Optimisation of GBS protocols for efficient genotyping of forage species

Bernadette Julier, Sébastien Blugeon, Sabrina Delaunay, Gaëtan Mappa, Tom Ruttink, Marie Pégard, Philippe Barre

Optimisation of GBS protocols for efficient genotyping of forage species

Bernadette Julier[1], Sébastien Blugeon[1], Sabrina Delaunay[1], Gaëtan Mappa[1], Tom Ruttink[2], Marie Pégard[1] and Philippe Barre[1]

[1] INRAE P3F, 86600 Lusignan, France
[2] ILVO Plant Science Unit, 9090 Melle, Belgium
bernadette.julier@inrae.fr

Abstract. GBS is an efficient genotyping tool for heterozygous and polyploid species, yet the fraction of missing data, inherent to the method, may be detrimental to genetic analyses. We tested several restriction enzymes to choose the ones that were likely to provide the maximum number of loci with little missing data considering a certain sequencing effort. For lucerne and cocksfoot, the enzymes that met the target of about 10 000 loci were the combination of *Pst*I-*Mse*I for lucerne and *Pst*I for cocksfoot. 1 066 lucerne accessions were genotyped with *Pst*I-*Mse*I, and more than 200 000 SNP with less than 5% missing data were obtained. Based on these results, we recommend performing such a test to optimize GBS genotyping efforts in forage species.

Keywords: cocksfoot, heterozygous, lucerne, polyploidy, SNP.

1 Introduction

GBS (Genotyping-by-Sequencing) previously showed to be a cost-efficient and reliable method to genotype several forage species [1; 2]. GBS is based on a reduction of genome complexity by the use of restriction enzymes that delimit genome portions (loci) which are amplified and sequenced [3]. In some cases, when the number of loci is high compared to the sequencing effort, genotyping matrices are generated with missing data for many loci. The optimization of GBS protocols is thus required to get the highest number of loci with the lowest fraction of missing data. We tested several restriction enzymes and evaluated the number of loci obtained in two species, lucerne (*Medicago sativa*) and cocksfoot (*Dactylis glomerata*). For lucerne, we also determined the number of SNP and missing data obtained on 1 066 accessions.

2 Material and methods

For a lucerne individual genotype, 16 libraries were prepared with different restriction enzymes: *Ape*KI, *Pst*I, *Eco*RI-*Bam*HI, *Eco*RI-*Hind*III, *Eco*RI-*Mse*I, *Eco*RI-*Msp*I, *Eco*RI-*Ape*KI, *Pst*I-*Ape*KI, *Pst*I-*Mse*I, *Pst*I-*Msp*I, *Sbf*I-*Eco*RI, *Sbf*I-*Ape*KI, *Sbf*I-*Bam*HI, *Sbf*I-*Hind*III, *Sbf*I-*Mse*I, *Sbf*I-*Msp*I. Based on the distribution of the size of

the fragments of each library, six were selected for sequencing (Fig. 1). The reads were mapped on a lucerne reference genome [4] and the number of loci that contained at least 10 reads was counted. Computational subsampling of data was performed to simulate varying numbers of reads per library. The number of loci was plotted as a function of library size. The objective was to select an enzyme or an enzyme pair that gave a saturation curve with a clear plateau of observed loci. From this curve, the number of loci at the plateau (here expected to be at least 10 000 loci) was used to determine the required number of reads per library. The same procedure was applied on cocksfoot with two single-enzymes and six enzyme pairs (Fig. 2). The reads were mapped on a draft reference genome sequence of cocksfoot (P. Barre & S. Buchman, unpubl.).

For lucerne, the most promising enzyme pair was used to genotype 1 066 accessions, each represented by a pool of 100 individuals. At this population level, the SNP were retained when the read depth reached at least 30.

3 Results

For lucerne (Fig. 1), the number of loci as a function of the number of forward reads showed a saturating curve shape for two enzyme pairs only (*Pst*I-*Ape*KI and *Pst*I-*Mse*I). With *Pst*I-*Mse*I, 12 600 loci (with at least 10 reads) were obtained with 10 million forward reads, which is economically affordable in a scientific project. This pair of enzymes was chosen for GBS genotyping of 1 066 lucerne accessions. For cocksfoot, *Pst*I revealed almost 10 000 loci (with at least 10 reads) (Fig. 2) with 8 million total reads per individual.

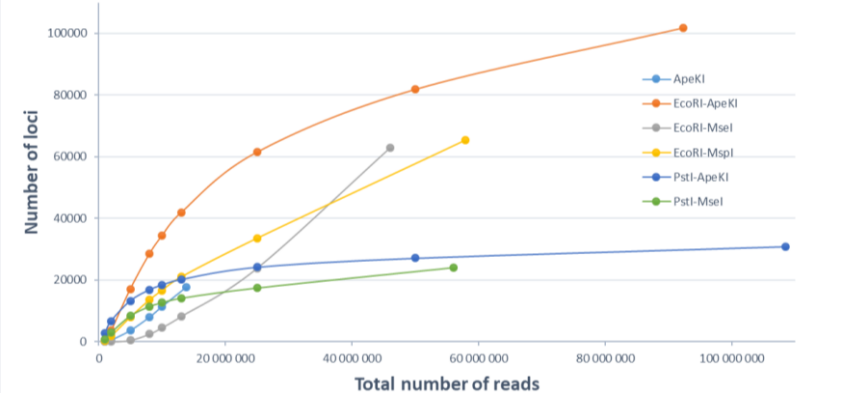

Fig. 1. Number of loci as a function of the number of reads obtained on one lucerne genotype, for *Ape*KI single-digest GBS and five combinations of two restriction enzymes.

For the 1 066 lucerne accessions, the mean number of forward reads obtained with *Pst*I-*Mse*I enzymes was 7.5 million. After trimming and filtering, reads were mapped on the reference genome [5], and we obtained 477 221 SNP for 1 061 accessions (5 accessions failed). With the objective to have less than 5% of missing data per SNP, a

total of 228 568 SNP in 31 743 loci were generated. This dataset was sorted again to isolate a subset of SNP without missing data, and 118 421 SNP were then retained.

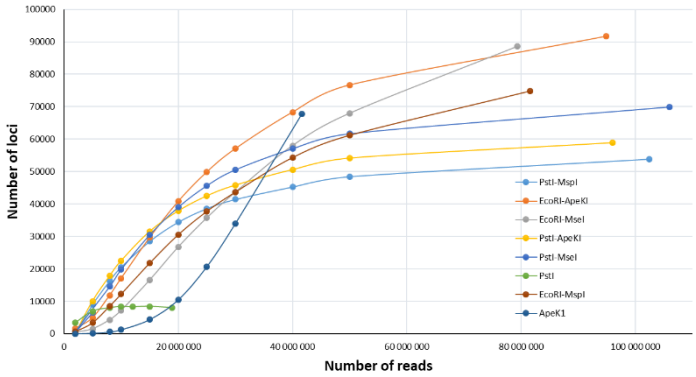

Fig. 2. Number of loci as a function of the number of reads obtained on one cocksfoot genotype, for *Pst*I, *Ape*KI single-digest GBS and six combinations of two restriction enzymes.

This study shows that a test of restriction enzymes is a way to optimize GBS protocols. The combination of *Pst*I-*Mse*I enzymes for lucerne and *Pst*I for cocksfoot were the most optimal to obtain a high number of markers with low proportion of missing data. A high-quality dataset comprising 228 568 SNP with less than 5% missing data on 1061 lucerne accessions has been obtained.

Acknowledgements

This study was supported by EUCLEG project (UE, H2020 n°727312) and Dactysec project (French Ministry of Agriculture Casdar n°C-2016-1).

References

1. Julier, B., Barre, P., Lambroni, P., Delaunay, S., Thomasset, M., Lafaillette, F., Gensollen, V.: Use of GBS markers to distinguish among lucerne varieties, with comparison to morphological traits. Mol. Breed. 38, 133 (2018).
2. Byrne, S., Czaban, A., Studer, B., Panitz, F., Bendixen, C., Asp, T.: Genome Wide Allele Frequency Fingerprints (GWAFFs) of populations via genotyping by sequencing. Plos One 8 (2013).
3. Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., Mitchell, S.E.: A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. Plos One 6 (2011).
4. Carrère, S., Gouzy, J., Debellé, F., Marande, W., Julier, B., Barre, P.: Medicago sativa cv. Mercedes genome sequence Available: https://hal.inrae.fr/hal-02993163v2/document (2020).
5. Chen, H.T., Zeng, Y., Yang, Y.Z., et al.: Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. Nature Communications 11 (2020).