Maize F2 Genotyping tGBS Project

Report Date: Aug 29, 2017

SUMMARY This report covers the genotyping of a total of 192 samples of Maize (*Zea mays*) using tGBS® Genotyping by Sequencing technology (2 base selection). Samples were sequenced using an Ion Proton instrument and a total of 534 million reads were generated with an average length of 110bp/read. Reads were aligned to the *Zea mays* AGPv2 reference genome. SNP calling was conducting using only those reads which uniquely aligned to a single location in the reference genome. The high quality SNP set includes 8,434 SNPs which were genotyped in at least 50% of the samples. This dataset is referred to in the rest of the report as the MCR50 SNP set. A second dataset included with this report is the "ALL SNPs" dataset, which includes 18,037 SNPs.

WORKFLOW

The client provided 192 Maize samples for genotyping using tGBS® technology (2 base selection). Samples were QCed prior to tGBS library construction and the client was informed if any samples failed to pass QC. Each sample was then used to construct one or more sequencing libraries. Raw sequence data were debarcoded to assign reads back to individual samples and quality trimming was used to remove low confidence portions of sequence reads. All reads were aligned to the reference genome. Aligned reads were used both to identify polymorphic SNP markers in the population, and to score the genotypes for these markers in individual samples. Figure 1 summarizes this workflow.

Figure 1 A generalized workflow for analyzing tGBS® data

REFERENCE GENOME

Freedom Markers used the *Zea mays* AGPv2 reference genome downloaded from **maizeGDB** (http://maizegdb.org/) as the reference genome for this project.

NOTE: We will typically store Materials and Results for up to 12 weeks after delivery of the Report, unless other arrangements have been made. If you would like your materials destroyed or your results purged sooner, please send your request via email to *support@freedommarkers.com*.

tGBS® READS SUMMARY

As part of this project, a total of 5 Ion Proton sequencing runs were conducted. The total amount of data generated, as well as the minimum, maximum, average and median numbers of raw reads per sample are provided in Table 1. At Freedom Markers we understand how much information can be lost in mean and median summary statistics, so information on the distribution of reads per sample is also presented as both a histogram and as a sorted barchart in Figure 2.

Table 1 Summary of tGBS Reads

DESCRIPTION	NO.
Number of DNA Samples	192
Total Reads *	533,580,676
Minimum Reads per Sample	1,069,794
Maximum Reads per Sample	6,838,292
Average Reads per Sample	2,779,066
Median Reads per Sample	2,481,882
25% Pctl. Reads per Sample	1,859,163

* Raw reads without any further processing

TRIMMING AND ALIGNMENT SUMMARY

Each individual sequence read was scanned for regions of low quality sequence, defined as having a PHRED quality score $\leqslant$15, which corresponds to an estimated error rate of $\leqslant$3%. Table 2 shows the updated statistics for total data and number of reads per sample after quality trimming. One thing to notice is that the total number of reads can sometimes increase as quality trimming can result in a single read being split into to reads, for example when a low quality region is removed from the middle of a read.

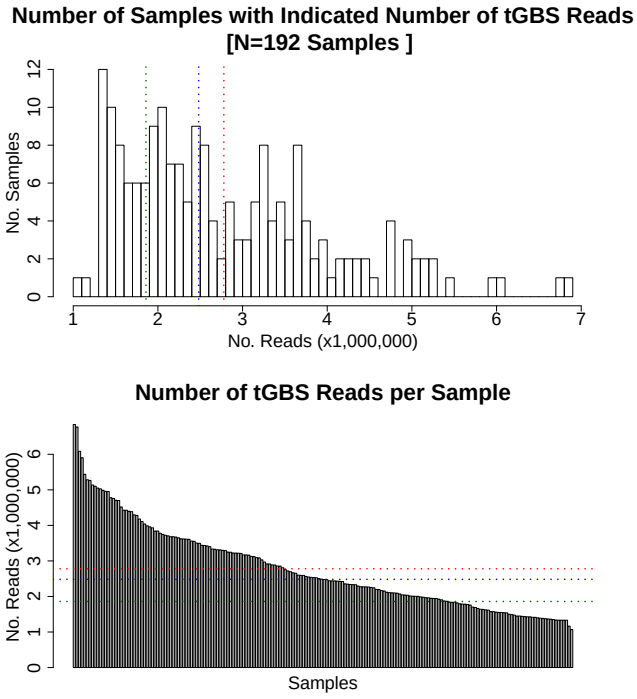

Figure 2 tGBS® Reads per Sample

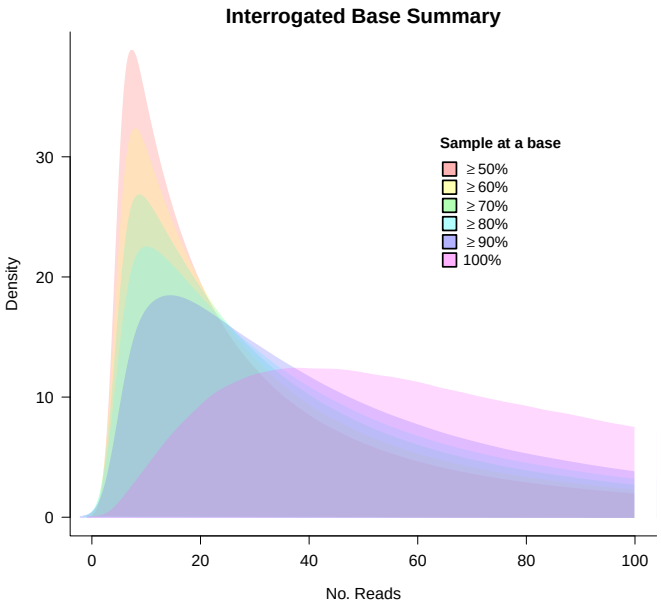

Figure 3 Read counts per interrogated base per sample

Quality trimmed sequence reads were aligned to the reference genome using GSNAP (WU AND NACU 2010). A summary of total, average, median numbers of reads that aligned (uniquely and non-uniquely) are provided in the last two columns of Table 3. However, only reads with a single unique alignment were used in subsequent analyses.

INTERROGATED BASES SUMMARY

An important thing to remember when comparing marker numbers between tGBS® and other genotyping technologies, such as microarrays, is that many genotyping technologies report the number of markers assayed, many of which will be nonpolymorphic in any given population. All 18,037 SNPs identified in this project are segregating among the samples genotyped. In order to provide a more informative basis for comparison to other genotyping technologies, Freedom Markers also reports the number of positions in the genome which were sequenced to sufficient depth in a sufficiently large number of samples that a SNP marker would have been detected and genotyped if a segregating polymorphism were present at that location in the genome among the samples genotyped as part of this project. Table 4 suammarizes these data as well as the median depth of coverage per interrogated site. The distribution of the number of aligned reads per interrogated site per sample is displayed in Figure 3 for the set of sites which were sequenced to sufficient depth to be genotyped in at least 50%, 60%, 70%, 80%, 90% or 100% of samples.

Table 4 Summary of interrogated bases in the indicated fraction of samples

Minimum % Samples Interrogated Per Base	Number Interrogated Bases	Percent Missing Data	Reads per Interrogated Base per Sample		
			25% *	Average	Median
⩾50% (MCR50)	2,792,056	21.9%	12	53	25
⩾60% (MCR60)	2,273,239	16.6%	14	59	29
⩾70% (MCR70)	1,797,933	11.7%	16	66	34
⩾80% (MCR80)	1,272,387	6.3%	19	76	40
⩾90% (MCR90)	898,595	2.4%	23	87	48
100% (MCR100)	318,649	0.0%	50	146	94

* 25% percentile

POLYMORPHIC SITES DISCOVERY AND tGBS® GENOTYPING

Freedom Markers generated several different SNP datasets as part of tGBS projects. The first set (”polymorphic sites”) included all sites that differ from the reference in at least one sample. This set was obtained after considering all reads that align to the reference genome (or consensus sequences if a reference genome is not available). We then examined, sample-by-sample, only those tGBS reads that meet certain tGBS genotyping criteria were used for the following ALL SNPs calling analyses.

Discovery of Polymorphic Sites

The coordinates of confident and single (unique) alignments that passed our filtering criteria were used for SNP discovery. Polymorphisms at each potential SNP site were carefully examined and putative homozygous and heterozygous SNPs were identified in each sample separately using the following criteria:

Homozygous SNPs Criteria:

- The most common allele must be supported by at least 80% of all the aligned reads covering that position.

Table 2 Trimming Summary

	Processed Reads			Quality Trimmed Reads		
	No. Reads	Base Pairs (bp)	Length (bp)	No. Reads	Base Pairs (bp)	Length (bp)
SUM	533,580,676	58,665,326,213	110	451,744,217 (84.7%)	50,921,303,922 (86.8%)	113
AVERAGE	2,779,066	305,548,574	110	2,352,834 (84.7%)	265,215,124 (86.8%)	113
MEDIAN	2,481,882	270,676,393	111	2,127,978 (85.7%)	237,714,380 (87.8%)	113

Table 3 Alignment Summary

	Quality Trimmed Reads			Alignment to Reference Genome	
	No. Reads	Base Pairs (bp)	Length (bp)	Alignments (⩾1 Location)	Unique Alignments (Single Location)
SUM	451,744,217	50,921,303,922	113	386,964,894 (85.7%)	254,876,608 (56.4%)
AVERAGE	2,352,834	265,215,124	113	2,015,442 (85.7%)	1,327,482 (56.4%)
MEDIAN	2,127,978	237,714,380	113	1,838,641 (86.4%)	1,196,903 (56.2%)

- At least 5 unique reads must support the most common allele.
- Polymorphisms in the first and last 3 bp of each read were ignored.
- Each polymorphic base must have at least a PHRED base quality value of 20 (⩽1% error rate).

Heterozygous SNPs Criteria:

- Each of the two most common alleles must be supported by at least 30% of all aligned reads covering that position.
- At least 5 unique reads must support each of the two most-common alleles.
- The sum of reads of the two most common alleles must account for at least 80% of all aligned reads covering that nucleotide position.
- Polymorphisms in the first and last 3 bp of each quality-trimmed read were ignored.
- Each polymorphic base must have at least a PHRED base quality value of 20 (⩽1% error rate).

Criteria for tGBS® Genotyping

Homozygous Call A SNP site was called as homozygous in a given diploid sample if at least 5 reads supported the major common allele at that site and at least 90% of all aligned reads covering that site shared the same nucleotide at that site.

Heterozygous Call A SNP was called as heterozygous in a given diploid sample if at least 2 read supported each of at least two different alleles and each of the two allele types separately comprised more than 20% of the reads aligning to that site. And when the sum of the reads supporting those two alleles at least equal to 5 and comprised at least 90% of all reads covering the site.

ALL SNPS

Freedom Markers filtered the SNPs sites that meet the tGBS® genotyping criteria further to obtain a subset of SNPs that was defined as ALL SNPs. The resulting ALL SNPs set contains 18,037 SNPs.

Filtering Criteria for ALL SNPs

- Minimum calling rate ⩾ 80%
- Allele number = 2
- Number of genotype ⩾ 2
- Minor allele Frequency ⩾ 10%
- Heterozygosity rate range: 30%-70%

Characteristics of the ALL SNPs Dataset

Distributions of various characteristics for the ALL SNPs dataset, including quantity of missing data, minor allele frequency, heterozygosity and genotype number are summarized in Figure 4.

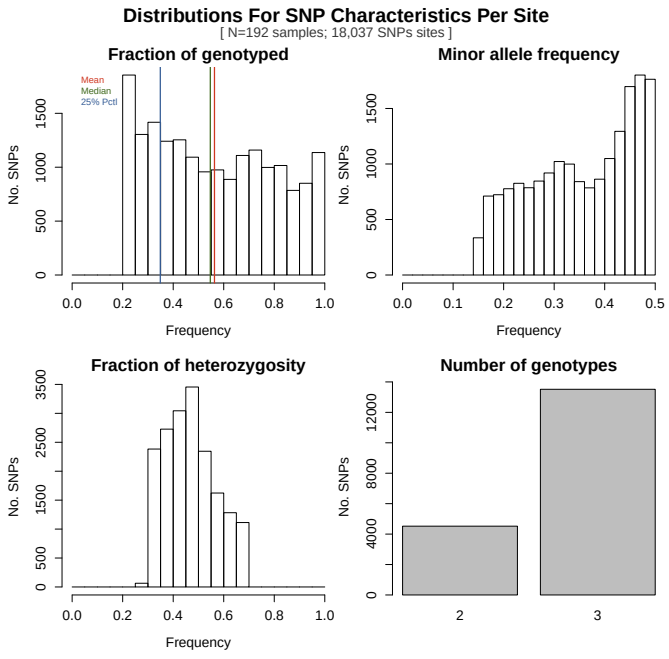

Figure 4 ALL SNPs Summary

The numbers of ALL SNPs per sample that are homozygous for the "REF" (reference) allele, homozygous for the

"ALT" (alternative) allele, heterozygous and missing are shown in the top panel of Figure 5. To allow for comparisons among samples unbiased by varying levels of missing data among samples, the bottom panel of Figure 5 shows the proportions of the SNPs per sample that are homozygous for the REF allele, homozygous for the ALT allele, or heterozygous among the non-missing data.

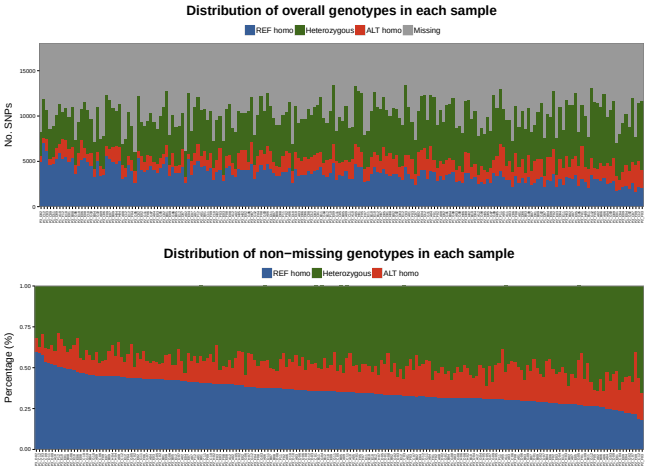

Figure 5 Distribution of genotypes of the ALL SNPs among samples

Missing Data Rate and Heterozygosity

The missing rate of SNPs in each sample was plotted against the fraction of SNPs that were heterozygous. Distributions of missing data rate and heterozygosity of these samples for ALL SNPs are shown in Figure 6.

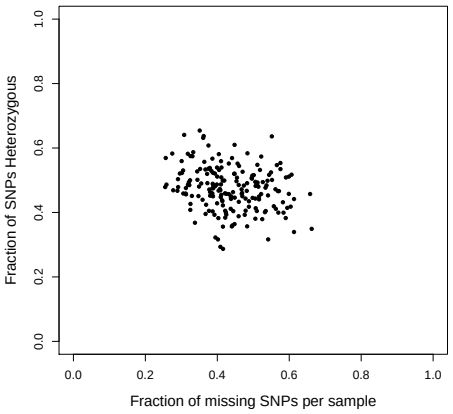

Figure 6 Missing Rate and Heterozygosity Summary per Sample – ALL SNPs

Average Missing Rate per SNP Site and Read Counts per SNP Site per Sample

The average missing data rate per SNP site across samples is provided in Figure 7. Figure 7 also presents the minimum, maximum, average and median numbers of reads per SNP site per sample. Only samples with data were considered here.

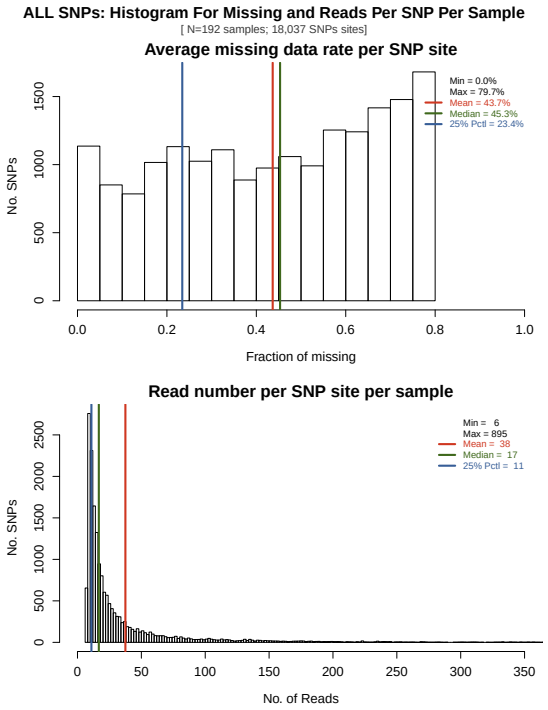

Figure 7 ALL SNPs: Average Missing Rate per SNP site and Read Counts per SNP Site per Sample

EXPLORATION OF DIFFERENT MISSING DATA RATES USED FOR GENOTYPING

Subsequently, Freedom Markers filtered the ALL SNPs set to explore different missing data rates used for genotyping across the samples that have reasonable missing rate and heterozygosity. The resulting number of SNPs remaining, missing data points and detected polymorphism rate for each MCR level are displayed in Table 5.

Table 5 SNPs Gentoyping Summary

	No. SNPs	Missing Data Points	% Polymorphisms *
MCR50	8,434	392,672 / 1,619,328 = 24.25%	8,434 / 2,792,056 = 0.3%
MCR60	6,831	253,780 / 1,311,552 = 19.35%	6,831 / 2,273,239 = 0.3%
MCR70	5,216	146,101 / 1,001,472 = 14.59%	5,216 / 1,797,933 = 0.29%
MCR80	3,477	62,335 / 667,584 = 9.34%	3,477 / 1,272,387 = 0.27%
MCR90	1,914	16,533 / 367,488 = 4.5%	1,914 / 898,595 = 0.21%

* Raw reads without any further processing

MCR50 SNPS

Subsequently, additional cut-offs are applied (e.g., a minimum percentage of call data, minimum and/or maximum heterozygosity rates, and/or minimum minor allele frequencies) to improve the utility of the selected sets of SNPs. These cut-offs are customized to identify a SNP set that best meets project needs. For example, defining an acceptable minimum call rate (MCR) per SNP across samples depends on project goals. In this project we used an MCR of $\geqslant$ 50% as the appropriate cut-off. This cut-off is not, however, cast in stone. One could easily define a set of SNPs for a given set of lines from the "ALL SNPs" table with a different missing data cut-off depending on project needs. Finally, 8,434 MCR50 SNPs were identified.

Filtering Criteria for MCR50 SNPs

- Minimum calling rate $\geqslant$ 50%
- Allele number = 2

- Number of genotype $\geqslant 2$
- Minor allele Frequency $\geqslant 10\%$
- Heterozygosity rate range: 35%-65%

Characteristics of the MCR50 SNP Dataset

Various characteristics of the MCR50 SNP dataset, including quantity of missing data, minor allele frequency, heterozygosity and genotype number are summarized in Figure 8. The numbers of MCR50 SNPs per sample that are homozygous for the REF allele, homozygous for the ALT allele, heterozygous and missing are shown in top panel of Figure 9, and bottom panel shows the proportions of the SNPs per sample that are homozygous for the REF allele, homozygous for the ALT allele, or heterozygous among the non-missing data.

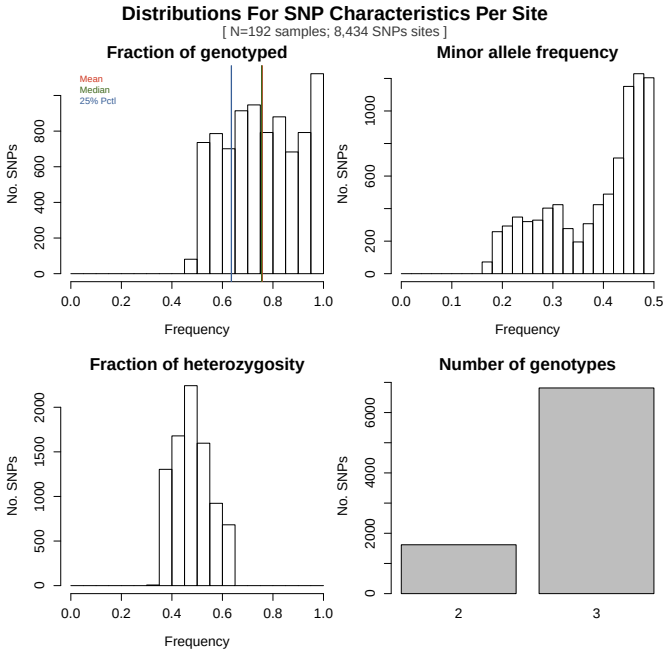

Figure 8 MCR50 SNPs Summary

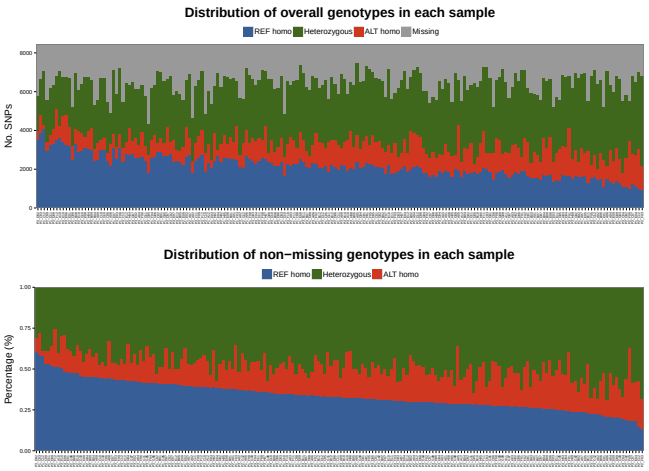

Figure 9 Distribution of genotypes of the MCR50 SNPs among samples

Missing Data Rate and Heterozygosity

The missing rate of SNPs in each sample was plotted against the fraction of SNPs that were heterozygous. Distributions of missing data rate and heterozygosity of these samples for MCR50 SNPs are shown in Figure 10.

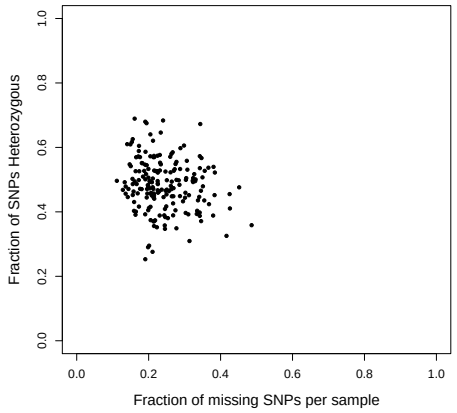

Figure 10 Missing Rate and Heterozygosity Summary per Sample – MCR50 SNPs

Average Missing Rate per SNP Site and Read Counts per SNP Site per Sample

The average missing data rate per SNP site across samples is provided in Figure 11. Figure 11 presents the minimum, maximum, average and median numbers of reads per SNP site per sample. Only samples with data were considered here.

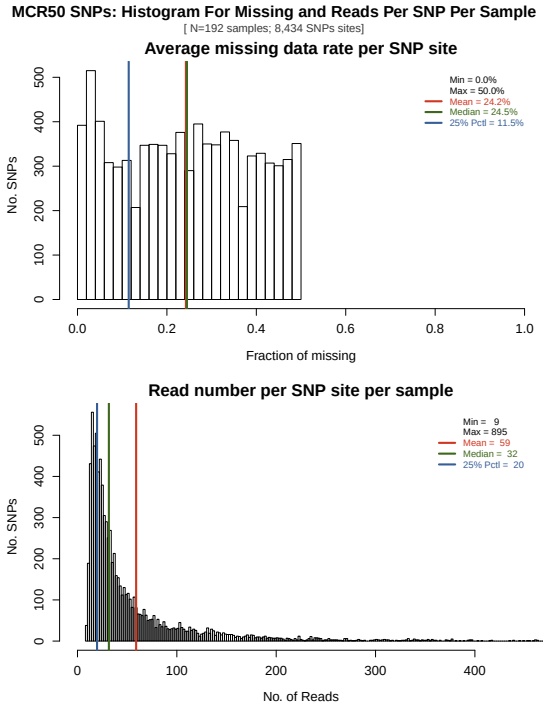

Figure 11 MCR50 SNPs: Average Missing Rate per SNP site and Read Counts per SNP Site per Sample

OUTPUTS OF ANALYSES

The numbers of reads obtained for each sample and genotyping calls are provided in tables provided along with this report. Listed below are descriptions of provided files in folder tables:

- **Maize.all.snps.genotype.txt**: This much larger table contains all the SNPs identified by Freedom Markers in the population and includes all the SNPs reported in the low missing data tables as well as many SNPs, which were genotyped in only a subset of the client's samples. The SNPs, genotype calls, and samples in this file were also formatted into Variant Calling Format (VCF) format and its included under the same directory with file extensions ending with *.vcf.
- **Maize.all.snps.Context201bp.txt**: Context sequence with at most 100 bp upstream and downstream of each of the filtered ALL SNPs sites.
- **Maize.MCR50.snps.genotype.txt**: This file contains markers for which data are available for at least 50% of the lines. This data set is a subset of ALL SNPs. The SNPs, genotype calls, and samples in this file were also formatted into Variant Calling Format (VCF) format and its included under the same directory with file extensions ending with *.vcf.
- **Maize.MCR50.snps.Context201bp.txt**: Context sequence with at most 100 bp upstream and downstream of each of the filtered LMD50 SNPs sites.
- **Maize.*.AlleleCounts**: Read counts per allele of each sample for each of the filtered SNPs sites.

We have also provided DNA sequence reads, after trimming off proprietary sequences that are added during tGBS® library preparation for each of the samples.

- **raw**: Sequencing reads generated by Freedom Markers after the removal of proprietary sequences.
- **trimmed**: Context sequence with at most 200 bp upstream and downstream of each of the filtered ALL SNPs sites.
- **genome**: generated assembled sequences in fasta format.
- **alignment.BAM**: Binary Alignment/Map format (*.bam) files required to visualize the alignments of your sequence data to the reference genome using the Integrated Genome Viewer (IGV)
- **alignment.SAM**: Uniquely aligned reads in SAM output format (http://samtools.sourceforge.net/)
- **figures**: All figures presented in the slide deck.

METHODS

tGBS®

Genomic DNA is digested with two restriction enzymes: NspI leaves a 3'overhang and BfuCI leaves a 5' overhang. Ligation. Two single-strand oligos are ligated to the complementary 3' and 5' overhangs . The oligo matching the 3' overhang contains a sample-specific internal barcode sequence for sample identification. The oligo matching the 5' overhang is universal and present in every reaction for later amplification. Selective PCR. Target sites are selected using a selective primer with variable selective bases ("CA") that match selective sites in the digested genome fragments and a non-selective primer. When properly amplified, the selective site is complementary to the selective bases. Final PCR. Primers matching the amplification primer and the selective primer which contain the full Proton adapter sequence are used for amplification of the final library. Final on-target sequence. The final sequence contains the 5' Proton adapter sequence, an internal barcode, the NspI restriction enzyme site, the target molecule, selective bases, the BfuCI restriction enzyme site and the 3' Proton adapter sequence. (Figure 12) (OTT ET AL. 2017).

Trimming of Sequencing Reads

Prior to alignment, the nucleotides of each raw read were scanned for low quality bases. Bases with PHRED quality

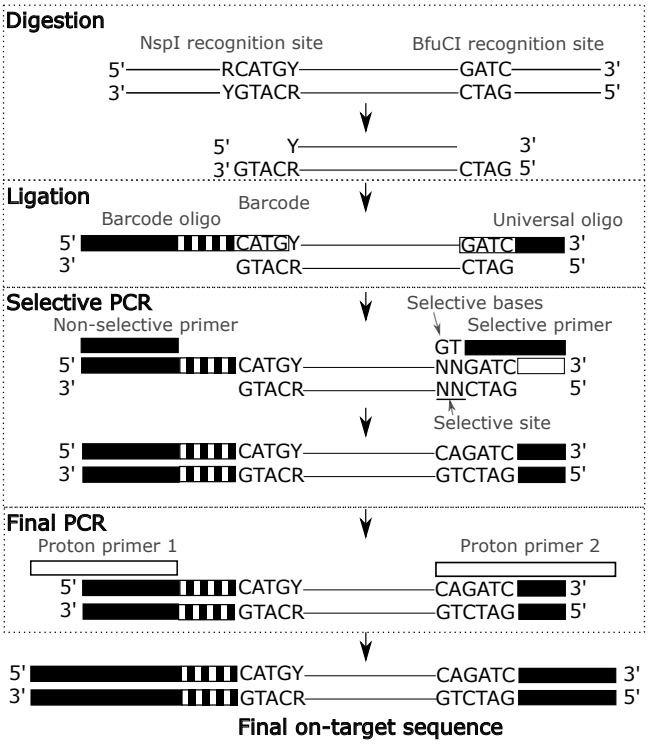

Figure 12 Diagram of tGBS. Digestion

value <15 (out of 40) (EWING ET AL. 1998; EWING AND GREEN 1998), i.e., error rates of ⩽3%, were removed by our trimming pipeline. Each read was examined in two phases. In the first phase reads were scanned starting at each end and nucleotides with quality values lower than the threshold were removed. The remaining nucleotides were then scanned using overlapping windows of 10 bp and sequences beyond the last window with average quality value less than the specified threshold were truncated. The trimming parameters were referred to the trimming software, Lucy (LI AND CHOU 2004).

Alignment of Reads to Reference Genome

Trimmed reads were aligned to the reference genome using GSNAP (WU AND NACU 2010)and confidently mapped reads were filtered if it mapped uniquely (⩽2 mismatches every 36 bp and less than 5 bases for every 75 bp as tails) and used for subsequent analyses.

Data Visualization

Freedom Markers provides a Youtube video to explain how to use IGV (THORVALDSDÓTTIR ET AL. 2013; ROBINSON ET AL. 2011)for data visualization (**Download IGV**). Files for IGV visualization are included under the 'BAM' folder in your Freedom Markersoutput:

Brief instructions for IGV visualization:

- **http://www.youtube.com/watch?v=tOV47_ogPWY**
- Install IGV
- Start IGV visualization program and select your reference genome from the drop-down box located on the top-left corner of the IGV window
- If your reference genome is not available, create a custom genome from the provided FASTA sequences file by going to File Import genome (refer to http://tinyurl.com/4kgjnms for detailed instructions)
- Copy aligment files (*.bam and *.bam.bai from your Freedom Markers output) to your local computer

- Make sure you save the *bam and *bam.bai files under the same folder
- Load *.bam files in the IGV. Now you can visualize the alignment results

REFERENCES

Thomas D. Wu and Serban Nacu. Fast and snp-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7):873–881, 2010. doi: 10.1093/bioinformatics/btq057. URL http://bioinformatics.oxfordjournals.org/content/26/7/873.abstract.

Alina Ott, Sanzhen Liu, James C Schnable, Cheng-Ting Yeh, Cassy Wang, and Patrick S Schnable. Tunable genotyping-by-sequencing (tgbs®) enables reliable genotyping of heterozygous loci. *bioRxiv*, page 100461, 2017.

Brent Ewing, LaDeana Hillier, Michael C. Wendl, and Phil Green. Base-calling of automated sequencer traces using-phred. i. accuracy assessment. *Genome Research*, 8(3):175–185, 1998. doi: 10.1101/gr.8.3.175. URL http://genome.cshlp.org/content/8/3/175.abstract.

Brent Ewing and Phil Green. Base-calling of automated sequencer traces using phred. ii. error probabilities. *Genome Research*, 8(3):186–194, 1998. doi: 10.1101/gr.8.3.186. URL http://genome.cshlp.org/content/8/3/186.abstract.

Song Li and Hui-Hsien Chou. Lucy2: an interactive dna sequence quality trimming and vector removal tool. *Bioinformatics*, 20(16):2865–2866, 2004. doi: 10.1093/bioinformatics/bth302. URL http://bioinformatics.oxfordjournals.org/content/20/16/2865.abstract.

Helga Thorvaldsdóttir, James T. Robinson, and Jill P. Mesirov. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2):178–192, 2013. doi: 10.1093/bib/bbs017. URL http://bib.oxfordjournals.org/content/14/2/178.abstract.

James T Robinson, Helga Thorvaldsdottir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. Integrative genomics viewer. *Nat Biotech*, 29(1):24–26, 01 2011. URL http://dx.doi.org/10.1038/nbt.1754.