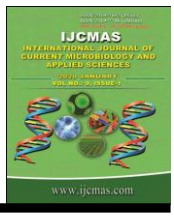
Int.J.Curr.Microbiol.App.Sci (2020) *9*(1): 1070-1077

International Journal of Current Microbiology and Applied Sciences
***ISSN: 2319-7706* Volume 9 Number 1 (2020)**
Journal homepage: http://www.ijcmas.com



Review Article https://doi.org/10.20546/ijcmas.2020.901.121

Genotyping-by-Sequencing (GBS): An Hi throughput Genotyping approach for Marker-Trait Association Analysis

Rahul Kumar* and J. Jorben

Division of Genetics, ICAR-Indian Agricultural Research Institute, New Delhi, India

**Corresponding author*

Keywords

Genotyping-by-sequencing, Marker-trait, RAD-seq, GBS

Article Info

Accepted:
15 December 2019
Available Online:
20 January 2020

A B S T R A C T

Exponential cost reduction in sequencing with advances in Next Generation Sequencing (NGS) technologies has contributed to rapid genotyping technology innovations. Genome Complexity Reduction Methods such as Restriction Associated DNA Sequencing (RAD-seq) and Genotyping-by-Sequencing (GBS) have emerged as a strong genotyping tool capable of identifying, sequencing and genotyping not hundreds but thousands of markers across almost any genome of interest, but also numbers of individuals in a single, simple experiment in a population. GBS currently uses a low coverage NGS-backed sequencing protocol for genotyping large populations and more accurately Genotype and phenotype association. Wide proportion of missing data points due to low coverage of sequencing, management and analysis of large amounts of sequence data are few potential drawbacks of GBS. But with further increase in sequencing performance, these techniques will be further improved by the availability of more reference genomes and advances in the bioinformatics field. GBS is however versatile, fast and low-cost, making it an ideal tool for many applications and addressing many plant breeding and genetics issues. The most popular application of the GBS method is linkage mapping, GWAS, marker-assisted and genomic selection, assembly of genomes and improvement and complexity of crops with different genome sizes.

Introduction

Genotyping is the process of determining differences in an individual's genetic makeup (genotype) by examining the DNA sequence of the individual using biological assays and comparing it with the sequence or reference sequence of another individual. It shows the alleles inherited from their parents by an adult. Current genotyping methods include genomic DNA (RFLP), genomic DNA (AFLP), (PCR), allele-specific oligonucleotide (ASO), DNA microarray or bead hybridization, and DNA sequencing. DNA sequencing is the method of determining a given DNA fragment's nucleotide sequence. Because of current technological constraints, nearly all genotypes

are partial. That is only a small fraction of genotype of an organism, such as GBS (genotyping by sequencing) or RADseq, is known. New technologies for mass sequencing aim to provide genotyping (or sequencing of the entire genome) in the future.

Methods for high throughput marker discovery

Amplicon sequencing

Amplicon sequencing is a highly targeted technique that helps researchers to examine genetic variation in different genomic regions. PCR products (amplicons) ultra-deep sequencing allows for efficient detection and characterization of variants. This method uses oligonucleotide samples designed to target and catch interesting areas, followed by next-generation sequencing (NGS). Amplicon sequencing is useful for finding unusual somatic mutations in complex samples (such as tumors combined with germline DNA). Another common application is multi-species sequencing of the bacterial 16S rRNA gene, a widely used tool for studies of phylogeny and taxonomy, particularly in various samples of metagenomy.

Transcriptome sequencing

Whole transcriptome shotgun sequencing (WTSS), also known as RNA-Seq (RNA sequencing), uses next-generation sequencing (NGS) to reveal the presence and quantity of RNA in a biological sample at a given time. RNA-Seq is used to analyze the cellular transcriptome that continues to change. In particular, RNA-Seq allows the ability to analyze alternate gene spliced transcripts, post-transcription modifications, gene fusion, mutations / SNPs and gene expression changes over time, or gene expression variations in different ways. Groups or therapies. In addition to mRNA transcripts, RNA-Seq may include total RNA, small RNA such as miRNA, tRNA, and ribosomal profiling in different populations of RNA. RNA-Seq can also be used to define exon / intron boundaries and to check or change previously annotated boundaries of 5' and 3' genes. Recent RNA-seq advances include single-cell sequencing and fixed tissue in-situ sequencing.

Whole genome sequencing

Whole genome sequencing (also known as WGS or complete genome sequencing) is the process of determining the complete DNA sequence of a genome. This involves sequencing all of the chromosomal DNA of an organism as well as the DNA in the mitochondria and in the chloroplast for plants. For fact, nearly complete genome sequences are also referred to as entire genome sequences.

Sequence capture

A sequencing library with selected biomarkers is designed and hybridized against a set of specific samples. Hybridization can be carried out either in solution (solution hybrid collection, SHS) with biotinylated samples collected by streptavidin-coated magnetic beads, or on a solid support (array-based hybrid selection, AHS) on which samples are detected. Non-target sequences after hybridization wash away and elucidate and sequence the enriched sample.

SNP Mining

Single nucleotide polymorphisms (SNPs) are the most abundant source of genetic variation on which most molecular markers can be based. For those regions or species that do not have verified SNPs in the public databases, mine them from DNA sequences is a good

alternative. The alignment of multiple sequence fragments from different genotypes on the genome representing the same region will enable sequence variants to be discovered. The public repositories contain a large number of sequence data to be collected. (These are both expressed sequence tags and genomic sequences) and are free to use without sequencing on a large scale. High-throughput sequencing is becoming widely available, however, with the advent of next-generation sequencing machines (Roche GS/454, Illumina GA / Solexa, SOLiD). This will allow polymorphic genotypes to be sequenced on specific target areas and the subsequent identification of SNP.

Reduced representation approaches

Exponential cost reduction with advances in Next Generation Sequencing (NGS) technology has resulted in rapid developments in genotyping technology. Genome complexity reduction methods Such as Restriction Associated DNA Sequencing (RAD-seq) and Genotyping-by-Sequencing (GBS) have become more popular genotyping tool capable of identifying, sequencing and genotyping thousands of markers across nearly every genome of interest and also the number of individuals in a single, simple experiment in a population. GBS currently uses NGS power-supported low coverage sequencing protocol for genotyping large populations and more reliable genotype and phenotype associations. The sequencing of reduced representation includes the following approaches:

a) Reduced Representation Libraries
b) Complexity Reduction of Polymorphic Sequences
c) Restriction Site –Associated DNA Sequencing
d) Genotyping By Sequencing
e) Multiplexed Shotgun Genotyping

Genotyping by sequencing approach

Elshire*et al.*, (2011) first identified the procedure. In the field of genetic sequencing, genotyping by sequencing, also known as GBS, is a technique for discovering single nucleotide polymorphisms (SNP) to perform genotyping studies, such as genome-wide association studies (GWAS). GBS uses enzymes to reduce the complexity of the genome and the genotype multiple samples of DNA.

PCR is performed after digestion to increase the pool of fragments and then sequencing GBS libraries using next-generation sequencing technologies, usually resulting in approximately 100 bp single-ends reads. It is the essential molecular technique to reduce the breeding cycle. It involves the discovery of markers, Development of assays and genotyping. It is a sequence-based genotyping that requires the joint discovery of markers and genotyping.

GBS is the modified form of DNA sequencing associated with the Restriction Site. It is less complex and cost-effective. GBS was originally developed for maize association studies of high resolution and was extended to a range of species with complex genomes, like RAD.

It works well with complex genomes and small genomes (organellar, microbial, and chloroplast). SNPs are the most abundant in a genome among different types of molecular markers and are suitable for analysis on a wide variety of genomic scales. GBS Provides a simple and low-cost method for genotype breeding populations, enabling plant breeders to carry out GWAS , genomic diversity research, genetic linkage review, molecular marker discovery and large scale plant breeding programs(GS) selection.

Requirements for GBS

DNA samples

From mapping populations such as RILs or DH, DNA extraction using standard CTAB method.

Restriction Enzyme

RE left 2-3 bp overhang, not regularly cut in the main repetitive fraction of the genome. e.g. Maize *ApeKI*. Two RE can also be used such as *PstI* and *MseI*. These are Type II endonuclease restriction which recognizes a degenerate 5 bp sequence (GCWGC where W is A or T) and creates a 5' overhang (3 bp) and has relatively few recognition sites in the major classes of maize retrotransposons. These are partially sensitive to methylation (will not be cut if the 3' base of the recognition sequence on both strands is5-methylcytosine).

Adapters

It uses two different adapter types, Barcode adapter and Common adapter. Adapters are constructed so that after ligation to genomic DNA, the *ApeKI* recognition site did not occur in any adapter series. "Barcode" adapter ends with a 4 to 8 bp barcode on the 3' end of its top strands and a 3 bp overhang on the 5' end of its bottom strand complementing the sticky ends produced by *ApekI* (CWG). The Second or Standard adapter only contains an *ApeKI*–compatible sticky ends.

Sequencers

It is possible to use next generation sequencers such as Illumina.

Sequence selection and program alignment

BLAST (Basic Local Alignment Search Tool), Burrows-Wheeler Alignment Tool (BWA), TASSEL, SAM tools can be used to align the sequences to the reference sequence.

GBS steps:

Sample preparation

DNA isolation and digestion with *PstI, MspI, or ApeKI* restriction enzyme.

NGS library construction

DNA fragments are connected to one barcoded and one specific adapter, leading to Illumina.

Flow cell PCR amplification. Only those segments which have forward adapter at one end and reverse adapter at other end are amplified to give a clone, other fragments are not amplified.

Reads mapped with the reference genome for calling after that fragments sequenced at one end.

Statistical analysis and SNP discovery

There is several software packages specifically designed for a reduced representation scheme, some of which are listed below:

PoPoolation

PoPoolation software is designed to analyze pooled NGS sequence data to detect nucleotide polymorphism.

The Burrows-Wheeler Alignment (BWA) tool is used to map the reads to a reference genome, convert the aligned reads to a pile-up file using SAM tools and this file is then used with the PoPoolation method for genome-wide study of polymorphism. The source code can be downloaded from the following link(http:/code.google.com/p/popoolation/)

RAD tools 1.2.4

RAD tools 1.2.4 software is designed to manage RAD sequence data to detect SNPs and structural variations. This program discovers Illumina sequence reads candidate genetic markers.

Stacks

Stacks package analyzes GBS sequence data efficiently. Analyze RAD-seq and GBS sequence data generated by any restriction enzyme input data in the FASTA or FASTQ format. Its source code can be downloaded at(http:/creskolab.uoregon.edu/stacks/).

TASSEL

It contains the TASSEL-GBS sequence data discovery pipeline for SNPs. Even an unfinished reference genome sequence can be used for the discovery of SNP. The discovery pipeline uses all available FASTQ files up to date to discover SNPs for the organisms. These SNPs are detected, sorted and stored in a physical map called TOPM, in a state. Download from the following link address (http:/www.maizegenetics.net/tassel).

SAMtools / BCFtools software

SAMtools / BCFtools package is designed to detect SNPs, short InDels and structural variations from NGS sequence data using a reference genome sequence. It can be downloaded from the following link address (http:/samtools.sourceforge.net/mpileup/shtml)

GBS application

Marker discovery and creation of high-density linkage maps

GBS markers can be used to enrich existing reference linkage maps or to create denovo GBS maps where reference maps are not available in a simple and straightforward manner (Poland and Rife 2012). In traditional mapping experiments in new populations, marker data is first used to generate linkage maps or integrate existing reference maps with high computational steps of recombination frequency calculation and then order markers. GBS markers can be ordered for organisms where reference genomes are available next, on reference maps based on their physical location on the genome.

Linkage / association mapping

GBS can be an excellent linkage / association-based mapping approach with more dense genetic maps generating more accurate and high-resolution gene / QTL mapping. For perennial, RAD sequencing was used to map the revolution gene sd-1 in a region of 100 kb in the RIL population showing the value of a high-resolution mapping method.

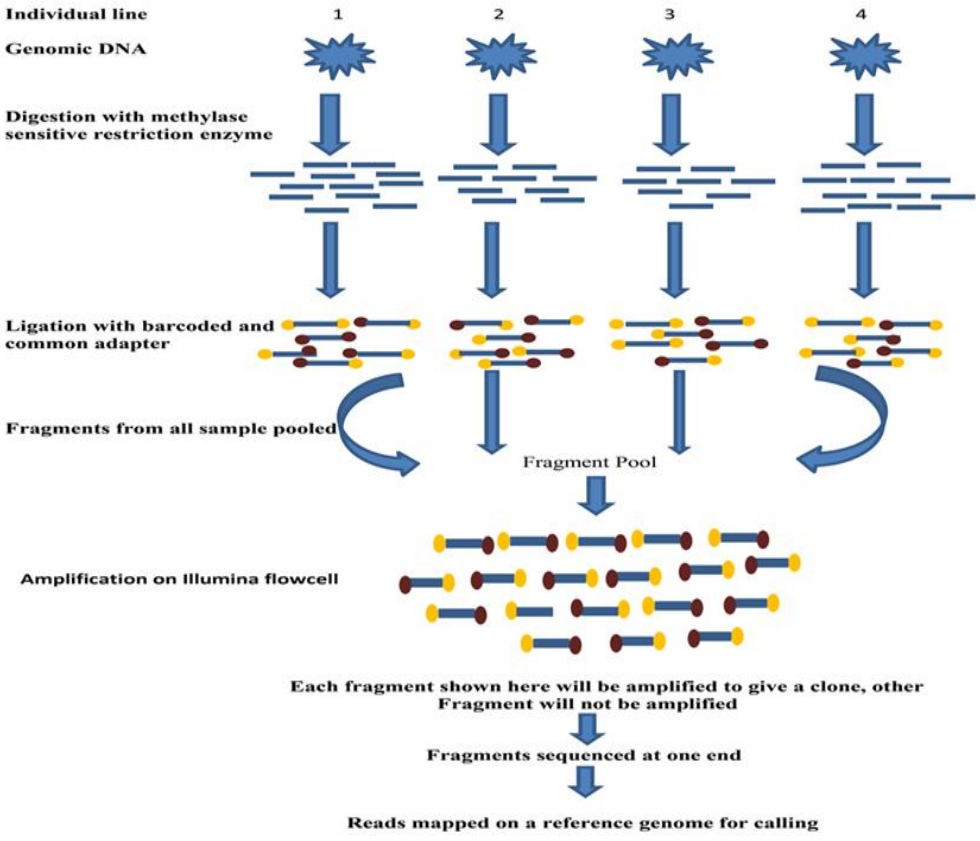
Improvement of the reference genome

GBS-developed high-density genetic maps can be used to anchor and order physical maps and to refine or correct unordered sequence contigs. In the case of *D. Simulans*, Andolfatto *et al.,* (2011) were able to assign 8 Mb to linkage groups, which comprised 30% of the unassembled *D. simulans* genome or about 6% of the total genome. This is a significant improvement in an already well-characterized genome. Gao *et al.,* (2013) enhanced the genome sequences of common Chinese hybrid Liang–You–Pei–Jiu paternal variety 93–11 and maternal variety PA64 using GBS re-sequencing dataset. Similarly, high-density GBS maps are used to assist with the anchoring and ordering of large numbers of assembled but unanchored and unordered contigs in much larger, more complex genomes, like barley and wheat (International Barley Sequencing Consortium 2012).

Table.1 A comparison among the reduced representation sequencing approaches

Feature	CRoPS	RRL	RAD-Seq	GBS	MSG
Amount of DNA needed	300 ng/sample	25ug	300 ng/sample	100 ng/sample	10 ng/sample
Reference genome	Not required	Not required	Not required	Preferable	Preferable
Fragments produced by	two RE	one RE	RE and Mechanical Shearing	one or two RE	one RE
Complexity reduction by	Selection nt in PCR primers	Fragment size selection	presence of restriction site	RE and Adapter ligation	Fragment size selection
Sequencing of	complete fragment	Fragment ends	Fragment ends	Fragment ends	complete fragment
Suitable for studies with	wild populations	wild populations	wild populations	experimental populations	experimental populations
QTL mapping and MAS	low suitability	low suitability	moderate suitability	high suitability	high suitability

Figure.1 A schematic representation of GBS for SNP discovery and genotyping



GS (genomic selection)

It involves methods using compact, genome-wide molecular markers to predict individuals ' GEBV (Genomic Estimated Breeding Value) and to select individuals based on GEBV without taking it. GS provides the ability to identify specific quantitative characteristics based solely on marker data and incorporates the benefits of high-throughput technologies such as GBS and advances in the statistical methods used for data analysis. GS will significantly accelerate the breeding cycle while also using marker knowledge to preserve genetic diversity and potentially expand advantage beyond phenotypic selection possibilities (Lorenz *et al.,* 2011). In several important crops, including maize and wheat, the accuracy of genomic prediction using GBS is currently under investigation.

Marker assisted selection

GBS is used to classify SNPs and these SNPs can also be used as markers for Marker Assisted Selection. GBS has been used for marker assisted selection in Grapevine for powdery mildew resistance (Saxena and Varshney, 2017).

In conclusion, GBS provides an alternative to complex and costly high-performance protocols for Marker discovery systems, the advantage of minimizing genome complexity with Restriction enzymes coupled with multiplex NGS performed for high-density SNP discovery and genotyping. Due to the fact that DNA fragments are processed more readily using a genome-wide approach (as opposed to a targeted approach where only a specific portion of the genome is sequenced), markers are uniformly spread across the genome. GBS is an ideal platform for studies from single gene markers to complete genome profiling. GBS is a quick and low-cost method for breeding populations of genotype, enabling plant breeders to carry out GWAS, genomic diversity studies, genetic association research, molecular marker discovery and genomic selection (GS) programs on a large scale.

References

Andolfatto P, Davison D, Erezyilmaz D, Hu TT, Mast J, Sunayama- Morita T, Stern DL (2011) Multiplexed shotgun genotyping for rapid and efficient genetic mapping. Genome Res 21: 610-617

Baird NA, Etter PD, Atwood TS *et al.,* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE 3: e3376 doi:10.1371/journal.pone.0003376.

Bastien M, Sonah H, François-Belzile F (2014) Genome wide association mapping of *Sclerotinia sclerotiorum* resistance in soybean with a genotyping-by-sequencing approach. The Plant Genome 7(1): 1-13.

Chen H, He H, Zhou F, Yu H, Deng XW (2013) Development of genomics-based genotyping platforms and their applications in rice breeding. CurrOpin Plant Biol 16: 247-254.

Crossa J, Beyene Y, Kassa S, Pérez P, Hickey JM, Chen C, Campos G de los, Burgueño J, Windhausen VS, Buckler E, Jannink JL and Babu R (2013).Genomic Prediction in Maize Breeding Populations with Genotyping-by-Sequencing.G3, 3: 1903–1926.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6: 193-199 doi:10.1371/journal.pone.0019379.

Godfray HCJ, Beddington JR, Crute IR, Haddad L, Lawrence D, *et al.,* (2010) Food security: The challenge of feeding

9 billion people. Science 327: 812-818.

Helentjaris TG, King G, Slocum M *et al.*, (1985) Restriction fragment length polymorphism as probes for plant diversity and their developments as tools for applied plant breeding. Plant MolBiol 5: 109-118.

Konieczny A, Ausubel FM (1993) A procedure for mapping *Arabidopsis* mutations using co-dominant ecotype-specific PCRbased markers. Plant J 4: 403-410.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics 25: 1754-1760.

Lorenz AJ, Chao S, Asoro FG, Heffner EL, Hayashi T, Iwata H, Smith KP, Sorrells ME, Jannink JL (2011) Genomic selection in plant breeding: knowledge and prospects. AdvAgron 10: 77-120.

Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nat Rev Genet 11: 499-511.

Marchini J, Howie B, Myers S, McVean G and Donnelly P (2007) A new multipoint method for genome-wide association studies via imputation of genotypes. Nat Genet 39: 906-913.

Oeveren van J, de Ruiter M, Jesse T, vanderPoel H, Tang J, *et al.*, (2011) Sequence-based physical mapping of complex genomes by whole genome profiling. Genome Res 21: 618-625.

Poland JA and Rife TW (2012).Genotyping-by-Sequencing for Plant Breeding and Genetics. The Plant Genome 5:92–102. doi: 10.3835/plantgenome2012.05.0005

Saxena RK. And Varshney RK (2017). Construction of genotyping by sequencing based high density genetic maps and QTL mapping for fusarium wilt resistance in pigeonpea.*Sci Rep* 7, 1911.doi:10.1038/s41598-017-01537-2

Stekhoven DJ, Bühlmann P (2011) MissForest-nonparametric missing value imputation for mixed-type data. Bioinformatics 28: 112- 118.

Stolle E, Moritz RFA (2013) RESTseq-Efficient benchtop population genomics with restriction fragment sequencing PLoS ONE. 8:179-185. doi: 10.1371/journal.pone.0063960.

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, *et al.*, (2001) Missing value estimation methods for DNA microarrays. Bioinformatics 17: 520-525.

How to cite this article:

Rahul Kumar and Jorben, J. 2020. Genotyping-by-Sequencing (GBS): An Hi throughput Genotyping approach for Marker-Trait Association Analysis. *Int.J.Curr.Microbiol.App.Sci.* 9(01): 1070-1077. doi: https://doi.org/10.20546/ijcmas.2020.901.121