Plant Methods

METHODOLOGY

Open Access

3D-GBS: a universal genotyping-by-sequencing approach for genomic selection and other high-throughput low-cost applications in species with small to medium-sized genomes

Maxime de Ronne[1,2,3], Gaétan Légaré[2], François Belzile[1,2,3], Brian Boyle[2] and Davoud Torkamaneh[1,2,3,4*]

Abstract

Despite the increased efficiency of sequencing technologies and the development of reduced-representation sequencing (RRS) approaches allowing high-throughput sequencing (HTS) of multiplexed samples, the per-sample genotyping cost remains the most limiting factor in the context of large-scale studies. For example, in the context of genomic selection (GS), breeders need genome-wide markers to predict the breeding value of large cohorts of progenies, requiring the genotyping of thousands candidates. Here, we introduce 3D-GBS, an optimized GBS procedure, to provide an ultra-high-throughput and ultra-low-cost genotyping solution for species with small to medium-sized genome and illustrate its use in soybean. Using a combination of three restriction enzymes (PstI/NsiI/MspI), the portion of the genome that is captured was reduced fourfold (compared to a "standard" ApeKI-based protocol) while reducing the number of markers by only 40%. By better focusing the sequencing effort on limited set of restriction fragments, fourfold more samples can be genotyped at the same minimal depth of coverage. This GBS protocol also resulted in a lower proportion of missing data and provided a more uniform distribution of SNPs across the genome. Moreover, we investigated the optimal number of reads per sample needed to obtain an adequate number of markers for GS and QTL mapping (500–1000 markers per biparental cross). This optimization allows sequencing costs to be decreased by ~ 92% and ~ 86% for GS and QTL mapping studies, respectively, compared to previously published work. Overall, 3D-GBS represents a unique and affordable solution for applications requiring extremely high-throughput genotyping where cost remains the most limiting factor.

Keywords Genotyping-by-sequencing, Ultra-high-throughput genotyping, Multiplexing, Next-generation sequencing, Genomic selection, Single-nucleotide polymorphism

*Correspondence:
Davoud Torkamaneh
davoud.torkamaneh.1@ulaval.ca
[1] Département de Phytologie, Université Laval, Quebec, Canada
[2] Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Quebec, Canada
[3] Centre de recherche et d'innovation sur les végétaux (CRIV), Université Laval, Quebec, Canada
[4] Institut intelligence et données (IID), Université Laval, Quebec, Canada

Introduction

Genome-wide genotyping of large populations, an essential component in quantitative trait loci (QTL) mapping or genomic selection (GS) studies, is constantly improving to minimize the cost of genotyping per individual sample. The identification of large numbers of molecular markers has been paralleled by the simultaneous development of high-throughput

approaches such as microarray- [18] or sequencing-based genotyping [50]. However, new needs related to applied breeding programs require the development of an ultra-high-throughput and cost-effective genotyping platform. SNP arrays are a popular approach (e.g. BARCSoySNP6K in soybean [55] and C7AIR in rice [42]) providing a robust genotype calling of multiple known polymorphic sites at the same time and across different populations allowing for a direct comparison of data between experiments, germplasm and studies [6, 26]. However, SNP arrays present ascertainment issues [41], an inability to target loci that were not included during the array development and need to be developed independently for each species and population [11]. In addition to these, the cost of genotyping using SNP arrays, even after development, is considerably higher than sequencing-based approaches [15].

While genotyping based on whole-genome sequencing (WGS) remains expensive and sometimes unnecessary in the context of large-scale studies, low- (< 5X) and very low-depth (0.1–0.5X) sequencing approaches, such as skimSeq, have been designed to decrease the sequencing cost for numerous applications in both model and non-model species [37, 58]. However, inferring genotypes from a random sampling of a small percentage of the genome is challenging because very low sequencing coverage often leads to inaccurate genotype calls, particularly for organisms with a high degree of paralogy and or heterozygosity [54]. In contrast, reduced-representation sequencing (RRS) approaches bypass this problem by focusing the sequencing effort on a smaller proportion of the genome that is constant between the genotyped samples (e.g., centered on the exome or on restriction fragments). Combined with high-throughput sequencing (HTS) of multiplexed samples, RRS approaches, allows for a cost-effective genotyping of millions of SNPs in large sets of individuals [23]. Among RRS approaches, genotyping-by-sequencing (GBS) is the most widely used method thanks to its speed, flexibility and cost-effectiveness [21, 43, 47]. In the last decade, GBS has been widely applied in animals [7, 38], plants [2, 72] and fungi [31], where other genotyping tools (e.g., SNP arrays; [18]) were not adapted [8]. The attractiveness of GBS has led to many optimizations related to the choice of enzymes [52], pipeline for calling SNPs [64], improved marker density (double-digest GBS [69] and high-density GBS [65]), and improved library-preparation procedure [62]. Although GBS is the most cost-effective genome-wide genotyping approach, it can still be expensive for routine screening of large populations as required in breeding programs [45, 50, 59]. Nevertheless, GBS could be optimized by focusing sequencing on a lower fraction of the genome allowing more samples to be multiplexed at a lower average sequencing coverage and thus reduce the sequencing cost per sample. Reducing the genome coverage through reduction of sequencing coverage will categorically result in a lower number of markers, however the uniform distribution of these markers is crucial for an efficient and effective genetic study. The appropriate choice of restriction enzymes can also be a challenging point as their recognition sites (based on the size of the enzyme, sensitivity to methylation, and its GC content) are not uniformly distributed across the genomes [24, 34, 39, 44].

The number of required reads is another determining factor in multiplexing and throughput. Despite the various improvements in GBS methods, the estimation of the number of reads for each sample required to achieve an efficient genotyping needs to be determined on a case-by-case basis [66]. An insufficient number of reads per sample will result in a high proportion of missing data, a reduced number of SNP loci at which genotypes can be successfully called and, possibly, an uneven distribution of markers across the genome [14, 25, 60]. In contrast, an excessive number of reads results in an inefficient use of the sequencing effort and therefore, unnecessarily increases per-sample cost [3]. Thus, finding an optimal number of reads per sample can also help minimize per-sample sequencing cost.

To optimize the multiplexing capacity of GBS, a novel combination of three restriction enzymes, hence 3D-GBS, was tested on soybean to reduce the initial number of digested DNA fragments (or sequencing coverage) while producing genotypic data as relevant as stdGBS. The use of this new enzyme combination has improved the distribution of markers across the genome in terms of uniformity and number of gaps compared to stdGBS. Finally, we investigated the optimal number of reads per sample to further maximize multiplexing capacity on a single sequencing run and thereby, significantly minimize the sequencing cost per sample. This approach will greatly facilitate the adoption of ultra-high-throughput genome-wide genotyping where the per-sample cost remains a limiting factor for various applications.

Materials and methods

Biological materials

To compare the GBS [15] and 3D-GBS methods, sixteen soybean accessions (QS4049, QS4054, QS4067, QS5008, QS4028, QS4043, QS5017, OAC Klondike, OAC Bright, Altesse, OAC Inwood, OAC Thames, OAC 08-18C, OAC Morris, OAC Embro and OAC McCall; provided by Dr. Louise O'Donoughue at CEROM, Quebec, QC, Canada) were used in this study. These accessions were selected based on the availability of GBS data [53]. For each accession, seeds

were grown in a growth chamber. Then, approximately 100 mg of young leaf tissues were collected for DNA extraction. Collected leaf tissues were dried for 4 days using a desiccating agent (Drierite; Xenia, OH, USA) and then ground with metallic beads in a RETSCH MM 400 mixer mill (Fisher Scientific, MA, USA). DNA was extracted using the DNeasy Plant Mini Kit (Qiagen, MD, USA) according to the manufacturer's protocol. DNA quantification was done with a Qubit fluorometer using the dsDNA HS assay kit (Thermo Fisher Scientific, MA, USA) and subsequently adjusted to 10 ng/µl for each sample.

3D-GBS library preparation

Choice of enzymes

The restriction enzymes for 3D-GBS were selected based on their sensitivity to methylation and the size of their recognition site compared to ApeKI, a standard GBS protocol for soybean. ApeKI is a 5 bp-cutter with one ambiguous site and 80% GC content (G*CWGC). Here, we used following enzymes: PstI, a 6-bp cutter with 66% GC content (CTGCA*G), NsiI, a 6-bp cutter with 33% GC content (ATGCA*T), and MspI, a 4-bp cutter with 100% GC content (C*CGG). *Ape*KI and *Pst*I are partially sensitive and sensitive to cytosine methylation, respectively, while NsiI and MspI are not sensitive to cytosine methylation.

Library preparation

3D-GBS libraries were prepared on a reduced scale (5 µL reaction volume) according to the NanoGBS protocol [62] with the three selected enzymes (PstI, NsiI and MspI). Briefly, a total of 10 ng of genomic DNA of each sample was used for digestion with the restriction-enzyme mix and then ligation with sample-specific barcoded adapters. The 5' adapters had an overhang compatible with the common overhang produced by PstI and NsiI, while the 3′ adapters had an overhang compatible with that produced by MspI. Then, individual libraries were pooled and a size-selection (50–350 bp) step was done using a BluePippin apparatus (Sage Science, MA, USA). PCR amplification (12 cycles), enrichment, and PCR clean-up were performed before quality control, quantification, and purity assessments of DNA libraries with a spectrophotometer (Nanodrop 1000, Fisher Scientific, MA, USA) and a Bioanalyzer 2100 (Agilent Technologies, CA, USA). The 3D-GBS libraries were then sequenced on an Ion Torrent instrument (Thermo Fisher Scientific, MA, USA) on Ion Proton 540 chips at the Genomic Analysis Platform of the Institut de Biologie Intégrative et des Systèmes (Université Laval, QC, Canada).

Data analysis

Sequencing and genotyping

Sequencing data were processed using the Fast-GBS v2.0 pipeline [64] and the Wm82.a2 soybean reference genome (Gmax_275_Wm82.a2.v1, [51]) for SNP calling. For GBS and 3D-GBS analyses, variant calls were filtered with VCFtools [9] to remove low-quality SNPs (QUAL < 10 and MQ < 30), variants residing on unassembled scaffolds and indels. Then, only biallelic markers with missing data < 0.8 and heterozygosity < 0.1 were retained. This filtering step resulted in the removal of approximately 70% of low-quality variants from both GBS (25,280 to 7904) and 3D-GBS (15,082 to 4826) data. The following statistical analysis were performed using filtered data. The genome coverage (fraction of the genome captured) was determined with the function 'coverage' in Samtools [10] while the mean depth of coverage (sequencing coverage) was calculated using VCFtools with the function '–depth'. The proportion of missing data and heterozygous calls, average minor allele frequency and nucleotide diversity (PiPerBP) were estimated using TASSEL v.5 [5]. In silico digestion analysis was performed using DepthFinder [66] to determinate the number of cutting site across the genome for different combinations of enzymes.

Distribution of markers on the physical and genetic maps

The distribution of markers across the physical map was based on the VCF files generated after Fast-GBS analysis and SNP filtration, using the rMVP package in R [70]. For genetic maps, the genetic position of each SNP was inferred from the closest corresponding SNP on the consensus genetic map based on GBS-derived SNP markers [16]. Then, the distribution of markers across the genetic maps was evaluated using the QTL IciMapping v4.1 software [40].

Random sampling of reads

Different subsets of reads (i.e., 50K, 100K, 200K and 300K reads) were randomly sampled three times for each of the 16 accessions using seqtk [32] with the function 'sample' [32]. Then, sequencing data as well as the number and distribution of SNPs were assessed as mentioned above to compare results generated from each read subgroup. To investigate 3D-GBS results for biparental crosses, the two genetically closest and most distant accessions were determined by using a matrix of pairwise distances generated with TASSEL v.5 [49].

Results and discussion

New enzyme combinations for an efficient and uniform capture of the genome

In this study, sixteen DNA samples that had been previously genotyped with the original ApeKI-based GBS protocol were used to produced 3D-GBS libraries. The 16-plex GBS and 3D-GBS libraires produced ~21.1M (ranging from ~800K to ~2.9M reads/sample) and ~10.4M (ranging from ~300K to ~800K reads/sample) reads, respectively. First, the distribution of the SNPs derived from PstI–MspI and NsiI–MspI reads was investigated to assess the relevance of this enzyme combination (Fig. 1a and Additional file 1: Fig. S1). We found 76.5% and 23.5% of NsiI–MspI and PstI–MspI reads, respectively, encompassing 4206 and 620 SNPs, respectively. The higher proportion of NsiI–MspI-derived fragments and SNPs could be expected because of the methylation insensitivity of NsiI and lower GC content compared to PstI. Nevertheless, PstI–MspI-derived fragments ensured the coverage of large gaps devoid of NsiI–MspI-derived fragments (e.g., on chromosomes 9, 11 and 20).

To perform a meaningful comparison, the same overall number of reads for each 16-plex library was used to compare the two protocols; as the number of reads per

Table 1 Sequencing and SNP-calling data generated from GBS and 3D-GBS libraries of 16 soybean samples

Steps	Measured parameters	GBS	3D-GBS
Sequencing	Mean read count (M)	0.6	0.6
	Coverage (%)[a]	4.7	1.2
	Mean depth of coverage (X)[b]	5.1	14.5
	Mean mapping quality	41	42
SNP calling	SNP count	7904	4826
	Proportion of missing data (%)	33.7	15.3
	Proportion of heterozygous genotypes (%)	4.4	3.8
	Average minor allele frequency (%)	33.8	31.3
	Nucleotide diversity (p per bp)	0.43	0.42
Physical map	SNP/Mb	8.3	5.1
	Number of gaps > 5 Mb	9	6
	Number of gaps > 10 Mb	1	0
Genetic map	SNP/cM[c]	3.7	2.3
	Number of gaps > 10 cM	7	10
	Number of gaps > 20 cM	0	1

[a] Fraction of the genome captured across all 16 libraries

[b] Average number of read at each sequenced position

[c] Inferred from the closest corresponding SNP on the consensus genetic map [16]

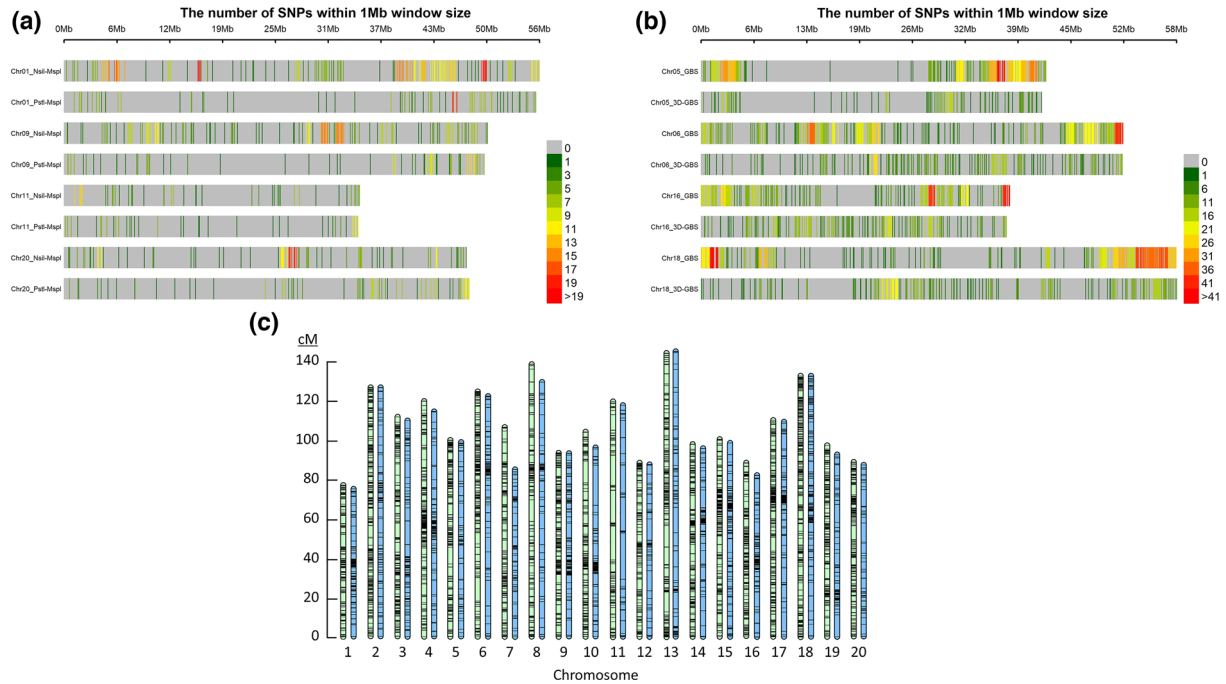

Fig. 1 Distribution of the GBS- and 3D-GBS-derived SNPs across the soybean genome. **a** Distribution of the SNPs derived from NsiI–MspI and PstI–MspI reads on selected chromosomes. The colors of the heatmap correspond to the number of SNPs within 1-Mb windows. **b** Distribution of the SNPs derived from GBS and 3D-GBS on selected chromosomes. **c** Distribution of the SNPs derived from GBS and 3D-GBS libraries across the soybean genetic map. Chromosomes in green and blue represent GBS and 3D-GBS, respectively

accession varied, an identical number of mapped reads for a given accession in each of the two libraries was used to compare GBS and 3D-GBS (Table 1). As expected, with 3D-GBS, a lower fraction of the genome was captured compared to GBS (genome coverage of 1.2% vs 4.7%, respectively). As the sequencing effort (i.e., the number of reads per sample) was focused on a smaller fraction of the genome, the mean depth of coverage was threefold higher in 3D-GBS compared to GBS (14.5X vs 5.1X, respectively) resulting in a lower proportion of missing data (15.3% vs 33.7%, respectively). Fortunately, while the genome coverage was 75% lower for 3D-GBS than GBS data, the number of SNPs identified was only 40% lower (4826 vs 7904 SNPs, respectively), showing that 3D-GBS either captures more polymorphic regions of the genome or improves the genotyping efficiency for the same sequencing effort. As expected, highly similar metrics were obtained for mapping quality, proportion of heterozygous genotypes, average minor allele frequency and nucleotide diversity in both datasets. This suggests that 3D-GBS data is as relevant as GBS data for performing different genetic analyses.

The density of SNPs captured by 3D-GBS (5.1 SNPs/Mb and 2.3 SNPs/cM with no gap > 30 cM) represents an adequate density to perform QTL mapping and GS analysis. To confirm this, the distribution of the SNPs across the physical and genetic maps has been evaluated (Fig. 1b, c, Additional file 2: Fig. S2). Compared to GBS-derived SNPs, the distribution of the 3D-GBS-derived SNPs was more uniform across the genome (Fig. 1b and Additional file 2: Fig. S2). This can be easily illustrated by (i) several regions > 5 Mb on chromosomes 1, 5, 6, 12, 16 and 18 that are missed by GBS while they were covered by 3D-GBS; and (ii) the more uniform distribution of SNPs which rarely exceeds 25 SNPs/Mb in 3D-GBS, compared to GBS where many regions are covered with an "excessive" number of SNPs (25 to more than 41 SNPs/Mb; e.g. on chromosomes 4, 5, 6, 16, 18, etc.). Finally, regarding the genetic map, the 3D-GBS SNPs were well distributed with only one gap close to 20 cM on Chr11, in a region that was also poor in GBS-derived SNPs (Fig. 1c), suggesting that 3D-GBS data are as efficient as GBS data to conduct genetic analyses such as GS or QTL mapping.

The appropriate choice of enzyme(s) is an essential step in developing a GBS protocol [20]. In the original GBS protocol [15], the ApeKI enzyme was used as frequent cutter with sensitivity to methylation to obtain SNPs mainly distributed in gene-rich regions (hypomethylated fraction of the genome) corresponding to a coverage of ~ 4–5% of the genome (Table 1). A two-enzyme strategy using a rare (e.g. PstI) and a frequent cutter (e.g. MspI) sensitive to methylation has also been developed to significantly reduce genome complexity in species with a very large genome (e.g., barley (5 Gb) [46]). However, this approach did not show enough efficiency with species with small to medium genome size [e.g., soybean (~1 Gb)] as it captured relatively few genomic regions [65]. Moreover, due to the palindromic nature of enzyme's restriction sites, this produces a bias in GC content, making the two-enzyme strategy using a rare cutter (none available with 50% GC content) impossible to obtain uniform distribution of fragments in the context of a universal use. Indeed, since there is natural variation in GC content across chromosomes [24, 34, 39, 44] and between species [29, 35], using a rare cutter with either 33% or 66% of GC will inevitably induce variable density of restriction fragments across chromosomes and species. On the other hand, frequent cutters can have a 50% GC content, such as MspI (CCGG) or BfaI (CTAG), allowing a more even distribution of restriction fragments throughout the genome, as illustrated by Torkamaneh et al. [65]. However, when they have been used alone, these frequent cutters induce too many restriction fragments across the genome, which is contrary to the objective of reducing genome coverage.

In light of the above, we explored the idea of improving the two-enzyme approach by using a second rare cutter, such as NsiI [17], with a cutting site differing in GC content and exploiting methylation insensitivity to capture hypermethylated regions missed by PstI. The combination of NsiI with PstI and MspI presented a good opportunity to obtain a sufficient and efficient low density of SNPs distributed more evenly in the genome. While ApeKI would be expected to cut every ~ 512 bp ($4^{4.5}$), here, a combination of three enzymes that include PstI and NsiI (two 6-bp-cutter with differing methylation sensitivity), with a predicted cutting frequency of one site every ~ 4096 bp (4^6), and MspI, a methylation-insensitive 4-bp cutter with an expected cutting frequency of one site every 256 bp (4^4) were used jointly to reduce the fraction of the genome that is captured. The high cutting frequency of MspI allows to generate more fragments of 100–400 bp [22] that are ideal for short-read sequencing. Together, these enzymes span a broad GC, 33% for NsiI, 66% for PstI and 100% for MspI, thus creating a suitable condition to reduce genome coverage and uniformly sample different genomic regions. Based on in silico digestion analysis, different combination of enzymes with similar cutting site criteria (size and GC content) could be considered to further reduce genome coverage (Additional file 3: Fig. S3). Finally, by focusing on fewer but well-distributed genomic regions, 3D-GBS offers an efficient and cost-effective approach for discovery and

genotyping of SNPs across the genome in species with small to medium-sized genome.

Optimizing the number of reads per sample to maximize multiplexing

Different numbers of reads (i.e. 50K, 100K, 200K and 300K reads) were randomly sampled three times for each accession from the 16-plex 3D-GBS library. For each metric investigated, the coefficient of variation between replicates based on the same number of reads was < 5% [not significantly different (Tukey HSD test p-value > 0.1)]. For this reason, the mean value (across all three replicates) for each metric is reported in Table 2. With increasing the sequencing effort from 50 to 300K reads per sample, the fraction of the genome captured increased from 0.6 to 1%, the number of SNPs increased from 1314 to 4082, and the proportion of missing data decreased from 37 to 20%. Even at the smallest value tested (50K reads/sample), the proportion of missing data was still reasonable and would allow for an accurate imputation [61]. For average minor allele frequency and nucleotide diversity values, equivalent results were provided across the entire range of reads per sample, suggesting that even with a very limited sequencing effort one can perform high-quality genetic diversity analysis.

The distribution of the SNPs on the genetic map was very similar from 100 to 300K reads while, with only 50K reads per sample, large gaps were detected (e.g., ~ 10 cM on Chr01, ~ 80 cM on Chr03, ~ 60 cM on Chr06) and some chromosome extremities were missed (Fig. 2).

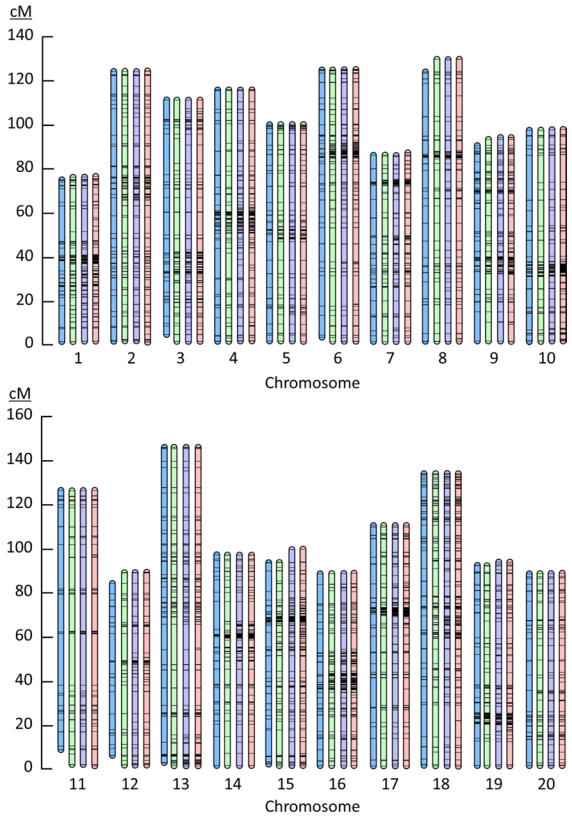

Fig. 2 Comparison between genetic maps based on different number of 3D-GBS reads. Genetic map in blue, green, purple and red were constructed based on 50K, 100K, 200K and 300K reads, respectively

Table 2 Variant calling using different subsets of reads derived from 3D-GBS on 16 soybean samples

Step	Measured parameters	50K reads	100K reads	200K reads	300K reads
Sequencing	Coverage (%)[a]	0.6	0.7	0.9	1
	Mean depth of coverage (X)[b]	2.7	4.1	6.3	8.4
SNP calling	SNP count	1,314	2,299	3,587	4,082
	Proportion of missing data (%)	37.1	29.3	23.3	20.3
	Proportion of heterozygous genotypes (%)	6.1	5.2	5	4.6
	Average minor allele frequency (%)	27.3	27.2	26.2	25.9
	Nucleotide diversity (p per bp)	0.36	0.36	0.35	0.35
Physical map	SNP/Mb	1.4	2.4	3.8	4.3
	Number of gaps > 5 Mb	23	9	5	7
	Number of gaps > 10 Mb	6	1	1	0
Genetic map	SNP/cM[c]	0.6	1.1	1.7	2
	Number of gaps > 10 cM	29	18	12	9
	Number of gaps > 20 cM	2	1	1	1

[a] Total genome fraction captured by the 16 libraries

[b] Average number of read at each sequenced position

[c] Inferred from the closest corresponding SNP on the consensus genetic map [16]

While the density of markers doubled between 100 and 300K reads, the distribution of the SNPs across the genetic map remained very similar with some regions that were denser in SNPs using 300K reads (e.g., ~40 cM on Chr01, ~60 cM on Chr04, ~90 cM on Chr06). This very promising result suggests that one can run 3D-GBS with only 100K reads per sample, a significant reduction in the sequencing cost, to achieve sufficient resolution (~2300 SNPs, 1.1 SNP/cM) to perform GS.

In the case of mapping studies using biparental populations (i.e. QTL mapping), the number of polymorphic marker loci can significantly vary based on the relatedness of parents. To ensure that the proposed number of reads would still offer a sufficient number of markers for biparental QTL mapping, we determined the number and distribution of SNPs between the least and most genetically distant pairs of accessions within this collection. A matrix of pairwise genetic distance among the 16 accessions was produced and identified QS4054 and OAC Bright as the most genetically similar, while QS5008 and QS4067 proved to be the most distant (Additional file 4: Table S1). The number of polymorphic markers using 100K and 300K reads varied from 426 to 669 for the closest lines and from 677 to 1325 for the most distant ones (Table 3). This means that for the closest lines, doubling or tripling the number of reads from 100K reads had only allowed the discovery of 32% and 36% more SNPs, respectively. In contrast, in the most distant lines, doubling or tripling of the number of reads from 100K has doubled the density of markers on the genetic map. Thus, as similar results were obtained between 200 and 300K, 200K reads per sample seems as suitable as 300K reads to perform QTL mapping in a biparental population. This represents a significant gain compared to current studies where ApeKI-based GBS protocol was used with over 1M reads per sample to conduct QTL mapping studies [12, 57].

Compared to other low- (<5X) and very low-depth (0.1–0.5X) sequencing approaches developed to reduce the cost of genotyping [37, 58], 3D-GBS can be considered as extremely low-depth sequencing (~0.01X). Low-depth sequencing methods suffer from genotype uncertainty as a limited amount of sequencing reads are normally used to cover the entire genome. As an example, 1M reads (100 bp) provide a mean depth of coverage of 0.1X of a medium size genome (e.g., soybean; 1 Gb). In 3D-GBS, 200K reads provide a mean depth of coverage of 6X as the complexity reduction allows to focus the sequencing effort on a small proportion of the genome. Furthermore, 3D-GBS offers markers that are better distributed across the genome. Finally, here 3D-GBS libraries were prepared with the least expensive NGS library preparation procedure and its data can also be processed with efficient and user-friendly bioinformatic pipelines [62, 64].

Maximizing multiplexing to minimize the sequencing cost per sample

Thanks to its efficiency and low cost, the GBS approach is commonly used to perform genome-wide genotyping for a large number of species (animal [19], plant [4], insect [13] and microorganism [31]) and different applications (association studies [63] and GS [27, 48]). Nevertheless, the cost associated with high-throughput screening for genome-wide markers remains the most limiting factor in the context of large-scale studies such as GS, genetic fingerprinting and genetic diversity studies. In association studies (GWAS), in general, the denser the catalog of SNPs, the higher the mapping resolution will be. However in contrast, in most of genetic studies (e.g., GS), linkage disequilibrium (LD) is very extensive and a low density SNP catalog is sufficient to capture linkage blocks and perform the analysis [49, 68]. Recent studies based on reducing the total number of SNPs by focusing on a subset with significant marker-trait associations [33, 56] or based on functional annotations [30], suggest that a lower-density catalog could generate prediction accuracies as high or better than dense catalogs (e.g., WGS-based genotyping) [36]. This has been well illustrated for GS in barley, where Abed et al. [1] showed that a catalog of 2K GBS-SNPs provided a very similar prediction accuracy compared to 35K SNPs.

As documented before [62], to reduce the genotyping cost, one can decide to increase the multiplexing level by decreasing the sequencing effort per sample, which can, however, lead to a higher proportion of missing

Table 3 Analysis of SNP density obtained with different number of reads for two hypothetical biparental crosses

Crossing	Closest accessions	Farthest accessions	Closest accessions	Farthest accessions	Closest accessions	Farthest accessions
Reads per sample (K)	100		200		300	
SNP count	426	677	630	1165	669	1325
SNP/1 Mb	0.5	0.7	0.7	1.2	0.7	1.4
SNP/cM	0.21	0.33	0.31	0.57	0.33	0.65

The genetically closest and farthest accessions were QS4054 and OAC Bright and, QS5008 and QS4067, respectively

data that need to be imputed correctly and a non-uniform distribution of SNPs across the genome [63, 64]. Here, using 3D-GBS, we showed that it is possible to produce a lower number of restriction fragments, well and uniformly distributed across the genome, to reduce the number of reads needed to provide sufficient read coverage to call genotypes efficiently. Here, we found that 100K reads is sufficient to conduct GS with 3D-GBS, and that is significantly lower compared to previous studies where GBS has been used (e.g., Qin et al. [48] with ~3.3M reads/sample, Jarquín et al. [27] with ~2.6M reads/sample and Jean et al. [28] with ~1.2M reads/sample). Similarly, we estimated the optimal number of reads per sample for an efficient genotyping of bi- and multi-parent populations. In the context of biparental populations, we estimated that 200K reads/sample is suitable for performing QTL mapping. 3D-GBS allowed a drastic reduction compared to equivalent studies using GBS where a much larger number of reads per sample were used (e.g., Yoon et al. [71] ~3.2M, Heim and Gillman [22] ~2.4M, St-Amour et al. [57] ~1.4M, de Ronne et al. [12] ~1.0M and Vuong et al. [67] ~843K).

To estimate the gain of 3D-GBS over the standard GBS approach, we selected two studies conducted internally, using ApeKI-based GBS protocol and with the lowest number of reads per sample for GS [28] and QTL mapping [12]. In these study cases, based on the optimal number of reads/sample estimated previously, with the same population, experimental design and goal, the application of 3D-GBS for GS and QTL mapping would have led to similar results with a significant reduction in per-sample sequencing cost: ~92% (~1.2M vs 100K reads/sample) and ~86% (~1.4M vs 200K reads/sample), respectively. All without taking into account the miniaturization of sequencing libraries which alone can reduce library preparation costs by 67% [62]. Overall, the combination of recent improvements in miniaturizing GBS library preparation procedure (i.e., NanoGBS [23]) and 3D-GBS provides a unique opportunity to dramatically reduce per-sample genotyping costs.

Conclusion

Recent advances in NGS technologies have enabled the massively parallel processing of hundreds of samples efficiently and cost-effectively, a prerequisite for genetic studies such as QTL mapping and GS. However, it still remains costly in the context of large-scale studies such as GS, as breeding programs typically produce many thousands of selection candidates each year. In the continuous objective of reducing the genotyping cost for scientific research and applied needs, 3D-GBS enables us to maximize the multiplexing capacity needed to achieve the ultra-high throughput that is needed in a wide range of applications and thus decreasing the sequencing cost per sample. While we demonstrated the efficiency of 3D-GBS using soybean samples, this method could easily be used across a wide range of species with small to medium genome size.

Abbreviations

NGS	Next-generation sequencing
HTS	High-throughput sequencing
SNP	Single-nucleotide polymorphism
QTL	Quantitative trait loci
GWAS	Genome-wide association study
MAS	Marker-assisted selection
GS	Genomic selection
WGS	Whole-genome sequencing
RRS	Reduced-representation sequencing
RAD	Restriction-associated DNA
CRoPS	Complexity reduction of polymorphic sequences
GBS	Genotyping-by-sequencing
LD	Linkage disequilibrium

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13007-023-00990-7.

Additional file 1: Figure S1. Distribution of the SNPs derived from NsiI–MspI and PstI–MspI reads across the physical map. The colors of the heatmap correspond to the number of SNPs within 1 Mb windows size.

Additional file 2: Figure S2. Distribution of the SNPs derived from GBS and 3D-GBS libraries across the physical map. The colors of the heatmap correspond to the number of SNPs within 1 Mb windows size.

Additional file 3: Figure S3. Predicted number of cutting site derived from in silico digestion with different restriction enzyme. These enzyme span a GC content of 33% for BglII, BclI and NsiI, 66% for BlnI, BamHI and PstI, and 100% for BfaI, BstUI and MspI. Each color represents a combination of different enzyme.

Additional file 4: Table S1. Heatmap of pairwise genetic distance between the 16 soybean accessions. Green to red reflect low to high genetic distance (smallest and largest values are 0.26 and 0.56, respectively).

Acknowledgements

The authors wish to thank Génome Québec, Genome Canada, the government of Canada, the Ministère de l'Économie et de l'Innovation du Québec, the Canadian Field Crop Research Alliance, Semences Prograin Inc., Sollio Agriculture, Grain Farmers of Ontario, Barley Council of Canada, and Université Laval.

Author contributions

MDR, BB, FB and DT conceptualized the concept of 3D-GBS. MDR and GL conducted the experiments. MDR conducted data analysis. MDR, DT and FB contributed to writing the manuscript. All authors read and approved the final manuscript.

Funding

This work was funded by Genome Canada [#6548] under Genomic Applications Partnership Program (GAPP).

Availability of data and materials

The VCF files generated from the sequencing data and used for the analyzes of this study are on FigShare.com and will be accessible after acceptance of the manuscript.

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 1 November 2022 Accepted: 31 January 2023
Published online: 05 February 2023

References
1. Abed A, Pérez-Rodríguez P, Crossa J, Belzile F. When less can be better: how can we make genomic selection more cost-effective and accurate in barley? Theor Appl Genet. 2018;131:1873–90. https://doi.org/10.1007/s00122-018-3120-8.
2. Begali H. A pipeline for markers selection using restriction site associated DNA sequencing (Radseq). J Appl Bioinform Comput Biol. 2018. https://doi.org/10.4172/2329-9533.1000147.
3. Beissinger TM, Hirsch CN, Sekhon RS, et al. Marker density and read depth for genotyping populations using genotyping-by-sequencing. Genetics. 2013;193:1073–81. https://doi.org/10.1534/genetics.112.147710.
4. Boudhrioua C, Bastien M, Torkamaneh D, Belzile F. Genome-wide association mapping of *Sclerotinia sclerotiorum* resistance in soybean using whole-genome resequencing data. BMC Plant Biol. 2020;20:1–24. https://doi.org/10.1186/s12870-020-02401-8.
5. Bradbury PJ, Zhang Z, Kroon DE, et al. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 2007;23:2633–5. https://doi.org/10.1093/bioinformatics/btm308.
6. Carvalho B, Bengtsson H, Speed TP, Irizarry RA. Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. Biostatistics. 2007;8:485–99. https://doi.org/10.1093/biostatistics/kxl042.
7. Chen Q, Ma Y, Yang Y, et al. Genotyping by genome reducing and sequencing for outbred animals. PLoS ONE. 2013;8: e67500. https://doi.org/10.1371/journal.pone.0067500.
8. da Fonseca RR, Albrechtsen A, Themudo GE, et al. Next-generation biology: sequencing and data analysis approaches for non-model organisms. Mar Genom. 2016;30:3–13. https://doi.org/10.1016/j.margen.2016.04.012.
9. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8. https://doi.org/10.1093/bioinformatics/btr330.
10. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021. https://doi.org/10.1093/gigascience/giab008.
11. Darrier B, Russell J, Milner SG, et al. A comparison of mainstream genotyping platforms for the evaluation and use of barley genetic resources. Front Plant Sci. 2019;10:544. https://doi.org/10.3389/fpls.2019.00544.
12. de Ronne M, Labbé C, Lebreton A, et al. Integrated QTL mapping, gene expression and nucleotide variation analyses to investigate complex quantitative traits: a case study with the soybean–*Phytophthora sojae* interaction. Plant Biotechnol J. 2020;18:1492–4. https://doi.org/10.1111/pbi.13301.
13. Dupuis JR, Brunet BMT, Bird HM, et al. Genome-wide SNPs resolve phylogenetic relationships in the North American spruce budworm (*Choristoneura fumiferana*) species complex. Mol Phylogenet Evol. 2017;111:158–68. https://doi.org/10.1016/j.ympev.2017.04.001.
14. Eaton DAR, Spriggs EL, Park B, Donoghue MJ. Misconceptions on missing data in RAD-seq phylogenetics with a deep-scale example from flowering plants. Syst Biol. 2017;66:399–412. https://doi.org/10.1093/sysbio/syw092.
15. Elshire RJ, Glaubitz JC, Sun Q, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE. 2011;6:1–46. https://doi.org/10.1371/journal.pone.0019379.
16. Fallah M, Jean M, Boucher St-Amour VT, et al. The construction of a high-density consensus genetic map for soybean based on SNP markers derived from genotyping-by-sequencing. Genome. 2022;65:413–25. https://doi.org/10.1139/gen-2021-0054.
17. Fu YB, Peterson GW, Dong Y. Increasing genome sampling and improving SNP genotyping for genotyping-by-sequencing with new combinations of restriction enzymes. G3 Genes Genomes Genet. 2016;6:845–56. https://doi.org/10.1534/g3.115.025775.
18. Ganal MW, Polley A, Graner EM, et al. Large SNP arrays for genotyping in crop plants. J Biosci. 2012;37:821–8. https://doi.org/10.1007/s12038-012-9225-3.
19. Gurgul A, Miksza-Cybulska A, Szmatoła T, et al. Genotyping-by-sequencing performance in selected livestock species. Genomics. 2019;111:186–95. https://doi.org/10.1016/j.ygeno.2018.02.002.
20. Hamblin MT, Rabbi IY. The effects of restriction-enzyme choice on properties of genotyping-by-sequencing libraries: a study in Cassava (*Manihot esculenta*). Crop Sci. 2014;54:2603–8. https://doi.org/10.2135/cropsci2014.02.0160.
21. He J, Zhao X, Laroche A, et al. Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. Front Plant Sci. 2014;5:1–8. https://doi.org/10.3389/fpls.2014.00484.
22. Heim CB, Gillman JD. Genotyping-by-sequencing-based investigation of the genetic architecture responsible for a ~sevenfold increase in soybean seed stearic acid. G3 Genes Genomes Genet. 2017;7:299–308. https://doi.org/10.1534/g3.116.035741.
23. Hirsch CD, Evans J, Buell CR, Hirsch CN. Reduced representation approaches to interrogate genome diversity in large repetitive plant genomes. Brief Funct Genom Proteom. 2014;13:257–67. https://doi.org/10.1093/bfgp/elt051.
24. Hodgkinson A, Eyre-Walker A. Variation in the mutation rate across mammalian genomes. Nat Rev Genet. 2011;12:756–66. https://doi.org/10.1038/nrg3098.
25. Huang H, Lacey Knowles L. Unforeseen consequences of excluding missing data from next-generation sequences: simulation study of rad sequences. Syst Biol. 2016;65:357–65. https://doi.org/10.1093/sysbio/syu046.
26. Hyten DL, Choi IY, Song Q, et al. A high density integrated genetic linkage map of soybean and the development of a 1536 universal soy linkage panel for quantitative trait locus mapping. Crop Sci. 2010;50:960–8. https://doi.org/10.2135/cropsci2009.06.0360.
27. Jarquín D, Kocak K, Posadas L, et al. Genotyping by sequencing for genomic prediction in a soybean breeding population. BMC Genom. 2014;15:1–10. https://doi.org/10.1186/1471-2164-15-740.
28. Jean M, Cober E, O'Donoughue L, et al. Improvement of key agronomical traits in soybean through genomic prediction of superior crosses. Crop Sci. 2021;61:3908–18. https://doi.org/10.1002/csc2.20583.
29. Karimi K, Wuitchik DM, Oldach MJ, Vize PD. Distinguishing species using GC contents in mixed DNA or RNA sequences. Evol Bioinform. 2018. https://doi.org/10.1177/1176934318788866.
30. Koufariotis LT, Chen YPP, Stothard P, Hayes BJ. Variance explained by whole genome sequence variants in coding and regulatory genome annotations for six dairy traits. BMC Genom. 2018. https://doi.org/10.1186/s12864-018-4617-x.
31. Leboldus JM, Kinzer K, Richards J, et al. Genotype-by-sequencing of the plant-pathogenic fungi *Pyrenophora teres* and *Sphaerulina musiva* utilizing ion torrent sequence technology. Mol Plant Pathol. 2015;16:623–32. https://doi.org/10.1111/mpp.12214.
32. Li H. seqtk: Toolkit for processing sequences in FASTA/Q formats. In: GitHub 767. 2012. https://github.com/lh3/seqtk/. Accessed 17 Aug 2022.
33. Li X, Guo T, Mu Q, et al. Genomic and environmental determinants and their interplay underlying phenotypic plasticity. Proc Natl Acad Sci USA. 2018;115:6679–84. https://doi.org/10.1073/pnas.1718326115.
34. Li XQ. Somatic genome variation in animals, plants, and microorganisms. Hoboken: Wiley; 2016. p. 1–419. https://doi.org/10.1002/9781118647110.
35. Li XQ. Genome variation in archaeans, bacteria, and asexually reproducing eukaryotes. In: Somatic genome variation in animals, plants, and microorganisms. Hoboken: Wiley; 2016. p. 253–66. https://doi.org/10.1002/9781118647110.ch10.

36. Li Y, Ruperao P, Batley J, et al. Genomic prediction of preliminary yield trials in chickpea: effect of functional annotation of SNPs and environment. Plant Genome. 2022;15: e20166. https://doi.org/10.1002/tpg2.20166.
37. Lou RN, Jacobs A, Wilder AP, Therkildsen NO. A beginner's guide to low-coverage whole genome sequencing for population genomics. Mol Ecol. 2021;30:5966–93. https://doi.org/10.1111/MEC.16077.
38. Luca F, Hudson RR, Witonsky DB, Di Rienzo A. A reduced representation approach to population genetic analyses and applications to human evolution. Genome Res. 2011;21:1087–98. https://doi.org/10.1101/gr.119792.110.
39. Melamed-Bessudo C, Shilo S, Levy AA. Meiotic recombination and genome evolution in plants. Curr Opin Plant Biol. 2016;30:82–7. https://doi.org/10.1016/j.pbi.2016.02.003.
40. Meng L, Li H, Zhang L, Wang J. QTL IciMapping: integrated software for genetic linkage map construction and quantitative trait locus mapping in biparental populations. Crop J. 2015;3:269–83. https://doi.org/10.1016/j.cj.2015.01.001.
41. Moragues M, Comadran J, Waugh R, et al. Effects of ascertainment bias and marker number on estimations of barley diversity from high-throughput SNP genotype data. Theor Appl Genet. 2010;120:1525–34. https://doi.org/10.1007/s00122-010-1273-1.
42. Morales KY, Singh N, Perez FA, et al. An improved 7K SNP array, the C7AIR, provides a wealth of validated SNP markers for rice breeding and genetics studies. PLoS ONE. 2020;15: e0232479. https://doi.org/10.1371/journal.pone.0232479.
43. Narum SR, Buerkle CA, Davey JW, et al. Genotyping-by-sequencing in ecological and conservation genomics. Mol Ecol. 2013;22:2841–7. https://doi.org/10.1111/mec.12350.
44. Nishida H. Genome DNA sequence variation, evolution, and function in bacteria and archaea. Curr Issues Mol Biol. 2008;15:19–24.
45. Pértille F, Guerrero-Bosagna C, Da SVH, et al. High-throughput and cost-effective chicken genotyping using next-generation sequencing. Sci Rep. 2016;6:1–12. https://doi.org/10.1038/srep26929.
46. Poland JA, Brown PJ, Sorrells ME, Jannink JL. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS ONE. 2012. https://doi.org/10.1371/journal.pone.0032253.
47. Poland JA, Rife TW. Genotyping-by-sequencing for plant breeding and genetics. Plant Genome. 2012. https://doi.org/10.3835/plantgenome2012.05.0005.
48. Qin J, Wang F, Zhao Q, et al. Identification of candidate genes and genomic selection for seed protein in soybean breeding pipeline. Front Plant Sci. 2022. https://doi.org/10.3389/fpls.2022.882732.
49. Quiroz M, Kohn R, Villani M, Tran MN. Speeding up MCMC by efficient data subsampling. J Am Stat Assoc. 2019;114:831–43. https://doi.org/10.1080/01621459.2018.1448827.
50. Rasheed A, Hao Y, Xia X, et al. Crop breeding chips and genotyping platforms: progress, challenges, and perspectives. Mol Plant. 2017;10:1047–64. https://doi.org/10.1016/j.molp.2017.06.008.
51. Schmutz J, Cannon SB, Schlueter J, et al. Genome sequence of the palaeopolyploid soybean. Nature. 2010;463:178–83. https://doi.org/10.1038/nature08670.
52. Sonah H, Bastien M, Iquira E, et al. An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of snp discovery and genotyping. PLoS ONE. 2013;8:1–9. https://doi.org/10.1371/journal.pone.0054603.
53. Sonah H, O'Donoughue L, Cober E, et al. Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. Plant Biotechnol J. 2015;13:211–21. https://doi.org/10.1111/pbi.12249.
54. Song K, Li L, Zhang G. Coverage recommendation for genotyping analysis of highly heterologous species using next-generation sequencing technology. Sci Rep. 2016;61(6):1–7. https://doi.org/10.1038/srep35736.
55. Song Q, Yan L, Quigley C, et al. Soybean BARCSoySNP6K: an assay for soybean genetics and breeding research. Plant J. 2020;104:800–11. https://doi.org/10.1111/tpj.14960.
56. Spindel JE, Begum H, Akdemir D, et al. Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. Heredity (Edinb). 2016;116:395–408. https://doi.org/10.1038/hdy.2015.113.
57. St-Amour VTB, Mimee B, Torkamaneh D, et al. Characterizing resistance to soybean cyst nematode in PI 494182, an early maturing soybean accession. Crop Sci. 2020;60:2053–69. https://doi.org/10.1002/csc2.20162.
58. Tachmazidou I, Süveges D, Min JL, et al. Whole-genome sequencing coupled to imputation discovers genetic signals for anthropometric traits. Am J Hum Genet. 2017;100:865–84. https://doi.org/10.1016/J.AJHG.2017.04.014.
59. Thomson MJ. High-throughput SNP genotyping to accelerate crop improvement. Plant Breed Biotechnol. 2014;2:195–212. https://doi.org/10.9787/pbb.2014.2.3.195.
60. Torkamaneh D, Belzile F. Scanning and filling: ultra-dense SNP genotyping combining genotyping-by-sequencing, SNP array and whole-genome resequencing data. PLoS ONE. 2015;10: e0131533. https://doi.org/10.1371/journal.pone.0131533.
61. Torkamaneh D, Boyle B, Belzile F. Efficient genome-wide genotyping strategies and data integration in crop plants. Theor Appl Genet. 2018;131:499–511. https://doi.org/10.1007/s00122-018-3056-z.
62. Torkamaneh D, Boyle B, St-Cyr J, et al. NanoGBS: a miniaturized procedure for GBS library preparation. Front Genet. 2020;11:1–8. https://doi.org/10.3389/fgene.2020.00067.
63. Torkamaneh D, Chalifour FP, Beauchamp CJ, et al. Genome-wide association analyses reveal the genetic basis of biomass accumulation under symbiotic nitrogen fixation in African soybean. Theor Appl Genet. 2020;133:665–76. https://doi.org/10.1007/s00122-019-03499-7.
64. Torkamaneh D, Laroche J, Belzile F. Fast-gbs v2.0: an analysis toolkit for genotyping-by-sequencing data. Genome. 2020;63:577–81. https://doi.org/10.1139/gen-2020-0077.
65. Torkamaneh D, Laroche J, Boyle B, et al. A bumper crop of SNPs in soybean through high-density genotyping-by-sequencing (HD-GBS). Plant Biotechnol J. 2021;19:860–2. https://doi.org/10.1111/pbi.13551.
66. Torkamaneh D, Laroche J, Boyle B, Belzile F. DepthFinder: a tool to determine the optimal read depth for reduced-representation sequencing. Bioinformatics. 2020;36:26–32. https://doi.org/10.1093/bioinformatics/btz473.
67. Vuong TD, Sonah H, Patil G, et al. Identification of genomic loci conferring broad-spectrum resistance to multiple nematode species in exotic soybean accession PI 567305. Theor Appl Genet. 2021;134:3379–95. https://doi.org/10.1007/s00122-021-03903-1.
68. Waldmann P, Hallander J, Hoti F, Sillanpää MJ. Efficient Markov chain Monte Carlo implementation of Bayesian analysis of additive and dominance genetic variances in noninbred pedigrees. Genetics. 2008;179:1101–12. https://doi.org/10.1534/genetics.107.084160.
69. Wang Y, Cao X, Zhao Y, et al. Optimized double-digest genotyping by sequencing (ddGBS) method with highdensity SNP markers and high genotyping accuracy for chickens. PLoS ONE. 2017. https://doi.org/10.1371/journal.pone.0179073.
70. Yin L, Zhang H, Tang Z, et al. rMVP: a memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. Genom Proteom Bioinform. 2021;19:619–28. https://doi.org/10.1016/j.gpb.2020.10.007.
71. Yoon MY, Kim MY, Ha J, et al. QTL analysis of resistance to high-intensity UV-B irradiation in soybean (*Glycine max* [L.] merr.). Int J Mol Sci. 2019;20:3287. https://doi.org/10.3390/ijms20133287.
72. Zhu WY, Huang L, Chen L, et al. A high-density genetic linkage map for cucumber (*Cucumis sativus* L.): based on specific length amplified fragment (SLAF) sequencing and QTL analysis of fruit traits in cucumber. Front Plant Sci. 2016;7:437. https://doi.org/10.3389/fpls.2016.00437.

Publisher's Note