où les bactéries pénètrent dans la plante. Une fois les bactéries au sein de la plante, elles se différencient en bactéroïdes. Les bactéroïdes sont des bactéries entourées d'une paroi péribactéroïdale formée par la plante **(Oldroyd et Downie, 2008)**.

Les bactéroïdes vont réduire l'azote atmosphérique (N_2) en azote ammoniacal (NH_3) grâce à la nitrogénase (enzyme bactérienne). Cet azote est alors assimilable par la plante pour être transformé en molécules organiques comme les protéines. En échange de cet azote la plante va apporter aux bactéroïdes de l'énergie sous forme de photosynthétats nécessaires à la synthèse des nodosités **(Voisin *et al*., 2015).**

Cette symbiose est très sensible aux stress biotiques et abiotiques. Les facteurs impactant la croissance de la plante vont jouer sur la symbiose comme par exemple la quantité d'azote présente dans l'agrosystème. En effet, si de l'azote est déjà présent dans l'agrosystème alors la plante ne va pas élaborer de symbiose (coup énergétique). Les conditions pédoclimatiques sont aussi très importantes pour favoriser la symbiose bactérienne. De plus, les ravageurs de cultures (bruches, sitones…) vont limiter la symbiose par phytophagie **(Voisin *et al*., 2015).**

5) La génétique d'association dans la recherche d'interactions pois/rhizobia

La génétique d'association est une analyse statistique des associations entre la variation quantitative pour un caractère mesuré sur une population d'individus et la variabilité génétique de ces individus. L'étude de la variabilité génétique peut se limiter à certains gènes candidats supposés avoir un impact sur le caractère étudié mais, depuis quelques années, elle est plus généralement réalisée à l'échelle du génome (GWAS, Genome Wide Association Study). C'est cette analyse qui a été choisie dans le cadre de notre étude des déterminants génétiques du choix des souches de Rlv chez le pois.

L'analyse de génétique d'association repose sur la mesure du déséquilibre de liaison (DL) entre des marqueurs génétiques (SNP, SSR, …) et des polymorphismes génétiques causaux de la variation phénotypique au sein d'un panel d'individus **(Yu and Buckler, 2006**). Le DL correspond à la mesure de la ségrégation non aléatoire entre deux loci. Dans le cas où, par exemple, A et B sont deux loci bi-alléliques avec leurs allèles respectifs A/a et B/b, le DL correspond à la différence observée entre la fréquence de l'haplotype AB et le produit des fréquences des allèles A et B **(Zhu *et al*., 2008)**.

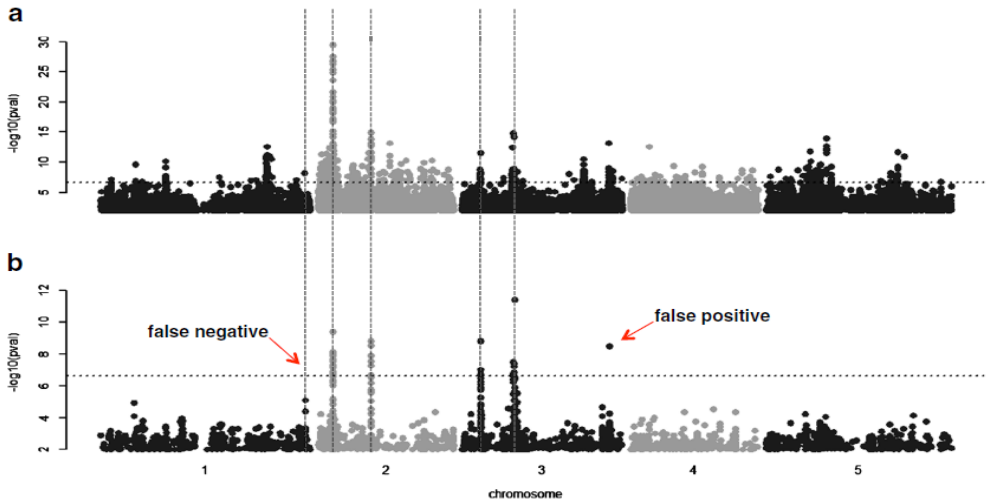


Figure 5 : Impact de la prise en compte de la structuration de l'échantillon sur la performance de GWAS (d'après **Korte and Farlow, 2013**)

Manhattan plots issus de l'analyse de GWAS pour un trait simulé : (a) modèle linéaire sans prise en compte de la structuration et de la kinship et (b) modèle mixte prenant en compte la structuration et la kinship. Le modèle corrigé (mixte) est plus stringent et élimine des faux positifs. A noter qu'il apparait un faux négatif.

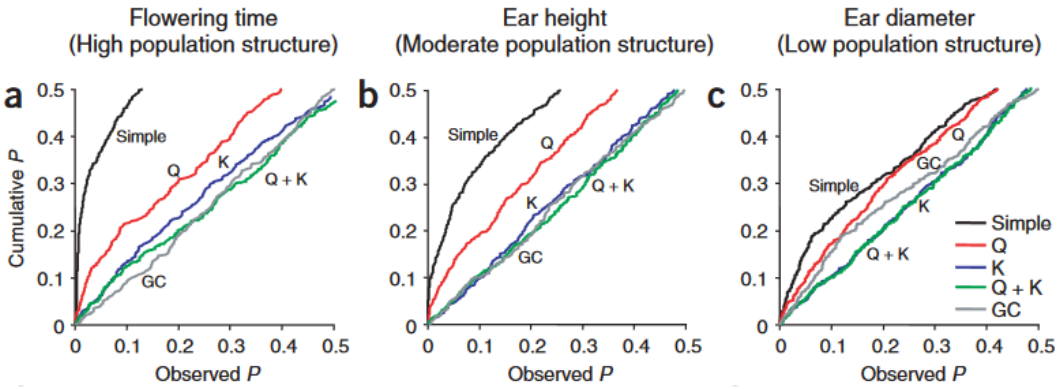


Figure 6 : Inflation du taux de faux-positifs en fonction de la structuration et de la kinship chez le maïs sur trois covariables différentes **(Yu *et al*., 2006).**

Analyses de SNP aléatoires (non-associés) sur la date de floraison (a), la longueur de l'épi (b) et le diamètre de l'épi (c).Sous l'hypothèse nulle de non-association, on attend une distribution uniforme des p-values (le long de la diagonale).

Les modèles prenant en compte seulement la kinship (K) sont plus efficaces que ceux prenant en compte la structure (Q) seule. Les modèles prenant à la fois la structure et la kinship sont plus efficaces contre les erreurs de la génétique d'association. La seule prise en compte de la kinship dans le modèle dépend du niveau de structuration de la population : plus la structuration est faible et plus la kinship suffit à elle-même pour contrôler les faux et vrais positifs.

La génétique d'association capitalise sur l'existence d'un nombre important de recombinaisons (nombreuses méioses) au sein de l'échantillon ce qui permet l'identification de blocs de DL plus courts que dans les analyses de liaisons classiques. Néanmoins, cela nécessite de bénéficier d'une densité en marqueurs suffisante. En conséquence, les panels de génétique d'association comportent généralement des individus plus divers et à l'histoire de recombinaison plus ancienne **(Korte and Farlow, 2013)**.

Les modèles de génétique d'association sont généralement basées sur le formalisme des modèles linéaires suivant : $y_i = \mu + x_i\beta + e_i$; où y_i est le phénotype du caractère d'intérêt pour l'individu i, μ la moyenne de la population (fixe), x_i le génotype au marqueur, β l'effet du marqueur (fixe) et e_i l'erreur résiduelle (aléatoire).

L'effet de la stratification des échantillons sur l'inflation du taux de faux-positifs dans les analyses de GWAS a été identifié depuis longtemps **(Astle and Baldingn, 2009).** Dans le cadre des analyses d'associations, on parle de faux positifs quand le résultat de l'analyse indique qu'il existe une association alors qu'il n'y en a pas. L'utilisation de modèles linéaires mixtes prenant en compte le degré d'apparentement entre individus et la structuration en populations permet de réduire la détection de faux positifs (Figure 5) de manière significative.

En effet, la structuration crée du DL tout le long du génome entre des loci non liés physiquement **(Yu *et al*., 2006**). L'hypothèse de distribution identique et indépendante des marqueurs indispensable aux analyses de GWAS n'est donc pas respectée en cas de très forte structuration. La prise en compte de la structuration permet de supprimer la covariance génétique due à la présence d'individus dans des populations distinctes et donc de limiter les faux positifs **(Astle and Balding, 2009)**. L'estimation de la structuration et de l'apparentement (kinship) se fait sur les bases de données génétiques **(Hoffman, 2013)**.

D'autres modèles d'analyses couplant modèles linéaires mixtes et régression stepwise mettent en évidence certains locus d'effets forts en covariables, ce qui peut permettre la détection de SNP dont l'effet aurait, par ailleurs, été masqué par les locus d'effet forts **(Segura *et al*, 2012)**.

Des recherches de Yu et de son équipe (2006) montrent que les modèles exprimant le moins de faux positifs sont ceux prenant en compte à la fois l'apparentement et la structuration du panel étudié dans le cas d'une structuration suffisamment forte (Figure 6). En cas de structuration faible, la kinship, seule, peut corriger les faux positifs.

Problématique : Quels sont les SNP significatifs en déséquilibre de liaison avec des loci causaux impliqués dans la nodulation chez les partenaires symbiotiques pois/rhizobia ? Quel est l'effet des différentes structurations issues des logiciels sur la GWAS ?

La recherche par GWAS de nouveaux gènes ou QTL impliqués dans la nodulation et la reconnaissance pois/bactérie est novatrice. En effet, d'autres gènes ou QTL peuvent impacter la nodulation et la reconnaissance pois/bactérie que ceux déjà connus. C'est dans cette démarche que s'inscrit mon stage. Pour cela, j'ai mené une étude de GWAS à partir de données phénotypiques de nodulation déjà acquises dans un projet précédent (SYMBIOPEA), dans lequel différents caractères phénotypiques avaient été mesurés sur 104 accessions de pois inoculées par un mélange de 5 souches de rhizobium. J'ai aussi participé à la mise en place et au suivi d'une nouvelle expérimentation, à plus grande échelle, sur la plateforme de phénotypage haut-débit (4PMI) de Dijon. Cette expérimentation menée dans le cadre du projet ANR GRaSP est destinée à réaliser une étude GWAS sur un panel de 337 accessions inoculées par un mélange de 28 souches.

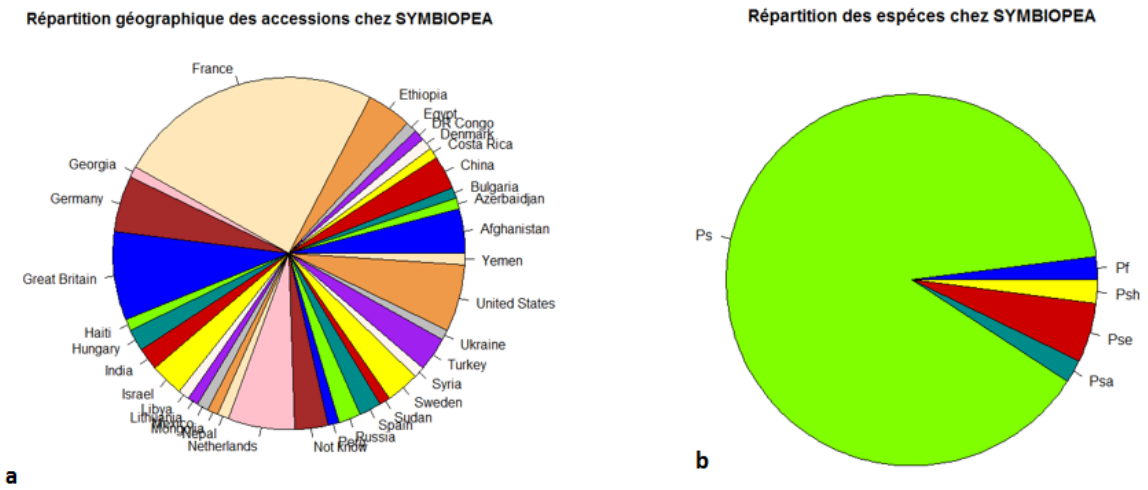


Figure 7 : Répartition par pays et par espèce des accessions composant SYMBIOPEA

La répartition géographique (a) du sous-ensemble de 98 accessions s'étend sur 36 pays et la répartition des espèces (b) répertorie 6 groupes. Avec Pf : *Pisum fulvum*, Psa : *Pisum* ***sativum*** subsp. *abyssinicum*, Ps *: Pisum sativum*, Pse : *Pisum sativum* subsp. *elatius*, Psh : *Pisum sativum subsp. Humile.*

Figure 8 : Expérimentation SYMBIOPEA

Le plan d'expérimentation se répartit en 3 blocs comprenant chacun deux pots par accession. Les conditions d'hygrométrie, de température et de luminosité sont contrôlées.

Matériels et Méthodes

1) Matériel végétal et bactérien

Le travail décrit dans ce rapport a été réalisé sur 104 accessions de pois issues de la collection de référence de l'INRA de Dijon **(http://www.thelegumeportal.net)**. Le but de l'expérimentation était d'étudier la diversité de choix de souches de Rlv par les différentes accessions de pois et de vérifier si ce choix est déterminé par l'efficience de fixation symbiotique obtenue **(Bourion *et al*., 2018)**. Parmi ces 104 accessions, 98 (cf. Annexe I) ont été génotypées avec 3,8 millions de marqueurs SNP (projets PeaMUST et GRaSP) en utilisant l'approche de capture d'exome. C'est sur ce sous-ensemble que j'ai réalisé les analyses de GWAS.

Les accessions de pois de ce projet proviennent de nombreux pays avec une majorité d'accessions originaires de France (figure 7a). Six espèces ou sous-espèces sont représentées, mais la majorité est *P. sativum* (figure 7b). Les accessions ont des statuts variés : ce sont des cultivars, des lignées, des variétés population ou des accessions sauvages.

Pour cette expérimentation SYMBIOPEA, cinq souches de Rlv d'origines géographiques diverses ont été utilisées : une souche provient d'Angleterre (SK), deux souches de France (SA et SD), une souche d'Algérie (SE) et une souche, connue comme spécifique à certaines accessions afghanes (SF), appelée souche TOM et collectée en Turquie. L'arbre phylogénétique réalisé sur un grand nombre de souches montre que ces 5 souches sont assez distantes génétiquement (Annexe II ; **Bourion *et al*., 2018**).

2) Plan d'expérimentation

L'expérimentation s'est déroulée sous serre (Figure 8) avec des conditions de températures, de luminosité et d'hygrométrie contrôlées. La température était de 21°C le jour et de 16°C la nuit avec un cycle jour-nuit de 16 heures sous rayonnement moyen photosynthétique actif (RPA) de 250 µmol photons m^{-2} s^{-1} et par l'utilisation de lampes à sodium à haute pression lorsque la lumière du jour diminuait.

Avant semis, 12 graines de chaque accession ont été choisies parmi un échantillon homogène de graines pour chacune d'elle. L'échantillon de graines choisies a été pesé de façon à déterminer le poids de mille grains de chacune des accessions. Les graines de certaines accessions ont dû être scarifiées pour permettre leur germination : il s'agit des graines appartenant à l'espèce *P. fulvum* et à certaines accessions de *P. sativum* subsp. *elatius ou P. sativum humile.* Ensuite, toutes les graines ont été désinfectées par traitement à 10% d'eau de Javel pendant 10 minutes suivi de cinq rinçages à l'eau osmosée ; ceci pour éliminer tout risque de contamination avec des souches de Rlv autres que les 5 souches de l'inoculum.

Les graines ont ensuite été semées dans des pots de deux litres stérilisés remplis de billes d'argile (2-6 mm de diamètre) et d'attapulgite également stérilisées. Trois semis successifs (espacés d'une semaine chacun) ont été réalisés et lors de chaque semis, chacune des 104 accessions a été semée dans 2 pots différents avec trois graines d'une même accession par pot. Chaque graine a été inoculée par 1 ml d'inoculum bactérien (comportant l'ensemble des 5 souches en proportions égales). Les semis ont ensuite été arrosés 4 fois par jour. Pendant les 8 premiers jours, les arrosages ont été effectués à l'eau osmosée puis, pendant les trois

semaines restantes, une solution nutritive comprenant une faible quantité d'azote (0.625 mM) a été apportée afin de ne pas limiter le développement des nodosités **(Bourion *et al*., 2018)**.

Cet azote contenait un faible taux de ^{15}N pour qu'une estimation du taux de fixation symbiotique soit possible. Quatre semaines après semis, les plantes de pois ont été dépotées avec précaution, les racines et les parties aériennes ont été phénotypées.

3) Variables mesurées

Trente-trois variables phénotypiques quantitatives ont été mesurées lors de l'expérimentation SYMBIOPEA. Ces variables concernent les parties aériennes ou les parties racinaires (Annexe III). Certaines ont été mesurées ou comptées directement au moment du dépotage, telles que la hauteur de la tige principale (ShootL) et la longueur du pivot (TapRootL), ainsi que le nombre de nodosités sur les racines latérales et sur le pivot (NNodTR). D'autres ont été estimées par analyse d'images avec le logiciel WinRhizo à partir des scans réalisés sur des systèmes racinaires précautionneusement lavés et étalés sur un support transparent **(Bourion *et al*., 2010)**. De plus, des pesées ont été réalisées après passage à l'étuve pendant 48h et à 80°C des échantillons décrits, permettant ainsi de déterminer le poids sec des parties aériennes, le poids sec des racines sans nodosités (RootB) et le poids sec des nodosités (NodB). Pour chacune des accessions, le pourcentage de nodosités avec chacune des 5 souches (SA, SD, SE, SF, SK) a été déterminé à Montpellier par l'équipe LSTM (Laguerre et Lepetit) à partir d'un sous-échantillon de 60 nodosités prélevées sur les racines des plantes aux trois dates de semis **(Bourion *et al*., 2018)**. Des analyses biochimiques concernant la teneur en azote (ShootNC) et le taux de fixation symbiotique (NDFA) ont été réalisées en laboratoire à partir d'un échantillon des parties aériennes broyées. Enfin, des variables ont été estimées par calcul à partir de ces variables mesurées. Le nombre de nodosités formées avec les souches A (NSA), D (NSD), E (NSE), F (NSF) et K (NSK) a été obtenu en calculant le produit du nombre de nodosités totales (NNod) par le pourcentage de souches (SA, SD, SE, SF, SK). La densité de racines latérales (LatRootdens) a été quant à elle calculée comme étant le quotient du nombre de racines latérales par la longueur du pivot (TapRootL). Concernant le poids sec moyen d'une nodosité (ANodB), le calcul s'est effectué en faisant le quotient du poids sec des nodosités (NodB) par le nombre total de nodosités (NNod). La variable de la longueur moyenne d'entre-nœuds (InterL) a été calculée en divisant la longueur de tige principale (ShootL) par le nombre d'étages de feuilles développées (NLeaf).

4) Analyses des données phénotypiques

Tout d'abord, les données phénotypiques ont été visualisées par analyse de dispersion (histogramme et boxplot) et les données aberrantes ont été supprimées. Une fois le filtrage des données phénotypiques réalisé, une analyse de variance (ANOVA) a été faite pour chaque variable avec le logiciel R. Ce test a permis de déterminer s'il existait un effet significatif de la date de semis et de vérifier s'il y avait bien un effet génotype significatif dans le modèle. Dans le cas d'un effet répétition (= date de semis), celui-ci a été corrigé par l'addition de moyenne phénotypique de chaque variable aux résidus du modèle. La normalité des données a été analysée par observation d'un histogramme de distribution et par le test de Shapiro-Wilk. Si les résidus ne suivaient pas une loi normale, les données ont été transformées (ln, e^x,√,boxcox …). En effet, même si la GWAS est une analyse robuste,

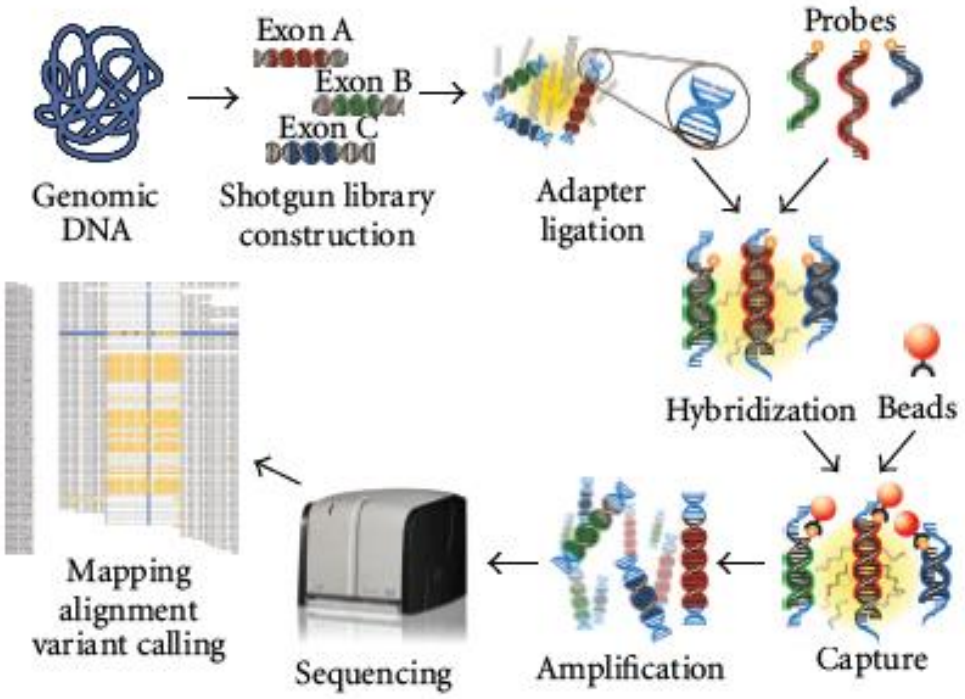


Figure 9 : Méthode de génotypage par capture d'exome (d'après **Schorderet *et al*., 2013)**

Cette méthode s'appuie sur l'hybridation de sondes (oligonucléotides) sur la partie de l'ADN à séquencer (ADN codant). Un fois les sondes hybridées, un lavage est effectué et le séquençage de l'ADN ciblé peut avoir lieu.

Il est préférable que les données phénotypiques suivent une loi normale pour de petits échantillons. A noter que pour de grands échantillons (environ 1 000 individus), cette mise en normalité des données n'est pas nécessaire **(Goh and Yap, 2009)**.

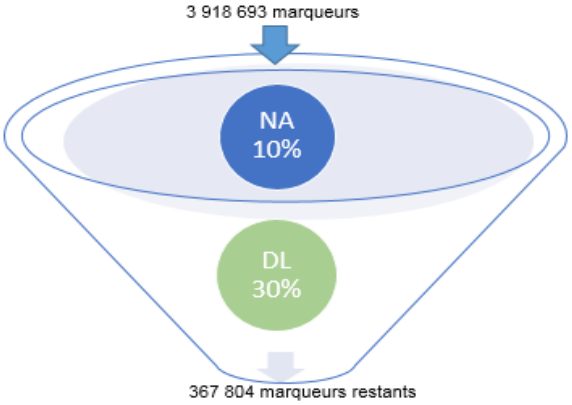
Des analyses de corrélation entre toutes les variables quantitatives mesurées ont été réalisées par calcul de corrélation de Pearson. De plus, une Analyse en Composantes Principales (ACP) a été réalisée sur les variables de choix de souches (SA, SD, SE, SF, SK) avec comme variable illustrative l'origine géographique des accessions de pois.

5) Le génotypage des accessions

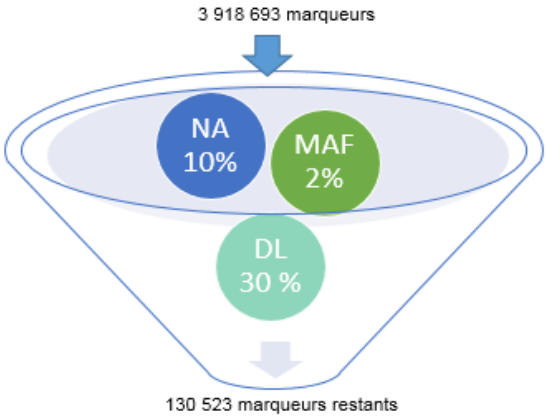
Un ensemble de 357 accessions à génotyper a été défini dans le cadre des projets PeaMUST (260 accessions) et GRaSP (97 accessions). Cet ensemble comprend également le sous-ensemble des 98 accessions de SYMBIOPEA avec lesquelles j'ai effectué l'approche de GWAS. Le génotypage de ces 357 accessions de pois a été réalisé par capture d'exome (Figure 9). Cette technique a été développée chez le pois sur différents panels de plantes dans le cadre du projet d'Investissement d'avenir PeaMUST, avec une forte contribution de Biogemma Génomique Amont. Des sondes de capture correspondant au transcriptome du pois ont été préalablement dessinées et synthétisées par Roche-Nimblegen en utilisant les séquences du transcriptome du pois publiées **(Duarte *et al.*, 2014 ; Alves-Carvalho *et al.*, 2015)**. Les ADN des 357 accessions ont ensuite été extraits à Dijon, à partir de prélèvements faits sur des plantes cultivées en serre ou au champ. L'extraction a été réalisée en utilisant le kit NucleoSpin® Plant II Midi (Macherey-Nagel) puis les ADN ont été fractionnés. Pour chaque échantillon, l'ADN fractionné a ensuite été hybridé aux sondes de capture et, après plusieurs lavages, la fraction enrichie en fragment correspondant aux sondes de capture a été récupérée, amplifiée et des adaptateurs ont été ajoutés préalablement au séquençage en utilisant le kit HyperCap Target Enrichment de Roche. Les banques ainsi créées ont ensuite été séquencées au Génoscope sur des séquenceurs Illumina. Les lectures produites ont été alignées sur une version préliminaire de la séquence du génome du pois **(Madoui *et al.*, 2016)**. Une matrice de génotypage contenant 3 918 693 positions polymorphes a ainsi été produite en filtrant sur un nombre maximal d'individus hétérozygotes à 10% par SNP. A noter qu'il y avait très peu de positions hétérozygotes (1,26%).

6) Analyses de structure et d'apparentement

Avant les analyses de la structuration des différentes accessions, un filtrage des marqueurs (Figure 10a) est effectué. Un filtre est appliqué pour éliminer : i) les marqueurs avec un taux de données manquantes (NA) supérieur à 10% et ii) les marqueurs avec un DL supérieur à 30% sur une fenêtre de 50 marqueurs à la fois avec un pas de 10 en 10. Ce filtrage est réalisé sous Plink avec les fonctions --geno 0.1 (filtrage des NA à 10%) et --indep-pairwise 50 10 0.3 (filtrage du DL à 30%). Le taux d'hétérozygotie parmi les marqueurs restants (367 804 marqueurs) étant inférieur à 2%, aucun filtre n'a été appliqué concernant ce critère. De façon à réduire le nombre de marqueurs pour pouvoir réaliser les études de structuration, 50 000 marqueurs sont choisis aléatoirement parmi les 367 804 par utilisation de la fonction --thin-count 50 000 de Plink. La structuration de la population est effectuée en utilisant ce même échantillon de marqueurs par trois méthodes différentes : FastStructure **(Raj *et al.*, 2014)**, Admixture **(Alexander *et al.*, 2009)** et DAPC (package adegenet de R).



(a)



(b)

Figure 10 : Filtrage des marqueurs pour la structuration (a) et avant étude de génétique d'association (b)

Le set de marqueurs (environ 3.9 millions) du départ a subi différents filtrages pour limiter les biais dans l'analyse.

NA = Filtre des SNP sur un nombre maximal de 10% de données manquantes, DL = Filtre des DL (supérieur à 30%) sur une fenêtre de 50 marqueurs se déplaçant de 10 en 10 tout au long de l'ensemble des marqueurs, MAF = Filtre des allèles minoritaires (inférieur à 2%) dans le set de marqueurs

Les logiciels FastStructure et Admixture utilisent le même modèle génétique sous-jacent pour estimer la proportion du génome de chaque individu provenant de différentes populations ancestrales à partir d'une information génétique multi-locus.

Le logiciel DAPC (package adegenet), quant à lui, s'appuie sur des mesures de variabilités et maximise la variation entre les groupes par la méthode de k-means. Il s'agit d'un algorithme qui forme des groupes maximisant la variabilité. K-means est exécuté séquentiellement avec des valeurs décroissantes de k formant différents clusters. Ces différents clusters sont comparés par un calcul de « Bayesian Information Criterion (BIC) » ; le meilleur groupement correspond à celui le plus bas dans la courbe de BIC.
Une autre structuration de type DAPC a été réalisée avec prise en considération, en plus des autres filtres, d'un filtrage sur la fréquence des allèles mineurs (MAF) avec un seuil fixé à 5%. Ceci permet de réduire le nombre de SNP de départ de 367 804 à 65 594 marqueurs (ce qui est un nombre analysable par DAPC) et de voir s'il y a un effet des allèles rares sur la structuration.

Chaque analyse de structuration a été lancée 2 fois au minimum pour vérifier la fiabilité des k donnés. Une courbe de maximum de vraisemblance permettant de visualiser le meilleur k pour FastStructure a été faite sur Python.

Les figures de sorties de FastStructure et d'Admixture ont été générées sous le logiciel R (package Pophelper et fonction hist(), respectivement), les figures de sorties d'Admixture ont été faites avec le logiciel R par la fonction hist(). La matrice de kinship a été analysée par le logiciel GEMMA (Genome-wide Efficient Mixed Model Association ; **Zhou and Stephens, 2012**).

7) Analyses de GWAS

Préalablement aux analyses de GWAS, un filtrage des SNP a été appliqué avec les mêmes filtres que ceux utilisés pour les analyses de structuration, auxquels a été ajouté un filtre supplémentaire sur la MAF (Figure 10 b). Le seuil choisi pour la MAF est de 2 % et le nombre de marqueurs après filtrage est de 130 523.

Les analyses de génétique d'association n'ont pas été réalisées sur la totalité des 33 variables mesurées. Les variables impliquées dans la nodulation de l'expérimentation SYMBIOPEA ont été choisies pour ces analyses. Les variables de choix de souches (SA, SD etc) n'ont pas été retenues car il s'agit de pourcentages et ces variables ne suivent pas une loi normale. Les 3 premières variables étudiées sont : le nombre de nodosités (NNod), la surface des nodosités (SurfNod) et la longueur de pivot portant des nodosités (LongNod). Une autre analyse de GWAS porte sur les coordonnées des accessions sur l'axe représentant le maximum de variabilité d'une ACP regroupant tous les choix de souche (SA, SD, SE, SF et SK). D'autres analyses de GWAS ont aussi été effectuées pour les variables du nombre de nodosités formées avec chacune des souches. Enfin, une analyse de GWAS a aussi été réalisée sur la variable de longueur de la tige principale (ShootL), qui a déjà été étudiée dans des expérimentations précédentes **(Desgroux *et al*, 2018)**.

Les GWAS ont été effectuées avec le logiciel GEMMA en modèle mixte linéaire univarié. Pour chaque variable, cinq analyses GWAS ont été réalisées : i) une en prenant en compte la kinship seule, ii) trois en prenant en compte la kinship et une structuration définie, à partir du sous-ensemble des 50K SNP choisis au hasard, soit par FastStructure, soit par DAPC, soit par Admixture ; et iii) une cinquième en prenant en compte la kinship et une structuration définie par DAPC à partir des 65 594 SNP restants après filtration avec un seuil de MAF à 5%.

Une héritabilité (pve ; *proportion of variance in phenotypes explained*) au sens large des caractères a été calculée dans GEMMA pour chacune des variables et pour chaque type de structuration ; l'équation est la suivante : pve =vg/(vg+ve) avec vg = variance génotypique et ve = variance environnementale. La somme de vg et ve correspond à la variation phénotypique.

Les sorties de GWAS ont ensuite été analysées dans R avec le package qqman où un seuil de Bonferonni a été appliqué sur les p-values. Seules les p-values supérieures au seuil de $-\log_{10}$ (0.05/nombre de marqueurs) seront significatives. Cette correction permet de tenir compte du fait qu'en raison du grand nombre de test réalisés (un test par marqueur), on s'attend à obtenir un nombre important de faux-positifs.

Un intervalle pour trouver les gènes sous-jacents a été fixé par l'étude du DL pour chaque chromosome. Cette étude porte sur 1 096 123 marqueurs correspondants aux marqueurs restants après filtrage des 3.9 millions de marqueurs avec une MAF de 5%. Le calcul du r^2, un estimateur classique du DL **(Hill and Robertson, 1968)** a été effectué sous Plink sur l'ensemble des chromosomes. La courbe de DL a été obtenue sous le package ggplot2 du logiciel R.

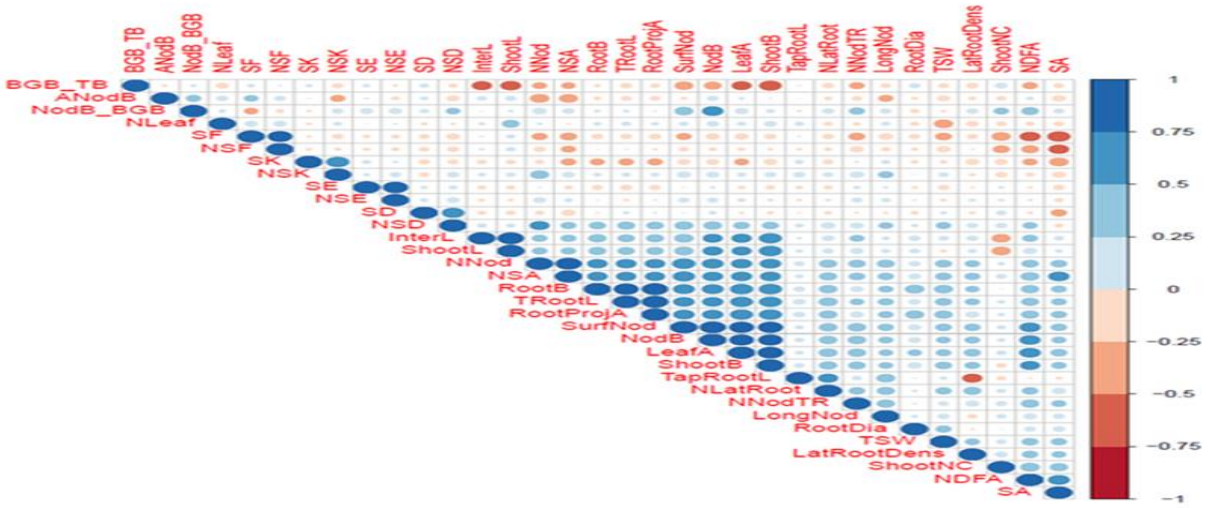
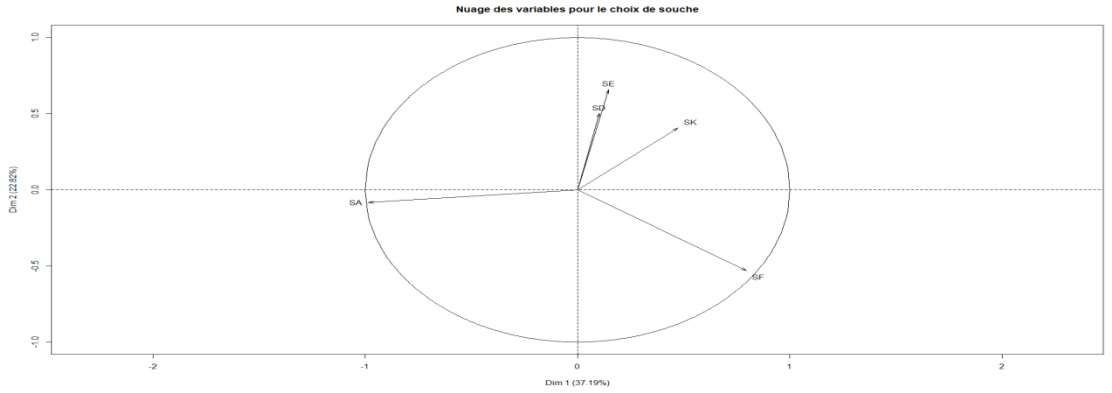
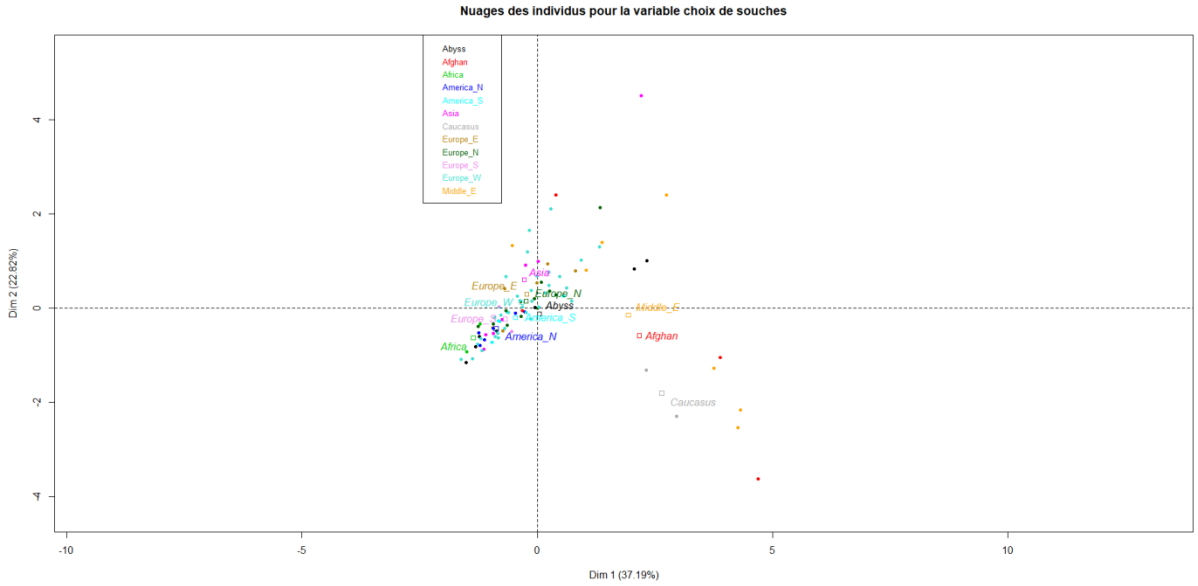


Figure 11 : Corrélations de Pearson entre les variables quantitatives mesurées sur les 98 individus analysés

Les cercles bleus correspondent aux corrélations positives et les cercles rouges, aux corrélations négatives. Plus le cercle est gros et foncé, plus les corrélations sont importantes. Le codage des variables est disponible dans l'annexe 3.



(a)



(b)

Figure 12 : Analyse multifactorielle des variables de nodulation sur les 98 individus analysés

(a) correspond au nuage des individus de l'ACP, (b) correspond au nuage des variables de l'ACP
La variable illustrative correspond aux continents d'origine.

Résultats

1) Analyses des données phénotypiques concernant la nodulation et la variable ShootL

Distribution des données

L'examen de la distribution des fréquences pour la variable du nombre de nodosités (NNod) montre qu'elle suit une loi normale (Annexe IV) ; ce qui est confirmé par le test de Shapiro-Wilk (p-value = 0.3445). Concernant les variables de choix de souches, seule la variable NSA qui correspond au nombre de nodosités formées avec la souche SA suit une loi normale (p-value = 0.0785). La variable ShootL, quant à elle, semble suivre une loi bimodale.

Analyse des corrélations

L'analyse des corrélations entre les variables (Figure 11) indique une corrélation significative très positive entre NNod et NSA. De plus, une forte corrélation significative négative entre les variables NSF et NSA est observée. Aucune autre corrélation n'est significative entre les variables de choix de souche. Une corrélation positive significative entre la surface des nodosités et la surface des racines est observée, tout comme une corrélation positive entre la surface des nodosités et la surface de feuille. D'autre part, une très forte corrélation significative entre la biomasse aérienne et la biomasse nodulaire est visible. Il en est de même entre la biomasse racinaire et la biomasse aérienne.

Analyse en Composante Principale

Les deux premiers axes de l'analyse par ACP des variables de choix de souches expliquent 60 % de l'inertie (Dim1 :37.19%, Dim2 : 22.82% ; Figure 12). Les variables SD et SK sont moyennement représentées sur les axes ($cos^2<0.5$), les variables SA, SF et SE sont plutôt bien représentées sur les axes ($cos^2>0.5$). Malgré cette représentativité moyenne des variables sur les deux premiers axes de l'ACP, on remarque que le nuage des variables montre une corrélation négative entre SA et SF (le cosinus carré de l'angle est proche de -1) ; les autres variables ne semblent pas être corrélées entre elles. Ces résultats sont en accord avec ceux obtenus par l'analyse précédente des corrélations entre variables.

Le nuage des variables indique aussi que le premier axe est représenté majoritairement par les variables SA et SF (contributions respectives de 52.232% et 34.024%), alors que l'axe des ordonnées est représenté majoritairement par la variable SE (contribution de 38.092%).

Concernant le nuage des individus, quatre groupes majeurs se distinguent sur l'ACP ; ces groupes correspondent à des accessions qui choisissent plus spécifiquement SF (individus en bas à droite de l'ACP), des accessions n'ayant pas de spécificité pour les souches (individus au milieu de l'ACP), des accessions spécifiques aux souches SD, SE et SK (individus en haut à droite de l'ACP) et des individus plus spécifiques à la souche SA (individus en bas à gauche de l'ACP).

Tableau 1 : Apparentements les plus significatifs entre les accessions du panel SYMBIOPEA

Les apparentements sont estimés par GEMMA et la matrice d'apparentement complète est disponible en Annexe VII.

Apparentement les plus significatifs entre les accessions			
Couple d'accessions	Indice de similarité	Espéce ou sous-espéces	Origine géographique
R037-R013	69%	abyssinicum-abyssinicum	Ethiopie-Yemen
R014-R015	63%	fulvum-fulvum	Israel-Syrie
R069-R029	27%	humile-sativum	Israel-Afghanistan
R029-R098	22%	sativum-sativum	Afghanistan-Afghanistan
R069-R098	22%	humile-sativum	Israel-Afghanistan
R002-R069	21%	sativum-humile	Afghanistan-Israel
R037-R015	21%	abyssinicum-fulvum	Yemen-Syrie
R013-R015	20%	abyssinicum-fulvum	Yemen-Syrie
R075-R066	20%	elatius-sativum	Turquie-Ukraine
R002-R029	19%	sativum-sativum	Afghanistan-Afghanistan
R099-R074	18%	sativum-sativum	Ethiopie-Soudan
R075-R014	18%	fulvum-elatius	Turquie-Israel
R075-R015	18%	elatius-fulvum	Turquie-Syrie
R013-R075	16%	abyssinicum-elatius	Yemen-Turquie
R037-R075	15%	abyssinicum-elatius	Ethiopie-Turquie
R087-R077	13%	sativum-sativum	France-Danemarque
R036-R098	13%	sativum-sativum	Nepal-Afghanistan

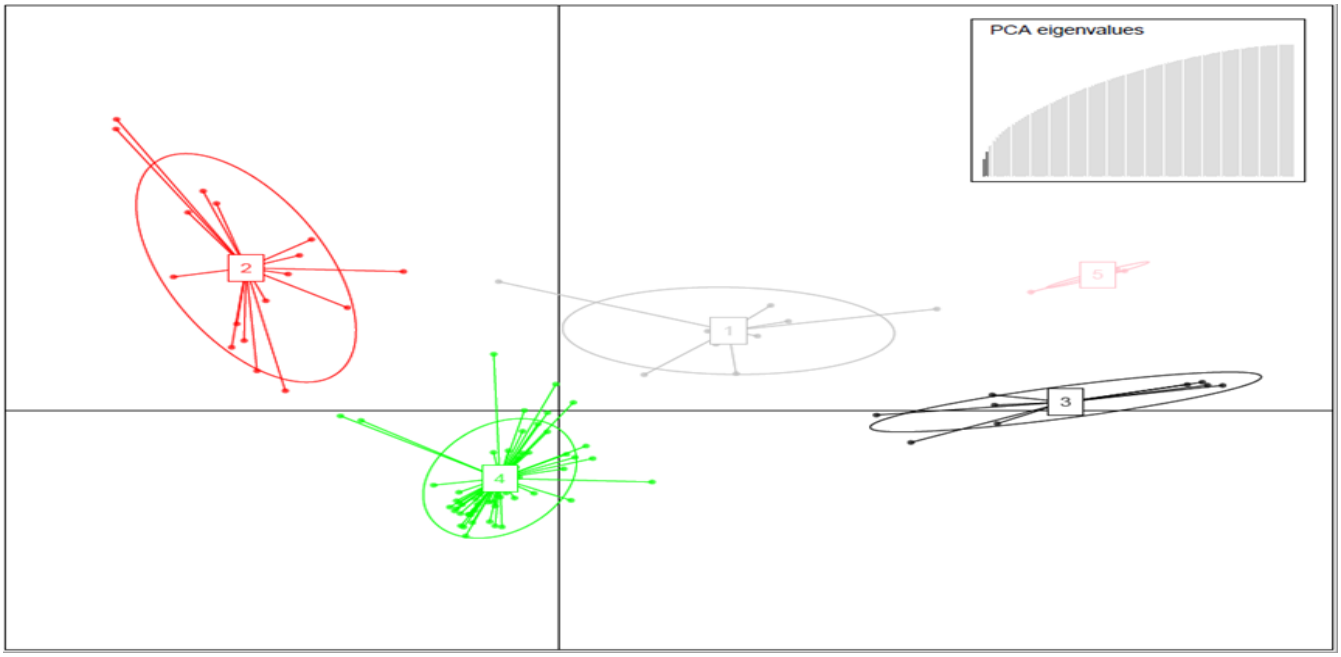


Figure 13 : Structuration du panel SYMBIOPEA par le package DAPC avec filtrage de la MAF à 5%

Le panel se structure en 5 groupes :

1 : groupe des pois d'Abyssinie, du Caucase et du Moyen-Orient - 2 : groupe des pois d'hiver - 3 : groupe des pois afghan - 4 : groupe des pois de printemps - 5 : groupe des *Pisum sativum fulvum/Pisum sativum abyssinicum*

La répartition des accessions selon leur choix de souches semble être liée à leur origine géographique. Ainsi, les accessions venant du Moyen-Orient et en particulier d'Afghanistan se répartissent de préférence vers la variable SF. C'est le cas en particulier de *Pisum sativum* Afghanistan JI86. Une accession provenant du continent asiatique (Kirin 40) est, quant à elle, plus spécifique des souches SE, SK et SD.

Analyse de variance

Le test d'ANOVA montre que l'effet génotype est significatif pour les 33 variables phénotypiques analysées (Annexe V). Les ANOVA indiquent également que l'effet date de semis est significatif pour la plupart des variables. L'effet date de semis est corrigé par des moyennes ajustée sur l'effet semis. Enfin, les résidus des ANOVA pour les 9 variables étudiées par GWAS suivent ou tendent vers une loi normale. Ainsi, les GWAS peuvent se faire sans modification à posteriori de ces variables (Annexe VI).

2) Les analyses génotypiques

L'apparentement dans le panel SYMBIOPEA

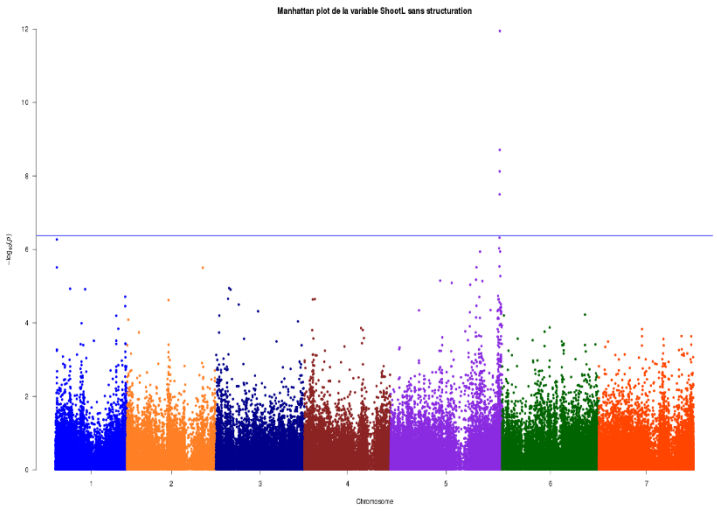
La matrice d'apparentement (kinship) est présentée en Annexe VII ; les apparentements les plus forts sont présentés dans le Tableau 1. Les plus forts indices de similarité calculés atteignent 69% entre les accessions R037 et R013 qui sont tous deux des *Pisum sativum subsp.abyssinicum* et 63% entre R014 et R015, qui sont des *Pisum fulvum*. D'autres apparentements moins prononcés sont observables pour 15 couples d'accessions allant de 27% de similarité à 13% de similarité. Les autres accessions du panel ne sont pas apparentées ou alors très faiblement.

La structuration du panel SYMBIOPEA

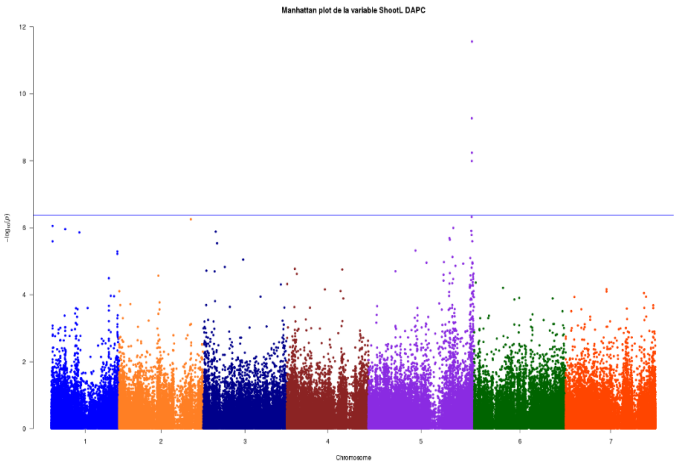
Les diverses méthodes de structuration montrent des images différentes. La courbe de log de FastStructure révèle une structuration en trois groupes (Annexe VIII). Dans cette structuration, un groupe rassemble les deux accessions *Pisum sativum abyssinicum* (R013 et R037) et les deux accessions *Pisum fulvum (*R014 et R015) (cluster 1), un autre groupe réunit les pois afghans ou d'Asie (cluster 2) et le troisième groupe regroupe le reste des accessions (cluster 3) (Annexe VIIII).

Concernant la structuration DAPC, la courbe de BIC (Annexe X) indique une structuration en cinq groupes. Ces cinq groupes (Annexe XI) sont le groupe des *P. fulvum* (cluster 3), le groupe des *P. s. abyssinicum* (cluster 5), le groupe des pois afghans ou d'Asie (cluster 2), le groupe des pois fourragers (cluster 1) et le groupe des pois potagers (cluster 4).

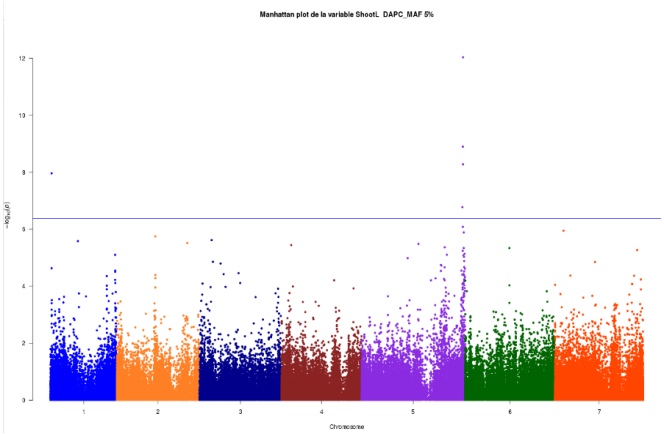
Cinq groupes se distinguent également en structuration de type DAPC après filtrage des données génotypiques pour une MAF supérieure ou égale à 5%. Cependant, ces groupes ne sont pas les mêmes que ceux obtenus précédemment. Les groupes de cette structuration (Figure 13) sont : un groupe comprenant les *P. fulvum* et les *P. s. abyssinicum* (cluster 5), le groupe des pois afghans ou d'Asie (cluster 3), le groupe des pois originaires d'Abyssinie, du



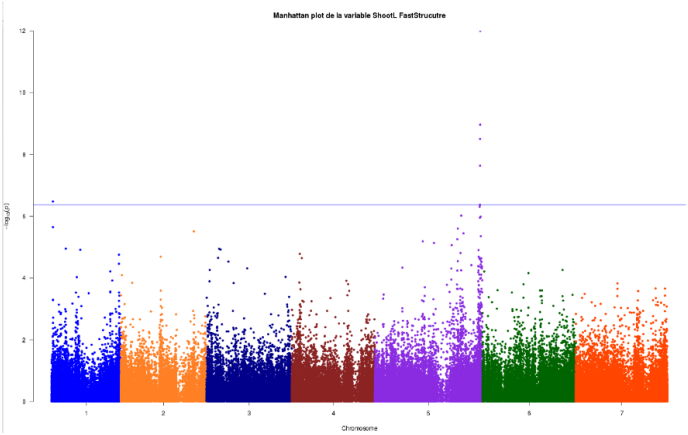
(a)



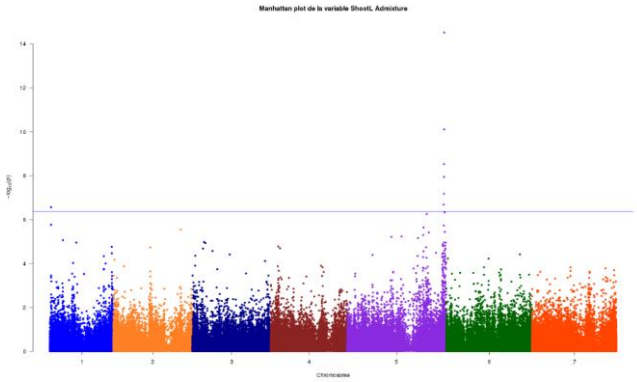
(b)



(c)



(d)



(e)

Figure 14 : Comparaison de la significativité des SNP en prenant en compte la structuration dans le modèle mixte pour la variable ShootL

(a) Modèle mixte de GWAS sans prise en compte de la structuration et avec prise en compte de la matrice d'apparentement.
Modèles mixtes de GWAS avec prise en compte de la structuration et de la matrice d'apparentement :
- (b) DAPC
- (c) DAPC avec une MAF à 5 %
- (d) FastStructure
- (e) Admixture
Le trait bleu correspond au seuil corrigé par la méthode de Bonferonni

Caucase ou du Moyen-Orient (cluster 1), le groupe des pois d'hiver (cluster 2) et le groupe des pois de printemps (cluster 4).

La courbe de cross-validation (Annexe XII) montre une structuration en quatre groupes avec l'outil Admixture (Annexe XIII). Ces groupes sont celui des *P. Fulvum* et *P.s. abyssinicum* (cluster 1), le groupe des pois afghans ou d'Asie (cluster 2), le groupe des pois protéagineux et potagers (cluster 3) et le groupe des pois fourragers et potagers (cluster 4). Ainsi, dans l'ensemble des structurations, les *P. fulvum* et *P. s. abyssinicum* et un *P. s. elatius* (R075) sont toujours regroupés ensemble sauf avec la structuration de type DAPC sans filtrage de la MAF. Il y a également pour chaque type de structuration un isolement des pois afghans ou d'Asie. Enfin, les admixtures sont différentes selon le type de structuration (Annexes VIIII, XII, XIV et XVI). L'admixture correspond à la proportion d'appartenance génétique d'individus dans une population d'origine. La structuration comportant le plus d'admixture est celle obtenue par l'outil Admixture.

Etude du déséquilibre de liaison

Après filtrage de l'ensemble des marqueurs avec une MAF à 5%, 1 096 123 marqueurs restants sont utilisés pour l'analyse du DL sur l'ensemble des chromosomes. On peut observer un patron de décroissance du déséquilibre de liaison à l'échelle de 250 kb où la valeur moyenne du r^2 n'est plus que de 0.09 (Annexe XVII).

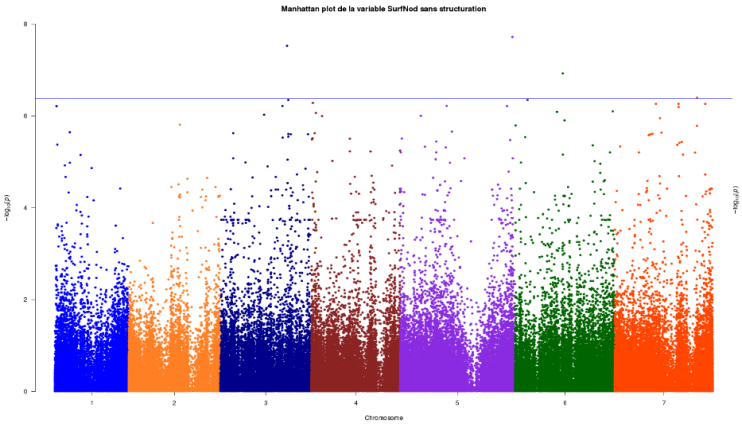
3) Les analyses de GWAS

Calcul d'héritabilité

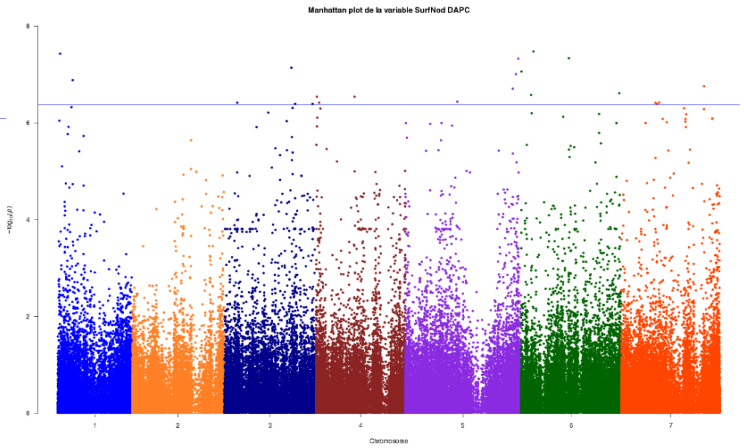
Avant toute analyse de GWAS, l'héritabilité des variables d'intérêt est étudiée pour quantifier la part de variabilité phénotypique expliquée par la variabilité génétique. Il y a une bonne héritabilité pour toutes les variables sauf pour les variables NSD et NSE (Annexe XVIII). Conformément à ce qui est attendu, la prise en compte de la structuration modifie peu l'héritabilité des variables.

Longueur de tige principale

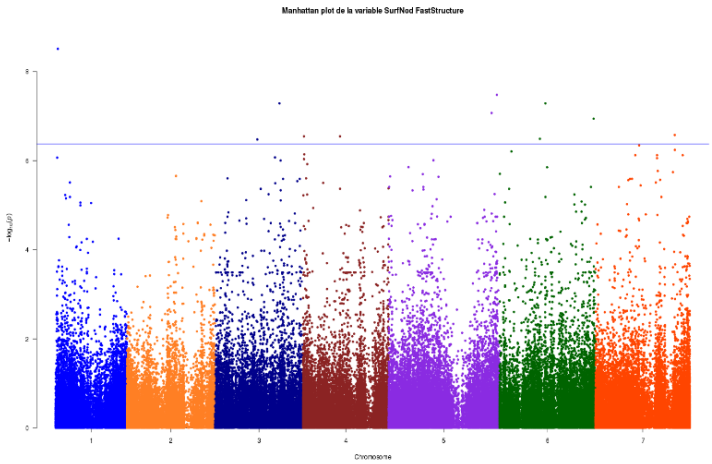
La variable ShootL, connue comme étant fortement déterminée par le gène majeur (=Ga3ox1), et codant pour une Gibbérelline 3 oxydase, est située sur le LGIII **(Wenden et al, 1996)**. L'étude de cette variable menée dans le cadre du stage permet de vérifier si le modèle utilisé et les notations phénotypiques sont corrects pour l'analyse de GWAS. On constate que la significativité des p-values n'est pas la même selon la matrice de structuration utilisée (Figure 14). Ainsi, pour les analyses concernant cette variable ShootL, il y a une amélioration de la puissance de détection (plus de p-values significatives) quand il y a une prise en compte de la structuration en plus de la kinship. En effet, la plus forte significativité des p-values est obtenue avec l'analyse de GWAS réalisée avec uniquement une prise en compte de la kinship, L'analyse donnant le plus de p-values significatives est celle réalisée en prenant en compte la structuration DAPC_MAF et la kinship.



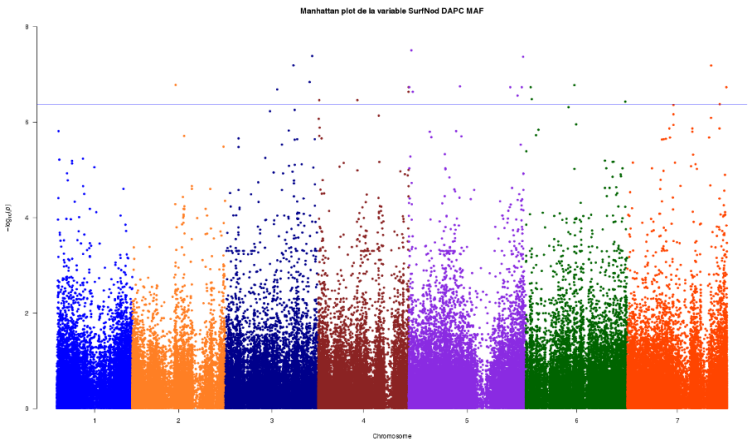
(a)



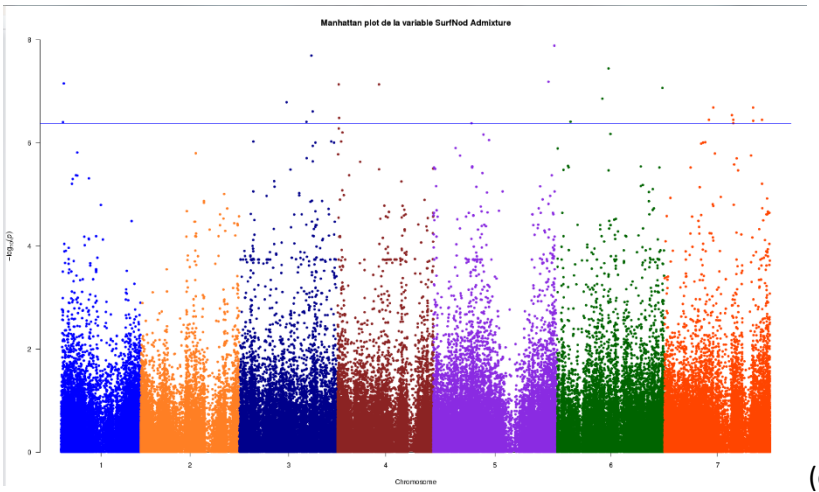
(b)



(c)



(d)



(e)

Figure 15 : Comparaison de la significativité des SNP en prenant en compte la structuration dans le modèle mixte pour la variable SurfNod.

(a) Modèle mixte de GWAS sans prise en compte de la structuration et avec prise en compte de la matrice d'apparentement.

Modèles mixte de GWAS avec prise en compte de la structuration et de la matrice d'apparentement :

- (b) DAPC

- (c) DAPC avec une MAF de 5%

- (d) FastStructure

- (e) Admixture

Le trait bleu correspond au seuil corrigé par la méthode de Bonferonni

Les GWAS réalisées avec les structurations FastStructure, Admixture et DAPC donnent des nombres de p-values significatives intermédiaires entre les 2 extrêmes. Pour l'ensemble des Manhattan plots, un pic majeur ressort sur le chromosome 5. Ce chromosome correspond au LGIII des cartes génétiques des analyses précédentes **(Tayeh *et al*., 2015** ; **Bourion *et al* 2010** ; **Desgroux *et al*, 2018)**. De plus avec la prise en compte de la structuration FastStructure, DAPC ou DAPC_MAF, un pic d'association devient significatif sur le chromosome 1 (=LGVI), avec plus ou moins de significativité selon le type de structuration. Concernant le marqueur en déséquilibre de liaison avec le locus causal putatif sur le chromosome 1 pour certaines structurations, celui-ci se positionne à 6 725 771 pb et l'allèle majeur a un effet de 10,39 cm sur la longueur de la tige.

Le SNP en déséquilibre de liaison sur le chromosome 5 avec le locus putatif impactant la croissance de la tige principale se situe à 567 367 193 paires de bases (pb), et l'effet de l'allèle majeur est une augmentation de 16,42 cm de la longueur de la tige.

4) Capacité à noduler

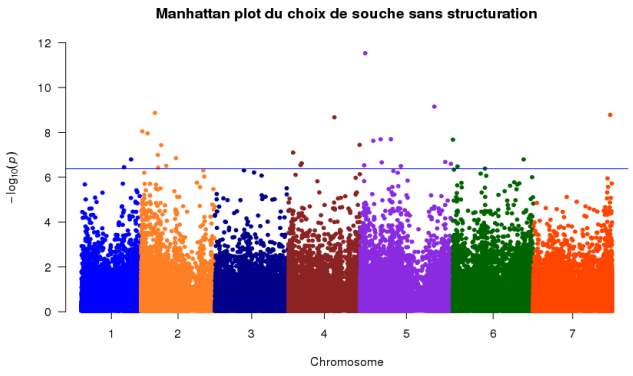
Le nombre de nodosités

Des analyses de GWAS ont été menées sur les variables du nombre de nodosités de la même façon que celles menées pour la longueur de la tige principale : une analyse avec uniquement une prise en compte de la kinship, et quatre analyses prenant en compte la kinship et un type de structuration donné. Aucune de ces analyses n'a permis de mettre en évidence des pics d'associations significatifs.

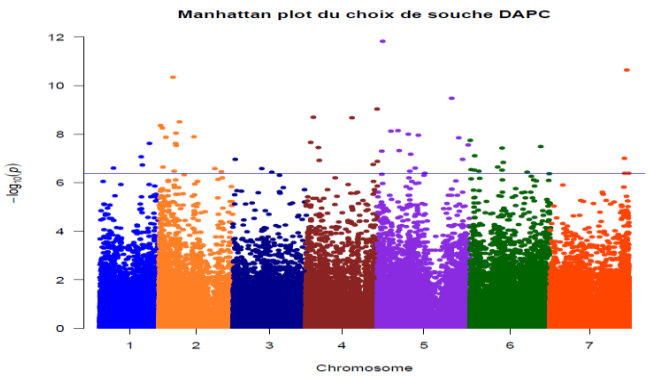
La surface de nodosités

Tout comme la variable ShootL, selon la structuration utilisée dans le modèle mixte, la significativité des p-values n'est pas la même. Une amélioration de la puissance de détection est observée en prenant en compte une structuration avec la kinship. La structuration donnant le plus de significativité est la structuration DAPC_MAF suivie d'Admixture, de DAPC et de FastStructure. Sans prise en compte de structuration, moins de p-values sont significatives.

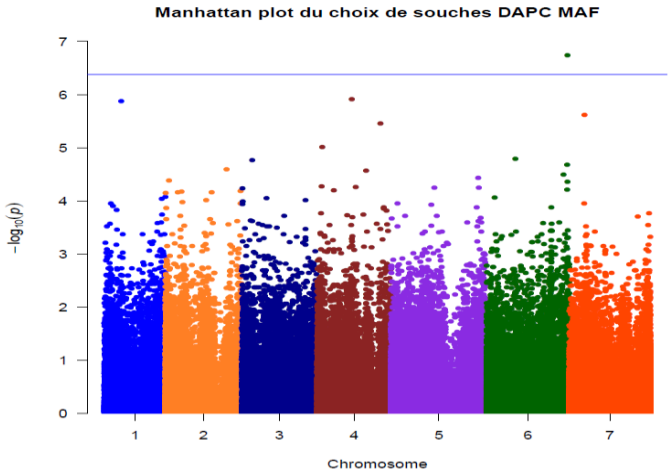
Dans l'ensemble des sorties de GWAS (Figure 15), 3 pics ressortent de façon significative. Un SNP en déséquilibre de liaison avec le locus causal impliqué dans la surface de nodulation se trouve sur le chromosome 3 à 337 174 537 pb, un sur le chromosome 5 situé à 565 557 662 pb et un autre sur le chromosome 6 à 239 691 015 pb. Les allèles majeurs de ces loci putatifs situés sur les chromosomes 3, 5 et 6 ont respectivement un effet de -2,32 mm, -2,29 mm et -3,18 mm sur la surface des nodosités. D'autres pics d'association, spécifiques à tel ou tel type de structuration pris en compte, ressortent. Ainsi un SNP à 553 969 166 pb en déséquilibre de liaison avec un locus causal putatif ressort sur le chromosome 5 ; l'allèle majeur de ce locus putatif a un effet de -2.36 mm sur la surface des nodosités. Concernant le chromosome 4, un SNP en déséquilibre de liaison avec un locus causal putatif ressort à 192 706 123 pb ; l'effet de l'allèle majeur du locus putatif est de -2,49 mm.



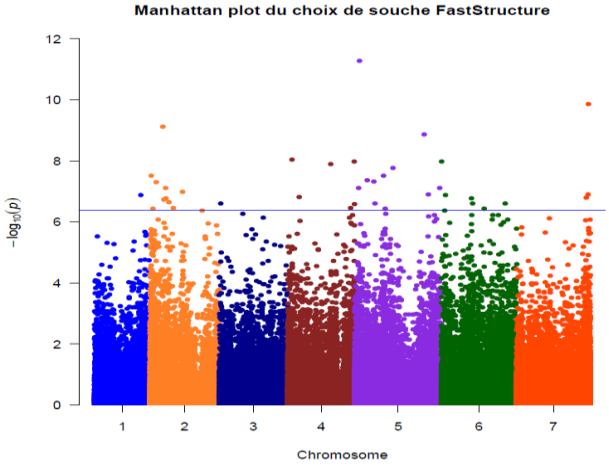
(a)



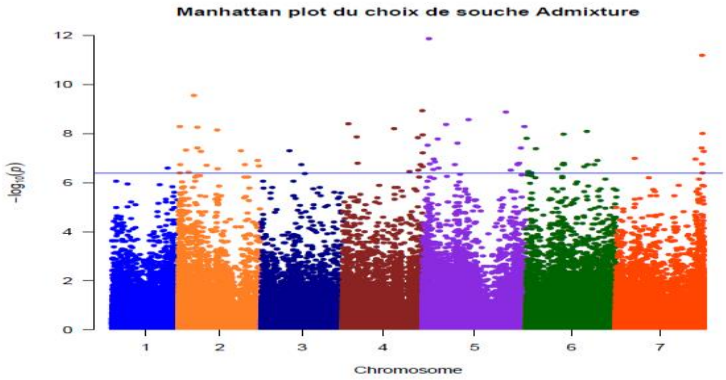
(b)



(c)



(d)



(e)

Figure 16 : Comparaison de la significativité des SNP en prenant en compte la structuration dans le modèle mixte pour la variable choix de souche.

(a) Modèle mixte de GWAS sans prise en compte de la structuration et avec prise en compte de la matrice d'apparentement.

Modèles mixte de GWAS avec prise en compte de la structuration et de la matrice d'apparentement :

- (b) DAPC
- (c) DAPC avec une MAF à 5%
- (d) FastStructure
- (e) Admixture

Le trait bleu correspond au seuil corrigé par la méthode de Bonferonni

La longueur de pivot portant des nodosités

Pour l'ensemble des modèles utilisés avec ou sans prise en compte de la structuration, un pic ressort sur le chromosome 6 (Annexe XIX). Ce SNP se situe à 487 113 647 pb. L'effet de l'allèle majeur du locus putatif en déséquilibre de liaison avec ce SNP significatif est de 14 cm sur la longueur de pivot portant des nodosités. Comme pour les autres variables étudiées précédemment, des SNP ressortent plus ou moins fortement selon les structurations et la prise en compte de la structuration. DAPC_MAF est celle qui permet de détecter le plus d'associations significatives.

Le choix de souches

Les variables NSD et NSE ne sont pas exploitables pour une étude de GWAS en raison de leurs héritabilités très faibles (Annexe XIIX). Concernant la variable NSA, aucune significativité ne ressort des modèles utilisés (Annexe XXI). Les analyses de génétique d'associations sur les variables NSF et NSK montrent un nombre conséquent de p-values significatives pour tous les modèles mixtes utilisés, ce qui est peu interprétable dans la recherche de loci putatifs favorisant la nodulation avec les souches NSF et NSK (Annexe XXII, XXIII).

L'ACP décrivant le choix de souches et montrant une exploitation de 60% de la variabilité totale sur les deux premiers axes est utilisée pour une étude générale de choix de souches. Les coordonnées sur l'axe 1 sont choisies pour l'analyse de GWAS. En effet, cet axe explique le plus de variabilité génétique (37.2%).

Les sorties de GWAS (figure 16) indiquent pour l'ensemble des modèles (sauf pour le modèle prenant en compte la structuration DAPC_MAF), quatre pics très significatifs en déséquilibre de liaison avec des loci causaux putatifs sur le chromosome 2, le chromosome 4, le chromosome 5 et le chromosome 7. Concernant le chromosome 2, le SNP se situe à une position de 85 212 876 pb et l'allèle majeur au locus putatif en déséquilibre de liaison avec ce SNP a un effet de 1, 63 sur le choix de souches. Le SNP sur le chromosome 4 se positionne à 244 943 262 pb et l'allèle majeur au locus en déséquilibre de liaison avec ce SNP à un effet de 2,1 sur le choix de souches. Pour le SNP sur le chromosome 5, celui-ci se situe à 32 222 495 pb et l'allèle majeur au locus putatif en déséquilibre de liaison avec ce SNP a un effet de 2,08 sur le choix de souches. Enfin, le SNP sur le chromosome 7 se situe à 479 786 412 pb et l'allèle majeur du locus en déséquilibre de liaison avec le SNP a un effet de 2,08 sur le choix de souches.

Un SNP ressort significativement sur le chromosome 6, pour tous les modèles, sauf pour le modèle avec une structuration DAPC_MAF. Ce SNP se situe à 147 528 pb et le locus putatif en déséquilibre de liaison avec ce SNP a un effet de 1,35 sur le choix de souches. D'autres SNP, plus ou moins significatifs, ressortent pour l'ensemble des structurations mais avec une significativité plus réduite. D'autres SNP ressortent plus spécifiquement selon le type de structuration utilisé.

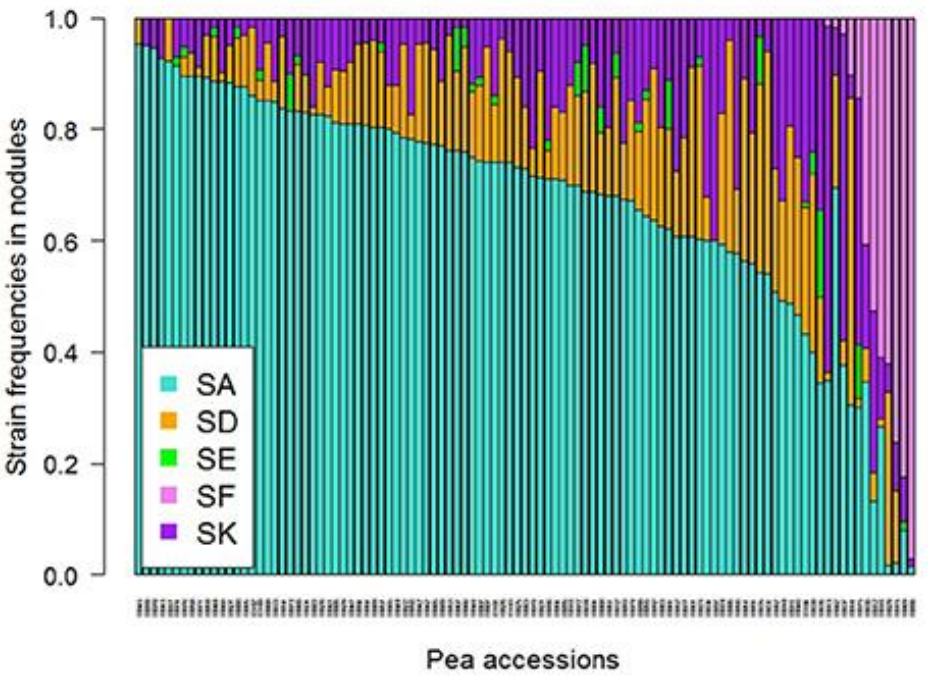


Figure 17 : Nodulation en fonction de la souche bactérienne sur 104 accesssions de pois **(Bourion *et al.*, 2018)**

Le nombre de nodosités est plus grand avec la souche SA par rapport aux autres souches, du fait de la meilleure spécificité de SA.

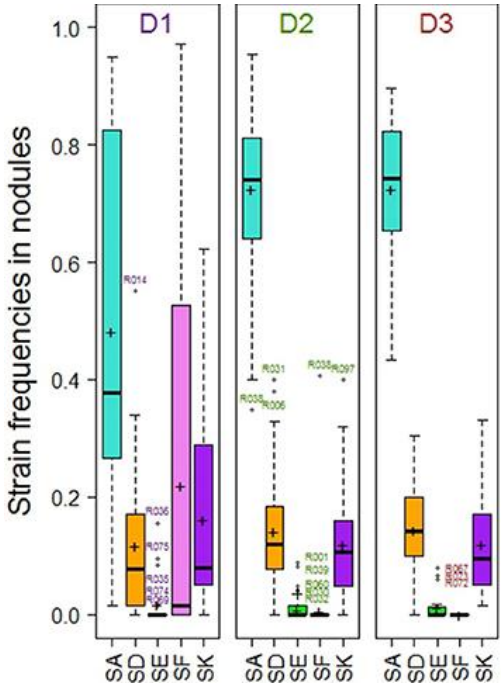


Figure 18 : Compétitivité entre les souches dans l'interaction plante/bactérie **(d'après Bourion *et al*., 2018)**

La compétitivité des souches est différente selon les espèces étudiées. Avec D1 correspondant au pois sauvages, D2 au pois de printemps et D3 au pois d'hiver.

Discussion

L'objectif fixé de cette étude était de trouver des gènes ou QTL putatifs impliqués dans la nodulation et le choix de souches de rhizobium chez le pois. L'étude s'est faite par différents modèles mixtes de génétique d'association sur un panel d'accessions de pois possédant une diversité géographique plutôt large mais en nombre limité d'individus pour des analyses de GWAS. Une forte variabilité pour des caractères de nodulation est présente au sein de ce panel.
Les différentes analyses portent sur des variables en lien avec la nodulation. Une étude a été aussi menée sur la variable de longueur de la tige principale. Cette étude nous a permis de retrouver un gène connu et ainsi de valider, au moins sur le principe, la validité de l'approche.

Les souches bactériennes : une histoire de compétitivité et d'efficience

L'étude menée s'inscrit dans un cadre plus général visant à connaître, d'une part, la capacité des génotypes de pois à favoriser les associations symbiotiques avec les souches de Rlv les plus performantes pour l'acquisition d'azote et, d'autre part, l'impact de la sélection végétale sur cette capacité. Il s'agit tout particulièrement d'acquérir une meilleure connaissance des déterminants génétiques du choix entre les partenaires symbiotiques pois et Rlv pour pouvoir, à terme, optimiser ce choix pour les conditions de culture en plein champ.

Des travaux antérieurs, menés par l'équipe ECP de l'UMR Agroécologie où j'ai effectué mon stage, ont montré que différents génotypes de pois bénéficient de manières différentes (qualitativement et quantitativement) de l'interaction avec les Rlv. La diversité et la structure génétique des populations naturelles des Rlv nodulant le pois fluctuent selon le sol **(Laguerre *et al*., 2003)** mais aussi selon le génotype végétal suggérant des associations préférentielles entre génotypes végétaux et bactériens **(Bourion *et al*., 2007; Depret & Laguerre 2008)**. La diversité génétique bactérienne se traduit par une diversité fonctionnelle au niveau de la plante, avec des effets sur son développement aérien, sur celui des racines et des nodosités racinaires **(Laguerre *et al*., 2007)**.

Un des premiers enjeux de l'expérimentation SYMBIOPEA a été d'évaluer la relation entre diversité génétique chez le pois et choix du symbiote bactérien, et d'étudier si ce choix était lié à meilleure efficience de fixation symbiotique. Dans la plupart des recherches menées auparavant, les souches de Rlv collectées à partir de piégeage, lors d'expérimentation dans les champs, étaient décrites pour la variabilité de leur efficience à fixer l'azote de l'air, variabilité mesurée par la biomasse produite par un génotype de pois de référence nodulé par chacune de ces souches. Elles étaient aussi, en complément, décrites pour leur compétitivité : c'est-à-dire leur capacité à noduler ce même génotype de pois lorsqu'elles étaient inoculées en mélange avec une souche Rlv de référence. L'étude SYMBIOPEA qui porte sur une collection de pois inoculée par un mélange de souches Rlv est donc une approche originale **(Bourion *et al*, 2018)**. Dans cette étude, il a été observé que les pois cultivés, pois de printemps ou pois d'hiver, choisissent de préférence la souche d'origine française SA (Figure 17, Figure 18). A l'opposé, les pois sauvages ou plus ancestraux ont une diversité de choix de souches beaucoup plus importante. Il a aussi été observé une résistance à la nodulation par les souches européennes au besoin spécifique pour la souche TOM (=SF) chez certains pois d'Afghanistan ou originaires du Moyen-Orient.

Dans notre étude réalisée sur le jeu de données de SYMBIOPEA, nous avons mesuré une forte corrélation négative entre les variables NSA et NSF ; ce qui est cohérent avec le fait

que des pois d'Afghanistan ou du Moyen-Orient choisissent quasi-spécifiquement la souche SF, ce qui n'est pas le cas de la majorité des autres accessions de pois qui choisissent préférentiellement SA. Les résultats de compétitivité différentielle des souches vis-à-vis des accessions se confirment avec l'ACP où il est visible que les pois venant du Moyen-Orient et plus spécifiquement d'Afghanistan choisissent préférentiellement SF. Les pois choisissant la souche A ont une origine géographique plus diversifiée (Europe, Amérique, Asie). Concernant la compétitivité des autres souches, SD, SE et SK, il est difficile de conclure qu'il existe une spécificité entre certaines accessions et l'une de ces souches. En effet, l'ACP représente mal ces variables et il n'y a pas de corrélations entre elles.

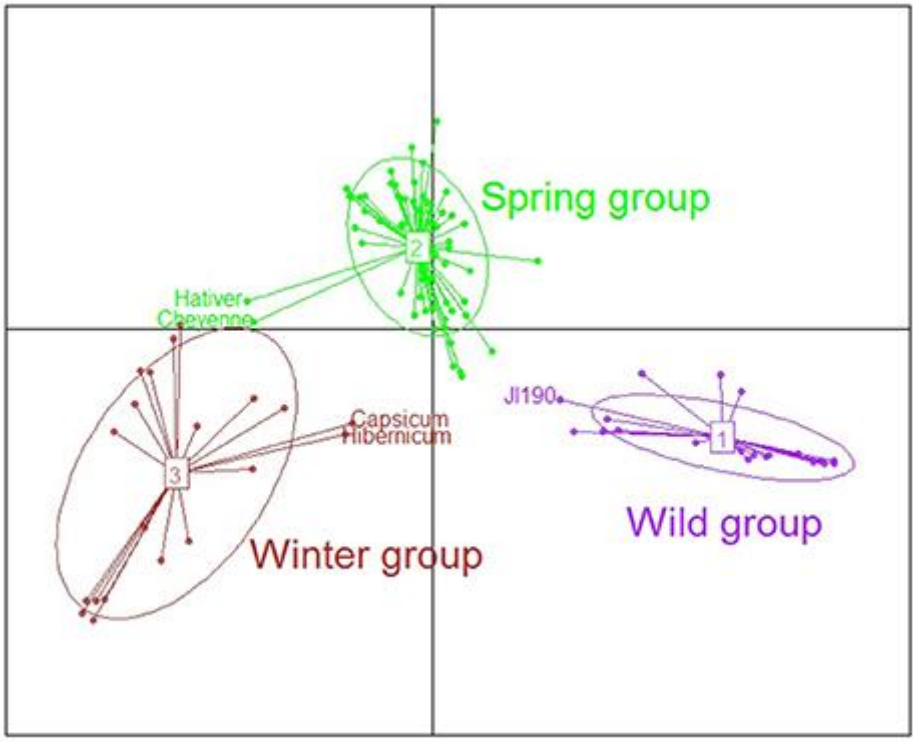
L'intérêt de notre analyse en ACP est une approche globale du choix parmi les cinq souches. Les cinq variables de choix de souches ne sont pas indépendantes ; en d'autres termes, si une souche est choisie préférentiellement, par conséquent, les autres souches sont moins choisies. Il existe des interactions entre les souches du mélange qui modifient les aptitudes de chacune des souches à noduler les pois. Ainsi, **Bourion *et al* (2018)** ont montré, sur un sous-ensemble de 18 accessions de pois parmi les 104 étudiées au départ, que la capacité à noduler de chacune des souches SA, SD, SE et SK, évaluée lors d'expérimentations en mono-inoculation, ne prédit pas sa capacité à noduler lorsqu'elle est en mélange (avec les quatre autres souches). Une corrélation a été observée uniquement pour SF, souche pour laquelle il existe une spécificité de choix. Les interactions entre souches sont de natures diverses ; elles peuvent être basées sur les interactions entre les différents facteurs Nod émis par les souches.

L'explication de la forte corrélation entre NSA et NNod s'explique aussi par le fait que la compétitivité de SA par rapport aux autres souches est plus forte et qu'il y a donc plus de nodosités avec cette souche. De plus, la corrélation très forte entre la surface des nodules et la surface des racines peut s'expliquer par le fait que, plus une racine est grande, plus les nodosités ont de surface pour se développer. En plus de la spécificité de certaines souches, l'efficience est un caractère important à prendre en compte car s'il n'y a pas d'efficience, il n'y a pas d'assimilation d'azote.

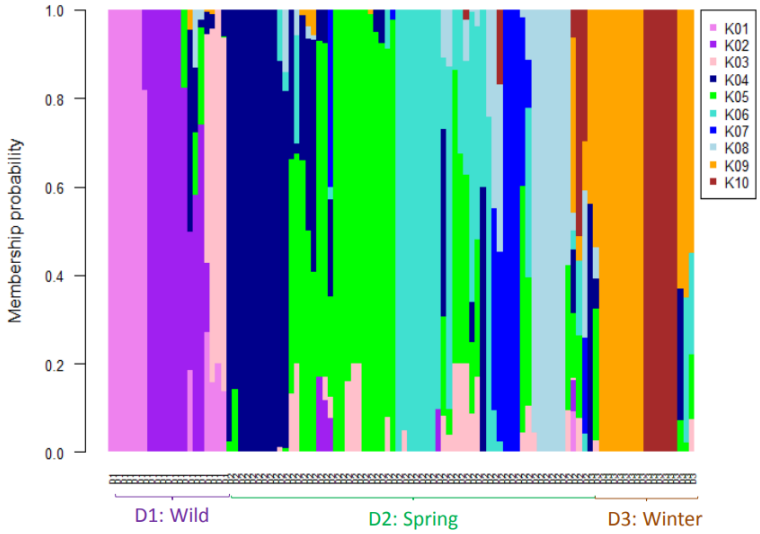
Les structurations et les apparentements des analyses complexes

La structuration du panel d'étude est très importante à considérer en génétique d'association, ceci permet de réduire le DL qui n'est pas d'origine physique et donc d'améliorer la précision des analyses de GWAS.

Les analyses d'apparentement réalisées sur le panel SYMBIOPEA ne mettent en évidence qu'un faible nombre d'accessions réellement apparentées. Les deux plus forts apparentements mesurés (supérieurs à 60%) sont i) entre les deux accessions de *Pisum fulvum* du panel et originaires de pays très proches (Israël et Syrie) ; ii) entre les deux accessions de *Pisum sativum subsp. abyssinicum* aussi géographiquement proches (Ethiopie et Yémen). On constate ainsi que les plus forts apparentements sont observés entre accessions appartenant à une même espèce de pois sauvage (*P. fulvum*) ou à une même sous-espèce (*P. s.* subsp*. abyssinicum*) de pois cultivé. Dans chacun de ces deux groupes, les accessions ont une origine géographique proche et ont des chances d'avoir un ou plusieurs ancêtres communs et donc d'être apparentées. A noter que le niveau d'apparentement entre les deux groupes est de 20%. Les analyses de structuration, quelle que soit la méthode utilisée, confirment que les *P. fulvum* et les *P. s. abyssinicum* sont éloignés génétiquement des autres pois. Ceci est en accord avec les observations que les *P. fulvum* et *P. s. abyssinicum* différent des *P. sativum* par plusieurs réarrangements chromosomiques qui les rendent presque incompatibles avec *P. sativum* **(Bogdanova *et al*., 2009)**.



(a)



(b)

Figure 19 : Structuration du panel SYMBIOPEA des 104 individus avec 13 000 marqueurs **(Bourion *et al.*, 2018)**

Trois groupes ressortent de cette structuration : le groupe des sauvages, le groupe des pois de printemps et le groupe des pois d'hiver.

(a) Structuration sur le package DAPC

(b) Structuration sur le logiciel FastStructure

Néanmoins, trois des quatre méthodes utilisées regroupent ces quatre accessions dans un même cluster. Seule la méthode de type DAPC à partir des 50 000 marqueurs choisis au hasard les identifie comme appartenant à deux groupes différents. Une cinquième accession, un *P. s. elatius* (R075) originaire de Turquie, est systématiquement regroupée avec les *P. s. abyssinicum*, en dépit d'un niveau d'apparentement moyen de 15% avec cette accession et les *P. s. abyssinicum*. Trois des méthodes de structuration utilisées montrent de l'admixture chez cette accession, avec une appartenance majoritaire au groupe des *P. s. abyssinicum* (et P*. fulvum*) et minoritaire à un groupe de *P. sativum*. Seule la méthode de type DAPC à partir des 50 000 marqueurs choisis au hasard ne détecte pas d'admixture chez cette accession et ne la différencie pas des *P .s .abyssinicum*. Ainsi, les analyses de structuration complètent celles d'apparentement mesurées par des analyses de kinship et permettent de mieux comprendre les phylogénies génétiques. Une explication de la proximité génétique entre les différentes accessions serait que les *P. s. abyssinicum* résulteraient d'hybridations très anciennes (mais très limitées car le pois est principalement autogame) entre un sous-ensemble de *P. s. elatius* et des pois sauvages considérés comme des plantes adventives **(Jing *et al*., 2010; Smýkal *et al*., 2011)**.

Les quatre différents types de structuration regroupent des pois provenant d'Afghanistan ou d'Asie (Népal, Inde et Chine) dans un même cluster. Ils sont assez proches entre eux (autour de 20% d'apparentement) et éloignés génétiquement des autres pois, *P. fulvum*, *P. s. abyssinicum* et aussi des autres *P. sativum*. L'hypothèse de la présence d'allèles rares les regroupant n'est pas fiable car la structuration DAPC avec filtrage des allèles mineurs à 5% montre toujours un isolement de ce groupe. L'autre hypothèse serait qu'il y a eu une différenciation de ce groupe (sélection) en raison de son isolement géographique.

Les structurations entre les *P. sativum* restants dépendent très nettement du type de structuration utilisée. Avec FastStructure, il n'y a pas de différenciation en plusieurs groupes mais il en existe avec les trois autres types de structuration. Selon le type de structuration, ces *P. sativum* sont regroupés selon leurs usages (potagers, fourragers ou protéagineux) ou selon leur date de semis (hiver ou printemps). Cette dernière structuration est basée sur la structuration de type DAPC avec élimination des allèles les plus rares et est très proche de celle obtenue par **Bourion *et al* (2018)**. Dans cette étude, l'ensemble du panel de SYMBIOPEA (104 accessions) a été analysé par une structuration de type DAPC, mais avec un nombre plus faible de marqueurs (13 000 ; Figure 19). 57 des 58 accessions (parmi les 98 communes aux deux études) identifiées par **Bourion *et al* (2018)**, comme faisant partie du groupe des pois de printemps, sont aussi identifiées dans notre étude comme étant des pois de printemps. De même, 18 des 19 accessions identifiées par **Bourion *et al* (2018)**, comme faisant partie du groupe des pois d'hiver, sont aussi identifiées dans notre étude comme étant des pois d'hiver. La particularité de notre structuration est qu'elle permet la différenciation par DAPC des pois identifiés comme « sauvages » dans **Bourion *et al* (2018)** en trois groupes distincts : les *P. fulvum*, *P. s. abyssinicum* et R075 dans un premier groupe, les pois afghans ou d'Asie dans un deuxième groupe, et enfin des pois d'Abyssinie, du Caucase ou des *P. s. elatius* (ou *humile*) dans un troisième groupe. A noter que cette structuration en trois groupes des « sauvages » est néanmoins très proche de la structuration plus fine obtenue par FastStructure dans la même étude de **Bourion *et al* (2018)**.

En conclusion, les quatre types de structuration que nous avons effectués n'ont pas structuré le panel de façon strictement identique. Mais chaque structuration a un sens et des similitudes avec des structurations menées précédemment existent.

La structuration en génétique d'association impacte la significativité

Les modèles utilisés dans GEMMA prennent tous en compte la kinship ; nous avons testé l'impact de la prise en compte en plus de chacun des quatre types de structuration. Selon la méthodologie utilisée, le nombre de groupes obtenus était de 3 à 5. Sur divers forums scientifiques dédiés aux études de la génétique d'association, il est expliqué que prendre une structuration subdivisant la population en peu de groupes reste, pour une étude de GWAS, la meilleure solution. En effet, le but est de corriger le DL dû à la stratification de la population **(Yu *et al*., 2006)** et d'éviter de sur-accumuler des corrections qui, à terme, biaisent le modèle.

Nous avons tout d'abord testé l'impact de la prise en compte de la structuration pour la variable de longueur de la tige principale. Toutes les analyses GWAS menées, celle sans structuration et les trois avec structuration, permettent de détecter un pic majeur d'association sur le chromosome 5 qui correspond au gène majeur *GA3ox1=Le* et aussi un autre pic d'association sur le chromosome 1. Les profils des Manhattan plots obtenus avec les cinq analyses différentes sont globalement très proches, mais les significativités des p-values changent en fonction des modèles de structuration utilisés. Ceci démontre que la prise en compte de la structuration impacte la puissance de détection des SNP.

Nous avons mené la même étude pour d'autres variables liées à la capacité à noduler ou au choix de souches. Pour toutes ces variables, en accord avec ce que l'on a observé pour la variable de longueur de tige principale, le fait de prendre en compte la structuration en plus de la kinship dans les modèles mixtes montre une augmentation des p-values significatives. Ceci semble contraire à ce qui est généralement présenté dans la bibliographie, notamment celle de **Yu *et al* (2006)**. Deux explications peuvent être données à cette augmentation de significativité : la première serait que des SNP sont associés au polymorphisme phénotypique mais avec des effets faibles, et donc peu significatifs. Le fait de corriger par la structuration perm*et al*ors de rendre ces SNP significatifs dans l'analyse de GWAS. La deuxième hypothèse serait que la kinship corrigerait à elle seule suffisamment le modèle ; le fait de prendre en compte la structure ajouterait artificiellement du bruit statistique dans les données.

En complément, il faut noter qu'après analyse de toutes les variables (sauf celle du choix de souches), les modèles prenant en compte la structuration avec méthode DAPC_MAF utilisant les 65 594 marqueurs restants, après filtration sur la MAF à 5%, sont ceux présentant le plus de significativité. On peut donc en conclure que ce set de marqueurs est plus adapté à des structurations pour les analyses de GWAS dans notre panel d'étude que le set de 50 000 marqueurs choisis au hasard parmi les 367 804 marqueurs restants après filtration. Il n'est cependant pas sûr que la structuration de type DAPC_MAF contrôle tous les faux positifs. Cela peut s'expliquer par l'histoire évolutive au sein du genre *Pisum* dont des représentants sont présents dans le panel.

Des analyses de variables complexes pour une analyse de GWAS

La GWAS concerne généralement des variables quantitatives. Dans notre étude, la majorité des variables est quantitative sauf celles impliquées dans le choix de souches. Les variables relatives au choix de SA ou SF sont plutôt qualitatives : les génotypes de pois choisissent majoritairement SA et dans ce cas pas SF, ou inversement SF et dans ce cas pas SA. Les autres souches sont moins majoritairement choisies et à un taux particulièrement réduit pour SD et SE. L'explication de la faible héritabilité des variables NSD et NSE peut s'expliquer par le fait que ces deux souches sont peu choisies. Ainsi, la part de variabilité phénotypique expliquée par le génotype transmise à la descendance est difficilement quantifiable. C'est ce qui explique aussi le fait de ne pas pouvoir conclure sur la recherche de gènes ou QTL putatifs pour ces variables de façon indépendante.

Les différentes analyses que nous avons menées concernant la variable ShootL permettent de mettre en évidence un pic d'association significatif sur le chromosome 5 ; ce pic déjà connu correspond au gène *Ga3ox1* connu chez le pois **(Desgroux *et al*., 2018)**. On peut alors considérer que les modèles de GWAS que nous avons utilisés sont valables pour les autres variables.

Aucun SNP significatif ne ressort pour la variable NNod malgré un nombre conséquent de marqueurs utilisés dans l'analyse. Pourtant, des analyses de génétique d'association concernant la nodulation chez *Medicago truncuntula* effectuées par Santon-Geddes *et al* (2013) ont permis de mettre en évidence 12 gènes impactant la nodulation chez *Medicago truncatula*. La synténie entre le *Pisum sativum* et *Medicago truncatula* peut aider à trouver ces gènes de nodulation chez le pois **(Tayeh *et al*., 2015)**. Le fait de ne pas trouver de SNP significatifs dans notre étude peut s'expliquer essentiellement en raison de la taille de la population qui n'est pas assez grande pour une analyse de GWAS concernant ce type de caractère, résultant de l'expression d'un grand nombre de gènes.

Une étude de génétique d'association sur l'orge (*Hordeum vulgare*) montre que l'effet de la taille de la population joue beaucoup sur la détection de gènes ou de QTL putatifs **(Wang *et al*., 2012)**. Augmenter la taille de la population semble alors capital pour détecter des gènes ou QTL putatifs et pour corriger les éventuels biais dans les analyses. C'est pour cela que le projet GRaSP a été développé, afin d'analyser 337 individus pour les mêmes variables. Ce projet va permettre d'obtenir plus de robustesse et de significativité dans les analyses de GWAS pour les caractères complexes que sont la nodulation et le choix de souches

En ce qui concerne les autres variables de nodulation, des SNP significatifs ont été trouvés par les analyses de GWAS. Les effets des allèles majeurs ont un fort impact sur la variation des données analysées, généralement de façon positive sauf pour la surface de nodosités. Les SNP en déséquilibre de liaison avec les loci concernant la variable SurfNod impactent de façon négative la surface des nodosités, ce qui ne semble pas intéressant d'un point de vue agronomique. En effet, plus une nodosité est petite, moins elle assimilera de l'azote atmosphérique. Cependant, une recherche parmi la population de mutants TILLING disponible à l'INRA pourrait permettre de trouver un mutant STOP pour le gène putatif **(Berbel *et al*., 2012)**. Il faudrait ensuite vérifier, après backcross, si la surface de nodosités et la fixation symbiotique de ce mutant est plus importante que celle du témoin « sauvage ».

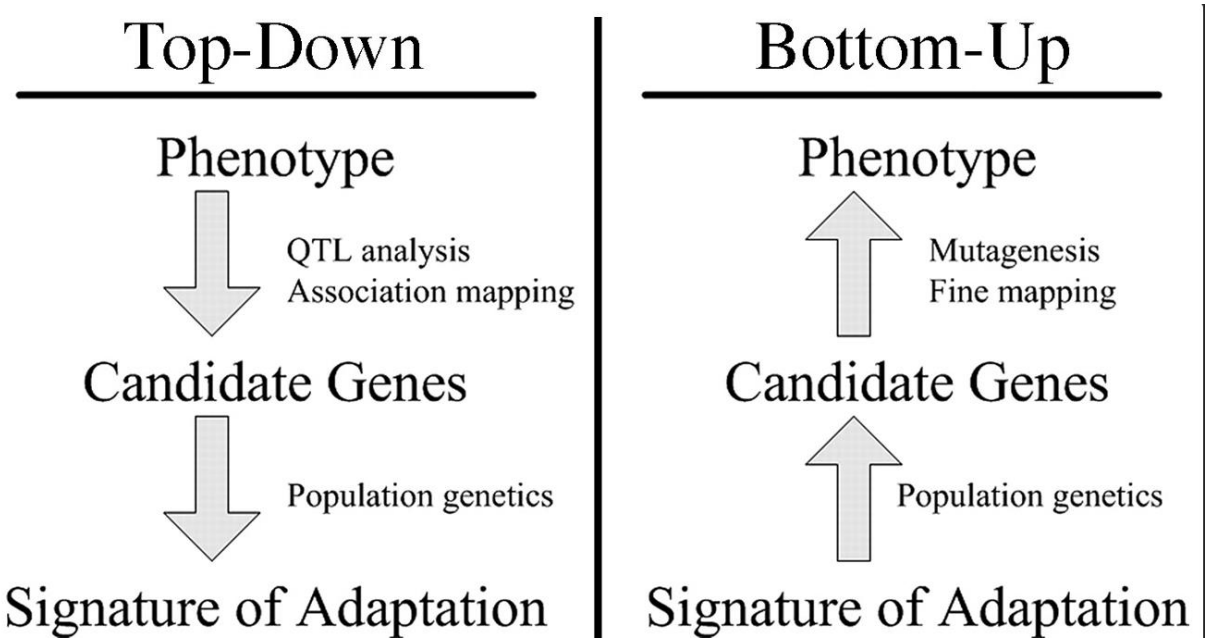


Figure 20 : Schéma de la hiérarchie phénotype-génotype représentée par des approches de « top-Down » et « bottom-up » **(Ross-Ibarra *et al.*, 2007)**

Ces deux approches font appel à des concepts différents. L'approche Top-Down utilise les concepts de la génétique quantitative alors que l'approche Bottom-up utilise les concepts de la génétique des populations.

Conclusion et perspectives

La recherche de gènes d'interaction entre bactéries et plantes est une approche innovante chez le pois. En effet, ces variables sont complexes car elles impliquent l'association entre deux organismes, et elles peuvent être à la fois quantitatives ou qualitatives.

La structuration est très compliquée à analyser et il existe une diversité d'outils bio-informatiques. Il est difficile de choisir la structuration la plus appropriée dans cette analyse de génétique quantitative. Pour avoir une meilleure idée de la structuration du panel, des analyses de FST peuvent être menées. En effet, le FST donne l'effet de subdivision entre les groupes de structuration et la population totale du panel par l'étude des différences des fréquences alléliques.

Les *P. fulvum* et les *P.sativum abyssinicum* étant vraiment à part dans les structurations, il serait intéressant de les supprimer pour évaluer les conséquences de leur absence, d'une part sur la structuration et d'autre part sur les résultats de GWAS. Peut-être cela permettrait-il de mettre en évidence d'autres allèles au sein de *P. sativum*.

Les analyses de GWAS pour de mêmes variables mais avec une structuration différente impactent la significativité des pics d'association. Dans chaque structuration, les pois Afghan ou d'Asie mais aussi les *P. fulvum* et *P.s. Abyssinicum* se retrouvent isolés.

Les analyses de GWAS sur le panel SYMBIOPEA ont permis de trouver des SNP en déséquilibre de liaison avec des gènes ou QTL putatifs causaux et ce, avec plus ou moins d'effet pour les variables impliquées dans la nodulation, notamment pour les variables de la longueur et de la surface des nodosités mais aussi dans le choix de souches en général.

L'amélioration des modèles utilisés dans les analyses de GWAS sur ce panel permettrait de confirmer les SNP trouvés mais pourrait aussi permettre de trouver d'autres SNP. Plusieurs modèles d'analyses de GWAS existent, comme le modèle MLMM prenant en covariable des SNP majeurs, permettant ainsi de visualiser les SNP cachés par ces SNP majeurs. D'autres outils plus performants que GEMMA, comme FarmCPU **(Zhang, 2014)**, existent et pourraient être utilisés pour ce type d'analyse. De plus, les analyses de GWAS peuvent être améliorées par optimisation du nombre d'individus dans le panel ; ce qui est fait pour le projet GRaSP.

Tous les SNP trouvés pour l'ensemble des variables vont faire l'objet de recherche de gènes sous–jacents à partir du calcul du déséquilibre de liaison (ici 200 kb). Ce déséquilibre de liaison permet de donner un intervalle de confiance dans la recherche de ces gènes putatifs **(Yu and Buckler, 2006)**. Ce travail de recherche de gènes potentiellement impliqués dans la variation des caractères est conséquent et sera réalisé après le stage. Une fois ces gènes trouvés, ils subiront une validation fonctionnelle. Cette validation peut se faire par TILLING.

Une fois ces analyses terminées, des essais aux champs (avec le mélange bactérien pour ces gènes mutants) vont ensuite être réalisés pour voir si les gènes ont toujours la même expression dans des milieux avec de fortes pressions biotiques et abiotiques. En complément, des analyses de « bottom-up » (figure 20), pour la recherche d'adaptations dans un ensemble de gènes, basées sur des méthodes traditionnelles de génétique des populations peuvent-être utilisées **(Ross-Ibarra *et al.*, 2007)**. Ces méthodes sont destinées à détecter les traces de la sélection en l'absence d'à priori sur la fonction des gènes. De telles approches, par exemple des génomes scans visant à identifier des zones génomiques présentant une différenciation excessive au regard du reste du génome, peuvent être intéressantes dans des cas où la structure génétique est confondue avec un trait phénotypique sous sélection.

Bibliographie

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Research **19**: 1655–1664

Alves-Carvalho S, Aubert G, Carrère S, Cruaud C, Brochot A-L, Jacquin F, Klein A, Martin C, Boucherot K, Kreplak J, et al (2015) Full-length *de novo* assembly of RNA-seq data in pea (*Pisum sativum* L.) provides a gene expression atlas and gives insights into root nodulation in this species. The Plant Journal **84**: 1–19

Berbel A, Ferrándiz C, Hecht V, Dalmais M, Lund OS, Sussmilch FC, Taylor SA, Bendahmane A, Ellis THN, Beltrán JP, et al (2012) VEGETATIVE1 is essential for development of the compound inflorescence in pea. Nature Communications. doi: 10.1038/ncomms1801

Bogdanova VS, Galieva ER, Kosterin OE (2009) Genetic analysis of nuclear-cytoplasmic incompatibility in pea associated with cytoplasm of an accession of wild subspecies Pisum sativum subsp. elatius (Bieb.) Schmahl. Theoretical and Applied Genetics **118**: 801–809

Borisov, Danilova, Koroleva, Naumkina, Pavlova, Pinaev, Shtark, Tsyganov, Voroshilova, Zhernakov, et al Pea (Pisum sativumL.) regulatory genes controlling development of nitrogen fixing nodule and arbuscular mycorrhiza: fundamentals and application. Biologia 13: 137—144, 2004

Bourion V, Heulin-Gotty K, Aubert V, Tisseyre P, Chabert-Martinello M, Pervent M, Delaitre C, Vile D, Siol M, Duc G, et al (2018) Co-inoculation of a Pea Core-Collection with Diverse Rhizobial Strains Shows Competitiveness for Nodulation and Efficiency of Nitrogen Fixation Are Distinct traits in the Interaction. Frontiers in Plant Science. doi: 10.3389/fpls.2017.02249

Bourion V, Laguerre G, Depret G, Voisin A-S, Salon C, Duc G (2007) Genetic Variability in Nodulation and Root Growth Affects Nitrogen Fixation and Accumulation in Pea. Annals of Botany **100**: 589–598

Bourion V, Rizvi SMH, Fournier S, de Larambergue H, Galmiche F, Marget P, Duc G, Burstin J (2010) Genetic dissection of nitrogen nutrition in pea through a QTL approach of root, nodule, and shoot variability. Theoretical and Applied Genetics **121**: 71–86

Broughton WJ, Jabbouri S, Perret X (2000) Keys to Symbiotic Harmony. J Bacteriol **182**: 5641–5652

Davis EO, Evans IJ, Johnston AWB (1988) Identification of nodX, a gene that allows Rhizobium leguminosarum biovar viciae strain TOM to nodulate Afghanistan peas. MGG Molecular & General Genetics **212**: 531–535

Depret Géraldine, Laguerre Gisèle (2008) Plant phenology and genetic variability in root and nodule development strongly influence genetic structuring of Rhizobium leguminosarum biovar viciae populations nodulating pea. New Phytologist **179**: 224–235

Desgroux A, Baudais VN, Aubert V, Le Roy G, de Larambergue H, Miteul H, Aubert G, Boutet G, Duc G, Baranger A, et al (2018) Comparative Genome-Wide-Association Mapping Identifies Common Loci Controlling Root System Architecture and Resistance to Aphanomyces euteiches in Pea. Frontiers in Plant Science. doi: 10.3389/fpls.2017.02195

Doležel J, Greilhuber J (2010) Nuclear genome size: Are we getting closer? Cytometry Part A **77A**: 635–642

Duarte J, Rivière N, Baranger A, Aubert G, Burstin J, Cornet L, Lavaud C, Lejeune-Hénaut I, Martinant J-P, Pichon J-P, et al (2014) Transcriptome sequencing for high throughput SNP development and genetic mapping in Pea. BMC Genomics **15**: 126

Ferguson BJ, Indrasumunar A, Hayashi S, Lin M-H, Lin Y-H, Reid DE, Gresshoff PM (2010) Molecular Analysis of Legume Nodule Development and Autoregulation. Journal of Integrative Plant Biology **52**: 61–76

Galloway JN, Townsend AR, Erisman JW, Bekunda M, Cai Z, Freney JR, Martinelli LA, Seitzinger SP, Sutton MA (2008) Transformation of the Nitrogen Cycle: Recent Trends, Questions, and Potential Solutions. Science **320**: 889–892

Geurts R, Fedorova E, Bisseling T (2005) Nod factor signaling genes and their function in the early stages of Rhizobium infection. Current Opinion in Plant Biology **8**: 346–352

Goh L, Yap VB (2009) Effects of normalization on quantitative traits in association test. BMC Bioinformatics **10**: 415

Gough C, Cullimore J (2011) Lipo-chitooligosaccharide Signaling in Endosymbiotic Plant-Microbe Interactions. MPMI **24**: 867–878

Hoffman GE (2013) Correcting for Population Structure and Kinship Using the Linear Mixed Model: Theory and Extensions. PLoS ONE **8**: e75707

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theoretical and Applied Genetics **38**: 226–231

International Conference on Legume Genetics and Genomics (2017) Book of abstracts: ICLGG 2017. s. n., s. l.

Jing R, Vershinin A, Grzebyta J, Shaw P, Smýkal P, Marshall D, Ambrose MJ, Ellis TN, Flavell AJ (2010) The genetic diversity and evolution of field pea (Pisum) studied by high throughput retrotransposon based insertion polymorphism (RBIP) marker analysis. BMC Evolutionary Biology **10**: 44

Kaló P, Seres A, Taylor SA, Jakab J, Kevei Z, Kereszt A, Endre G, Ellis THN, Kiss GB (2004) Comparative mapping between<Emphasis Type="Italic"> Medicago sativa</Emphasis> and<Emphasis Type="Italic"> Pisum sativum</Emphasis>. Mol Genet Genomics **272**: 235–246

Korte A, Farlow A (2013) The advantages and limitations of trait analysis with GWAS: a review. Plant Methods **9**: 29

Laguerre G, Louvrier P, Allard M-R, Amarger N (2003) Compatibility of Rhizobial Genotypes within Natural Populations of Rhizobium leguminosarum Biovar viciae for

Nodulation of Host Legumes. Applied and Environmental Microbiology **69**: 2276–2283

Laguerre Gisèle, Depret Géraldine, Bourion Virginie, Duc Gérard (2007) Rhizobium leguminosarum bv. viciae genotypes interact with pea plants in developmental responses of nodules, roots and shoots. New Phytologist **176**: 680–690

Madoui M, Labadie K, Agata L, Aury J, Kreplak J, Gali KK (2016) Assembly of the pea genome by integration of high throughput sequencing (PacBio and Illumina) and whole genome profiling (WGPTM) data. San Diego

Maj D, Wielbo J, Marek-Kozaczuk M, Skorupska A (2010) Response to flavonoids as a factor influencing competitiveness and symbiotic activity of Rhizobium leguminosarum. Microbiological Research **165**: 50–60

Oldroyd GED, Downie JA (2008) Coordinating Nodule Morphogenesis with Rhizobial Infection in Legumes. Annual Review of Plant Biology **59**: 519–546

Peix A, Ramírez-Bahena MH, Velázquez E, Bedmar EJ (2015) Bacterial Associations with Legumes. Critical Reviews in Plant Sciences **34**: 17–42

Prudent M, Vernoud V, Girodet S, Salon C (2016) How nitrogen fixation is modulated in response to different water availability levels and during recovery: A structural and functional study at the whole plant level. Plant and Soil **399**: 1–12

Raj A, Stephens M, Pritchard JK (2014) fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. Genetics **197**: 573–589

Ross-Ibarra J, Morrell PL, Gaut BS (2007) Plant domestication, a unique opportunity to identify the genetic basis of adaptation. Proceedings of the National Academy of Sciences **104**: 8641–8648

Satgé C, Moreau S, Sallet E, Lefort G, Auriac M-C, Remblière C, Cottret L, Gallardo K, Noirot C, Jardinaud M-F, et al (2016) Reprogramming of DNA methylation is critical for nodule development in Medicago truncatula. Nature Plants. doi: 10.1038/nplants.2016.166

Schorderet DF, Iouranova A, Favez T, Tiab L, Escher P (2013) IROme, a New High-Throughput Molecular Tool for the Diagnosis of Inherited Retinal Dystrophies. BioMed Research International **2013**: 1–9

Segura V, Vilhjálmsson BJ, 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. Nature Genetics **44**: 825–830

Smýkal P, Kenicer G, Flavell AJ, Corander J, Kosterin O, Redden RJ, Ford R, Coyne CJ, Maxted N, Ambrose MJ, et al (2011) Phylogeny, phylogeography and genetic diversity of the Pisum genus. Plant Genetic Resources **9**: 4–18

Tayeh Nadim, Aluome Christelle, Falque Matthieu, Jacquin Françoise, Klein Anthony, Chauveau Aurélie, Bérard Aurélie, Houtin Hervé, Rond Céline, Kreplak Jonathan, 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. The Plant Journal **84**: 1257–1273

Voisin A., Cellier P, Jeuffoy M. (2015) Fonctionnement de la symbiose fixatrice de N2 des légumineuses à graines :Impacts Agronomiques et Environnementaux. Innovations Agronomiques 43 139–160

Wang H, Smith KP, Combs E, Blake T, Horsley RD, Muehlbauer GJ (2012) Effect of population size and unbalanced data sets on QTL detection using genome-wide association mapping in barley breeding germplasm. Theoretical and Applied Genetics **124**: 111–124

Wenden N., Ellis TH., Timmerman-Vaughan G., Dirlewanger E (1996) The current pea linkage map. Pisum Genetics

Willems A (2006) The taxonomy of rhizobia: an overview. Plant and Soil **287**: 3–14

Young JM, Park D-C, Weir BS (2004) Diversity of 16S rDNA sequences of *Rhizobium* spp. implications for species determinations. FEMS Microbiology Letters **238**: 125–131

Yu J, Buckler ES (2006) Genetic association mapping and genome organization of maize. Current Opinion in Biotechnology **17**: 155–160

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, et al (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics **38**: 203–208

Zhou X, Stephens M (2012) Genome-wide efficient mixed-model analysis for association studies. Nature Genetics **44**: 821–824

Zhu C, Gore M, Buckler ES, Yu J (2008) Status and Prospects of Association Mapping in Plants. The Plant Genome **1**: 5–20

Zhukov V, Radutoiu S, Madsen LH, Rychagova T, Ovchinnikova E, Borisov A, Tikhonovich I, Stougaard J (2008) The Pea Sym37 Receptor Kinase Gene Controls Infection-Thread Initiation and Nodule Development. MPMI **21**: 1600–1608

Zohary D, Hopf M, Weiss E (2012) Domestication of Plants in the Old World: The Origin and Spread of Domesticated Plants in Southwest Asia, Europe, and the Mediterranean Basin. OUP Oxford

Sitographie

FAO, Organisation des Nations Unies pour l'Alimentation et l'Agriculture, 2016. 2016 année internationale des légumineuses [en ligne].Disponible sur http://www.fao.org/pulses-2016/fr/ (consulté le 20/01/2018).

Terres Univia, 2017. Les espèces cultivées, le pois [en ligne].Disponible sur http://www.terresunivia.fr/cultures-utilisation/les-especes-cultivees/pois (consulté le 21/01/2018).

FranceAgriMer, 2017. Les protéagineux, le pois [en ligne]. Disponible sur http://www.franceagrimer.fr/Donnees-de-reference/Termes-de-classement/Produits-vegetaux/proteagineux/pois (consulté le 22/01/2018).

BASF, Badische Anilin- & Soda Fabrik, 2018. Cours du pois [en ligne].Disponible sur https://www.agro.basf.fr/agroportal/fr/fr/services_et_outils/infoservices/cours_et_marches/cours_et_marches_pois/cotation_pois.html (consulté le 21/01/2018)

NZ Rhizobia,Bactérial and fungal systematics research,2016.The current taxonomy of rhizobia [en ligne]. Disponible sur https://www.rhizobia.co.nz/taxonomy/rhizobia (consulté le 24/01/2018).

INRA de Dijon, Institut National de la Recherche Agronomique de Dijon, 2012. A portal to INRA Dijon Legume genetic and genomic resources [en ligne]. Disponible sur http://www.thelegumeportal.net (consulté le 20/03/2018).

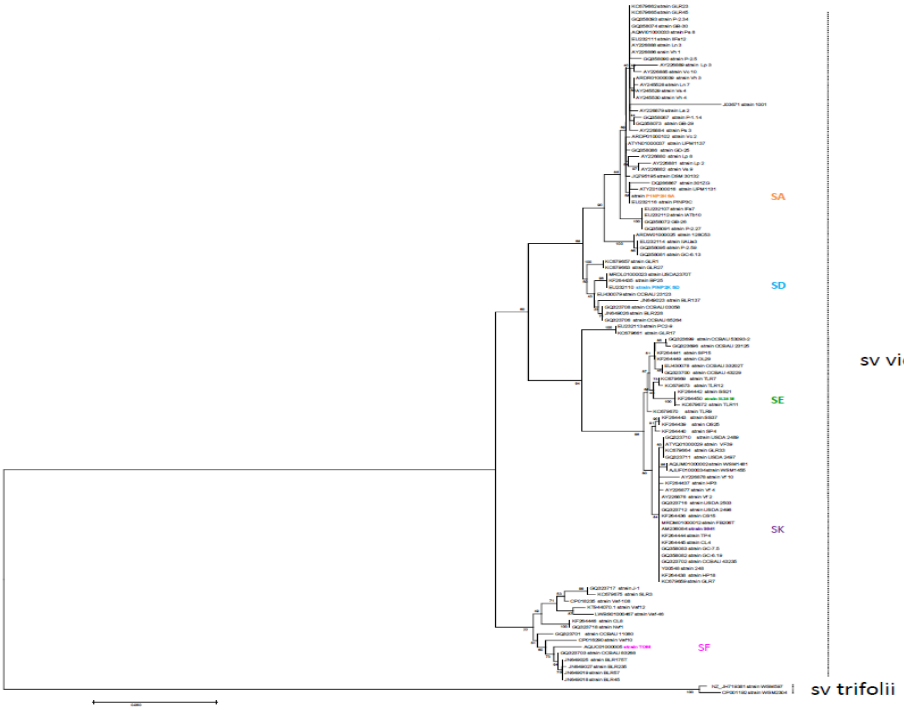
Annexes

Annexe I : Tableau des accessions du projet SYMBIOPEA ; d'après **Bourion *et al*, 2018**

Codage	Nom commun	Espéce	Statut	Pays d'origine
R014	PISUM FULVUM JI2473	Pf	Wd	Israel
R015	PISUM FULVUM JI2523	Pf	Wd	Syria
R037	ABYSSINICUM VAVILOVANIUM	Psa	Lr	Ethiopia
R013	PISUM ABYSSINICUM JI2202	Psa	Lr	Yemen
R075	PISUM ELATIUS JI261	Pse	Semi-Wd/Wd	Turkey
R010	PISUM ELATIUS JI1075	Pse	Wd	Turkey
R069	PISUM HUMILE JI241	Psh	Semi-Wd/Wd	Israel
R098	271-134	Ps	Lr	Afghanistan
R029	PI212112	Ps	Lr	Afghanistan
R028	SHRAT	Ps	Lr	India
R002	PISUM SATIVUM-AFGHANISTAN JI86	Ps	Lr	Afghanistan
R036	NEPAL A	Ps	Lr	Nepal
R005	PISUM SATIVUM-INDIA JI1267	Ps	Lr	India
R035	AFGHANISTAN ASIATICUM	Ps	Lr	Afghanistan
R054	CHINA JI1491	Ps	Lr	China
R012	PISUM HUMILE JI1794	Psh	Wd	Israel
R066	K4088	Ps	Lr	Ukraine
R070	PISUM ELATIUS JI1089	Pse	Semi-Wd/Wd	Turkey
R017	PISUM SPECIOSUM-LIBYA JI2605	Ps	Semi-Wd/Wd	Libya
R074	WIRAIG JI190	Ps	Lr	Sudan
R099	PISUM SATIVUM-ETHIOPIA JI281	Ps	Lr	Ethiopia
R016	PISUM TRANSCAUCASICUM JI2546	Ps	Wd	Georgia
R095	FRISSON	Ps	Cv	France
R104	TIL336/11	Ps	Bl	France
R101	ISARD	Ps	Cv	France
R071	CE101=FP	Ps	Bl	France
R020	MISTRAL	Ps	Cv	France
R072	CHAMPAGNE	Ps	Lr	France
R073	DP	Ps	Gm	France
R081	HOLLY11	Ps	Lr	Hungary
R067	PISUM SATIVUM-HIBERNICUM JI1846	Ps	Cv	Egypt
R085	WINTERBERGER	Ps	Lr	Germany
R007	PISUM SATIVUM-MEXICO JI1844	Ps	Lr	Mexico
R082	KARNOBAT	Ps	Cv	Bulgaria
R019	COTE D'OR	Ps	Lr	France
R083	KAZAR	Ps	Cv	France
R022	WNC Z61Z SPP ARVENSE Z009	Ps	Lr	NA
R080	GLACIER	Ps	Cv	United States
R076	MELROSE	Ps	Cv	United States
R061	PISUM SATIVUM-ETHIOPIA JI1431	Ps	Lr	Ethiopia
R096	L1073	Ps	Bl	Sweden
R038	CAPSICUM	Pse	Lr	Azerbaidjan
R011	PISUM ELATIUS JI1703	Pse	Wd	NA
R062	COSTA RICA JI975	Ps	Lr	Costa Rica
R026	DARK SKIN PERFECTION	Ps	Cv	Great Britain
R056	MESSIRE	Ps	Cv	France
R049	AMINO	Ps	Cv	France
R077	TERESE	Ps	Cv	Denmark
R100	PUGET	Ps	Cv	Great Britain
R084	TORSDAG	Ps	Cv	Sweden
R055	SOMMETTE	Ps	Cv	Netherlands
R051	CAMEOR	Ps	Cv	France
R064	K1666	Ps	Lr	Russia
R063	K4269	Ps	Lr	Lithuania
R065	K8290/NORD	Ps	Cv	Russia
R058	ZP126	Ps	Lr	Spain
R057	ZP141	Ps	Lr	Spain
R068	YANGWAN	Ps	Lr	China
R053	CERISE-ce,CR	Ps	Gm	Netherlands
R079	CHEYENNE	Ps	Cv	France
R052	BALLET	Ps	Cv	Great Britain
R097	AUSTIN	Ps	Cv	France
R094	90-2079	Ps	Bl	United States
R027	90-2131	Ps	Bl	United States
R001	KOROZA	Ps	Cv	Netherlands
R024	PI180693	Ps	Cv	Germany
R008	PISUM SATIVUM-ZAIRE JI2376	Ps	Semi-Wd/Wd	DR Congo
R103	ASTRONAUTE	Ps	Cv	France
R102	KAYANNE	Ps	Cv	France
R078	BACCARA	Ps	Cv	France
R021	WNC 23Z SPP ARVENSE 1809	Ps	Lr	NA
R025	552	Ps	Bl	United States
R004	ALASKA	Ps	Cv	United States
R086	DU CHEMIN LONG	Ps	Cv	France
R087	CLAMART HATIF	Ps	Cv	France
R041	FIN DE LA BIEVRE	Ps	Cv	France
R039	HATIVER	Ps	Cv	Netherlands
R042	MERVEILLE D'ETAMPES	Ps	Cv	France
R088	MICHAUX DE PARIS	Ps	Cv	Netherlands
R018	AURALIA	Ps	Cv	Germany
R089	CHAMPION D'ANGLETERRE	Ps	Cv	Great Britain
R046	LIVIOLETTA	Ps	Cv	Germany
R045	DESIREE	Ps	Cv	Netherlands
R031	IREGI SARGA	Ps	Cv	Hungary
R032	KIRIN 40	Ps	Bl	China
R033	CUZCO 1	Ps	Lr	Peru
R044	PETIT PROVENCAL	Ps	Cv	Great Britain
R090	PLEIN LE PANIER	Ps	Cv	Great Britain
R091	SERPETTE D'AUVERGNE	Ps	Cv	France
R040	TELEPHONE A RAMES	Ps	Cv	Great Britain
R092	CAROUBY DE MAUSSANE	Ps	Cv	France
R093	CORNE DE BELIER	Ps	Cv	France
R034	HAITI COLORE	Ps	Lr	Haiti
R047	BINGEFORS	Ps	Bl	Sweden
R059	PISUM SATIVUM-ETHIOPIA JI1594	Ps	Lr	Ethiopia
R060	PISUM SATIVUM-MONGOLIA JI1345	Ps	Lr	Mongolia
R043	NFG KRUPP PELUSCHKE	Ps	Cv	Germany
R030	ENGLISH	Ps	Bl	Great Britain

Pf: *Pisum fulvum*, Psa: *Pisum sativum* subsp. *abyssinicum*,
Ps: *Pisum sativum*, Pse: *Pisum sativum* subsp. *elatius*, Psh: *Pisum sativum* subsp. *humile*;
Bl: Breeding line, Cv: Cultivar, Gmp: germplasm, Ld: landrace, Semi-Wd: Semi-wild accession,
Wd: wild accession

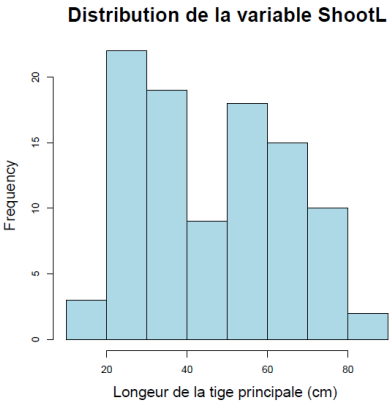
Annexe II : Analyse phylogénétique des séquences nodD par la méthode de Maximun Likelihood **(Bourion *et al*., 2018)**

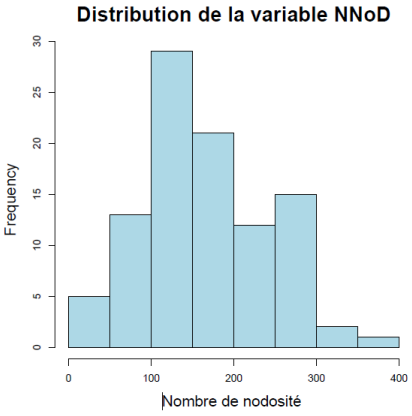


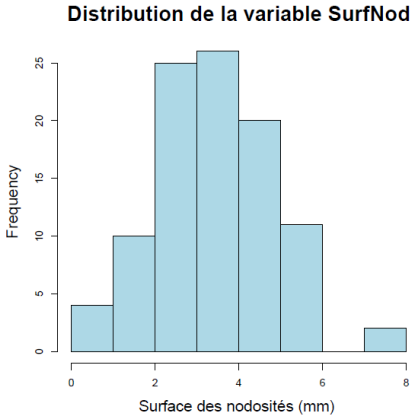
Annexe III : Codage des variables quantitatives phénotypiques

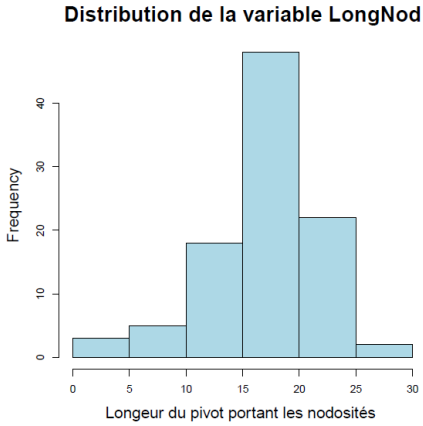
Variable	Signification	Quantification
NLeaf	Stade en nb d'étages de feuilles développées	mesure manuelle
InterL	Longueur moyenne entrenoeud (cm)	=ShootL/Nleaf
ShootL	Longueur tige principale jusqu'à apex (cm)	mesure manuelle
LeafA	Surface totale de feuilles (cm²)	analyse image scan feuilles prélevées
ShootB	Poids sec partie aérienne (mg)	pesée
TaRootL	Longueur pivot (cm)	mesure manuelle
TRootL	Longueur totale racines (cm)	analyse image système scan système racinaire entier
RootProjA	Surface totale de racines (cm²)	analyse image système scan système racinaire entier
RootDia	Diamètre moyen des racines (mm)	analyse image système scan système racinaire entier
RootB	Poids sec racines sans nodosités (mg)	pesée
NLatRoot	Nombre racines latérales premier ordre	mesure manuelle
LatRootDens	Densité racines latérales premier ordre	= NLatRoot/longueur pivot
NNodTR	Nombre nodosités sur pivot	mesure manuelle
LongNod	Longueur pivot avec nodosités (cm)	mesure manuelle
NNod	Nombre total nodosités	mesure manuelle
SurfNod	Surface totale de nodosités (cm²)	analyse image scan nodosités prélevées
NodB	Poids sec nodosités (mg)	pesée
ANodB	Poids sec moyen d'une nodosité	=NodB/NNod
NodB_BGB	Poids sec nodosités /Poids sec Racines + Nod	=NodB/(RootB+NodB)
BGB_TB	Poids sec nodosités + Racines/Poids sec total	=(NodB+RootB)/(NodB+RootB+ShootB)
ShootNC	Teneur en azote des parties aériennes	analyse labo
NDFA	Taux fixation symbiotique	analyse labo
TSW	Poids de 1000 grains	pesée
SA	% nodosités avec souche A	mesure à partir échantillon de 60 nodosités prélevées
SD	% nodosités avec souche D	mesure à partir échantillon de 60 nodosités prélevées
SE	% nodosités avec souche E	mesure à partir échantillon de 60 nodosités prélevées
SF	% nodosités avec souche F	mesure à partir échantillon de 60 nodosités prélevées
SK	% nodosités avec souche K	mesure à partir échantillon de 60 nodosités prélevées
NSA	Nombre de nodosités avec Souche A	=NNod x SA
NSD	Nombre de nodosités avec Souche D	=NNod x SD
NSE	Nombre de nodosités avec Souche E	=NNod x SE
NSF	Nombre de nodosités avec Souche F	=NNod x SF
NSK	Nombre de nodosités avec Souche K	=NNod x SK

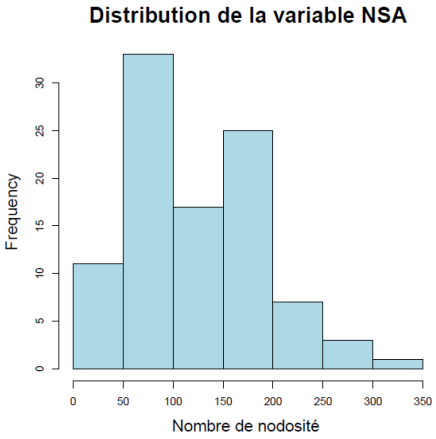
Annexe IV : Distribution des variables (ShootL et variable de nodulation)

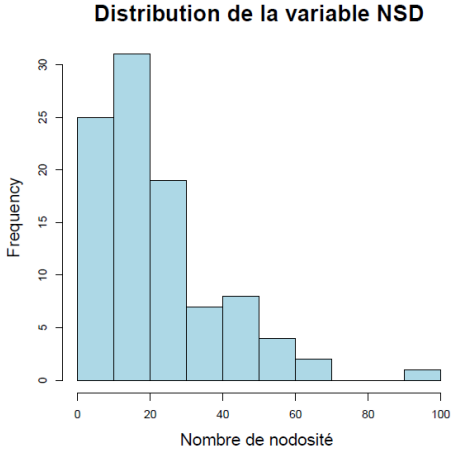


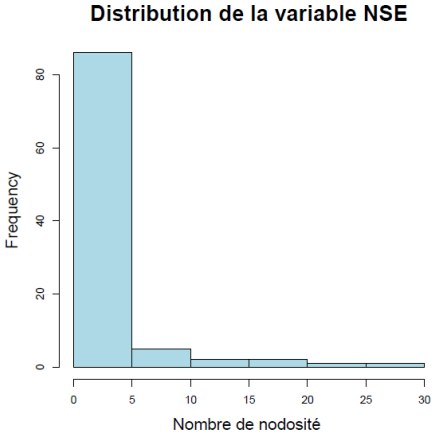


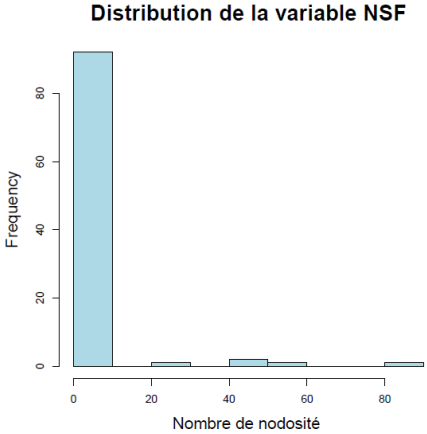


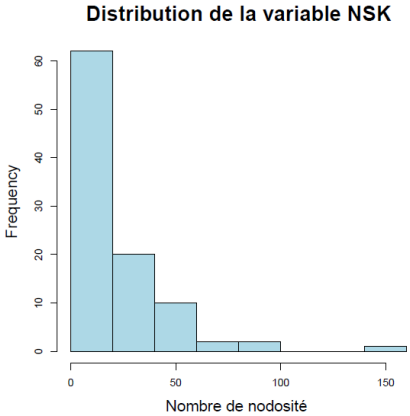








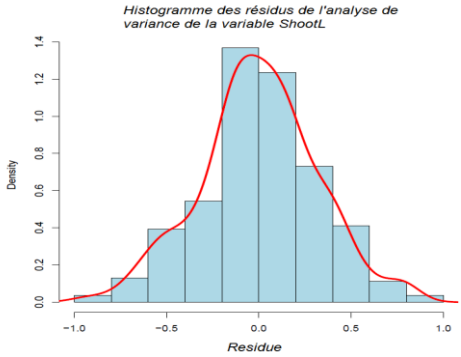


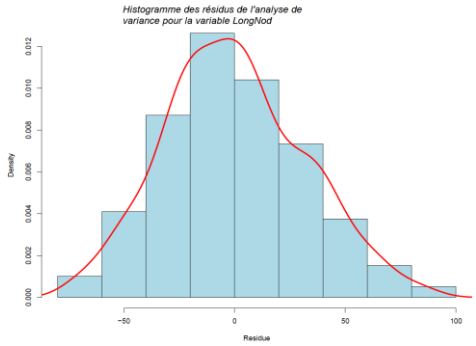


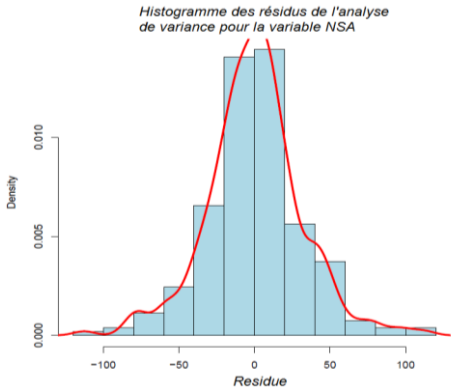
Annexe V : Analyse de Variance prenant en compte l'effet date de semis (Sowing) et l'effet génotype (AccNb)

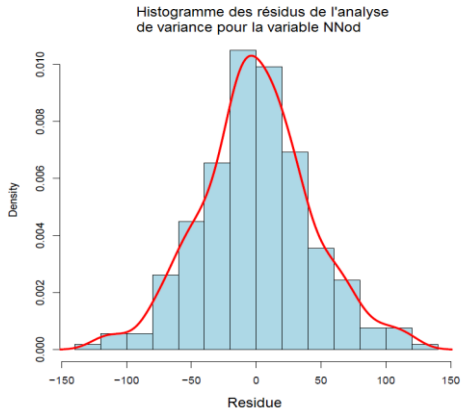
Variable/Significativité	AccNb	Sowing
NNod	<2.2e-16***	<2.2e-16***
ShootL	<2.2e-16***	5.971e-12***
LongNod	4.103e-05***	0.009035**
InterL	<2.2e-16***	0.4636
LeafA	<2.2e-16***	1.839e-15***
ShootB	<2.2e-16***	<2.2e-16***
TapRootL	0.04134*	0.52674
TrootL	<2.2e-16***	1.089e-12***
RootProjA	<2.2e-16***	<2.2e-16***
RootDia	2.983e-10***	1.613e-08***
RootB	<2.2e-16***	<2.2e-16***
NLatRoot	5.33e-15***	0.2121
LatRootDens	2.769e-10***	0.0615.
NNodTR	<2.2e-16***	0.008123**
SurfNod	<2.2e-16***	2.357e-11***
NodB	<2.2e-16***	0.001713**
ANodB	1.306e-09***	0.000537***
NodB_BGB	<2.2e-16***	1.307e-14***
BGB_TB	<2.2e-16***	8.774e-06***
ShootNC	<2.2e-16***	0.002407**
NDFA	<2.2e-16***	2.426e-06***
TSW	<2.2e-16***	0.1662
SA	<2.2e-16***	X
SD	<2.2e-16***	X
SE	<2.2e-16***	X
SF	<2.2e-16***	X
SK	<2.2e-16***	X
NSA	<2.2e-16***	2.35e-16***
NSD	<2.2e-16***	8.573e-11***
NSE	<2.2e-16***	0.003677**
NSF	<2.2e-16***	0.06959.
NSK	<2.2e-16***	1.016e-09***

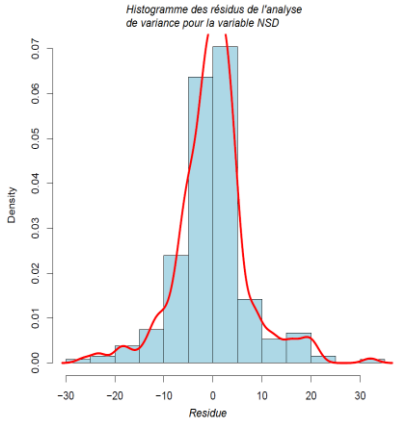
Annexe VI : Distribution des résidus des analyses de variances des variables (ShootL et variable de nodulation)

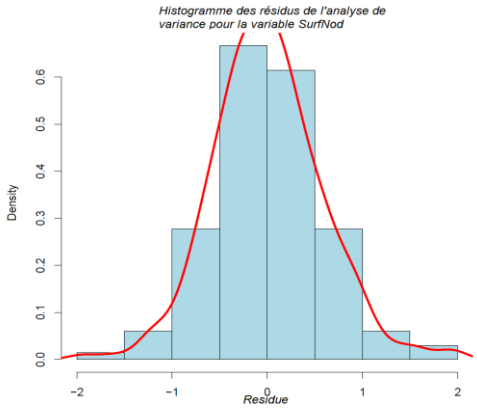


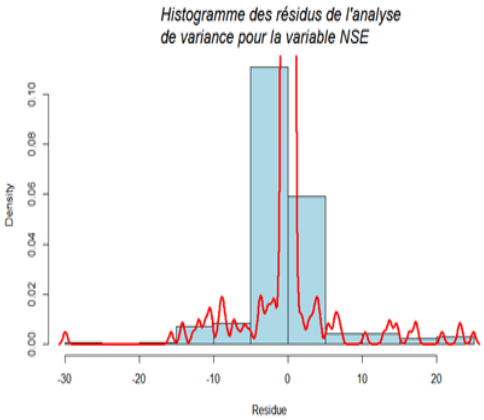


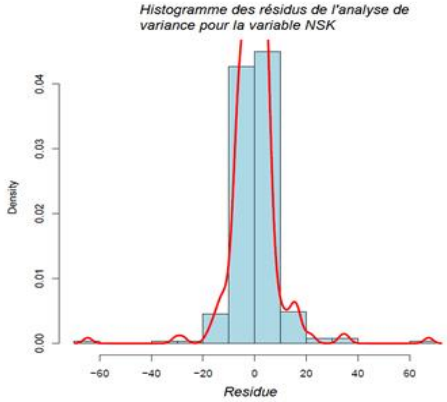


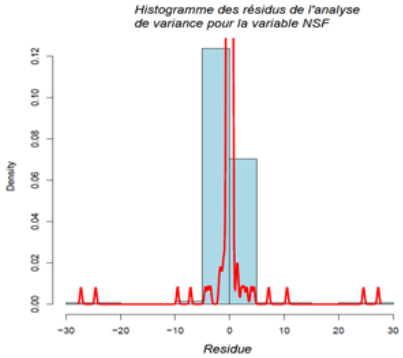






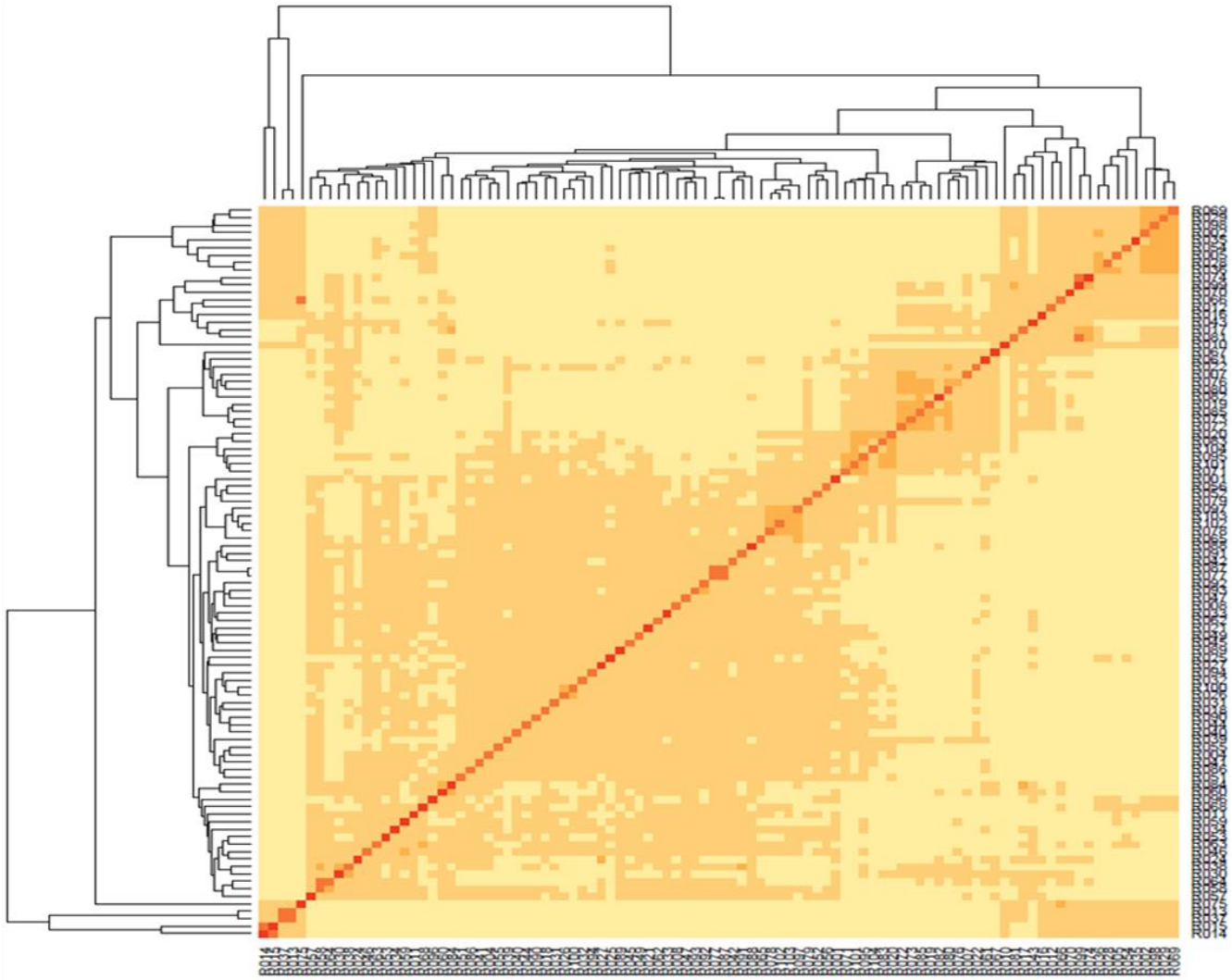




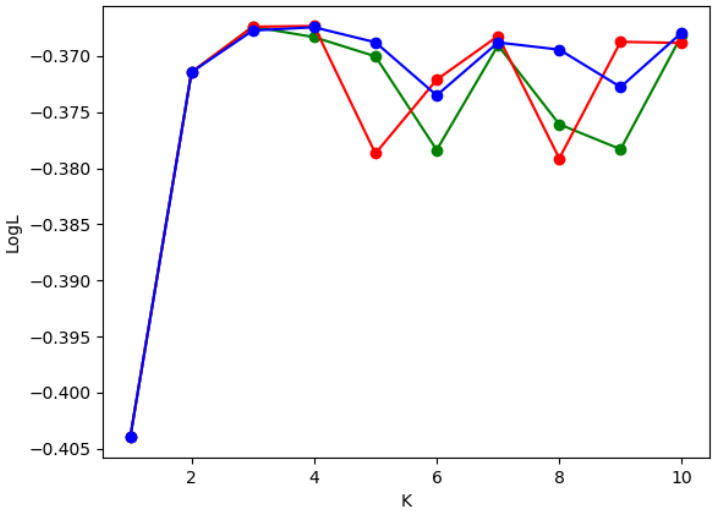


Annexe VII : Matrice d'apparentement entre les individus du panel SYMBIOPEA

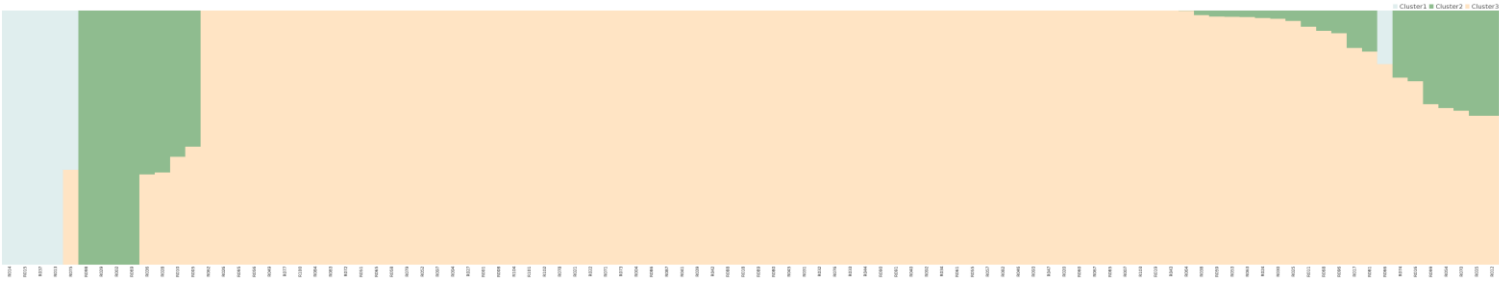
Les apparentements les plus forts sont de couleur rouge alors que les apparentements les plus faibles sont de couleur jaune.



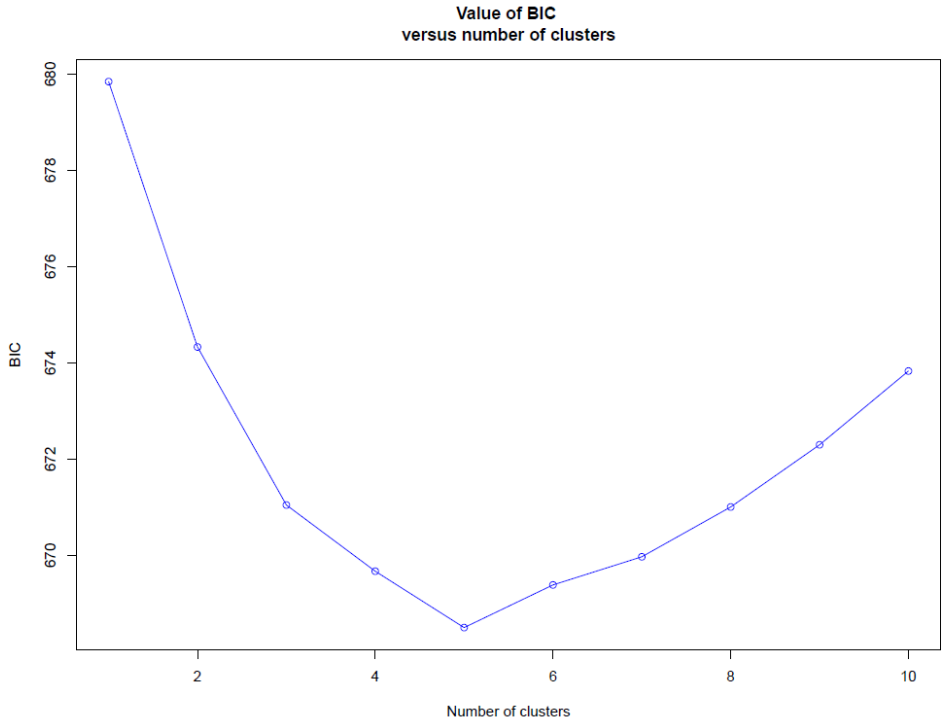
Annexe VIII : Courbe de log de FastStructure



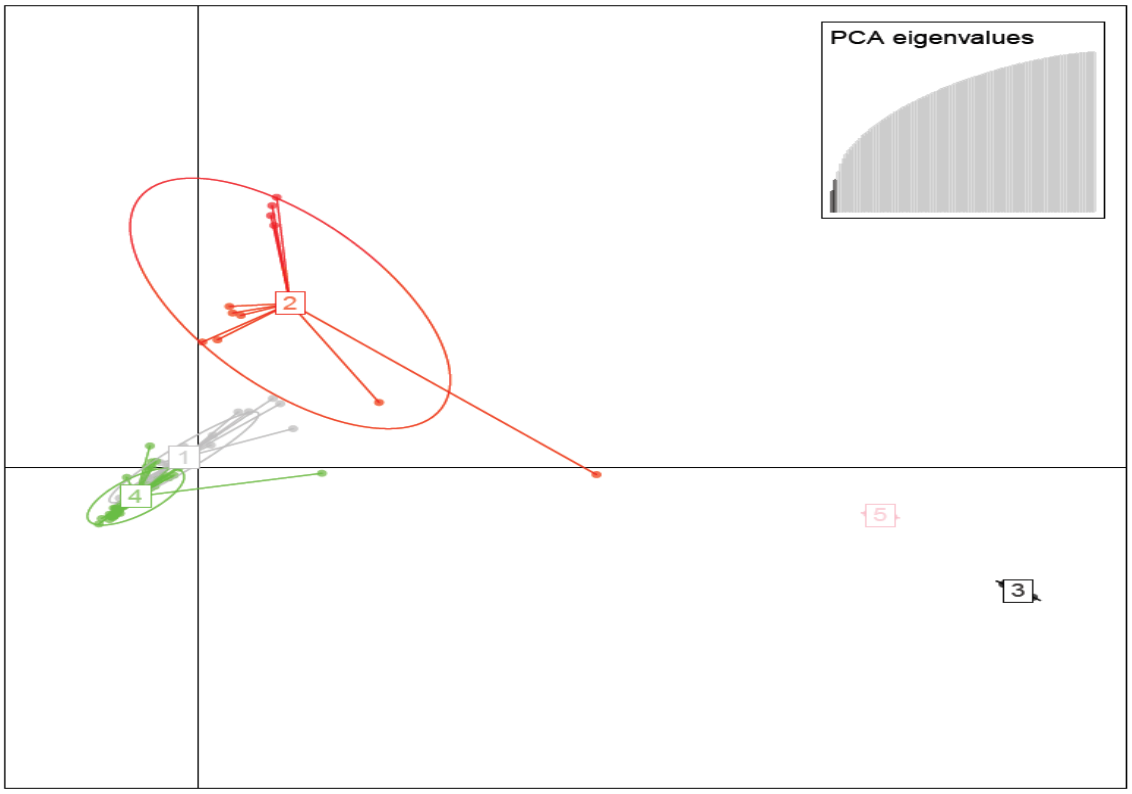
Annexe VIIII : Sortie de FastStructure sur le panel SYMBIOPEA



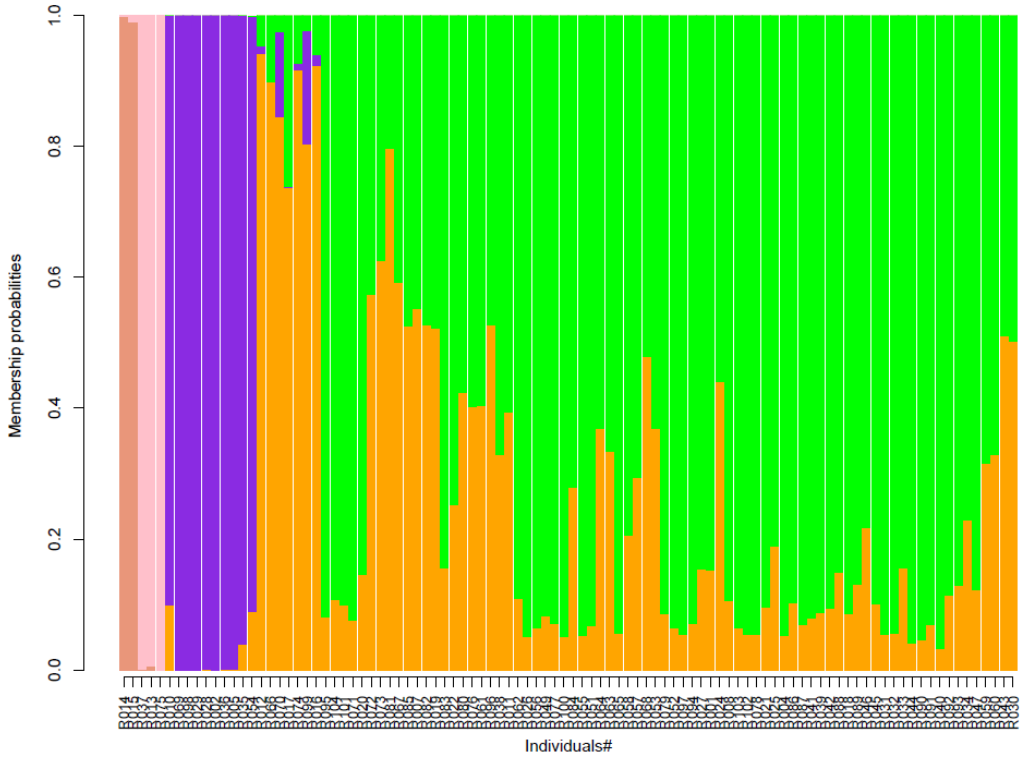
Annexe X : Courbe de BIC de DAPC



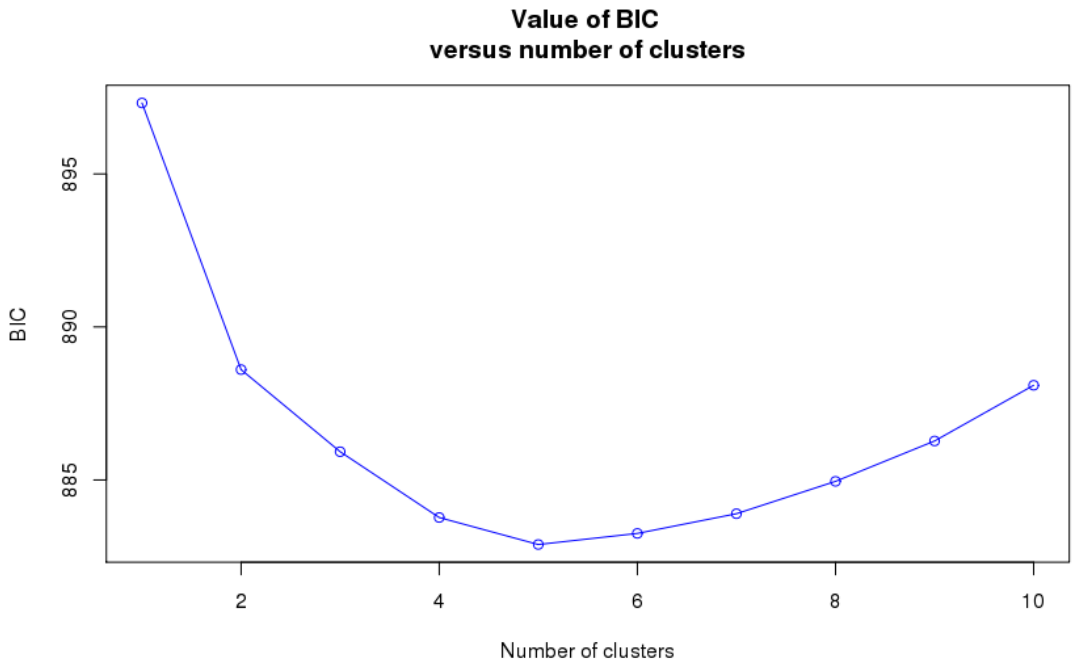
Annexe XI : Sortie de DAPC sur le panel SYMBIOPEA.



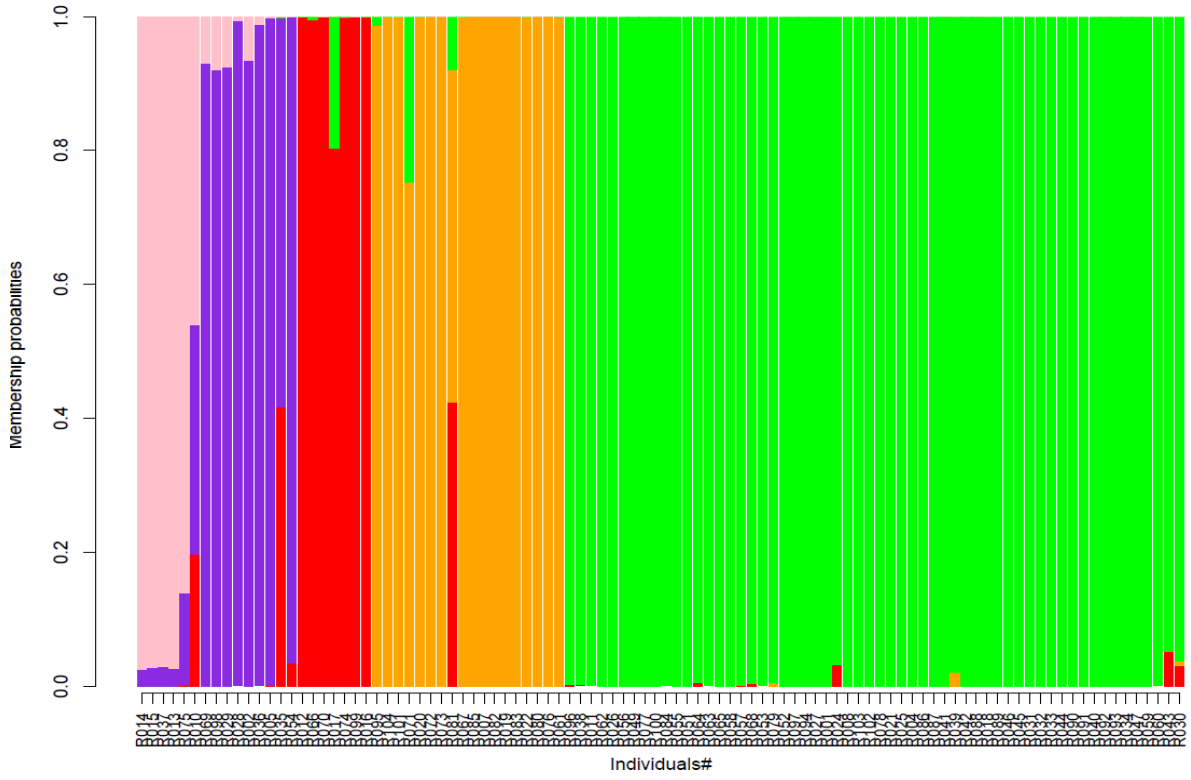
Annexe XII : Admixture entre les accessions par étude de structure DAPC sur le panel SYMBIOPEA.



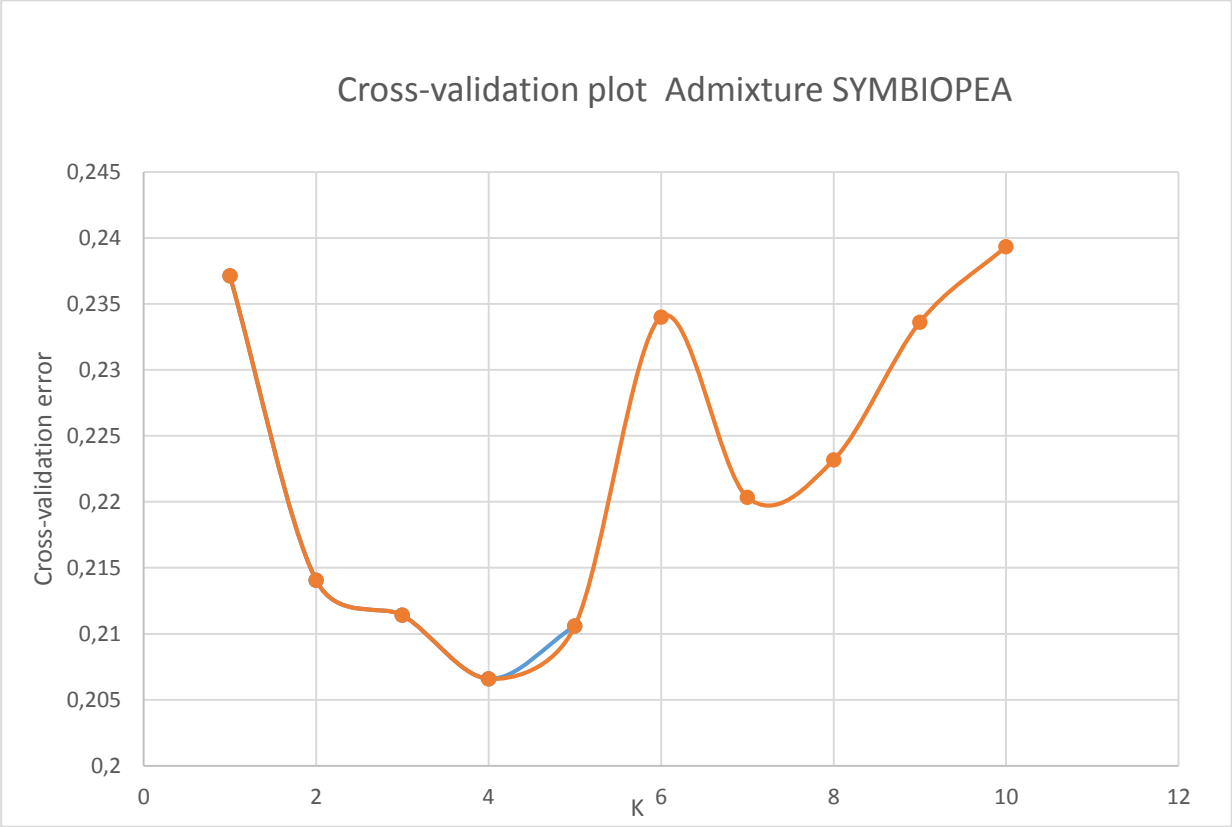
Annexe XIII : Courbe de BIC de DAPC avec filtrage de la MAF à 5% sur le panel SYMBIOPEA



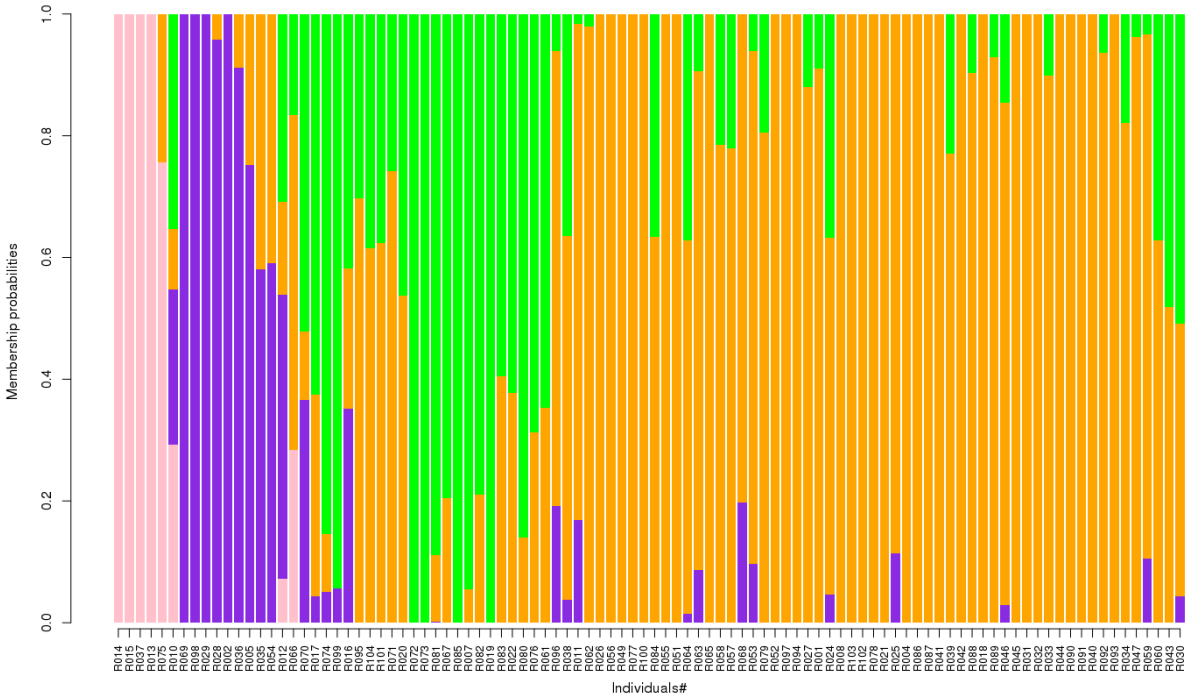
Annexe XIV : Admixture entre les accessions par étude de structure DAPC avec filtrage à la MAF à 5% sur le panel SYMBIOPEA



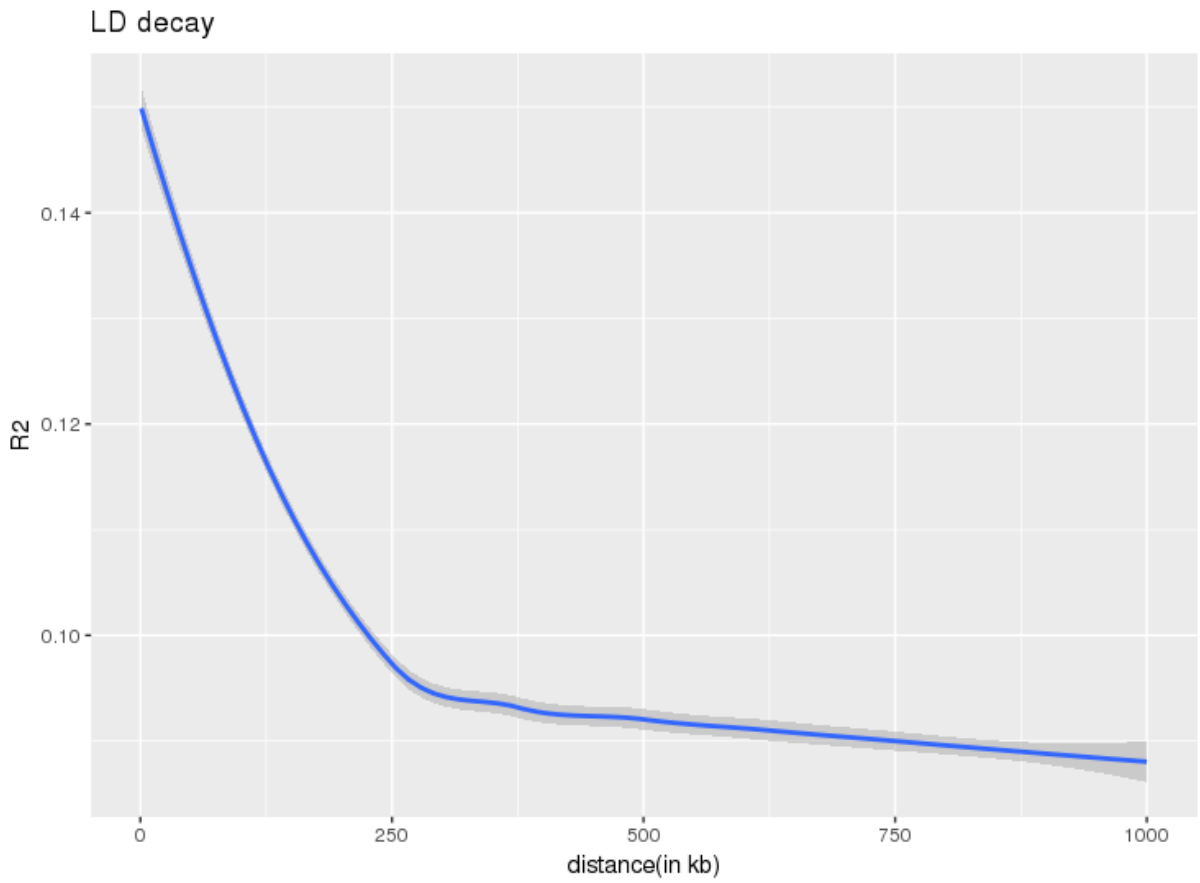
Annexe XV : Courbe de cross-validation de la structuration du panel SYMBIOPEA par le logiciel Admixture



Annexe XVI : Sortie de la structuration par le logiciel Admixture sur le panel SYMBIOPEA



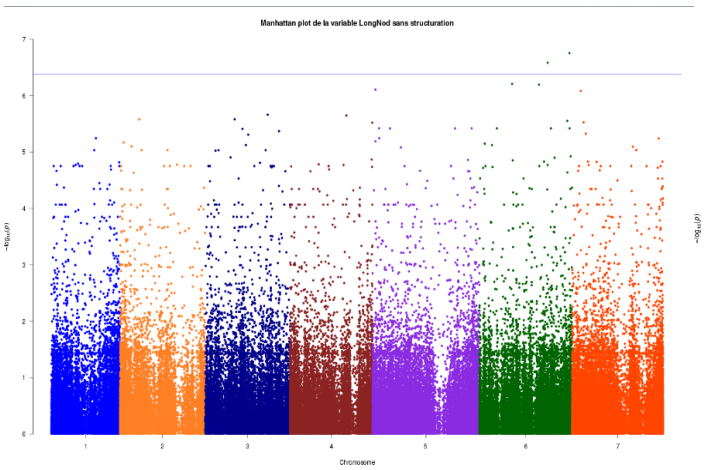
Annexe XVII : Courbe de déséquilibre de liaison sur les marqueurs de SYMBIOPEA

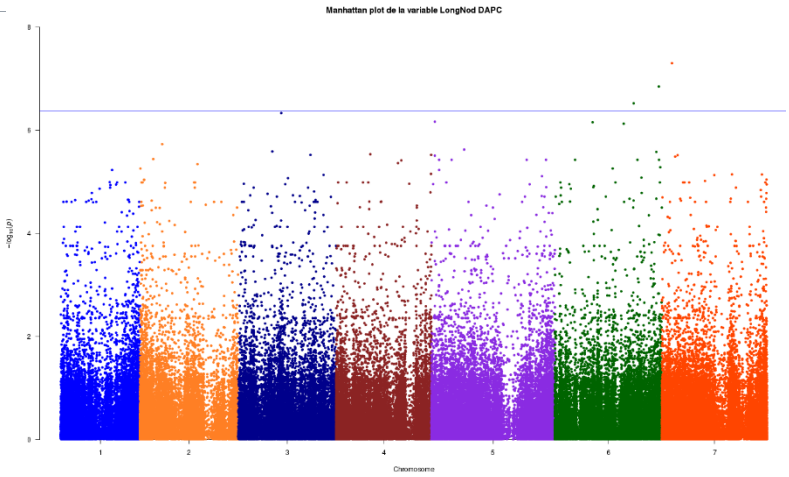


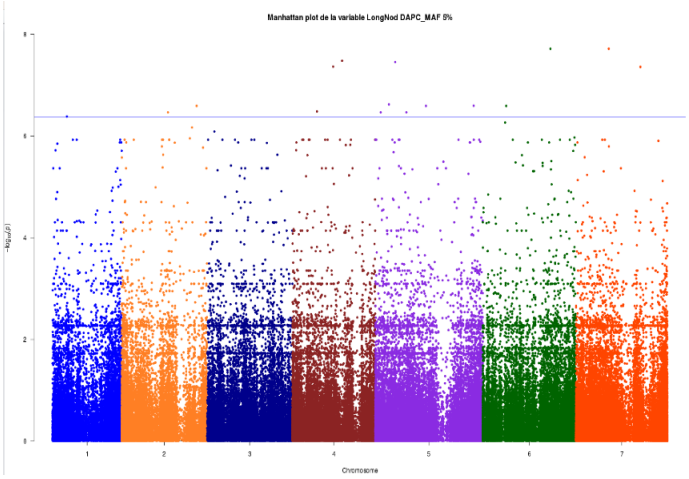
Annexe XVIII : Héritabilité des variables analysées par GWAS

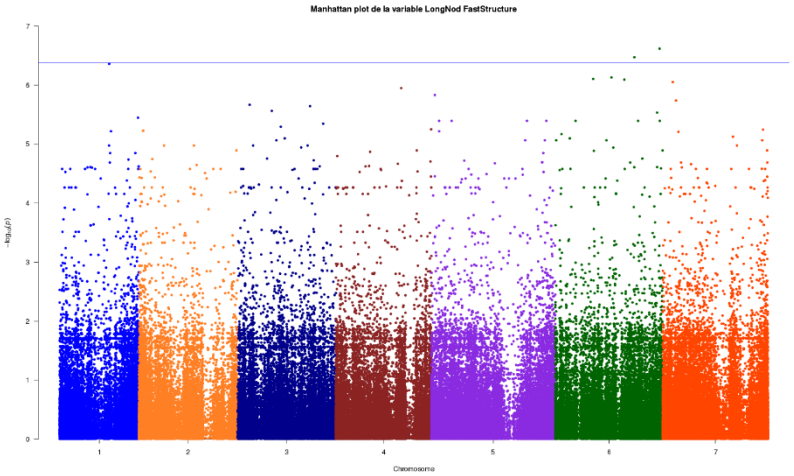
Structuration\variable phénotypique	ShootL	Nnod	LongNod	SurfNod	NSA	NSD	NSE	NSF	NSK
Sans structuration	0.632991	0.779608	0.53485	0.856394	0.856603	0.161509	1.99E-06	0.572451	0.462956
Faststructure	0.652129	0.75785	0.573153	0.836507	0.800021	0.113672	1.99E-06	0.382551	0.435016
Admixture	0.661364	0.787132	0.616842	0.861448	0.850856	0.0170032	1.99E-06	0.520237	0.451148
DAPC	0.63947	0.804824	0.585184	0.864767	0.864279	0.0986047	0.0267238	0.448423	0.447261
DAPC_MAF	0.64767	0.691125	0.556789	0.837201	0.876344	1.99E-06	1.99E-06	0.609276	0.449498

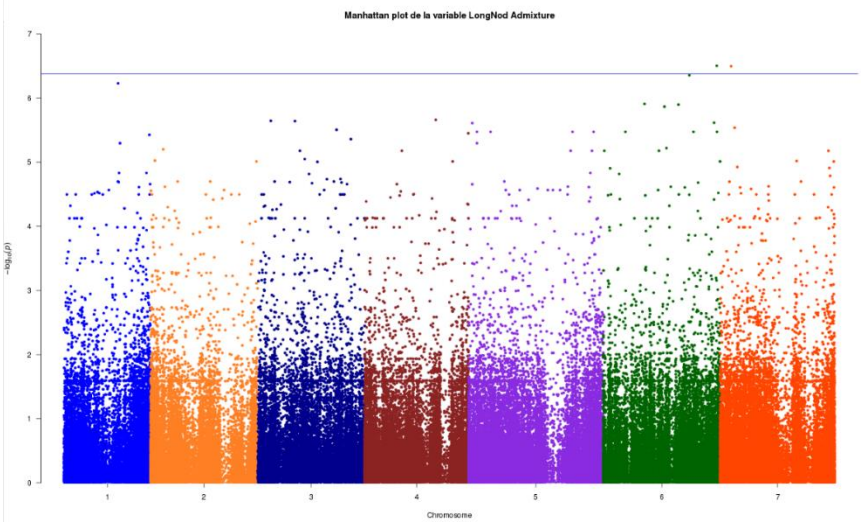
Annexe XIX : Comparaison de la significativité des SNP en prenant en compte la structuration dans le modèle mixte pour la variable LongNod



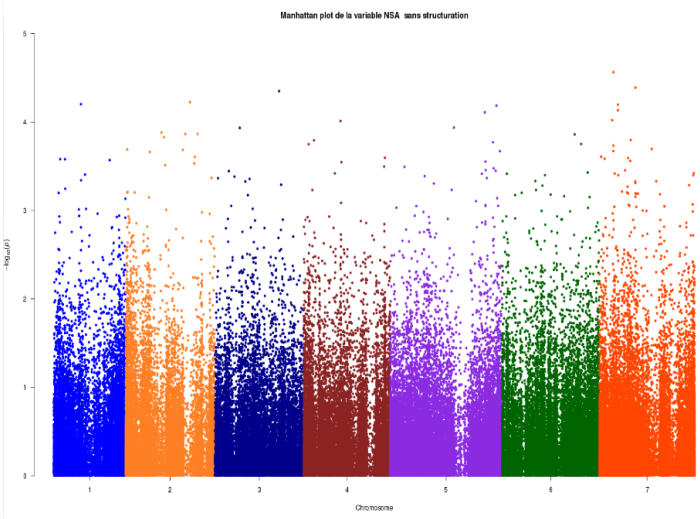


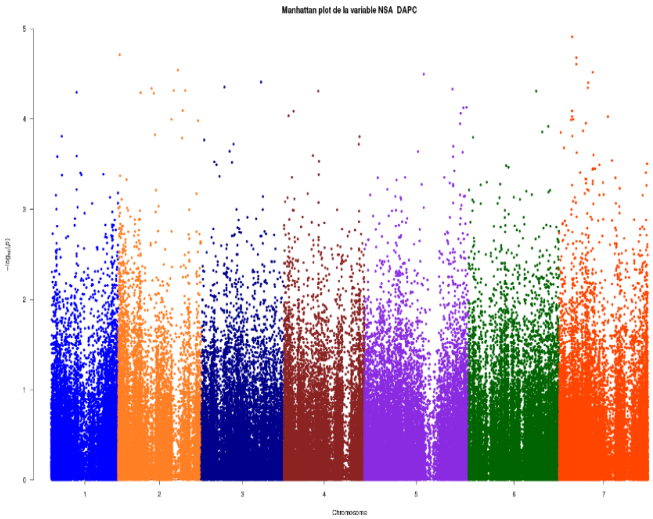


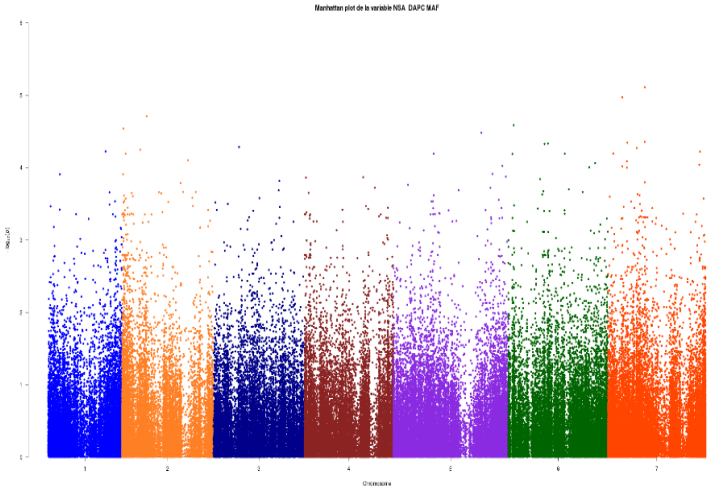


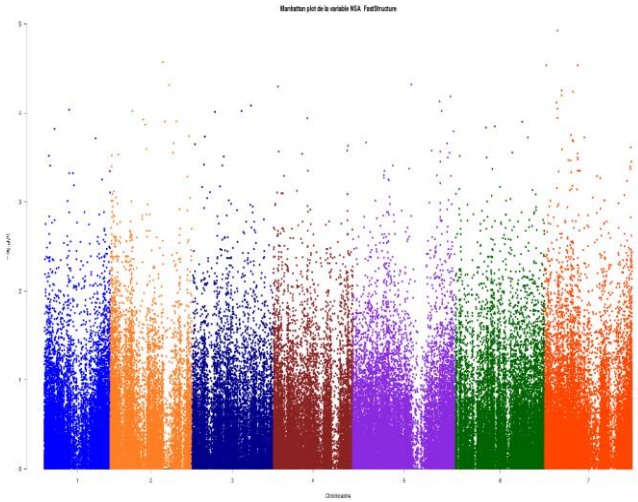


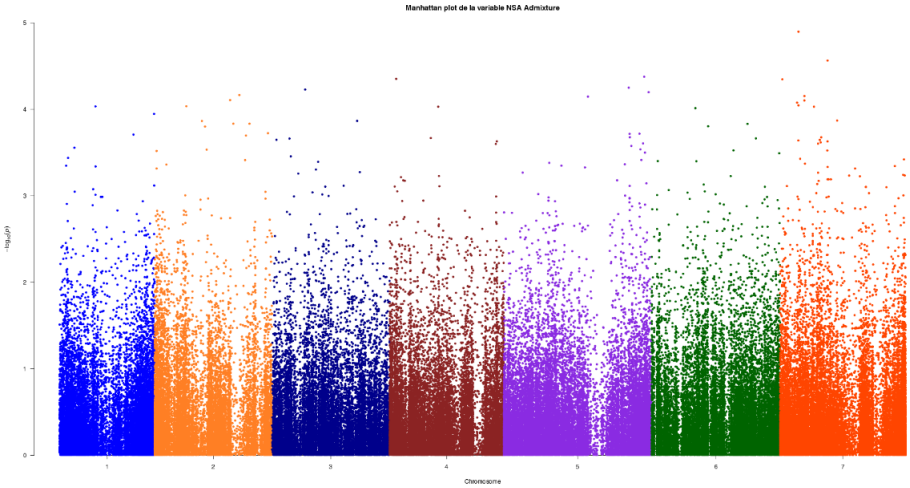
Annexe XX : Comparaison de la significativité des SNP en prenant en compte la structuration dans le modèle mixte pour la variable NSA



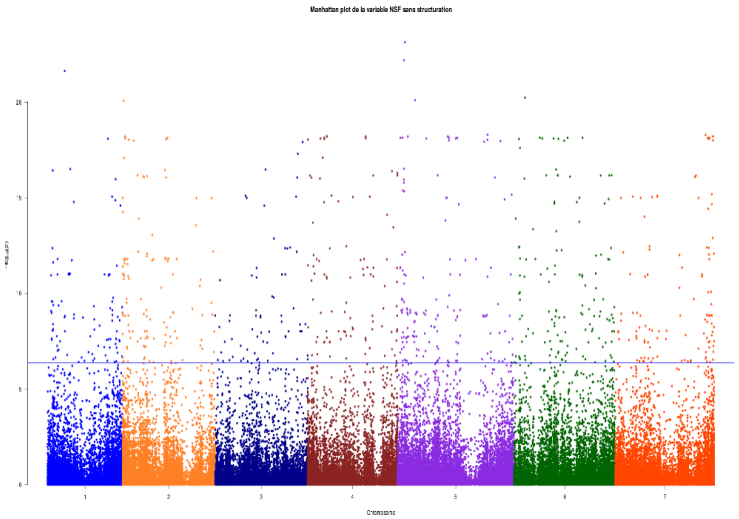


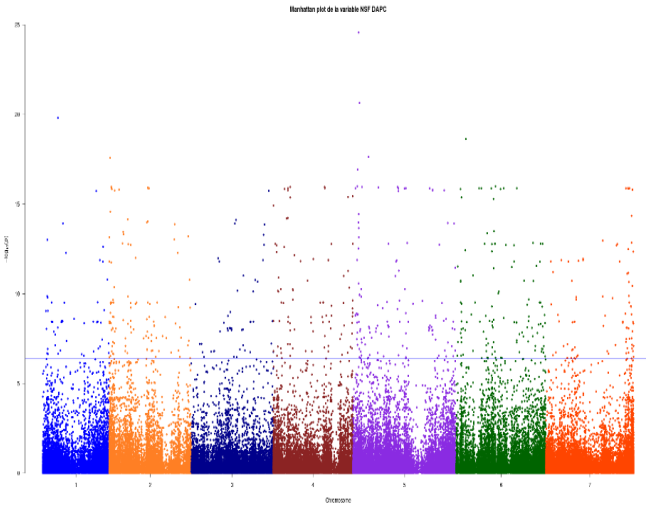


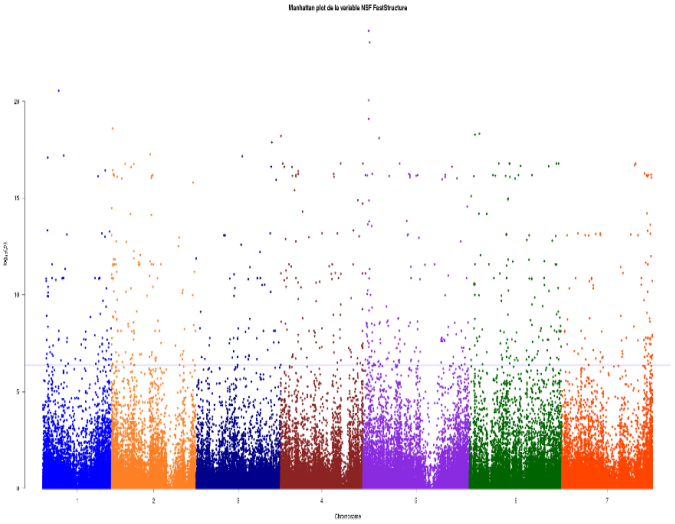


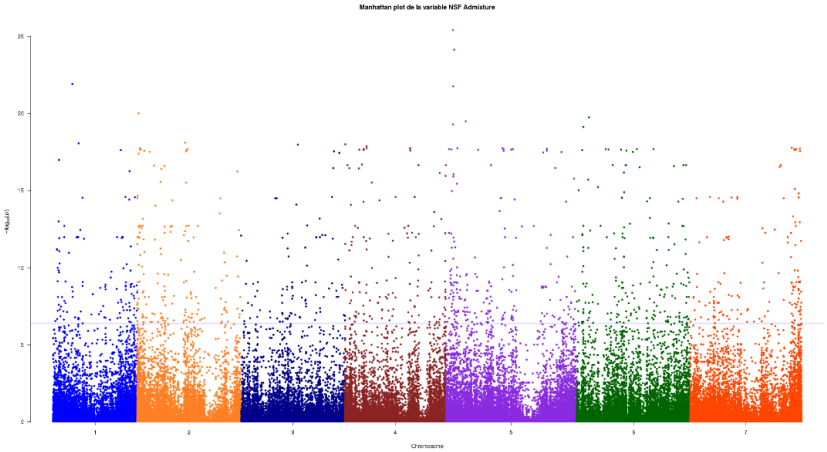


Annexe XXI : Comparaison de la significativité des SNP en prenant en compte la structuration dans le modèle mixte pour la variable NSF

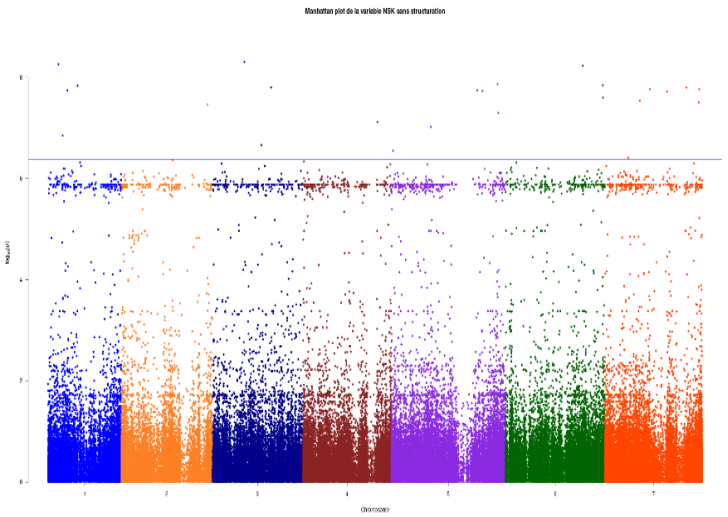


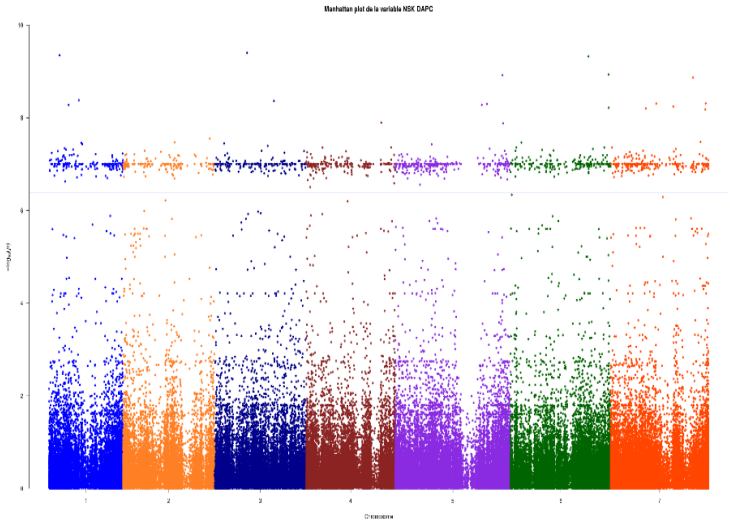


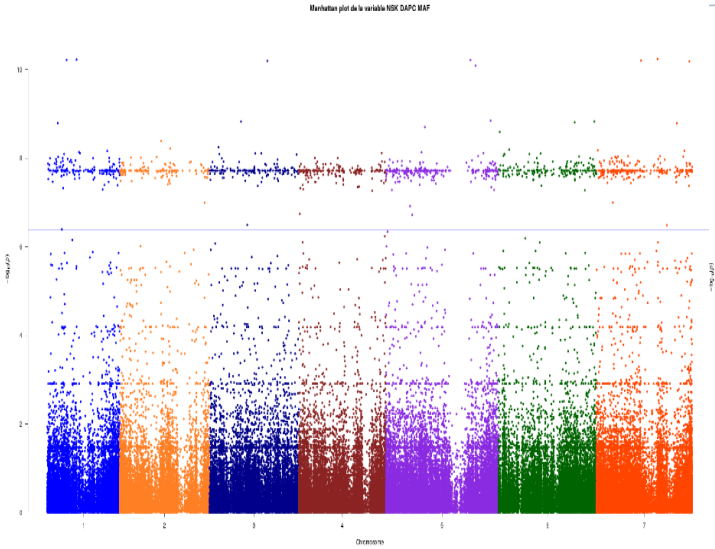


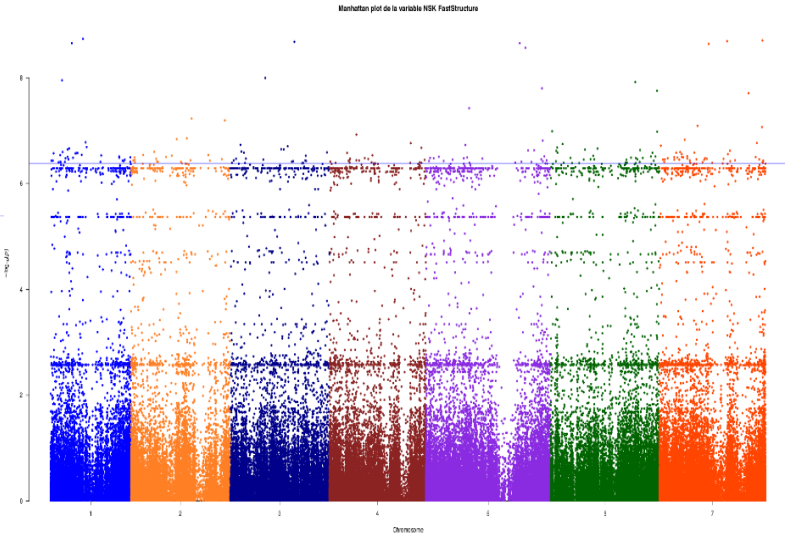


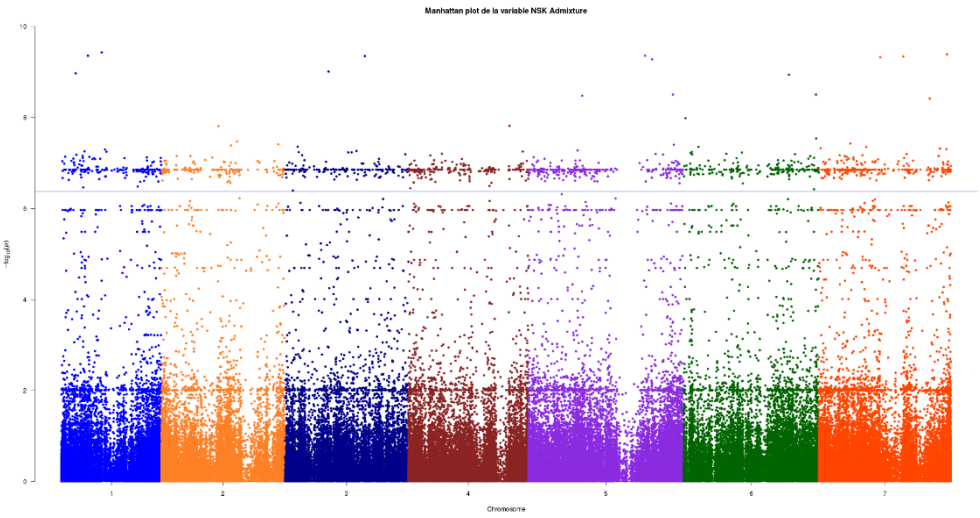
Annexe XXII : Comparaison de la significativité des SNP en prenant en compte la structuration dans le modèle mixte pour la variable NSK











Diplôme : Master de l'institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage

Spécialité : APVV (Amélioration, Production, Valorisation du végétal)

Spécialisation / option : GGAP (Génétique, Génomique et Amélioration des Plantes)

Enseignant référent : Mélanie Jubault

Auteur : Valentin Delefortrie

Date de naissance* : 10/06/1995

Organisme d'accueil : INRA de Dijon

Adresse : 17 Rue Sully, 21000 Dijon

Maître de stage : Virginie Bourion

Nb pages : 25 **Annexes** : 22

Année de soutenance : 2017-2018

RECHERCHE PAR GENETIQUE D'ASSOCIATION DE GENES IMPLIQUES DANS L'INTERACTION POIS x RHIZOBIUM

RESEARCH BY GENOME WIDE ASSOCIATION OF GENES INVOLVED IN PEA x RHIZOBIUM INTERACTION

Résumé :

L'azote est une ressource importante en agronomie ; son utilisation massive dans l'agronomie actuelle a un impact négatif sur l'environnement. Les légumineuses présentent le double intérêt de permettre une production de graines à haute teneur en protéines sans nécessité d'apport d'engrais azoté. La culture des légumineuses, y compris du pois, l'espèce la plus cultivée en France, reste cependant limitée, ceci principalement en raison de l'irrégularité de leur rendement. L'enjeu du projet SYMBIOPEA est d'améliorer la symbiose entre le pois et son partenaire symbiotique fixateur d'azote Rhizobium leguminosarum sv. viciae (Rlv). Ce projet se base sur des analyses de génétique d'association et a pour but de trouver des gènes ou QTL ayant un impact sur la nodulation et l'interaction entre pois et rhizobium par le biais d'une étude d'un panel diversifié d'accessions de pois inoculées par un mélange de souches Rlv. Pour chacune des variables étudiées, différents modèles mixtes basés chacun sur un type différent de structuration ont été testés. Toutes les structurations testées étaient cohérentes. Les choix pour chacune des souches Rlv sont des variables difficiles à analyser par génétique d'association car elles ont un profil plutôt qualitatif. En conséquence, un profil de choix parmi les souches Rlv a été établi à partir d'une analyse en composantes principales. Des SNP en déséquilibre de liaison avec des gènes putatifs d'intérêt ont été identifiés pour la croissance de la plante, la longueur de pivot porteuse de nodosités et la variable de profil de choix de souches.

Abstract :

Nitrogen is an important resource in agronomy; its massive use in the current agronomy has a negative impact on the environment. Legumes have the dual benefit of allowing the production of high-protein seeds without the need for nitrogen fertilizer. The cultivation of legumes, including pea, the most cultivated species in France, however, remains limited, mainly because of the irregularity of their yield. The challenge of the SYMBIOPEA project is to improve the symbiosis between pea and its symbiotic nitrogen-fixing partner Rhizobium leguminosarum sv. viciae (Rlv). This project is based on association genetic analyzes and aims to find genes or QTLs with an impact on nodulation and interaction between pea and rhizobium through a study of a diverse panel of pea accessions inoculated with a mixture of Rlv strains. For each of the variables studied, different mixed models each based on a different type of structuring were tested. All the structures tested were consistent. The choices for each of the Rlv strains are variables that are difficult to analyze by association genetics because they have a rather qualitative profile. As a result, a profile of choice among the Rlv strains was established from a principal component analysis. SNPs in linkage disequilibrium with putative genes of interest were identified for plant growth, the pivot length bearing nodules, and strain selection profile variable.

Mots-clés: Azote, Spécificité, Génétique Quantitative, Structuration, Modèle mixtes

Key Words: Nitrogen, Specificity, Quantitative genetic, Structuring, Mixed model