doi 10.18699/bgrs2024-06-55

Cost-effective estimation of optimal number of reads in GBS sequencing

Zamalutdinov A.*, Boldyrev S., Ben C., Gentzbittel L.
Skolkovo Institute of Science and Technology, Moscow, Russia
** Alexey.Zamalutdinov@skoltech.ru*

Key words: GBS; read number; soybean

Motivation and Aim: Genotype-by-sequencing (GBS) is a cost-effective approach for large-scale genotyping that has been widely employed for different species, particularly those with large genomes. In order to reduce genome complexity, GBS uses digestion of the genome by one or more restriction enzymes [1]. They are different in their properties and affect the size, number, and genome localization of the fragments. That is why, the choices of the restriction enzymes for library preparation are a crucial step in GBS. Several protocols and different enzyme combinations were evaluated in order to reduce costs and increase accuracy [2, 3]. However, quality and genotyping cost also depend on number of reads used [4]. Here, we addressed this question and suggested cost-effective empirical approach to define range for optimal number of reads.
Methods and Algorithms: 12 soybean accessions were genotyped using GBS. HindIII-NlaIII enzyme combination was used for library preparation. The resulting library was sequenced using the Illumina HiSeq4000 in a 2*150bp paired-end mode with 111 million paired-end reads. In order to evaluate the dependency between number of reads and number of SNPs, random sampling of reads was performed in triplicate, from 1 million reads per library up to 111 million. Demultiplication and SNP calling were performed using stacks with default settings [5]. Trimming and filtering were conducted using the bbduk tool [6] with following parameters: “k=31 ref=artifacts,phix,adapters,lambda,pjet,mtst,kapa ordered cardinality qtrim=rl trimq=20 maq=25 minlen=50 tbo mink=11 ktrim=r minlen=50”. The mapping was performed using BWA-MEM [7]. The vcf files were subjected to filtering using Bcftools 1.18 [8], with genotypes with depth < 3 set to missing. Subsequently, biallelic single-nucleotide polymorphisms (SNPs) with an average depth per genotype of greater than five, a minor allele frequency (MAF) of 0.05, and a fraction of missing data of less than 0.4 were used for further analysis. In addition, we estimated the fraction of genome regions covered by average number of reads per accession in each library. Fraction of covered regions was calculated using Bedtools 2.30.0 [9].
Results: Based on real sequencing data obtained through the use of HindIII-NlaIII combination, libraries of different sizes were obtained by random read selection. It is expected that more reads are used, more SNPs are obtained. However, this dependency is not linear and three regions on the curve can be identified (Fig. 1A). Low number of reads will output extremely low number of SNPs. In the middle linear dependency between reads and SNPs count is observed. High number of reads will generate similar number of SNPs. So, here it is clear how many SNPs will be called using this number of reads. However, it requires preparation of small library of several accessions with high sequencing depth.

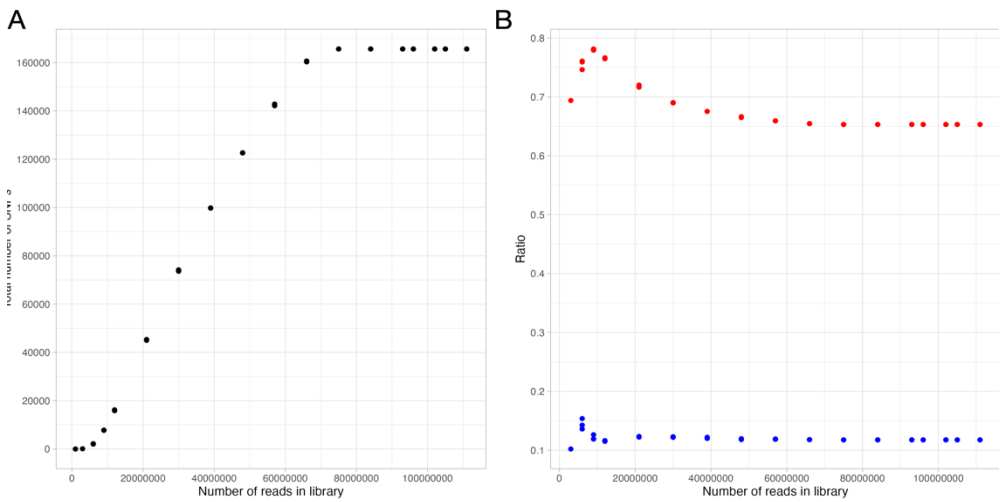

Fig. 1. A – Dependence of the total number of SNPs on the number of reads for library sequencing. Three regions of interest can be identified on the curve. From 1 to 6 million reads there is no increase in the number of SNPs. From 9 to 66 million – linear dependence, more reads are used, more SNPs are obtained. From 75 to 111 there is no increase in the number of SNPs, only an increase in depth and accuracy. B – Dependency of the fraction of SNPs in repetitive regions and in genes on the number of reads for library sequencing. The proportion of SNPs in repeats is depicted in red, while the proportion of SNPs in genes is shown in blue. The ratio of SNPs in genes is stable from 9 million reads per library. In contrast, the proportion of SNPs in repeats decreases slightly as the number of reads used increases

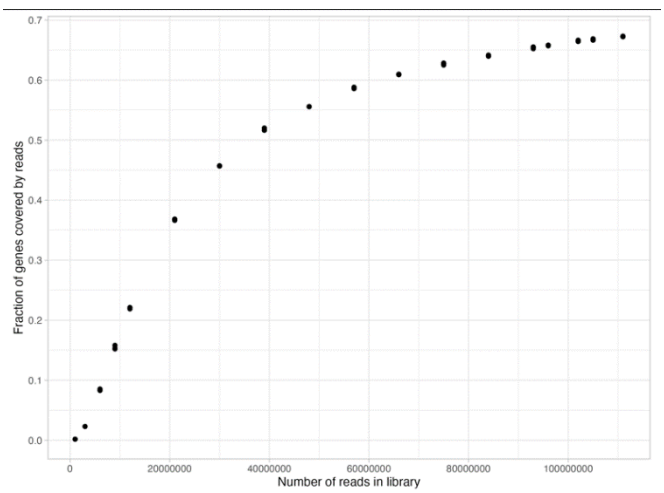

Fig. 2. Dependency of fraction of genes covered on number of reads for library sequencing. First additions of read number rapidly increase the fraction of covered genes. However, the maximum possible number of covered genes is predefined by library design, or enzyme choice. So, at high numbers of used reads fraction of covered genes increases slowly

We also examined the localization of SNPs in genome and find that fraction of SNPs in genes is quite stable in different read numbers. The fraction of SNPs in repeats is decrease slowly and reach plateau at 75 million reads, together with total number of reads (Fig. 1B).

Nevertheless, SNP calling requires relatively high coverage for each sample and depends on collection diversity. We evaluated which information can be used to estimate required number of reads without sequencing of libraries, but using only one sample based on genome features. Reads were sampled according to average number of reads per accession in examined libraries and the fraction of genes covered by different numbers of reads was estimated. Figure 2 demonstrates that it is again non-linear dependency on read number. At small number of reads rapid growth of covered genes can be observed. It corresponds to the first region form Figure 1A. However, saturation of covered genes is rather smooth and no sharp bend can be observed. Huge number of reads will not significantly increase the fraction of covered genes as maximum number of possibly covered genes is predefined by enzyme choice.

Conclusion: GBS genotyping is well-established approach and widely used for many plant species. However, the results depend on many factors, such as enzyme choice, library preparation, and read number. Here, we demonstrated that read number affects final number of SNPs as well as their localization in soybean using a real dataset with the HindIII-NlaIII enzyme combination. However, estimating the optimal number of reads by genotyping the test GBS library is either costly when high coverage is utilized or inaccurate when low numbers of reads are used. Alternative approach involving high coverage genotyping of the one sample without SNP calling was suggested. Such indirect approach allows to define the range of read numbers where costs will be adequate to output. We consider that this approach provides easy way to estimate adequate read number per library and can be extended to other species and enzyme combinations.

References

1. Elshire R.J. et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*. 2011;6(5):e19379. doi 10.1371/journal.pone.0019379
2. de Ronne M., Légaré G., Belzile F., Boyle B., Torkamaneh D. 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. *Plant Methods*. 2023;19(1):13. doi 10.1186/s13007-023-00990-7
3. Hamblin M.T., Rabbi I.Y. The effects of restriction-enzyme choice on properties of genotyping-by-sequencing libraries: A study in Cassava (*Manihot esculenta*). *Crop Sci.* 2014;54(6):2603-2608. doi 10.2135/cropsci2014.02.0160
4. Beissinger T.M. et al. Marker density and read depth for genotyping populations using genotyping-by-sequencing. *Genetics*. 2013;193(4):1073-1081. doi 10.1534/genetics.112.147710
5. Catchen J., Hohenlohe P.A., Bassham S., Amores A., Cresko W.A. Stacks: an analysis tool set for population genomics. *Mol Ecol*. 2013;22(11):3124-3140. doi 10.1111/mec.12354
6. Bushnell B., Rood J., Singer E. BBMerge – Accurate paired shotgun read merging via overlap. *PLoS One.* 2017;12(10):e0185056. doi 10.1371/journal.pone.0185056
7. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*. 2013. doi 10.48550/arXiv.1303.3997
8. Danecek P. et al. Twelve years of SAMtools and BCFtools. *Gigascience*. 2021;10(2):giab008. doi 10.1093/gigascience/giab008
9. Quinlan A.R., Hall I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841-842. doi 10.1093/bioinformatics/btq033