Analyse de la variation nucléotidique et structurale chez le soja par une approche de re-séquençage

Thèse

Davoud Torkamaneh

Doctorat en biologie végétale
Philosophiae doctor (Ph. D.)

Québec, Canada

Analyse de la variation nucléotidique et structurale chez le soja par une approche de re-séquençage

Thèse

Davoud Torkamaneh

Sous la direction de :

François Belzile, directeur de recherche

Richard Bélanger, codirecteur de recherche

Résumé

Le séquençage de nouvelle génération (NGS) a révolutionné la recherche chez les plantes et les animaux de plusieurs façons, y compris via le développement de nouvelles méthodes de génotypage à haut débit pour accélérer considérablement l'étude de la composition des génomes et de leurs fonctions. Dans le cadre du projet SoyaGen, financé par Génome Canada, nous cherchons à mieux comprendre la diversité génétique et l'architecture sous-jacente régissant les principaux caractères agronomiques chez le soja. Le soja est la plus importante culture oléagineuse au monde en termes économiques. Dans cette étude, nous avons cherché à exploiter les technologies NGS afin de contribuer à l'élucidation des caractéristiques génomiques du soja. Pour ce faire, trois axes de recherche ont formé le cœur de cette thèse : 1) le génotypage pan-génomique à faible coût, 2) la caractérisation exhaustive des variants génétiques par reséquençage complet et 3) l'identification de mutations à fort impact fonctionnel sur la base d'une forte sélection au sein des lignées élites.

Un premier défi en analyse génétique ou génomique est de rendre possible une caractérisation rapide et peu coûteuse d'un grand nombre de lignées à un très grand nombre de marqueurs répartis sur tout le génome. Le génotypage par séquençage (GBS) permet d'effectuer simultanément l'identification et le génotypage de plusieurs milliers de SNP à l'échelle du génome. Un des grands défis en analyse GBS est d'extraire, d'une montagne de données issues du séquençage, un grand catalogue de SNP de haute qualité et de minimiser l'impact des données manquantes. Dans une première étape, nous avons grandement amélioré le GBS en développant un nouveau pipeline d'analyse bio-informatique, Fast-GBS, conçu pour produire un appel de génotypes plus précis et plus rapide que les outils existants. De plus, nous avons optimisé des outils permettant d'effectuer l'imputation des données manquantes. Ainsi, nous avons pu obtenir un catalogue de 60K marqueurs SNP au sein d'une collection de 301 accessions qui se voulait représentative de la diversité du soja au Canada. Dans un second temps, toutes les données manquantes (~50%) ont été imputées avec un très grand degré d'exactitude (98 %). Cette caractérisation génétique a été réalisée pour un coût modique, soit moins de 15$ par lignée.

Deuxièmement, pour caractériser de manière exhaustive les variations nucléotidiques et structurelles (SNV et SV, respectivement) dans le génome du soja, nous avons séquencé le génome entier de 102 accessions de soja au Canada. Nous avons identifié près de 5M de variants nucléotidiques (SNP, MNP et Indels) avec un haut niveau d'exactitude (98,6 %). Ensuite, en utilisant une combinaison de trois approches différentes, nous avons détecté ~92K SV (délétions, insertions, inversions, duplications, CNV et translocations) et estimé que plus

de 90 % étaient exacts. C'est la première fois qu'une description complète de la diversité des haplotypes SNP et du SV a été réalisée chez une espèce cultivée.

Enfin, nous avons mis au point une approche analytique systématique pour faciliter grandement l'identification de gènes dont des allèles ont fait l'objet d'une très forte sélection au cours de la domestication et de la sélection. Cette approche repose sur deux progrès récents en génomique : 1) le séquençage de génomes entiers et 2) la prédiction des mutations entraînant une perte de fonction (LOF pour « loss of function »). En utilisant cette approche, nous avons identifié 130 gènes candidats liés à la domestication ou à la sélection chez le soja. Ce catalogue contient tous les gènes de domestication précédemment caractérisés chez le soja, ainsi que certains orthologues chez d'autres espèces cultivées. Cette liste de gènes fournit de nombreuses pistes d'investigation pour des études visant à mieux comprendre les gènes qui contribuent fortement à façonner le soja cultivé.

Cette thèse permet ultimement une meilleure compréhension des caractéristiques génomiques du soja. En outre, elle fournit plusieurs outils et références génomiques qui pourraient facilement être utilisés dans de futures recherches en génomique chez le soja de même que chez d'autres espèces.

Abstract

Next-generation sequencing (NGS) has revolutionized plants and animals research in many ways, including the development of new high-throughput genotyping methods to accelerate considerably the composition of genomes and their functions. As part of the SoyaGen project, funded by Genome Canada, we are seeking to better understand the genetic diversity and underlying architecture governing major agronomic traits in soybeans. Soybean is the world's largest oilseed crop in economic terms. In this study, we sought to exploit NGS technologies to help elucidate the genomic characteristics of soybeans. To this end, three main research topics have formed the core of this thesis: 1) low-cost genome-wide genotyping, 2) exhaustive characterization of genetic variants by whole-genome resequencing, and 3) identification of mutations with high functional impact on the basis of a strong selection within the elite lines.

A first challenge in genetic or genomic analysis is to make possible a rapid and inexpensive characterization of a large number of lines with a very large number of markers distributed throughout the genome. Genotyping-by-sequencing (GBS) allows simultaneous identification and genotyping of several thousand SNPs on a genome-wide scale. One of the major challenges in GBS analysis is to extract a large catalog of high quality SNP from a mountain of sequencing data and minimize the impact of missing data. As a first step, we have greatly improved the GBS by developing a new bio-informatics analysis pipeline, Fast-GBS, designed to produce a more accurate and faster call of genotypes than existing tools. In addition, we have optimized tools for imputing missing data. For example, we were able to obtain a catalog of 60K SNP markers from a collection of 301 accessions that were representative of soybean diversity in Canada. Second, all missing data (~ 50%) were imputed with a very high degree of accuracy (98%). This genetic characterization was performed at a low cost, less than $ 15 per line.

Second, to fully characterize the nucleotide and structural variations (SNV and SV, respectively) in the soybean genome, we sequenced the whole genome of 102 Canadian soybean accessions. We have identified nearly 5M of nucleotide variants (SNP, MNP and Indels) with a high level of accuracy (98.6%). Then, using a combination of three different approaches, we detected ~ 92K SV (deletions, insertions, inversions, duplications, CNVs and translocations) and estimated that more than 90% were accurate. This is the first time that a complete description of the diversity of SNP and SV haplotypes has been carried out in a cultivated species.

Finally, we have developed a systematic analytical approach to greatly facilitate the identification of genes whose alleles have undergone a very strong selection during domestication and selection. This approach is based on two recent advances in genomics: (1) whole-genome sequencing and (2) predicting mutations resulting in loss of function (LOF). Using this approach, we identified 130 candidate genes related to domestication or selection in soybean. This catalogue contains all of the previously well-characterized domestication genes in soybean, as well as some orthologues from other domesticated crop species. This list of genes provides many avenues of investigation for studies aimed at better understanding the genes that contribute strongly to shaping cultivated soybeans.

This thesis ultimately leads to a better understanding of the genomic characteristics of soybeans. In addition, it provides several tools and genomic resources that could easily be used in future genomic research in soybeans as well as in other species.

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations et des sigles

NGS	Next-generation sequencing
WGS	Whole-genome sequencing
GBS	Genotyping-by-sequencing
RAD-seq	Restriction site associated DNA sequencing
SNP	Single nucleotide polymorphism
MNP	Multiple nucleotide polymorphism
Indel	Insertion-deletion
SV	Structural variants
LOF	Loss-of-function
LD	Linkage disequilibrium
MaxMD	Maximum missing data
MinMAF	Minimum minor allele frequency
GWAS	Genome-wide association study

Remerciements

Le succès de mon projet de doctorat est l'aboutissement d'innombrables réflexions générées et alimentées par plusieurs personnes qui m'entourent : ma famille, mes amis, collègues et mentors. Je tiens à adresser mes remerciements les plus sincères à tous ces êtres humains qui m'ont aidé et supporté.

Je tiens d'abord à remercier mon directeur de thèse, François Belzile. François, merci pour ta générosité, merci pour ton ouverture d'esprit, merci pour ton côté humain, merci pour ta compréhension, merci de m'avoir fait grandir, dans tous le sens du terme, merci de m'avoir apprendre à faire de la science utile à la communauté, juste, transparente et rigoureuse.

François, tu m'as enseigné la théorie et la philosophie du métier de scientifique. Tu as été l'un des meilleurs modèles que j'aurai pu avoir, merci tellement pour ton support, tes encouragements, nos discussions et débats sur pas juste la science, mais la société en général aussi.

François, tu crois sincèrement en tes étudiants et tu prends toujours le temps de les écouter et de considérer leurs opinions et leurs idées. À mon avis, c'est une qualité remarquable, je te lève mon chapeau.

François, merci d'avoir passé dans ma vie.

Je tiens également à remercier mon codirecteur de thèse, Richard Belanger. Richard, tu n'étais pas obligé d'accepter la codirection de mon projet, mais tu l'as fait. J'ai toujours grandement apprécié ton sens de l'humour.

Je tiens ensuite à remercier très chaleureusement Martine Jean. Je m'estime chanceux d'avoir pu bénéficier de tes réflexions (parfois très pessimistes), tes conseils et tes idées toujours à point et d'une justesse désarmante.

Je voudrais également remercier Jérôme Laroche et Brian Boyle. Jérôme, j'ai toujours grandement apprécié ton aide et support bio-informatique. Brian, j'ai eu des discussions extrêmement formatrices avec toi au cours de nos échanges pendant ces quelques années. Votre travail et votre aide sont plus qu'appréciés.

Merci également aux membres du jury, François et Richard bien sûr, ainsi que Roger C. Levesque, Rupesh Deshmukh et Scott Jackson qui devront lire et commenter les nombreuses pages de ce travail. Votre travail et votre aide sont plus qu'appréciés.

Merci aussi à tous mes collaborateurs et co-auteurs qui ont rendu possible la publication des travaux qui seront présentés dans les prochaines pages.

Merci aux membres du laboratoire François Belzile. J'ai toujours eu énormément de plaisir avec vous tous. J'ai plein de souvenirs inoubliables avec vous tous.

Je tiens à formuler un dernier remerciement à ma famille : mon amour (Leila), maman, mes frères et ma sœur. Papa, je souhaite que tu sois ici avec moi! Leila, t'as toujours été là pour moi, dans n'importe quelle situation. Je t'aime si fort!

Avant-propos

Cette thèse est organisée en huit chapitres. Le premier chapitre consiste en une introduction générale, alors que le second chapitre présente l'état des connaissances d'une manière plus spécifique sur le sujet abordé par la thèse et expose les différentes méthodes de génotypage reposant sur les technologies de séquençage. Ce chapitre de ma thèse (Chapitre II) a été rédigé en vue d'être publié sous forme d'article de revue. Les chapitres III, IV et V ont été publiés dans des revues scientifiques et le chapitre VI a été accepté. Le chapitre VII a été soumis dans une revue scientifique. Finalement, le chapitre IX est un chapitre de conclusions générales et réflexions personnelles. Voici l'état de ces publications :

Le chapitre III est publié sous la référence : Torkamaneh D., Laroche J., Bastien M., Abed A., Belzile F. (2017a). Fast-GBS: a new pipeline for the efficient and highly accurate calling of SNPs from genotyping-by-sequencing data. *BMC Bioinformatics*, 18:5

DT a développé l'idée du projet. FB a supervisé le projet. AA et MB ont fourni les données et analyses pour l'orge et la pomme de terre. DT et JL ont contribué à la programmation. DT et FB ont écrit le manuscrit.

Le chapitre IV est publié sous la référence : Torkamaneh D., Laroche J., Belzile F. (2016). Genome-Wide SNP Calling from Genotyping by Sequencing (GBS) Data: A Comparison of Seven Pipelines and Two Sequencing Technologies. *PLoS ONE,* 11(8): e0161333.

DT a développé l'idée du projet. FB a supervisé le projet. DT et JL ont contribué à la programmation et à l'analyse des données. DT et FB ont écrit le manuscrit.

Le chapitre V est publié sous la référence : Torkamaneh, D., and Belzile, F. (2015). Scanning and Filling: Ultra-Dense SNP Genotyping Combining Genotyping-By-Sequencing, SNP Array and Whole-Genome Resequencing Data. *PLoS ONE*, 10(7): e0131533.

DT et FB ont développé l'idée du projet. FB a supervisé le projet. DT a réalisé le travail de laboratoire, produit et analysé les données. DT et FB ont écrit le manuscrit.

Le chapitre VI est accepté sous la référence : Torkamaneh, D., Laroche, J., Tardivel, A., O'Donoughue, L., Cober, E., Rajcan, I., Belzile, F. 2017b. Comprehensive Description of Genome-Wide Nucleotide and Structural Variation in Short-Season Soybean. *Plant Biotechnology Journal*.

DT et FB ont développé l'idée du projet. DT et JL ont contribué à la programmation et à l'analyse des données. LO, EC et IR ont contribué à la sélection des échantillons. AT a évalué les variants structuraux par PCR. DT et FB ont écrit le manuscrit.

Chapitre I
Introduction générale

I.1 Séquençage de nouvelle génération en phytogénétique

Des avancées technologiques ont rendu possible le séquençage de nouvelle génération (NGS), lequel ouvre la voie à une caractérisation rapide et exhaustive du génome des plantes. Le NGS a révolutionné la recherche sur les plantes et les animaux de plusieurs façons. Tout d'abord, il a permis aux chercheurs de décoder le génome entier de nombreux organismes. Actuellement, les génomes de centaines d'eucaryotes ont été séquencés et, pour certaines espèces, de nombreux individus, cultivars ou accessions d'une même espèce ont également été séquencés. L'avènement des technologies de séquençage NGS a fourni une occasion exceptionnelle de détecter systématiquement les variations génétiques chez les plantes (El-Metwally et al. 2014; Hall, 2007). Les variations génétiques constituent la matière première de l'évolution, car certaines d'entre elles améliorent l'adaptabilité et la survie d'une population face aux conditions environnementales changeantes et à d'autres circonstances imprévues (Hedrick, 2011; Dobzhansky, 1970). La variation génétique peut être divisée en deux grandes catégories: les variations nucléotidiques et structurelles. Les variants nucléotidiques sont généralement définis comme englobant des variants de nucléotides simples ou multiples (SNP, MNP) et de petites insertions/délétions (indels), tandis que les variants structurels (SV) représentent des réarrangements plus importants de différents types [délétions, insertions, inversions, translocations, duplications et variations du nombre de copies (CNV)]. Dans ce travail (Chapitre VI), nous avons séquencé le génome entier de 102 lignées canadiennes de soja pour décrire de manière exhaustive les variations génétiques qui existent au sein de ce matériel.

Le NGS a également facilité grandement le développement de méthodes de génotypage à haut débit pour détecter un très grand nombre de marqueurs moléculaires tels que les SNP (« single nucleotide polymorphism »). Dans une telle approche, le séquençage à grande échelle a permis aux chercheurs d'explorer la diversité des nucléotides dans des collections d'individus pour découvrir des sites polymorphes, puis développer des puces de génotypage (« SNP chips »). Ces puces de génotypage peuvent être utilisées pour déterminer le génotype d'une ligne individuelle à des milliers jusqu'à des millions de sites SNP (Ha et al. 2014). À ce jour, des puces de génotypage avec plus de 40K SNP ont été développées et utilisées pour diverses applications en génétique et dans des programmes d'amélioration génétique chez plusieurs cultures comme le riz, le maïs, le tournesol, le soja, l'avoine, le coton et le blé (Rasheed et al. 2017). D'un autre côté, des méthodes de génotypage qui exploitent la puissance des technologies NGS ont également été développées pour rendre possible l'identification et le génotypage simultané de milliers à des millions de sites SNP. Le génotypage par séquençage (GBS) est un exemple d'une telle approche de génotypage SNP

qui s'appuie sur le NGS (Elshire et al. 2011). Cette méthode a été grandement utilisée chez des plantes.

I.2 Génotypage par séquençage (GBS)

Chez les plantes, le GBS a été développé comme une approche rapide et robuste pour le séquençage partiel du génome (« reduced-representation sequencing ») qui permet la découverte et le génotypage de marqueurs moléculaires à l'échelle du génome chez un grand nombre de lignées (Davey et al. 2011). Le GBS est une approche hautement flexible avec un faible coût qui en fait un excellent outil pour de nombreuses applications et questions de recherche en génétique et en élevage. De telles avancées modernes permettent le génotypage de milliers de SNP et, ce faisant, augmente la probabilité d'identifier des SNP associés avec des caractères d'intérêt. Cependant, lors de l'utilisation d'approches telles que le GBS, laquelle repose sur l'examen d'une fraction du génome, certains défis sont rencontrés. En général deux problématiques majeures doivent être surmontés : 1) comment transformer une masse d'informations de séquence en un catalogue de marqueurs SNP ; et 2) comment surmonter le problème posé par les données manquantes.

I.2.1 Analyse bio-informatique des données GBS

Le principal défi du GBS, pour la plupart des utilisateurs, est l'analyse bio-informatique de la grande quantité d'informations de séquence dérivées du séquençage des librairies GBS en vue d'appeler les allèles chez les locus SNP (Davey et al. 2011). Il est clair que les pipelines bio-informatiques d'analyse sont devenus une nécessité pour filtrer, trier et aligner ces données de séquence. Un pipeline pour le GBS doit inclure des étapes pour filtrer et retirer les lectures de mauvaise qualité, classer les lectures par pool ou les individus en fonction des séquences des code-barres, identifier les locus et les allèles *de novo* ou aligner les lectures sur un génome de référence pour découvrir des polymorphismes et attribuer le génotype à chaque locus pour chaque individu (Glaubitz et al. 2014). Pour répondre à ces besoins, de nombreux pipelines de bio-informatiques ont été développés. Dans ce travail (Chapitre III), nous décrivons un nouveau pipeline, Fast-GBS, qui utilise le génome de référence. Il est facile à utiliser avec différentes espèces, dans différents contextes, et fournit une plate-forme d'analyse qui peut être exécutée avec différents types de données de séquençage et avec des ressources de calcul modestes. Nous avons évalué Fast-GBS en fonction d'une analyse à grande échelle chez trois espèces. Nous avons aussi comparé de manière exhaustive les principaux pipelines d'analyse GBS existants en fonction du nombre de SNP appelés, de la précision des génotypes résultants ainsi que de la rapidité et de la facilité d'utilisation de ces

pipelines. Nous avons également comparé les résultats obtenus à l'aide des lectures Illumina et Ion Torrent (Chapitre IV).

I.2.2 L'imputation de données manquants généré par GBS

Comme décrit, le GBS est une approche balayage ou d'échantillonnage du génome basé sur l'enzyme de restriction utilisée (Elshire et al. 2011). Ainsi, lors l'analyse de données GBS, une quantité considérable de données manquantes peut être rencontrée. Une question importante qui restait sans réponse à ce stade est la mesure dans laquelle les données manquantes peuvent être tolérées (Jarquín et al. 2014). Et aussi dans quelle mesure elles affectent la précision du processus d'imputation. En général, il existe deux types de données manquantes dans de grands ensembles de données. Le plus évident est lorsque nous ignorons le génotype de certains individus à un locus qui a été génotypé avec succès chez les autres individus dans une population. Dans une autre situation, qui se pose lorsque différents ensembles de données (par exemple, obtenus à l'aide de différentes technologies de génotypage) sont combinés, il peut y avoir des locus qui ne sont pas génotypés au sein de la population, c'est-à-dire qu'il n'y a pas d'information pour un locus SNP chez tous les individus de la population, sauf pour quelques individus qui peuvent être communs aux deux ensembles de données. On appelle ce premier cas de figure une « donnée manquante » tandis que le second est appelé « génotype manquant » (Hao et al. 2009). Il y a eu un intérêt considérable à imputer ces données manquantes en fonction des données disponibles. Beaucoup d'outils utilisés dans l'analyse génétique nécessitent des jeux de données complets et il existe donc deux possibilités: ne retenir que des locus SNP dépourvus de données manquantes (ce qui réduit considérablement le nombre de SNP disponibles en GBS) ou imputer ces données manquantes à travers diverses stratégies.

Dans ce travail (Chapitre V), nous avons exploré la précision et l'efficacité de différents outils d'imputation à la fois pour l'imputation des données manquantes dans le contexte du GBS et des génotypes manquant dans le contexte de la combinaison des ensembles de données SNP obtenues au moyen de différentes approches de génotypage (GBS, puce et re-séquençage). Enfin, nous avons examiné l'impact de l'utilisation de ces ensembles de données SNP améliorés dans les analyses d'association.

I.3 Pourquoi le soja?

Le soja (*Glycine max* L. Merr.), est une légumineuse annuelle de la famille des légumineuses (Fabaceae) avec des graines comestibles (Lam et al. 2010). Le soja est économiquement la légumineuse la plus importante au monde, fournissant des protéines végétales à des millions

de personnes et des ingrédients pour des centaines de produits chimiques (Mian, 2006). De nombreux botanistes pensent qu'il a été domestiqué pour la première fois en Chine centrale en 7000 avant l'Ère Commune. Le soja est utilisé en Chine, au Japon et en Corée depuis des milliers d'années comme un aliment et un composant de médicaments (Mian, 2006). Le soja a été introduit aux Canada au 19ème siècle et est devenu particulièrement important. Au niveau mondial, les États-Unis, le Brésil et l'Argentine sont les trois plus grands producteurs de soja (FAOSTAT).

I.3.1 Le génome du soja

Le génome du soja (1,1 gigabase) a été séquencé en 2010 (Schmutz et al. 2010). Plus de 50K gènes codant pour des protéines ont été prédits, soit 70% de plus que la plante modèle Arabidopsis, aussi une dicotylédone comme le soja. Le génome du soja porte les traces d'une polyploïdie ancienne (paléopolyploïde) et il aurait subi deux duplications complètes. Les duplications du génome se seraient produites il y a environ 59 et 13 millions d'années, ce qui a donné lieu à un génome hautement dupliqué avec près de 75 % des gènes présents en plus d'une copie. Les deux événements de duplication ont été suivis d'une diversification, d'une perte de gènes et de nombreux réarrangements chromosomiques (Schmutz et al. 2010).

I.3.2 Domestication du soja

Au cours des 12 000 dernières années, les humains ont domestiqué des centaines d'espèces végétales et animales pour plusieurs fins: surtout en tant que source d'aliments et de matériaux (p. ex. textiles et peaux) ou encore en tant qu'espèces compagnes ou pour leur valeur esthétique (Zeder 2015). Il est largement admis que le soja cultivé moderne a été domestiqué du soja sauvage (*Glycine soja* Sieb & Zucc.) en Asie de l'Est il y a 6000-9000 ans (Carter et al. 2004; Kim et al. 2012b.). La dissection de l'architecture génétique des caractères de domestication chez les plantes cultivées et la nature de la sélection ont été un sujet d'étude en génétique moléculaire au cours des deux dernières décennies. Récemment, Sedivy et col. (2017), en utilisant des données de séquençage des génomes entiers ont montré que le soja cultivé provenait de multiples événements de domestication, mais l'identification des gènes liés à la domestication reste un travail très difficile et laborieux. Dans ce travail (Chapitre VII), on présente une nouvelle approche basée sur les données de re-séquençage et les mutations de perte de fonction (LOF) pour la détection efficace et hautement précise de gènes candidats liés à la domestication.

I.8 Objectifs spécifiques de la thèse

Tel que présenté et décrit, les technologies de séquençage de nouvelle génération (NGS) ont révolutionné la recherche sur les plantes. L'objectif global de cette thèse était de développer de nouveaux outils génomiques pour permettre d'étudier le génome du soja et de faciliter l'utilisation de ces outils pour les sélectionneurs. En s'appuyant sur une grande culture (le soja), cette thèse visait plus spécifiquement à : 1) permettre l'amélioration de la plate-forme de génotypage en développant de nouveaux outils et approches analytiques, 2) réaliser un séquençage complet du génome entier d'une collection de lignées représentatives de la diversité génétique du soja cultivé au Canada en vue d'identifier et de génotyper des variations génétiques, 3) d'utiliser ces données génétiques pour identifier de manière rapide des gènes ayant joué un rôle important dans la domestication et l'adaptation du soja.

Le prochain chapitre (Chapitre II) constitue une revue de littérature (rédigée sous forme de manuscrit) visant à décrire plus en détail les différentes stratégies et méthodes de génotypage fondées sur les technologies NGS, lesquelles permettant de génotyper un très grand nombre de marqueurs moléculaires chez un grand nombre d'individus. Ce chapitre décrit les principaux défis existants dans les approches génotypage actuel qui ont été mises au point dans la présente thèse. Les chapitres III, IV et V sont trois chapitres très proches sur le plan thématique, car ils relatent nos travaux visant à améliorer la méthode de génotypage par séquençage (GBS). Le chapitre III présente le point de départ de notre développement et amélioration du GBS. On y décrit un nouveau pipeline d'analyse bio-informatique des données GBS, appelé Fast-GBS, pour l'appel efficace et très précis des SNP à partir des données de GBS. Ensuite, dans le chapitre IV, on rapporte le fruit d'une analyse comparée des principaux pipelines analytiques en usage. Enfin, dans le chapitre V, on présente une approche d'imputation des données manquantes chez le GBS, laquelle permet de maximiser les données génotypiques tirées du GBS. Au chapitre VI, nous avons généré un catalogue exhaustif de la variation génétique, tant nucléotidique que structurale, rencontrée au sein du génome du soja cultivé au Canada. Dans le cas des variations structurales, il s'agit de la première description complète de ce type de variation chez une plante cultivée. Jusqu'alors, les analyses de la variation génétique ont été largement concentrées sur les variations nucléotidiques. Dans le chapitre VII, nous avons développé une approche analytique systématique visant à identifier des gènes liés à la domestication chez le soja. On y décrit comment, à partir de nombreuses données génomiques chez les espèces cultivées et leurs ancêtres sauvages, on peut identifier des gènes qui sont fixés pour des allèles distincts chez ces deux groupes de plantes. Finalement, dans un ultime chapitre, nous tentons de présenter une vue d'ensemble des contributions apportées par ces travaux à l'état des

connaissances, mais surtout d'en décrire les retombées dans tous les domaines de la phytogénétique.

Chapitre II

NGS-based genome-wide genetic marker discovery and genotyping methods in crop plants

Davoud Torkamneh[1,2] and François Belzile[1,2]

[1]Département de Phytologie, Université Laval, Québec City, QC, Canada
[2]Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Quebec City, QC, Canada

II.1 Résumé

Les technologies de séquençage de la nouvelle génération (NGS) fournissent des méthodes de génotypage puissantes et flexible aux sélectionneurs et aux chercheurs. Ces méthodes offrent une large gamme d'applications allant de l'analyse pan-génomique au dépistage de routine avec un haut niveau de précision et de reproductibilité. En outre, ils fournissent un flux de travail direct pour identifier, valider et afficher des variants génétiques en peu de temps avec un faible coût. Ici, on passe en revue et aussi nous discutons les avantages et les défis de plusieurs méthodes NGS pour le développement de marqueurs génétiques à l'échelle du génome et le génotypage chez les plantes cultivées. Ces méthodes comprennent le ré-séquençage du génome entier, la puce de génotypage et le génotypage par séquençage, qui sont largement appliqués chez les plantes. Nous discutons également les méthodes d'imputation qui peuvent être utilisées pour remplacer les données manquantes dans les ensembles de données génotypiques et aussi pour intégrer les ensembles de données obtenus à l'aide de différents outils de génotypage. Nous espérons que cette vision synthétique des méthodes de génotypage aidera les généticiens et les sélectionneurs à intégrer ces méthodes qui sont basées sur le NGS dans les programmes d'amélioration génétique et la recherche sur les plantes cultivées.

II.2 Abstract

Next-generation sequencing technologies provide powerful and flexible genotyping methods to plant breeders and researchers. These methods offer a wide range of applications from genome-wide analysis to routine screening with a high level of accuracy and reproducibility. Furthermore, they provide a straightforward workflow to identify, validate, and screen genetic variants in a short time with a low cost. Here we review and discuss the advantages and challenges of several NGS methods for genome-wide genetic marker development and genotyping in crop plants. These methods include whole-genome re-sequencing, SNP arrays and genotyping-by-sequencing, which are widely applied in crops. We also discuss how imputation methods can be used to both fill in missing data in genotypic datasets and to integrate datasets obtained using different genotyping tools. It is our hope that this synthetic view of genotyping methods will help geneticists and breeders integrate these NGS-based methods in crop plant breeding and research.

II.3 Introduction

Since the Green Revolution in the 1960s (Swaminathan, 2009), plant breeding efforts have been supported and facilitated by new technologies and approaches. The genomic tools and resources that facilitate genotype-phenotype studies (Pérez-de-Castro et al. 2012), in particular for complex traits, are leading to a second revolution. Next-generation sequencing (NGS) technologies (Metzker, 2010), known as high-throughput parallel (HTP) DNA and RNA sequencing technologies, have revolutionized plant research in many ways (Figure II.1). Firstly, deep sequencing and *de novo* assembly have enabled the decoding of the entire genome in many plant species (>100 plant species, to date) (NCBI, "www.ncbi.nlm.nih.gov/projects/WGS/WGSprojectlist.cgi"). In addition to providing valuable insights into crop genome organization and evolution, such reference genomes represent a foundational resource for transcriptome analysis, sequence mapping and genetic marker development (Church, 2006). Secondly, NGS has also allowed to quickly and exhaustively assess genetic diversity at the intraspecific level thanks to low- to mid-depth sequencing (whole-genome re-sequencing (WGR)) of the entire genome of numerous cultivars or accessions of the same species. WGR provides the most comprehensive approach for genome-wide discovery of genetic variants (nucleotide and structural variants (NVs and SVs, respectively)) (Goodwin et al. 2016). Finally, NGS has enabled researchers to develop cost-effective high-throughput genotyping methods such as genotyping-by-sequencing (GBS) (Davey et al. 2011).

In crop genetics and breeding, genotyping obviously plays a critical role in both the identification of genomic regions controlling traits of interest (genes or QTLs) but also in marker-assisted selection (MAS) used to expedite the development of advanced lines with the desired traits (Varshney et al. 2014). As highlighted above, NGS has made major contributions to genotyping through the discovery of polymorphic sites in a genome. These polymorphic sites can then serve to develop genotyping arrays ("SNP chips") that allow one to interrogate these genomic positions in a high-throughput fashion (Ha et al. 2014). In a third step (after variant discovery and array design), such SNP chips can be used to characterize the genotype of an individual at thousands to millions of SNPs (Kumar et al. 2012). Alternatively, NGS technologies, coupled with complexity-reduction methods, have been used to simultaneously identify large numbers of variant positions and determine an individual's genotype at these SNPs. In plants, such approaches, in particular genotyping-by-sequencing (GBS), have been the tool of choice to discover and type SNPs using NGS (Deschamps et al. 2012). GBS is a particularly attractive complexity reduction method that offers a simple, robust, low-cost, and high-throughput method for genotyping in both model

and non-model species (Elshire et al. 2011). Despite the tremendous opportunities brought about by NGS in crop genetics and breeding, these technologies bring new challenges. The typically very large raw datasets (e.g. billions of sequence reads) are devoid of value on their own and only become intelligible and useful once they have been subjected to some form of bioinformatics analysis, sometime within a very short time (Nielsen et al. 2011). The efficient and accurate computational processing, variant and genotype calling of large-scale NGS data requires the development of new bioinformatics tools (algorithms, software and pipelines) (Pirooznia et al. 2014). Given the large scale of such data, there is a natural, yet dangerous, tendency to trust the outcomes of these analyses. In our view, this is one of the great dangers of this revolution: the insufficient critical assessment of the reliability of the resulting processed data.

In this review, we present and discuss the most relevant advances in genotyping methods for crop plants. We introduce the most widely-used genotyping approaches and illustrate their potential contributions to HTP genotyping in several crop plants. Furthermore, we discuss the limitations and challenges of each method along with proposed solutions. The objective is to provide geneticists and breeders with an updated synthetic view of the NGS-based genotyping tools available for the improvement of the efficiency of crop breeding programs.

II.4 Identification of genetic variants through WGR

For the most part, WGR experiments have been conducted to comprehensively identify the differences between the genomes of individual samples of interest and a reference genome (Li et al. 2009). Several recent reviews have comprehensively discussed the bioinformatics analytical tools and pipelines that have been developed for discovery and genotyping of genetic variants through WGR (Hwang et al. 2015). The genetic variants provide an extremely valuable insight into the genetic background of the individuals (Hedrick, 2011). Generally, genetic variants are divided in two main categories, nucleotide variants (NVs) and structural variants (SVs). In the following sections, we introduce and discuss these two categories of genetic variants with several examples in crop plants.

II.4.1 Nucleotide variants (NVs)

Nucleotide variants (NVs) reflect variation in a single or multiple neighboring nucleotides (SNVs and MNVs) that occurs at these positions in the genome. Small insertions or deletions (InDels), generally smaller than 50 bp, are also typically called by tools designed to call SNVs and MNVs. NVs may arise within the coding or non-coding regions of genes, or in the intergenic regions. In most crop species, as the portion of the genome that codes for a protein

is a small part of the whole, NVs occur more frequently in non-coding than in coding regions (Varela and Amos, 2010). The subset of NVs located within coding regions may or may not change the amino acid sequence, due to the degeneracy of the genetic code (Karki et al. 2015). Up to date, several large WGR projects have been conducted in Arabidopsis, maize, rice, soybean and tomato. For example, the WGR of 2,029 *A. thaliana* accessions unveiled 11M biallelic SNVs and 1.4M InDels (up to 40 bp). A genome-wide association analysis based on this extensive dataset allowed researchers to gain insight on the evolution of Arabidopsis from the glacial age to modern (The 1001 Genomes Consortium, 2016). In 2013, a rice WGR project was conducted on 3,000 accessions to create a public rice genetic/genomic database for global rice community. The 18.9M NVs derived from this project showed that the *O. sativa* gene pool is differentiated into five varietal groups – *indica*, aus/boro, basmati/sadri, tropical *japonica* and temperate *japonica* (The 3,000 rice genomes project, 2014). The WGR of 302 wild and cultivated soybean (*G. max* and *G. soja*) accessions revealed 10M SNVs and 1M InDels. This dataset has allowed researchers to detect more than two hundred domestication and improvement sweeps in soybean genome (Zhou et al. 2015). In addition to genetic diversity studies based on WGR data, several WGR projects performed to gain insights into the genetic architecture of agronomic traits in crop plants. A genome-wide association study (GWAS) for 14 agronomic traits in 517 accessions of *O. sativa indica* using 3.6M SNVs derived from WGR allowed to detect 80 strong genotype-phenotype associations (Huang et al. 2010). Genomic analyses based on the WGR of 360 tomato accessions provided insights into the history of tomato breeding. It showed that 18 QTLs related to fruit mass in tomato are located within domestication and improvement sweeps (Lin et al. 2014). As exemplified above, WGR studies generate the most comprehensive catalogues of NVs that provide key genetic insights into complex traits in crop plants.

II.4.2 Structural variants (SVs)

Structural variants (SVs) represent larger genetic rearrangements (>50 bp) that comprise various types of variants: deletions, insertions, inversions, translocations, duplications, and copy number variations (CNVs) (Figure II.2) (Tattini et al. 2015). To date, for the identification of SVs from NGS reads, three major strategies have been exploited: i) depth of coverage, ii) paired-end mapping and iii) split reads (Alkan et al. 2011). For many SVs (excluding inversions), the rearrangement causes a change in the number of reads that map to a given region in the reference genome. This is most straightforward in the case of deletions, duplications and CNVs as, in all three cases, the reference genome contains the affected region that either is lacking in the re-sequenced sample (resulting in a lack of read coverage)

or present in more copies (leading to abnormally deep coverage) (Campbell et al. 2008). Translocations represent a special case where read coverage is absent at the original location of the translocated segment in the reference genome, but appears as an insertion in a new position in the reference genome. Insertions cannot be detected through depth of coverage analysis for lack of the corresponding sequence in the reference genome on which sequence reads are mapped. Finally, inversions cannot be detected in this way as they do not result in a change in read coverage (except at the breakpoints of the inverted segment). The second approach, paired-end reads, relies on reads derived from the two ends of the same DNA fragment originally obtained after fragmentation of the genomic DNA (Hormozdiari et al. 2011). Because of the known mean distance between these paired reads and the expectation that they should map to opposite strands of the reference genome, deviations from these expectations provide evidence of a SV in an individual sample compared to reference genome (Ye et al. 2009). As illustrated below, deletions and insertions result in abnormal spacing between the paired reads. Inversions, duplications, translocations and CNVs only cause abnormal read pairs at the junctions between the rearranged segment and the rest of the reference genome. Finally, split-read mapping is specifically aimed at detecting SV breakpoints (Mills et al. 2011). This strategy exploits the fact that SVs generate breakpoints that are analogous to "scars". These "scars" generate sequence reads that are not contiguous in the reference genome. The alignment of the two portions of the sequence in two different regions of the reference genome provides evidence for the existence of SV in an individual sample (Figure II.2).

To date, numerous studies have illustrated the functional importance of the SVs in crop plants where these have been associated with diverse phenotypes ranging from adaptation to disease resistance. One such example is resistance to soybean cyst nematode (SCN) in soybean. Cook et al. (2012) showed that copy number variation at the *Rhg1* locus determines nematode resistance in soybean, with a high copy number resulting in greater resistance. CNVs have been extensively characterized in maize (Springer et al. 2009). Maron et al. (2013) found that an increased number of copies of the *MATE1* gene is associated with superior Al tolerance in maize. Wang et al. (2015) reported that CNV at the GL7 locus contributes to grain size diversity in rice. Nishida et al. (2013) identified a deletion in the 5' upstream region of photoperiod-insensitive alleles *Ppd-A1a* and *Ppd-B1a* in hexaploid wheat (*Triticum aestivum* L.). They also showed the functional effect of this deletion on wheat heading time. Furthermore, it has been comprehensively documented that genes encoding miRNAs in plants originated by inverted duplication of target gene sequences (Fenselau et al. 2008). Additionally, several studies have shown the impact of translocations on neutral and functional

genetic diversity within and among plant populations (Saxena et al. 2014). These examples provide ample evidence that SVs often contribute to allelic variation at loci of great functional significance.

Despite their involvement in the generation of allelic diversity in crop plants, the identification of SVs on a genome-wide scale remains very challenging and limited. The identification of SVs in crop plants, using WGR data or comparative genomic hybridization (CGH) arrays (Pinkel, 2005), has mostly been limited to the identification of large deletions, insertions and sometimes CNVs (Redon et al. 2009). Recently, Torkamaneh et al. (2017) identified 92K SVs among a collection of 102 elite soybean accessions using WGR data and three SV discovery approaches. More importantly, they showed that 34.5% of SVs or their breakpoints (close to 32k SVs) overlapped completely or partially with genic regions. This indicates that a substantial proportion of SVs would be expected to impact the function of one or more genes. In contrast, of the ~5M SNPs and small indels identified in the same collection of lines, only a very small proportion resided in coding regions (2%) and a still smaller subset (0.01%) were predicted to impact gene function. Thus, despite the much lower abundance of SVs compared to SNPs and small indels, their "functional footprint", e.g. the number of genes impacted by such variants, is relatively similar.

II.4.3 Challenges of WGR in crop breeding

Despite the significant reductions in cost experienced over the last few years, WGR of every accession remains too costly to be performed routinely on thousands of individual lines assessed each year within a breeding program. Fortunately, it is usually unnecessary as the amount of recombination encountered within the progeny of a cross is relatively limited such that large segments of the genome remain unaffected by recombination (Esch et al. 2007). If one has captured the alleles and their association (in the form of haplotypes) in a collection of parental accessions through WGR, it is sufficient to simply scan the progeny of a cross and impute missing genotypes (Howie et al. 2011). A second challenge is that WGR generates a huge amount of sequencing data that should be analyzed and stored. The analysis of WGR data requires high-performance computing systems (computers with a large number of processors (CPUs) and large amounts of memory) (Muir et al. 2016). Typically, breeders are not well equipped either for storing this volume of data or performing the various bioinformatics analyses. Although the use of WGR is limited in crop breeding programs, it has greatly facilitated the development of the high-throughput genotyping methods for crop plants.

II.5 NGS-based SNP arrays for crop genotyping

Large-scale sequencing, made possible and affordable thanks to NGS technologies, has allowed researchers to probe nucleotide diversity in panels of individuals to discover genetic variants (mostly SNPs and small indels). The identification of large numbers of molecular markers in crops has allowed the development of high-throughput genotyping tools such as SNP arrays (Ganal et al. 2012). Array-based genotyping methods are based on two strategies: i) the use of solid-phase bound oligonucleotide probes diagnostic for the respective alleles and subsequent hybridization of genomic DNA onto such arrays (Affymetrix) (Adessi et al. 2000), and ii) the use of single-base primer extension (SBE) technologies to determine the specific allelic state for a given SNP (Illumina) (Giusto and King, 2003). To date, SNP arrays with >40K SNPs have been developed for several crops such as rice (44K, 50K and 700K), maize (50K and 600K), soybean (50K, 180K and 355K), rye (600K), pepper (640K), canola (60K), cotton (63K) and wheat (90K, 660K and 820K) (Rasheed et al. 2017). Generally, on arrays capable of interrogating fewer than 100K SNPs, more than 80% of markers are present in genic regions.

These SNP arrays are widely used for genetic diversity analysis (Song et al. 2013), evolutionary studies (identification of domestication and improvement sweeps), GWAS (Gao et al. 2016), and marker-assisted selection (MAS) programs (Collard et al. 2008). As

described, SNP chips have been or are currently being developed for a large number of important crop plants that can be used to get more precise insights into their genetic constitution and for the improvement of breeding programs.

II.5.1 Limitations of SNP arrays

SNP arrays have greatly reduced the time and effort spent on genotyping, but the development of new markers or new arrays still requires significant investments (Tennessen et al. 2011). It has been largely shown that an increase in SNP density results in a higher resolution in large samples for genome-wide association studies (GWAS), bulk segregant analysis (BSA) and genomic selection (GS) (Deschamps et al. 2017). The development of a new SNP array requires prior generation of sequence information, identification of polymorphisms, validation and array production that can be seriously restricted by cost and time (Tennessen et al. 2011). Furthermore, current array-based technologies have clear limitations for different application, because the markers are often specific to the population in which they were developed, and the resulting allelic bias can be problematic in some divergent populations and species (Lachance, and Tishkoff, 2013). In other words, SNP loci that are polymorphic in one set of accessions may not be informative in another and vice versa. On the other hand, it has been documented that several biological factors in crop plants can affect the quality of SNP arrays such as polyploidy, high structural polymorphism, significant sequence diversity and a high proportion of repetitive regions (Deschamps et al. 2017). For example, in a hexaploid species such as bread wheat, interrogating SNPs located in genic regions (that are typically more highly conserved) increases the odds of capturing DNA fragments originating from the various homeologues (Deschamps et al. 2017). This will greatly complicate the calling of genotypes at such a SNP locus. Also, several studies have reported that most causal SNPs (i.e. ones responsible for a change in phenotype) are located in regulatory regions and not in the coding region (Edwards et al. 2013). As a majority of markers on SNP chips are present in genic regions, this can reduce the odds of capturing such causal mutations on arrays.

II.6 Genotyping by sequencing (GBS)

Genotyping by sequencing (GBS), is a genotyping approach that relies on sequencing to simultaneously discover nucleotide positions that are polymorphic within a collection of samples and call genotypes at these informative sites (Elshire et al. 2011). It does not examine all nucleotide positions in a genome, but relies on a complexity reduction approach to inspect a relatively small and constant subset of the genome (Rosato et al. 2012). This is

achieved through the use of restriction enzymes (one or a combination of enzymes) that cut the genome at the same position in most samples (Davey et al. 2011). Once the genomic DNA has been digested with the chosen enzyme (or enzyme combination), the resulting restriction fragments will be sequenced in part (typically 100-150 bp) to provide sequence information on the region immediately flanking the restriction sites (Elshire et al. 2011; Sonah et al. 2013). GBS provides the ability of exceptional multiplexing of individual samples through barcoding. High levels of multiplexing and consistently reduced genome representations have been achieved via GBS, thus allowing a significant reduction in cost. The GBS approach and it applications in crop breeding have been greatly described and discussed in several reviews (Poland and Rife, 2012). The key factors that must be considered in any GBS experiment are the analytical pipeline and missing data imputation. In the following section, we will discuss these two aspects of GBS.

II.6.1 GBS data analysis

GBS data analysis can be complex owing to both biological and technical factors (Nielsen et al. 2011; Gompert et al. 2010; Lynch et al. 2009; Hohenlohe et al. 2010). Among the former, we can note the number of detected variants, the complexity of the genome, the degree of heterozygosity, the proportion of repetitive sequences throughout the whole genome, the level of polymorphism and divergence among populations. Among the latter, we need to consider the degree of sample multiplexing, the total number of reads per sample, the length of reads, and the sequencing error rate. To overcome these challenges and extract SNP genotypes from a large number of GBS reads, efficient and accurate bioinformatics analytical pipelines are required. In these pipelines, several steps must be included to filter out poor-quality reads, categorize reads by pool or individual (based on barcodes), align reads to a reference genome to uncover polymorphisms, and finally score genotypes for each individual at each polymorphic locus (Glaubitz et al. 2014; Torkamaneh et al. 2017a). Early in the development of GBS in crop plants, Illumina was the most commonly used sequencing technology (with fixed read length of ~100 bp) and TASSEL was the main GBS bioinformatics analytical pipeline (Bradbury et al. 2007). Later, the Ion Torrent sequencing technology also started to be used for GBS. It differs from Illumina sequencing in that it produces reads of variable length (50 to 150-bp) (Mascher et al. 2012). Recently, several custom packages such as Stacks (Catchen et al. 2013), IGST (Sonah et al. 2013) and Fast-GBS (Torkamaneh et al. 2017a) have been developed specifically for the processing of reads produced by GBS technologies. All of these GBS pipelines were developed using different combinations of tools for demultiplexing, trimming, mapping, and variant calling. Mascher et al. (2013) performed

GBS on barley RILs using three sequencing technologies (Illumina, Ion PGM and Ion Proton) and using two GBS bioinformatics pipelines (TASSEL and IGST) found ~53% overlap between SNP calls derived from Illumina and Ion Proton reads. More recently, a greater overlap (69%) between Ion Proton and Illumina SNP calls was reported in soybean (Torkamaneh et al. 2016). Both studies reported a high level of concordance between shared SNPs (~99%). Recently, Torkamaneh et al. (2016) have comprehensively compared seven GBS pipelines (Stacks, Stacks *de novo*, TASSEL-GBS v1, UNEAK, IGST, TASSEL-GBS v2 and Fast-GBS) and two sequencing technologies (Illumina and Ion Proton) for variant calling from GBS data. They found more than 87% overlap between different GBS pipelines, with the sole exception of TASSEL-GBSv1 that showed the lowest overlap (36.7%). Furthermore, they showed that SNPs called by more than one pipeline were typically highly accurate. They also documented that the main source of errors in GBS SNP calls was the presence of paralogues and/or repetitive regions. Typically, all such pipelines offer a certain number of parameters that can be adjusted by the user based on the specific properties of the genetic materials being studied. It is impossible to develop a universal pipeline that would be equally suited to every situation. Ultimately, users need to adjust pipeline parameters to suit their chosen sequencing platform (Illumina vs. Ion Proton) and the characteristics of their species in terms of genome complexity (genome size and proportion of repetitive regions), ploidy and level of heterozygosity.

II.6.2 Missing data in GBS

As described above, GBS is a genome-wide scanning or sampling approach. By nature, GBS will generate sizeable amounts of missing data because sequence reads are not necessarily obtained for the same region (flanking a restriction site) in all individuals subjected to GBS (Rutkoski et al. 2013; Jarquín et al. 2014). Also, because the GBS sequence reads are distributed across a very large number of loci, the mean depth of coverage at each site is relatively thin (often less than 10). Several studies have shown that the quantity of missing data generated by GBS can be substantial, thus the final number of informative SNPs obtained from GBS data can be greatly affected by the chosen tolerance towards missing data (Beissinger et al. 2013; Crossa et al. 2013). Typically, such missing data will need to be imputed as many tools used in genetic analysis require complete datasets. Among a panel of 301 soybean accession, the number of markers increased five fold (from 12K to 62K) when increasing the amount of missing data tolerated at each locus, from 20% to 80% (Torkamaneh and Belzile, 2015). It is important to note, however, that this criterion refers to the maximal proportion of missing data per locus. When SNP loci with up to 20% of missing

data were retained, the overall proportion of missing data in the SNP dataset was 7%. When this maximal allowance was increased to 80%, the overall proportion of missing data was 51%. The question that needs to be answered (and the answer may vary in different crops) can be framed in this way: Is it better to impute a small amount of missing data (e.g. 7% in the example given) at a limited number of loci (12K SNPs) or to impute a larger amount of missing data (e.g. 51%) at a much larger number of loci (62K SNPs) for which there are some data? This requires some understanding of how imputation works.

Generally, imputation is the substitution of some value for missing data, in other words, 'filling in' missing data with plausible values through various strategies (Hao et al. 2009). Several imputation algorithms were designed for imputation in ordered markers such as Hidden Markov Models (HMH), linear models and pedigree-based haplotyping (Glodziket al. 2013; Cheung et al. 2013; Kong et al. 2008; Pei et al. 2008). Most current imputation tools used for the imputation of missing GBS data are based on the HMH algorithm. These tools rely on linkage disequilibrium (LD), i.e. non-random or favored occurrences of certain combinations of alleles at different loci (Li et al. 2009). SNPs residing close together on a chromosome are often inherited together as a unit known as a haplotype (Slatkin, 2008). In this approach, missing alleles can be inferred from the available data in other samples sharing the same haplotype (Figure II.3). In principle, a larger number of SNP markers (even with half of the data missing) could provide a better opportunity to capture the haplotypes than a smaller number of markers (albeit with fewer missing data). In the two contrasting scenarios described above (12K SNPs with 7% missing data and 62K SNPs with 51% missing data), the accuracy of imputation of missing data was higher with more SNPs and missing data (96%) than with fewer SNPs and missing data (12K with 7% missing data) (Torkamaneh and Belzile 2015).

As described above, imputation success is related to how LD blocks and haplotypes are captured by SNP data. It thereby stands to reason that imputation accuracy increases with an increasing density of markers. Unfortunately, LD patterns are not homogenous across species. In some, such as soybean and rice, LD extends over long stretches (soybean: ~150 kb; rice: <65–180) (Lam et al. 2010; Zhu et al. 2007). In others, such as maize (<1 kb) or *Arabidopsis thaliana* (~3–4 kb) (Gore et al. 2009; Kim et al. 2007), LD decays much faster and a larger number of markers will be needed to adequately capture the underlying haplotypes. Thus, it is impossible to define an optimal level of tolerance for missing data in different species or even in different collections of accessions from the same species if they differ in the extent of LD between markers. Nonetheless, as GBS usually allows for the identification of large sets of informative SNP loci, the imputation accuracy of missing data

reported in different species (maize, rice, wheat, barley and soybean) has generally been high (92–98%; Crossa et al. 2013; Huang et al. 2014; Jarquín et al. 2014; Torkamaneh and Belzile 2015).

II.7 Integrating data obtained using different genotyping tools

II.7.1 Combining two SNP datasets via imputation

Genotyping platforms differ in the set of loci on which they can provide information and updated content is continuously being added as new products and datasets become available (LaFramboise, 2009). Several SNP chips with different SNP sets have been developed for different species (e.g. 6K, 50K, 180K and 355K in soybean) (Wang et al. 2016). There is therefore a need to be able to combine these available datasets via imputation. The process for combining two genotypic datasets via imputation is schematically illustrated in Figure II.4. Here two sets of samples were genotyped with two different genotyping platforms (SNP array and GBS). As can be seen, there are two categories of SNP loci: i) platform-specific SNP loci (blue or yellow in the figure), and ii) common SNP loci present in both datasets (green). A partial overlap can also exist with regards to the samples for which data are available, i.e. data for some samples may be available only for one of the two genotyping technologies. Merging these two datasets will provide partial information for all samples (Hao et al. 2009). Then missing data in the combined dataset can be imputed. For example, Torkamaneh and Belzile (2015) used imputation to combine SNP catalogues derived from two high-throughput genotyping techniques in soybean: GBS and a SNP array. The GBS-derived dataset (301 samples, 60K SNPs) was merged with a SNP array dataset (25 samples, 40K polymorphic SNPs), where these 25 samples were a subset of the larger collection of 301. Despite the limited overlap between GBS and SNP array SNP loci (7% of common loci), the resulting catalogue (301 samples, >100K SNPs) was highly accurate with ~95% of the missing data having been correctly imputed. Combining SNP datasets derived from different genotyping tools can thus be successfully performed and can enhance the power of genetic analysis in crop plants.

II.7.2 Genotype imputation using a reference panel

Genotype imputation using a reference panel refers to the situation in which a reference panel of haplotypes (generally derived from WGR projects) with a genome-wide exhaustive set of SNPs can be used to impute onto a set of samples that have been genotyped at a subset of the SNPs (derived from GBS or a SNP array). An overview of this process is given in Figure

II.5. Here, imputation algorithms use the correlation (LD) between SNPs present in the reference panel for making predictions of the genotypes present in the samples genotyped using a low-density (low-cost) method. These algorithms use both the dense information from reference panel and less-dense genotype information from samples to infer genotypes at SNP loci that are missing (Marchini and Howie 2010). To date, such reference panels have been developed for Arabidopsis, maize, rice and soybean (Cao et al. 2011; Bukowski et al. 2015; The 3,000 rice genomes project, 2014; Torkamaneh et al. 2017b). In Arabidopsis, a high level (>98%) of missing data imputation accuracy has been reported (Cao et al. 2011) using a reference panel with 80 samples. In maize, a set of 35M SNPs discovered by WGR of 1,268 inbred lines, was imputed on a large collection (>10,000) previously genotyped with 500K GBS-derived SNPs, again with a high level of accuracy (98%) (Swarts et al. 2016). In humans, it has been documented that population structure, the properties of the reference panel (comprehensiveness of haplotype diversity) and the chosen low-density genotyping platform (GBS or SNP array) will all influence performance, and performance may vary between rare and common alleles (Marchini and Howie 2010). In the coming years, we expect that imputation based on reference panels, due to the ever-increasing availability of WGR data, will become a key tool in crop genomics.

II.8 Conclusion

Genotyping technologies have become an essential component in many crop breeding programs. Continuous reductions in the cost of sequencing and rapid advances in data processing suggest that sequencing-based genotyping approaches will become increasingly advantageous. Similarly, decreases in the cost of sequencing will spur an important increase in the use of WGR as a means to provide exhaustive characterization of nucleotide and structural variation in core collections in view of capturing a significant portion of this extant variation. Such in-depth characterization of genetic variation in core collections will also provide exceptional data for genotype-phenotype association studies (e.g. GWAS). Cost-effective, genome-wide genotyping platforms (e.g. GBS and SNP chips) will remain the main tool in breeding programs. Despite all these impressive technological advances, the uptake of these new tools will likely require a significant effort in user training and in the development of analytical tools capable of extracting information that is the most relevant to breeders from these very large datasets.

II.9 Acknowledgements

The authors wish to acknowledge the financial support received from Génome Québec, Genome Canada, the government of Canada, the Ministère de l'Économie, Science et Innovation du Québec, Semences Prograin Inc., Syngenta Canada Inc., Sevita Genetics, Coop Fédérée, Grain Farmers of Ontario, Saskatchewan Pulse Growers, Manitoba Pulse & Soybean Growers, the Canadian Field Crop Research Alliance and Producteurs de grains du Québec.

II.10 Figures

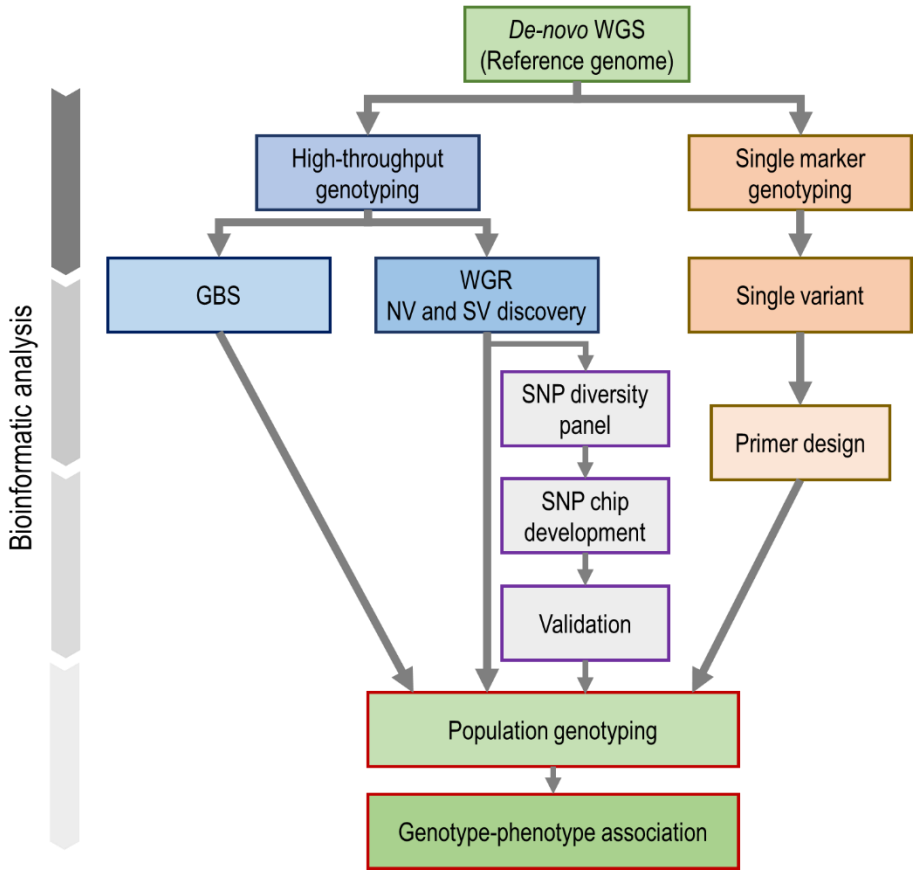


Figure II.1. The position of NGS and bioinformatic analysis in crop breeding program. One of the main aims of modern crop breeding is development of genetic markers related to agronomic traits. Application of different genotyping platforms and approaches is related to breeding program.

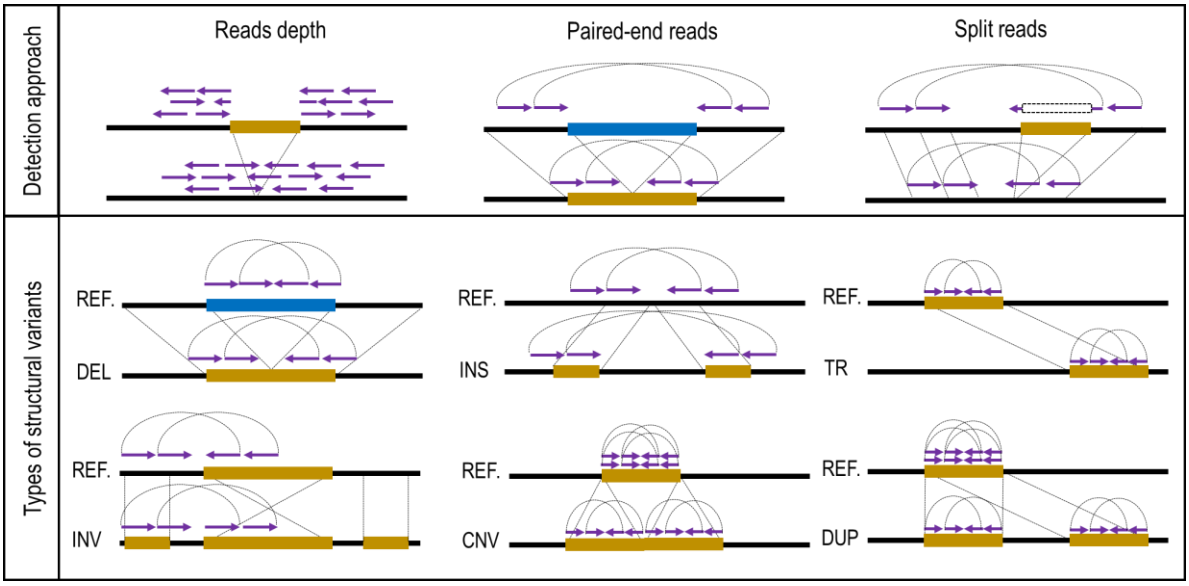


Figure II.2. Identification of structural variants through the analysis of NGS reads. Top, three different approaches used in SV detection. Bottom, different types of SVs and identification strategies, DEL: deletion; INS: insertion; TR: translocation; INV: inversion; CNV: copy-number variation; DUP: duplication.

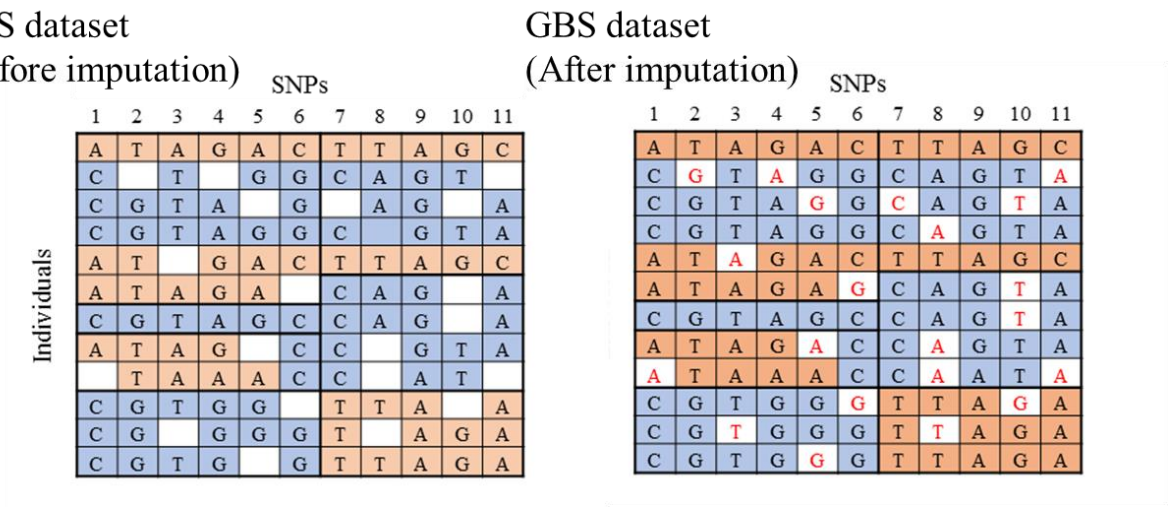


Figure II.3. Phase-based imputation of missing data. Left, GBS raw genotype table with missing data (white blocks). Right, imputed dataset (white blocks with green imputed genotype values). Markers located within the same LD block are shaded in the same tone of purple.

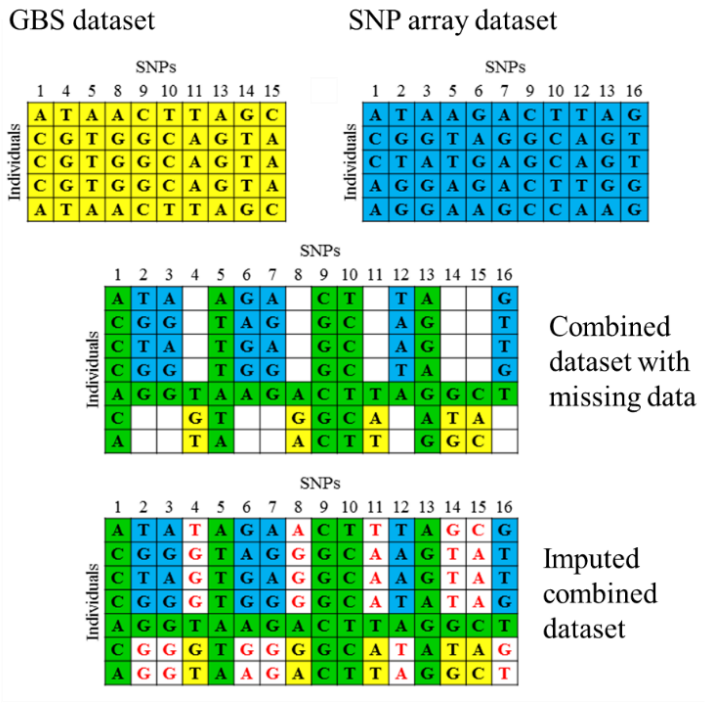


Figure II.4. Integration of different genotype dataset via imputation.

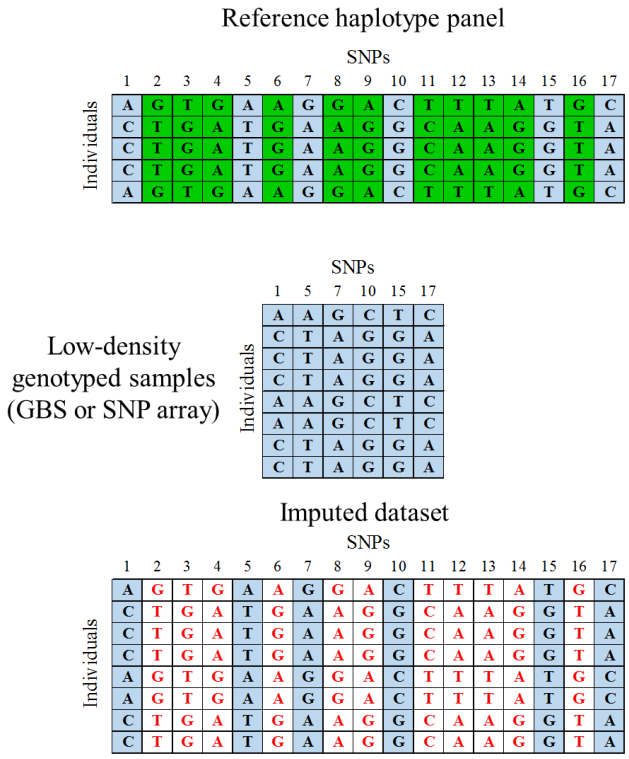


Figure II.5. Untyped-genotype imputation using haplotype reference panel.

Chapitre III

Fast-GBS: a new pipeline for the efficient and highly accurate calling of SNPs from genotyping-by-sequencing data

Davoud Torkamneh[1,2], Jérôme Laroche[2], Maxime Bastien[1,2], Amina Abed[1,2], and François Belzile[1,2]

[1]Département de Phytologie, Université Laval, Québec City, QC, Canada
[2]Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Quebec City, QC, Canada

BMC Bioinformatics
2017, 18:5

III.1 Résumé

Les technologies de séquençage de la nouvelle génération (NGS) ont considérablement accéléré l'étude de la composition des génomes et de leurs fonctions. Génotypage par séquençage (GBS) est une approche de génotypage qui fait usage de NGS pour balayer rapidement et économiquement d'un génome. Il a été démontré qu'il permet la découverte simultanée et le génotypage de milliers à des millions de SNP à travers un large éventail d'espèces. Pour la plupart des utilisateurs, le principal défi de GBS est l'analyse bioinformatique de la grande quantité d'informations de séquence dérivées du séquençage des librairies GBS en vue d'appeler les allèles au locus SNP. Nous décrivons ici un nouveau pipeline d'analyse bioinformatique GBS, appelé Fast-GBS, conçu pour fournir un génotypage très précis, nécessiter des ressources informatiques modestes et offrir une facilité d'utilisation. Fast-GBS est basé sur le langage et les formats de fichiers bioinformatiques standard, capable de gérer les données à partir de différentes plates-formes de séquençage. En plus il est capable de détecter différents types de variants (SNP, MNP et Indels). Pour illustrer sa performance, nous avons appelé des variants chez trois collections d'échantillons (soja, orge et pomme de terre) qui couvrent une gamme de différentes au termes de tailles de génome, les niveaux de complexité du génome et de ploïdie. Au sein de ces petits ensembles d'échantillons, nous avons appelé 35k, 32k et 38k SNP pour le soja, l'orge et la pomme de terre, respectivement. Pour évaluer la précision du génotype, nous avons comparé ces génotypes de SNP dérivés de GBS avec des ensembles de données indépendants obtenus à partir de séquençage de génome entier ou la puce de SNP. Cette analyse a donné des précisions estimées de 98,7, 95,2 et 94% pour le soja, l'orge et la pomme de terre, respectivement. Nous concluons que Fast-GBS fournit un outil hautement efficace et fiable pour appeler des SNP à partir de données GBS.

III.2 Abstract

Next-generation sequencing (NGS) technologies have accelerated considerably the investigation into the composition of genomes and their functions. Genotyping-by-sequencing (GBS) is a genotyping approach that makes use of NGS to rapidly and economically scan a genome. It has been shown to allow the simultaneous discovery and genotyping of thousands to millions of SNPs across a wide range of species. For most users, the main challenge in GBS is the bioinformatics analysis of the large amount of sequence information derived from sequencing GBS libraries in view of calling alleles at SNP loci. Herein we describe a new GBS bioinformatics pipeline, Fast-GBS, designed to provide highly accurate genotyping, to require modest computing resources and to offer ease of use. Fast-GBS is built upon standard bioinformatics language and file formats, is capable of handling data from different sequencing platforms, is capable of detecting different kinds of variants (SNPs, MNPs, and Indels). To illustrate its performance, we called variants in three collections of samples (soybean, barley, and potato) that cover a range of different genome sizes, levels of genome complexity, and ploidy. Within these small sets of samples, we called 35k, 32k and 38k SNPs for soybean, barley and potato, respectively. To assess genotype accuracy, we compared these GBS-derived SNP genotypes with independent data sets obtained from whole-genome sequencing or SNP arrays. This analysis yielded estimated accuracies of 98.7, 95.2, and 94% for soybean, barley, and potato, respectively. We conclude that Fast-GBS provides a highly efficient and reliable tool for calling SNPs from GBS data.

III.3 Introduction

Currently, genomics lies at the heart of an extraordinary number of discoveries, innovations and applications. This revolution is a direct result of the rise of next-generation sequencing (NGS) technologies (Metzker 2010; Edwards *et al.* 2013; Kilpinen & Barrett 2013; Kumar et al. 2012). In the area of genotyping, the combination of NGS and reduced representation methods, which focus the sequencing effort on a small subset of the genome, has made it possible to simultaneously perform genome-wide single nucleotide polymorphism (SNP) discovery and genotyping in a single step even in species with large genomes (Davey et al. 2011) This has facilitated greatly the genotyping of very large numbers of SNPs using a number of related methods (e.g. CRoPS, RAD-seq, GBS, double-digest RAD-seq, and 2bRAD) (van Orsouw *et al.* 2007; Etter et al. 2007; Elshire et al. 2011; Etter et al. 2011; Peterson et al. 2012; Wang et al. 2012). These various methods make it possible to study important questions in molecular breeding, population genetics, ecological genetics and evolution using thousands to millions of genetic markers in a wide array of species (Davey et al. 2011). Genotyping-by-sequencing (GBS) is a particularly attractive complexity reduction method that offers a simple, robust, low-cost, and high-throughput method for genotyping in both model and non-model species (Elshire et al. 2011).

Advanced sequencing technologies (NGS) have reduced both the cost and the time required to generate sequence data. The efficient and accurate computational processing, variant and genotype calling, of large-scale NGS sequence data is the new bottleneck in genomics. To meet this need, numerous bioinformatics pipelines have been developed (Nielsen et al. 2011; Bradbury et al. 2007; Glaubitz et al. 2014; Catchen et al. 2013; Lu et al. 2013) and all need to accomplish a similar set of steps such as: 1) acquiring raw sequence data, 2) demultiplexing pooled sequence read data, 3) filtering out low-quality reads, 4) assembling or aligning reads, and finally 5) discovering polymorphic loci and inferring actual genotypes at these loci. Each step has its own set of associated challenges and uncertainties. These arise from genomic attributes such as the number of loci identified, genome complexity, degree of heterozygosity, abundance of repetitive sequences throughout the genome, and the level of polymorphism and divergence among populations (Nielsen et al. 2011). These biological factors also interact with technical factors such as the quality of the DNA, the degree of sample multiplexing, the total number and length of reads, and the sequencing error rate (Gompert et al. 2010; Lynch et al. 2009; Hohenlohe et al. 2010a). Key decisions therefore need to be made at each step regarding parameters such as the required depth of coverage or allowable nucleotide distance between reads for assembly. Finally, because of biological and sequencing sampling variation, the use of statistical models will often be necessary.

Conventionally, bioinformatics pipelines for handling GBS data are categorized in two groups: *de novo*-based and reference-based. In the presence of a reference genome, the reads from reduced-representation sequencing can be mapped to the reference genome and SNPs can be called (Nielsen et al. 2011; Li & Durbin 2009). Up to now, several reference-based GBS analysis pipelines have been developed. The most widely used reference-based GBS analysis pipelines are: TASSEL-GBS (v1 and v2), Stacks, and IGST (Bradbury et al. 2007; Glaubitz et al. 2014; Catchen et al. 2013; Sonah et al. 2013). But when a reference genome is not available, pairs of nearly identical reads (presumed to represent alternative alleles at a locus) need to be identified. The most highly used pipelines for such a *de novo*-based approach are UNEAK and Stacks (Catchen et al. 2013; Lu et al. 2013).

Herein, we describe a new reference-based pipeline, Fast-GBS, and we benchmark the pipeline based upon a large-scale, species-wide analysis of soybean, barley and potato. It is easy to use with various species, in different contexts, and provides an analysis platform that can be run with different types of sequencing data and modest computational resources.

III.4 Test dataset

To test the performance of Fast-GBS, we used existing sequence datasets of association mapping panels for three species covering a range of genomic situations: soybean (Torkamaneh & Belzile 2015), barley (Abed et al. unpublished), and potato (Bastien et al. unpublished). Table III.1 shows the species which we used in this study. These vary in terms of their ploidy, genome size and mode of reproduction (which relates to the expected zygosity). We used sequence datasets composed of 24 samples for each species.

III.5 Genotype validation

To estimate genotype accuracy for Fast-GBS calls, we compared the called SNPs with independently derived genotypic data resulting from either whole-genome resequencing (soybean and barley) or genotyping on a SNP array (potato) for the same samples. For soybean, we compared the GBS-called SNPs with whole genome resequencing data for the same 24 samples. In the case of barley, GBS-derived genotypic data for one of the 24 barley samples (cv. Morex) was compared to the barley reference genome produced using this same cultivar. For potato, we compared the GBS-derived genotypes with those obtained for the same 24 samples at a set of 122 SNPs that were in common with the SolCAP Infinium Chip (8.3k SNPs) (Felcher et al. 2012).

III.6 Implementation

The Fast-GBS analysis pipeline was developed by integrating public packages with internally developed tools. The public packages include Sabre (demultiplexing), Cutadapt (read

trimming and cleaning) (Martin 2011), BWA (read mapping) (Li & Durbin 2010), SAMtools (file conversion and indexing) (Li 2011), and Platypus (post-processing of reads, haplotype construction and variant calling) (Rimmer et al. 2014). Fast-GBS functions and software tools are presented in Figure III.1.

III.6.1 Creating directory structure

We developed a Bash script to create the directory structure before running the Fast-GBS pipeline. This command line creates the directories for data (FASTQ files), barcodes (key file), reference genome, and results (Fast-GBS outputs).

III.6.2 Input

The input data are sequenced DNA fragments from any restriction enzyme–based GBS protocol. Fast-GBS handles raw sequencing data in FASTQ format.

III.6.3 Preparing the parameter file

The parameter file is a text file containing key information about the analysis including the path to the FASTQ files, barcodes and reference genome. It also contains information about the type of sequence (paired or single-end), the adaptor sequence and the sequencing technology. In this file we can define critical filtering options such as the minimal quality scores for reads, minimal number of reads required to call a genotype, and maximal amount of missing data allowed. Number of CPU, names of output files are also defined in this file. This file comes with the Fast-GBS pipeline.

III.6.4 Data demultiplexing

The cost efficiency of GBS is partly due to the multiplexing of samples and the resulting pooled reads will need to be demultiplexed prior to SNP calling. Fast-GBS uses Sabre to demultiplex barcoded reads into separate files. It simply compares the provided barcodes with the 5' end of each read and separates the reads into the appropriate barcode files after having clipped the barcode from the read. If a read does not have a recognized barcode, it is put into an "unknown" file. Sabre also has an option (-m) to allow mismatches within barcodes. Sabre supports gzipped input files. After demultiplexing, Sabre outputs a BC summary log file of how many reads went into each barcode file.

III.6.5 Trimming and cleaning

After demultiplexing, Fast-GBS uses Cutadapt to find and remove adapter sequences, primers, and other types of unwanted sequence from high-throughput sequencing reads.

III.6.6 Read mapping algorithms

Fast-GBS uses the MEM (maximal exact matches) algorithm implemented in BWA that works by seeding alignments and then extending seeds with the Smith-Waterman (SW) algorithm using an affine gap penalty. This algorithm can perform local alignment for reads of 70 bp up to 1Mbp. This algorithm can perform parallel alignment, thus markedly increasing the speed of the analysis. The ability to align reads of variable size allows the use of data obtained using different sequencing platforms (Illumina, Ion Torrent, etc). Aligned reads may be gapped to allow for Indels.

III.6.7 Post-processing of mapped reads

After initial alignment, the mapped reads are further processed by Platypus in order to improve the sensitivity and specificity of variant calling. This post-processing seeks to improve the quality of mapping by performing a re-examination of poorly mapped reads and reads mapping to multiple locations. Platypus classifies poorly mapped reads in three categories: 1) reads with numerous mismatches (high level of sequencing errors), 2) reads mapping to multiple locations in the genome, and 3) any remaining linker or adaptor sequences (causing poor mapping). Variants called using such potentially incorrectly mapped reads (see next step) are highlighted using a BadReads flag.

III.6.8 Haplotype construction and variant calling

In Fast-GBS, variants are called using Platypus. Unlike alignment-based variant callers which focus on a single variant type (SNP or indel), Platypus uses multi-sample variant calling that helps to exploit information from multiple samples to call variants that may not look reliable in a single sample. This approach decreases the errors around indels and larger variants (MNPs). At first, the local assembler looks at a small window (~few kb) at a time and uses all the reads in the window to generate a colored de Bruijn graph, then using all candidate variants, it generates an exhaustive list of haplotypes. Candidate haplotypes are generated by clustering the candidate alleles across windows. Haplotype frequencies are estimated by the expectation-maximization (EM) algorithm. Then variants are called using the estimated haplotype frequencies. This approach works on the local haplotype level rather than on the

level of individual variants and does well on highly divergent regions. This also decreases computational requirements.

III.6.9 Variant and individual-level filtering

Platypus was originally designed and used to detect variants in human, mouse, rat and chimpanzee samples. To optimize Platypus options in the context of the analysis of GBS-derived single-end reads, we modified several options (see https://wiki.gacrc.uga.edu/wiki/Platypus-Sapelo for details of Platypus options). Some of the filters used in Fast-GBS variant calling steps are: number of reads (NR) per locus (default=2), mapping quality score of reads to call a variant (MQ≥10), minimum base quality (default=10), MNPs distance (minFlank=5), and maximum missing data (MaxMD) allowed (default≤80%). See Fast-GBS user manual for a full description of all filtering options.

III.6.10 Output data

The main output file of Fast-GBS is a .vcf file (Danecek et al. 2011) containing detailed information on each of the variants. In addition, Fast-GBS also generates a simple text file containing only the genotypic data. The Fast-GBS log file contains the completed steps of the pipeline as it is running. In cases where an error occurs and prematurely terminates the running of the pipeline, the log file shows the step at which the analysis stopped. An analysis can be started at any point on the existing intermediate files simply by creating a log file in which the previously completed steps are listed. Fast-GBS will re-initiate the analysis starting from that point onwards.

III.7 Results and discussion

III.7.1 Performance of Fast-GBS

To assess the performance of the Fast-GBS analysis pipeline, we used it to analyze existing GBS-derived read data from sets of 24 soybean, barley, and potato samples. Table III.2 presents a summary of this analysis. As can be seen, a total of 35k SNPs were called using 42M 100-bp Illumina reads on *Ape*KI-digested DNA from 24 different soybean lines. Similarly, for barley, 32k SNPs were successfully called from 72M Ion Torrent reads (50 – 150 bp in length) derived from a 24-plex *Msp*I/*Pst*I library. Finally, in potato, 38k SNPs were obtained from sequencing a 24-plex *Msp*I/*Pst*I library (43 million 100-bp Illumina reads).

GBS was originally demonstrated for soybean by Sonah et al. (2013) using the IGST pipeline. Using 8 diverse soybean lines, they called ~10k SNPs. Later work by the same group lead to

the calling of 45k SNPs on a large collection of 304 soybean lines for the purpose of conducting a GWAS study (Sonah et al. 2014). Analysis of this dataset using IGST took four days while the same analysis using Fast-GBS took only 11 hours and called ~60k SNPs (data not shown). As can be seen Fast-GBS present a high level of performance for soybean samples.

Barley has one of the larger genomes (>5 Gb) among cultivated plant species. Because of the huge size and high level of complexity of its genome, complexity reduction is highly recommended in barley, an important crop species for which a draft genome has been published (The International Barley Genome Sequencing Consortium 2012). Mascher et al. (2013) genotyped 94 barley RIL lines using GBS (*Msp*I/*Pst*I-digested library) and they called 34k and 19k SNPs using either the reference genome (with SAMtools) or a *de novo* pipeline (TASSEL), respectively. In this study we used Fast-GBS for SNP calling in barley and, as can be seen in Table III.2, Fast-GBS called 32k SNPs for a small number of samples (24). This showed the capability of Fast-GBS to run with large and complex genomes.

Because of the high level of ploidy and heterozygosity, potato is a challenging species for genotyping. The most often used method for genotyping in potato is a SNP array. Two SNP arrays have been developed so far, the SolCAP 8k and 20k arrays (Felcher et al. 2012; Peter et al. 2015; Prashar et al. 2014). Recently, Endelman (2015), genotyped 96 F2 diploid potato samples using GBS. Using an R-based bioinformatics pipeline to filter the GBS variants, they identified 11k SNPs. In this study, we called 38k SNPs from 24 samples which had also been genotyped using the SolCAP 8k SNP array. Of these, 5.5k SNPs on the array were polymorphic among this set of 24 potato samples. As can be seen, using Fast-GBS, we called around almost seven times more polymorphisms than using a SNP array (38k vs 5.5k SNPs).

III.7.2 Validation of Fast-GBS data

An important aspect to consider for any variant calling tool is the accuracy of called genotypes. In this study, we estimated the accuracy of genotypes called by Fast-GBS (Table III.2) by comparing them to the "true" genotypes (obtained from either whole-genome resequencing or SNP array data). For soybean, for all 24 samples, we compared the SNP genotypes called by Fast-GBS to the genotypes assigned to the same loci following whole-genome sequencing. We found a very high level of concordance, as almost all genotypes (98.7%) proved identical. For barley, we compared the SNP genotypes called by Fast-GBS with the true genotypes for one of the 24 lines (cv. Morex), the only one for which we had whole genome sequencing data. Again, a high degree of agreement between the two datasets (97%) was obtained. Finally, for potato, we used data obtained on the SolCAP 8k Infinium Chip for the same 24

samples used to perform GBS. These two datasets shared 122 SNP loci. In our initial comparison, only 87.7% were in agreement. When we examined the proportion of concordant calls, we discovered that more than 50% of all discordant calls came from only three samples and the degree of discordance in these was so great that it suggested we were not comparing the same clones. After removing these outliers from the analysis, 94% of genotypes called by Fast-GBS and the SNP array were in agreement in the remaining 21 clones. We conclude that Fast-GBS can accurately call SNPs in species with different characteristics (genome size, ploidy, zygosity).

III.7.3 Flexibility to run different sequencing platforms

In this study, to assess the performance of Fast-GBS, we used both Illumina and Ion Torrent reads. Soybean and potato samples were sequenced using an Illumina Hiseq platform and barley samples on an Ion Torrent (Proton) platform. Typically for GBS, Illumina sequencing generates reads of uniform length (100 bp), while Ion Torrent reads are in 50 to 150 bp. Ion Torrent sequencing usually leads to a higher rate of sequencing errors (Golan & Medvedev 2013; Bragg et al. 2013). Thus, it is preferable for an analytical pipeline to be versatile and capable of using reads derived from either technology (or new technologies in development). Most GBS bioinformatics pipelines are able to proceed with Ion Torrent reads, but often need to be modified to be suitable for this type of read data. TASSEL, UNEAK, and Stacks generate tags of a fixed length (e.g. 64 bp). This will lead to an important loss of sequence information and can lead to inaccurate or ambiguous mapping of reads. Also, because of the increased amount of sequencing errors, these pipelines can generate false tags which produce false SNPs. As shown above, Fast-GBS proved the capacity of accurately proceed maximum SNP calling using reads obtained from both sequencing platforms (Ion Proton and Illumina).

III.8 Conclusions

GBS provides an extremely powerful and versatile tool for identifying and calling genetic markers to be used by researchers working in numerous species and fields of study. This genotyping approach, like all applications based on NGS, generates a huge amount of raw data. These data need to be analyzed as quickly and efficiently as possible, all the while yielding SNP data that is highly accurate. Fast-GBS showed itself to be a powerful pipeline to generate large numbers of highly accurate SNPs using sequence read data obtained from different sequencing platforms and diverse species characterized by different levels of ploidy, zygosity, and genome complexity. By combining efficiency and accuracy in this way, Fast-GSB constitutes a useful tool for a broad array of users in different research communities.

III.9 Acknowledgements

This work has been made possible by the University of Laval (UL) and Institut de Biologie Intégrative et des Systèmes (IBIS). Funding for this research was provided by Agriculture and AgriFood Canada and the Canadian Field Crop Research Alliance (Grant no. AIP-CL23).

III.10 Tables

Table III.1. List of species genotyped using a GBS approach and analyzed using Fast-GBS. Description of three different species representing essential factors (ploidy, genome size and reproduction mode) influencing GBS analysis.

Name	*Species*	*Ploidy*	*Genome size (Mb)*	*Mode of reproduction*	*Number of chromosomes*
Soybean	*Glycine max*	Paleotetraploid	1,100	Selfing	20
Barley	*Hordeum vulgare*	Diploid	5,300	Selfing	7
Potato	*Solanum tuberosum*	Autotetraploid	844	Clonal	12

Table III.2. Number of variants detected among 24 soybean, barley, and potato samples. The sequencing platform, number of reads, filtering options, and genotype accuracy for each dataset are also provided.

				Filtering options*				
Name	**Sequencing platform**	**Restriction enzyme**	**Number of reads**	**minNR**	**MinMAF**	**MaxMD (%)**	**Number of variants**	**Accuracy (%)**
Soybean	Illumina	*Ape*KI	42 M	2	0.04	80	35k	98.7
Barley	Ion Torrent	*Msp*I/*Pst*I	72 M	2	0.04	80	32k	95.2
Potato	Illumina	*Msp*I/*Pst*I	43 M	11	0.04	20	38k	94.0

*Filtering options: minNR; minimum number of reads to call a variant (depth), MinMAF; minimum minor allele frequency, and MaxMD; maximum missing data allowed.

III.11 Figures

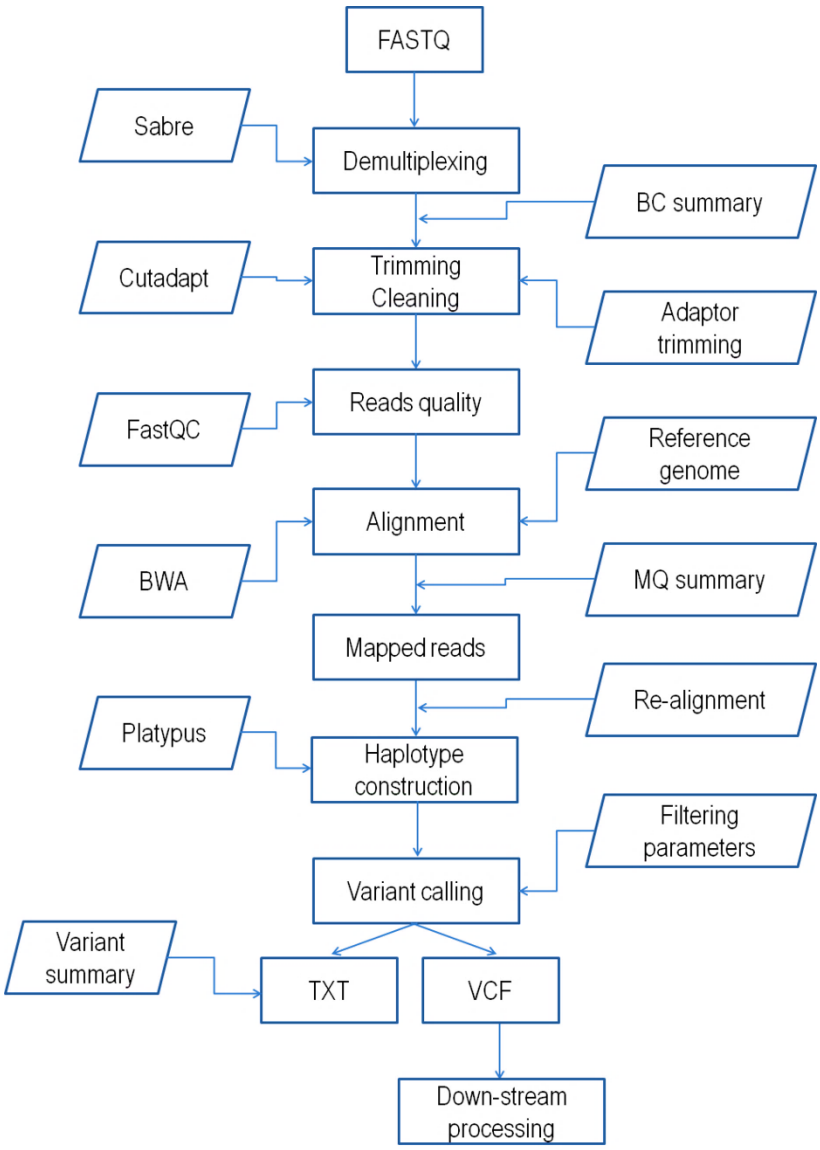


Figure III.1. Schematic representation of the analytical steps in the Fast-GBS pipeline. Showing implemented tools at left and inputs and outputs of each steps at right.

Chapitre IV

Genome-Wide SNP Calling from Genotyping by Sequencing (GBS) Data: A comparison of seven pipelines and two sequencing technologies

Davoud Torkamaneh[1,2], Jérôme Laroche[2], François Belzile[1,2]

[1]Département de Phytologie, Université Laval, Quebec City, QC, Canada

[2]Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Quebec City, QC, Canada

PLoS ONE

2016, 11(8): e0161333

IV.1 Résumé

Le séquençage de la nouvelle génération (NGS) a révolutionné la recherche sur les plantes et les animaux de nombreuses façons, y compris de nouvelles méthodes de génotypage à haut débit. Le génotypage par séquençage (GBS) s'est révélé être une méthode de génotypage robuste et rentable qui est capable de produire des milliers à des millions de SNP dans un large éventail d'espèces. Sans aucun doute, le plus grand obstacle à son utilisation plus large est le défi de l'analyse des données. Nous décrivons ici une comparaison complète de sept pipelines bioinformatique de GBS développés pour traiter les données brutes de séquence GBS pour génotypage de SNP. Nous avons comparé cinq pipelines qui nécessitent un génome de référence (TASSEL-GBS v1 et v2, Stacks, IGST et Fast-GBS) et deux pipelines *de novo* qui ne nécessitent pas de génome de référence (UNEAK et Stacks). En utilisant les données de séquence d'Illumina pour un ensemble de 24 lignes de soja dont leur génome a déjà entièrement séquencé, nous avons effectué des appels des SNPs avec ces pipelines et comparé les appels de SNP avec les données de séquençage afin d'évaluer leur précision. Le nombre de SNP appelés sans génome de référence était inférieur (13k à 24k) qu'avec un génome de référence (25k à 54k SNP) alors que la précision était élevée (92.3 à 98.7%) pour toutes les pipelines sauf une (TASSEL-GBSv1, 76.1%). Parmi les pipelines offrant une grande précision (> 95%), Fast-GBS a appelé le plus grand nombre de polymorphismes (près de 35 000 SNP + Indels) et a donné la plus haute dégrée d'exactitude (98,7%). En utilisant les données de séquence d'Ion Torrent pour les mêmes 24 lignes, nous avons comparé les performances de Fast-GBS avec celles de TASSEL-GBSv2. Il a encore appelé plus de polymorphismes (25,8 K contre 22,9 K) et ceux-ci se sont révélés plus précis (95,2 vs 91,1%). En règle générale, les catalogues SNP appelés à partir des mêmes données de séquençage utilisant différentes pipelines ont abouti à des catalogues SNP très chevauchants (chevauchement de 79 à 92%). En revanche, le chevauchement entre les catalogues SNP obtenus à l'aide du même pipeline, mais différentes technologies de séquençage étaient moins étendues (~ 50-70%).

IV.2 Abstract

Next-generation sequencing (NGS) has revolutionized plant and animal research in many ways including new methods of high throughput genotyping. Genotyping-by-sequencing (GBS) has been demonstrated to be a robust and cost-effective genotyping method capable of producing thousands to millions of SNPs across a wide range of species. Undoubtedly, the greatest barrier to its broader use is the challenge of data analysis. Herein we describe a comprehensive comparison of seven GBS bioinformatics pipelines developed to process raw GBS sequence data into SNP genotypes. We compared five pipelines requiring a reference genome (TASSEL-GBS v1& v2, Stacks, IGST, and Fast-GBS) and two *de novo* pipelines that do not require a reference genome (UNEAK and Stacks). Using Illumina sequence data from a set of 24 re-sequenced soybean lines, we performed SNP calling with these pipelines and compared the GBS SNP calls with the re-sequencing data to assess their accuracy. The number of SNPs called without a reference genome was lower (13k to 24k) than with a reference genome (25k to 54k SNPs) while accuracy was high (92.3 to 98.7%) for all but one pipeline (TASSEL-GBSv1, 76.1%). Among pipelines offering a high accuracy (>95%), Fast-GBS called the greatest number of polymorphisms (close to 35,000 SNPs + Indels) and yielded the highest accuracy (98.7%). Using Ion Torrent sequence data for the same 24 lines, we compared the performance of Fast-GBS with that of TASSEL-GBSv2. It again called more polymorphisms (25.8K vs 22.9K) and these proved more accurate (95.2 vs 91.1%). Typically, SNP catalogues called from the same sequencing data using different pipelines resulted in highly overlapping SNP catalogues (79-92% overlap). In contrast, overlap between SNP catalogues obtained using the same pipeline but different sequencing technologies was less extensive (~50-70%).

IV.3 Introduction

Next-generation sequencing (NGS) has facilitated greatly the development of methods to genotype very large numbers of molecular markers such as single nucleotide polymorphisms (SNPs). NGS offers several approaches that are capable of simultaneously performing genome-wide SNP discovery and genotyping in a single step, even in species for which little or no genetic information is available (Davey et al. 2011). This revolution in genetic marker discovery enables the study of important questions in molecular breeding, population genetics, ecological genetics and evolution. The most highly used methods of genotyping relying on NGS use restriction enzymes to capture a reduced representation of a genome (Miller et al. 2007; Baird et al. 2008; Van Orsouw et al. 2007; Andolfatto et al. 2011; Elshire et al. 2011; Peterson et al. 2012; Parchman et al. 2012; Sonah et al. 2013). New approaches such as restriction site-associated DNA sequencing (RAD-seq) and genotyping-by-sequencing (GBS) have been developed as rapid and robust approaches for reduced-representation sequencing of multiplexed samples that combines genome-wide molecular marker discovery and genotyping (Davey et al. 2011). This family of reduced representation genotyping approaches generically called genotyping-by-sequencing (GBS) (Davey et al. 2011). The flexibility and low cost of GBS makes this an excellent tool for many applications and research questions in genetics and breeding. Such modern advances allow for the genotyping of thousands of SNPs, and, in doing so, the probability of identifying SNPs correlated with traits of interest increases (Kumar et al. 2012). Even with advancement of NGS to produce millions of sequence reads per run, data analysis for these new approaches can be complex owing to using restriction enzymes, sample multiplexing, different fragment length and variable read depth (Davey et al. 2011). It is crystal clear that advanced analysis pipelines have become a necessity to filter, sort and align this sequence data. A pipeline for GBS must include steps to filter out poor-quality reads, classify reads by pool or individuals based on sequence barcodes, either identify loci and alleles *de novo* or align reads to an index reference genome to discover polymorphisms, and often score genotypes for each individual included in the study. Generally, pipelines for handling GBS data are categorized in two groups; *de novo*-based and reference-based. When a reference genome is available, the reads from reduced-representation sequencing can be mapped to the reference genome and SNPs can be called as for whole-genome resequencing projects (Li & Durbin 2009; Nielsen et al. 2011). Up to now, several reference-based GBS analysis pipelines have been developed. The most widely used reference-based GBS analysis pipelines are: TASSEL-GBS (v1 and v2) (Bradbury et al. 2007; Glaubitz et al. 2014), Stacks (Catchen et al. 2013), IGST (Sonah et al. 2013), and Fast-GBS (the most recent pipeline (Torkamaneh et al. 2017a)). In the absence of a reference

genome, pairs of nearly identical reads (presumed to represent alternative alleles of a locus) need to be identified. The most highly used pipelines for such a *de novo*-based approach are UNEAK and Stacks (Catchen et al. 2013; Lu et al. 2013).

Finally, different NGS sequencing platforms are currently available and offer different advantages. For example, whereas the Illumina technology offers very high throughput and read quality, this usually comes at the expense of speed as close to two weeks are required to complete a run. In contrast, the Ion Torrent technology (Rothberg et al. 2011) offers great speed (4 hours) at the expense of lower throughput and read quality. Depending on the constraints, one or the other technology may prove more suitable. Ideally, one would like SNP calling pipelines to perform equally well with both types of read data.

In this study, we comprehensively compared existing GBS analysis pipelines on the basis of the number of SNPs called, the accuracy of the resulting genotypes as well as the speed and ease of use of these pipelines. We also compared the results obtained using Illumina and Ion Torrent reads. Finally, we examined the amount of overlap in the SNP loci that were called using different pipelines.

IV.4 Materials and methods

IV.4.1 Samples and sequencing platform

Soybean (*Glycine max* L.) is a diploid species with 20 pairs of chromosomes and it has a medium-sized genome (1.1 Gb). Because it is an autogamous species, soybean lines/cultivars breed true and are highly homozygous. A set of 23 Canadian soybean lines and one plant introduction (PI) was subjected to GBS analysis. These same lines were resequenced as previously described by Torkamaneh and Belzile (2015). Using the same DNA, two GBS libraries were constructed following *Ape*KI digestion: one for Illumina sequencing (as per Elshire et al. (2011)) and the other for Ion Torrent sequencing (as per Mascher et al. (2013)). Single-end sequencing was performed either on an Illumina HiSeq 2000 at the McGill University-Génome Québec Innovation Center in Montreal, Canada, or on an Ion Proton machine at the Institut de Biologie Intégrative et des Systèmes (IBIS) of Université Laval, Quebec, Canada. A total of 42 million 100-bp reads were generated on the Illumina platform and 38 million 50- to 135-bp reads were obtained on the Ion Torrent platform. All data (GBS and WGS) are available in NCBI Sequence Read Archive (SRA) with the SRP Study accession, SRP059747 (Illumina sequences) and SRP073237 (Ion Torrent sequences).

IV.4.2 GBS analysis pipelines

We used two *de novo* variant callers and five reference-based pipelines (Williams82 reference genome; (Schmutz et al. 2010)) to call SNPs. We ran all pipelines in the same conditions of depth of coverage (minDP ≥2), maximum mismatch for alignment (n=3), Maximum Missing Data (MaxMD=80%), and Minimum Minor Allele Frequency (MinMAF ≥0.05). Below, we briefly describe the processes for each pipeline. For computation, we used a Linux system with 10 CPU and 25G of memory. In addition to the descriptions provided below, a summary of the different components of each pipeline is provided in Supplementary Table IV.1 and we provide all command lines used in this work as supporting information.

IV.4.2.1 Fast-GBS

The Fast-GBS analysis pipeline has been developed by integrating public packages with internally developed tools. The core functions include: (1) demultiplexing and cleaning of raw sequence reads; (2) read quality assessment and mapping; (3) filtering of mapped reads and estimation of library complexity; (4) re-alignment and local haplotype construction; (5) fit population frequencies and individual haplotypes; (5) raw variant calling; (6) variant and individual-level filtering; (7) identification of highly consistent variants. Since researchers may not always have immediate access to cluster resources, this pipeline allows either parallel processing of a large number of samples in a cluster or serial processing of multiple samples on a single machine.

IV.4.2.2 IGST (IBIS Genotyping-by-Sequencing Tool)

A pipeline implemented in Perl programming language was developed for the processing of Illumina sequence read data. The steps involved in the pipeline were executed in separate shell scripts. This pipeline uses different publicly available software tools (FASTX toolkit, BWA, SAMtools, VCFtools) as well as some in-house tools (Li & Durbin 2009; Li et al. 2009; Danecek et al. 2011). The raw SNPs obtained were further filtered using VCFtools based on read depth, missing data in genotypes and minor allele frequency. Heterozygous correction is performed by an in-house Python script.

IV.4.2.3 TASSEL-GBS (version 1 and 2)

TASSEL-GBS pipelines are implemented in Java programming language. Currently, two versions are available: TASSEL-GBS v1 (TASSEL 3.0) and TASSEL-GBS v2 (TASSEL 5.0). Both pipelines function in a similar manner and require that all reads be trimmed to an identical length (64 bp in v1, up to 92 bp in v2) and identical reads are collapsed into tags. These tags

are then aligned against the reference genome and SNPs are called from aligned tags. The main changes implemented in TASSEL-GBS v2 are: 1) the possibility to use longer tags to improve the accuracy of alignment to the reference genome and 2) an enhanced SNP discovery and production step.

IV.4.2.4 UNEAK (Universal Network Enabled Analysis Kit)

The general design of UNEAK is as follows: 1) reads are trimmed to 64 bp; 2) identical 64-bp reads are collapsed into tags; 3) pairwise alignment identifies tag pairs having a single base pair mismatch. These single base pair mismatches are candidate SNPs. A "network filter" is employed to discard repeats, paralogs and sequencing errors, resulting in a collection of reciprocal tag pairs, or SNPs.

IV.4.2.5 Stacks (reference-based and de novo)

The raw input data to Stacks are sequenced DNA fragments from any restriction enzyme–based GBS protocol. Stacks can handle raw sequencing data to identify loci *de novo* or via alignment against a reference genome. Regardless of whether the data are assembled *de novo*, or aligned against a reference genome, many subsequent steps in Stacks are shared. The pipeline can be described as follows: (1) Raw sequence reads are demultiplexed and cleaned (process_radtags). (2) Data from each individual are grouped into loci, and polymorphic nucleotide sites are identified (ustacks or pstacks for unaligned or aligned data, respectively). (3) Loci are grouped together across individuals and a catalogue is written (cstacks). (4) Loci from each individual are matched against the catalogue to determine the allelic state at each locus in each individual (sstacks). (5) Allelic states are either converted into a set of mappable genotypes (for a genetic map) using genotypes or subjected to population genetic statistics via populations, with the results being written in one or several output files.

IV.4.3 Genotype accuracy

For the estimation of the accuracy of genotype calls, we used an in-house script to compare the genotypes called using GBS with the genotypes called at the same loci following WGS. The sequencing and calling of SNPs in this collection of 24 soybean lines was previously described in Torkamaneh and Belzile (2015). Briefly, soybean lines were sequenced to a mean depth of coverage of 9x and a genome coverage of 96% was achieved. Illumina paired-end reads were aligned onto the soybean reference genome (Williams82) using BWA and the

genotypes at polymorphic loci were called using SAMtools. Variants with two or more alternative alleles were removed. A total of 3.6M SNPs were thus called among these lines. As a complementary means to measure genotype quality, we estimated the proportion of missing data and heterozygous calls produced with each analysis pipeline. For *de novo* pipelines, we aligned the tags supporting SNPs against reference genome to find the physical position and then we compared them with WGS dataset.

IV.5 Results

IV.5.1 Variant calling with different pipelines using Illumina read data

To assess the performance of different GBS analysis pipelines, we analyzed publicly available GBS data (100-bp Illumina reads) from a set of 24 previously studied soybean lines. We compared five reference-based analysis pipelines: TASSEL-GBS v1 and v2, Stacks, IGST, and Fast-GBS. We also compared two widely used *de novo* variant callers: UNEAK and Stacks. We used the same number of reads for all analyses (42M reads) and attempted to select parameters that would be as similar as possible for all the pipelines (see M&M for details). As shown in Table IV.1, large differences in the number of SNPs called were seen with both *de novo* and reference-based pipelines. Among the former, Stacks called the fewest SNPs, ~2 fold fewer than UNEAK (13,303 vs 24,743). The number of SNPs called by UNEAK was not too far below the mean number of SNPs called by reference-based pipelines (32,423). Among reference-based pipelines, the number of SNPs called varied between 18,941 (Stacks) and 54,412 (TASSEL-GBS v1), a 2.8-fold difference. The other three reference-based pipelines were much closer to the mean, calling between roughly 25k and 35k SNPs. In addition to calling SNPs, IGST and Fast-GBS were also able to call indels. In both cases, these contributed an extra 12-13% to the tally of variants.

Fast-GBS and TASSEL-GBS v1 proved to be the fastest running among the reference-based pipelines (~1h45), whereas IGST proved the slowest, requiring almost 13h to complete the analysis. Among *de novo* pipelines, UNEAK was almost three times faster than Stacks (1h11 vs 3h07) and proved the fastest of all pipelines. In terms of memory required, here also, very large differences were observed. Among *de novo* pipelines, UNEAK required almost three times as much disk space compared to Stacks (20 Gb vs 7 Gb). Among the reference-based pipelines, the differences were even greater as IGST required 17.1-fold more memory (240 Gb) than Stacks (14 Gb).

IV.5.2 Accuracy and efficacy of GBS bioinformatics pipelines

To examine the quality of the SNP data obtained using reference-based pipelines, we first measured the amount of missing data and then estimated genotype accuracy by comparing the GBS-derived genotypes with the true genotypes uncovered through whole-genome resequencing of the same lines. Assessments of the accuracy of GBS-called SNPs were performed on all SNPs for all pipelines at the same levels of tolerance for missing data (≤80%) and minor allele frequency (≥0.05). As can be seen in Table IV.2, among reference-based pipelines, the proportion of missing data varied from as little as 28% (TASSEL GBS v1) to as much as 57.3% (Stacks). Among the *de novo* pipelines, the proportion of missing data was less variable, ranging from 39.4% (Stacks) to 41.3% (UNEAK).

When we compared the genotypes obtained using each pipeline with the genotypes derived from resequencing, we found that 98.7% of SNP genotypes called using the Fast-GBS pipeline matched the true genotypes. Similar levels of accuracy were found for SNPs called with IGST (98.4%). With a single exception, all reference-based pipelines achieved levels of accuracy >92%. TASSEL-GBS v1 proved the least accurate of these pipelines, as only 76.1% of the genotypes it called were identical to the resequencing data. Among *de novo* pipelines, the accuracy of genotype calls was only slightly lower (93.7%, on average) than that obtained with the reference-based pipelines other than TASSEL-GBS v1 (95.6%, on average).

Among plants, recent or ancient polyploidization events can generate paralogs that can be mistaken to represent alleles of a single locus based on short sequence reads. We therefore examined both the overall number of heterozygous genotype calls and the number of loci containing a large proportion (>50%) of heterozygous calls. As can be seen in Table IV.2, *de novo* pipelines called a similar proportion of heterozygous genotypes (~3.7 and 5.3% for Stacks and UNEAK, respectively), and did not retain any loci with a large proportion of heterozygotes. Among reference-based pipelines, Fast-GBS and TASSEL-GBS v1 called the fewest and the most heterozygous genotypes (3.4 and 11.5%, respectively). Additionally TASSEL-GBS v1 called the largest number of loci with a large proportion of heterozygous genotypes (1125), while Stacks only called 65 loci with more than 50% heterozygotes.

IV.5.3 Overlap between SNP catalogues

We then determined the degree of overlap between the SNP catalogues obtained using the different pipelines and their accuracy. We selected Fast-GBS as the basis for comparison because of its ability to very accurately call a large number of SNPs. As demonstrated in Table

IV.3, among reference-based pipelines, the most overlap was observed between Fast-GBS and Stacks (>96%), and 92% of SNPs called with IGST were also found in the Fast-GBS dataset. In contrast, TASSEL-GBS v1 showed the lowest overlap (36.7%) with Fast-GBS. The *de novo* pipelines showed similar levels of overlap with Fast-GBS (Stacks= 89.1% and UNEAK= 87.5%). In an additional analysis (not shown in Table IV.3), we measured the overlap between the two *de novo* pipelines; around 67% of SNPs called by Stacks were also found in the UNEAK dataset. These two *de novo* pipelines therefore seem to identify fairly distinct subsets of the more extensive SNP catalog obtained using Fast-GBS.

To gain a deeper understanding of the genotypic accuracy among different subsets of shared or unique SNPs, we prepared two separate Venn diagrams, each comprising only four pipelines (for clarity), with Fast-GBS included in both panels (Figure IV.1). What stands out in this figure is that SNPs called by more than one pipeline were typically highly accurate (weighted mean accuracy = 94.8%). In contrast, with the sole exception of Fast-GBS, SNPs called by a single pipeline were typically much less accurate (weighted mean accuracy = 66.3%). Most strikingly, we note that TASSEL-GBS v1 called a very large number of unique SNPs (over 30,000) that show a low accuracy (65%). Unique SNPs called by other pipelines also typically showed low accuracy but were far fewer in number and thus had less impact overall.

IV.5.4 Reasons for poor performance of some pipelines

Given the observed variation in the number of called SNPs and their accuracy, we chose to investigate the causes of erroneous calls. To conduct this investigation, we followed a systematic approach illustrated in Figure IV.2. We divided the catalogue of SNPs in two categories, accurate and inaccurate, based on the comparison of the GBS-derived calls and the calls resulting from WGS. Inaccurate SNPs were then classified as being either unique to a single pipeline or shared between at least two pipelines. To investigate unique "weaknesses" of pipelines, we focused our attention on unique inaccurate SNPs. The first step in this investigation was to classify these inaccurate SNPs as being supported by reads mapping to a unique position in the genome or by reads mapping to multiple positions. In the first case, genotyping errors were attributed to a fault by the variant caller (e.g. due to sequencing or PCR amplification errors). In the second case, we reasoned that the mapping of reads to more than one location in the genome could result from these reads originating from either paralogues or repetitive regions. To resolve this, we mapped the reads against the masked reference genome v1.1 to estimate the proportion of inaccurate SNPs originating from repetitive regions. Means that repetitive parts of the reference genome are hidden away (turned into n's), so they won't be aligned to. In this reference genome 29.1% of the sequence

have been masked. SNPs that were no longer present in the catalogue derived from mapping to the masked reference genome were taken to be due to repetitive sequences. The remaining reads that successfully mapped to multiple sites in the masked reference genome were analyzed via a BLAST search to detect paralogy. A read was deemed to derive from a paralogue when we encountered at least 2 hits with 100% coverage and minimum of 96% identity. On average, we found 2.4 hits per read deemed to originate from paralogous loci defined in this fashion.

The results of this analysis are shown in Table IV.4. As most pipelines provided a largely accurate (>92%) set of SNPs, only a few hundred unique inaccurate SNPs were called by each pipeline with the sole exception of TASSEL-GBS v1 (9,828 unique inaccurate SNPs). A minority (11.5 to 29.7%) of the unique inaccurate SNPs were supported by reads mapping to a single position in the genome and deemed to result from an error in variant calling. The majority (70.3 to 88.5%) of inaccurate SNPs were supported by reads mapping to more than one region in the genome. Among these, the vast majority were due to reads mapping to paralogous regions (74 to 93%). We therefore conclude that most genotyping errors in soybean could be attributed to the presence of paralogs and that TASSEL-GBS v1 proved to be, by far, the pipeline most subject to making erroneous calls because of this.

Another result that begged investigation was the relatively low number of SNPs called by Stacks, as both *de novo* and reference-based versions of Stacks had called the fewest SNPs. We investigated the efficacy of the demultiplexing step as this had already been described as problematic. In our analyses, we found that 19.7% of Illumina reads failed to be assigned to a specific barcode file, a number that is much higher than that seen with the other pipelines. To measure the impact of such a decrease in the number of reads available to call SNPs, we used an alternative demultiplexing tool (Sabre), instead of the one provided in Stacks. The proportion of missing reads decreased to ~2% and the number of SNPs called using this more extensive set of reads increased by 12 and 24% (21,456 and 17,342) for Stacks reference-based and Stacks *de novo*, respectively. We conclude that the poor performance of the Stacks demultiplexing tool is an important contributor to the decreased number of SNPs called by Stacks.

IV.5.5 GBS using different sequencing platforms

To compare SNP calling using different sequencing technologies, we performed GBS on the same 24 soybean samples on an Ion Torrent platform. In contrast to Illumina reads that are all exactly the same length (100 bp), Ion Torrent reads varied in length from 50 to 135 bp.

In this analysis, we used only two reference-based pipelines that had performed best in the tests described above (Fast-GBS and TASSEL-GBS v2) using 38 million Ion Torrent reads. As seen in Table IV.5, the number of SNPs called with each pipeline at the same levels of tolerance for missing data (≤80%) and minor allele frequency (≥0.05) was highly similar (~23K in both cases). As above, Fast-GBS called a greater number of variants as it called a total of over 2,000 indels in addition to the SNPs. In terms of computing time, Fast-GBS was more than two-fold faster than TASSEL-GBS v2 (1h31 vs 3h29), while it used 15% more disk space (20 Gb vs 17 Gb).

In a second analysis, we measured the amount of missing data and estimated the accuracy of genotypes both by comparing GBS-called genotypes to the ones obtained through resequencing and by assessing the amount of heterozygosity in these lines that are presumed homozygous. As can be seen in Table IV.6, the proportion of missing data was relatively similar for the two pipelines (37% vs 33%). In this analysis, TASSEL-GBS v2 called more heterozygous genotypes than Fast-GBS (6.6% vs 4.5%). Also, TASSEL-GBS v2 called many more loci with a large proportion (>50%) of heterozygous genotypes than Fast-GBS (4,831 vs 861). In this analysis, Fast-GBS again achieved the highest accuracy in calling genotypes (95.2%), compared to 91.1% using TASSEL-GBS v2.

Finally, we compared the overlap among SNP catalogues obtained using the two sequencing platforms (Illumina vs Ion Torrent). As illustrated in Figure IV.3, when using Fast-GBS, we found that 69% (16,416 of 23,792 SNPs) of the SNPs derived from Ion Torrent reads were also present in the catalogue of SNPs obtained using Illumina reads. Conversely, of all the SNPs called using Illumina reads (34,953 SNPs), 47% were in common with the Ion Torrent catalogue. Using TASSEL-GBS v2, a slightly lower proportion (54%) (12,377 of 22,921 SNPs) of SNPs called from Ion Torrent reads were also obtained using Illumina reads. Conversely, a similar proportion (44%) of SNPs called using Illumina reads were in common with those called using the Ion Torrent reads. We found that using Ion Torrent reads leads more inaccurate SNPs compared to Illumina reads. Using Illumina reads only 23.7% and 12.9% of inaccurate SNPs called by TASSEL-GBS v2 and Fast-GBS had unique position, while using Ion Torrent reads this proportion increased to 76% and 87% for TASSEL-GBS v2 and Fast-GBS, respectively. This result suggested the higher level of sequencing error for Ion Torrent reads compared to Illumina. On the other hand, proportion of inaccurate SNPs with origin of paralogy and repetitive regions were similar for both of two sequencing technologies.

In conclusion, the amount of overlap across sequencing platforms was similar using both pipelines but much lower than the overlap seen across pipelines using the same sequencing platform.

IV.6 Discussion

The flexibility and low cost of genotyping methods relying on NGS make these excellent tools for many applications and research questions in genetics, breeding, and biodiversity (Baird et al. 2008; Elshire et al. 2011; Sehgal et al. 2015; Truong et al. 2012; Poland et al. 2012). Currently, GBS appears to be favored in the agricultural sciences (plant and animal breeding) whereas RAD-Seq seems to be the more prevalent approach in the field of ecology (Davey et al. 2011). Whatever library preparation approach is chosen to achieve complexity reduction prior to sequencing, bioinformatics must be used to extract useful information on SNP loci and genotypes from a vast amount of short sequence reads (Davey et al. 2011; McCormack et al. 2013). It is at this stage that the choice of an analytical method will have the greatest impact on the amount and quality of the resulting genotypic information. Unfortunately, to date, few studies have systematically compared SNP-calling pipelines for GBS and compared their efficiency, accuracy and degree of overlap.

The first question that arises concerns the use of *de novo* vs reference-based methods. In the absence of a reference genome, there is little choice but to use one of the two currently widespread tools, UNEAK and Stacks. Although they use different algorithms to do so, these two pipelines are conceptually similar in that they seek to first establish catalogues of identical reads and then to search for highly related reads that are potentially alleles at the same locus. Under the conditions used in this work, UNEAK greatly outperformed Stacks in that it generated 82% more SNPs (~25k vs ~13k). From a qualitative perspective, both *de novo* pipelines performed similarly well in terms of missing data (~40%) and genotypic accuracy (~94%). This is comparable to the results reported by Lu et al. (2013) in maize where it was estimated that 92% of genotype calls were accurate and that this proportion could be increased to 96.2% by filtering for SNPs with a MAF > 0.3 in a segregating biparental population. Both *de novo* pipelines can be run quite quickly and are relatively conservative in their SNP calls resulting in a dataset of high quality. Thus, for the vast majority of species for which no reference genome is available currently or in the foreseeable future, the *de novo* SNP calling tools perform extremely well in terms of accuracy, but UNEAK will yield almost two-fold more SNPs.

The picture painted of the performance of *de novo* pipelines in this comparison may be too rosy, however. Indeed, for the sake of uniformity, we used the same filtering options

(MinMAF≥0.05, MaxMD=80%, and minDP≥2) for both *de novo* and reference-based pipelines. But this high tolerance towards missing data may not be realistic in the case of *de novo* pipelines. We have shown previously that missing data imputation is very efficient and accurate on a dense set of SNPs obtained using a reference-based pipeline (Torkamaneh & Belzile 2015). In the case of *de novo* pipelines, in the absence of positional information on the different SNPs and the haplotype structure, imputation is much more challenging. For this reason, most users of *de novo* pipelines will set a lower ceiling for the maximal amount of missing data, typically between 20% and 50% at most (Lu et al. 2013; Mascher et al. 2013; Larson et al. 2014). With the GBS sequence data used in this work, tolerating up to 20% of missing data substantially decreases the number of SNPs that can be called using both *de novo* pipelines (~5k SNPs; data not shown). Under these more realistic conditions (in view of the necessary imputation of missing data), we find that reference-based pipelines yielded about 5- to 7-fold more high-quality SNP markers (~5k vs 25k to 35k markers).

Given the increasing availability of reference genomes in economically important crops and animals, we then need to ask which of the available reference-based pipelines produces the best catalogue of SNPs both in terms of abundance of markers and their accuracy. Among the five reference-based pipelines, Fast-GBS can be run quickly, resulted in the highest genotyping accuracy for a very large number of SNP loci (close to 35,000) in addition to almost 4,000 indels. Based on these considerations, it seems to be the pipeline of choice, at least in the case of soybean and likely also for other species with similar genomic and reproductive characteristics.

Of the pipelines tested, TASSEL-GBSv1 stood out from the rest of the group in terms of the number of SNP loci called (50-100% more than the others), but this came at the cost of accuracy as it was the only pipeline whose genotypic calls were accurate in less than 90% of cases (76.1%). As it is not easy to distinguish true from false genotypes, we would argue that TASSEL-GBSv1 is insufficiently accurate to be used on its own. In previous work, the large resulting catalogue of SNPs was often "filtered" by discarding markers that did not behave as expected in a segregating population (Elshire et al. 2011). This presumably helped to discard "false" markers that resulted from confounding alleles (at a single locus) and reads derived from paralogous loci. We hypothesized that the main reason for this decreased accuracy is the fact that TASSEL-GBSv1 clips all reads to a uniform length of 64 bases, thus producing short tags that are at increased risk of mapping to multiple or erroneous locations. Pipelines using longer reads did not exhibit this problem and typically had at least 10-fold fewer reads mapping to multiple locations. For example, despite sharing much in common with TASSEL-

GBS v1, when TASSEL-GBS v2 was run under conditions that allow for longer tags (92 bases in our case), the reliability of the genotypes increased considerably.

The reference-based version of Stacks is the other pipeline that stood out in that it called much fewer SNPs than the others. In investigating the different steps needed to go from sequences to SNPs, we found that Stacks lost ~20% of reads at the demultiplexing step, i.e. some barcoded reads were not attributed to a sample and were simply discarded from the ensuing steps. This obviously resulted in a concomitant decrease in the number of SNPs called (~19k vs ~25k). This poor performance of the Stacks demultiplexing step has been previously reported by Herten et al. (Herten et al. 2015).

In our view, the genome-wide measurement of the accuracy of GBS datasets derived from different bioinformatics pipelines represents an important and key contribution of this work. It was assessed by comparing directly to whole genome resequencing data. In many previous studies, estimates of genotypic accuracy were often achieved by indirect measurement (Lu et al. 2013) or performed on a very small subset of SNP loci (Sonah et al. 2013). Typically, levels of genotype accuracy ranging between 92 and 98% have been reported with slight differences being observed between species and types of population (Sonah et al. 2013; Lu et al. 2013; Mascher et al. 2013). The advantage of using resequencing data in this fashion is that we can directly assess the accuracy of GBS data yielded by different pipelines.

Another important consideration is whether the SNP catalogues produced using different pipelines and different sequencing technologies are concordant. When using a single sequencing technology (Illumina), we found that ~80% or more of SNPs called by most pipelines were also present in the SNP catalogue derived from Fast-GBS. Thus, these pipelines largely agree on the loci that are polymorphic within a given set of germplasm. The only exception was TASSEL-GBS v1, as, only a quarter of the SNPs present in the resulting catalogue was also present in the set derived using Fast-GBS. This is likely due to the shorter sequences used (only 64 bp) and a large number of "false" SNPs as this pipeline proved the least accurate of all. When using the same pipeline to analyze data derived from two sequencing technologies (Illumina and Ion Torrent), we typically found that the overlap between SNP catalogues varied between roughly 50 and 70%. Thus, the choice of sequencing technology used resulted in a greater variability in the catalogue of SNPs produced than did the choice of pipeline used on a single set of reads. At first glance, this would seem to contradict the conclusions drawn by Mascher et al. (2013) who found that the SNP catalogues produced using two pipelines (TASSEL-GBS v1 and SAMtools) differed more than the catalogues obtained using different sequencing technologies (Illumina and Ion Torrent)

(Mascher et al. 2013). In our view, this is more a reflection of the limitations of TASSEL-GBS v1 (due to its short tags). When we consider a broader array of reference-based pipelines, these generally provide a very good overlap in SNP loci uncovered.

IV.7 Conclusion

The conclusions drawn from this work are likely to extend to other organisms sharing similar genomic features (medium-sized genome, diploid). It can be anticipated that species having experienced recent whole genome duplication events will represent a greater challenge as the risk of confounding alleles at the same locus and paralogs will likely increase in such cases. In species where such events occurred in the more distant past, there will have been more opportunity for paralogs to diverge, thus facilitating the correct mapping of reads.

As such, it is impossible to devise a single pipeline that will be equally suited to every situation. This is where it becomes important for users to be able to change various parameters in the SNP calling process. Unfortunately, not all pipelines are equally "transparent" in this regard and offer the same opportunity to be altered. At one end of the spectrum, UNEAK and TASSEL-GBS offer very good performance, but rely on some purpose-built tools or algorithms that a user cannot easily alter (e.g. for demultiplexing and variant calling). Also, the intermediate data files are not always easily accessible and this makes it more difficult to investigate specific problems. At the other end of the spectrum, IGST and Fast-GBS string together a set of existing tools for which the user can alter parameters/options at will, and the intermediate files are easily accessible. In this spectrum, in our view, Stacks offers an intermediate level of transparency.

Finally, although whole-genome sequencing of entire populations is rapidly approaching, we believe that the methods described here are likely to remain invaluable for years to come in population genomics, breeding, mapping studies and reference genome sequence assembly, particularly for non-model organisms.

IV.8 Tables

Table IV.1. Number of SNPs and indels detected among 24 soybean lines using seven different bioinformatics pipelines on Illumina reads. The time and amount of memory needed to run each pipeline are also provided.

		Variants			
Approach	**Pipeline**	**SNPs**	**Indels**	**Time* (h:m)**	**Memory (Gb)**
de novo	Stacks	13,303	ND	3:07	7
	UNEAK	24,743	ND	1:11	20
Reference - based	TASSEL-GBSv1	54,412	ND	1:45	15
	Stacks	18,941	ND	3:30	14
	IGST	25,650	3,170	12:59	240
	TASSEL-GBSv2	28,158	ND	4:16	18
	Fast-GBS	34,953	3,921	1:47	27

***** Using a Linux system with 10 CPU and 25G of memory

Table IV.2. Accuracy of GBS SNP data derived from Illumina platform using different bioinformatics pipeline.

Approach	***de novo***		**Reference-based**				
Parameter/Pipeline	**Stacks**	**UNEAK**	**TASSEL-GBS v1**	**Stacks**	**IGST**	**TASSEL-GBS v2**	**Fast-GBS**
Number of SNPs	13,303	24,743	54,412	18,941	25,650	28,158	34,953
Number of genotypes	319,272	593,832	1,305,888	454,584	615,600	675,792	838,872
Missing data (%)	41.3	39.4	28	57.3	44	35.6	46
Heterozygotes (%)	3.7	5.3	11.5	4.4	5.9	5.7	3.4
Loci with >50% heterozygotes*	0	0	1125	65	324	551	184
Accuracy (%)	93.6	93.9	76.1	93.2	98.4	92.3	98.7

*These were eliminated from the final catalogue used to estimate accuracy

Table IV.3. Degree of overlap among SNP loci called using Fast-GBS and six other bioinformatics pipelines.

		SNPs			
Approach	**Pipeline**	**Total**	**Common (in %)**	**Other pipeline only**	**Fast-GBS only**
de novo	Stacks	13,303	89.1	1,450	23,100
	UNEAK	24,743	87.5	3,172	13,382
Reference-based	TASSEL-GBS v1	54,412	36.7	34,420	14,961
	Stacks	18,941	96.2	1,709	16,721
	IGST	25,650	92.4	1,950	11,253
	TASSEL-GBS v2	28,158	88.3	3,295	10,090

Table IV.4. Number and characteristics of unique inaccurate SNPs called by different pipelines.

Approach	***de novo***		**Reference-based**				
Pipeline	Stacks	UNEAK	TASSEL GBS v1	Stacks	IGST	TASSEL GBS v2	Fast-GBS
Unique inaccurate SNPs	495 (3.7% of 13.303)	533 (2.2% of 24,743)	9,828 (18.1% of 54,412)	103 (0.5% of 18,941)	207 (0.8% of 25,650)	558 (2.0% of 28,158)	272 (0.8% of 34,953)
Inaccurate SNPs with unique position (%)	146 (29.7)	72 (13.5)	1,126 (11.5)	20 (19.4)	46 (22.2)	132 (23.7)	35 (12.9)
Inaccurate SNPs with multiple positions (%)	349 (70.3)	461 (86.5)	8,702 (88.5)	83 (80.6)	161 (77.8)	426 (76.3)	237 (87.1)
Repetitive region (%)	45 (13)	120 (26)	1,828 (21)	9 (11)	15 (9)	60 (14)	17 (7)
Paralogues (%)	304 (87)	341 (74)	6875 (79)	74 (89)	146 (91)	366 (86)	220 (93)

Table IV.5. Number of SNPs and indels detected among 24 soybean lines using Ion Torrent reads and two different bioinformatics pipelines.

		Variants			
Approach	**Pipeline**	**SNP**	**Indels**	**Time* (h:m)**	**Memory (Gb)**
Reference-based	TASSEL-GBSv2	22,921	ND	3:29	17
	Fast-GBS	23,792	2,054	1:31	20

***** Using a Linux system with 10 CPU and 25G of memory

Table IV.6. Accuracy of SNP data derived using Ion Torrent reads and two different bioinformatics pipelines.

Stat type/Pipeline	TASSEL-GBSv2	Fast-GBS
Number of SNPs	22,921	23,792
Missing data (%)	37	33
Loci with >50% heterozygotes*	4,831	861
Residual heterozygotes (%)	6.6	4.5
Accuracy (%)	91.1	95.2

*These were eliminated from the final catalogue used to estimate accuracy

IV.9 Figures

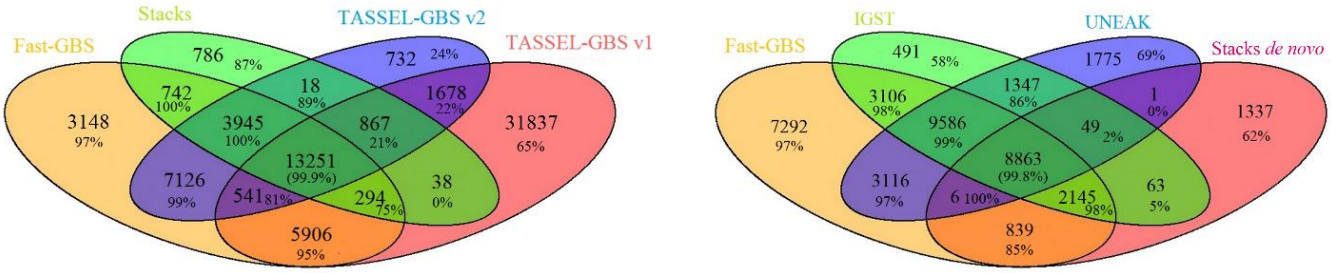


Figure IV.1. Venn diagram representing the degree of overlap among SNP loci called using seven bioinformatics pipelines. The percentages showed estimated accuracy for all groups of SNPs (unique and shared).

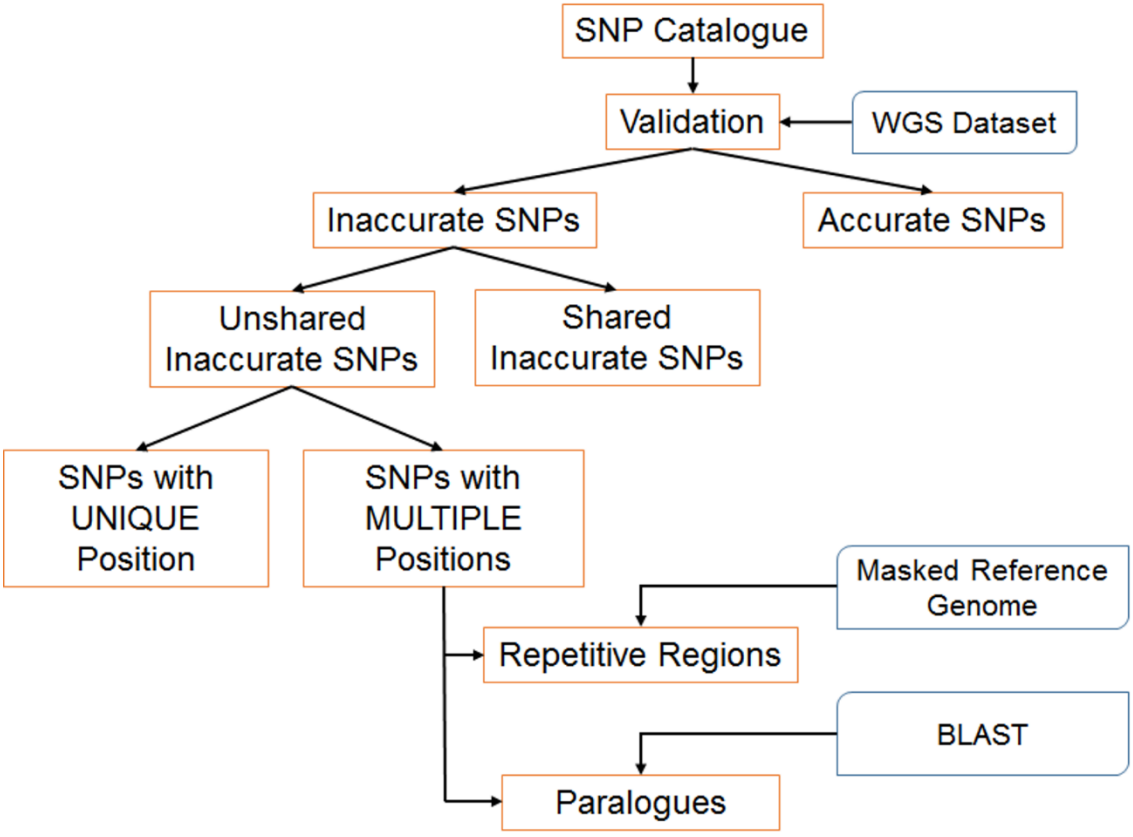


Figure IV.2. Systematic approach used to investigate the possible causes of unique inaccurate SNP calls.

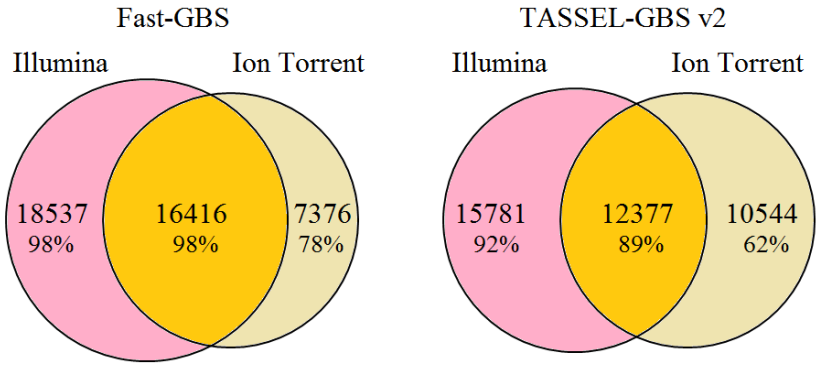


Figure IV.3. Venn diagram for overlap of the SNPs called using two different bioinformatics pipelines (a) Overlap of SNPs called with Fast-GBS using Illumina and Ion Torrent reads. (b) Overlap of SNPs called with TASSEL-GBS v2 using Illumina and Ion Torrent reads.

IV.10 Supplementary files

Supplementary files listed and described below can be found online at

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0161333#sec018

Additional file IV.1 Supplementary Table. Summary of five reference based GBS pipelines.

Additional file IV.1 Supplementary Text. Command lines for seven pipelines used in this study.

Chapitre V

Scanning and filling: ultra-dense SNP genotyping combining genotyping-by-sequencing, SNP array and whole-genome resequencing data

Davoud Torkamaneh and Francois Belzile

Département de Phytologie and Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Quebec City, QC, Canada G1V 0A6

PLoS ONE

2015, 10(7): e0131533

V.1 Résumé

Le génotypage par séquençage (GBS) représente une approche de génotypage à haut débit hautement rentable. Par nature, cependant, GBS est soumise de générer des quantités importantes de données manquantes et celles-ci devront être imputées pour de nombreuses analyses en aval. La mesure dans laquelle ces données manquantes peuvent être tolérées lors de l'appel des SNP n'a pas été largement explorée. Dans ce travail, nous explorons d'abord l'utilisation de l'imputation pour compléter les génotypes manquants dans les ensembles de données GBS. Il est important de noter que nous utilisons des données de re-séquençage du génome complet pour évaluer l'exactitude des données imputées. À l'aide d'un panel de 301 accessions de soja, nous montrons que plus de 62 000 SNP peuvent être appelés lorsqu'ils tolèrent jusqu'à 80% de données manquantes, une augmentation de cinq fois par rapport au nombre appelé tolérant jusqu'à 20% de données manquantes. À tous les niveaux de données manquantes examinées (entre 20% et 80%), les jeux de données SNP résultants étaient d'une précision uniformément élevée (96 à 98%). Nous avons ensuite utilisé l'imputation pour combiner des ensembles de données SNP complémentaires dérivés de GBS et une puce de SNP (SoySNP50K). Nous avons donc produit un ensemble de données amélioré de>100 000 SNP et les génotypes dans les loci qui était précédemment absent ont encore été imputés avec un haut niveau de précision (95%). Sur les 4 000 000 de SNP identifiés par re-séquençage 23 accessions (parmi les 301 utilisés dans l'analyse GBS), 1,4 million de tags SNP ont été utilisés comme référence pour imputer ce grand ensemble de SNP sur l'ensemble du panel de 301 accessions. Ces loci précédemment absent pourraient être imputés avec une précision d'environ 90%. Enfin, nous avons utilisé l'ensemble de données SNP 100K (GBS + SoySNP50K) pour effectuer un GWAS sur la teneur en huile de graines dans cette collection d'accessions de soja. Le nombre d'associations importantes de marqueurs-caractères et les niveaux de signification maximale ont été considérablement améliorés en utilisant ce catalogue amélioré de SNP par rapport à un catalogue plus petit résultant de GBS seul à ≤20% de données manquantes. Nos résultats démontrent que l'imputation peut être utilisée pour remplir à la fois les génotypes manquants et les loci absent avec une précision très élevée et que cela ment à des analyses génétiques plus puissantes.

V.2 Abstract

Genotyping-by-sequencing (GBS) represents a highly cost-effective high-throughput genotyping approach. By nature, however, GBS is subject to generating sizeable amounts of missing data and these will need to be imputed for many downstream analyses. The extent to which such missing data can be tolerated in calling SNPs has not been explored widely. In this work, we first explore the use of imputation to fill in missing genotypes in GBS datasets. Importantly, we use whole genome resequencing data to assess the accuracy of the imputed data. Using a panel of 301 soybean accessions, we show that over 62,000 SNPs could be called when tolerating up to 80% missing data, a five-fold increase over the number called when tolerating up to 20% missing data. At all levels of missing data examined (between 20% and 80%), the resulting SNP datasets were of uniformly high accuracy (96-98%). We then used imputation to combine complementary SNP datasets derived from GBS and a SNP array (SoySNP50K). We thus produced an enhanced dataset of >100,000 SNPs and the genotypes at the previously untyped loci were again imputed with a high level of accuracy (95%). Of the >4,000,000 SNPs identified through resequencing 23 accessions (among the 301 used in the GBS analysis), 1.4 million tag SNPs were used as a reference to impute this large set of SNPs on the entire panel of 301 accessions. These previously untyped loci could be imputed with around 90% accuracy. Finally, we used the 100K SNP dataset (GBS + SoySNP50K) to perform a GWAS on seed oil content within this collection of soybean accessions. Both the number of significant marker-trait associations and the peak significance levels were improved considerably using this enhanced catalog of SNPs relative to a smaller catalog resulting from GBS alone at ≤20% missing data. Our results demonstrate that imputation can be used to fill in both missing genotypes and untyped loci with very high accuracy and that this leads to more powerful genetic analyses.

V.3 Introduction

Next generation sequencing (NGS) has revolutionized plant and animal research in many ways. Firstly, it has allowed researchers to decode the whole genome of many organisms. Currently, hundreds of eukaryotic genomes have been sequenced (NCBI, "www.ncbi.nlm.nih.gov/projects/WGS/WGSprojectlist.cgi") and, for some species, numerous individuals, cultivars or accessions of the same species have also been sequenced (Huang et al. 2012; Aflitos et al. 2014; Daetwyler et al. 2014). Next generation sequencing has also facilitated greatly the development of methods to genotype very large numbers of molecular markers such as single nucleotide polymorphisms (SNPs). In one such approach, large-scale sequencing has allowed researchers to probe nucleotide diversity in panels of individuals to discover polymorphic sites and then to develop genotyping arrays ("SNP chips") that can subsequently be used to determine the genotype of an individual line at thousands to millions of such SNPs (Kumar et al. 2012). In soybean, an example of this approach is the SoySNP50K array that was constructed to interrogate over 52K SNPs of which 47,337 were found to be polymorphic among a set of 288 elite cultivars, landraces and wild soybean accessions (Ha et al. 2014). Alternatively, genotyping methods exploiting the power of NGS technologies have also been developed to simultaneously identify and genotype SNPs. RAD-Seq and genotyping-by-sequencing (GBS) are two examples of such SNP genotyping approaches relying on NGS (Song et al. 2013; Davey et al. 2011).

In soybean, GBS has been developed as a rapid and robust approach for reduced-representation sequencing of multiplexed samples that combines genome-wide molecular marker discovery and genotyping (Donato et al. 2013). The flexibility and low cost of GBS makes this an excellent tool for many applications and research questions in genetics and breeding. Such modern advances allow for the genotyping of thousands of SNPs, and, in doing so, the probability of identifying SNPs correlated with traits of interest increases (Sonah et al. 2013). However, when using approaches such as GBS that perform a scan or a sampling of the genome, the quantity of missing data can be substantial. An important question that remains unanswered at this point is the degree to which missing data can be tolerated and to what extent they affect the accuracy of the imputation process.

Conceptually, there are two types of missing data in large datasets. The most obvious is when some individuals are missing a genotype value at a locus that is otherwise successfully typed in the other individuals of a population. In another situation, which arises when different datasets (e.g. obtained using different genotyping technologies) are combined, there can be loci that are not typed at all within a population, i.e. there is no information for a SNP locus in all individuals of the population except for a few individuals that can be common to both

datasets. The first type of missing data can be termed a "missing genotype" while the second is termed an "untyped locus". There has been considerable interest in imputing such missing data based on the available data (Li et al. 2009). Many tools used in genetic analysis require complete datasets and there are thus two possibilities: work only with SNP loci devoid of any missing data (thereby considerably reducing the number of SNPs available) or impute these missing data through various strategies.

Imputation is the substitution of some value for missing data, in other words, 'filling in' missing data with plausible values. Generally, methods of genotype imputation are based on the concept that SNPs close together on a chromosome are often inherited together. The resulting correlations among SNPs are referred to as linkage disequilibrium (LD), or association, in the genetic literature (Ardlie et al. 2002). Many methods for imputing missing genotypes have been suggested and tested. Generally, two methodological classes are considered: regression and phasing.

A first approach is to use regression models to impute the missing genotypes by using flanking SNPs as covariates (Li et al. 2009). Regression-based methods face a common problem in variable selection; it can be difficult to select which available SNPs should be included as covariates. One reason for this is that LD patterns are not homogenous across the genome (Ardlie et al. 2002); for example, lower LD would be expected among SNPs located in recombination hotspots than those in low recombination regions (high LD regions). Therefore, fewer SNPs may be useful as covariates in lower LD regions. These limitations made regression methods less attractive and less accurate.

Phase-based methods consider haplotype structure and common descent patterns (Li et al. 2009). As humans, animals, and plants are (often) diploid, a genotype is the combination of maternal and paternal alleles. Alleles close together on a chromosome are typically inherited together in a whole unit as a haplotype. Phase-based algorithms try to split genotypes at SNPs into haplotypic phases. Here, a "phase" is simply an inferred parental haplotype. Once phased, missing alleles can be estimated from neighboring haplotype alleles through their LD relationship, and the inferred alleles are then combined to impute the missing genotype. Currently, many popular genotype imputation methods are phase-based.

In this work, we explored the accuracy and efficiency of different imputation tools for both the imputation of missing genotypes in the context of GBS and of untyped loci in the context of combining SNP datasets obtained through different genotyping approaches (GBS, SNP array and resequencing). Finally, we examined the impact of using such enhanced SNP datasets in genome-wide association analyses.

V.4 Materials and methods

V.4.1 Samples and SNP datasets

A set of 301 Canadian soybean lines was subjected to GBS analysis (with *Ape*KI digestion) and a total of 450 million 100-bp reads (~1.5M reads/line) were processed through our analytical pipeline that relies on SAMtools to call SNPs as described previously in Sonah et al. (2013) and Sonah et al. (2014). The SoySNP50K iSelect BeadChip (Song et al. 2013) has been used to genotype the USDA Soybean Germplasm Collection (Song et al. 2015). The complete dataset for 19,652 *G. max* and *G. soja* accessions genotyped with 42,508 SNPs are publicly available on Soybase (www.soybase.org). Of these 19,652 accessions, 25 were in common with the 301 Canadian lines used for GBS. Finally, on the basis of geographic distribution and genotypic diversity, we chose 23 soybean lines from the set of 301 mentioned above to undergo whole genome resequencing (described below).

V.4.2 DNA extraction and whole genome resequencing

Seeds were planted in individual two-inch pots containing a single Jiffy peat pellet (Gérard Bourbeau & fils inc. Quebec, Canada). First trifoliate leaves from 12 day-old plants were harvested and immediately frozen in liquid nitrogen. Frozen leaf tissue was ground using a Qiagen TissueLyser. DNA was extracted from approximately 100 mg of ground tissue using the Qiagen Plant DNeasy Mini Kit according to the manufacturer's protocol. DNA was quantified on a NanoDrop spectrophotometer. Illumina Paired-End libraries were constructed for DNA samples using the Illumina Tru-seq DNA Library Prep Kit (Illumina, San Diego CA, USA) following the manufacturer's instructions. DNA library quality was verified on an Agilent Bioanalyzer with a High Sensitivity DNA chip. Samples were sequenced using the Illumina HiSeq 2000 platform at the McGill University-Génome Québec Innovation Center in Montreal, QC, Canada.

V.4.3 Alignment and variant calling

Illumina paired-end reads were aligned using the Burrows-Wheeler Aligner (BWA) (Li & Durbin 2009) onto the soybean reference genome (Williams82) (Schmutz et al. 2010). Variants were called using SAMtools 0.1.18 (Li et al. 2009). BAM files were pooled for variant calling. Variants were then removed if they had two or more alternative alleles, no observation of the alternative allele on either forward or reverse reads, an overall quality (QUAL) score of <20,

a mapping quality (MQ) score <30, a read depth of <2, or were suspected of representing false heterozygotes (based on unequal read depth of the two alleles). For tag SNP selection, we used PLINK (Purcell et al. 2007) to calculate linkage disequilibrium (LD) between each pair of SNPs within a sliding window of 50 SNPs and we removed all but one SNP that were in perfect LD (LD=1); the remaining SNPs were deemed tag SNPs.

V.4.4 Imputation methods

We used three software tools to impute missing data: fastPHASE (Scheet et al. 2006), BEAGLE v4.0 (Browning & Browning 2007), and IMPUTE2 (Howie et al. 2009). As recommended by Delaneau and Marchini (2014) we used SHAPEIT2 (Delaneau et al. 2013) to first infer the haplotypes among the set of genotypes studied, and then used the resulting output to perform the imputation of untyped loci using IMPUTE2. All three software tools were used to impute missing genotypes while only the last two were used to impute untyped loci. The parameters for fastPHASE were: `fastPHASE –T 20 –E 10 –M 0 –o output_name fastPHASE_input_file`. The command line for BEAGLE read as follows for missing data imputation: `java --Xmx5000m --jar unphased=phased.input.bgl missing=0 niterations=10 out=out_file`, and for untyped genotype imputation: `java --Xmx5000m --jar phased=phased.input.bgl unphased=unphased.input.bgl markers=marker.ids missing=0 niterations=10 out=out_file`. Finally, the command line for IMPUTE2 was: `impute –h phased_file --l legend_file --g geno_file –m genetic_map_chr*.txt --call_-thresh 0.0 --Ne 11418 --i info_file –o out_file`. Finally, both BEAGLE and IMPUTE2 were used to assess the impact of the number of lines composing the reference panel on the accuracy of imputation at untyped loci.

V.4.5 Genotype accuracy

For the initial estimation of the accuracy of genotype calls in GBS analysis, we compared the called genotypes at all loci on a single chromosome (Gm03; 3326 SNP loci) for the 23 lines common to both the GBS and WGS datasets. These GBS-derived genotypes were directly compared with the true genotypes (revealed by WGS) using an in-house script. Similarly, all imputed genotype calls (initially missing data) on Gm03, following imputation (three imputation methods, as described above, at the different levels of MaxMD and MinMAF), were compared with the true genotypes (WGS). To verify that this chromosome was representative of the broader genome, we estimated the overall genotype accuracy (GBS-derived and

imputed SNPs) for all chromosomes (Gm01 to Gm20) using BEAGLE only and at MaxMD≤80% and MinMAF=0.003.

To assess the accuracy of imputation at untyped loci when combining GBS and SoySNP50K datasets, i.e. when the SoySNP50K data were used as a reference panel to impute genotypes at loci not common to both datasets we extracted the genotypes at all loci on chromosome Gm03 for three lines (Maple Presto, Mandarin, and Evans) for which WGS, GBS, and SoySNP50K data were available. Imputed SNP genotypes were compared with the true genotypes revealed by WGS.

Similarly, to assess the accuracy of imputation at untyped loci that were imputed using the WGS dataset, we used the WGS SNP data from 22 of the 23 resequenced lines as a reference panel to impute these SNPs onto the GBS or GBS + SoySNP50K data. The remaining line was kept for validation of the imputed SNPs. We performed three permutations where a single line was kept aside to estimate imputation accuracy (Supplementary Table V.3). We then extracted the genotypes at all loci on chromosome Gm03 for the remaining line and we directly compared with the true genotypes.

V.4.5 Genome-wide association study

A subset of 139 soybean lines were used in the GWAS analysis. Phenotypic data (seed oil content) for these lines were originally described by Sonah et al. (2014). All the analyses were performed using the Genomic Association and Prediction Integrated Tool (GAPIT) (Lipka et al. 2012). A general linear model (GLM) was used with or without the covariate P from principal component analysis (PCA) and a kinship matrix was calculated either using the VanRaden method (K) or the EMMA method (K*) to determine relatedness among individuals (Lipka et al. 2012). A multi-locus mixed model (MLMM) incorporating a kinship matrix (K or K*) along with a P or Q matrix was used to test for marker-trait association (Segura et al. 2012). The negative log(1/p) was used to establish a significance threshold (Wang et al. 2012; Yang et al. 2013).

V.5 Results

V.5.1 Factors that affect number of SNPs in GBS analysis

We first explored the impact of two key filtering steps central to the production of SNP catalogs derived from GBS analysis: the maximal amount of missing data allowed (MaxMD, in %) and

the minimal minor allele frequency (MinMAF). A set of 301 Canadian soybean lines was subjected to GBS analysis and a total of 450 million 100-bp reads (mean of ~1.5M reads/line) were processed through our analytical pipeline that calls SNPs using Samtools (see materials and methods for details). Using a minimum of one read to call a genotype, we obtained an initial catalog of 247,851 SNPs. We then filtered this set of SNPs for both MaxMD (between 0 and ≤80% missing data) and for MinMAF (0.003, 0.05 and 0.1). As can be seen in Figure V.1a, the amount of missing data allowed had a very large impact on the number of SNPs retained. At a MinMAF of 0.003 (i.e. a single line carrying a different allele among 301 lines), the number of SNPs increased steadily from only 1 (0% missing data) up to 62,643 (≤80% missing data). At the other MinMAF values, SNP numbers similarly increased markedly between 0 and 41,024 (MinMAF=0.05) and between 0 and 32,035 (MinMAF=0.1).

As the MaxMD filter only reflects the maximal proportion of missing data that are tolerated for an individual SNP marker to be retained, it does not accurately reflect the actual mean amount of missing data that characterizes a SNP dataset. To better capture this, we plotted the mean proportion of missing data at each of the MaxMD and MinMAF levels described above (Figure V.1b). As can be seen, the proportion of missing data in an entire dataset was hardly affected by the MinMAF threshold used but was heavily impacted by the chosen MaxMD level. Even at MaxMD of 80%, the mean amount of missing data was around 50%, while at more stringent MaxMD levels (e.g. 20%), the mean proportion of missing data became quite low (<10%).

We then examined the distribution of these SNPs based on the amount of missing data (in successive increments of 10%) at the most permissive MinMAF level (0.003). As can be seen in Figure V.1c, over 13,000 SNPs were called with >70% and ≤80% missing data, while around 7,000 were called with ≤10% missing data. Globally, approximately half of the SNPs could be called with ≤50% missing data while the other half were called with between 50% and 80% missing data. We therefore conclude that it is possible to quite significantly increase the number of called SNPs by allowing for more missing data, but this will only be attractive if these missing data can be accurately imputed.

V.5.2 Accuracy and efficacy of imputation for missing genotypes

To examine the quality of the SNP data obtained using GBS, we first assessed the accuracy of the SNP genotypes initially called by GBS, prior to any imputation. To achieve this, we performed whole-genome resequencing on a representative subset of 23 soybean lines at a

mean depth of coverage of 9x (genome coverage of 96%). A total of 3.6M SNPs were called among these lines and this dataset was presumed to represent the true genotype at variant positions. Assessments of the accuracy of called or imputed SNPs were performed on SNPs located on a single chromosome (Gm03) for all methods at different levels of MaxMD and MinMAF. At a MaxMD of 80% and MinMAF of 0.003, we found that 98.4% of SNP genotypes called by our GBS pipeline proved to be identical to the true genotypes. Similar levels of accuracy were found for called SNPs under all filtering conditions (data not shown).

In a second step, to estimate the accuracy of imputed SNP data (i.e. formerly missing genotypes), we performed imputation at all levels of MaxMD and MinMAF on the entire set of 301 lines. Once again, we used the resequencing data as a reference and, as shown in Figure V.2a and detailed in Table V.1, we found that imputation accuracy was hardly affected by the chosen minor allele frequency and only moderately affected by the amount of missing data. Somewhat surprisingly, the accuracy of imputation actually increased with increasing missing data. Indeed, while the imputation accuracy was 86% at MaxMD=20%, it rose steadily to reach 94% at MaxMD=80%. Therefore, allowing for a greater amount of missing data not only yielded a larger number of SNP markers, but this also proved beneficial in terms of the accuracy of imputed genotypes.

As illustrated above, the proportions of called and imputed SNP genotypes did vary at different MaxMD levels and thus impacted the overall accuracy of the resulting SNP catalog. The accuracy of the entire GBS-derived SNP dataset (after imputation) was measured and is illustrated in Figure V.2b and detailed in Table V.1. This includes both the SNP genotypes initially called and those resulting from imputation. Overall genotype accuracy ranged between 96% (MaxMD=80%) and 98% (MaxMD=20%), with hardly any impact of the MinMAF level. To determine if Gm03 was representative of the entire set of 20 chromosomes, we measured overall genotype accuracy for all chromosomes using a single imputation tool (BEAGLE) at a single level of MaxMD and MinMAF (80% and 0.003, respectively). The imputation accuracy differed very little between chromosomes, ranging between 95.3% and 96.3% (mean = 95.84% ± 0.28%).

Finally, although all three software tools performed equally well in terms of accuracy of imputation, computational speed varied considerably (Table V.1). Whereas it took fastPHASE 14h to impute the missing data, BEAGLE completed the task in only 30 minutes. In conclusion, we find that large amounts of missing data do not have a significant detrimental impact on the overall accuracy thanks to a highly accurate imputation.

V.5.3 Accuracy of imputation at untyped loci

The existence of multiple genotyping approaches offers the opportunity to exploit already existing haplotype information to further enhance marker density and to facilitate the integration of data obtained from different genotyping platforms. We first wanted to test whether the publicly available SoySNP50K array data obtained on 19,562 USDA soybean accessions could be used to impute additional ("untyped") SNPs in our GBS-derived catalog of SNPs. In a first step, we identified 25 accessions common to our set of 301 lines and the USDA collection. By comparing the SNP data for these common accessions, we found that only 7% of markers (2,975 of 42,508; at MAF=0.05) were shared between the GBS and SoySNP50K data. As these two datasets have a limited overlap, this offered the potential of adding a large number of untyped SNP loci through imputation. In a second step, we used the SoySNP50K data as a reference panel to perform imputation of genotypes at the untyped loci in our GBS-derived catalog. As shown in Table V.2, both BEAGLE and IMPUTE2 performed very well resulting in a high accuracy of the imputed genotypes (94.9 and 95.3%, respectively). The successful imputation of these untyped loci increased the number of SNP markers from 62,643 to 102,175, all the while maintaining a high level of accuracy of the combined catalog of SNPs.

Another source of haplotype information resided in our WGS data on the subset of 23 resequenced Canadian lines. We therefore tested how useful this information could be in terms of imputing an even larger set of untyped loci. As described above, a total of 3.6M SNPs were identified among these 23 lines. We removed all redundant markers, i.e. SNPs that were in perfect LD with at least one other SNP, thus reducing this reference panel to 1.4M tag SNPs. We then used BEAGLE and IMPUTE2 for imputation using the SNP data from 22 lines as a reference panel and keeping the last line (Gaillard) for the estimation of accuracy. As shown in Table V.2, the accuracy of imputed genotypes ranged from as low as 88% to as high as 91.8%. Again, differences in computation time were observed with BEAGLE proving to be the most efficient.

Finally, to ensure that these results were broadly applicable to the larger set of 23 lines, two additional permutations were done where a different set of 22 lines was used as a reference panel and the remaining line (Mandarin or OAC-Lakeview) used for validation. Here again, the accuracy of imputation proved highly similar to the results described above, ranging between 87.9% and 92.4%.

To explore the impact of the size of this reference panel on the accuracy of imputed SNPs, we performed imputation with reference panels representing subsets of 5, 10, 15 or 22 of the 23 lines for which WGS data were available. As can be seen in Figure V.3, the accuracy of imputation was highly affected by the number of lines used in the reference panel. With only 5 lines included in the reference panel, imputation accuracy was low (60% with BEAGLE and 59% with IMPUTE2) while it increased (to 88% with BEAGLE and 91.8% with IMPUTE2) using the maximum number of lines available (22). This suggests that a further increase in the number of lines included in the reference panel could provide an increase in the accuracy of the imputation of untyped loci.

V.5.4 Power of association test using imputed data

To determine if the enhanced SNP catalogs obtained through imputation could provide increased power in genome-wide association scans, a subset composed of 139 soybean lines was used to perform an association analysis for seed oil content. This subset was used because phenotypic data were available only for these lines. One analysis was conducted using a "basic" GBS catalog of 7,152 SNPs obtained at MaxMD=20% and MinMAF=0.05, while the other was performed using an enhanced catalog resulting from imputation of missing GBS data (at MaxMD=80%) and untyped loci from the SoySNP50K dataset. At MAF ≥0.05, a total of 83,532 SNPs were retained within this combined dataset. As can be seen in Figure V.4a, using the "basic" SNP catalog, a single SNP marker on Gm19 showed a significant association ($p = 9.6\times10^{-3}$ and $q = 0.09$) with seed oil content. In contrast, using the enhanced SNP catalog (Figure V.4b) and a multi-locus mixed-model implementation, a total of 11 markers were in significant association with this trait despite the fact that the significance threshold increased from 3.4 to 5.3 ($-\text{Log}_{10}$ p-value). Interestingly, the peak SNP in both cases was the same (Gm19_41742182), but its association with oil content exhibited a much higher p-value (3.1×10^{-7}) and lower q-value (0.01). This demonstrates that the increased number of informative SNP loci, obtained through the imputation of both missing GBS data and untyped loci from additional sources of SNP haplotype information, can prove highly beneficial in studying the genetic architecture of complex traits.

V.6 Discussion

A first key element to come out of this work is that MaxMD is the most important factor determining the number of SNPs in GBS analysis. As seen in this study when increasing

MaxMD from 20 to 80% incrementally, the number of SNPs called increased from 12,712 to 62,643. As previously described, one of the unique features associated with GBS is the generation of highly incomplete SNP genotype data (Fu et al. 2011; Fu et al. 2012; Poland et al. 2012; Fu et al. 2014), largely due to low coverage sequencing (Davey et al. 2011). The incompleteness could be up to 90% of observations missing (Fu et al. 2011; Elshire et al. 2011). As described in several GBS studies in different species (maize, rice, wheat, soybean, and barley), increasing the amount of missing data allows to capture more SNPs (Huang et al. 2014; Rutkoski et al. 2013; Fu et al. 2014; Jarquín et al. 2014; Crossa et al. 2014). In the most closely related work, Jarquin et al. (2014) observed a 4-fold increase in the number of SNPs scored in elite soybean breeding lines when increasing the percentage of missing data allowed from 5% to 80%. These data confirm that with increasing MaxMD the number of SNPs called through GBS can be increased substantially.

As described, the number of SNPs is also affected by MinMAF, but the overall proportion of missing data is hardly affected. The effect of MinMAF on the number of SNPs has been described in several reports. The number of SNP increases as the minor allele frequency decreases (Jarquín et al. 2014; Crossa et al. 2014; Howie et al. 2011). These authors, however, did not show the relation between MinMAF and the proportion of missing data. In this study, we demonstrated that the proportion of missing data is largely independent of the chosen MinMAF. In a practical context, however, there is a more limited scope for using a broad range of MinMAF values, as these are usually constrained by the need to have an adequate representation of the minor allele state. Typically, in GWAS and other similar genetic studies, the most frequently encountered MinMAF values are 0.05 and 0.10. In contrast, the amount of missing data that is tolerated is much more variable across studies and is mostly constrained by the quality of the imputation that can be achieved when filling in these missing data.

Somewhat counterintuitively, a second key result of this work was that imputation of missing data was more accurate when performed on datasets with a higher proportion of missing data. Indeed, at MaxMD=80%, 94% of SNP genotypes were correctly imputed, whereas at MaxMD=20%, the accuracy decreased to 86%. Upon reflection, however, it seems logical that a larger number of SNP markers (albeit with more missing data) better captures the diversity of haplotypes that are present within a collection of lines. Increased imputation accuracy at MaxMD=80% is likely achieved through increased LD between markers. As documented by Zheng et al. (2011), imputation accuracy increases with increasing density of markers. Soybean has high levels of LD and the average distance over which LD decays to half of its

maximum value in soybean is substantially longer than that of many plants and animals analyzed to date (cultivated soybean: ~150 kb; wild soybean: ~75 kb; cultivated rice: <65-180; wild rice <10 kb; maize: <1 kb; and *Arabidopsis thaliana*: ~3-4 kb; humans <5kb; cattle <10kb) (Gore et al. 2009; Kim et al. 2007; Zhu et al. 2007; Xu et al. 2012; Shifman et al. 2003; Porto-Neto et al. 2014). High levels of LD will decrease the haplotype diversity and as a result facilitate the imputation of missing data even over long distances. This suggests that imputation accuracy will vary with differing levels of LD in different species.

A novel aspect of this work is that the measurement of the accuracy of imputation was assessed by comparing directly to whole genome resequencing data obtained for a subset of the lines. In many previous studies, estimates of the accuracy of imputation have been achieved by masking a subset of the data, imputing these missing genotypes, and then comparing the imputed genotype with the original data (Huang et al. 2014; Jarquín et al. 2014; Crossa et al. 2013; Howie et al. 2011). For the most part, similarly high levels of imputation accuracy (92-98%) have been reported with slight differences being observed between species and types of population (related or unrelated individuals). The advantage of using resequencing data in this fashion is that we can assess the accuracy of imputation at a specific level of missing data without having to add to this by masking a subset of the available data.

Furthermore, although the threshold for retaining a SNP marker at MaxMD=80% would suggest a tremendous amount of missing data, we showed that, averaged across all markers kept at this threshold, a mean of 50% missing data was obtained. When we considered jointly both the called and imputed markers comprising the final dataset at the various missing data levels, all were highly accurate (96-98%). This is because the genotypes initially called via GBS analysis are themselves highly accurate (98.4%). At MaxMD=20%, these high-quality SNPs are combined with a small proportion (7%) of SNPs imputed with what we term a "good" accuracy (84%). At the other end of the missing data spectrum (MaxMD=80%), the original set of GBS-called SNPs is combined with an equal amount (~50%) of SNPs derived from imputation with an only slightly lower accuracy (94%). Thus, catalogs of called and imputed SNPs retain a constant, high level of accuracy (~97%) across a broad range of missing data thresholds.

A third key finding of this work is that different and highly complementary marker datasets can be successfully combined via imputation at untyped loci. We showed that SNP catalogs derived from two high-throughput genotyping techniques, GBS and a SNP array

(SoySNP50K), could be fused through the imputation of a large number of untyped loci. Because of the different composition of the two initial catalogs, only 7% of the GBS markers were present in the SoySNP50K set. This is because most (90%) of the SoySNP50K markers are present in genic regions (Song et al. 2013), while most of the GBS markers are present in intergenic regions (29.8%) or downstream regions (20.2%) (Sonah et al. 2013). We nonetheless successfully imputed ~40K SNPs from the array that were absent from the GBS dataset with a high level of accuracy (95%). By doing so, our catalog of SNPs for the collection of 301 Canadian soybean lines was enhanced and exceeded 100K SNPs. This analysis shows that GBS and SNP arrays are highly complementary approaches that can be used in parallel and combined. As the SoySNP50K has been used by the USDA to characterize close to 20,000 lines of soybean, and because these data are public, any researcher anywhere in the world can make use of this data, in combination with their own GBS-derived data obtained at a very low cost, to achieve excellent genome coverage. Similarly, Pei et al. (2008) and Hao et al. (2009) used imputation to combine data from two human genotyping arrays: the Affymetrix 500k SNP chip and the Illumina 550k chip with HapMap SNPs. They showed that the accuracy of imputation at such untyped loci using various tools (BEAGLE, fastPHASE, and IPMUTE2) ranged between 92 and 94%. We suggest that the higher level of imputation accuracy observed in this study compared to the human dataset is because of the high level of LD in soybean. Again this result suggests that the accuracy of genotype imputation at untyped loci will vary in different species because of stark differences in the extent of LD. Overall, a competing genotyping platforms are developed, it is good to know that researchers can produce high-quality integrated data sets offering better genome coverage by such imputation of untyped loci.

Although all imputation softwares use the same fundamental phenomenon of LD across the genome, the algorithms employed by each package differ. Likewise, each package offers differing strengths and weaknesses. Therefore, it is a good idea to use more than one software package, compare results, and investigate any major discrepancies (Ellinghaus et al. 2009). To perform genotype imputation, we used three imputation softwares and found that these showed approximately the same level of accuracy for missing data imputation. In our view, BEAGLE proved the most attractive, as it ran very quickly and was the most user friendly. As reference panels for the imputation of untyped loci become larger and larger, thanks to the increasing availability of data derived from the resequencing of an increasing number of soybean lines, these tools will gain further utility. In the context of this work, genotype imputation using the SoySNP50K data as a reference, both BEAGLE and IMPUTE2 showed the

same accuracy (95%). Contrary to most previous work, we did not assess the accuracy of our imputation through the masking of a subset of available data. Rather, we performed whole-genome resequencing of a subset (23 lines) from our study population (301 lines) and we compared directly the imputed genotype and the true genotype. This analysis showed the high level of imputation accuracy.

When performing imputation at a much larger scale, using the 1.4M tag SNPs identified in our resequencing effort, the accuracy of imputation of this large number of untyped loci was dependent on the number of lines included in the reference panel. When increasing the number of lines composing the reference panel from only 5 to a maximum of 22, imputation accuracy increased from ~60% to close to 90%. Similarly, in humans, Li et al. (2009) showed that increasing the number of individuals in the reference panel from 60 to 500 improved the accuracy of imputation (from 85% to more than 95%, respectively). Interestingly, even a small number of soybean lines (22) resulted in higher imputation accuracy than was achieved with 60 human samples. As LD is much more extensive in soybean than in humans, this again illustrates how important this factor will be in determining imputation accuracy. In future, to achieve a level of accuracy similar to that seen using the SoySNP50K data (95%), more lines from the Canadian germplasm collection would likely need to be sequenced.

A final key finding of this work is that the much increased marker coverage achieved through a better exploitation of available GBS and SoySNP50K data is highly useful in the genetic dissection of complex traits. The availability of higher density marker coverage enables researchers to more accurately determine which regions to investigate further and actually narrow down each region on which they should perform fine mapping. As illustrated in our analysis of seed oil content, the use of an enhanced SNP catalog (~6 fold larger) allowed us to capture more significant marker-trait associations around candidate QTLs and the significance level of such associations was also much higher. These results are consistent with recent work in both animals and plants that have demonstrated the benefits of marker imputation for GWAS (Santana et al. 2014; He et al. 2015). In the latter case, the authors compared the benefits of marker imputation on the accuracy of measures of relatedness, the accuracy of genomic selection and the power to detect QTLs through GWAS. In this work, these authors concluded that "association mapping profited most from imputing missing values".

V.7 Conclusion

As seen in this study, genotype imputation represents an essential tool in the analysis of high-throughput genotypic data. One of the most common criticisms regarding GBS is the presence of a substantial amount of missing data. Our data show that this can largely be overcome in soybean thanks to highly accurate imputation of missing genotypes. Furthermore, genotype imputation is particularly useful for combining results across studies that rely on different genotyping platforms. As different groups may use different genotyping tools, it is highly important to be able to produce integrated datasets that include all such markers to facilitate the exchange of knowledge and information. It is important to remember, however, that imputation accuracy will be affected by the extent of LD in the population/species studied. Finally, a further benefit of such imputation is that it increases the power of individual scans thanks to more extensive marker coverage. In the coming years, we expect these imputation-based analyses will become a key tool in the analysis of massively parallel shotgun sequence data enabling geneticists to rapidly deploy these technologies to analyze large samples and dissect the genetic basis of complex traits.

V.8 Tables

Table V.1. Accuracy of imputed GBS SNP data and computational speed of three imputation methods at different levels of missing data (MaxMD) and minor allele frequency (MinMAF).

			Missing data imputation accuracy (%)						
			MAF 0.003		MAF 0.05		MAF 0.1		
Method	Dataset	MaxMD (%)	Missing data	Overall*	Missing data	Overall	Missing data	Overall	Computing Time
fastPHASE	GBS**	80	93.2	95.8	93.9	96.4	94.1	96.5	14 hours
		20	85.6	97.5	86.5	98.1	87.5	98.1	
BEAGLE	GBS	80	92.9	95.6	94.0	96.5	94.2	96.6	30 minutes
		20	85.6	97.5	86.7	98.1	87.6	98.1	
IMPUTE2	GBS	80	93.0	95.6	93.5	96.2	94.3	96.6	2 hours
		20	86.1	97.5	86.9	98.1	88.1	98.2	
Number of SNPs	GBS	80	62,643		41,024		32,035		
		20	12,712		7,152		5,657		

* Includes both genotypes originally called by GBS and following imputation

** 301soybean lines

Table V.2. Accuracy and computational efficiency of imputation at untyped loci. SNP data from a SNP array (SoySNP50K) or whole-genome resequencing (WGS) were used as a reference to impute missing data at loci that were untyped in an initial dataset (GBS data only or GBS +SoySNP50K data).

Dataset	Imputation method	Reference panel	Untyped loci imputation accuracy (%)	Number of markers	Computing Time
	BEAGLE				
GBS	Beagle	SoySNP50K	94.9	102,175	71 hours
GBS	Beagle	WGS	80.0	1,414,925	2 hours
GBS+ SoySNP50K	Beagle	WGS	88.1	1,312,760	2 hours
	IMPUTE2				
GBS	pre-Phasing by SHAPIT2	SoySNP50K	95.3	102,175	91 hours
GBS	pre-Phasing by SHAPIT2	WGS	90.0	1,414,925	7 hours
GBS+ SoySNP50K	pre-Phasing by SHAPIT2	WGS	91.8	1,312,760	8 hours

V.9 Figures

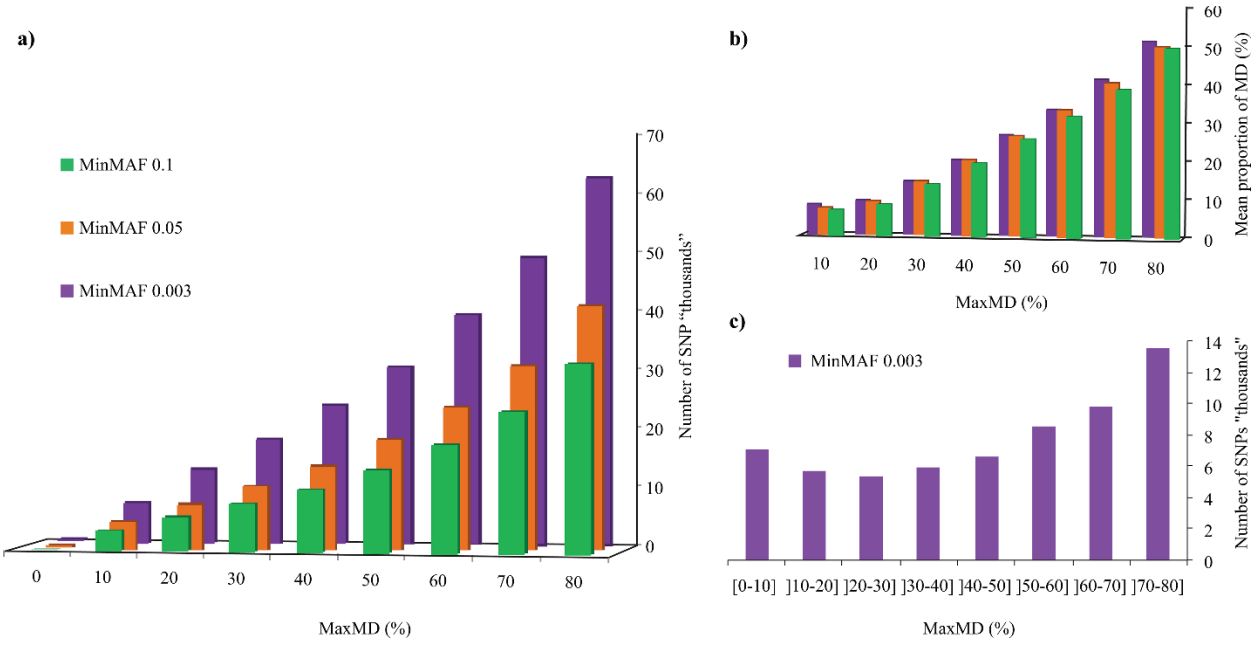


Figure V.1. Impact of missing data and minor allele frequency on the number of SNPs. (a) The number of SNPs (in '000's) is plotted as a function of the maximal proportion (in %) of missing data tolerated (MaxMD) at three levels of minimal minor allele frequency (MinMAF). (b) Overall mean proportion of missing data (in %) for datasets obtained at different levels of MaxMD and MinMAF. (c) Distribution of SNPs called at different levels of missing data.

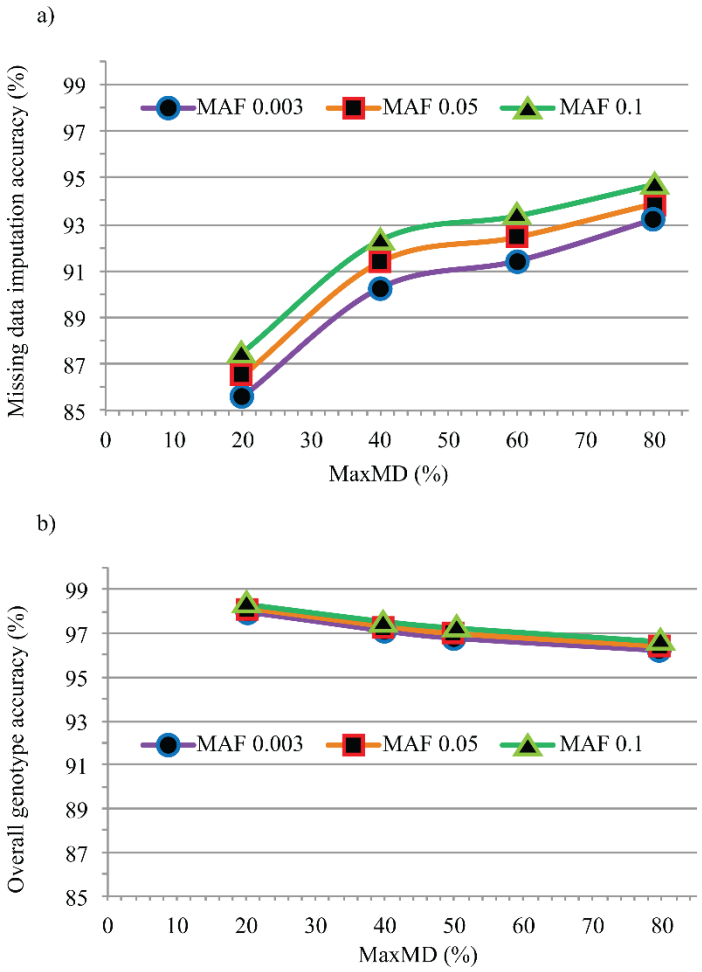


Figure V.2. Missing data imputation accuracy. (a) The accuracy of imputed missing data (in %) is plotted against the proportion of missing data (in %) tolerated (MaxMD) at three levels of minimal minor allele frequency (MinMAF). (b) Accuracy of overall GBS dataset (in %) after imputation at different levels of MaxMD and MinMAF.

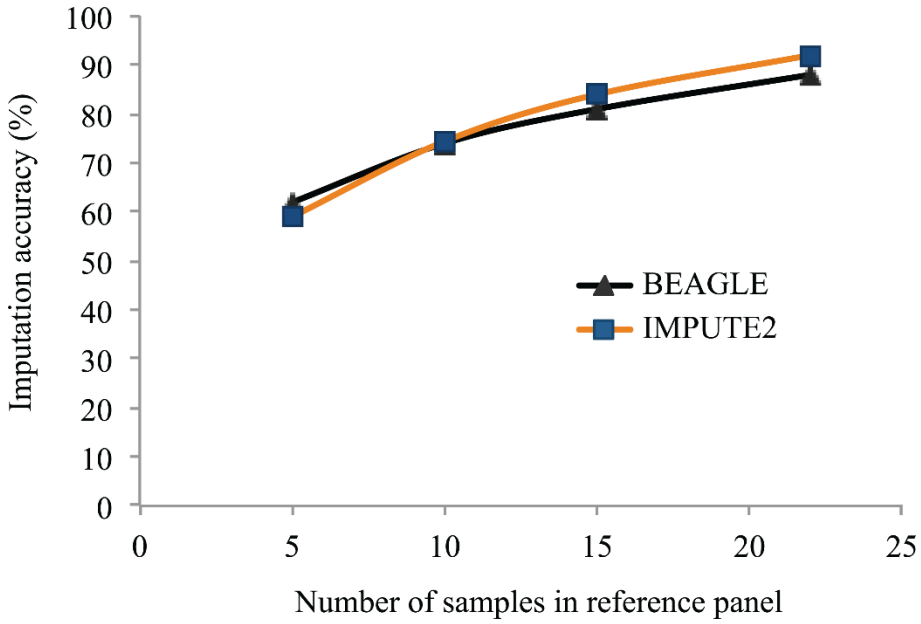


Figure V.3. Imputation accuracy at untyped SNPs using reference panels of different sizes. SNPs identified through resequencing of a varying number (5 to 22) soybean accessions were used as a reference panel to impute the genotypes at these SNP loci in a set of 301 soybean accessions using two different imputation softwares (BEAGLE and Impute2).

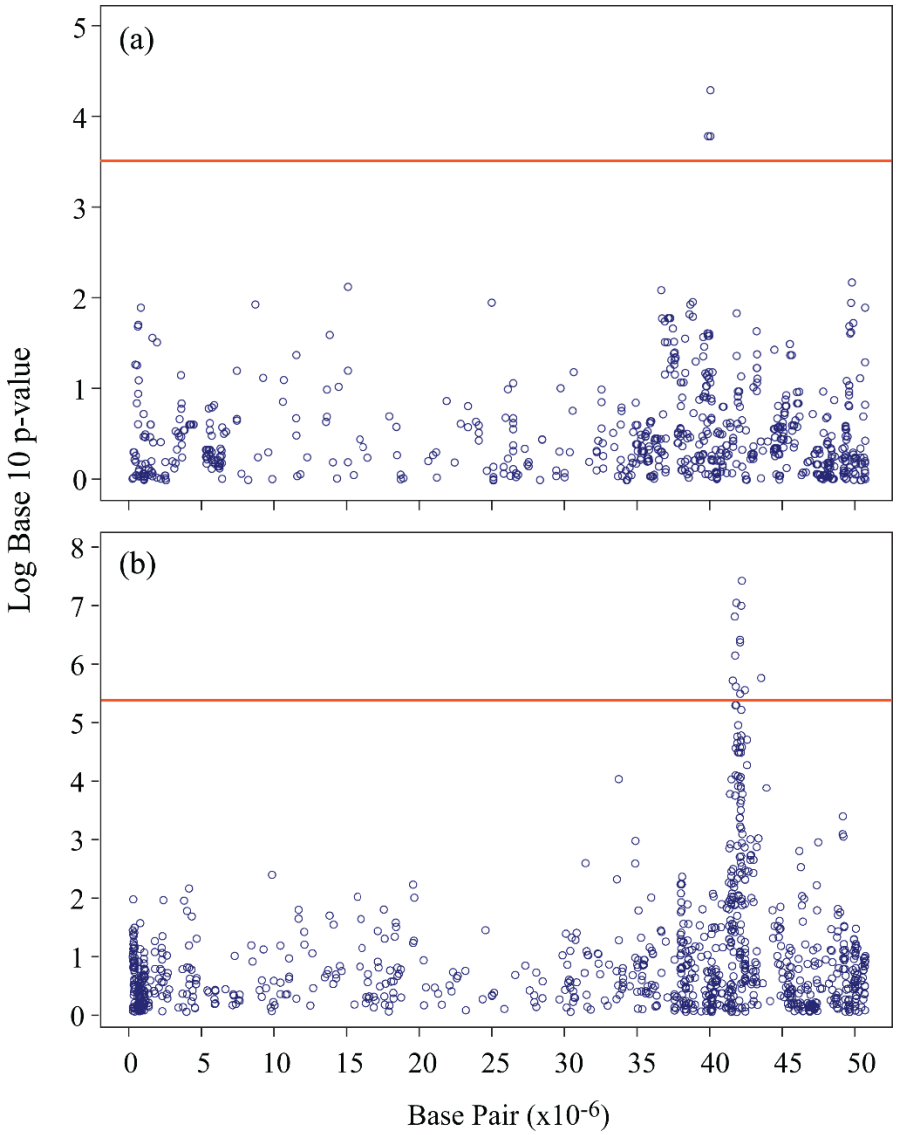


Figure V.4. Association analysis for seed oil content on chromosome 19 (Gm19) in soybean. Negative log10 *p*-values from a genome-wide scan are plotted against marker positions on chromosome 19. (a) Association analysis with the original GBS dataset (~7K SNPs). (b) Association analysis with the enhanced SNP dataset (>83K SNPs) after combining GBS and SoySNP50K data via imputation.

V.10 Supplementary files

Supplementary files listed and described below can be found online at

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0131533#sec015

Additional file V.1 Supplementary Table. List of resequenced samples with the number of reads and bases.

Additional file V.2 Supplementary Table. Overall accuracy of genotypic data following GBS analysis and imputation of missing data for all 20 soybean chromosomes.

Additional file V.3 Supplementary Table. Imputation accuracy of genotypes at untyped loci using whole-genome sequence data as a reference panel.

Chapitre VI

Comprehensive Description of Genome-Wide Nucleotide and Structural Variation in Short-Season Soybean

Davoud Torkamaneh[1,2], Jérôme Laroche[2], Aurélie Tardivel[1,2,3], Louise O'Donoughue[3], Elroy Cober[4], Istvan Rajcan[5] and François Belzile[1,2]

[1]Département de Phytologie, Université Laval, Quebec City, QC, Canada

[2]Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Quebec City, QC, Canada

[3]CÉROM, Centre de recherche sur les grains inc., Saint-Mathieu de Beloeil, QC, Canada

[4]Agriculture and Agri-Food Canada, Ottawa, ON, Canada

[5]Department of Plant Agriculture, Crop Science Bldg., University of Guelph, Guelph, ON, Canada

Plant Biotechnology Journal

2017

VI.1 Résumé

Les séquençages de la nouvelle génération (NGS) et les outils de bioinformatique ont grandement facilité la caractérisation des variations nucléotidique; néanmoins, une description exhaustive de la diversité haplotypique et de la variation structurelle reste limité dans la plupart des espèces. Dans cette étude, nous avons séquencé un ensemble de 102 accessions de soja hâtif qui a permis à attendre une couverture étendue de la diversité nucléotidique et des variations structuraux (SV). Nous avons détecté proches de 5M (SNP, MNP et Indels) des variants de séquences. Nous avons remarqué que le nombre d'haplotypes uniques avait plafonné dans cet ensemble de germoplasme (1.7M tag SNPs). Cet ensemble de données s'est avéré très précis (98,6%) en fonction d'une comparaison des génotypes appelés à loci partagé avec une puce de SNP. Nous avons utilisé ce catalogue de SNP en tant que panneau de référence pour imputer les génotypes manquants dans les loci absent dans les ensembles de données dérivés d'outils de génotypage à faible densité (150K GBS-SNPs dérivés / 530 échantillons). Après imputation, 96,4% des génotypes manquants imputés de cette manière se sont révélés exacts. À l'aide d'une combinaison de trois pipelines bioinformatique, nous avons découvert ~92K SV (délections, insertions, inversions, duplications, CNV et translocations) et estimé que plus de 90% étaient exacts. Enfin, nous avons remarqué que la duplication de certaines régions génomiques expliquait une grande partie de l'hétérozygotie résiduelle des loci SNP dans les accessions de soja très consanguines. C'est la première fois d'une description complète de la diversité haplotypique et du SV a été réalisée chez un grand culture.

VI.2 Abstract

Next-generation sequencing (NGS) and bioinformatics tools have greatly facilitated the characterization of nucleotide variation; nonetheless, an exhaustive description of both SNP haplotype diversity and of structural variation remains elusive in most species. In this study, we sequenced a representative set of 102 short-season soybeans and achieved an extensive coverage of both nucleotide diversity and structural variation (SV). We called close to 5M sequence variants (SNPs, MNPs, and Indels) and noticed that the number of unique haplotypes had plateaued within this set of germplasm (1.7M tag SNPs). This dataset proved highly accurate (98.6%) based on a comparison of called genotypes at loci shared with a SNP array. We used this catalogue of SNPs as a reference panel to impute missing genotypes at untyped loci in datasets derived from lower density genotyping tools (150K GBS-derived SNPs/530 samples). After imputation, 96.4% of the missing genotypes imputed in this fashion proved to be accurate. Using a combination of three bioinformatics pipelines, we uncovered ~92K SVs (deletions, insertions, inversions, duplications, CNVs, and translocations), and estimated that over 90% of these were accurate. Finally, we noticed that the duplication of certain genomic regions explained much of the residual heterozygosity at SNP loci in otherwise highly inbred soybean accessions. This is the first time that a comprehensive description of both SNP haplotype diversity and SV has been achieved within a regionally relevant subset of a major crop.

VI.3 Introduction

Genetic variation describes the occurrence of DNA sequence differences among individuals of the same species (Hedrick 2011). Genetic variation is highly advantageous in an evolutionary sense as it enhances adaptability and survival of a population in the face of changing environmental conditions and other unexpected circumstances (Hedrick 2011; Dobzhansky 1970). Genetic variation can be broadly divided into two major categories: nucleotide and structural variations. Nucleotide variants are usually defined as encompassing single or multiple nucleotide variants (SNPs, MNPs) and small insertions/deletions (indels), whereas structural variants (SVs) represent larger rearrangements of various types [deletions, insertions, inversions, translocations, duplications, and copy number variations (CNVs)] (Tuzun et al. 2005). The advent of Next-Generation Sequencing (NGS) technologies have provided an exceptional opportunity to systematically detect both nucleotide and structural variants in plant and animal genomes (El-Metwally et al. 2014; Hall 2007; Church 2006).

NGS has facilitated greatly the development of methods to genotype very large numbers of nucleotide variants such as single nucleotide polymorphisms (SNPs) (Goodwin et al. 2016). In a complementary approach, NGS has been exploited to simultaneously identify and genotype informative SNPs, without the need for any prior knowledge of these polymorphic loci, using complexity reduction approaches such as genotyping-by-sequencing (GBS) (Davey et al. 2011). Finally, decreased whole-genome sequencing (WGS) costs have made it possible to sequence entire genomes of numerous individuals, cultivars or accessions of the same species (Zhang et al. 2001; Zhou et al. 2015; Gudbjartsson et al. 2015).

NGS technologies now allow large quantities of high-quality DNA sequence data to be generated at modest cost (Zhang et al. 2001). However, despite considerable advances in algorithm development, the processing of these massive amounts of sequence data into high-quality variant calls remains challenging (Muir et al. 2016). To date, several tools have been developed to discover and genotype nucleotide variants, while SV detection and calling algorithms are relatively recent (Hwang et al. 2015). Decoding the raw sequencing data into a catalogue of nucleotide variants and genotype calls requires two essential steps: read mapping and variant/genotype calling. First, reads are aligned against a reference genome, variable sites are identified and genotypes at those sites are determined (Nielsen et al. 2011). In addition to calling SNPs and small indels, however, bioinformatics tools have been developed to allow the discovery and genotyping of larger sequence variants (Layer et al. 2014; Chen et al. 2009; Abyzov et al. 2011). To date, three major strategies have been exploited to identify structural variants from aligned reads: depth of coverage, paired-end

mapping and split read mapping. Depth of coverage is designed to detect changes in the number of reads that align to a given region in the genome. A reduction or an increase in this coverage can suggest that a deletion or an increase in the copy number of a sequence has occurred in a given individual compared to the reference genome. When paired-end sequencing is used, it can be assumed that the two sequences that form a pair originate from a single DNA fragment, and thus lie in close proximity on an opposite strand of the reference genome. In the paired-end mapping approach, when paired reads deviate from this expectation, either because they map to sites that are too far apart or are no longer on opposite strands, this suggests that the individual sample from which these paired reads were generated differed from the reference genome in some structural fashion. Finally, in the case of split reads, this strategy exploits the fact all structural rearrangements generate breakpoints that are analogous to "scars". The "scars" produce sequence reads that contain base pairs that are not contiguous in the reference genome. If two portions of a single sequence read align to different places in the reference genome, this suggests that a rearrangement has occurred (Marroni et al. 2014).

To date, the genetic dissection of complex traits in plants and animals has relied almost exclusively on nucleotide variants either as markers of a closely-associated mutation or as the direct causal mutation. In recent years, several studies have illustrated the functional impact of SVs in human disease, plant phenotypes and disease resistance (Carvalho et al. 2016; Cook et al. 2012). Therefore, no characterization of genetic diversity is complete without the description of both nucleotide and structural variation.

In this study, we describe the WGS of 102 short-season soybean accessions to identify both nucleotide and structural variants using a combination of several bioinformatics tools. We then measure the accuracy of these variants through validation experiments and describe their distribution in the soybean genome. We also show the impact of joint analysis of nucleotide and structural variants in elucidating the cause of residual heterozygous genotypes observed in inbred lines that are expected to be fixed at all loci.

VI.4 Materials and methods

VI.4.1 Soybean accessions

In this study, we used three collections of soybean samples. A first panel of 441 accessions (cultivars/advanced breeding lines) was subjected to genotyping-by-sequencing (GBS; *Ape*KI protocol) (Elshire et al. 2011; Sonah et al. 2013) and SNPs were called using the Fast-GBS pipeline (Torkamaneh et al. 2017a). Based on a cladogram produced using these data, a second panel comprising 102 elite accessions (Supplementary Table VI.1) were selected to capture the diversity among this collection of short-season soybean and were used for WGS (Supplementary Figure VI.1). Finally, a set of 89 accessions (mostly advanced breeding lines harboring traits of interest) was genotyped by GBS, as described above, and added to the collection of 441 accessions to produce a third panel totaling 530 soybean accessions on which we tested the accuracy of imputation at untyped loci (see below for details).

VI.4.2 Whole-genome sequencing

Illumina Paired-End libraries were constructed for 102 elite accessions (panel 2 described above) using the KAPA Hyper Prep Kit (Kapa Biosystems, Wilmington, Massachusetts, USA) following the manufacturer's instructions (KR0961 – v5.16). Samples were sequenced using the Illumina HiSeq 2500 platform at the Centre Hospitalier de l'Université Laval (CHUL) in Quebec, QC, Canada.

VI.4.3 Choice of WGS analytical pipeline

Two SNP-calling pipelines were used: SOAPsnp (Li et al. 2009) and Fast-WGS, a new pipeline that we have developed (see details in Supplementary Text 1). Every effort was made to call SNPs under comparable conditions. The final variant catalogue was prepared using Fast-WGS. Then we downloaded the catalogue of sequence variants of *Glycine* spp. from dbSNP (build 147), to compare and identify the novel variants detected in this study.

VI.4.4 Genotype accuracy

The SoySNP50K iSelect BeadChip has been used to genotype the entire USDA Soybean Germplasm Collection (Song et al. 2015). The complete dataset for 19,652 *G. max* and *G. soja* accessions genotyped with 42,508 SNPs was downloaded from Soybase (Grant et al.

2010). Of these accessions, 19 were in common with the collection of 102 short-season soybean lines characterized here via WGS. For these 19 accessions, we extracted their genotype calls at all SNP loci for which data were available. This large set of SoySNP50K genotype calls (>600K) was directly compared with the WGS-derived SNP calls (obtained using one or the other pipeline) using an in-house script.

VI.4.5 Imputation

To impute missing data in the WGS dataset, we used BEAGLE v5 (Browning and Browning 2007) with the parameters described in Torkamaneh et al. (2015). Imputed genotypes at loci in common with the SoySNP50K array were directly compared to those called using the chip. The WGS SNP data from 101 of the 102 resequenced lines were also used as a reference panel to impute missing data onto a collection of 530 accessions (panel 3) previously genotyped with ~150K GBS-derived SNPs. The remaining line was kept out of the reference panel to determine how accurately data at untyped loci (present in the WGS data but absent from the GBS catalogue) could be imputed in this line. We performed five such permutations where a single line was kept aside to estimate imputation accuracy. For these lines purposely excluded from the reference panel, we compared the imputed genotypes against the genotypes called at these same loci following WGS.

VI.4.6 Population genetics, LD, and tag SNP selection

Population structure was estimated using the Bayesian inference implemented in fastSTRUCTURE (Raj et al. 2014). Five runs were performed for each number of populations (K) set from 1 to 12. The most likely K value was determined by the log probability of the data (LnP(D)) and delta K, based on the rate of change in LnP(D) between successive K values. A neighbour- joining tree was built using MEGA6 (Tamura et al. 2013) with 100 bootstraps. Principal-component analysis (PCA) was performed using TASSEL v5 and GAPIT (Bradbury et al. 2007; Lipka et al. 2012) in three dimensions. For tag SNP selection, we used PLINK (Purcell et al. 2007) to calculate linkage disequilibrium (LD) between each pair of SNPs within a sliding window of 50 SNPs and we removed all but one SNP that were in perfect LD (LD = 1); the remaining SNPs were deemed tag SNPs.

VI.4.7 Annotation and GO analysis

Functional annotation of nucleotide variation was done by SnpEFF and SnpSift (Cingolani et al. 2012) using *G. max* reference genome [*Gmax*_275 (Wm82.a2.v1)] (Schmutz et al. 2010). Genes containing variants predicted to have a large functional impact were selected from the annotation file. To obtain the description of these genes we used Phytozome (Goodstein et al. 2012) and SoyBase (Grant et al. 2010). For gene ontology (GO) analysis we used the Singular Enrichment Analysis (SEA) method implemented in agri-GO (Zhou et al. 2010).

VI.4.8 Structural variant calling and genotyping

To discover a comprehensive catalogue of SVs from WGS data we used three tools: LUMPY (Layer et al. 2014), BreakDancer (Chen et al. 2009) and CNVnator (Abyzov et al. 2011). We used SVtyper (Chiang et al. 2015) and svtools (Larson et al. 2016) for calling the presence or absence of SVs in individual accessions. The raw calls were filtered for 1) the estimated read-depth ratio (<0.75), 2) the number of spanning read pairs (>10), 3) regions around centromeres (+/- 1Kb) and 4) regions around assembly gaps (+/- 50bp). The read-depth (RD) ratio was calculated as the average RD of the samples that supported the SV divided by the average RD of the samples that did not support the SV. The site list was prepared by using an 80% reciprocal overlap (RO) threshold, a maximum breakpoint offset of 250 bp and a genotype quality (phred scaled) >30. Inversions were filtered such that the minimum ratio of genotyped to ungenotyped samples was >0.4 and the fraction of inversions supporting pairs in carriers was >0.3. The translocation calls located in syntenic regions were removed.

VI.4.9 Annotation of structural variants

Functional annotation of SVs was done using an in-house Python script. We used the *G. max* v2 annotation file to create a genic reference panel in which we recorded the genomic region spanned by each gene. Similarly, we created a file for each SV in which the positions of both breakpoints (start and end) were noted. To detect SVs that had a likely functional impact on genes, we proposed four possible scenarios; (1) a SV was located inside a gene, (2) a SV began in an intergenic region (upstream) and terminated in a gene, (3) a SV began in a gene and terminated in an intergenic region (downstream), and (4) a SV encompassed the gene completely (Supplementary Figure VI.5). Using this program, we compared the intervals spanned by SVs with genic intervals to identify partial or complete overlaps.

VI.4.10 Validation of structural variants

We selected two known SVs in known maturity genes (E3 and E4) and 38 random SVs with a focus on translocations and inversions for a PCR-based validation. Primers were designed using Primer3Plus (Untergasser et al. 2007), and their specificity was examined using BLAST on the NCBI and SoyBase databases (Supplementary Table VI.5). Williams82 was used as the reference (control) for PCR. For estimation of breakpoint precision, the PCR products were sequenced using Sanger sequencing.

VI.5 Results

VI.5.1 Nucleotide variation

VI.5.1.1 Discovery and genotyping

We selected 102 Canadian short-season elite soybean accessions for whole-genome sequencing based on a prior genetic analysis containing a larger set of accessions (n=441) that had been genotyped with ~80K SNPs using a genotyping-by-sequencing (GBS) approach (Supplementary Figure VI.1). This collection of 102 samples was selected based on genetic distance to cover genetic diversity of short-season soybean germplasm. The accessions were sequenced using Illumina short-read technology (100- or 125-bp reads) to a median depth of 11x (Supplementary Table VI.1). A total of 1.02×10^9 high-quality trimmed reads (Phred quality score > 32) were used to call nucleotide variation in this dataset. On average, a coverage of at least 1x was achieved for 956 Mb (excluding gaps), thus covering 97.6% of the *G. max* genome sequence.

To date, all variant calling from WGS data in soybean has been performed using the SOAPsnp pipeline. Prior to conducting large-scale variant calling on all accessions, however, we first tested the performance and speed of four genotyping pipelines/tools: Fast-WGS (developed in-house, see description in Supplementary Text 1), SOAPsnp, GATK HC and SAMtools on a subset of only 10 accessions. All four called a similar number of SNPs (~1.7M) and indels (~270K), but vast differences were observed in terms of the time needed to complete this analysis (23h, 61h, 581h and 238h, respectively) on the same server (Linux, 48 CPU, 1 Tb RAM). Based on these results, we chose to conduct an analysis on the entire set of accessions only with the two fastest pipelines: Fast-WGS and SOAPsnp. We then analyzed the complete set of reads (for all accessions) with these two pipelines under the same variant-calling conditions. As shown in Table VI.1, Fast-WGS called slightly more (7.2%) total variants due either to base substitutions (SNPs and MNPs) or small indels (4,998,229 vs 4,636,634). Of

these, close to 1M variants were identified as novel polymorphisms not previously recorded in dbSNP among the *Glycine* spp. (Supplementary Text 2).

To assess and compare the quality of genotype calls, we compared our WGS data with the SoySNP50K array data for 19 accessions for which these data were available. Globally, more than 600K genotype calls (35,481 SNP loci × 19 samples) could be compared in this fashion, of which 0.25% were presumed to be indels when no genotype (missing data) was indicated for a given site in a given accession in the SoySNP50K data. As can be seen in Table VI.2, the quality of the genotype calls made using Fast-WGS was higher for all three types of genotype calls; the degree of concordance with the calls made on the SoySNP50K array increasing by between 2.6 and 6.8% relative to those observed for the SOAPsnp data. This analysis suggests that a higher level of genotypic accuracy could be obtained for the soybean SNP datasets currently available by using the Fast-WGS pipeline.

The SNP dataset obtained using Fast-WGS contained 9% missing data. We wanted to test how accurately these could be imputed. After imputation of these missing data, we compared the imputed genotypes with the subset of corresponding genotypes obtained using the SoySNP50K array. As can be seen in Table VI.3, there were 635 shared genotypes which could be compared in this fashion, 41 of which were heterozygous while the remainder (594) were homozygous. We found a high level of concordance between these two datasets, with 98.8 and 92.7% of homozygous and heterozygous genotypes having been correctly imputed, respectively. Taken together (original calls + imputed calls) across all three types of variants, we found that 99.6% (672,005/674,139) of the genotypes obtained using the Fast-WGS pipeline (including imputed data) proved to be in agreement with the genotypes obtained at loci in common with the SoySNP50K array.

VI.5.1.2 Variant annotation and prediction of their functional impact

We grouped sequence variants into five categories based on the observed minor allele frequency (MAF). As can be seen in Figure VI.1a, 35% of sequence variants were present in up to 10 samples ([0.0-0.1[) and 14% were present at an almost equal frequency with the other allele ([0.4-0.5[). Almost half of these variants were present in the immediate vicinity of genes (up/downstream regions (5 kb before and after gene), 47%) or further removed from genes (intergenic regions, 40%), while exonic and intronic regions contained only 2% and 9% of variants, respectively (Figure VI.1b). Also splice sites contained very few variants with only 0.1% of the total.

We then grouped all observed sequence variants into four categories based on the predicted functional impact of the observed mutation: i) high (0.071%) variants, which are predicted to have a disruptive impact on the protein, probably leading to protein truncation, loss of function or triggering nonsense-mediated decay; ii) moderate (1.341%), non-disruptive variants that might change the protein effectiveness (missense variants and in-frame deletions); iii) low (1.1%), mostly harmless or unlikely to change protein behavior (synonymous variants); and iv) modifier (97.48%), non-coding variants. Figure VI.2 presents the frequency distribution of these four predicted functional impact categories of the mutant (alternative) allele. All four of these categories of mutations showed a similar distribution with most mutations being present at relatively low frequency (< 20%) and only a small subset being present at high frequency (>80%).

From a functional standpoint, we were particularly interested in the subset of mutations predicted to have a large impact. Although these represent only a small fraction of all sequence variants (0.071%), this still corresponds to 4,113 variants in 3,064 genes. Of these variants 2,279 were SNPs, 230 MNPs, and 1,604 indels. Although only 12% of the sequence variants were indels, they were over-represented in this category, owing to their tendency to shift the reading frame when they occur in exons. Thus, indels represented 39% of the 4,113 functionally high impact variants. In total, we detected 1,418 frameshift, 1,378 splice receptor/donor, 1,251 stop-gained, and 185 start/stop lost variants. As expected, the largest proportion of these variants (35.5%, 1,461/4,113) were present at a low frequency (<10%). On the other hand, a total of 331 mutations in 238 genes (7.8%) were present in the vast majority of these soybean lines (frequency ≥0.8) (Supplementary Figure VI.2). Owing to the lack of any significant enrichment in terms of GO annotation (data not shown), we investigated the functional annotation of these genes individually using public databases (Supplementary Table VI.2). Using the SoyBase and Phytozome databases we found that of 238 genes, 31 had no annotation nor evidence of expression, we considered these genes as possible pseudogenes. Among the remaining 207 genes, which had annotation and expression profile, we found at least one other functional copy for 177 genes, while the final 30 genes seemed to be unique genes. We suggest that nonsynonymous mutations in these 30 unique genes for which there was evidence of transcriptional activity would be expected to impact plant function significantly in short-season soybean. Indeed, *Glyma.10g221500* (*GmGIa*) is the gene underlying the maturity locus *E2*. The mutation in exon 10 of this gene is the known causal variant for the *e2* allele. As the lines characterized in this work are all adapted to a

short growing season, it makes perfect sense that these are fixed for a non-functional allele that contributes to earliness.

VI.5.1.3 Population genetics, LD, haplotypes and untyped-genotype imputation

To provide a comprehensive understanding of the population structure among this set of short-season soybean lines, we performed three analyses using SNP data: 1) a phylogenetic tree (neighbor-joining method) with *G. soja* as an outlier; 2) a principal component analysis (PCA); and 3) a STRUCTURE analysis using different K values to detect evidence of admixture in this collection (Supplementary Figure VI.3). The neighbor-joining tree, based on all pairwise genetic distances among the 102 soybean accessions, showed many distinct branches with *G. soja* as a clear outlier (Supplementary Figure VI.3a). Principal component analysis (PCA) also showed that the accessions seemed to form approximately five divergent groups (circled) (Supplementary Figure VI.3b). Similarly, using fastSTRUCTURE, the most likely number of subpopulations (K) was five, with most accessions showing some degree of admixture (Supplementary Figure VI.3c). This collection of soybean accessions is composed of lines belonging to different maturity groups (MGs ranging from 000 to I). We tested whether these defined subpopulations could correspond to different MGs, but this did not prove to be the case (data not shown).

The extent of linkage disequilibrium (LD) can provide a measure of haplotype diversity in a population. We calculated all pairwise LD (r^2 and D´) for sequence variants and we found high levels of LD among short-season soybeans. The average distance over which LD decayed below 0.2 in this population was ~150 kb. Using these LD data, we identified 1.7 million tag SNPs based on haplotypes. To determine if a good level of saturation of both variants and haplotypes had been achieved among elite short-season soybean using this collection of accessions, we analyzed randomly selected subsets of samples of increasing size (N=12, 24, 44, 64, 84, and 102). As illustrated in Figure VI.3, the number of variants discovered did not increase much beyond 80 accessions. Interestingly, the number of tag SNPs (haplotypes) reached a plateau much faster; the vast majority of haplotypes having been discovered within the first set of approximately 40-50 accessions. These results suggest that the current dataset offers an exhaustive characterization of the variants and haplotypes present in the elite Canadian soybean germplasm.

To test how well this reference panel of variants could serve as a reference panel to impute missing data in datasets derived from lower density genotyping tools, we used a set of ~150K

GBS-derived SNPs called on a set of 530 short-season soybean accessions from Canada. This set of 530 included all 102 accessions characterized by WGS. All tag SNPs that were present in the reference panel but were absent from the GBS-derived dataset (~1.5M SNPs) were imputed onto the GBS dataset. To allow us to estimate the accuracy of this imputation at previously untyped loci, the WGS data from a single accession (among the 102) were left out of the reference panel. Then, the imputed genotypes at untyped loci (not present in the GBS dataset) were compared to the actual genotypes revealed through WGS. Five such permutations were done by randomly selecting one accession for removal from the reference panel and imputation. On average, 96.4% of the missing genotypes imputed in this fashion proved to be imputed correctly. As for the 3.6% that were inaccurately imputed, these variants were located in regions with a high degree of haplotype diversity (i.e. low level of LD) and included several rare haplotypes that are difficult to correctly impute. Overall, this dataset provides an excellent reference panel for highly accurate imputation of untyped loci in elite short-season soybean.

VI.5.2 Structural variation

VI.5.2.1 Exploration and characterization

To produce a comprehensive catalogue of large SVs (deletions, duplications, inversions, translocations, and CNVs), we used a combination of three bioinformatics tools: LUMPY, BreakDancer and CNVnator. LUMPY using jointly multiple SV signals (read-pair, split-read and read-depth) was able to identify nearly all SV classes except interchromosomal translocations, while BreakDancer (paired-end SV detection method) was unable to detect small inversions and tandem duplications. CNVnator precisely discover and genotype CNVs (deletions, insertions and duplication) from depth-of-coverage by mapped reads. Using a combination of different tools allowed us to detect all classes of SVs, and also to do a cross-validation between outputs of these tools. Among the four types of SVs that were called by three tools (deletions, insertions, inversions, and duplications), 91, 87, 86, and 83% of all SVs were called by at least two tools. Thanks to the large predominance and high degree of concordance of deletions and insertions, the mean weighted concordance for these variants reached 89.6%. This result suggests that this catalogue of SVs is highly reproducible using various SV-calling tools. We produced a unified catalogue of SVs called by at least two of these three bioinformatics tools and these are described in Table VI.4. This catalogue comprises 63,556 deletions, 16,442 insertions, 2,865 duplications, 4,221 inversions, 1,435 copy-number variants, and 3,313

translocations (intra- or interchromosomal). Despite the fact that the size of these SVs spanned a broad range (10 bp to 3 Mb), these rearrangements were typically rather small. Indeed, the median size of the SVs varied between 106 bp (deletions) to 5.6 kb (CNVs). The breakpoints for these SVs could be defined with a variable level of resolution (ranging from 0 to 35 bp) depending on the type of SV. We estimated that deletions, the most abundant type of SV, affected 11.2 Mb (1.1%) of the soybean genome across all accessions examined. This catalogue of SVs is the first comprehensive characterization and classification of SVs in soybean and it illustrates the significance of the "footprint" of SVs on the soybean genome.

VI.5.2.2 Distribution and annotation of SVs

For illustrative purposes, we plotted the distribution of SVs on a single representative soybean chromosome, Chr 10 (Figure VI.4). To capture the full range of variant densities (no. of variants/window), a logarithmic scale was used. While the most abundant variants were distributed all along the length of this chromosome, CNVs seemed to cluster in certain regions. On the other hand, we saw no correlation between the number of SVs per chromosome and chromosome length (Supplementary Figure VI.4). To annotate and identify the potential functional impact of these SVs, we used an in-house script to identify genes residing within intervals defined by the SV breakpoints (for deletions, duplications and CNVs) or genes in which breakpoints were located (for inversions and translocations) (see M&M and Supplementary Figure VI.5 for details). Table VI.5 shows the number and proportion of the SVs which affected genic regions. In total, 19,424 deletions, 6,762 insertions, 2,023 duplications, 2,286 inversions, 995 CNVs, and 246 translocations impacted genic regions. Overall, 34.5% (31,735/91,832) of SVs were identified as affecting genes and all or almost all of these would be expected to have a strong impact on the function of these genes. Of this number, duplications and CNVs most often affected genic regions (70.6% and 69.3%, respectively), while translocations were the least likely to affect genes (8.2%). These results show that a much higher proportion of SVs are likely to have functional consequences than was the case for smaller variants (SNPs, MNPs and small indels).

VI.5.2.3 Validation of SVs and breakpoint

To estimate the sensitivity and the precision of the results, we selected 40 SVs of different sizes and frequencies within the population for PCR-based experimental validation. The SVs called on the basis of WGS reads were confirmed by PCR in 80% (32/40) of the cases

(Supplementary Table VI.3). In all eight cases where we could not confirm a SV by PCR, these were relatively rare, occurring in less than 7% of the lines. The mean size of these rare and unconfirmed SVs was also much larger than that of the successfully validated SVs (815 kp vs 8 kp). Interestingly, four PCR-validated SVs were shared by all 102 lines of this collection, suggesting one of three possibilities: 1) these variants are fixed in this particular set of short-season soybean, 2) the cultivar used to produce the reference genome (Williams 82) is atypical in its genome structure in these areas, or 3) the reference genome is imperfectly assembled in these regions. We examined the predicted breakpoints defining these SVs by performing Sanger sequencing on PCR amplicons spanning such breakpoints. Sanger sequencing results also confirmed the identified breakpoints at the nucleotide level.

Finally, we sought to examine if we could detect previously described SVs and if these were accurately called in the various accessions. At the E3 (GmPhyA3) locus, some early-flowering accessions are known to carry the e3-tr allele characterized by a 15.5-kb deletion that leads to a truncated and non-functional phytochrome. Similarly, at the E4 (GmPhyA2) locus, many early accessions carry the e4(SORE-1) allele characterized by the insertion of a 6.2-kb retroelement. In previous work, allele-specific primers had been used to precisely identify the alleles present at these two loci for 50 of the soybean lines used here and, in all cases, the SVs called on the basis of the WGS reads coincided perfectly with the PCR results (Supplementary Table VI.4, Supplementary Figure VI.6). In addition to a large degree of overlap between the SVs discovered by the three tools used, we were able to perform direct validation of some SVs that are highly relevant to breeders of short-season soybean.

VI.5.2.4 SVs and residual heterozygosity in soybean

Soybean elite lines are presumed to be highly inbred and, therefore, homozygous. Nonetheless, 3.2% of all genotypes were called heterozygous and, interestingly, a similar proportion was also called as heterozygous using the SoySNP50K array. We wanted to investigate the source of these heterozygous genotypes. Based on their distribution in the genome, these heterozygous genotypes could be qualified as dispersed or clustered. The latter group was almost systematically called heterozygous by both WGS and the array. In contrast, dispersed heterozygous genotypes, although less abundant (~25% of all heterozygous calls), tended not to be in agreement. Therefore, it was possible that some genomic feature could cause both WGS and the array to falsely call heterozygotes. We hypothesized that duplications and CNVs could be involved. As shown in Figure VI.5a, we saw that in the genomic regions showing a cluster of heterozygous calls, evidence of a duplication or CNV could be found in

the form of "excess" read coverage and extended across the same interval affected by heterozygosity. Accessions with the duplicated (or more) genomic segment invariably showed an abnormally high level of heterozygosity, while accessions with a single copy of this segment (as in the reference genome) showed a very low "background" level of heterozygosity (<1%) as seen elsewhere in the genome (Figure VI.5b). These results show that most residual heterozygosity observed in inbred lines is likely artefactual and the result of duplicated regions leading both the WGS and arrays to make erroneous heterozygous calls. The remaining (dispersed) heterozygous calls (<1% of all called genotypes) are likely a specific artefact of SNP calling based on WGS data. This observation of a tight link between duplicated regions and the occurrence of heterozygous SNP calls provided us with yet another opportunity to test the validity of our SV calls. We found that 89% of genomic regions that were indicated as being duplicated in specific accessions (based on SV-calling tools) coincided with regions showing a high level of heterozygosity in the same accessions. This result suggests that close to 90% of the duplications/CNVs called existed in the same set of accessions as those for which heterozygous calls were made.

VI.6 Discussion

A first key element to come out of this work is that SVs are a highly important contributor to DNA sequence differences in the soybean genome. We identified ~5M nucleotide and only ~92K SVs among 102 soybean accessions. At the first glance, there were 54-fold more nucleotide variants than SVs. In terms of the extent of their "fingerprint" or impact on the genome, however, SVs accounted for a greater proportion of the total nucleotide differences compared to nucleotide variants. Considering only "large" deletions (>10 bp), the former affected more than 1% of the soybean genome compared to less than 0.5% (4.35M SNPs and MNPs/1.1 Gb) for the nucleotide variants. Thus, the large deletions seem to affect two times more bases compared to all nucleotide variants in the soybean genome. Similarly, Sudmant et al. (2015) demonstrated that, in human genomes, a median of 8.9Mb of sequence are affected by SVs, compared to 3.6Mbp for SNPs. This illustrates the importance of characterizing SV, in addition to the nucleotide variants, in the sequenced genomes as these collectively make a very large contribution to the differences that distinguish various accessions within a species.

Beyond the simple quantitative contribution of SNPs and SVs, in terms of nucleotides affected per genome, it is also important to consider the functional impact of these various types of polymorphism. As described in this study (Figure VI.1 and Table VI.5), only 2% of nucleotide variants are located in coding regions, and barely 0.071% (4,113) were predicted to have a

high functional impact. In striking contrast, 34.5% of SVs or their breakpoints (close to 32k SVs) overlapped completely or partially with genic regions. As a result, a much larger number of genes may be affected functionally by SVs compared to SNPs. Currently, this very significant portion of functionally relevant genomic variation has been, for the most part, ignored in work aiming to identify variants underlying or in close proximity to variants responsible for the phenotypes of interest. Recently, in humans, Sudmant et al. (2015) demonstrated that SVs are enriched in haplotypes identified by genome-wide association studies and exhibit up to 50-fold enrichment among expression quantitative trait loci. In addition, these estimates of the impact of SVs on gene function are likely conservative as Lower et al. (2009) showed that SVs can affect the expression of genes up to 300 kb away from the variant whereas the effect of SNPs is generally much more local. We suggest that the collection of SVs identified in this study will help to dissect the genetic basis of important agronomic traits in soybean.

With the increasing cost-effectiveness of whole-genome sequencing projects, the amount of sequence information available to call variants can only increase with time. This requires a constant improvement in the efficiency and speed of SNP-calling tools to allow for the timely analysis of increasingly large amounts of sequence data. In addition, while many studies have reported on nucleotide variation in soybean and numerous other species, in our opinion, too little emphasis has been placed on assessing the accuracy of the resulting data. In this study, we used and compared a new bioinformatics analytical pipeline, Fast-WGS, that is able to efficiently and highly accurately call all three types of nucleotide variants (SNPs, MNPs and indels). In addition to being significantly more rapid (3.2 fold) than SOAPsnp, it resulted in a significantly more accurate dataset, especially with regards to small deletions. In previous studies, lower levels of genotype-calling accuracy (92–98%) have been reported, and only for SNPs (Hwang et al. 2015), whereas using Fast-WGS achieved similar or higher levels of accuracy for MNPs and indels. We suggest that using Fast-WGS to process existing WGS data would represent an improvement in the quality and quantity of nucleotide variants available to the research community.

In spite of extensive advancement of sequencing technologies and bioinformatics tools for sequence variant detection, the study of SVs has remained limited to human research (Sudmant et al. 2015; Stankiewicz et al. 2010; Lam et al. 2010). The main reason for this limitation is the fact that SVs are large-scale DNA rearrangements that present computational and bioinformatics challenges (Ye et al. 2016). We called SVs using a combination of three different tools; LUMPY, BreakDancer and CNVnator. These tools use one or a combination of

two to three major referenced-based mapping approaches (read depth, paired read or split read) to detect SVs (Layer et al. 2014; Chen et al. 2009; Abyzov et al. 2011). It is likely that none of these approaches by itself is sufficient to uncover all SVs (Carvalho et al. 2016). As reported previously, each approach has different strengths and weaknesses in SV detection, which depends on the type of SV or the properties of the underlying sequence at the SV locus (Tattini et al. 2015). Using a combination of different tools is important for several reasons; i) algorithms using a split-read approach can define rearrangement breakpoints, ii) algorithms exploiting read-depth data have the highest breakpoint resolution for smaller SVs, iii) a paired-read approach is highly powerful, but lower quality mapping assignments in repetitive regions is challenging and accurate prediction of SV breakpoints depends on very tight fragment size distributions (Quinlan et al. 2010). Alkan et al. (2011) showed that paired-read and split-read methods had the greatest extent of overlap (~67%) in terms of the SVs called, while read-depth and split-read approaches were the most discordant, with fewer than 20% of SVs detected by one approach detected by the other. It was found that the main differences in SV detection between these approaches were primarily in duplication- and repeat-rich regions, consistent with what we found in this study. We used these three complementary approaches to overcome the weakness of each approach.

As was done for nucleotide variation, we attempted to assess the reproducibility and the accuracy of SV dataset, although this is inherently much more challenging than for nucleotide variation due to the complex and large-scale nature of many rearrangements and the lack of an independent source of data on structural variation (such as CGH data) for this collection of accessions. A first indication of the quality of the SV data was the observation that close to 90% of all variants were called with more than one tool. In a second approach, we examined if the characteristics of the SVs uncovered in this work were similar to those reported in other species. In terms of the size and type of SVs, we found that 93% were less than 1 kb in size and that 69% of all SVs were deletions. Similarly, Mills et al. (2011) sequenced 185 human genomes and created a SV map that encompassed 22,025 deletions and 6,000 additional SVs, including insertions and tandem duplications. Furthermore, they reported that more than 90% of the discovered events were less than 1 kb in size and most of these were deletions rather than insertions. In a third approach, somewhat limited in scope, we performed a direct validation on a subset of SVs using PCR and Sanger sequencing. Here, again a high rate of validation was achieved as 80% of the 40 tested SVs were confirmed, with unconfirmed SVs being typically rare events. Finally, we exploited the fact that clusters of residual heterozygosity could be explained by duplication of the corresponding genomic regions to

perform a validation of duplications and CNVs. By using heterozygosity as a hallmark of duplicated regions, we found that close to 90% of predicted duplications and CNVs were validated in this fashion.

A final key finding of this work is that the joint study of nucleotide and structural variation not only can reveal biological but also technical complications. A frequent question that has been raised in previous studies on inbred lines or strains was the origin of the small fraction (2-5%) of heterozygous genotypes in genotype data. In this study we observed that the SVs (particularly duplications and CNVs) are the main reason for artefactual heterozygous genotype calls in soybean inbred lines. Duplicated regions can diverge and thus generate reads that are almost identical and that convincingly map onto regions that are present in single copy in the reference genome. This apparent diversity at specific positions in these mapped reads is erroneously taken to indicate heterozygosity. We feel it is highly likely that such artefactual heterozygotes will be encountered in many an inbred species and even in haploid organisms in which one would not expect to see any heterozygosity.

VI.7 Conclusion

We sequenced 102 elite soybean lines from Canada, the largest collection of elite soybean germplasm from a defined geographic region to be sequenced to date. This study is groundbreaking for several reasons: i) for the first time, we characterized all classes of structural variants in soybean; ii) we have presented a new analytical pipeline (Fast-WGS) that can facilitate and improve SNP-calling using WGS data; iii) the SNP haplotype collection shown in this study can be used as a reference panel to accurately impute missing genotypes at untyped loci in short-season soybean (the first such reference panel in soybean); iv) we have found an explanation for the residual heterozygosity at SNP loci; v) this resource combining both nucleotide and structural variants will help investigate phenotype-genotype associations in a more complete fashion in soybean.

VI.8 Acknowledgements

The authors wish to acknowledge the Génome Québec, Genome Canada, the government of Canada, the Ministère de l'Économie, Science et Innovation du Québec, Semences Prograin Inc., Syngenta Canada Inc., Sevita Genetics, Coop Fédérée, Grain Farmers of Ontario, Saskatchewan Pulse Growers, Manitoba Pulse & Soybean Growers, the Canadian Field Crop Research Alliance and Producteurs de grains du Québec.

VI.9 Tables

Table VI.1. Number of detected variants using two different WGS variant-calling pipelines (Fast-WGS and SOAPsnp).

Pipeline/Variants	SNPs	MNPs	Indels	Computing time*
Fast-WGS	4,071,378	284,836	642,015	81 hours
SOAPsnp	4,124,216	ND	512,418	261 hours

*Analysis was done using a Linux server with 64 CPU and 1Tb of RAM.

Table VI.2. Accuracy of genotype calls made using two WGS variant-calling pipelines (Fast-WGS and SOAPsnp). WGS-derived SNP genotypes were compared to the genotypes called at loci in common with the SoySNP50K array for the same samples.

Variants/Pipeline	Fast-WGS	Concordance (%)	SOAPsnp	Concordance (%)
Shared genotypes*	674,139		645,070	
Homozygous	668,672	99.7	641,215	97.1
Heterozygous	3,842	98.6	2,152	91.8
Indels	1,625	96.1	1,703	89.5

*Shared genotypes with the SoySNP50K dataset

Table VI.3. Accuracy of imputed missing data in the WGS SNP dataset. Imputed genotypes were compared to the genotypes called at loci in common with the SoySNP50K array for the same samples.

Variants	***WGS dataset***	***Imputation accuracy (%)***
Number of homozygous genotypes	594	***98.8***
Number of heterozygous genotypes	41	***92.7***
Total	635	***98.6***

Table VI.4. List of structural variant types identified in short-season soybeans and their characteristics.

SV type	Number of SV sites	SV size	Median size of SV (bp)	SV site breakpoint precision (bp)
Deletion	63,556	10bp-3Mb	106	±3*
Insertion	16,442	32bp-3Mb	144	±4*
Duplication (disperse duplication)	2,865	66bp-3Mb	2,513	±15†
Inversion	4,221	33bp-2.8Mb	116	±12‡
CNV (tandem duplication)	1,435	500bp-1.5Mb	5,623	-
Translocation (intrachromosomal)	3,011	30bp-2Mb	112	±6
Translocation (interchromosomal)	302	100bp-3Mb	4,523	±35

*Ascertained with split-reads

†Estimated for tandem duplications

‡Estimated for inversions with paired-end support from both breakpoints.

Table VI.5. Number of SVs located in genic regions based on their span or breakpoints.

SV type	Deletion	Insertion	Duplication*	Inversion	CNV†	Translocation‡
In gene	15,365	3,201	71	1,949	71	164
Upstream and gene	1,653	1,652	513	147	213	35
Downstream and gene	1,714	1,579	617	175	267	32
Whole gene	692	329	821	15	443	15
Total	19,424	6,762	2,023	2,286	995	246.6
Percent of all SVs affecting genes (%)	30.6	41.1	70.6	54.2	69.3	8.2

*Non-tandem duplication, †Tandem duplication, ‡Intrachromosomal translocation

VI.10 Figures

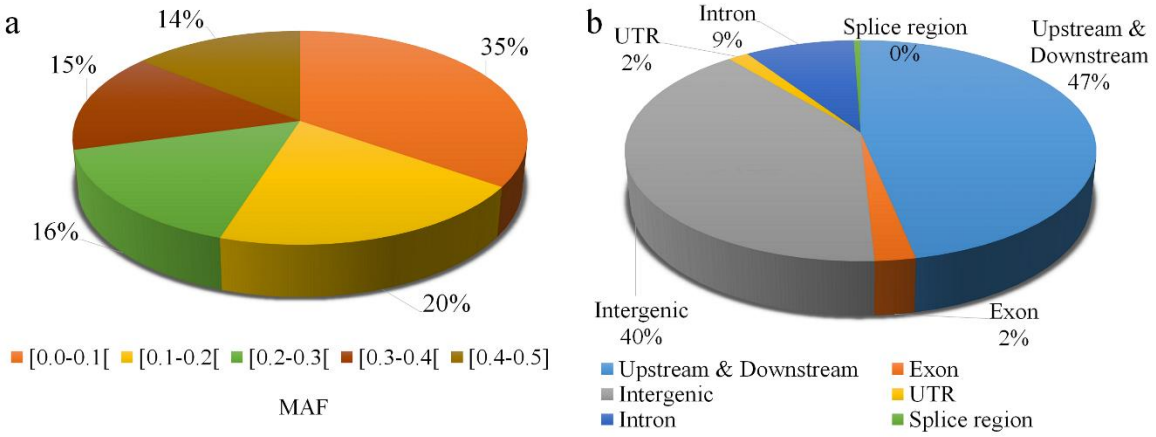


Figure VI.1. (a) Minor allele frequency (MAF) of variants. **(b)** Location of variants within the genome.

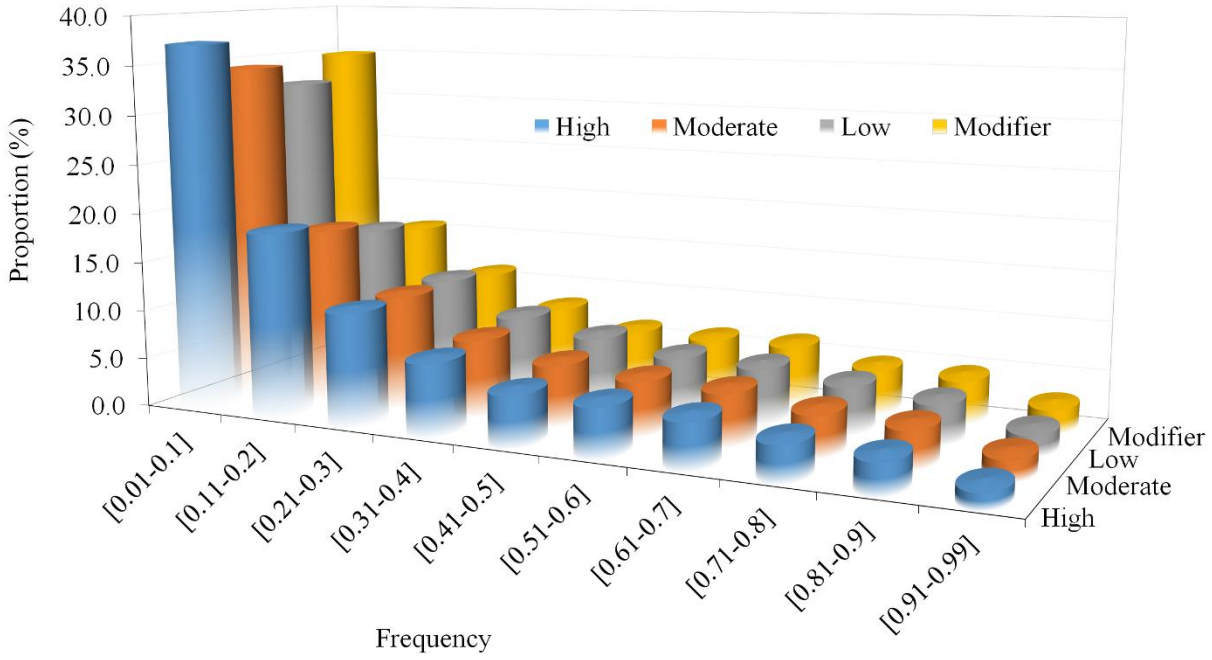


Figure VI.2. Distribution of variants with different degrees of predicted functional impact based on mutant allele frequency.

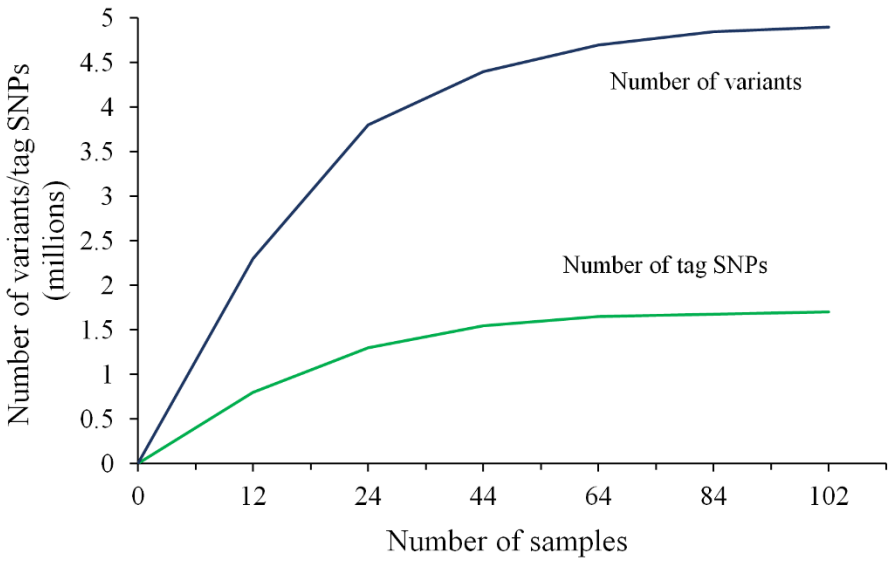


Figure VI.3. Number of variants (blue) and tag SNPs (green) based on different number of samples.

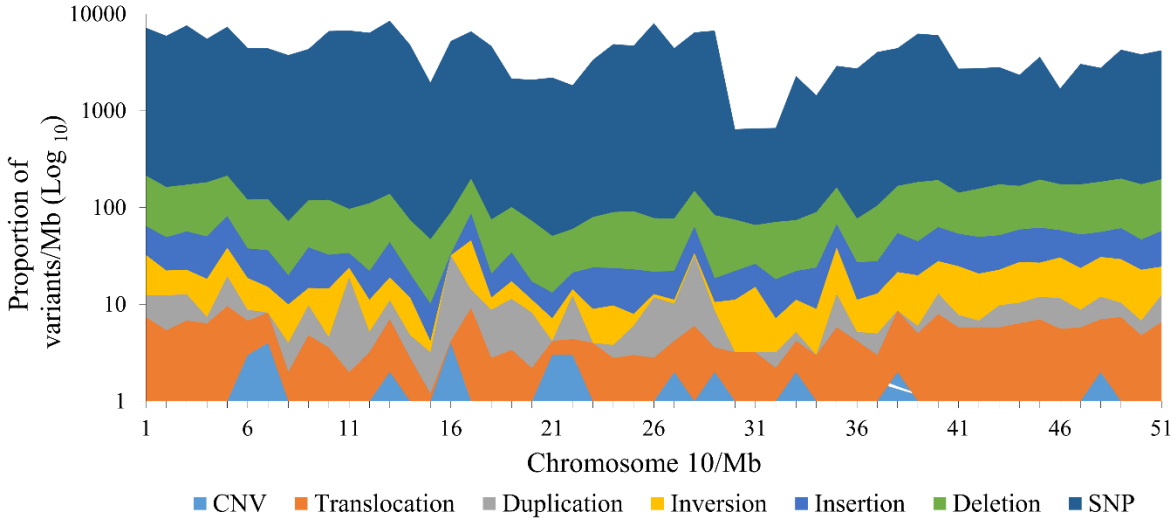


Figure VI.4. Distribution of SNPs and SVs on chromosome Chr10.

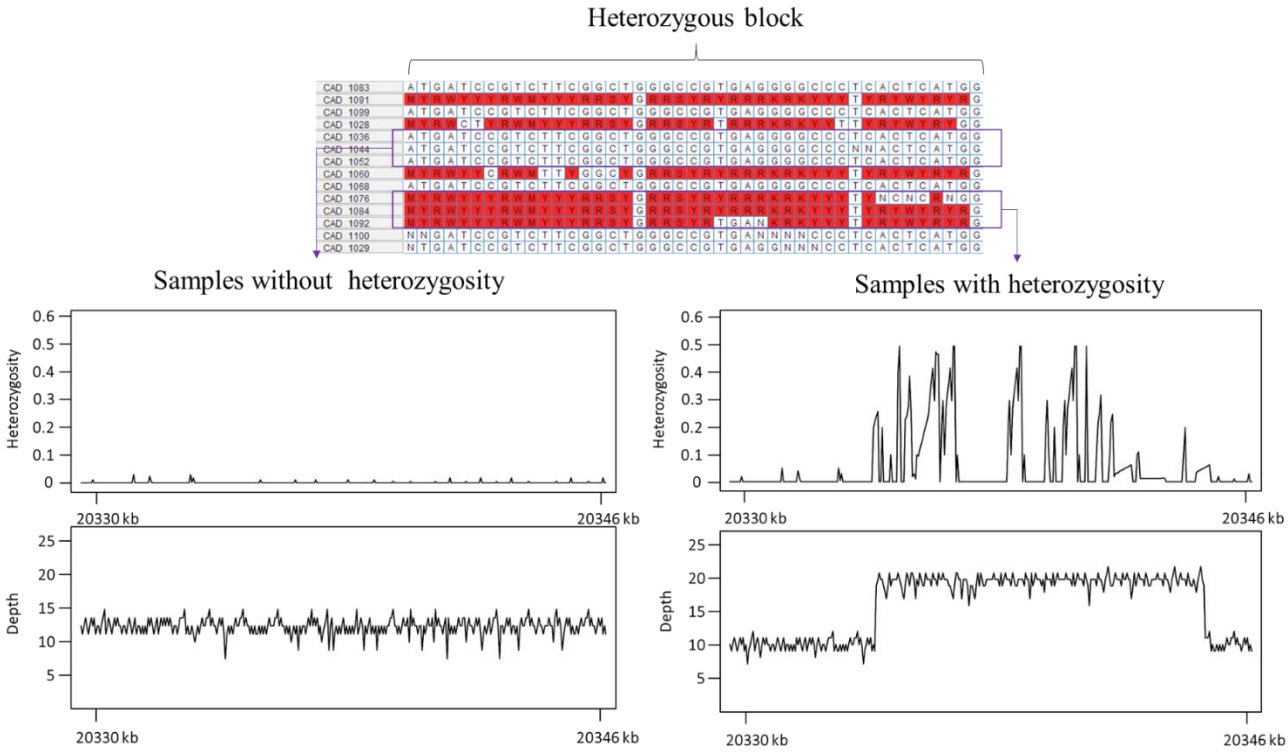


Figure VI.5. Plot of mapped-read depth and heterozygosity in a segment of chromosome Chr10 for which some lines exhibited clusters of heterozygous calls while other lines were homozygous.

VI.11 Supplementary files

Supplementary files listed and described below can be found online at

Additional file VI.1 Supplementary Text 1. Description of Fast-WGS. Bioinformatics analytical pipeline for whole-genome sequencing analysis.

Additional file VI.2 Supplementary Text 2. Significant contribution to the public SNP dataset (dbSNP) for *Glycine* spp.

Additional file VI.3 Supplementary Figure 1. Cladogram of 441 short-season soybean accessions from Canada produced using a set of close to 80k SNP markers. Arrows indicate the samples selected for whole-genome sequencing.

Additional file VI.4 Supplementary Figure 2. Distribution of allele frequency for sequence variants located in coding regions and predicted to have a high impact on gene function

Additional file VI.5 Supplementary Figure 3. Population genetics analysis. **a)** Phylogenetic tree using Neighbour Joining method, a *Glycine soja* line's used as outlier. b) Population STRUCTURE analysis using WGS SNPs dataset, representing the existence of five sub-populations in this collection. c) Principal component analysis (PCA) also represented five sub-groups (circled) which are correlated by five sub-population derived from STRACTURE analysis.

Additional file VI.6 Supplementary Figure 4. Correlation between number of SVs and chromosome length. Deletions (DEL), insertions (INS), copy-number variations (CNV), duplications (DUP), inversions (INV), and translocations (TRANS).

Additional file VI.7 Supplementary Figure 5. Different cases used to identify structural variants that could directly impact the function of a gene. (1) the SV resides entirely within a gene, (2 and 3) a SV encompasses at least part of a gene or one of its breakpoints lies within a gene (4) the SV completely encompasses a gene.

Additional file VI.8 Supplementary Figure 6. Visualized example of PCR-based genotyping of 10 samples for *E4* gene. *E4* is the wild type form and *e4* resides an insertion. These results also confirmed the WGS SV genotypes dataset.

Additional file VI.9 Supplementary Table 1. Information of sequenced short-season soybean accessions with name and number of trimmed reads (Phred score >32).

Additional file VI.10 Supplementary Table 2. List of genes containing variants predicted to have a high impact on gene function

Additional file VI.11 Supplementary Table 3. PCR-based validation of SVs called on the basis WGS data.

Additional file VI.12 Supplementary Table 4. Concordance of WGS-based genotyping and PCR-based genotyping results for a deletion in *E3* gene and an insertion in *E4* gene.

Additional file VI.13 Supplementary Table 5. Primers used for PCR-based SV validation.

Chapitre VII

A Systematic Analytical Approach to Rapidly Identify Candidate Domestication-Related Genes

Davoud Torkamaneh[1,2], Jérôme Laroche[2] and François Belzile[1,2]

[1]Département de Phytologie, Université Laval, Québec City, QC, Canada

[2]Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Québec City, QC, Canada

VII.1 Résumé

La domestication est un processus co-évolutif clé par lequel les humains ont considérablement modifié le composition génomique et l'apparence des plantes et des animaux. L'identification des gènes liés à la domestication reste très ardue. Dans cette étude, nous présentons une approche analytique systématique qui exploite deux progrès récents dans la génomique, le séquençage du génome entier et la prédiction des mutations de perte de fonction (LOF), afin de faciliter grandement l'assemblage d'un catalogue exhaustif de gènes candidats liés à la domestication. En utilisant les données de séquençage du génome entier pour 296 lignées cultivés (*G. max*) et 64 accessions de soja sauvage, nous avons identifié 10 792 variants LOF et 193 gènes qui sont uniquement fixés pour l'allèle LOF chez le soja domestiqué. Les données transcriptomiques existantes du soja nous ont amené à surmonter les défis analytiques liés aux duplications du génome entier et à identifier les gènes néo ou sous-fonctionnalisés. Cette approche systématique nous a permis d'identifier 130 gènes candidats liés à la domestication de manière efficace et rapide. Ce catalogue contient tous les gènes de domestication précédemment caractérisés chez le soja, ainsi que certains orthologues d'autres espèces cultivées. En outre, il comprend de nombreux autres nouveaux gènes candidats liés à la domestication. En générale, cette collection de gènes candidate liés à la domestication chez le soja est presque deux fois plus grande que la somme de tous les gènes candidats précédemment rapportés chez toutes les autres cultures. Nous croyons que cette approche systématique pourrait facilement être utilisée chez de nombreuses espèces.

VII.2 Abstract

Domestication is an important key co-evolutionary process through which humans have extensively altered the genomic make-up and appearance of both plants and animals. The identification of domestication-related genes remains very arduous. In this study, we present a systematic analytical approach that harnesses two recent advances in genomics, whole-genome sequencing and prediction of loss-of-function (LOF) mutations, to greatly facilitate the assembly of an exhaustive catalogue of domestication-related candidate genes. Using whole-genome sequencing data for 296 cultivated (*G. max*) and 64 wild soybean accessions, we identified 10,792 LOF variants, and 193 genes that are uniquely fixed for the LOF allele in domesticated soybeans. Existing soybean transcriptomic data led us to overcome analytical challenges associated with whole-genome duplications and to identify neo- or sub-functionalized genes. This systematic approach allowed us to identify 130 candidate domestication-related genes in an efficient and rapid way. This catalogue contains all of the previously well-characterized domestication genes in soybean, as well as some orthologues from other domesticated crop species. In addition, it comprises many additional novel candidate domestication genes. Overall, this collection of candidate domestication-related genes in soybean is almost twice as large as the sum of all previously reported candidate genes in all other crops. We believe this systematic approach could readily be used in wide range of species.

VII.3 Introduction

Domestication in a broad sense (including domestication, diversification and selective breeding) constitutes an ongoing 12,000-year-old evolutionary experiment that has vastly enhanced the reproductive output of crop plants and livestock, far beyond that of their wild ancestors to meet human needs (Zeder 2015). Up to date, humans, directly or indirectly, have domesticated more than 2,500 crop species (Meyer and Purugganan 2013) and the traits selected during domestication have varied depending on the species, as well as on the nature of human needs (Zeder 2012; Zeder 2015). At first, humans initiated domestication of food crops a part of a behavioral switch from food gathering to agriculture, a profound mutation in human civilization known as the Neolithic revolution (Meyer and Purugganan 2013).

The dissection of the genetic architecture of domestication traits in crop plants and livestock and the nature of selection have been a major subject of molecular genetic studies over the past two decades (Olsen and Wendel 2012). Most studies of domestication genes to date have been limited to obvious characters such as behavior (aggression), morphology (size, architecture) and physiology (maturity) (Zeder 2012; Olsen and Wendel 2012). In most crop species, very few genes and mutations underlying domestication have been described (Meye and Purugganan 2013). Furthermore, domestication has left molecular footprints in the genome of domesticated species. Thus far, two main approaches have been used to identify candidate loci involved in domestication.

In a first approach, a single causal gene controlling a specific domestication trait at a unique locus is cloned, often through tedious positional cloning, and these candidates are then functionally validated. Such studies, due to the requirements for high-density genetic maps and molecular markers, as well as considerable genetic resources, have been carried out in very few species (Meyer and Purugganan 2013). In crop plants, a review of over 60 such cloned domestication-related genes revealed, from 24 plant species, that most [51 genes (~85%)] alleles present in domesticated or elite lines were the result of loss-of-function (LOF) mutations due to single nucleotide variants (SNVs) and small insertions/deletions (InDels) (Meyer and Purugganan 2013). Additionally, from 51 genes, 31 genes (~68%) affected by multiple LOF mutations, that these authors concluded that, multiple LOF mutations suggests multiple processes of domestication.

Whole-genome duplication (WGD) or polyploidization is a common event in plants that has occurred multiple times over the past 200 million years of crop evolution (Panchy et al. 2016). In addition to WGDs, gene duplication, on the other hand, led to an abundance of duplicated

genes in plant genomes. On average, in plant genomes 65% of annotated genes have a duplicate copy (Panchy et al. 2016). Retention of extant pairs of duplicated genes to revert back to single copy occurred by LOF mutations due to SNVs and InDels. Distinguishing the genes affected by LOF due to WGD from domestication events represents many challenges and ambiguities.

In a second approach, researchers have conducted genome-wide scans to identify "domestication regions", regions exhibiting a marked decrease in genetic diversity among domesticated individuals or lines relative to their wild progenitors (Hufford et al. 2012; Axelsson et al. 2013). Furthermore, QTL mapping analysis, using genome-wide scans on a population derived from domesticated and wild progenitors for specific domesticated traits revealed the genomic regions controlling these traits (Palaisa et al. 2004). These are presumed to contain domestication-related genes, however, such genome-wide scans do not typically result in the identification of a specific candidate gene.

Evolutionary molecular biologists have proposed several criteria for the identification of domestication genes. According to these criteria, the function of a candidate domestication-related gene can be related to a domestication trait, furthermore it should present evidence of positive selection and a complete or near-complete fixation of at least one causal mutation (Meyer and Purugganan 2013). Here, we propose a systematic analytical approach that relies on the mining of whole-genome sequencing data to identify LOF mutations derived from domestication to rapidly uncover an extensive catalogue of candidate genes associated with soybean domestication with respecting to the domestication-related gene calling criteria.

VII.4 Materials and methods

VII.4.1 Whole-Genome Sequencing Data

Raw WGS data (100-bp, paired-end sequences, Illumina HiSeq) for all cultivated soybean samples was downloaded from the NCBI Sequence Read Archive (SRA) where it is stored under three accession numbers; SRP062245 (Valliyodan et al. 2016), SRP045129 (Zhou et al. 2015), SRP094720 (Torkamaneh et al. 2017b) and PRJNA294227 (Maldonado dos Santos et al. 2016). To cover world-wide soybean genetic diversity, we selected samples from diverse origins and cultivation areas (China, Japan, North/South Korea, Vietnam, Nepal, Russia, Sweden, Serbia, Brazil, Canada and United States) (Supplementary Table VII.1). In addition, we analyzed the WGS data for 64 *G. soja* samples from Zhou et al. (2015). All raw sequence data was processed with the Fast-WGS pipeline (Torkamaneh et al. 2017b), using Williams82

(*Gmax*_275_v2) as a reference genome. Variants were removed if 1) they had two or more alternate alleles, 2) an allele was supported only by reads mapping to one of the two strands, 3) the overall quality (QUAL) score of was <32, 4) the mapping quality (MQ) score was <30, 5) read depth of was <2, or 6) support for the two alleles was highly unequal (0.7).

VII.4.2 Variant Calling Validation

A subset of samples (35) had been previously genotyped using the SoySNP50K array (Song et al. 2013). We used the genotypes derived from the two different genotyping approaches (WGS and SNP array) and did a direct comparison of the genotypes called at these shared loci (>1.6M data points) to assess the accuracy of genotype calls.

VII.4.3 Variant Annotation

LOF variants were called using SnpEFF (Cingolani et al. 2012). Three groups of variants were called: stop-gain variants (premature stop codons), frameshift InDels and essential splice site–disrupting variants.

VII.4.4 Duplicated Gene Identification

We detected putative duplicated genes, presumably derived from WGD or gene duplication, using BLAST, DAGChainer, PAML and gene family member analysis (homology), as described by Grant et al.(2010) and Goodstein et al. (2012).

VII.4.5 Transcriptome Data

The complete transcriptome dataset for 26 tissues was downloaded from Phytozome database (Goodstein et al. 2012) for the genes affected by LOF and their duplicated copies identified in this study. We measured the expression level for these genes using FPKM (fragments per kilobase of exon per million fragments mapped) values. We declared that a gene was unexpressed when its FPKM value was equal to 0 or $-2\sigma_{mean}=\frac{\sigma}{\sqrt{N}}$ (defined for each tissue) (Supplementary Figure VII.3 and Supplementary Table VII.6). The same approach was applied to the other copies of the genes affected by LOF to determine their unique or similar expression pattern.

VII.4.6 Domestication Sweeps and QTLs

We identified 121 domestication sweeps reported in soybean based on WGS analysis of *G. max* and *G. soja* using F_{st} and XP-CLR tests (Zhou et al. 2015; Lam et al. 2010; kim et al. 2010; Li et al. 2013) (Supplementary Table VII.4). Furthermore, 31 previously reported domestication-related QTL regions in soybean (Zhao et al. 2015) (Supplementary Table VII.5) were also considered as regions of interest.

VII.5 Results

VII.5.1 Whole-Genome Variant Identification

The whole-genome sequencing (WGS) data (4.3×10^9 100- or 125-bp reads) for 296 cultivated soybean (*G. max*) accessions from Brazil (28), Canada (113), Asia (China, Korea, Japan) (96), and the USA (59), and 64 wild soybean (*G. soja*) accessions, with a median depth of coverage of 14× (Supplementary Table VII.1) were proceeded for variant calling. On average, a coverage of at least 1x was achieved for 958 Mb (excluding gaps), thus covering 98% of the *G. max* genome sequence. We identified more than 9 million sequence variants including: 7.5M SNVs (80%), 590K MNVs (6.3%) and 1.3M InDels (13.7%). To assess and compare the quality of genotype calls, we used a subset of samples which were previously genotyped using SoySNP50K array (details in Methods). The quality assessment of dataset represented an accuracy of 99.7% for SNVs and 96.1% for InDels.

VII.5.2 Prediction of Loss-Of-Function Variants

The functional impact of the sequence variants located in the 54,174 protein-coding genes of soybean were predicted using SnpEff. We observed 10,792 loss-of-function (LOF) variants (0.098%) that are predicted to severely impair protein synthesis or function (stop-gain variants, frameshift InDels and essential splice site–disrupting variants) (MacArthur et al. 2012) in 6,689 genes (12.3% of all genes). These mutations are the result of 4,087 SNVs (37.9%), 148 MNVs (1.3%) and 6,557 InDels (60.8%). Frameshift variants (6,147) were the predominant category, representing 57% of LOF mutations and affecting 4,126 genes. InDels (ranging from -50 bp to +32 bp) were, understandably, over-represented (4.5-fold) in the LOF category due to their high probability of resulting in a LOF allele. We found 4,586 genes with a single LOF and 2,103 genes with 2 or more LOF mutations (Table VII.1 and *SI Appendix, Fig. S1a*).

Protein enrichment analysis for these genes showed an enrichment for 15 protein domains, many of which were derived from transposable elements (Supplementary Table VII.2). On

the other hand, gene ontology (GO) enrichment analysis showed a significant enrichment in five GO groups: biological regulation, response to stimulus, cellular component organization, signaling and signaling process (*SI Appendix, Fig. S2*)**.**

In view of the fact that 31.4% of genes were affected by ≥2 LOFs, we estimated the cumulative frequencies of all LOF alleles for such genes. This is consistent with the recent work of Sedivy et al. (2017) showing that cultivated soybean originated from multiple domestication events. As a consequence, different LOF alleles could be predominant in different regions where soybean domestication and cultivation occurred. To identify candidate domestication genes, we estimated the cumulative frequency of all LOF alleles for each individual gene. We then classified this collection of genes affected by LOF variants in three major groups; i) 4,769 genes (71.3%) with a low cumulative frequency of LOF alleles (<20%); ii) 1,479 genes (22.1%) with an intermediate cumulative frequency (20%<F<80%); and iii) 441 genes (6.6%) fixed or nearly fixed for LOF alleles (≥80%) (*SI Appendix, Fig. S1b*). In total, most LOF mutations were rare, with 87% having a minor allele frequency (MAF) below 0.5%. In point of fact, LOF mutations were enriched for low-frequency alleles compared to synonymous and missense mutations.

VII.5.3 Identification of Domestication-Related Candidate Genes

We followed a systematic analytical approach illustrated in Figure VII.1 to identify candidate domestication genes based on the assumption that such genes would exhibit two important features: fixation for LOF allele/s only in domesticated samples and localization of LOF allele/s in single copy or uniquely expressed genes. From 6,689 genes, we identified 284 genes with fixed LOF allele/s and then excluded 91 genes with LOF mutations that were similarly abundant in *G. soja* accessions, as these would have played no role in the differentiation of domesticated soybeans. We then examined whether the remained 193 genes with fixed LOFs in domesticated soybeans were unique (single copy) or not in the soybean genome. Because of whole-genome duplication events (~59 and 13 million years ago) and tandem duplication events, most (75%) soybean genes have more than one copy (Schmutz et al. 2010). We categorized the genes affected by LOF mutations into two groups (unique vs. duplicated). We reasoned that a LOF mutation in a unique gene would necessarily result in phenotypic consequences. We found that from 193 genes, 32 genes (16.9%) affected by LOF in the domesticated soybean genome were unique. Conversely, we found that 161 genes (83.1%) with fixed LOF in domesticated soybeans had at least one other copy. This constitutes a significant enrichment (P < 0.001) compared to the genome-wide occurrence of gene

duplication. Of these genes 95 (59%) had a paralogue and 66 (41%) were the result of tandem duplication. From 161 genes, 30 genes were highly duplicated (more than 15 copies in the genome). We excluded these highly duplicated genes and conserved 131 genes for following steps. In the case of duplicated genes, LOF alleles could also have functional consequences if the mutated copy was uniquely expressed (through neo- or subfunctionalization) (*SI Appendix, Fig. S3*). We assessed this by examining transcriptomic data from 26 tissues (Severin et al. 2010; Libault et al. 2010) (*SI Appendix, Fig. S4*). The 98 genes having met all these criteria were deemed to be good candidates for domestication genes strictly on the basis of fixation for alleles predicted to result in a loss of function without possibility of complementation through another copy of the same gene. In total, we identified 130 domestication-related candidate genes (32 unique and 98 duplicated genes) (Supplementary Table VII.3), the most extensive catalogue of domestication-related candidate genes to date.

VII.5.4 Validation of Domestication-Related Candidate Genes

At first, we asked if these candidate domestication genes were located in genomic regions previously reported as harbouring potential domestication genes. To do this, we identified a set of 152 regions previously reported to contain domestication-related genes in soybean (Supplementary Table VII.4 & 5). Of these, 121 were domestication sweeps identified through genome-wide analysis of panels of unrelated lines. A further 31 QTL were identified through QTL mapping in crosses between wild and domesticated soybean accessions. These domestication-related regions span a total of ~110Mb (~10%) of the soybean genome and contain a very large number of genes (~9,695; 17.9% of soybean genes). Of the 130 candidate domestication genes identified previously, 60 resided within such genomic regions (Supplementary Table VII.3).

On the other hand, we would expect to find many of the already known domestication genes in soybean within this set. This is indeed the case. As shown in Table VII.2, the catalogue of putative domestication-related genes (a subset) produced through our systematic approach contains all known (i.e. functionally validated) domestication genes in soybean. These are involved in pod shattering (Glyma.16G019400 (NST1/2) and Glyma.16G141500 (GmPdh1)), pod color (Glyma.19G101700 (L1(MYB))) and growth habit (Glyma.19G194300 (GmTFL1b)) (Dong et al. 2014; Funatsuki et al. 2014; He et al. 2015; Bollman et al. 2003). Most importantly, we found and confirmed the known non-functional alleles in these genes. Furthermore, orthologues of well-characterized domestication genes identified in other

species such as the Arabidopsis GSL gene involved in growth habit (Glyma.04G192300) (Qian et al. 2014) and the maize Dwarf 1 gene involved in plant architecture (Glyma.04G168800) (Axelsson et al. 2013) are also part of this catalogue. Finally, we found that 16 genes not only had a known orthologue but also, they resided in domestication regions. We conclude that these 71 genes which were completely fixed for LOF alleles and have a known orthologue or resided in domestication regions constitute the strongest candidates for putative domestication genes. Above all, however, the other genes in this list comprise highly promising novel candidate genes on which future research can focus.

VII.6 Discussion

Here we describe a systematic analytical approach for the efficient and accurate identification of domestication-related genes. We assembled the most comprehensive catalogue of candidate domestication-related genes to date, with 130 soybean genes. This catalogue of genes is two-fold greater (130 vs. 60) than all known domestication genes in crop plants (as reviewed in Meyer and Purugganan 2013.), and five-fold greater (130 vs. 26) compared to the most highly studied domesticated crops (maize and rice). In the past, the identification of domestication genes has mostly been achieved via fine mapping and positional cloning within the progeny of a biparental cross (Meyer and Purugganan 2013). This presents three major limitations: i) it is restricted to the characterization of very few domestication traits per segregating progeny, ii) it is extremely demanding in terms of labor, cost and time to narrow down the genetic interval to one or a few candidate genes and iii) the validation of a candidate gene through functional complementation is challenging in species that are not easily transformed (Ross-Ibarra et al. 2007). In contrast, the approach developed here builds on often existing genomic data to systematically identify a highly-enriched catalogue of candidate domestication genes as well as providing, for each gene, an allelic series that can be helpful in further characterizing these candidate genes.

Two questions can be asked about this catalogue: 1) Is it missing any domestication-related genes (false negatives)? 2) Does it contain any genes that are not really related to domestication (false positives)? The fact that we captured all previously described, functionally validated soybean domestication genes, as well as all known alleles of these genes, suggests a low false-negative rate. Admittedly, given the small number of such genes in soybean (four), there is limited scope to answer this question more definitively. As for false positives, of the 126 candidate genes that were not already known to be related to

domestication, 27 (21.4%) are orthologues of known domestication genes in other crop species (Table VII.2). It can reasonably be argued that these played a similar role in soybean domestication. Assuming that the catalogue of known domestication genes in all crops is highly incomplete, the fact that not all of our soybean candidate domestication genes could be associated with an orthologue from another species should come as no surprise. Of the remaining 99 candidate genes, more than half (55, 56%) were found to reside in previously reported domestication-related regions (identified via the analysis of selection sweeps of QTL analysis). This constitutes a massive enrichment (p value = 1.7×10E-28) compared to the null hypothesis. Here again, we cannot assume that all domestication regions have been successfully identified. There is extensive literature which shows the limitations of classical population genetics methods, based on diversity, to detect selected sites (Bustamante et al.2001; Neher, 2013; Messer and Petrov, 2013; Good et al. 2014). It has also been shown that these models break down when the density of selected polymorphisms increases (Pritchard et al. 2010; Good et al. 2014). Taken together, these lines of evidence suggest that it is warranted to focus future work on this small, but promising set of candidate genes.

As described, we observed multiple LOF variants (mean = 3.4 LOFs/gene) in domestication-related genes, a number that is 5.5-fold higher than the genome-wide average (0.62 LOFs/gene). This suggests one of three possibilities: 1) these genes were under strong selection pressure, 2) multiple domestication events occurred in soybean, 3) both of the above. In many previous studies, it has been shown that crop plants experienced a strong selection pressure during the domestication process (Zohary, 2004; Gepts, 2004; Purugganan, and Fuller, 2009). Furthermore, it has been shown that strong artificial selection often results in the independent arisal of multiple spontaneous adaptive mutations, most of which are base substitutions (Hall, 1988). Furthermore, MacArthur et al. (2012) argued that gene inactivation occurred through the accumulation of multiple LOF variants rather than the increased frequency of a single LOF allele. Recent evidence suggests that cultivated soybean originated from multiple domestication events (Sedvy et al. 2017). Therefore, we propose that the third possibility is likely for soybean and helps explain the observation of multiple LOF variants in its domestication genes.

An intriguing, but as yet unexplored finding, is that several genes were fixed for LOF variants present in both wild and domesticated soybean. In principle, these could represent genes whose inactivation contributed to the speciation, either of *G. max* and *G. soja* away from other members of the *Glycine* genus or at earlier stages of speciation. Such earlier events in plant speciation have so far received little attention.

We recognize at least three limitations to our approach. First, this approach is based only on LOF variants. Although this class of variant is known to be predominant, this single type of variant cannot cover all possible types of mutations. For example, it is estimated that around 7% of known domestication genes are characterized by gain-of-function variants (Meyer and Purugganan 2013). These are much more difficult to identify as the acquisition of a new function does not require that the pre-existing form (found in wild relatives) was itself non-functional. Such domestication-related mutations would not be detected using our approach. Secondly, extensive transcriptome data is not always available for all species. In the absence of such information, it can be difficult to determine if duplicate copies of a gene are expressed in the same tissues or whether there is evidence for neo- or sub-functionalization. In crop plants, as polyploidization and whole-genome duplication have played major roles in evolution, this could represent an important limitation until transcriptomic data become available in sufficient quantity. Finally, this approach is based on nucleotide variants and small indels. Structural variants have been presented as large genetic variants (>50bp) that can play a role in crop domestication. For example, it has been comprehensively demonstrated that an insertion of a transposable element in the regulatory region of *teosinte branched1* (tb1) gene is the main contributor to the increase in apical dominance during maize domestication (Tsiantis, 2011). Furthermore, pan-genome analysis of soybean accessions (one *G. max* and seven *G. soja*), based on *de novo* assembly, showed a total of 1.86 Mb of *G. max*–specific present/absent variants (>100 bp) that can be related to soybean domestication (Li et al. 2014). Overall, large structural variants related to the domestication would not be detected using our approach.

VII.6 Conclusion

In conclusion, we propose here a robust and rapid approach to detect putative domestication-related candidate genes that relies on the ever-increasing amount genomics data that is accumulating for a large number of domesticated species. The genes identified in this study comprise highly promising novel candidates on which future research and further characterization can focus. Using soybean as a model for this approach allowed us to resolve many of the challenges one would expect to encounter, such as polyploidy and a high level of gene duplication. For these reasons, we believe this approach could readily be used in a wide range of species.

VII.8 ACKNOWLEDGMENTS

The authors wish to acknowledge the Génome Québec, Genome Canada, the government of Canada, the Ministère de l'Économie, Science et Innovation du Québec, Semences Prograin Inc., Syngenta Canada Inc., Sevita Genetics, Coop Fédérée, Grain Farmers of Ontario, Saskatchewan Pulse Growers, Manitoba Pulse & Soybean Growers, the Canadian Field Crop Research Alliance and Producteurs de grains du Québec.

VII.9 Tables

Table VII.1. Number of loss-of-function variants by sequence ontology (SO).

SO term	SNVs	MNVs	INSs	DELs	Total variants	Genes
Splice donor	599	17	110	75	801	741
Splice acceptor	853	22	77	87	1,039	949
Stop gain	2,482	94	24	1	2,601	2,150
Frameshift	0	0	1,773	4,377	6,150	4,130
Start loss	150	15	21	12	198	185
Stop loss	3	0	0	0	3	3
Total	**4,087**	**148**	**2,005**	**4,552**	**10,792**	**6,689**

SNVs: single-nucleotide variants; MNVs: multiple-nucleotide variants; INSs: small insertions; DELs: small deletions.

Table VII.2. The list of domestication-related candidate genes with known orthologues in soybean and other species.

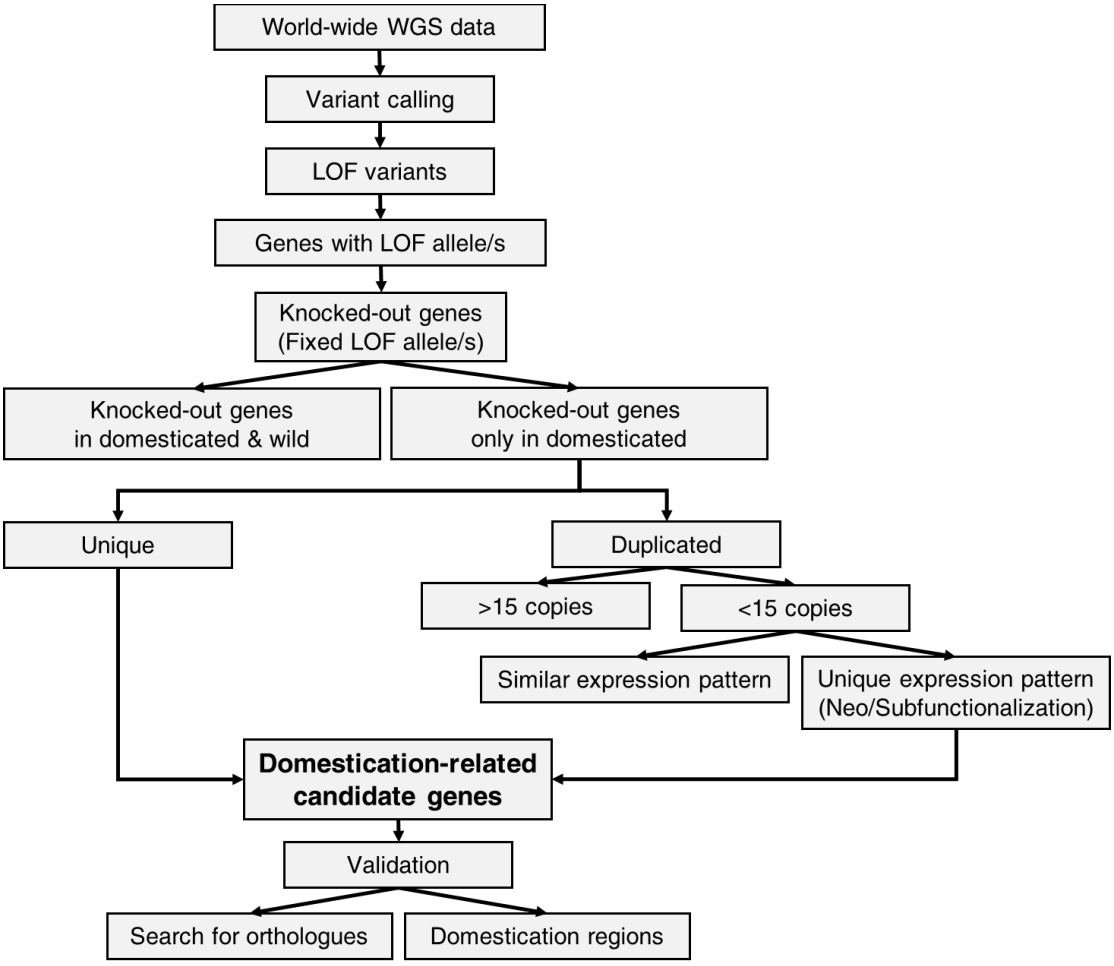
Domestication-related trait	*Domestication-derived trait possible function*	*Gene ID*	*Gene function*	*Orthologues*
Plant growth habit	Determinate growth	Glyma.04G192300	Callose synthase	GSL
		Glyma.02G182100	DNA helicase	PIF1 helicase
		Glyma.01G209200	Regucalcin	RGPR
		Glyma.05G193900	Regulator	SWI
		Glyma.01G232400	CoA synthase	ACLB-2
		Glyma.01G205300	Vesicle associated protein	VAMP
	Maturity	Glyma.02G286700	Pectinesterase	PGR95–094
	Flowering time	Glyma.06G206500	Transcription factor	tfiid
		Glyma.02G215000	Vernalization	VIN3
		Glyma.03G019900	Hydroxypyruvate	HPR1
Plant architecture	Plant height	Glyma.06G205500	COBRA-like protein	Culm1
		Glyma.02G292200	MutS homolog	MSH1
		Glyma.05G057800	Lipid transferase	PLTP
		Glyma.07G091700	ATP binding	Kinesin
	Maximum internode length	Glyma.04G168800	Translation factor	Dwarf 1
	Stem determinacy	Glyma.19G194300*	Transcription cofactor	GmTFL1b
		Glyma.02G005900	BED zinc finger	zf BED
		Glyma.01G202400	GTPase	Rab5
	Twining habit	Glyma.07G078700	Xyloglucan	PsXTH1
Seed	Shattering	Glyma.16G019400*	Transcriptional regulator	NST1/2
		Glyma.16G141500*	Dirigent protein	Pdh1
	Size	Glyma.05G160000	Transferase	WD40

		Glyma.06G113700	Ribosomal protein	s3a
		Glyma.05G051900	Abscisic acid	Phaseic acid
	Pod color	Glyma.19G101700*	Transcriptional regulator	L1 (MYB)
Physiology	Stress adaptation	Glyma.09G048100	Arabinogalactan	FLAs
		Glyma.03G227300	Regulation of transcription	PAS fold
		Glyma.20G159300	Oxalyl-coa synthetase	AAE3
		Glyma.04G195600	Cation	Ca2+ chanel
		Glyma.03G056600	Replication factor	RFA1
		Glyma.06G200200	Sugar transporter	Sweet16

* Known functionally validated domestication genes in soybean.

VII.10 Figures

Figure VII.1. Systematic approach used to investigate the possible impact of LOF mutations in domestication process.



Chapitre VIII
Conclusion générale

Au cours des dernières années, les technologies de séquençage de nouvelle génération (NGS) ont joué un rôle clé dans la révolution de la phytogénétique. La phytogénétique moderne et la génomique reposent sur une connaissance approfondie des gènes et de leur fonctionnement dans les cellules que la génétique antérieure, y incorporant la contribution des facteurs environnementaux et épigénétiques qui jouent un rôle important dans le développement des caractères génétiques. En général, plusieurs axes de recherche différents ont été proposés pour la phytogénétique moderne : le génotypage du génome, les études d'association génomique, la sélection génomique, l'adaptation et la re-domestication. Le travail présenté dans cette thèse nous a permis d'étudier tous les axes de la phtytogénétique moderne en exploitant les plus récentes avancées méthodologiques dans le domaine des NGS.

L'utilisation optimale des ressources génomiques est une question importante en phytogénétique. Une ressource génomique pourrait être créée par séquençage de centaines d'échantillons. Cette ressource, maintenant disponible pour toute la communauté scientifique du soja, fournit des informations critiques sur toutes les catégories de variants génétiques et leur impact fonctionnel. Cette ressource a déjà été utilisée dans plusieurs recherches appliquées différentes chez le soja, comme nous le décriront brièvement dans les sections suivantes.

Pour illustrer de quelle manière les ressources génomiques développées au cours de cette thèse ouvrent la voie à des recherches novatrices, nous allons fournir quelques exemples tirés de travaux réalisés au sein du projet SoyaGen, un projet pan-canadien en génomique fonctionnelle du soja. Un premier exemple porte sur l'identification des allèles et leur caractérisation. Tardivel et al. (en rédaction) ont utilisé les données GBS et de re-séquençage des lignées canadiennes pour identifier les allèles présents au sein de cette collection chez quatre gènes importants contrôlant la maturité chez le soja. On y démontre une approche analytique systématique permettant d'extraire, à partir d'un catalogue de SNP issu d'un génotypage GBS peu coûteux, une liste des allèles pour chaque gène. Une telle approche, centrée sur un gène d'intérêt, a permis non seulement de découvrir de nouveaux allèles, jusqu'alors inconnus, mais aussi de déterminer les allèles présents chez chaque lignée. Cela fournit aux sélectionneurs une information qui est beaucoup plus pertinente et utile qu'un simple catalogue contenant des milliers ou des millions de marqueurs SNP.

Un autre champ d'investigation où les outils et ressources développés seront très utiles est celui de la recherche de QTL et de gènes contrôlant les caractères d'intérêt. Bien que la cartographie QTL soit une approche analytique employée depuis des décennies, sa résolution

a été grandement augmentée avec l'avènement des analyses d'association pan-génomiques ou GWAS (pour « Genome-wide association analysis ») (Brachi et al. 2011.). Comme nous l'avons décrit précédemment, l'analyse d'un grand nombre de lignées non-apparentées avec un nombre très élevé de marqueurs SNP permet d'identifier des variants qui montrent une association très étroite avec le caractère étudié. En raison de la résolution accrue des analyses GWAS, il devient possible parfois d'identifier un gène candidat. Oeuvrant toujours au sein de l'équipe SoyaGen, Boudhrioua et al. (en rédaction) ont profité des données de re-séquençage des lignées canadiennes pour réaliser une imputation des génotypes manquants au sein d'une collection de lignées qui avaient été génotypées par GBS et caractérisées pour leur réaction à un agent pathogène, *Sclerotinia sclerotiorum*, l'agent responsable de la sclérotiniose chez le soja. L'imputation exhaustive des marqueurs a produit un catalogue de plus d'un million de marqueurs SNP et a offert une couverture dense et complète du génome. Grâce à cette couverture exhaustive, de nouveaux QTL très prometteurs ont été identifiés. Ici encore, les sélectionneurs pourront bénéficier des retombées de ces travaux sous la forme de marqueurs SNP associés à la tolérance à cette maladie et cela facilitera grandement le développement de nouvelles variétés dotées d'une résistance accrue à cette maladie.

En plus de cataloguer de manière complète les variants génétiques présents au sein de collections de lignées de soja, j'ai présenté l'impact fonctionnel de ces variants. Une des conclusions importantes tirées de ces travaux est que les variants structuraux, bien que beaucoup moins nombreux que les variants nucléotidiques, ont potentiellement un impact beaucoup plus grand sur la fonction des gènes. À ce jour, l'analyse génétique chez les plantes ou les animaux était basée presque exclusivement sur des variants nucléotidiques, tandis que, comme on le voit, les variants structuraux ont un impact fonctionnel plus important que les variants nucléotidiques. Ce résultat souligne l'importance et l'urgence de mieux décrire et de mieux tenir compte de ces variants en analyse génétique. Par exemple, en ce moment, les analyses GWAS se font strictement à l'aide de variants nucléotidiques alors qu'il est clair que les variants structuraux sont cause de nombreux allèles qu'on cherche à découvrir. Pour cela, il sera nécessaire de développer de nouveaux outils de bioinformatique et statistiques pour incorporer les informations sur les variants structuraux.

Un autre domaine « chaud » en phytogénétique est celui de la sélection génomique ; un concept selon lequel il serait possible de prédire la performance d'une plante sur la seule base de son bagage génétique (Bhat et al. 2016.). Les méthodes de sélection génomique sont conçues pour prédire des caractères tels que le rendement en grains, la qualité et la résistance aux stress abiotiques et biotiques. La précision de la prédiction est importante à chaque étape

du programme de sélection et les méthodes de reproduction sont conçues pour améliorer la précision de ces prédictions. Les nouvelles stratégies de sélection sont guidées par la technologie et les nouvelles connaissances, et la prédiction de la performance d'un individu repose maintenant sur des données génotypiques. Comme on peut l'imaginer, ce travail fait appel à l'analyse d'un très grand nombre d'individus (des milliers ou des dizaines de milliers) et nécessite des outils rapides et efficaces pour l'analyse d'une quantité énorme de données NGS. En partie grâce aux travaux réalisés lors de cette thèse, les outils sont maintenant disponibles pour toute la communauté scientifique. Le pipeline Fast-GBS n'est pas limité au soja et a déjà commencé à être utilisé chez une large gamme d'espèces végétales et animales pour lesquelles il existe un génome de référence. À titre d'exemples, nous pouvons citer les travaux de Tardivel et al. (en rédaction) sur le soja, d'Abed et al. (en rédaction) chez l'orge et de Tekeu et al. (en rédaction) chez le blé. Ainsi, des travaux en sélection génomique comptent déjà parmi les retombées immédiates de cette thèse.

Finalement, en dehors de la phytogénétique appliquée, de nombreux champs d'investigation en génétique végétale plus fondamentale pourront bénéficier des fruits de cette thèse. Comme nous l'avons décrit, il y a plusieurs questions liées au processus de la sélection naturelle et de la domestication des espèces cultivées. Une des principales questions est de savoir si les allèles sélectionnés lors de l'adaptation (domestication) étaient les meilleurs (dans le contexte de l'agriculture par exemple) et si les mutations donnant lieu à ces allèles ont été capturées dans des fonds génétiques optimaux ? Pour pouvoir répondre à ce genre de questions, il faut d'abord identifier quels sont ces gènes qui ont contribué à la profonde transformation des espèces sauvages pour en faire des espèces domestiquées. Ici encore, les travaux décrits dans cette thèse rapportent une avancée majeure des connaissances. L'ensemble des gènes découverts au cours de ce travail représente une ressource unique pour commencer à répondre aux différentes questions liées à l'évolution du soja. De plus, l'approche développée ici peut facilement s'appliquer à un large éventail d'espèces. Nous espérons que ce travail contribuera à une amélioration significative de l'état des connaissances de la communauté scientifique.

Bibliographie

1. Abyzov A, Urban AE, Snyder M, Gerstein M (2011) CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Research. 21(6):974-984.
2. Adessi C, Matton G, Ayala G, Turcatti G, Mermod JJ, Mayer P, Kawashima E (2000) Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. Nucleic Acids Research. 28(20), e87.
3. Aflitos et al (2014) Exploring genetic variation in the tomato (Solanum section Lycopersicon) clade by whole-genome sequencing. The Plant Journal.doi.10.1111/tpj.12616.
4. Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. Nature Reviews Genetics. 12(5):363-376. doi:10.1038/nrg2958.
5. Andolfatto P et al (2011) Multiplexed shotgun genotyping for rapid and efficient genetic mapping. Genome Res. 21, 610–617.
6. Ardlie KG, Kruglyak L, Seielstad M (2002) Patterns of linkage disequilibrium in the human genome. Nat Rev Genet. 3(4):299-309.
7. Axelsson E et al (2013) The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature 495, 360–364.
8. Baird NA, Etter PD, Atwood TS, et al. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE, 3, e3376.
9. Beissinger TM, Hirsch CN, Sekhon RS, Foerster JM, Johnson JM, Muttoni G, … de Leon N. (2013). Marker Density and Read Depth for Genotyping Populations Using Genotyping-by-Sequencing. Genetics, 193(4), 1073–1081. http://doi.org/10.1534/genetics.112.147710
10. Bollman KM et al (2003) HASTY, the Arabidopsis ortholog of exportin 5/MSN5, regulates phase change and morphogenesis. Development. 130(8):1493-504.
11. Bhat JA, Ali S, Salgotra RK, et al (2016) Genomic Selection in the Era of Next Generation Sequencing for Complex Traits in Plant Breeding. Frontiers in Genetics. 7:221.
12. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007). TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 23(19):2633–5.
13. Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW (2013) Shining a Light on Dark Sequencing: Characterising Errors in Ion Torrent PGM Data. PLoS Comput Biol. 9(4): e1003031. doi:10.1371/journal.pcbi.1003031.
14. Brachi B, Morris GP, Borevitz JO (2011) Genome-wide association studies in plants: the missing heritability is in the field. Genome biology. 12(10):1–8.
15. Browning S, and Browning B (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81, pp. 1084–97.
16. Bustamante CD, Wakeley J, Sawyer S, Hartl DL (2001) Directional selection and the site-frequency spectrum. Genetics 159: 1779–1788.
17. Campbell PJ et al. (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nature Genet. 40, 722–729.

18. Carvalho CMB, Lupski JR (2016) Mechanisms underlying structural variant formation in genomic disorders. Nature Genetics. 2016; doi:10.1038/nrg.2015.25.
19. Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. Mol Ecol. Jun 22(11):3124–40. doi: 10.1111/mec.12354.
20. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS et al. (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat. Methods. 6:677-681.
21. Cheung CY, Thompson EA, Wijsman EM (2013) GIGI: an approach to effective imputation of dense genotypes on large pedigrees. Am J Hum Genet. 92: 504–516. pmid:23561844.
22. Chiang C, Layer RM, Faust GG et al. (2015) SpeedSeq: Ultra-fast personal genome analysis and interpretation. Nature Methods. 12(10):966-968. doi:10.1038/nmeth.3505.
23. Church GM (2006) Genomes for all. Sci. Am. 294 (1): 46–54. doi:10.1038/scientificamerican0106-46.
24. Cingolani P, Platts A, Wang LL et al. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly. 6(2):80-92. doi:10.4161/fly.19695.
25. Collard BC, Mackill DJ (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philosophical Transactions of the Royal Society B: Biological Sciences, 363(1491), 557–572. http://doi.org/10.1098/rstb.2007.2170.
26. Cook DE, Lee TG, Guo X, Melito S, Wang K, Bayless AM et al. (2012) Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. Science. 30;338(6111):1206-9. doi: 10.1126/science.1228746.
27. Crossa J, Beyene Y, Kassa S, Pérez P, Hickey JM, Chen C, de los Campos G, Burgueño J, Windhausen VS, Buckler E, Jannink J-L, Cruz MAL, Babu R (2013) Genomic Prediction in Maize Breeding Populations with Genotypingby. Sequencing. G3 3:1903–1926. doi: 10.1534/g3.113.008227.
28. Daetwyler HD et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nature Genetics. 2014; 46,858–865. doi:10.1038/ng.3034.
29. Danecek P, Auton A, Abecasis G, Albers CA, Banks E et al. (2011) The Variant Call Format and VCFtools. Bioinformatics. doi: 10.1093/bioinformatics/btr330.
30. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature. doi:10.1038/nrg3012.
31. Delaneau O, Marchini J (2014) The 1000 Genomes Project Consortium. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. Nature Communications. 5 3934.
32. Delaneau O, Zagury JF, Marchini J (2013) Improved whole chromosome phasing for disease and population genetic studies. Nature Methods.10 (1): 5–6. doi:10.1038/nmeth.2307.

33. Deschamps S, Llaca V, May GD (2012) Genotyping-by-Sequencing in Plants. Biology, 1(3), 460–483. http://doi.org/10.3390/biology1030460.
34. Di Giusto D, King GC (2003) Single base extension (SBE) with proofreading polymerases and phosphorothioate primers: improved fidelity in single-substrate assays. Nucleic Acids Research, 31(3), e7.
35. Dobzhansky T (1970) Genetics of the evolutionary process. New York: Columbia Univ. Press. ISBN 0-231-02837-7.
36. Donato MD, Peters SO, Mitchell SE, Hussain T, and Imumorin IG (2013) Genotyping-by-Sequencing (GBS): A Novel, Efficient and Cost-Effective Genotyping Method for Cattle Using Next-Generation Sequencing. PLoS One. 8(5): e62137. PMCID: PMC3656875.
37. Dong, Y et al. (2014) Pod shattering resistance associated with domestication is mediated by a NAC gene in soybean. Nature Communications 5: 3352.
38. Edwards D, Batley J, Snowdon RJ (2013) Accessing complex crop genomes with next-generation sequencing. Theor Appl Genet. 126: 1–11.
39. Edwards SL, Beesley J, French JD, Dunning AM (2013) Beyond GWASs: Illuminating the Dark Road from Association to Function. American Journal of Human Genetics, 93(5), 779–797. http://doi.org/10.1016/j.ajhg.2013.10.012.
40. Ellinghaus D, Schreiber S, Franke A, Nothnagel M (2009) Current software for genotype imputation. Human Genomics. vol 3. no 4. 371–380.
41. El-Metwally S, Ouda OM, Helmy M (2014) New Horizons in Next-Generation Sequencing. Next Generation Sequencing Technologies and Challenges in Sequence Assembly. Springer Briefs in Systems Biology. Volume 7. pp 51-59.
42. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES et al. (2011) A robust, simple genotyping- by-sequencing (GBS) approach for high diversity species. PLoS ONE. 6: e19379. doi: 10.1371/journal.pone.0019379. doi: 10.1371/journal.pone.0019379 PMID: 21573248.
43. Endelman J (2015) Genotyping-By-Sequencing of a Diploid Potato F2 Population. https://pag.confex.com/pag/xxiii/webprogram/Paper15683.html
44. Esch E, Szymaniak JM, Yates H, Pawlowski WP, Buckler ES (2007) Using Crossover Breakpoints in Recombinant Inbred Lines to Identify Quantitative Trait Loci Controlling the Global Recombination Frequency. Genetics, 177(3), 1851–1858. http://doi.org/10.1534/genetics.107.080622.
45. Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA et al. (2007) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. PLoS ONE. 3(10): e3376. doi:10.1371/journal.pone.0003376.
46. Etter PD, Bassham S, Hohenlohe PA, Johnson EA, Cresko WA(2011) SNP discovery and genotyping for evolutionary genetics using RAD sequencing. Methods in Molecular Biology. 772, 157–178.
47. Felcher KJ, Coombs JJ, Massa AN, Hansey CN, Hamilton JP et al. (2012) Integration of Two Diploid Potato Linkage Maps with the Potato Genome Sequence. PLoS One. 7(4): e36347.
48. Fenselau de Felippes F, Schneeberger K, Dezulian T, Huson DH, Weigel D (2008) Evolution of Arabidopsis thaliana microRNAs from random sequences. RNA, 14(12), 2455–2459. http://doi.org/10.1261/rna.1149408.

49. Fu YB (2014) Genetic diversity analysis of highly incomplete SNP genotype data with imputations: an empirical assessment. G3 (Bethesda). 13;4(5):891-900. doi:10.1534/g3.114.010942.
50. Fu YB, Cheng B, Peterson GW (2014) Genetic diversity analysis of yellow mustard (Sinapis alba L.) germplasm based on genotyping by sequencing. Genet. Resour. Crop Evol. 61: 579–594.
51. Fu YB, Peterson GW (2011) Genetic diversity analysis with 454 pyrosequencing and genomic reduction confirmed the eastern and western division in the cultivated barley gene pool. Plant Gen. 4: 226–237.
52. Fu YB, Peterson GW (2012) Developing genomic resources in two Linum species via 454 pyrosequencing and genomic reduction. Mol. Ecol. Resour. 12: 492–500.
53. Funatsuki H et al. (2014) Molecular basis of a shattering resistance boosting global dissemination of soybean. Proceedings of the National Academy of Sciences, USA 111: 17797–17802.
54. Ganal MW, Polley A, Graner EM, Plieske J, Wieseke R, et al. (2012) Large SNP arrays for genotyping in crop plants. J Biosci. Nov;37(5):821-8.
55. Gao L, Turner MK, Chao S, Kolmer J, Anderson JA (2016) Genome Wide Association Study of Seedling and Adult Plant Leaf Rust Resistance in Elite Spring Wheat Breeding Lines. PLoS ONE, 11(2), e0148671. http://doi.org/10.1371/journal.pone.0148671.
56. Gepts P (2004) Crop domestication as a long-term selection experiment. Plant Breed. Rev. 24, 1–44.
57. Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, et al. (2014) TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. PLoS ONE 9(2): e90346. doi: 10.1371/journal. pone.0090346.
58. Glodzik D, Navarro P, Vitart V, Hayward C, McQuillan R, et al. (2013) Inference of identity by descent in population isolates and optimal sequencing studies. Eur J Hum Genet 21: 1140–1145. pmid:23361219.
59. Golan D, Medvedev P (2013) Using state machines to model the Ion Torrent sequencing process and to improve read error rates. Bioinformatics. 29 (13): i344-i351. doi: 10.1093/bioinformatics/btt212.
60. Gompert Z, Forister ML, Fordyce JA, Nice CC, Williamson RJ, Buerkle CA (2010) Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of Lycaeides butterflies. Molecular Ecology. 2010; 19, 2455–2473.
61. Good BH, Walczak AM, Neher RA, Desai MM (2014) Genetic Diversity in the Interference Selection Limit. PLoS Genet 10(3): e1004222.
62. Goodstein DM, Shu S, Howson R et al. (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Research. 40(Database issue):D1178-D1186. doi:10.1093/nar/gkr944.
63. Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics. 17,333–351. doi:10.1038/nrg.2016.49
64. Gore MA, Chia JM, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL et al. (2009) A first-generation haplotype map of maize. Science. 326, 1115–1117 doi: 10.1126/science.1177837 PMID: 19965431.

65. Grant D, Nelson RT, Cannon SB, Shoemaker RC (2010) SoyBase, the USDA-ARS soybean genetics and genomics database. Nucl. Acids Res. D843-D846. doi: 10.1093/nar/gkp798 .
66. Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A et al. (2015) Large-scale whole-genome sequencing of the Icelandic population. Nature Genetics. 47, 435–444. doi:10.1038/ng.3247.
67. Ha NT, Freytag S, Bickeboeller H (2014) Coverage and efficiency in current SNP chips. European Journal of Human Genetics. 22, 1124–1130; doi: 10.1038/ejhg.2013.304.
68. Hall BG (1988) Adaptive Evolution That Requires Multiple Spontaneous Mutations. I. Mutations Involving an Insertion Sequence. Genetics. 120(4):887-897.
69. Hall N (2007) Advanced sequencing technologies and their wider impact in microbiology. J. Exp. Biol. 209 (Pt 9): 1518–1525. doi:10.1242/jeb.001370.
70. Hao K, Chudin E, McElwee J, Schadt E (2009) Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies. BMC Genet. 10: 27. doi: 10.1186/1471-2156-10-27.
71. He S, Zhao Y, Mette MF, Bothe R, Ebmeyer E, Sharbel TF, Reif JC, Jiang Y (2015) Prospects and limits of marker imputation in quantitative genetic studies in European elite wheat (Triticum aestivum L.). BMC Genomics 16:168 doi:10.1186/s12864-015-1366-y.
72. He, Q. et al. (2015) Fine mapping of the genetic locus L1 conferring black pods using a chromosome segment substitution line population of soybean. Plant Breeding 134: 437–445.
73. Hedrick, P (2011) Genetics of populations. 4th ed. Boston: Jones & Bartlett Learning Press. ISBN 978-0-7637-5737-3.
74. Herten K, Hestand MS, Vermeesch JR, Van Houdt JKJ (2015) GBSX: a toolkit for experimental design and demultiplexing genotyping by sequencing experiments. BMC Bioinformatics. DOI 10.1186/s12859-015-0514-3.
75. Hohenlohe PA, Phillips PC, Cresko WA (2010a) Using population genomics to detect selection in natural populations: key concepts and methodological considerations. International Journal of Plant Sciences. 171, 1059–1071.
76. Hormozdiari F, Hajirasouliha IAM, Eichler EE, Sahinalp SC (2011) Simultaneous structural variation discovery in multiple paired-end sequenced genomes. Proc. RECOMB.
77. Hottes AK et al. (2013) Bacterial Adaptation through Loss of Function. PLoS Genet 9(7): e1003617.
78. Howie B, Marchini J, Stephens M (2011) Genotype Imputation with Thousands of Genomes. G3: Genes, Genomes, Genetics. 1(6) 457–470.
79. Howie BN, Donnelly P, and Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genetics 5(6): e1000529.
80. Huang BM, Raghavan C, Mauleon R, Broman KW, Leung H (2014) Efficient Imputation of Missing Markers in Low-Coverage Genotyping-by-Sequencing Data from Multiparental Crosses. Genetics Society of America. 2014; doi: 10.1534/genetics.113.158014.

81. Huang et al. (2012) A map of rice genome variation reveals the origin of cultivated rice. Nature. 490: 497-501. doi:10.1038/nature11532.
82. Huang X, Wei X, Sang T, Zhao Q, Feng Q et al. (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nature Genetics 42, 961–967 doi:10.1038/ng.695.
83. Hufford MB et al. (2012) Comparative population genomics of maize domestication and improvement. Nat. Gen. 44, 808–811.
84. Hwang S, Kim E, Lee I, Marcotte EM (2015) Systematic comparison of variant calling pipelines using gold standard personal exome variants. Scientific Reports. 17875. 693 doi:10.1038/srep17875.
85. Jarquín D, Kocak K, Posadas L, Hyma K, Jedlicka J, Graef G, Lorenz A (2014) Genotyping by sequencing for genomic prediction in a soybean breeding population. BMC Genomics. 15(1), 740. http://doi.org/10.1186/1471-2164-15-740.
86. Karki R, Pandya D, Elston RC, Ferlini C (2015) Defining "mutation" and "polymorphism" in the era of personal genomics. BMC Medical Genomics, 8, 37. http://doi.org/10.1186/s12920-015-0115-z.
87. Kilpinen H, Barrett JC (2013) How next-generation sequencing is transforming complex disease genetics. Trends Genet. 29: 23–30.
88. Kim MY et al. (2010) Whole-genome sequencing and intensive analysis of the undomesticated soybean (Glycine soja Sieb. and Zucc.) genome. Proc. Natl. Acad. Sci. 107, 22032–22037.
89. Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, et al. (2007) Recombination and linkage disequilibrium in Arabidopsis thaliana. Nat. Genet. 39, 1151–1155. PMID: 17676040.
90. Kong A, Masson G, Frigge ML, Gylfason A, Zusmanovich P, et al. (2008) Detection of sharing by descent, long-range phasing and haplotype imputation. Nat Genet 40: 1068–1075. pmid:19165921.
91. Kumar S, Banks TW, Cloutier S. (2012) SNP discovery through next-generation sequencing and its applications. International Journal of Plant Genomics. doi: 10.1155/2012/831460.
92. Lachance J, Tishkoff SA (2013) SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology, 35(9), 780–786. http://doi.org/10.1002/bies.201300014
93. Lam HM, Xu X, Liu X, Chen WB, Yang GH, Wong FL, Li MW, He WM, Qin N, Wang (2010) B: Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet. 42: 1053-1059. 10.1038/ng.715.
94. Lam HYK, Mu XJ, Stutz AM, Tanzer A, Cayting PD, Snyder M et al. (2010) Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. Nature Biotechnology. 28 (1): 47–55. doi:10.1038/nbt.1600.
95. Larson D, Chiang C, Badve A, Eldred J, Morton D (2016) svtools: svtools v0.3.0 [Data set]. Zenodo. http://doi.org/10.5281/zenodo.167453.
96. Larson WA, Seeb LW, Everett MV, Waples RK, Templin WD, Seeb JE (2014) Genotyping by sequencing resolves shallow population structure to inform conservation of Chinook

salmon (Oncorhynchus tshawytscha). Evolutionary Applications.7(3):355-369. doi:10.1111/eva.12128.
97. Layer RM, Chiang C, Quinlan AR, Hall IM (2014) LUMPY: A probabilistic framework for structural variant discovery. Genome Biology. 15:R84.
98. Lek M et al. (2016) Analysis of protein-coding genetic variation in 60,706 humans. Nature. 536, 285–291.
99. Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 1;27(21):2987-93.
100. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 25, 1754–1760.
101. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer J, Marth G, Abecasis G, Durbin R (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25, 2078-2079
102. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K (2009) SNP detection for massively parallel whole-genome resequencing. Genome Res. 19 (6): 1124-1132. 10.1101/gr.088013.108.
103. Li Y, Willer C, Sanna S (2009) Genotype Imputation. Annu. Rev. Genomics Hum. Genet. 10:387–406. doi: 10.1146/annurev.genom.9.081307.164242.
104. Li YH et al. (2013) Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. BMC Genomics 14, 579.
105. Li YH, Zhou G, Ma J, Jiang W, Jin LG, Zhang Z, et al. (2014) De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat Biotechnol. 32:1045–52.
106. Libault M et al. (2010) An integrated transcriptome atlas of the crop model Glycine max, and its use in comparative analyses in plants. Plant J. 1;63(1):86-99.
107. Lin T, Zhu G, Zhang J, Xu X, Yu Q et al. (2014) Genomic analyses provide insights into the history of tomato breeding. Nature Genetics 46, 1220–1226 doi:10.1038/ng.3117.
108. Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ et al. (2012) GAPIT: genome association and prediction integrated tool. Bioinformatics. 28 (18): 2397–2399. doi: 10.1093/bioinformatics/bts444.
109. Lower KM, Hughes JR, De Gobbi M et al. (2009) Adventitious changes in long-range gene expression caused by polymorphic structural variation and promoter competition. Proceedings of the National Academy of Sciences of the United States of America. 106(51):21771-21776. doi:10.1073/pnas.0909331106.
110. Lu F, Lipka AE, Glaubitz J, Elshire R, Cherney JH, et al. (2013) Switchgrass Genomic Diversity, Ploidy, and Evolution: Novel Insights from a Network-Based SNP Discovery Protocol. PLoS Genet 9(1): e1003215. doi:10.1371/journal.pgen.1003215.
111. Lynch M (2009) Estimation of allele frequencies from high coverage genome sequencing projects. Genetics.182:295–301.
112. MacArthur DG et al. (2012) A systematic survey of loss-of-function variants in human protein-coding genes. Science 335, 823–828.
113. Maldonado dos Santos JV et al. (2016) Evaluation of genetic variation among Brazilian soybean cultivars through genome resequencing. BMC Genomics 17:110.

114. Maron LG, Guimarães CT, Kirst M, Albert PS, Birchler JA et al. (2013) Aluminum tolerance in maize is associated with higher MATE1 gene copy number. Proceedings of the National Academy of Sciences of the United States of America. 5241–5246, doi: 10.1073/pnas.1220766110
115. Marroni F, Pinosio S, Morgante M (2014) Structural variation and genome complexity: is dispensable really dispensable? Current Opinion in Plant Biology. Volume 18, Pages 31–36
116. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. DOI: http://dx.doi.org/10.14806/ej.17.1.200.
117. Mascher M, Wu S, Amand PS, Stein N, Poland J (2013) Application of Genotyping-by-Sequencing on Semiconductor Sequencing Platforms: A Comparison of Genetic and Reference-Based Marker Ordering in Barley. PLoS ONE 8(10): e76925. doi: 10.1371/journal.pone.0076925.
118. McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2013) Applications of next-generation sequencing to phylogeography and phylogenetics. Molecular Phylogenetics and Evolution, 66, 526–538.
119. Messer PW, Petrov DA (2013) Frequent adaptation and the mcdonald-kreitman test. Proc Natl Acad Sci 110: 8615–8620.
120. Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet. 31–46. doi:10.1038/nrg2626.
121. Meyer RS, Purugganan MD (2013) Evolution of crop species: genetics of domestication and diversification. Nat. Rev. Gen. 10,1038.
122. Mian N (2006) Soy Applications in Food. Boca Raton, FL: CRC Press. ISBN 0-8493-2981-7.
123. Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res. 17, 240–248.
124. Mills RE, Walter K, Stewart C et al. (2011) Mapping copy number variation by population scale genome sequencing. Nature. 470(7332):59-65. doi:10.1038/nature09708.
125. Muir P, Li S, Lou S, Wang D, Spakowicz DJ et al. (2016) The real cost of sequencing: scaling computation to keep pace with data generation. Genome Biology. 17:53. doi: 10.1186/s13059-016-0917-0.
126. Neher RA (2013) Genetic draft, selective interference, and population genetics of rapid adaptation. Annual Review of Ecology, Evolution, and Systematics 44: 195–215
127. Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. Nature Rev Genet.12:443–51. doi:10.1038/nrg2986.
128. Nishida H, Yoshida T, Kawakami K, Fujita M, Long B, Akashi Y, Laurie DA, Kato K (2013) Structural variation in the 5′ upstream region of photoperiod-insensitive alleles Ppd-A1a and Ppd-B1a identified in hexaploid wheat (Triticum aestivum L.), and their effect on heading time. Mol. Breed. 31: 27–37.
129. Olsen KM, Wendel JF (2013) A bountiful harvest: genomic insights into crop domestication phenotypes. Annu. Rev. Plant Biol. 64:47–70.

130. Palaisa K, Morgante M, Tingey S, Rafalski A (2004) Long-range patterns of diversity and linkage disequilibrium surrounding the maize Y1 gene are indicative of an asymmetric selective sweep. Proc. Natl Acad. Sci. USA 101, 9885–9890.
131. Panchy N, Lehti-Shiu M, Shiu S-H (2016) Evolution of Gene Duplication in Plants. Plant Physiology.171(4):2294-2316. doi:10.1104/pp.16.00523.
132. Parchman TL, Gompert Z, Mudge J, Schilkey FD, Benkman CW, Buerkle CA (2012) Genome-wide association genetics of an adaptive trait in lodgepole pine. Molecular Ecology, 21, 2991–3005. doi:10.1111/j.1365-294X.2012.05513.x.
133. Pei YF, Li J, Zhang L, Papasian CJ, Deng HW (2008) Analyses and comparison of accuracy of different genotype imputation methods. PLoS ONE. 3: e3551. doi: 10.1371/journal.pone.0003551.
134. Pérez-de-Castro AM, Vilanova S, Cañizares J, Pascual L, Blanca JM, Díez MJ, Picó B (2012) Application of Genomic Tools in Plant Breeding. Current Genomics, 13(3), 179–195. http://doi.org/10.2174/138920212800543084.
135. Peter GV, Uitdewilligen JGAML, Voorrips E, Visser RGF, van Eck HJ (2015) Development and analysis of a 20K SNP array for potato (Solanum tuberosum): an insight into the breeding history. Theoretical and Applied Genetics. DOI: 10.1007/s00122-015-2593-y.
136. Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest radseq: an inexpensive method for de novo SNP discovery and genotyping in model and nonmodel species. PLoS ONE, 7, e37135.
137. Pinkel D, Albertson DG (2005) Comparative genomic hybridization. Annu Rev Genomics Hum Genet. 6:331-354.
138. Pirooznia M, Kramer M, Parla J, Goes FS, Potash JB, McCombie WR, Zandi PP (2014) Validation and assessment of variant calling pipelines for next-generation sequencing. Human Genomics, 8(1), 14.
139. Platypus-Sapelo: https://wiki.gacrc.uga.edu/wiki/Platypus-Sapelo
140. Poland J, Endelman J, Dawson J, Rutkoski J, Wu S et al. (2012) Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Gen. 5: 103–113.
141. Poland JA, Brown PJ, Sorrells ME, Jannink J-L (2012) Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. PLoS ONE 7(2): e32253.
142. Poland JA, Rife TW (2012).Genotyping-by-sequencing for plant breeding and genetics. Plant Genome 5 92–102 10.3835/plantgenome2012.05.0005
143. Porto-Neto LR, Kijas JW, Reverter A (2014) The extent of linkage disequilibrium in beef cattle breeds using high-density SNP genotypes. Genet Sel Evol. Mar 24;46:22. doi: 10.1186/1297-9686-46-22.
144. Prashar A, Hornyik C, Young V, McLean K, Sharma SK, Dale MF, Bryan GJ (2014) Construction of a dense SNP map of a highly heterozygous diploid potato population and QTL analysis of tuber shape and eye depth. Theor Appl Genet. 127(10):2159-71. doi: 10.1007/s00122-014-2369-9.
145. Pritchard JK, Pickrell JK, Coop G (2010) The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation. Current biology. CB;20(4):R208-R215. doi:10.1016/j.cub.2009.11.055.

146. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D et al. (2007) PLINK: a tool set for whole genome association and population-based linkage analyses. Am. J. Hum. Genet. 81(3):559–75.
147. Purugganan MD, Fuller DQ (2009) The nature of selection during plant domestication. Nature. 457, 843–848.
148. Quinlan AR, Clark RA, Sokolova S et al. (2010) Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. Genome Research. 20(5):623-635. doi:10.1101/gr.102970.109.
149. Raj A, Stephens M, and Pritchard JK (2014) fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. Genetics. 197:573-589.
150. Rasheed A, Hao Y, Xia X, Khan A, Xu Y, Varshney RK, He Z (2017) Crop Breeding Chips and Genotyping Platforms: Progress, Challenges, and Perspectives. Mol Plant. 7;10(8):1047-1064. doi: 10.1016/j.molp.2017.06.008.
151. Redon R, Carter NP (2009). Comparative Genomic Hybridization: microarray design and data interpretation. Methods in Molecular Biology (Clifton, N.J.), 529, 37–49. http://doi.org/10.1007/978-1-59745-538-1_3.
152. Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, et al. (2014) Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nature Genetics. doi:10.1038/ng.3036.
153. Rosato C, Etter P, Kamps-Hughes N, Johnson E (2012) Genotyping on High Throughput Sequencers: Preparation and Analysis of Reduced Representation Genomic Libraries. Journal of Biomolecular Techniques: JBT, 23(Suppl), S20.
154. 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 of the United States of America. 104, 8641–8648.
155. Ross-Ibarra J, Morrell PL, Gaut, BS (2007) Plant domestication, a unique opportunity to identify the genetic basis of adaptation. Proc. Natl Acad. Sci. USA 104, 8641–8648.
156. Rothberg JM., Hinz W, Rearick TM, Schultz J, Mileski W et al. (2011) An integrated semiconductor device enabling non-optical genome sequencing. Nature 475, 348–352. doi:10.1038/nature10242.
157. Rutkoski JE, Poland J, Jannink JL, and Sorrells ME (2013) Imputation of unordered markers and the impact on genomic selection accuracy. Genes Genomes Genetics. 3: 427–439. doi: 10.1534/g3.112.005363.
158. Sabre-barcode-demultiplexing: https://github.com/najoshi/sabre
159. Santana MH, Utsunomiya YT, Neves HH, Gomes RC, Garcia JF, Fukumasu H et al. (2014) Genome-wide association analysis of feed intake and residual feed intake in Nellore cattle. BMC Genet.15:21.
160. Saxena RK, Edwards D, Varshney RK (2014). Structural variations in plant genomes. Briefings in Functional Genomics, 13(4), 296–307. http://doi.org/10.1093/bfgp/elu016.
161. Scheet P, and Stephens M (2006) A fast and flexible statistical model for large scale population genotype data: application to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. 78, pp. 629–44.

162. Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., et al. (2010) Genome sequence of the palaeopolyploid soybean. Nature. 463(7278):178–183. doi: 10.1038/nature08670.
163. Sedivy EJ et al. (2017) Soybean domestication: the origin, genetic architecture and molecular bases. New Phytologist.
164. Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren U, Long Q, Nordborg M (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat. Genet. doi:10.1038/ng.2314.
165. Sehgal D, Vikram P, Sansaloni CP, Ortiz C, Pierre CS, Payne T et al. (2015) Exploring and Mobilizing the Gene Bank Biodiversity for Wheat Improvement. PLoS ONE 10(7): e0132112.
166. Severin AJ et al. (2010) RNA-Seq Atlas of Glycine max: A guide to the soybean transcriptome. BMC Plant Biology 10:160.
167. Shifman S, Kuypers J, Kokoris M, Yakir B, and Darvasi A (2003) Linkage disequilibrium patterns of the human genome across populations. Hum. Mol. Genet. 12 (7): 771-776. doi: 10.1093/hmg/ddg088.
168. Slatkin M (2008) Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. Nature Reviews. Genetics, 9(6), 477–485. http://doi.org/10.1038/nrg2361.
169. Sonah H, Bastien M, Iquira E, Tardivel A, Légaré G, Boyle B et al. (2013) An Improved Genotyping by Sequencing (GBS) Approach Offering Increased Versatility and Efficiency of SNP Discovery and Genotyping. PLoS ONE 8(1): e54603. https://doi.org/10.1371/journal.pone.0054603.
170. Sonah H, Bastien M, Iquira E, Tardivel A, Legare G, et al. (2013) An Improved Genotyping by Sequencing (GBS) Approach Offering Increased Versatility and Efficiency of SNP Discovery and Genotyping. PLoS ONE 8(1): e54603. doi:10.1371/journal.pone.0054603 PMID: 23372741.
171. Sonah H, O'Donoughue L, Cober E, Rajcan I, Belzile F (2014) Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. Plant Biotechnol. J. doi: 10.1111/pbi.12249.
172. Song Q et al. (2015), Fingerprinting soybean germplasm and its utility in genomic research.; In press.
173. Song Q, Hyten DL, Jia G et al. (2015) Fingerprinting Soybean Germplasm and Its Utility in Genomic Research. G3: Genes|Genomes|Genetics. 5(10):1999-2006. doi:10.1534/g3.115.019000.
174. Song Q, Hyten DL, Jia G, Quigley CV, Fickus EW, Nelson RL et al. (2013) Development and Evaluation of SoySNP50K, a High-Density Genotyping Array for Soybean. PLoS ONE 8(1): e54985. https://doi.org/10.1371/journal.pone.0054985.
175. Springer NM, Ying K, Fu Y, Ji T, Yeh C-T, Jia Y et al. (2009) Maize Inbreds Exhibit High Levels of Copy Number Variation (CNV) and Presence/Absence Variation (PAV) in Genome Content. PLoS Genet 5(11): e1000734. https://doi.org/10.1371/journal.pgen.1000734
176. Stankiewicz P, Lupski JR (2010) Structural variation in the human genome and its role in disease. Annu Rev Med. 61:437-55. doi: 10.1146/annurev-med-100708-204735.

177. Sudmant PH, Rausch T, Gardner EJ et al. (2015) An integrated map of structural variation in 2,504 human genomes. Nature. 526(7571):75-81. doi:10.1038/nature15394.
178. Swaminathan MS (2009). "Obituary: Norman E. Borlaug (1914–2009) Plant scientist who transformed global food production". Nature. 461 (7266): 894–894.
179. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Molecular Biology and Evolution. 30(12):2725-2729. doi:10.1093/molbev/mst197.
180. Tattini L, D'Aurizio R, Magi A (2015) Detection of Genomic Structural Variants from Next-Generation Sequencing Data. Frontiers in Bioengineering and Biotechnology. 3:92. doi:10.3389/fbioe.2015.00092.
181. Tennessen JA, O'Connor TD, Bamshad MJ, Akey JM (2011). The promise and limitations of population exomics for human evolution studies. Genome Biology, 12(9), 127. http://doi.org/10.1186/gb-2011-12-9-127.
182. The 1001 Genomes Consortium. Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., … Zhou, X. (2016). 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell, 166(2), 481–491. http://doi.org/10.1016/j.cell.2016.05.063
183. The 3,000 rice genomes project. (2014) The 3,000 rice genomes project. GigaScience. 3:7 https://doi.org/10.1186/2047-217X-3-7
184. The International Barley Genome Sequencing Consortium (2012) A physical, genetic and functional sequence assembly of the barley genome. Nature. 491, 711 – 716. doi:10.1038/nature11543
185. Torkamaneh D, Belzile F (2015) Scanning and Filling: Ultra-Dense SNP Genotyping Combining Genotyping-By-Sequencing, SNP Array and Whole-Genome Resequencing Data. PLoS ONE 10(7): e0131533. doi:10.1371/journal.pone.0131533.
186. Torkamaneh D, Laroche J, Bastien M, Abed A, Belzile F (2017a) Fast-GBS: a new pipeline for the efficient and highly accurate calling of SNPs from genotyping-by sequencing data. BMC Bioinformatics. doi: 10.1186/s12859-016-1431-9.
187. Torkamaneh D, Laroche J, Tardivel A, O'Donoughue L, Cober E, Rajcan I, Belzile F. (2017) Comprehensive Description of Genome-Wide Nucleotide and Structural Variation in Short-Season Soybean. Plant Biotechnology Journal.
188. Truong HT, Ramos AM, Yalcin F, de Ruiter M, van der Poel HJA et al. (2012) Sequence-Based Genotyping for Marker Discovery and Co-Dominant Scoring in Germplasm and Populations. PLoS ONE 7(5): e37565.
189. Tsiantis M (2011) A transposon in tb1 drove maize domestication. Nature Genetics 43, 1048–1050
190. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM et al. (2005) Fine-scale structural variation of the human genome. Nature Genetics. 37 (7): 727–32. doi:10.1038/ng1562. PMID 15895083.
191. Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JAM (2007) Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Research. 35(Web Server issue):W71-W74. doi:10.1093/nar/gkm306.
192. Valliyodan B et al. (2016) Landscape of genomic diversity and trait discovery in soybean. Sci Rep. 6: 23598.

193. Van Orsouw NJ, Hogers RCJ, Janssen A, Yalcin F, Snoeijers S, Verstege E et al. (2007) Complexity reduction of polymorphic sequences (CRoPS): a novel approach for large-scale polymorphism discovery in complex genomes. PLoS ONE. 2, e1172.
194. Varela MA, Amos W (2010) "Heterogeneous distribution of SNPs in the human genome: Microsatellites as predictors of nucleotide diversity and divergence". Genomics. 95 (3): 151–159. PMID 2002. 6267. doi:10.1016/j.ygeno.2009.12.003.
195. Varshney RK, Terauchi R, McCouch SR (2014) Harvesting the Promising Fruits of Genomics: Applying Genome Sequencing Technologies to Crop Breeding. PLoS Biol 12(6): e1001883. https://doi.org/10.1371/journal.pbio.1001883.
196. Wang M, Yan J, Zhao J, Song W, Zhang X, Xiao Y, and Zheng Y (2012) Genome-wide association study (GWAS) of resistance to head smut in maize. Plant Sci. 196, 125–31.
197. Wang S, Meyer E, McKay JK, Matz MV (2012) 2b-RAD: a simple and flexible method for genome-wide genotyping. Nature Methods. 9, 808–810.
198. Wang Y, Xiong G, Hu J, Jiang L, Yu H, Xu J et al. (2015) Copy number variation at the GL7 locus contributes to grain size diversity in rice. Nat. Genet. 47, 944–948. doi: 10.1038/ng.3346.
199. Xu et al. (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nature Biotechnology. 30,105–111. doi:10.1038/nbt.2050.
200. Yang Z, Li Z, and Bickel DR (2013) Empirical Bayes estimation of posterior probabilities of enrichment: a comparative study of five estimators of the local false discovery rate. BMC Bioinformatics, 14, 87.
201. Ye K, Hall G, Ning Z (2016) Structural Variation Detection from Next Generation Sequencing. Next Generat Sequenc & Applic. S1:007. doi:10.4172/2469-9853.S1-007.
202. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics (Oxford, England) 25, 2865–2871.
203. Zeder MA (2012) The domestication of animals. Journal of Anthropological Research. 68: 161–190.
204. Zeder MA (2015) Core questions in domestication Research. Proceedings of the National Academy of Sciences of the United States of America. 112: 3191–8.
205. Zhang J, Chiodini R, Badr A, Zhang G (2001) The impact of next-generation sequencing on genomics. Journal of Genetics and Genomics. 38(3):95-109. doi:10.1016/j.jgg.2011.02.003.
206. Zhao S et al. (2015) Impacts of nucleotide fixation during soybean domestication and improvement. BMC Plant Biology. 15, 81.
207. Zheng J, Li Y, Abecasis GR, Scheet P (2011) A comparison of approaches to account for uncertainty in analysis of imputed genotypes. Genet Epidemiol. 35:102-110.
208. Zhou D, Zhou X, Ling Y, Zhang Z, Su Z (2010) agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Research. doi:10.1093/nar/gkq310.

209. Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J et al. (2015) Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nature Biotechnology. 33, 408–414. doi:10.1038/nbt.3096.
210. Zhu Q, Zheng X, Luo J, Gaut BS, Ge S. (2007) Multi locus analysis of nucleotide variation of Oryza sativa and its wild relatives: severe bottleneck during domestication of rice. Mol. Biol. Evol. 2007; 24, 875–888 PMID: 17218640.
211. Zohary D (2004) Unconscious selection and the evolution of domesticated plants. Econ. Bot. 58, 5–10.