



# Soybean (*Glycine max*) Haplotype Map (GmHapMap): a universal resource for soybean translational and functional genomics

Davoud Torkamaneh<sup>1,2,3</sup> , Jérôme Laroche<sup>2</sup>, Babu Valliyodan<sup>4</sup>, Louise O'Donoghue<sup>5</sup>, Elroy Cober<sup>6</sup>, Istvan Rajcan<sup>3</sup>, Ricardo Vilela Abdelnoor<sup>7,8</sup>, Avinash Sreedasyam<sup>9</sup>, Jeremy Schmutz<sup>9,10</sup>, Henry T. Nguyen<sup>4</sup>  and François Belzile<sup>1,2,\*</sup>

<sup>1</sup>Département de Phytologie, Université Laval, Québec City, QC, Canada

<sup>2</sup>Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Québec City, QC, Canada

<sup>3</sup>Department of Plant Agriculture, University of Guelph, Guelph, ON, Canada

<sup>4</sup>National Center for Soybean Biotechnology and Division of Plant Sciences, University of Missouri, Columbia, MO, USA

<sup>5</sup>CÉROM, Centre de recherche Sur Les Grains Inc., Saint-Mathieu de Beloeil, QC, Canada

<sup>6</sup>Agriculture and Agri-Food Canada, Ottawa, ON, Canada

<sup>7</sup>Brazilian Corporation of Agricultural Research (Embrapa Soja), Warta County, PR, Brazil

<sup>8</sup>Londrina State University (UEL), Londrina, PR, Brazil

<sup>9</sup>Institute for Biotechnology, HudsonAlpha, Huntsville, AL, USA

<sup>10</sup>Department of Energy, Joint Genome Institute, Walnut Creek, CA, USA

Received 2 October 2019;

revised 24 July 2020;

accepted 7 August 2020.

\*Correspondence (Tel +1-418-656-2131  
ext. 405763; fax +1-418-948-5487;  
email francois.belzile@fsaa.ulaval.ca)

**Keywords:** soybean, whole-genome sequencing, genetic variants, haplotype map, imputation, haplotype, loss-of-function mutation.

## Summary

Here, we describe a worldwide haplotype map for soybean (GmHapMap) constructed using whole-genome sequence data for 1007 *Glycine max* accessions and yielding 14.9 million variants as well as 4.3 M tag single-nucleotide polymorphisms (SNPs). When sampling random subsets of these accessions, the number of variants and tag SNPs plateaued beyond approximately 800 and 600 accessions, respectively. This suggests extensive coverage of diversity within the cultivated soybean. GmHapMap variants were imputed onto 21 618 previously genotyped accessions with up to 96% success for common alleles. A local association analysis was performed with the imputed data using markers located in a 1-Mb region known to contribute to seed oil content and enabled us to identify a candidate causal SNP residing in the *NPC1* gene. We determined gene-centric haplotypes (407 867 GCHs) for the 55 589 genes and showed that such haplotypes can help to identify alleles that differ in the resulting phenotype. Finally, we predicted 18 031 putative loss-of-function (LOF) mutations in 10 662 genes and illustrated how such a resource can be used to explore gene function. The GmHapMap provides a unique worldwide resource for applied soybean genomics and breeding.

## Introduction

Soybean (*Glycine max* [L.] Merr.) is a unique crop with substantial economic value. It is the largest plant source of both animal feed protein and edible oil. It also plays a key role in sustainable agriculture as it fixes atmospheric nitrogen with the help of microorganisms (Hungria and Mendes, 2015). Diverse evolutionary processes and forces (including cycles of polyploidization and subsequent diploidization), along with domestication and modern breeding, have shaped the soybean genome (Schmutz *et al.*, 2010). The detection of the molecular footprints of these processes is essential for understanding how genetic diversity is generated and maintained and for identifying allelic variants responsible for phenotypic variation (Torkamaneh *et al.*, 2018).

The global production of soybean has increased substantially in recent years (Ray *et al.*, 2013), but the rate of annual yield gains (without considering symbiotic nitrogen fixation) has lagged behind that of maize (FAOSTAT Database). In addition, with

increased fluctuations in climatic conditions, next-generation soybean cultivars must not only be higher yielding but also more resilient to multiple abiotic and biotic stresses (Djanaguiraman *et al.*, 2019). However, in the main soybean-growing areas of the world (North and South America), soybean is an introduced crop and, typically, small numbers of *G. max* accessions are thought to have made large contributions to the gene pool of improved cultivars currently grown in these regions (Gizlice *et al.*, 1994; Hyten *et al.*, 2006; Maldonado dos Santos *et al.*, 2016). Continued genetic improvement in soybeans will require a better understanding of the genetic and especially allelic diversity within worldwide resources (Qiu *et al.*, 2013).

Decreased whole-genome sequencing (WGS) costs due to the advent of next-generation sequencing (NGS) technologies has provided an exceptional opportunity to systematically detect genetic variants from the entire genomes of numerous individuals of the same species (Wang *et al.*, 2018a). In soybean, several comprehensive descriptions of nucleotide variants, using WGS,

have been achieved within different populations (e.g. China (Fang *et al.*, 2017; Lam *et al.*, 2010; Zhou *et al.*, 2015), USA (Song *et al.*, 2017; Valliyodan *et al.*, 2016), Korea (Chung *et al.*, 2014), Canada (Torkamaneh *et al.*, 2017a) and Brazil (Maldonado dos Santos *et al.*, 2016)). All these different studies used similar short-read sequencing technologies, but the variant calling was made using different bioinformatics pipelines and different versions of the soybean reference genome. To construct a comprehensive worldwide catalogue of genetic variants in the form of a haplotype map, several criteria should be met including: (i) uniformity of variant calling (single bioinformatics pipeline and single reference genome), (ii) accuracy (low rate of erroneous genotype calls) and (iii) extensiveness (representative of the relevant germplasm).

A comprehensive haplotype map can be used to determine the common pattern of DNA sequence variation in the genome of a species of interest by characterizing sequence variants, their frequencies and the correlation between sequence variants (The International HapMap Consortium, 2003). A high degree of correlation known as linkage disequilibrium (LD) and coinheritance of sequence variants create specific genomic blocks called haplotypes (Pääbo, 2003). Therefore, a subset of representative sequence variants (known as tag SNPs) is enough to identify the haplotypes. These tag SNPs can be used to guide the development of SNP arrays, perform genome-wide association studies (GWAS) and determine the underlying alleles (Nicolas *et al.*, 2006). WGS-derived sequence variants can also be used to build what is called a Practical Haplotype Graph (PHG) (Jensen *et al.*, 2020). A PHG captures key haplotypes in a specific collection of accessions and makes it possible to predict genotypes from low-coverage sequence data (mapped to the PHG) and to impute missing genotypes.

Another benefit of such extensive catalogues of existing genetic variation is to provide tools to better explore gene function. In addition to the development of model-based approaches (Kono *et al.*, 2018), prediction tools have made it possible to identify and classify variants that are likely to alter gene function and, in extreme cases, lead to loss of function (LOF) (Cingolani *et al.*, 2012). A comprehensive catalogue of informative and functionally important variants can accelerate the efforts to dissect the genetic basis of physiological and agronomic traits in soybean.

Here, we present a comprehensive haplotype map for *Glycine max* (GmHapMap) assembled from DNA resequencing data for a collection of 1,007 worldwide soybean accessions. We explore the use of this GmHapMap for (i) imputation of untyped variants to create high-density genotype data required for gene-level resolution of genome-wide association studies (GWAS); (ii) construction of gene-centric haplotypes (GCHs) for the entire set of soybean genes; and (iii) identification of close to 11 K genes in which at least one LOF allele has been documented in at least one of the studied accessions. The GmHapMap provides a unique resource for translational and functional genomics for the worldwide soybean community.

## Results

### Genomic variation in GmHapMap

To establish a comprehensive haplotype map for *Glycine max* (GmHapMap), a total of 1007 resequenced soybean accessions were analysed. These included 727 accessions (representative of national and regional core collections) that had been

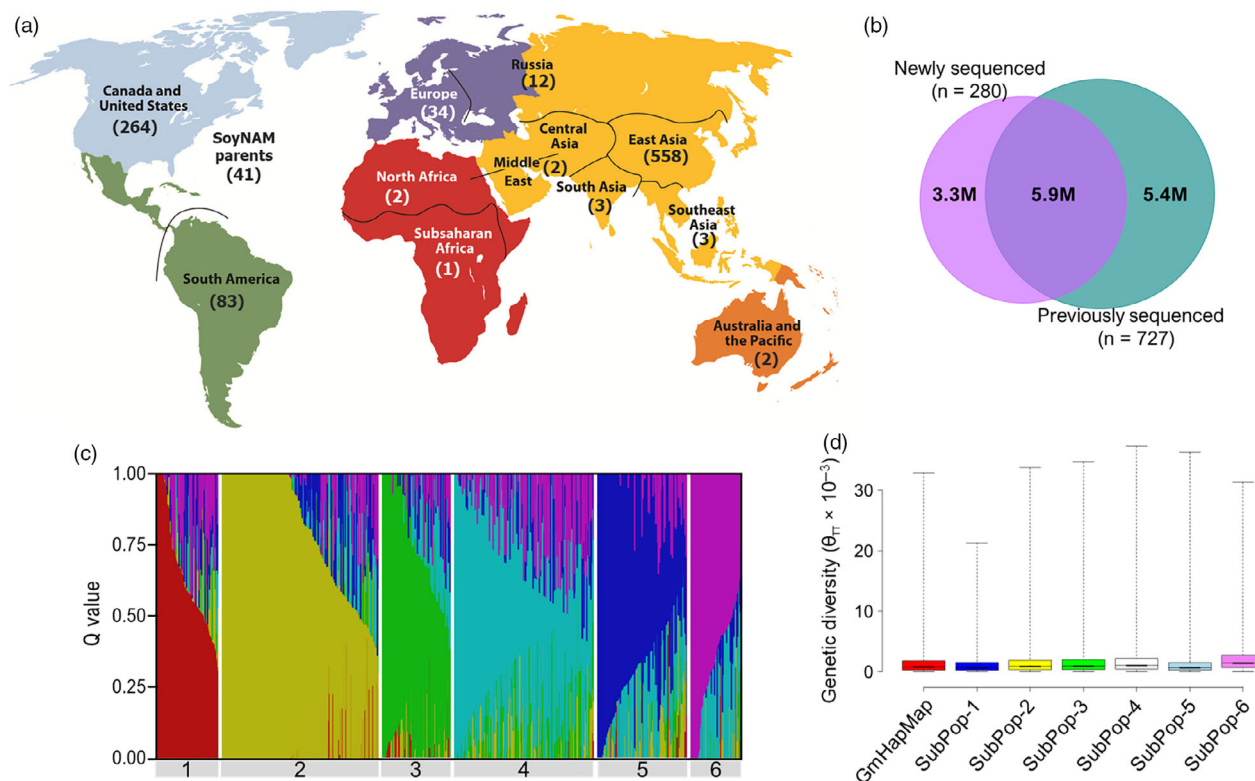
sequenced in previous work and 280 newly sequenced accessions as part of this study (Data S1). Collectively, these are thought to be representative of the worldwide cultivated germplasm (Figure 1a). In total, 165 billion paired-end reads (100–150 bp; total of 19.2 trillion bp) provided an average depth of coverage of 14 × and these were analysed using a single pipeline (Fast-WGS) to ensure uniform variant calling. After applying a set of supervised filters to exclude low-quality variant calls, we identified 14 872 592 nucleotide variants (publicly available at [www.SoyBase.org](http://www.SoyBase.org)) with an average nonsynonymous/synonymous ratio ( $\omega$ ) of 1.49. Of the close to 15 M variants, 3.3 M variants were derived from the sequencing of 280 accessions newly sequenced in this work (Figure 1b). GmHapMap includes 12 197 920 single-nucleotide polymorphisms and 801,373 multiple-nucleotide polymorphisms (SNPs and MNPs) and 1 873 299 small insertions/deletions (indels) (–50 bp to +32 bp), mostly (85.6%) located in non-genic regions. Approximately 45% of variants in GmHapMap dataset were rare (minor allele frequency (MAF) <5%) (Figure S1). Missing data comprised less than 8% of the data, and these were subsequently imputed with high accuracy. In the resulting dataset, 99.7% of the called or imputed genotypes matched the genotype indicated for the same SNP and accession in the SoySNP50K data. The GmHapMap accessions were grouped into six subpopulations (Figure 1c and Figure S2) with some admixture and exhibited a consistent level of genetic diversity (mean of  $\theta_{\pi} = 1.36 \times 10^{-3}$ ) (Figure 1d). This constitutes an extensive and highly accurate set of foundational data for a soybean haplotype map.

### Extensiveness of GmHapMap

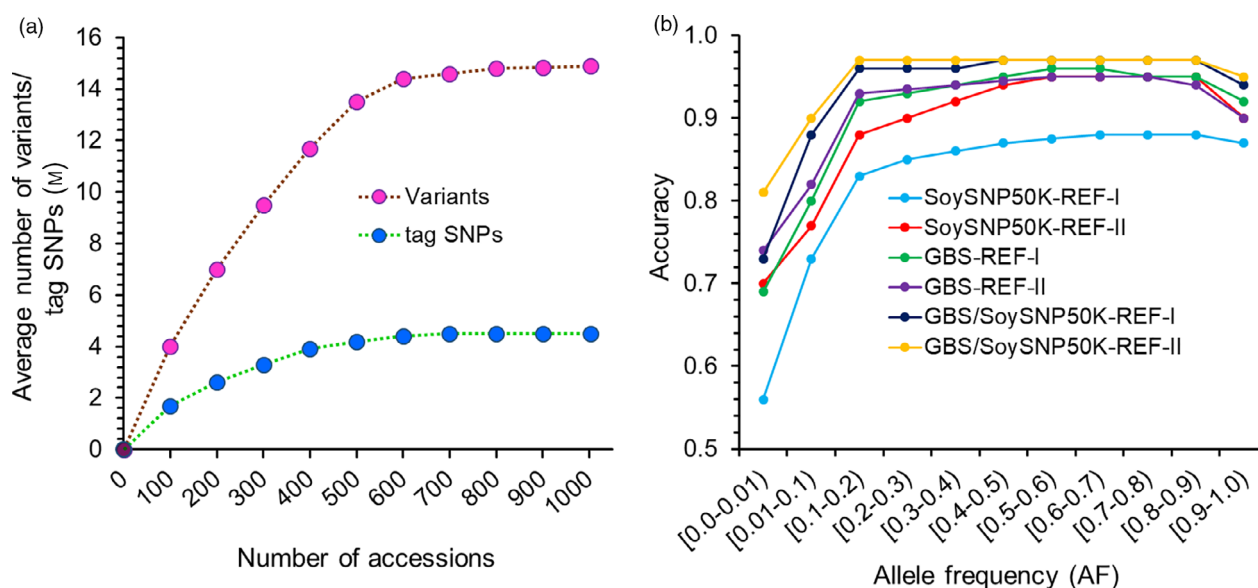
The extensiveness of the GmHapMap was measured based on nucleotide diversity and haplotype diversity. Previously, the SoySNP50K array has been used to genotype the entire USDA soybean germplasm collection (20 087 accessions of *G. max* and *G. soja*; Song *et al.*, 2013). We found that GmHapMap includes nearly all polymorphisms (99.4%; 31 K) with a MAF > 1%, as well as ~89% (15 K) of rare SNPs (MAF < 1%) documented within these *G. max* accessions. Haplotype diversity (pairwise LD using both  $r^2$  and  $D'$ ) was calculated for sequence variants, and the average distance over which LD decayed to 0.2 was ~138 kb (Figure S3). We identified 4.3 million haplotype-based tag SNPs and, to determine whether a good level of saturation of both variants and tag SNPs had been achieved, we randomly selected subsets of samples of increasing size ( $n = 100, 200, \dots$ , and 1007). As illustrated in Figure 2a, the number of variants discovered did not increase significantly beyond ~800 accessions, while the number of tag SNPs reached a plateau much faster (within the first ~600 accessions). Together, these results suggest that the GmHapMap dataset offers an exhaustive characterization of the variants and tag SNPs present in improved soybean cultivated worldwide.

### Large-scale imputation of untyped variants using GmHapMap

The determination of haplotype phase is important because of its applications such as the imputation of untyped variants (imputed sites which are present in the GmHapMap data but absent from the low-density SNP catalogue). We created two reference panels: REF-I comprising all SNPs and REF-II containing 1.9 M haplotype-based tag SNPs that reside in genic regions. Three lower density genotype datasets, SoySNP50K (20 087 accessions



**Figure 1** Description of GmHapMap. (a) Geographical distribution of GmHapMap accessions. (b) Venn diagram representing the degree of overlap among variants called using the two collections of sequenced soybean accessions. (c) Population structure analysis using all SNPs representing six different subpopulations ( $K = 6$ ) in the GmHapMap collection. (d) Distribution of genetic diversity among subpopulations of GmHapMap. [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]



**Figure 2** (a) Average number of variants (pink) and tag SNPs (blue) detected in random subsets of  $N$  accessions (where  $n = 100, 200$  etc.). This average was derived from subsampling 20 times. (b) Imputation accuracy as a function of allele frequency for 6 different scenarios; three different experimentally derived genotype datasets (SoySNP50K, GBS and GBS/SoySNP50K) and two reference panels (REF-I and REF-II). [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

genotyped with 43 K SNPs), genotyping-by-sequencing (GBS; 1531 accessions genotyped with 210 K SNPs) and combined GBS/SoySNP50K (1531 accessions genotyped with 250 K SNPs), were used for untaped-variant imputation with each of the two

reference panels. In all but one case, the accuracy (concordance between imputed and known genotypes, see M&M for details) ranged between 92% and 96% for common variants (allele frequency (AF)  $> 0.2$ ) in each dataset, while decreasing gradually

with allele frequency (Figure 2b). In the case of the SoySNP50K dataset using REF-I, the accuracy of imputed untyped variants was significantly lower (80–85% for common alleles). Given the observed variation in the accuracy of imputation using different reference panels and datasets, we investigated the causes of erroneous inferred calls. Several characteristics were tied to inaccurately imputed SNPs: These were commonly rare variants (low AF), located in short LD blocks (with LD decaying to  $r^2 \leq 0.2$  after only 10 kb, compared to 138 kb genome-wide) or in genomic regions with structural variants. Furthermore, the initial marker density in the experimentally derived dataset had a large impact on imputation accuracy. GBS and SNP array datasets are two highly complementary marker datasets because most (~90%) of the SoySNP50K markers are present in genic regions, while most of the GBS markers (~60%) are present in intergenic regions (Figures S4 & S5). Therefore, combining GBS and SoySNP50K datasets increases the density and uniformity of distribution of SNPs across the genome. The joint use of such commonly available SNP data increased the level of accuracy of imputation of untyped variants (Figure 2b).

### Impact of Imputation Using GmHapMap on Association Analysis

Imputation of untyped variants greatly boosts variant density, allowing fine-mapping studies of loci underlying phenotypic variation and large-scale meta-analysis. To demonstrate the benefits of untyped-variant imputation on marker–trait association analysis, imputation was performed on a 1 Mb-region harbouring a QTL previously identified for seed oil content on chromosome 14. We used the REF-I panel to perform imputation on an initial dataset of 64 K GBS-derived SNPs (genome-wide) among 139 soybean lines that had been characterized for their seed oil content (Sonah *et al.*, 2015). Using this enhanced SNP catalogue and a mixed linear model (MLM) implementation, we performed a regional association mapping (Sosso *et al.*, 2015) and the strongest association ( $P$ -value =  $6.3 \times 10^{-14}$  and  $q$ -value < 0.001) was detected with a SNP residing in the *Glyma.14g001500* gene (Figure S6). This gene codes for a Niemann-Pick C1 (*NPC1*) protein that has been annotated as a lipid transporter (SoyBase). Feldman *et al.* (2015) have documented that an Arabidopsis mutant of this gene (*npc1*) exhibits a 58% higher fatty acid content making this gene a likely candidate contributing to total oil content in soybean. This demonstrates that the increased number of informative SNP loci, obtained through the imputation of untyped variants, can prove highly beneficial in studying the genetic architecture of complex agronomic traits in soybean.

### Gene-centric haplotypes for soybean translational genomics and breeding

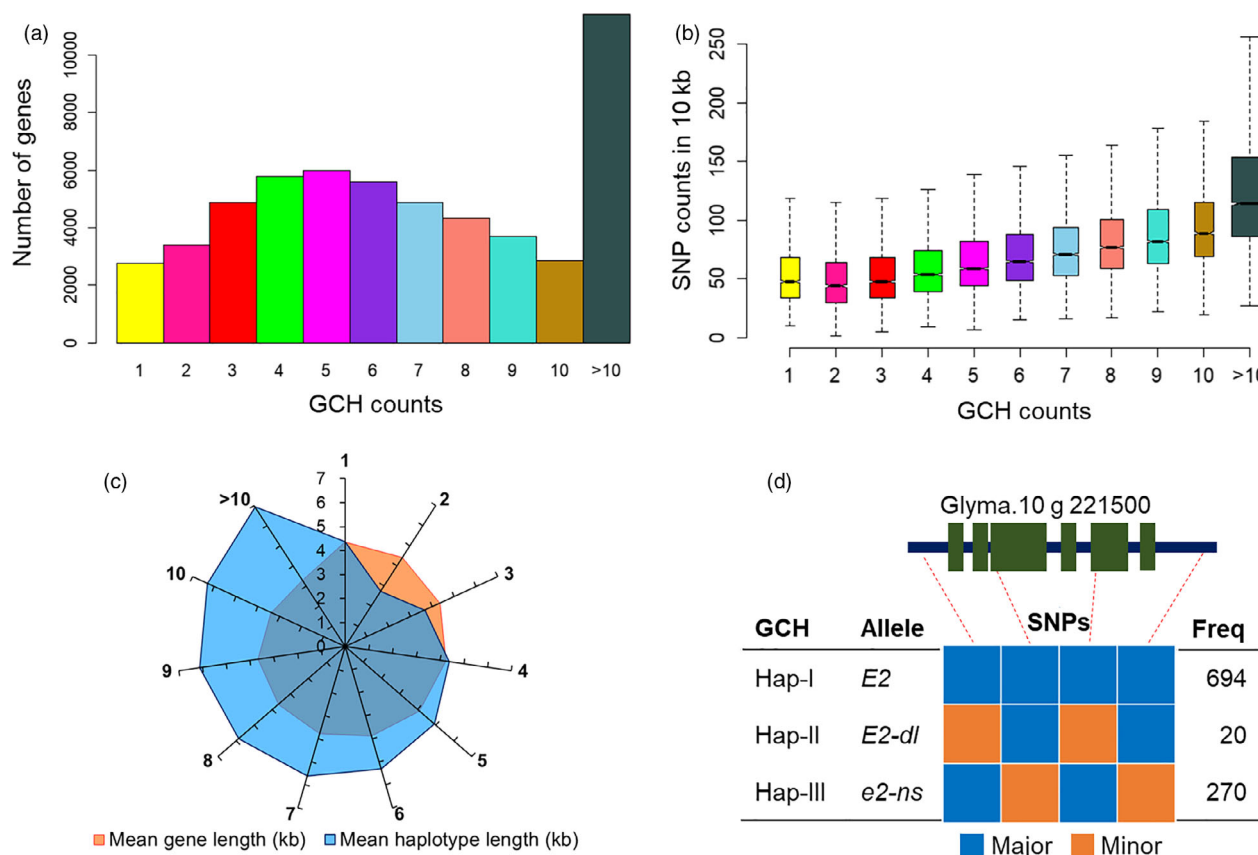
A succession of SNP loci that are in high LD in and around a gene form a haplotype that can provide a useful guide to the rational exploration of genetic diversity. In principle, all accessions sharing the same SNP haplotype in a genic region share the same allele and the characterization of the phenotypes associated with different alleles can be facilitated by selecting representative subsets of accessions that share the same haplotype. We used an LD-based haplotyping method (HaplotypeMiner) to identify gene-centric haplotypes (GCHs) for the complete set of soybean genes. In total, we identified 407,867 GCHs for all soybean genes (55 589; Data S2). The number of GCHs per gene ranged

between 1 and 43, while averaging ~7 (Figure 3a). Overall, 11 407 genes had more than 10 GCHs with 71% (8082 genes) of these harbouring 11–15 GCHs. Such genes were typically located in very short LD blocks with a high degree of nucleotide diversity (mean  $\theta_\pi = 4.5 \times 10^{-3}$ ; Figure 3b). A slight negative correlation was observed between gene length and the number of GCHs. However, we found a positive correlation between GCH counts (number of GCHs) and haplotype size (distance between two most distant SNPs defining a GCH) (Figure 3c). We also found 2766 genes with a single haplotype, of which 353 were located in highly conserved genomic regions where no genetic variation observed in coding regions (Data S3). Such genomic regions present an exceptionally low (115-fold lower) level of nucleotide diversity (mean  $\theta_\pi = 6.15 \times 10^{-6}$ ) compared to genome-wide genic regions (mean  $\theta_\pi = 7.1 \times 10^{-4}$ ). Gene ontology (GO) enrichment analysis for these genes showed no significant enrichment.

An example of GCHs for the *GmGla* (*Glyma.10g221500*) gene (E2 locus controlling maturity) (Tsubokura *et al.*, 2014), an orthologue of the arabidopsis *GIGANTEA* (*Gl*) gene, is presented in Figure 3d. We found three GCHs for *GmGla*, which is consistent with the number of alleles that have been previously reported for this gene (Tsubokura *et al.*, 2014). Additional examples of the correspondence between SNP haplotypes and functionally defined alleles of soybean genes are presented in Figure S7. Knowledge of the GCHs (and possibly alleles) in all soybean genes can greatly facilitate the establishment of a functional link between the various alleles of a gene and the associated phenotype.

### LOF mutations in GmHapMap

Using SnpEff, a subset of variants located inside the coding regions were predicted to have a large functional impact. Of these variants, 18 031 putative loss-of-function (LOF) mutations are predicted to severely impair protein synthesis or function through disruption of splicing, introduction of a premature stop codon, shifts in the coding frame and alterations to the start/stop codons (MacArthur *et al.*, 2012), and these were identified in a total of 10 662 genes (19.3% of all soybean genes; Table 1). These mutations are the result of 5987 SNPs (33.2%), 279 MNPs (1.5%) and 11 765 indels (65.3%). Frameshift-inducing variants (10 754) were the predominant category, representing 59.6% of LOF mutations and affecting 6718 genes. InDels (ranging from –50 bp to +32 bp) were, understandably, over-represented (4-fold) in the LOF category due to their high probability of resulting in a LOF allele. Overall, most of the LOF mutations were present at low frequency, with 78% having an allele frequency below 10% (Figure S8). Genes harbouring one or more LOF mutations were categorized into two groups: unique and multi-copy. We reasoned that a LOF mutation in a unique gene would necessarily result in phenotypic consequences. We found that only 706 (6.6%) of genes were single copy, while the remaining 9957 (93.4%) had at least one other copy. This constitutes a significant enrichment ( $P < 0.001$ ) compared to the genome-wide occurrence of gene duplication. LOF mutations in duplicated genes could also have functional consequences if the mutated copy was uniquely expressed as a consequence of neo- or sub-functionalization (Roulin *et al.*, 2013). We assessed this by examining transcriptomic data from 26 tissues (see details in M&M) and found that 9570 of the 9957 duplicated genes (96%) exhibited a unique



**Figure 3** Description of GCHs characterized in the GmHapMap dataset. (a) Distribution of the number of genes that have a given number of predicted GCHs. (b) Distribution of the number of SNPs residing in a 10-kb window in and around genes in soybean according to the number of gene-centric haplotypes (GCHs) defined using HaplotypeMiner. (c) Distribution of the mean length of genes and gene-centric haplotypes (GCHs) according to the number of GCHs defined by HaplotypeMiner. Haplotype length is defined as the distance between the two retained SNP markers that reside to one side and the other (relative to the middle of the gene) and are the furthest apart from one another. (d) Schematic representation of predicted GCHs for *GmGla*. [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

expression pattern (Data S4 & Figure S9). Thus, despite the fact that the vast majority of LOF mutations occur in genes for which there is more than one copy, a large proportion of these genes exhibit unique expression patterns, thus increasing the likelihood that a LOF will result in a detectable phenotype.

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

SO term	SNP	MNP	INS	DEL	Total variants	Genes
Splice site-disrupting (donor)	1270	38	247	205	1760	1640
Splice site-disrupting (acceptor)	1546	52	207	146	1951	1803
Stop codon-introducing	2826	149	100	7	3082	2418
Frameshift-inducing	0	0	4158	6596	10 754	6718
Start/Stop codon-disrupting	345	40	54	45	484	452
Total	5987	279	4766	6999	18 031	13 031
Total number of genes affected by LOF variants*						10 662

\*Some of the genes were affected with more than one LOF mutation; therefore, the total number of genes is lower than the sum of the all genes.

### Application of LOF mutation in soybean functional genomics and breeding

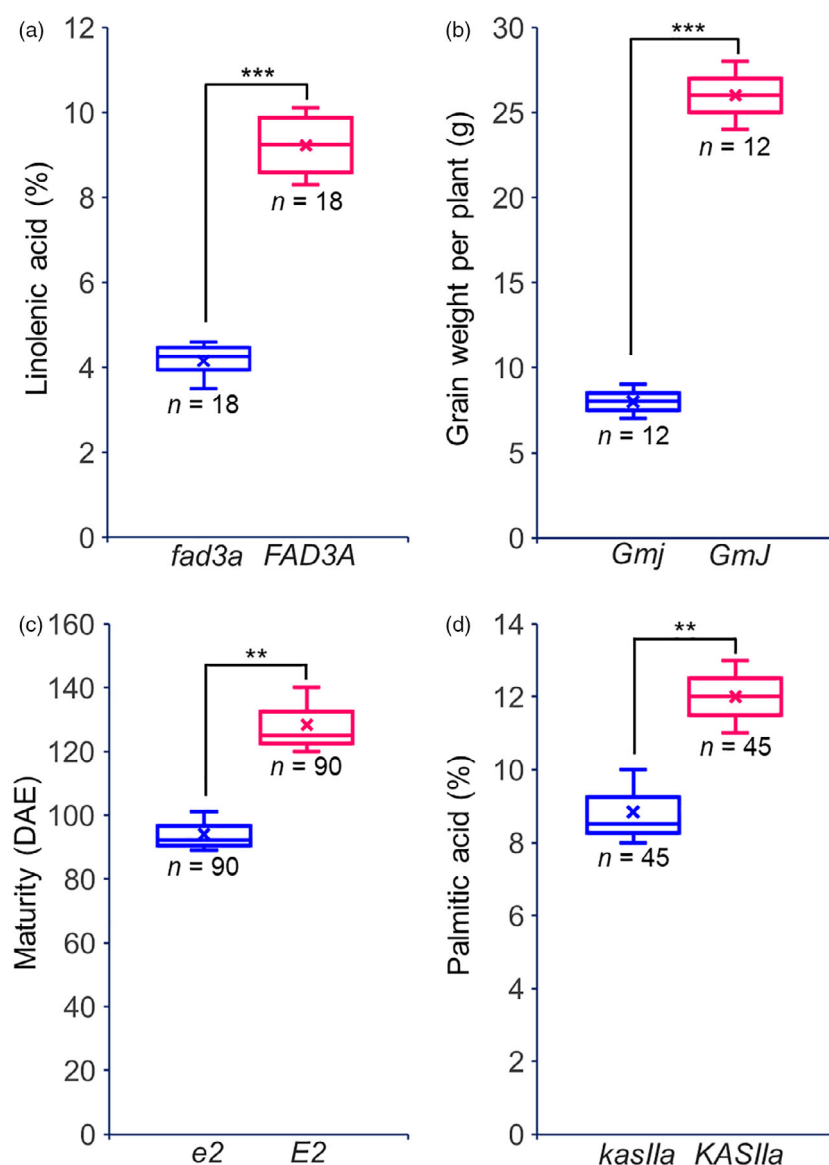
To assess the quality of this catalogue of mutations, we first inspected it for genes already known (i.e. functionally validated) to harbour an LOF mutation. This is indeed the case; all known genes in the literature were found within the catalogue (Data S5). Then, we investigated and confirmed the phenotypic impact of some of these LOF mutations in GmHapMap accessions (Figure 4 and Data S6). A frameshift mutation (frequency = 0.003) in the microsomal omega-3 fatty acid desaturase (*FAD3A*), a key gene for linolenic acid synthesis in soybean seeds (Reinprecht and Pauls, 2016), was found in three accessions. Near-infrared spectroscopy (NIRS) analysis of 36 soybean samples (with or without this LOF mutation) showed a significant ( $P < 0.001$ ) decrease in linolenic acid content in the mutant lines (4%) compared to the wild type (10%) (Figure 4a). A mutation ( $f = 0.005$ ) in *Glyma.04G050200*, the gene underlying the J locus controlling the Long Juvenile trait (Lu et al., 2017), resulted in a significant difference ( $P < 0.002$ ) in grain weight per plant (8 g in the mutant compared to 25 g in the wild type; Figure 4b). The introduction of a premature stop codon ( $f = 0.02$ ) due to a SNP in *GmGla/E2* (Watanabe et al., 2012) significantly ( $P < 0.002$ ) reduced the number of days from emergence to the appearance of the first open flower (DAE) (from 125 in wild-

type lines to 95 in the mutant; Figure 4c). Finally, a SNP ( $f = 0.009$ ) resulted in the disruption of splicing in the gene coding for the 3-ketoacyl-ACP synthase II (KASII) enzyme, a key gene in the oil biosynthesis pathway (Goettel *et al.*, 2016). NIRS analysis of palmitic acid levels showed a significant ( $P < 0.04$ ) decrease in the mutant lines (9%) compared to the wild type (12%) (Figure 4d). The development of a catalogue of LOF mutations represents a valuable resource for functional genomics.

## Discussion

Using whole-genome sequencing data from a large collection of 1,007 soybean accessions, we developed a haplotype map of soybean (GmHapMap), a valuable resource for soybean genetic studies and breeding. A first challenge was to create a uniform and accurate catalogue of nucleotide variation using a common version of the reference genome and a single bioinformatics pipeline (Lek *et al.*, 2016; Roy *et al.*, 2018). The GmHapMap produced here is not only uniform but also achieved a high level of genotype accuracy (>99.7%). To create a representative

haplotype map, a good level of saturation of both variants and haplotypes is required. Close to 15 M sequence variants (SNPs, MNPs and indels) were called that captured nearly all polymorphisms with MAF > 1% in the USDA *G. max* germplasm collection (Song *et al.*, 2013). The number of sequence variants did not increase significantly beyond the first 600 accessions, suggesting that a collection of this size has succeeded in capturing a sizeable fraction of worldwide nucleotide variation within cultivated soybean. Similarly, the number of unique haplotypes (4.3 M tag SNPs) also plateaued relatively early within this collection of soybean germplasm. Together, these data suggest that the 15 M variants captured in GmHapMap are both highly accurate and comprehensive of the genetic diversity within cultivated soybean at a worldwide level. GmHapMap brings more resolution to the within-species diversity of *G. max*. A lower level of genome-wide genetic diversity was observed here in soybean (mean  $\theta_\pi = 1.36 \times 10^{-3}$ ) compared to other major crops such as rice ( $\theta_\pi = 4.0 \times 10^{-3}$ ) (Wang *et al.*, 2018a) and corn ( $\theta_\pi = 6.6 \times 10^{-3}$ ; Chia *et al.*, 2012). It is presumed that several genetic bottlenecks, as well as strong selection pressure, have reduced genetic



**Figure 4** Phenotypic variation observed between accessions with (blue) and without (red) a predicted LOF mutation in four different genes. (a) *FAD3A*, a key gene for linolenic acid synthesis; (b) *Gmj*, a key gene of Long Juvenile trait; (c) *GmGla*, a key gene controlling maturity; (d), *KASIIa*, a key gene in the oil biosynthesis pathway. In each case, the number of accessions sharing the same allele (and for which phenotypic data were at hand) is indicated. [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

diversity in soybean (Hyten *et al.*, 2006). In addition, modern soybean breeding is founded on a very limited number of the founder accessions (Hymowitz and Harlan, 1983).

The GmHapMap was used as a reference panel and more than 21 K accessions that had been previously genotyped using common approaches (SNP array and/or GBS) and obtained an imputation accuracy of 92%–96% for common variants and ~80% for rare variants. The accuracy levels, obtained here, are comparable to the 98% reported by Bukowski *et al.* (2018) in maize and Wang *et al.* (2018a) in rice. The success of untyped-genotype imputation depends critically on how well a reference panel has captured the relevant haplotype diversity, as well as the marker density of the experimental dataset (Browning and Browning, 2016). Here, we document that GmHapMap provides an extensive capture of SNP and haplotype diversity within cultivated soybeans worldwide. It is likely that the lower imputation accuracy observed for the SNP array dataset can be attributed to the relatively low marker density of this dataset.

Enhanced datasets resulting from large-scale imputation can improve the efficacy of GWAS analysis (Hao *et al.*, 2009; Marchini and Howie, 2010; Wang *et al.*, 2018b). To illustrate the benefits of the GmHapMap resource for GWAS, we performed a local association analysis on soybean seed oil content using imputed SNPs. A strong association with an imputed SNP residing in the *NPC1* gene was detected, and its orthologue in *Arabidopsis* is known to contribute to seed oil content (Feldman *et al.*, 2015). The nearest significantly associated GBS-derived variant (i.e. in the absence of the imputation made possible thanks to the GmHapMap data) was located 100 kb upstream of this gene and exhibits a relatively low degree of LD ( $r^2 = 0.5$ ). This shows that an enhanced dataset, obtained through the imputation of untyped variants, can improve the power of GWAS analysis. Several studies in human (Li *et al.*, 2009), cattle (Santana *et al.*, 2014), pig (Yan *et al.*, 2017), maize (Yang *et al.*, 2014) and rice (Wang *et al.*, 2018b) have demonstrated the capacity of imputation to improve the power of GWAS analysis. In the coming years, we expect that soybean researchers will deploy GmHapMap for imputation and more precise dissection of the genetic basis of complex traits in soybean.

This is the first time that a comprehensive description of GCHs, for the complete set of genes (55 589), has been achieved for a species. This catalogue of GCHs was obtained using HaplotypeMiner (Tardivel *et al.*, 2019). Tardivel *et al.* (2019) reported that HaplotypeMiner allowed the identification of SNP haplotypes for which 97.3% of lines sharing a same haplotype were correctly identified as having the same allele. It has been well documented that haplotypes are more informative than single biallelic SNPs (Stephens *et al.*, 2001). Knowledge of the GCHs (and possibly alleles) can greatly facilitate the establishment of a functional link between the various alleles of a gene and the associated phenotype. Haplotype–phenotype association revealed the functional alleles of several genes in wheat (Jiang *et al.*, 2015), maize (Yang *et al.*, 2013), rice (Si *et al.*, 2016) and soybean (Langewisch *et al.*, 2014). Knowledge of the alleles present at one or many genes can be tremendously important to breeders. Epistatic interactions between specific alleles and the effects of alleles at neighbouring loci (carried along via linkage drag) can be very important when considering which combinations of alleles will be most desirable to achieve a given phenotype.

A final aspect of this work is that the identification of LOF mutations in soybean protein-coding genes. GmHapMap includes

a set of nearly 11K knocked-out genes. We suggest that this catalogue of knocked-out genes is highly advantageous for soybean functional genomics for investigation of gene function, and application as genetic makers in soybean breeding programs.

The next challenge will be to link genetic variation, GCHs and LOFs derived from GmHapMap with agronomic traits. This will need an extensive effort to measure phenotypes under multiple field and laboratory conditions. We believe that GmHapMap will lead and accelerate the soybean breeding efforts and future sustainable agriculture.

## Experimental procedures

### GmHapMap sequencing data

Two collections of soybeans were used: (i) a set of 727 accessions for which whole-genome sequencing had been previously released for representative diverse accessions (core collection) for multiple countries (Zhou *et al.*, 2015), Brazil (Maldonado dos Santos *et al.*, 2016), USA (Valliyodan *et al.*, 2016), China (Fang *et al.*, 2017), soybean nested association mapping (NAM) parents (Song *et al.*, 2017) and Canada (Torkamaneh *et al.*, 2017a), and (ii) a set of 280 accessions which were sequenced as part of this work. The latter accessions were chosen to provide a more balanced representation of various soybean-growing areas in the world. 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 280 accessions using the KAPA Hyper Prep Kit (Kapa Biosystems, Wilmington, Massachusetts, USA) following the manufacturer's instructions (KR0961 – v5.16). Samples were sequenced on an Illumina HiSeq X10 platform at the McGill University-Génome Québec Innovation Center in Montreal, QC, Canada.

### Identification of nucleotide variants

Sequencing reads from all 1007 accessions were processed using a single analytical bioinformatics pipeline (Fast-WGS; Torkamaneh *et al.*, 2017a) to create a uniform catalogue of genetic variants. In brief, the 100–150-bp paired-end reads were mapped against the *G. max* reference genome [Gmax\_275 (Wm82.a2)] (Schmutz *et al.*, 2010). Then, we removed variants if (i) they had more than two alleles, (ii) an allele was not supported by reads on both strands, (iii) the overall quality (QUAL) score was <32, (iv) the mapping quality (MQ) score was <30, (v) read depth (minNR) was <2 and (vi) the minor allele frequency (MinMAF) was <0.0009.

### Determining the accuracy of nucleotide variants

The SoySNP50K iSelect BeadChip has been used to genotype the entire USDA soybean germplasm collection (Song *et al.*, 2013). The complete dataset for 20 087 *G. max* and *G. soja* accessions genotyped with 42 508 SNPs was downloaded from SoyBase (Grant *et al.*, 2010). We extracted the genotype calls at all SNP loci for which data were available for 420 accessions which were in common with the GmHapMap collection. This large set of SoySNP50K genotype calls (~18 M genotypes or data points) was directly compared with the WGS-derived SNP calls (obtained using the Fast-WGS pipeline) to assess genotype accuracy.

## Determining the effects of nucleotide variants

The functional impact of nucleotide variants was performed using SnpEff and SnpSift (Cingolani *et al.*, 2012). To identify LOF mutations, a database was built using 55 K soybean protein-coding genes (Gmax\_275\_Wm82.a2.v1.gene.gff3, from Phytozome on Jan. 2016) for SnpEff. Variants were mapped on to transcripts annotated as 'protein\_coding' and containing an annotated 'START' codon, and then classified as synonymous, missense and non-sense (stop codon-introducing, start/stop codon-disrupting or splice site-disrupting (canonical splice sites)). In this work, we excluded transcripts labelled as NMD (predicted to be subject to non-sense-mediated mRNA decay). We also applied a second filtering step, based on annotation, to identify high-confidence knocked-out genes. The genes with LOF mutations were removed if (i) the 'REF' field in the input VCF file did not match the reference genome, (ii) they had an incomplete transcript or (iii) they did not have a proper START codon.

## Population structure and genetic diversity

Population structure was estimated using variational Bayesian inference implemented in fastSTRUCTURE (Raj *et al.*, 2014). Five runs were performed for each number of populations (K) set from 1 to 15. Then, a ChooseK analysis was conducted to determine the number of subpopulations. A principal component analysis (PCA) (Jolliffe, 2011) was conducted in PLINK (Purcell *et al.*, 2007). Total variation per vector (calculated from the 100-eigenvalue proportion of each vector) was used to generate a scree plot. A neighbour-joining unrooted phylogenetic tree (Saitou and Nei, 1987) was constructed in MEGA7 (Kumar *et al.*, 2016). The taxa were clustered together, and the reliability of these clusters was assessed by bootstrapping (1000 replicates; Felsenstein, 1985). All population structure analyses were conducted using complete set of GmHapMap SNPs (12M). We measured the nucleotide diversity ( $\pi$ ) (Nei and Li, 1979) in sliding windows of 1000 bp across the genome using --window-pi option of VCFtools (Danecek *et al.*, 2011). Similarly, the pairwise  $\pi$  was calculated among different subpopulations.

## Linkage disequilibrium and tag SNP identification

Genome-wide pairwise linkage disequilibrium (LD) analysis ( $r^2$  and  $D'$ ) was performed using all nucleotide variants from the GmHapMap dataset. The average  $r^2$  value was calculated for sliding windows of 1000 bp, and LD decay was calculated using PopLDdecay (Zhang *et al.*, 2019). For tag SNP selection, we used PLINK (Purcell *et al.*, 2007) to calculate 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. We then randomly selected subsets of samples of increasing size ( $n = 100, 200, \dots$ , and 1007) and calculated cumulative number of variants and tag SNPs with 20 iterations.

## Imputation of untyped variants

We used two reference panels for untyped-variant imputation. The 'REF-I' panel includes 1,007 accessions from GmHapMap with the entire SNP dataset, while the 'REF-II' panel includes 1007 accessions and only 1.9 M tag SNPs from genic regions (tag SNPs in genic regions or within 2 kb of a gene). The latter panel (REF-II) was developed for use in imputing missing loci from SNP datasets derived from the SoySNP50K array. The SNPs included on this array are essentially located within genes and do not offer

sufficient coverage of the intergenic regions to make imputation with the entire set of SNPs (REF-I) reliable. These two reference panels were created for all 20 chromosomes of soybean and were phased using BEAGLE v4.1 (Browning and Browning, 2016) with 100 iterations.

As initial lower density datasets, we used three collections of soybean accessions genotyped with commonly used genotyping tools. A first set of 20 087 accessions (the entire USDA Soybean Germplasm Collection) had been characterized using the SoySNP50K iSelect Bead Chip (Song *et al.*, 2013) to yield a set of 43 K polymorphic markers. A second set comprised 1531 accessions which had been subjected to genotyping-by-sequencing (GBS; ApeKI protocol; Sonah *et al.*, 2013) and in which SNPs had been called using the Fast-GBS pipeline (Torkamaneh *et al.*, 2017b). Finally, the same set of 1531 accessions (GBS set) was used, but the original catalogue of GBS-derived SNPs was complemented with additional SNPs from the SoySNP50K array.

Phasing and imputation were performed using BEAGLE v4.1 (Browning & Browning, 2016) for each chromosome with the following parameters: (i) nthreads = 10 (number of threads); (ii) window = 100 000 (number of markers in a sliding window); (iii) overlap = 50 000 (number of overlapping markers between adjacent windows); (iv) iterations = 100 (number of phasing iterations); and (v) err = 0.00001 (the allele miscall rate).

## Determining the imputation accuracy

The WGS SNP data from 1006 of the 1007 resequenced accessions were used as a reference panel to impute untyped variants. The remaining line was kept out of the reference panel (leave-one-out design; Ramnarine *et al.*, 2015) to determine how accurately data at untyped loci (present in the GmHapMap data but absent from the low-density genotype catalogue) could be imputed in this accession. We performed three such permutations where a single accession (Gm\_H083, Gm\_H059 and Gm\_H586) was kept aside to estimate imputation accuracy. These three accessions were randomly selected from a set of accessions for which both GBS and SNP array data were available. Imputation accuracy was assessed as the degree of concordance between the imputed genotypes in the accession that had been left out of the reference panel and the true genotypes.

## Marker-trait association analysis

Sonah *et al.* (2015) described a set of QTLs from GWA analysis on a subset of 139 soybean accessions. These accessions were genotyped via GBS. We imputed untyped variants on this low-density genotype dataset from GmHapMap in 1 Mb of chromosome 14, encompassing a QTL for seed oil content. GWA analysis was conducted using GAPIT R package (Lipka *et al.*, 2012) using an MLM model. A candidate gene was identified using SoyBase database (Grant *et al.*, 2010) and The Arabidopsis Information Resource (TAIR) [<https://www.arabidopsis.org/servlets/TairObject?type=gene&name=AT4G38350.1>].

## Identification of gene-centric haplotypes

The identification of GCHs was performed using the HaplotypeMiner R package (<https://github.com/malemay/HaplotypeMiner>; Tardivel *et al.*, 2019) with the entire SNP dataset on 55,381 protein-coding genes in the soybean genome. In brief, the following parameters were used: (i) R2\_measure = 'r2s' (the estimation of linkage disequilibrium between markers was measured based on corrected  $r_{vs}^2$  which takes into account information related to genetic relatedness and population

structure); (ii) cluster\_R2 = 'r2s' (LD measure to use in the clustering step); (iii) max\_missing\_threshold = 0.05 (the maximum proportion of missing genotypes allowed for a marker); (iv) max\_het\_threshold = 0.01 (the maximum proportion of heterozygous genotypes allowed for a marker); (v) min\_allele\_count = 4 (the minimum number of times the minor allele has to be seen for a marker to be retained); (vi) cluster\_threshold = 0.9 (the minimum LD beyond which markers were clustered); (vii) max\_flanking\_pair\_distance = 10 000 (the maximum distance (in bp) that can separate two markers in LD at the final selection step); (viii) max\_marker\_to\_gene\_distance = 6000 (the maximum distance (in bp) from a marker to the centre of the gene of interest); and (ix) marker\_independence\_threshold = 0.8 (the minimum LD for two markers to be considered in LD at the final selection step).

### Identification of duplicated genes

We detected putative duplicated genes, presumably derived from WGD or gene duplication, using protein homology analysis integrated in the Phytozome (Goodstein *et al.*, 2012) database. Protein homologs were identified using dual-affine Smith–Waterman alignments (Smith and Waterman, 1981) between the predicted translation product of the selected transcript (aka query gene) and all other predicted proteins in the soybean genome. We identified duplicated genes with 90% identity ( $ID \geq 90$ ), 90% coverage ( $CV \geq 90$ ) and 5% size difference ( $SD \leq 5$ ) threshold. Putative duplicated genes, presumably derived from WGD or gene duplication, were identified using the soybean reference genome (Gmax\_275) (Wm82.a2; Schmutz *et al.*, 2010) and protein homology analysis integrated in the Phytozome (Goodstein *et al.*, 2012) database. We then downloaded the complete transcriptome dataset for 26 tissues from the Phytozome database (Goodstein *et al.*, 2012). We extracted transcriptome datasets for the genes affected by LOF mutations and their duplicated copies identified in this study. The expression level was measured for these genes using FPKM (fragments per kilobase of exon per million fragments mapped) values (Conesa *et al.*, 2016). As the expression level varied among different tissues, we declared that a gene was not expressed when its FPKM value was equal to 0 or below  $-2\sigma_{\text{mean}} = \frac{\sigma}{\sqrt{N}}$  (defined for each tissue; Figure S8). In general, we observed three different situations: (i) both/all gene copies showed a similar expression level in all tissues; (ii) the gene bearing an LOF mutation and its duplicate(s) showed a quantitatively different expression level in all tissues; and (iii) the gene copy with a LOF allele showed a distinct expression pattern in some tissues. In the latter cases, we expect that the LOF allele will result in a phenotypic difference that could be uncovered.

### Acknowledgements

This work was supported by the SoyaGen grant (www.soyagen.ca) awarded to F. Belzile and funded by 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. The work conducted by the US DOE Joint Genome Institute is supported by the Office of Science of the US DOE under Contract DE-AC02-05CH11231. We thank the Joint Genome Institute and collaborators for

pre-publication access to the Glycine max W82 V2 genome sequence. We thank Dr. Brain Boyle to help for DNA library construction. We thank Dr. Gary Stacy and Dr. Shakhawat Hossain for pre-publication access to the soybean transcriptome atlas. We also thank SoyBase team (Dr. Steven Cannon, Dr. Rex Nelson and Dr. Anne Brwon) for integrating GmHapMap datasets in SoyBase.

### Competing interests

The authors declare that they have no competing interests.

### Author contributions

DT and FB conceived the project and contributed to writing the manuscript. DT and JL contributed to programming and analysis of genomic data. DT carried out genetic diversity analysis, imputation, GCHs and identification of LOFs. BV and HN provided sequence data for American accessions. RA provided data for Brazilian accessions. DT, FB, EC, IR and LO provided data for Canadian accessions and also carried out NIRS analysis. AS and JS carried out the identification of gene duplication.

### Data and code availability

The complete datasets and the bioinformatics codes and scripts are publicly available at SoyBase (<https://soybase.org/projects/SoyBase.C2020.01.php>) and FigShare ([https://figshare.com/projects/Soybean\\_Haplotype\\_Map\\_GmHapMap\\_A\\_Universal\\_Resource\\_for\\_Soybean\\_Translational\\_and\\_Functional\\_Genomics/56921](https://figshare.com/projects/Soybean_Haplotype_Map_GmHapMap_A_Universal_Resource_for_Soybean_Translational_and_Functional_Genomics/56921)).

### References

- Browning, B.L. and Browning, S.R. (2016) Genotype Imputation with Millions of Reference Samples. *Am. J. Hum. Genet.* **98**, 116–126.
- Bukowski, R., Guo, X., Lu, Y., Zou, C., He, B., Rong, Z., Wang, B. *et al.* (2018) Construction of the third-generation Zea mays haplotype map. *GigaScience*, **7**, gix134.
- Chia, J.M., Song, C., Bradbury, P.J., Costich, D., de Leon, N., Doebley, J., Elshire, R.J. *et al.* (2012) Maize HapMap2 identifies extant variation from a genome in flux. *Nat. Genet.* **44**, 803–807.
- Chung, W.H., Jeong, N., Kim, J., Lee, W.K., Lee, Y.G., Lee, S.H., Yoon, W. *et al.* (2014) Population structure and domestication revealed by high-depth resequencing of Korean cultivated and wild soybean genomes. *DNA Res.* **21**, 153–167.
- Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J. *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**, 80–92.
- Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szczesniak, M.W. *et al.* (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13.
- Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.
- Djanaguiraman, M., Schapaugh, W., Fritschi, F., Nguyen, H., Prasad, P.V.V. (2019) Reproductive success of soybean (*Glycine max* L. Merrill) cultivars and exotic lines under high daytime temperature. *Plant Cell Environ.* **42**, 321–336.
- Fang, C., Ma, Y., Wu, S., Liu, Z., Wang, Z., Yang, R., Hu, G. *et al.* (2017) Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. *Genome Biol.* **18**, 161.
- Feldman, M.J., Poirier, B.C. and Lange, B.M. (2015) Misexpression of the Niemann-Pick disease type C1 (NPC1)-like protein in Arabidopsis causes sphingolipid accumulation and reproductive defects. *Planta*, **242**, 921.

- Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791.
- Gizlice, Z., Carter Jnr, T.E. and Burton, J.W. (1994) Genetic base for North American public soybean cultivars released between 1947 and 1988. *Crop Sci.* **34**, 1143–1151.
- Goettel, W., Ramirez, M., Upchurch, R.G. and An, Y.C. (2016) Identification and characterization of large DNA deletions affecting oil quality traits in soybean seeds through transcriptome sequencing analysis. *Theoret. Appl. Genet.* **129**, 1577–1593.
- Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T. *et al.* (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, D1178–D1186.
- Grant, D., Nelson, R.T., Cannon, S.B. and Shoemaker, R.C. (2010) SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucl. Acids Res.* **38**, D843–D846.
- Hao, K., Chudin, E., McElwee, J. and Schadt, E. (2009) Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies. *BMC Genet.* **10**, 27.
- Hungria, M. and Mendes, I.C. (2015) Nitrogen fixation with soybean: the perfect symbiosis? In: *Biological Nitrogen Fixation* (de Bruijn F.J., ed), pp. 1005–1019. New Jersey: John Wiley & Sons, Inc (ISBN: 978-1-118-63704-3).
- Hymowitz, T. and Harlan, J.R. (1983) Introduction of soybean to North America by Samuel Bowen in 1765. *Econ. Bot.* **37**, 371–379.
- Hyten, D.L., Song, Q., Zhu, Y., Choi, I.-Y., Nelson, R.L., Costa, J.M., Specht, J.E. *et al.* (2006) Impacts of genetic bottlenecks on soybean genome diversity. *Proc. Natl. Acad. Sci. USA*, **103**, 16666–16671.
- Jensen, S.E., Charles, J.R., Kebede, M., Bradbury, P.J., Casstevens, T., Deshpande, S.P., Gore, M.A. *et al.* (2020) A sorghum practical haplotype graph facilitates genome-wide imputation and cost-effective genomic prediction. *Plant Genome*, **13**, e20009.
- Jiang, Y., Jiang, Q., Hao, C., Hou, J., Wang, L., Zhang, H., Zhang, S. *et al.* (2015) A yield-associated gene TaCWI, in wheat: its function, selection and evolution in global breeding revealed by haplotype analysis. *Theor. Appl. Genet.* **128**, 131–143.
- Jolliffe, I. (2011) Principal component analysis. In *International Encyclopedia of Statistical Science* (Lovric, M., ed.), Berlin, Heidelberg: Springer.
- Kono, T.J.Y., Lei, L., Shih, C.H., Hoffman, P.J., Morrell, P.L. and Fay, J.C. (2018) Comparative genomics approaches accurately predict deleterious variants in plants. *G3 (Bethesda)*, **8**, 3321–3329. Published 2018 Oct 3.
- Kumar, S., Stecher, G. and Tamura, K. (2016) MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33**, 1870–1874.
- Lam, H.M., Xu, X., Liu, X., Chen, W., Yang, G., Wong, F.-L., Li, M.W. *et al.* (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat. Genet.* **42**, 1053–1059.
- Langewisch, T., Zhang, H., Vincent, R., Joshi, T., Xu, D. and Bilyeu, K. (2014) Major soybean maturity gene haplotypes revealed by SNPviz analysis of 72 sequenced soybean genomes. *PLoS One*, **9**, e94150.
- Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H. *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 285–291.
- Li, Y., Willer, C. and Sanna, S. (2009) Genotype Imputation. *Annu. Rev. Genomics Hum. Genet.* **10**, 387–406.
- Lipka, A.E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P.J., Gore, M.A. *et al.* (2012) GAPIT: genome association and prediction integrated tool. *Bioinformatics*, **28**, 2397–2399.
- Liu, S., Kandath, P., Lakhssassi, N., Kang, J., Colantonio, V., Heinz, R., Yeckel, G. *et al.* (2017) The soybean GmSNAP18 gene underlies two types of resistance to soybean cyst nematode. *Nat. Commun.* **8**, 14822.
- Lu, S., Zhao, X., Hu, Y., Liu, S., Nan, H., Li, X., Fang, C. *et al.* (2017) Natural variation at the soybean J locus improves adaptation to the tropics and enhances yield. *Nat. Genet.* **49**, 773–779.
- MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L. *et al.* (2012) A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, **335**, 823–828.
- Maldonado dos Santos, J.V., Valliyodan, B., Joshi, T., Khan, S.M., Liu, Y. *et al.* (2016) Evaluation of genetic variation among Brazilian soybean cultivars through genome resequencing. *BMC Genom.* **13**, 110.
- Marchini, J. and Howie, B. (2010) Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511.
- Nei, M. and Li, W.H. (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. USA*, **76**, 5269–5273.
- Nicolas, P., Sun, F. and Li, L.M. (2006) A model-based approach to selection of tag SNPs. *BMC Bioinformatics* **7**, 303.
- Pääbo, S. (2003) The mosaic that is our genome. *Nature*, **421**, 409–412.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D. *et al.* (2007) PLINK: a tool set for whole genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575.
- Qiu, L.J., Xing, L.L., Guo, Y., Wang, J., Jackson, S.A. and Chang, R.Z. (2013) (2013) A platform for soybean molecular breeding: the utilization of core collections for food security. *Plant Mol Biol.* **83**, 41–50.
- Raj, A., Stephens, M. and Pritchard, J.K. (2014) fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*, **197**, 573–589.
- Ramnarine, S., Zhang, J., Chen, L.S., Culverhouse, R., Duan, W., Hancock, D.B. *et al.* (2015) When does choice of accuracy measure alter imputation accuracy assessments? *PLoS One*, **10**, e0137601.
- Ray, D.K., Mueller, N.D., West, P.C. and Foley, J.A. (2013) Yield trends are insufficient to double global crop production by 2050. *PLoS One*, **8**, e66428.
- Reinprecht, Y. and Pauls, K.P. (2016) Microsomal omega-3 fatty acid desaturase genes in low linolenic acid soybean line RG10 and validation of major linolenic acid QTL. *Front. Genet.* **7**, 38.
- Roulin, A., Auer, P.L., Libault, M. *et al.* (2013) The fate of duplicated genes in a polyploid plant genome. *Plant J.* **73**, 143–153.
- Roy, S., Coldren, C., Karunamurthy, A. *et al.* (2018) Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: A joint recommendation of the Association for Molecular Pathology and the College of American Pathologists. *J. Mol. Diagn.* **20**, 4–27.
- Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406.
- Santana, M.H., Utsunomiya, Y.T., Neves, H.H., Gomes, R.C., Garcia, J.F., Fukumasu, H. *et al.* (2014) Genome-wide association analysis of feed intake and residual feed intake in Nellore cattle. *BMC Genet.* **15**, 21.
- Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W. *et al.* (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.
- Si, L., Chen, J., Huang, X., Gong, H., Luo, J., Hou, Q. *et al.* (2016) OsSPL13 controls grain size in cultivated rice. *Nat. Genet.* **48**, 447–456.
- Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197.
- Sonah, H., Bastien, M., Iquira, E., Tardivel, A., Legare, 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**, e54603.
- Sonah, H., O'Donoghue, L., Cober, E., Rajcan, I. and Belzile, F. (2015) Identification of loci governing eight agronomic traits using a GBS–GWAS approach and validation by QTL mapping in soya bean. *Plant Biotech. J.* **13**, 211–221.
- Song, Q., Hyten, D.L., Jia, G., Quigley, C.V., Fickus, E.W., Nelson, R.L. *et al.* (2013) Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. *PLoS One*, **8**, e54985.
- Song, Q., Yan, L., Quigley, C., Jordan, B.D., Fickus, E., Schroeder, S., Song, B.H. *et al.* (2017) Genetic characterization of the soybean nested association mapping population. *Plant Genome*. **10**(2). <https://doi.org/10.3835/plantgenome2016.10.0109>
- Sosso, D., Luo, D., Li, Q.-B., Sasse, J., Yang, J. *et al.* (2015) Seed filling in domesticated maize and rice depends on SWEET-mediated hexose transport. *Nat. Genet.* **47**, 1489–1493.
- Stephens, J.C., Schneider, J.A., Tanguay, D.A., Choi, J., Acharya, T. *et al.* (2001) Haplotype variation and linkage disequilibrium in 313 human genes. *Science*, **293**, 489–493.
- Tardivel, A., Torkamaneh, D., Lemay, M.A., Belzile, F. and O'Donoghue, L.S. (2019) A systematic gene-centric approach to define haplotypes and identify alleles on the basis of dense single nucleotide polymorphism datasets. *Plant Genome*, **12**, 180061.

- The International HapMap Consortium (2003) The International HapMap Project. *Nature*, **426**, 789–796.
- Torkamaneh, D., Laroche, J., Tardivel, A., O'Donoghue, L., Cober, E., Rajcan, I. and Belzile, F. (2017a) Comprehensive description of genome-wide nucleotide and structural variation in short-season soybean. *Plant Biotechnol. J.* **16**, 749–759.
- Torkamaneh, D., Laroche, J., Bastien, M., Abed, A. and Belzile, F. (2017b) Fast-GBS: a new pipeline for the efficient and highly accurate calling of SNPs from genotyping-by-sequencing data. *BMC Bioinformatics*, **18**, 5.
- Torkamaneh, D., Laroche, J., Rajcan, I. and Belzile, F. (2018) Identification of candidate domestication-related genes with a systematic survey of loss-of-function mutations. *Plant J.* **96**, 1218–1227.
- Subokura, Y.S., Watanabe, Z., Xia, H., Kanamori, H., Yamagata, A., Kaga, Y. et al. (2014) Natural variation in the genes responsible for maturity loci E1, E2, E3 and E4 in soybean. *Ann. Bot.* **113**, 429–441.
- Valliyodan, B., Qiu, D., Patil, G., Zeng, P., Huang, J., Dai, L., Chen, C. et al. (2016) Landscape of genomic diversity and trait discovery in soybean. *Sci. Rep.* **6**, 23598.
- Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Taiet, S., Wu, Z., Li, M. et al. (2018a) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*, **557**, 43–49.
- Wang, D.R., Agosto-Pérez, F.J., Chebotarov, D., Shi, Y., Marchini, J., Fitzgerald, M., McNally, K.L. et al. (2018b) An imputation platform to enhance integration of rice genetic resources. *Nat. Commun.* **9**, 3519.
- Watanabe, S., Harada, K. and Abe, J. (2012) Genetic and molecular bases of photoperiod responses of flowering in soybean. *Breed. Sci.* **61**, 531–543.
- Yan, G., Qiao, R., Zhang, F., Xin, W., Xiao, S., Huang, T., Zhang, Z. et al. (2017) Imputation-based whole-genome sequence association study rediscovered the missing QTL for lumbar number in suture pigs. *Sci. Rep.* **7**, 615.
- Yang, Q., Li, Z., Li, W., Ku, L., Wang, C., Ye, J., Li, K. et al. (2013) CACTA-like transposable element in ZmCCT attenuated photoperiod sensitivity and accelerated the post domestication spread of maize. *Proc. Natl. Acad. Sci. USA*, **110**, 16969–16974.
- Yang, N., Lu, Y., Yang, X., Huang, J., Zhou, Y., Ali, F., Wen, W. et al. (2014) Genome wide association studies using a new nonparametric model reveal the genetic architecture of 17 agronomic traits in an enlarged maize association panel. *PLoS Genet.* **10**, e1004573.
- Zhang, C., Dong, S.S., Xu, J.Y., He, W.M. and Yang, T.L. (2019) PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics*, **35**, 1786–1788.
- Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., Yu, Y. et al. (2015) Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* **33**, 408–414.

## Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**Figure S1.** Frequency distribution of nucleotide variants in the GmHapMap dataset.

**Figure S2.** Population structure of GmHapMap accessions. Left, a scree plot representing the estimated number of subpopulations based on a principal component analysis (PCA). Right, unrooted phylogenetic tree of all accessions inferred from whole-genome SNPs representing existing genetic diversity and admixture among GmHapMap accessions.

**Figure S3.** Decay of linkage disequilibrium (LD) in the soybean genome.

**Figure S4.** Distribution of SNPs derived from three different genotyping platforms (whole-genome sequencing (WGS), genotyping-by-sequencing (GBS) and SNP array (SoySNP50K)) on chromosome 3 (Chr03) of soybean.

**Figure S5.** Distribution of genome-wide SNPs in genic and intergenic regions of soybean for four different genotype datasets (soybean haplotype map (GmHapMap), genotyping-by-sequencing (GBS), SNP array (SoySNP50K) and combined (combination of GBS and SoySNP50K)).

**Figure S6.** Manhattan plot of a regional association mapping for seed oil content using imputed data. The significance threshold was drawn at an FDR-adjusted *p*-value (or *q*-value) below 0.01.

**Figure S7.** Identified GCHs for (a) *E1* (a soybean maturity gene) (Langewisch et al., 2014), (b) *E3* (a soybean maturity gene) (Langewisch et al., 2014), (c) *Gmdt1* (a plant architecture gene) (Langewisch et al., 2014) and (d) *GmSNP18* (a key gene controlling soybean-cyst nematode (SCN) resistance) (Liu et al., 2017). Left is the name of known allele, and right is the number of accessions harbouring the haplotype. ND stands for not determined.

**Figure S8.** Frequency distribution of LOF alleles.

**Figure S9.** Schematic representation of expression analysis for duplicated genes. Three broad situations are illustrated: (a) Both/all gene copies showed a similar expression level in all tissues; (b) the gene bearing an LOF mutation and its duplicate(s) show a quantitatively different expression level in all tissues; and (c) the gene copy with a LOF allele shows a distinct expression pattern in some tissues. In the latter cases, we expect that the LOF allele will result in a phenotypic difference that could be uncovered.

**Data S1.** Information of sequenced soybean accessions with name, origin and depth of coverage.

**Data S2.** Description of number of GCHs, haplotype size and number of SNPs used to define GCHs for each gene.

**Data S3.** Description of genes identified in genomic regions with an exceptionally low level of nucleotide diversity.

**Data S4.** Description of LOF mutation/s.

**Data S5.** List of known genes with LOF mutation.

**Data S6.** Description of phenotypic data used to associate with LOF mutations.