Technical note
AN-161104.01

Efficient high-throughput SNP discovery and genotyping using normalised Genotyping-by-Sequencing (nGBS)

Samuel Arvidsson, Berthold Fartmann, Silke Winkler and Wolfgang Zimmermann
LGC, Ostendstraße 25, 12459 Berlin, Germany

Abstract

Large-scale genotyping programs require efficient tools for high-density SNP discovery. The commonly used technique involving whole genome shotgun sequencing of parental genotypes is often prohibitively expensive and resource intensive. Here we present an optimised, self-tuning Genotyping-by-Sequencing (GBS) method, called normalised GBS (nGBS), which efficiently reduces the genome complexity of any species to a few hundred thousand loci across the complete genome. No prior knowledge of the genome is required and repetitive regions are avoided. The selected loci are reliably sequenced across large sample batches using next generation sequencing. Simultaneous SNP discovery and genotyping is carried out for between forty-eight and several thousand samples without prior assay design effort and bias.

Background

The majority of most recent projects in industry and academia involving large scale breeding programs, trait mapping, germplasm characterisation and quality control now rely on the tools offered by molecular genetic markers (He et al. 2014). Traditionally, microsatellite markers have been used in such applications but, in the last decade, single nucleotide polymorphism (SNP) markers have become increasingly important. SNP markers are particularly efficient due to their specificity (as they target a single base pair difference) and to the availability of simple and robust genotyping methods.

Development of SNP markers is, however, relatively expensive and time-consuming, especially when very dense genetic maps are necessary. Although high-throughput sequencing costs are decreasing rapidly, the effort required for sequencing and data analysis is still prohibitive. This is a particular issue when performing whole genome shotgun (WGS) sequencing of large genomes, where many genotypes need to be analysed to fully cover the genetic diversity within the species. Furthermore the data analysis involved (de novo assembly, variant calling) in comparing the different varieties and developing accurate SNP PCR assays can be very time consuming and require expensive computational resources.

Traditional Genotyping-by-Sequencing

Instead of analysing the complete genome sequence of each sample, one can focus on reducing the complexity of the genome sequence to certain regions or features which can be cost-efficiently sequenced for a large number of samples. One way to achieve this, with little or no prior knowledge of the genome in question, is to use restriction enzyme digestion of the genomic DNA. This involves the use of one or two restriction enzymes, and results in a reproducible set of fragments for each sample. These non-randomly selected DNA pieces represent genomic loci, and typically cover between several tens of thousands and a few hundred thousand loci in the genome. Sample fragments are then subjected to highly multiplexed sequencing, producing up to several million sequence reads for each sample on standard next generation sequencing instruments. The resulting sequence data is analysed in order to determine the polymorphic loci (typically every 10th to 1,000th of the covered loci) and subsequently the genotype of each polymorphic locus for each sample.

There are several protocols for this technique, with the following being the most common in the literature:

- RAD-Seq (Restriction Associated DNA sequencing) – single enzyme digestion combined with random fragmentation, sequencing adapter ligation (Baird et al. 2008)
- GBS – single enzyme digestion, size selection and inline barcode addition with custom PCR protocol; typically using the ApeKI enzyme (Elshire et al. 2011)
- ddRAD-Seq – double enzyme digestion, size selection and inline barcode addition with custom PCR protocol; often PstI-MspI (Poland et al. 2012)

The number of genomic loci covered per sample for a given method can be theoretically estimated if the

average guanosine-cytosine (GC) content is known for the genome. However, there are other technical factors affecting the final restriction fragment profile, such as methylation pattern and fragment size selection during library preparation as well as clustering bias on the sequencing flow cell. Sequence diversity across the samples also influences the restriction fragment pattern; diversity is high in ecological studies of wild populations and low in inbred populations. This latter sequence diversity factor also influences the main outcome of the experiment, which is the number of reliably covered polymorphic loci across a majority of the samples.

The optimised normalised Genotyping-by-Sequencing method

When dealing with a new species, it is usually necessary to perform extensive experimental trials involving several enzyme combinations and various protocol modifications. The restriction enzyme(s) used are usually chosen to be sensitive to the most abundant types of methylation in the target genome and thus typically avoid repetitive regions (mainly retrotransposons). Some high copy number DNA in the sample (e.g. mitochondrial DNA, chloroplasts and ribosomal repeats) are not methylated, and can result in a reduced number of sequencing reads from the lower copy number regions.

In order to simplify the protocol development stage and to reduce the sequencing required due to the presence of non-methylated high copy number DNA, LGC has developed a self-tuning protocol applicable to any kind of genome, regardless of size, methylation pattern, and type and abundance of repetitive elements. This method is called normalised Genotyping-by-Sequencing (nGBS) and an overview is given in Figure 1. The normalisation step is adapted from hybridisation kinetics resulting in reduction of abundant fragments. nGBS uses the MslI restriction enzyme to produce blunt end fragments and an enzyme treatment in a subsequent normalisation step. Although MslI digestion is inhibited by overlapping CNG and CHH methylation (common methylation types in plants), the number of fragments produced is generally too high for cost-efficient sequencing. Therefore, the normalisation step is important to reduce any remaining high copy number fragments.

Based on prior genome knowledge and the availability of GBS data for comparison, a suitable sequencing protocol is chosen. This can range from single-sided 75 bp sequencing (at minimal cost per sample) for species with a genome of low complexity and known sequence, to paired 150 bp sequencing which is the protocol of choice when working with species for which no reference sequence is available. Up to 384 samples are multiplexed and sequenced on an Illumina NextSeq flow cell; this offers the flexibility of generating a few hundred thousand reads to a few million read pairs per sample. Generally an average depth of 1.5 million read pairs per sample is sufficient for a well-tuned protocol.

The appropriate data analysis is again dependent on prior knowledge of the genome and availability of previous data:

- With an available reference sequence or a well-saturated reference of GBS tags, the newly acquired sequence data is aligned, observed alleles called, and genotypes determined for each sample according to the read counts seen for each allele.
- When no reference sequence, GBS tags or fragment cluster collection is available the newly acquired sequence data is clustered to generate a reference. For 75 bp protocols, this is performed using all unique sequence tags from the first 64 bases as input. For paired 150 bp protocols, where the typical insert sizes are below 270 bp, all unique sequences from paired end overlap consensus fragments are used. The polymorphic tag clusters are then determined and genotypes called for each sample according to the read counts seen for each allele.

Either data analysis pathway results in a genotype table with the samples as columns, and the SNP locus as rows. Filters are then applied to remove low-quality loci, typically resulting in a set of 5,000 - 50,000 polymorphic loci with high-quality genotype calls in ≥ 75% of the samples (dependent on population diversity and sequence depth per sample). The entire data analysis process can be carried out with publically available GBS analysis packages e.g., Stacks, TASSEL or PyRAD, or an LGC-developed pipeline optimised for nGBS data.

Low sampling rate in sequencing of the libraries or the presence of existing polymorphisms in the restriction sites (presence/absence variants) result in missing genotype data. As certain downstream applications (e.g. GWAS, QTL) typically expect complete datasets, missing data imputation (e.g. Swarts et al. 2014) can then be used to correct and complete the dataset, a process that is facilitated by previous knowledge of the organism (published tag sets) and the population studied.

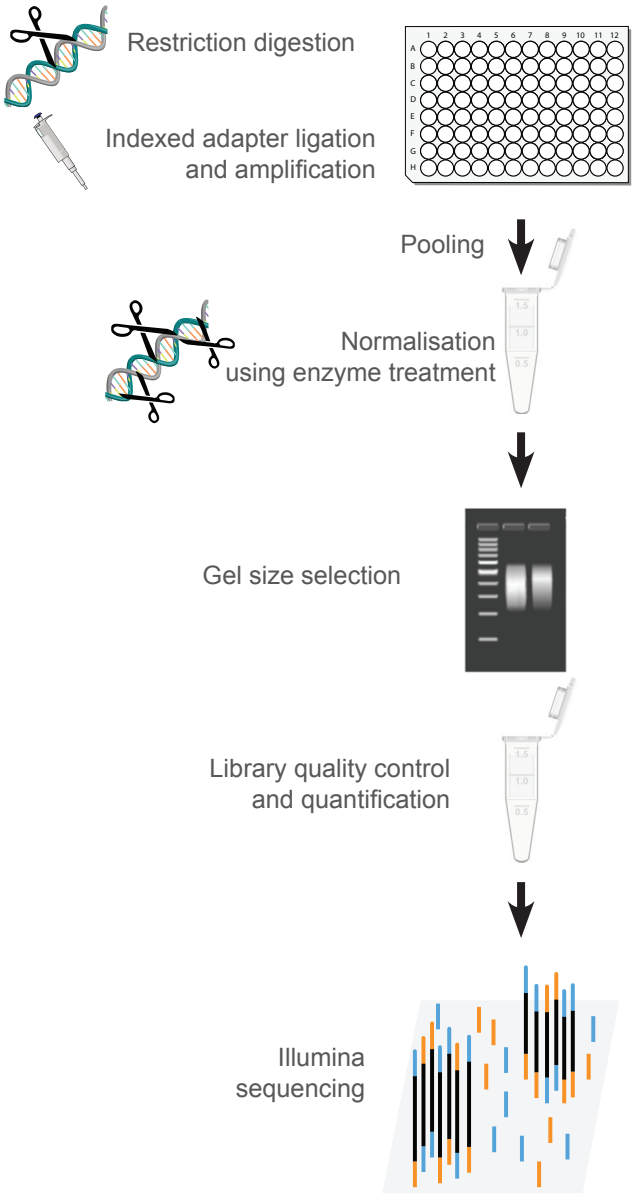

Figure 1. The nGBS protocol

Conclusions

LGC has established the improved nGBS protocol as a very efficient tool in SNP discovery for marker development. The identified polymorphic locus sequences can be used directly to develop PCR-based SNP assays. The method can also be used directly for routine high-throughput genotyping, as the cost per sample is lower than SNP microarray genotyping ("SNP chips") with an optimised nGBS protocol. Additionally, the experimental design is faster with a lower associated cost and there is no design bias, which is especially important when working in species with very high polymorphism rates e.g. maize (see Elshire et al. 2011). All these features make the nGBS method an ideal solution for rapid and high quality SNP discovery and genotyping in any organism, with a reasonable price tag.

References

- He J, Zhao X, Laroche A, Lu Z-X, Liu H and Li Z (2014) Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. Front. Plant Sci. 5:484. http://dx.doi.org/10.3389/fpls.2014.00484
- Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, et al. (2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. PLoS ONE 3(10): e3376. http://dx.doi.org/10.1371/journal.pone.0003376
- Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE 6(5): e19379. http://dx.doi.org/10.1371/journal.pone.0019379
- Poland JA, Brown PJ, Sorrells ME, Jannink J-L (2012) Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. PLoS ONE 7(2): e32253. http://dx.doi.org/10.1371/journal.pone.0032253
- Swarts, K., H. Li, J. A. Romero Navarro, D. An, M. C. Romay, S. Hearne, C. Acharya, J. C. Glaubitz, S. Mitchell, R. J. Elshire, E. S. Buckler, and P. J. Bradbury. 2014. Novel Methods to Optimize Genotypic Imputation for Low-Coverage, Next-Generation Sequence Data in Crop Plants. Plant Genome 7. http://dx.doi.org/10.3835/plantgenome2014.05.0023