SEMINAR REPORT

Genotyping by Sequencing (GBS)

By

Sreekutty S. S.

2018-11-003

Presented on 25/01/2020

Submitted in partial fulfilment of requirement of the course

MBB 591 Masters' Seminar (0+1)

CENTRE FOR PLANT BIOTECHNOLOGY AND MOLECULAR BIOLOGY

COLLEGE OF HORTICULTURE

KERALA AGRICULTURAL UNIVERSITY

VELLANIKKARA

THRISSUR, KERALA- 680656

CERTIFICATE

This is to certify that the seminar report entitled **"Genotyping by Sequencing (GBS)"** has been solely prepared by **Ms. Sreekutty S. S. (2018-11-003)**, under my guidance and has not been copied from seminar reports of any seniors, juniors or fellow students.

Vellanikkara

Dr. Deepu Mathew

Assistant Professor,

Centre for Plant Biotechnology
and Molecular Biology,

College of Horticulture,

KAU, Vellanikkara - 680656

DECLARATION

I, Sreekutty S. S. (2018-11-003) declare that the seminar entitled **"Genotyping by Sequencing"** has been prepared by me, after going through various references cited at the end and has not been copied from any of my fellow students.

Vellanikkara   Sreekutty S. S.
                                                               (2018-11-003)

CERTIFICATE

This is to certify that the seminar report entitled **"Genotyping by Sequencing"** is a record of seminar presented by **Ms. Sreekutty S. S. (2018-11-003)** on 16th January, 2020 and is submitted for the partial requirement of the course MBB 591.

Dr. Anil Kuruvila
Associate Professor
Department of Agricultural Economics
College of Horticulture, Vellanikkara

Dr. Reshmy Vijayaraghavan   Dr. Sangeethakutty
Assistant Professor   Assistant Professor
Department of Plant Pathology   Department of Vegetable science
College of Horticulture, Vellanikkara   College of Horticulture,Vellanikkara

INDEX

LIST OF TABLES

LIST OF FIGURES

1. Introduction

World population is estimated at 9 billion by 2050 and by the time yield of major crops should increase by 2.4% per year to feed the entire population. The primary aim of next generation genomics in agriculture is to connect phenotype to genotype with great precision and use this knowledge to make crop improvements in less time and cost. Genotyping by sequencing (GBS) is currently a powerful technique backed by power of Next Generation Sequencing (NGS) for genotyping large populations and more precise association of genotype and phenotype. GBS uses enzyme-based complexity reduction method coupled with DNA barcoded adapters to produce multiplex libraries of samples ready for NGS sequencing (Bhatia *et al*.,2013).

The advent of next generation sequencing technologies has led to the development of genome wide single nucleotide polymorphism detection applications in many plant genomes. whole genome sequencing is found to be applicable in almost all organisms. But in many cases it is just as effective as to obtain information from the subset of the genomes. Recent improvements in sequencing throughput combined with an overall decrease in costs per giga base of sequence is allowing NGS to be applied to not only the evaluation of small subsets of parental inbred lines, but also the mapping and characterization of traits of interest in much larger populations. Such an approach, where sequences are used simultaneously to detect and score SNPs, therefore bypassing the entire marker assay development stage, is known as Genotyping by Sequencing (GBS). It provides an ideal platform for studies ranging from single gene marker to whole genome profiling (Deschamps *et al*.,2012).

GBS is a method to identify genetic variants and quickly genotype samples, reduces genome complexity by using restriction enzymes to divide the genome into fragments whose ends are sequenced on short-read sequencing platforms. It is a genotyping approach that makes use of NGS to rapidly and economically scan a genome. It has been shown to allow the simultaneous discovery and genotyping of thousands to millions of SNPs across a wide range of species. GBS is a particularly attractive complexity reduction method that offers a simple, robust, low-cost, and high-throughput method for genotyping (Torkamaneh *et al*., 2016).

This method is to identify genetic variants and quickly genotype samples, reduces genome complexity by using restriction enzymes to divide the genome into fragments whose ends are sequenced on short-read sequencing platforms. Recent advance of GBS offers an ultimate marker assisted selection tool to accelerate plant breeding and crop improvement. The use of single nucleotide polymorphism as DNA markers for plant genotyping has increased the

potential to score variation in specific DNA targets. GBS is becoming increasingly important as a cost-effective and unique tool for genomics-assisted breeding in a range of plant species. GBS technology by summarizing the steps needed for any plant species and some potential application of the results (He *et al*., 2014).

The GBS procedure is demonstrated with maize (IBM) and barley (Oregon Wolfe Barley) recombinant inbred populations where roughly 200,000 and 25,000 sequence tags were mapped, respectively. An advantage in species like barley that lack a complete genome sequence is that a reference map need only be developed around the restriction sites and this can be done in the process of sample genotyping. In such cases, the consensus of the read clusters across the sequence tagged sites becomes the reference. Alternatively, for kinship analyses in the absence of a reference genome, the sequence tags can simply be treated as dominant markers. Future application of GBS to breeding, conservation and global species and population surveys may allow plant breeders to conduct genomic selection on a novel germplasm or species without first having to develop any prior molecular tools or conservation biologists to determine population structure without prior knowledge of the genome or diversity in the species (Elshire *et al.,* 2011).

2. Sequencing methods

Sequencing methods mainly include whole genome sequencing, resequencing, resequencing reduced representation sequencing and targeted sequencing. Whole genome sequencing is the process of determining the complete DNA sequence of an organism's genome at a single time (Deschamps *et al*.,2012). In resequencing, sequencing is done on a particular part of an individual genome to detect the sequence variations present in the individual genome and standard genome of that species. In reduced representation Sequencing, sequencing is done on randomly selected parts of a whole genome using molecular markers. It is divided into two types, Restriction site Associated DNA sequencing (RAD sequencing) and Genotyping by Sequencing (GBS). In Targeted Sequencing, sequencing is done on a targeted genome or a DNA fragment by using either probe method, PCR methods or any next generation sequencing methods. Among GBS and RAD, GBS is the most advanced form because steps are minimum, only less amount of DNA is needed and size selection step is neglected. Barcode sequence length is more in GBS than RAD. It increases the distinction of samples while pooling and thus increases the accuracy (Wickland *et al*., 2017).

Genotyping by Sequencing (GBS)

Genotyping by sequencing is a genetic screening method for discovering novel plant SNPs and performing genotyping studies (He *et al*., 2014). GBS is the cost-effective method of choice for genome wide SNP discovery without prior knowledge of the genome sequence. It is a genotyping approach that makes use of Next Generation Sequencing (NGS) to rapidly and economically scan a genome. It is first developed by Elshire and his co-workers by using recombinant inbred lines of maize and double haploid lines of barley in 2011.
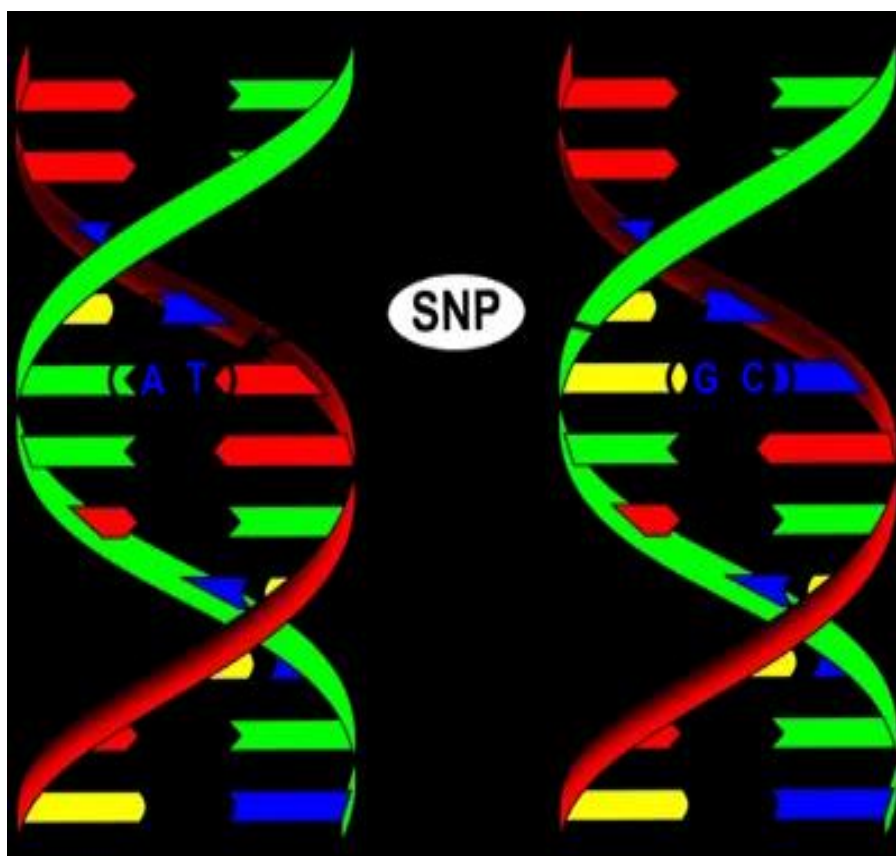

Fig 1: SNP variation present among two genomes

In conventional DNA sequencing the procedure is marker discovery, assay designing and genotyping (Gore *et al*., 2007). But in GBS simultaneous marker discovery and genotyping takes place. Since the advent of high-throughput sequencing technologies, the abundant and heritable single nucleotide polymorphisms (SNPs) have emerged as the most widely used genotyping markers. The versatility of SNPs has also led to their widespread use in phylogenetics and phylogeography. A major advantage of the single-base resolution of SNPs is that it allows better detection of 'perfect' markers, which are causally linked to agronomic traits (Scheben *et al*., 2017).

3 GBS Methodology

The first step is to create an inexpensive, robust multiplex sequencing protocol and low coverage Sequencing has to be done using an Illumina genome analyzer. Informatics pipelines are used to anchor markers across the genome to impute missing data and finally Combine genotypic and phenotypic data (Elshire *et al*., 2011)

3.1 Materials

1) DNA samples

DNA has to be isolated from plant materials using different extraction methods. The exact method for collecting DNA samples varies by organism, but should be performed in a way that yields enough DNA for analysis while minimizing contamination from other samples. Since the default GBS protocol involves relatively low coverage per individual, bulking heterogeneous individuals (i.e., pooling DNA from multiple, genetically distinct individuals) is not recommended since there most likely will not be enough read depth at any given locus to accurately call allele frequencies. Even when working with inbred organisms, sampling only a single individual, if possible, is recommended to minimize chance of contamination (Wallace and Mitchell, 2017)

2) Restriction Endonuclease

It mainly involves two strategies, one enzyme and two enzyme strategies, in which one enzyme strategy involves one methylation sensitive restriction enzyme which do not cut frequently on repetitive genomic fraction. Most commonly used one enzyme is *ApeK1* restriction enzyme. It is proved to be applicable in crops like barley, maize, wheat *etc* and it has a restriction site GCWGC, W is nothing but A or T base (Elshire *et al*., 2011). The two enzyme strategy involves one common cutter and a rare cutter enzyme for the uniform and complete distribution of markers across the genome, *Pst1* is the most commonly used rare cutter and *Msp1* is the common cutter enzyme (Poland and Rife., 2012). In GBS, we can adopt different reading strategies according to the restriction enzyme. Low coverage sequencing can be obtained from one enzyme strategy and if we need intermediate number of fragments we can adopt two enzyme strategy. But in some cases, we may need reads having more sequence depth and it is possible only by using rare cutter enzyme alone. If the case is, more number of fragments rather than reading depth common cutter enzyme can be used (Sonah *et al.,*2013).
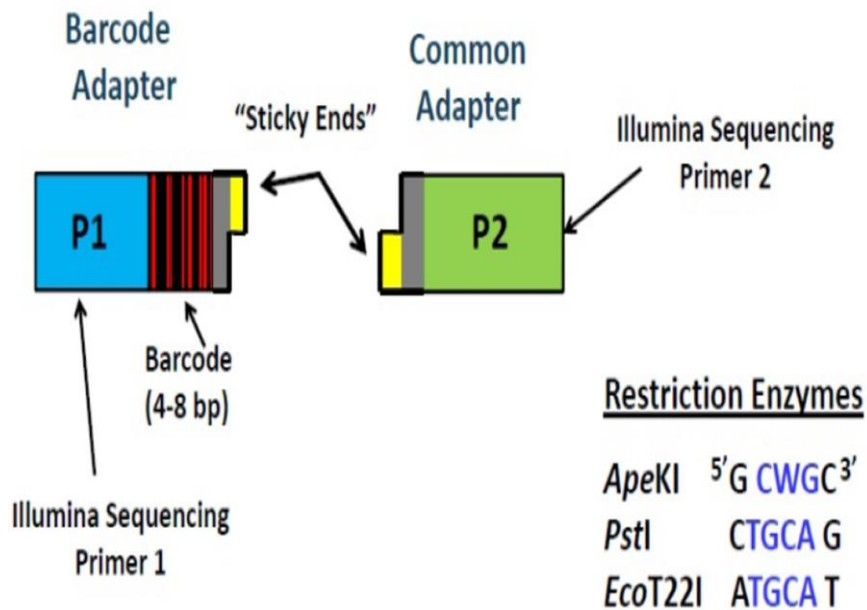
3) Adapters and primers

Fig 2: Adapters and Primers

Commonly used adapters are barcode and common adapter. Barcode adapter contains 4-8 bp unique barcode sequence and complementary sequence for the sticky end generated by restriction enzyme and common adapter only contains the compatable sticky end for the restriction enzyme. Adapters were designed so that the ApeKI recognition site did not occur in any adapter sequence and was not regenerated after ligation to genomic DNA. Primers mainly include Illumina sequencing primer 1 and primer 2, which is used for bridge amplification and PCR primer 1 and primer 2. PCR primer 1 will be always complementary to the barcode adapter sequence and primer 2 will be complementary to common adapter sequence. This complementarity creates an easy platform for PCR reaction (Elshire *et al*.,2011)
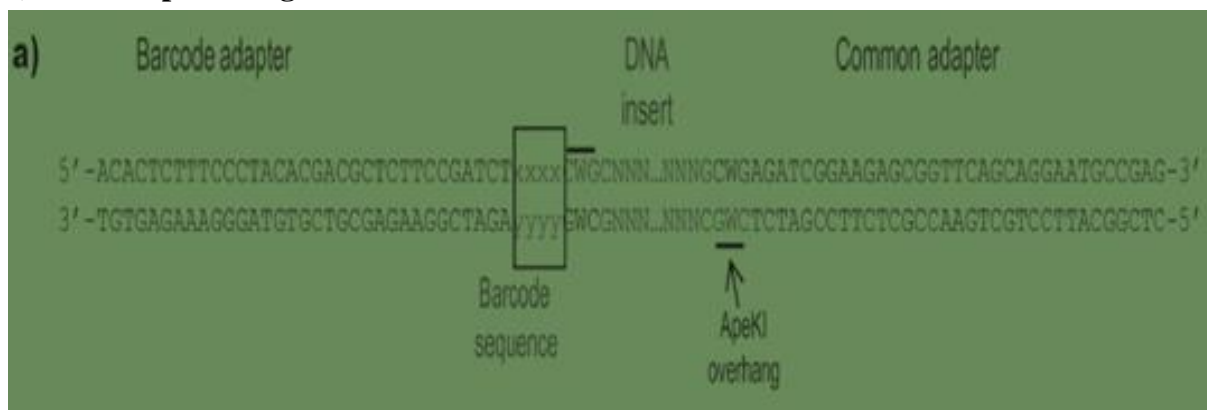
4) GBS adapter design

Fig 3: GBS Adapter design for a typical DNA construct

The figure shows a typical DNA construct having DNA insert in the middle and ApeK1 overhang has its complementary common adapter sequence. Barcode adapter contains unique four base pair sequence which is specific for each sample.

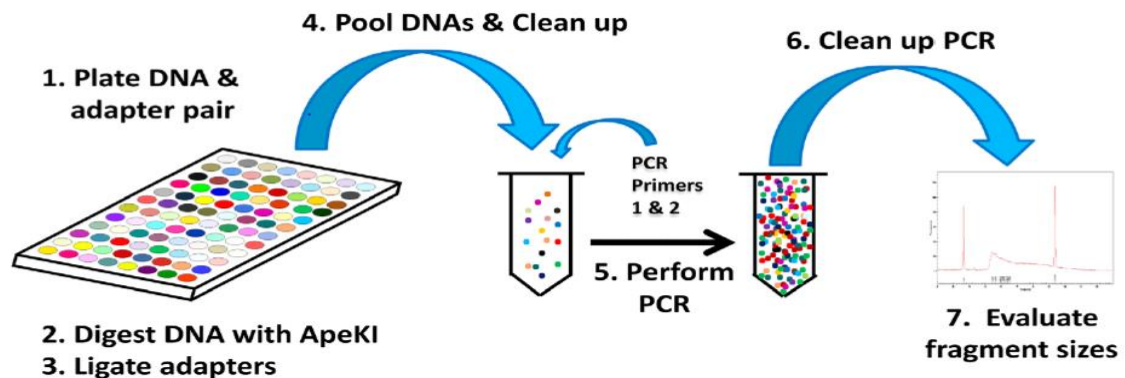## 3.2 Steps involved

Fig 4: Steps involved for GBS protocol   (Elshire *et al*., 2011)

Oligonucleotides comprising the top and bottom strands of each barcode adapter and a common adapter were diluted (separately) in TE (50 mM each) and annealed in a thermocycler. Barcode and common adapters were then quantified using an intercalating dye (PicoGreenH ; Invitrogen, Carlsbad, CA), diluted in water to 0.6 ng/mL (,02 pmol/ml), mixed together in a 1:1 ratio, and 6 mL (,0.06 pmol each adapter) of the mix was aliquoted into a 96-well PCR plate and dried down. DNA samples (100 ng in a volume of 10 mL) were added to individual adapter-containing wells and plates were again dried. Samples (DNA plus adapters) were digested for 2 h at 75uC with ApeKI (New England Biolabs, Ip switch, MA) in 20 ml volumes containing 16 NEB Buffer 3 and 3.6 U ApeKI. Adapters were then ligated to sticky ends by adding 30 ml of a solution containing 1.666 ligase buffer with ATP and T4 ligase solution containing 1.666 ligase buffer with ATP and T4 ligase cohesive end units) (New England Biolabs) to each well. Samples were incubated at 22uC for 1 h and heated to 65uC for 30 min to inactivate the T4 ligase. Sets of 48 or 96 digested DNA samples, each with a different barcode adapter, were combined (5 ml each) and purified using a commercial kit (QIA quick PCR Purification Kit; Qiagen, Valencia, CA) according to the manufacturer's instructions. DNA samples were eluted in a final volume of 50 mL (Elshire *et al*., 2011).

3.3 Sequencers

Sequencing Platform	Sequencing Chemistry	Detection Chemistry	Run Time	Read Length(bp)
Roche 454	Sequencing by Synthesis	Light	23 hours	~800
Illumina Miseq	Sequencing by Synthesis	Fluorescence	39 hours	2×250 b
Illumina Hiseq2500	Sequencing by Synthesis	Fluorescence	11 days (high output)/27 hrs (rapid run)	2×100 b
Life Technologies 5500xl	Sequencing by Synthesis	Fluorescence	8 days	$75 + 35$ b
Ion Torrent PGM	Sequencing by Synthesis	pH	4 hours	100

(Deschamps *et al*., 2012)

Table 1: Commonly used Sequencing platforms

Different sequencing platforms are available like Roche 454, Illumina sequencing platform, life technologies, ion torrent etc, all having a sequencing chemistry of sequencing by synthesis. Among them Illumina is the mostly used next generation platform due to its low cost and high efficiency of up to 98 per cent.
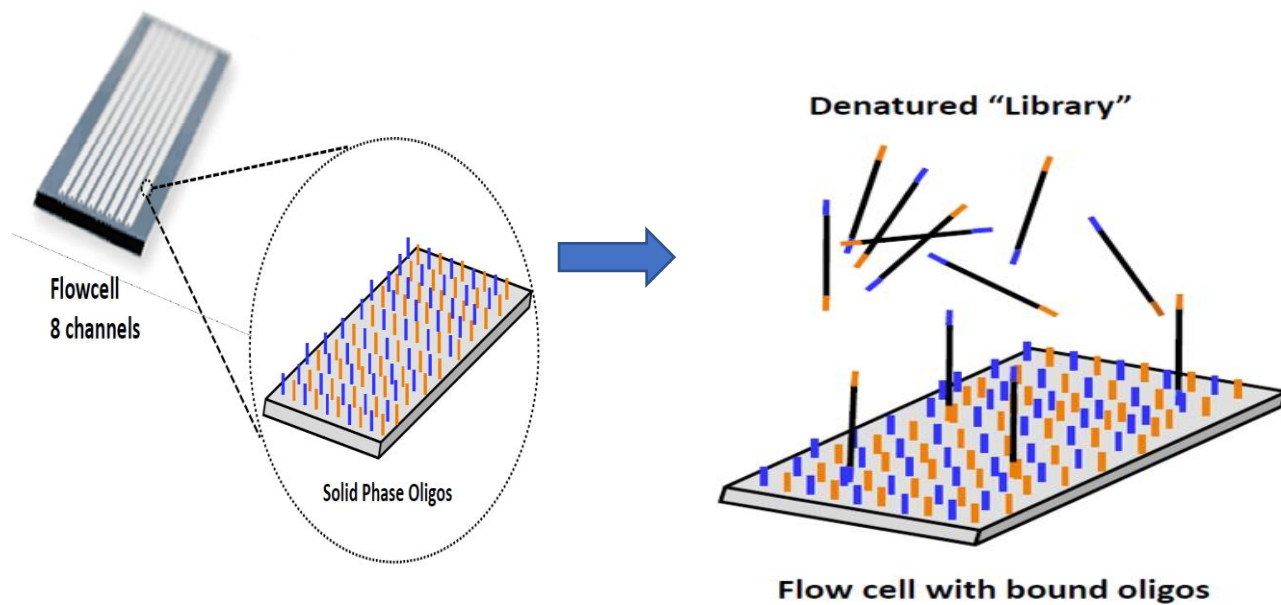
4. Illumina sequencing

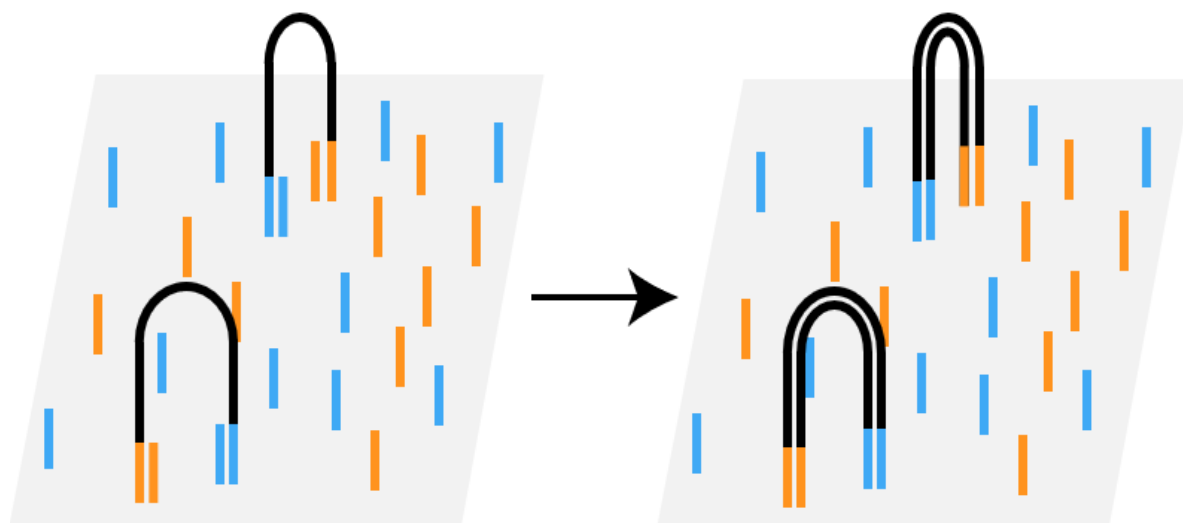

Fig 5: flow cell with bound oligos
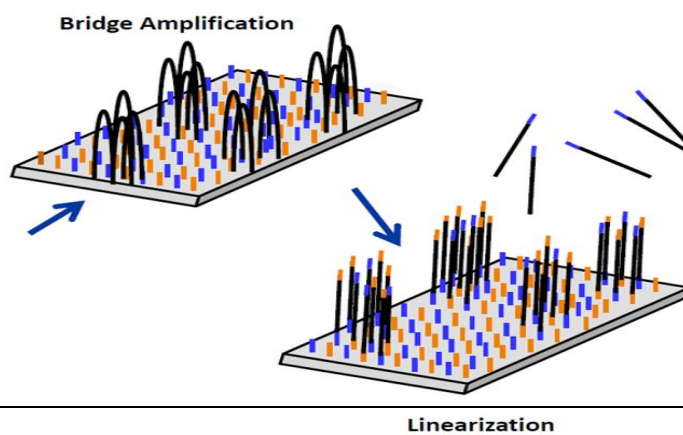

Fig 6: Bridge amplification

Fig 7: Linearization

For Illumina NGS sequencers, DNA molecules and primers are first attached on a slide and amplified with polymerase so that local clonal DNA colonies are formed. To determine the sequence, four types of reversible terminator bases (RT-bases) are added and non-incorporated nucleotides are washed away. A camera takes images of the fluorescently labelled nucleotides, then the dye, along with the terminal 3' blocker, is chemically removed from the DNA, allowing for the next cycle to begin. Unlike pyrosequencing, the DNA chains are extended one nucleotide at a time and image acquisition can be performed at a delayed moment, allowing for very large arrays of DNA colonies to be captured by sequential images taken from a single camera (He *et al*., 2014).

5. Selection of reads

A huge amount of sequence data obtained from the Illumina sequencing platform and selection process involves removal of overlapping and out of phase regions. The first step is to ignore reads with N in the first 72 bp, N is nothing but the out of phase regions. At the sequencing time if any one of the DNA cloned fragment showing a different time of sequencing than the rest it is said to be in out of phase state and computer detected it as 'N'. After selection, the selected reads keep matching one of the barcodes and the cut site remnant. After selection of the reads contain the barcodes, we can trim off these barcode sequences. After trimming off, the remaining sequence again trim off into 64 bp sequence and these sequences are called sequence tags (Elshire *et al*, 2011).

6. Bioinformatic tools

The bioinformatic tools used are

✓ Quality control

For filtering and trimming reads. PRINSEQ is capable of filtering and trimming reads. The FASTX-Toolkit is a collection of command line tools providing quality reports and read trimming. MultiQC and Qualimap 2, now enable multi sample quality assessments. If low sequence quality or contamination is identified, Trimmomatic and Adapter Removal 2 offer the highest throughput and high overall performance for removing contamination and low-quality bases from single- and paired-end FASTQ files (Scheben *et al*., 2017).

✓ To aligns the tags in fasta format Burrows-Wheeler Aligner is using

- ✓ Tags On Physical Map (TOPM) file Contains information to interpret tags present in a species

- ✓ SNPBy Alignment SNPs and small indels are identified from each tag alignment

(Glaubitz *et al*., 2014)

Some of the softwares used are TASSEL, Stacks, SEGMAP etc all having multiple functions such as association study, evaluating evolutionary relationships, analysis of linkage disequilibrium, principal component analysis, cluster analysis, missing data imputation and data visualization (Bhatia *et al*., 2013).

7. Highlights of GBS

It involves the factors responsible for low cost of GBS.

1) Restriction enzyme

   Restriction enzyme involves one methylation sensitive restriction enzyme which do not cut frequently on repetitive genomic regions are used. So can concentrate on the lower copy regions which contains most of the SNPs and traits of interest. Thus there is a reduction in complexity of the genome and the cost.

2) Multiplex sequencing

   It is the process of sequencing multiple samples in a single sequencing run. If the plate have 96 wells, we can add different 96 plant samples into the well and can sequence all in a single Illumina sequencing platform.

3) Illumina sequencing technology

   It itself is a cost reducing technology.

4) Use of bioinformatics

   Rather than other next generation sequencing methods GBS use more bioinformatic tools. So there is a reduction in lab work and reagent cost.

8. Case study

Imputation accuracy of wheat genotyping by sequencing (GBS) data using barley and wheat genome references

(Alipour *et al*., 2019)

This is the latest available paper on GBS published by Alipour and his coworkers in 2019. High throughput SNP genotyping platforms have been successfully used for diploid crops such as maize and barley. wheat however is polyploid crop and has a huge genome (17gb)

with abundant repetitive DNA which present major challenges to direct sequencing the genome for developing high density SNP maps. Poland and his co-workers done an experiment of genotyping by sequencing in wheat genome, but missing data was their. To minimize this missing data, we can either use genome references or can increase the reading depth. Here they calculated the imputation accuracy of wheat genome using genome references. Here the genome references include three wheat references (Chinese spring survey sequence, W7984, IWGSC ref seq v1.0) and a barley reference genome.

Genomic DNA of each sample was double-digested with *PstI* (CTGCAG) and *MspI* (CCGG) restriction enzymes and ligated to barcoded adapters using T4 ligase (New England BioLabs Inc.). All the ligated products were pooled and cleaned up using the QIAquick PCR Purification Kit (Qiagen Inc., Valencia, CA, USA). Primers complementary to both adaptors were used for PCR. PCR amplification started at 95 ˚C for 5 min, followed by 16 cycles of 95 ˚C for 30 s, 62 ˚C for 20 s and 68 ˚C for 1 min and ended by a final extension step at 72 ˚C for 5 min. The PCR product was then cleaned up again using the QIA quick PCR Purification Kit and quantified using Bioanalyzer 7500 Agilent DNA Chips (Agilent Technologies, Inc.). After size-selection for 250-350 bp fragments in an E-gel system (Life Technologies Inc.), concentration of the library was evaluated using a Qubit 2.0 fluorometer and Qubit dsDNA HS Assay Kit (Life Technologies Inc.). The size-selected library was sequenced on an Ion Proton system.

Sequence reads were first trimmed to 64 bp and identical reads were grouped into sequence tags. Unique sequence tags were aligned internally to identify SNPs within the tags allowing mismatches of up to 3 bp. SNPs were called using the Universal Network Enabled Analysis Kit (UNEAK) GBS pipeline [33] in TASSEL 3.0 bioinformatics analysis package [34]. Tags with low quality score (< 15) were removed. SNPs with heterozygotes or a minor allele frequency > 10% were discarded to reduce the false positive markers. Only SNPs with lower than 80% missing data were used for this study. BLASTn analysis was carried out to align sequence tags to the four genome references including one from the barley reference genome [28] and three from wheat reference genomes, the flow-sorted Chinese Spring survey sequence (CSSS) [29], the Popseq W7984 sequence reference [30] and IWGSC RefSeq v1.0. The purpose of using the barley reference genome is to show efficiency of using reference genomes of closely related species to impute missing data in cases where reference genome sequence is absent in some species. If a SNP could be mapped in multiple chromosome positions, the position with the lowest E-value was used to represent the SNP location. In this study, imputation was performed using BEAGLE v3.3.2 [20] and the four genome reference

genomes. BEAGLE used a phasing algorithm to determine haplotype phase for each individual and to impute the missing values based upon allele frequencies. This was done by constructing local haplotype clusters and then sampling a number of haplotypes for each individual from a special class of HMM.

Results include, Two GBS libraries were generated for the 384 wheat accessions with 276 landraces and 12 cultivars in library 1 and 96 cultivars in library 2. To minimize missing data, an average of two sequencing runs was performed for each plate of 96 samples, therefore, library 1 with three plates of samples was run a total of six times and library 2 with one plate of samples was run twice. Eight sequencing runs generated a total of 566,439,207 reads from the two libraries with 81% (458,363,607) of high-quality barcoded reads, from which 133,039 unique SNPs were identified including 16,506, 38,642 and 65,560 SNPs with <20%, <50% and <80% missing data, respectively.
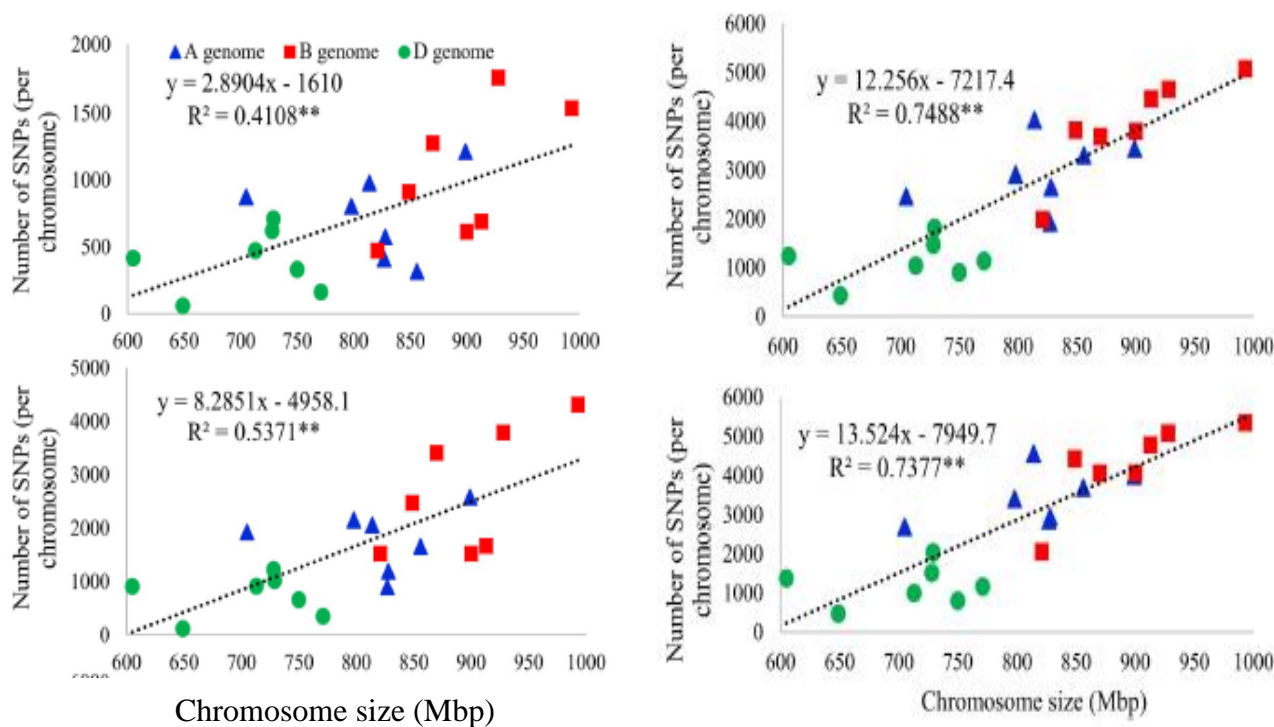

Fig 8: Graphical representation of number of SNPs v/s chromosome size

The numbers of SNPs per chromosome were significantly correlated to the chromosome sizes (Mbp) in all four references. Although they were all significant, the correlations were much lower for the barley reference genome (Fig a) and wheat CSSS assembly (Fig b) than those for W7984 assembly (Fig c) and IWGSC RefSeqv1.0 (Fig d).
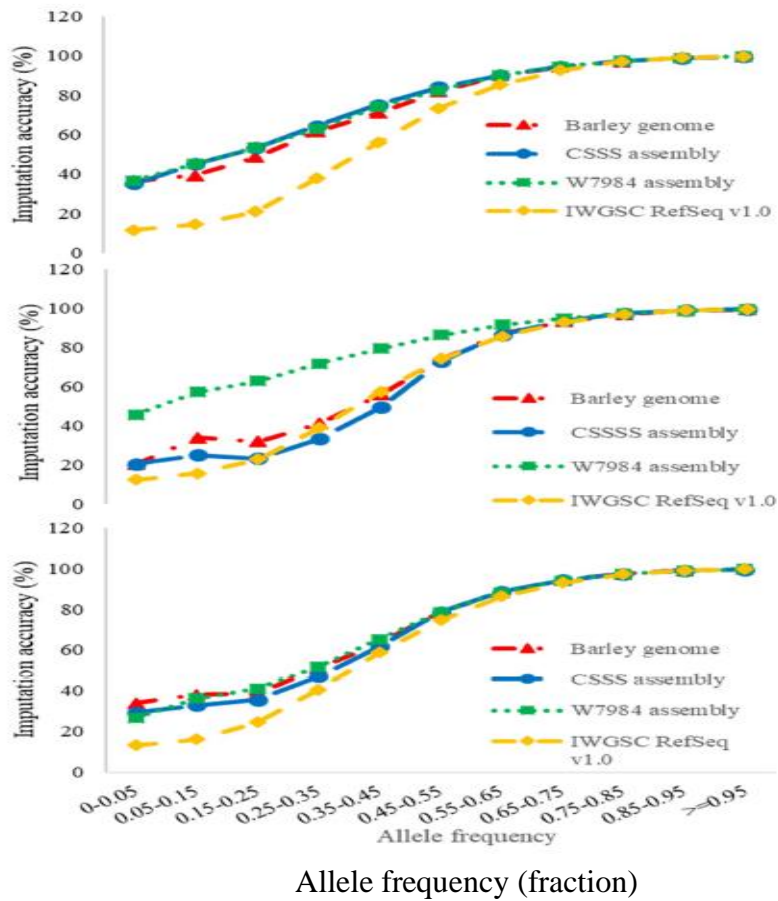
Allele frequency (fraction)

Fig 9: Graphical representation of imputation accuracy against allele frequency

4.9%
IWGSC

12.5%
W7984

38.0%
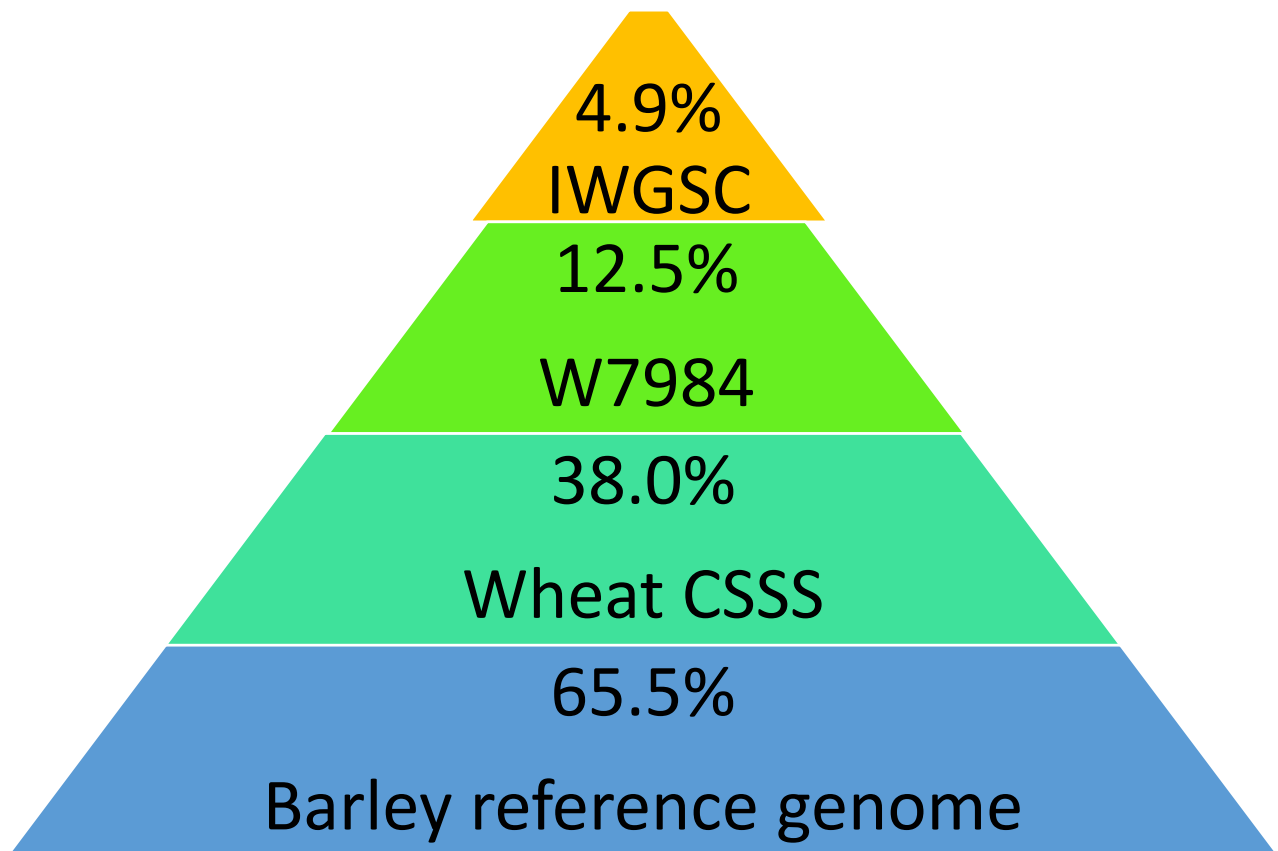Wheat CSSS

65.5%
Barley reference genome

Fig 10: Missing data imputation in genome references

IWGSC contain the highest amount of sequence tags and lowest amount of missing data. So this genome give best results for increasing accuracy of wheat genome. Thus we can conclude that imputation and increasing sequencing depth can quickly fill up missing data, imputation can reduce more missing data and therefore detect more SNPs.
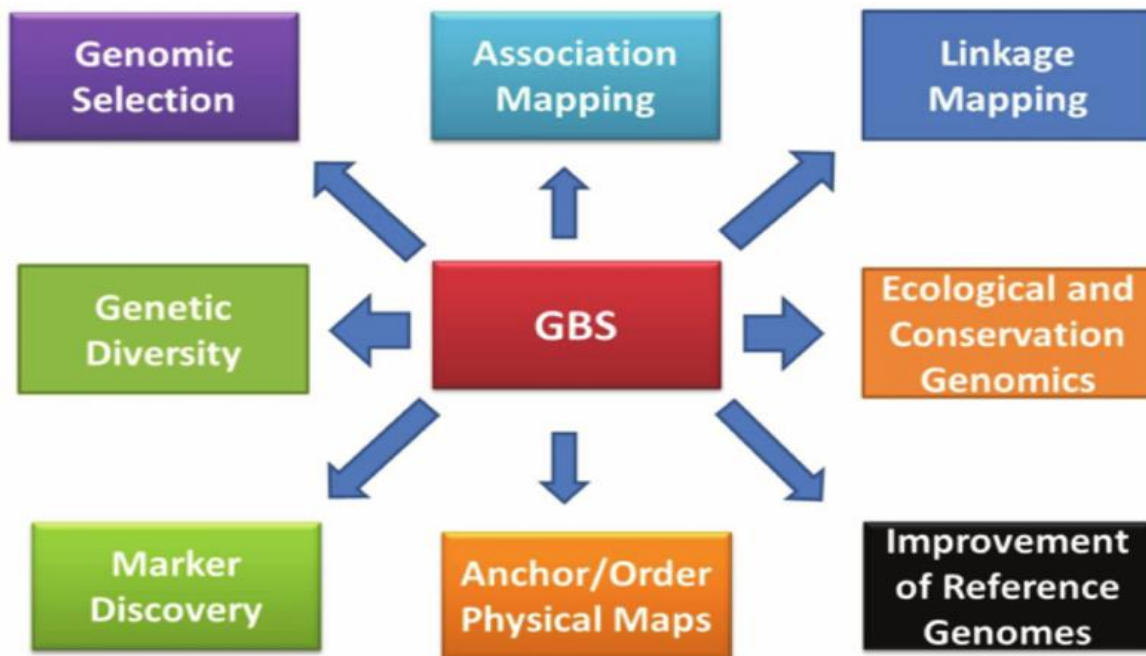
9. Applications of GBS

Fig 11: Applications of GBS

(Chung *et al*., 2017)

10. Why GBS ??

Discussed about why we are using GBS or why we are not using other NGS methods. The reasons are

- Rapid and low cost tool to study
- Large amount of SNPs can be discovered
- Identification of minor allele or quantitative trait loci
- No requirement for prior knowledge of the species genome
- SNP discovery and genotyping are completed together

(Telfer *et al*., 2019)

11. Future prospects

- Plant breeders may conduct marker assisted selection or genomic selection on a novel germplasm or species without first having to develop any prior molecular tools

- Though challenges are there in maintaining and analysing huge datasets, but developments in this field looks promising in future with emerging field of bioinformatics

12. Summary

So far we discussed about

- ❖ Genetic markers and its importance

- ❖ Pooling of DNA samples

- ❖ Illumina sequencing technology

- ❖ Bioinformatic tools

- ❖ Case study on GBS

- ❖ Applications, advantages and highlights of GBS

- ❖ Future prospects

13. Conclusion

- ❑ GBS offers a cost-effective alternative for simultaneous SNP discovery and genotyping across individual lines within a population

- ❑ GBS analysis is the optimal strategy to obtain the largest and accurate SNP data even from a low coverage sequence

15. Discussion

1.What is multiplex sequencing?

It is the process of pooling multiple samples in a single well and sequence all in a single sequencing run.

2.How it become cost effective?

It will take a cost of up to 30$ per sample. It is one by tenth that of whole genome sequencing and can sequence multiple samples.

3.Is it well established in India?

No. It is not commercially used in India. It is on the developing stage and some research works were conducted in India regarding GBS.

4.When doing these types of sequencing in private companies, what is its reliability?

The lab protocol step we can do in our lab and can sequence the restricted fragments by carefully eluting that part. This eluted part can be sequenced with the help of private companies.so there is no such issues of reliability occurs.

5.Which are the molecular markers used in this?

SNPs are constructed by checking the variation between two genomes. No other molecular markers are used in this. This SNP variation is considered as a marker and can be used for wide applications.

6.Sequencing will takes place within a fraction of seconds. Then how light will show one by one in a flow cell?

Actually sequencing is a slow process. Here we are using a 3'blocker in the reverse terminator nucleotide. It will delay the process of adding nucleotides one after other and so we can detect the correct colour.

7.Is GBS can be used for marker assisted selection?

Yes, it is the ultimate tool of Marker Assisted Selection (MAS). In conventional MAS we can access only few number of markers. So the traits under consideration is very low. In GBS we are using thousands of markers and can identify the superior genotypes.

8.How much time will take for GBS?

It only takes 1-2 days and its mainly depending on our handling. If we are doing it by hand it will take more time that done by a liquid handling robot. For ligation we can incubate the samples for hours or over night.

9.Other detection methods are also available for SNP discovery. Then why GBS?

In other detection methods most commonly used is probe method, but its costly and laborious due to the designing process of the probes. Only thousands of SNPs can be discovered from probe method. But in GBS more number can be detected.

References

Alipour, H., Bai, G., Zhang, G., Bihamta, M. R., Mohammadi, V., and Peyghambari, S. A. 2019. Imputation accuracy of wheat genotyping by sequencing data using barley and wheat genome references. *PLoS ONE* 14(1): e0208614.

Bhatia, D., Wing, R. A., and Singh, K. 2013. Genotyping by sequencing, its implications and benefits. *Crop Improv.* 40(2): 101-111.

Chung, Y. S., Choi, S. C., Jun, T. H., and Kim, C. 2017. Genotyping by sequencing: a promising tool for plant genetics research and breeding. *Hortic. Environ. Biotechnol*. 58(5):425-431.

Deschamps, S., Llaca, V., and May, G. 2012. Genotyping by sequencing in plants. *Biology* 1: 460-483

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, A., and Buckler, E. S. 2011. A robust, simple genotyping by sequencing approach for high diversity species. *PLoS ONE* 6(5): e19379.

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire1, R. J., Sun, Q., and Buckler, E. S.2014. TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. *PLoS ONE* 9(2): e90346.

Gore, M., Bradbury, P., Hogers, R., Kirst, M., Verstege, E., Oeveren, J. V., Peleman, J., Buckler, E., and Eijk, M. V. 2007. Evaluation of target preparation methods for single-feature polymorphism detection in large complex plant genomes. *Crop Sci* 47: 135–148.

He, J., Zhao, X., Laroche, A., Lu, Z., 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: 00484.

Poland, J. A. and Rife, T. W. 2012. Genotyping by sequencing for plant breeding and genetics. *The Plant Genome* 5(3): 92–102.

Scheben, A., Batley, J., and Edwards, D. 2017. Genotyping by sequencing approaches to characterize crop genomes: choosing the right tool for the right application. *J. Plant Biotechnol.* 15: 141-161.

Sonah, H., Bastien, M., Iquira, E., Tardivel, A. L., Legare, G., Boyle, B., Normandeau, E., Laroche, B., Larose, S., and Jean, M.2013. An improved genotyping by sequencing approach offering increased versatility and efficiency of SNP discovery and genotyping. *PLoS ONE* 8(1): e54603.

Telfer, E., Graham, N., Macdonald, L., Li, Y., Klapste, J., Resende, M., Neves, L. G., Dungey, H., and Wilcox, P.2019. A high-density exome capture genotype-bysequencing panel for forestry breeding in Pinus radiata. *PLoS ONE* 14(9): e0222640.

Torkamaneh, D., Laroche, J., and Belzile, F.2016.Genome-wide snp calling from genotyping by sequencing (gbs) data: a comparison of seven pipelines and two sequencing technologies. *PLoS ONE* 10: 0161333.

Wallace, J. G. and Mitchell, S. E. 2017. Genotyping by sequencing. *Curr. Protoc. Plant Biol*. 2:64-77.

Wickland, D. P., Battu, G., Hudson, K. A., Diersand, B. W., and Hudson, M. E. 2017. A comparison of genotyping by sequencing analysis methods on low-coverage crop datasets shows advantages of a new workflow, GB-eaSy. *BMC BIOINFORMATICS* 18:586.