*Published online 25 September 2017*

*Nucleic Acids Research, 2017, Vol. 45, No. 21* ***e178***
*doi: 10.1093/nar/gkx853*

# tGBS® genotyping-by-sequencing enables reliable genotyping of heterozygous loci

**Alina Ott[1,†], Sanzhen Liu[1,2,3,*,†], James C. Schnable[3,4], Cheng-Ting 'Eddy' Yeh[1,3], Kai-Sin Wang[3] and Patrick S. Schnable[1,3,*]**

[1]Department of Agronomy, Iowa State University, Ames, IA 50011-3650, USA, [2]Department of Plant Pathology, Kansas State University, Manhattan, KS 66506, USA, [3]Data2Bio LLC, Ames, IA 50011-3650, USA and [4]Department of Agriculture and Horticulture, University of Nebraska-Lincoln, Lincoln, NE 68588, USA

Received May 22, 2017; Revised August 09, 2017; Editorial Decision September 12, 2017; Accepted September 13, 2017

## ABSTRACT

**Conventional genotyping-by-sequencing (cGBS) strategies suffer from high rates of missing data and genotyping errors, particularly at heterozygous sites. tGBS® genotyping-by-sequencing is a novel method of genome reduction that employs two restriction enzymes to generate overhangs in opposite orientations to which (single-strand) oligos rather than (double-stranded) adaptors are ligated. This strategy ensures that only double-digested fragments are amplified and sequenced. The use of oligos avoids the necessity of preparing adaptors and the problems associated with inter-adaptor annealing/ligation. Hence, the tGBS protocol simplifies the preparation of high-quality GBS sequencing libraries. During polymerase chain reaction (PCR) amplification, selective nucleotides included at the 3′-end of the PCR primers result in additional genome reduction as compared to cGBS. By adjusting the number of selective bases, different numbers of genomic sites are targeted for sequencing. Therefore, for equivalent amounts of sequencing, more reads per site are available for SNP calling. Hence, as compared to cGBS, tGBS delivers higher SNP calling accuracy (>97–99%), even at heterozygous sites, less missing data per marker across a population of samples, and an enhanced ability to genotype rare alleles. tGBS is particularly well suited for genomic selection, which often requires the ability to genotype populations of individuals that are heterozygous at many loci.**

## INTRODUCTION

A fundamental goal of biology is to link variation in genotype with variation in phenotype. Achieving this goal requires accurate methods for measuring both genotypes and phenotypes. The development of polymerase chain reaction (PCR) made feasible assays of genotypic variation between individuals on a scale never before achieved (1). The introduction of fluorescent dyes and hybridization technology have enhanced the reliability, improved the sensitivity and increased the throughput of genotyping assays (2–4). In the last decade, advances in DNA sequencing technologies and substantial cost reduction have made it possible to genotype individual organisms via sequencing (5,6). Genotyping using sequence data can incorporate marker discovery and marker scoring into a single process, reducing the ascertainment bias inherent in many other PCR- or hybridization-based genotyping approaches which are designed to score a pre-defined set of markers.

The most comprehensive form of genotyping using sequence data is complete resequencing of the genomes of individuals of interest at sufficient depth to identify polymorphisms. However, for many eukaryotic species this approach is still cost prohibitive given their genome sizes. Various genome reduction strategies have been developed to target only a subset of an organism's genome for sequencing, thereby reducing the total amount of sequence data needed per individual. The most common genome reduction approach is to sequence genomic loci flanked by restriction enzymes (REs). Other methods substitute amplification for enzymatic digestion, e.g. SLAF-seq (7) and NextRAD (8).

One well-known next generation sequencing-based genotyping strategy that utilizes REs as a method of genome reduction is RAD-Seq (9). While RAD-Seq and related methods such as CRoPS (10), MGS (11), GBS (12), double digest RADseq (13), 2b-RAD (14) and RESTSeq (15) repre-

*To whom correspondence should be addressed. Tel: +1 515 294 0975; Fax: +1 515 294 5256; Email: schnable@iastate.edu
Correspondence may also be addressed to Sanzhen Liu. Tel: +1 785 532 1379; Fax: +1 785 532 5692; Email: liu3zhen@ksu.edu
†These authors contributed equally to the paper as first authors.

**Disclaimer:** Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

sented a significant advance in reducing cost and increasing throughput relative to whole genome resequencing, the initial protocols included labor intensive and costly steps such as physical shearing of DNA molecules and enzymatic end repair to process DNA. A number of modified protocols focused on increasing the stringency of genome reduction. Even so, current methods often still target hundreds of thousands to millions of sites per genome. As a result, given a reasonable amount of sequencing, read depths per site are often quite low, resulting in any given site not being sequenced in a subset of individuals, leading to high levels of missing data and problems detecting rare alleles (16). Low read depths also result in higher error rates especially at heterozygous loci where low numbers of aligned reads increase the risk that only one of the two alleles present will be represented (17). This limits the use of these methods primarily to inbred lines, or requires more sequencing per individual to increase read depths, thereby reducing the advantages gained from genome reduction.

In practice, the ideal level of genome reduction varies depending on the size of the target genome, the nature of the population being sequenced, the prevalence of polymorphic loci in a population, and the research goals. Ascertaining phylogenetic relationships can often be achieved using only a few hundred markers. Mapping QTLs within an $F_2$ or RIL population will generally benefit from genotyping several thousand markers. Genome-wide association studies (GWAS) may require anywhere from tens of thousands to millions of markers depending on the level of linkage disequilibrium in the population. In principle each of these needs could be addressed by separate genome reduction methods. However, such an approach would mean very few markers would be shared across different datasets generated for different initial aims, limiting interoperability and data reusability.

Here, we describe a new method of genotyping-by-sequencing. This method provides the ability to adjust the number of targeted sites according to research goals by modifying a single primer in the protocol. In addition, unlike the genome reduction methods described above, this method utilizes oligonucleotides (oligos, which by definition consist of a single strand of DNA) in place of adaptors, which by definition consist of two annealed olgios and are therefore always double-stranded. The use of adaptors requires the careful annealing of the two oligos prior to use, and accurate quantification to obtain the proper ratio of adaptors to DNA templates to achieve efficient ligation. In contrast, the preparation and accurate quantification of oligos are simple, substantially enhancing the reliability of tGBS library preparation.

Our results demonstrate that sequencing reads from tGBS libraries are highly enriched at target sites and produce higher average read depths per target site given the same number of reads per sample employed by conventional genotyping-by-sequencing (cGBS) strategies. As a consequence of the high average read depth per site, a low fraction of missing data and high repeatability in single nucleotide polymorphism (SNP) calls among individuals is obtained, avoiding the need for extensive imputation. Finally, tGBS exhibits high accuracy in genotyping both homozygous (>97%) and heterozygous (>98%) loci, which makes it possible to accurately genotype non-inbred populations such as $F_1BC_1$s and $F_2$s which are widely used in both genetic research and selective breeding programs, including those involving genomic selection (18).

## MATERIALS AND METHODS

### Extraction of DNA samples

DNA samples from the inbred lines B73, Mo17 and the nested association mapping (NAM) founders (19) were extracted from 6-day-old seedling tissue using the DNeasy Plant Maxi Kit [QIAGEN (Valencia, CA, USA), No. 68163]. The 232 B73xMo17 recombinant inbred lines (IBM RILs) (20) and the 192 $F_2$ individuals were extracted from 6-day-old seedling leaf tissue using the MagAttract 96 DNA Plant Core Kit [QIAGEN (Valencia, CA, USA), No. 67163]. Samples were normalized using the Qubit dsDNA Broad Range Assay [ThermoFisher (Waltham, MA, USA), no Q32853].

### tGBS procedure

Approximately 120 ng of genomic DNA from each sample was digested with 100 units of NspI [New England Biolabs (Beverly, MA, USA), No. R0602L] and 400 units of BfuCI [New England Biolabs (Beverly, MA, USA), No. R0636L] in NEB CutSmart Buffer 4 in a 30-μl volume at 37°C for 1.5 h following the manufacturer's protocol. Unique, barcoded oligos (100 μM) and a universal single-strand oligo (100 μM) were added to each sample for ligation with T4 DNA ligase [New England Biolabs (Beverly, MA, USA), No. R0602L]. Ligation was performed at 16°C for 1.5 h in a 60 μl volume following the manufacturer's protocol. The T4 DNA ligase was inactivated by incubation at 80°C for 20 min. All digestion-ligation products were pooled and 1 ml of pooled product was purified using the QiaQuick PCR purification kit [QIAGEN (Valencia, CA, USA), No. 28106]. The pooled, purified digestion-ligation product was used as the template for a single selective PCR reaction using a selective primer (100 μM), an amplification primer (100 μM) and Phusion High-Fidelity PCR Master Mix with HF Buffer [New England Biolabs (Beverly, MA, USA), No. M0531L]. The PCR program consisted of 95°C for 3 min; 15 cycles of 98°C for 15 s, 65°C for 20 s, 72°C for 20 s; and a final extension at 72°C for 5 min. The selective PCR product was purified using a 1:1 ratio of Agencourt AMPure XP Beads [Beckman Coulter, Inc. (Brea, CA, USA), No. A63880]. The purified selective PCR product was used as the template for a single, final PCR reaction using primers for the Proton platform and Phusion High-Fidelity PCR Master Mix with HF Buffer [New England Biolabs (Beverly, MA, USA), No. M0531L]. The PCR program consisted of 98°C for 3 min; 10 cycles of 95°C for 15 s, 65°C for 20 s, 72°C for 20 s; and a final extension at 72°C for 5 min. The final PCR product was purified using Agencourt AMPure XP Beads [Beckman Coulter, Inc. (Brea, CA, USA), No. A63880]. The purified final PCR product underwent size selection for a target of 200–300 bp using the 1.5% Agarose DNA cassette for the BluePippin [Sage Science (Beverly, MA, USA), No. HTC2010]. The size-selected

final PCR product was run on a Bioanalyzer High Sensitivity DNA chip to quantify and ensure proper size selection [Agilent Technologies (Santa Clara, CA, USA), No. 5067–4626]. Oligo and primer sequences for the Proton and Illumina sequencing platforms are provided in Supplementary Tables S1 and 2, respectively.

### Sequencing on the Ion proton

tGBS libraries were sequenced on Life Technologies' Ion Proton Systems following the Ion PI Hi-Q Sequencing 200 Kit User Guide (Revision C.0) at Iowa State University's Genomics Technologies Facility. Template preparation was performed with the Ion PI Hi-Q OT2 200 Kit [Thermo Fisher (Waltham, MA, USA), No. A26434] on the Ion Onetouch 2 System. Sequencing runs were performed using the Ion PI Hi-Q Sequencing 200 Kit [Thermo Fisher (Waltham, MA, USA), No. A26433] and the Ion PI Chip Kit v3 [Thermo Fisher (Waltham, MA, USA), No. A26771] at 300 flows.

### Debarcoding and cleaning of tGBS reads

Sequencing reads were analyzed with a custom Perl script (available at https://github.com/orgs/schnablelab) which assigned each read to a sample and removed the associated barcode. Each debarcoded read was further trimmed to remove Proton adaptor sequences using Seqclean (sourceforge.net/projects/seqclean) and to remove potentially chimeric reads harboring internal restriction sites of NspI or BfuCI. Only reads with the correct barcodes and RE sites were retained for further processing. Retained reads were subjected to quality trimming. Bases with PHRED quality value $<15$ (out of 40) (21,22), i.e. error rates of $\leq 3\%$, were further removed with another custom Perl script. 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 as referred to in the trimming software, Lucy (23,24).

### Alignment of reads to reference genome

Cleaned reads were aligned to the B73 reference genome (AGP v2) (25) using GSNAP (26). Only confidently mapped reads were used for subsequent analyses, which are uniquely mapped with at least 50 bp aligned, at most two mismatches every 40 bp and tail of $<3$ bp for every 100 bp of read.

### SNP discovery

The resulting confident alignments were used for SNP discovery. Reads at each potential SNP site were counted. A site was considered interrogated if it was covered by at least five reads. At each interrogated site, each sample was genotyped individually using the following criteria: an SNP was called as homozygous in a given sample if at least five reads supported the genotype at that site and at least 90% of all aligned reads covering that site shared the same nucleotide; an SNP was called as heterozygous in a given sample if at least two reads supported each of at least two different alleles, each of the two read types separately comprised more than 20% of the reads aligning to that site, and the sum of the reads supporting those two alleles comprised at least 90% of all reads covering the site. To compare samples with equal data, SNP discovery was performed in subsets of the data where equal numbers of randomly selected trimmed reads were processed from each sample individually.

### Determination of selectivity

Sequencing reads obtained from Life Technology's Proton instrument are single-end and only include the barcode, NspI digestion site and the adjacent sequence. For this reason, selectivity could not be directly determined from reads. Based on the closest BfuCI site of uniquely aligned reads in the B73 genome, the complementary bases that target the selective sequences in each read were predicted. On-target and off-target reads were categorized based on this selected sequence prediction. 'On-target sites' in the genome are defined as those that contain both an NspI RE recognition site and a BfuCI RE recognition site which are separated by 100–300 bp and that contain the appropriate selected sequence adjacent to the BfuCI recognition site. 'On-target reads' align to on-target sites. The number of interrogated sites was determined by identifying all the bases in the reference genome that had $\geq 5$ reads uniquely aligned to that site.

*In silico* digestion of the B73 reference genome was performed to identify all possible NspI and BfuCI RE fragments. Reads were aligned to this digested genome to determine which fragments have coverage.

### Accuracy of tGBS calls

Concordance of genotyping calls among methods was used as a proxy for accuracy. The accuracy of tGBS calls in the NAM founders was determined by identifying concordant and non-concordant genotypes between tGBS calls and calls from TASSEL SNPs (27) and RNA-sequencing SNPs (28) (SRA050790 and SRA050451). The HapMap2 TASSEL SNP genotypes from Panzea were used directly, while RNA-sequencing SNPs were called as described for tGBS SNP calling. Polymorphic sites (i.e. at least one of the NAM founders has a non-reference allele) that were in common across the three SNP calling methods were compared. For each sample with no missing data at that site, the genotyping calls from each method were compared. If the call in one method disagreed, then the method in disagreement was considered non-concordant.

To assess accuracy of tGBS SNP calls in the IBM RILs, tGBS SNP calls were compared to genotypes from RNA-sequencing (29) and Sequenom data in the IBM RILs (30). Because the RILs are expected to have low levels of heterozygosity and be segregating $\sim$1:1 for B73-like versus Mo17-like alleles, the tGBS and RNA-sequencing SNPs were filtered independently for sites with minor allele frequencies $>0.3$ and heterozygosity $<0.05$. A total of 67 RILs were genotyped with all three methods and could

be compared for accuracy. To increase the number of sites that could be compared between the tGBS and RNA-sequencing genotyping, segmentation was performed on each set of SNPs to identify B73-like and Mo17-like regions in each RIL. Segments were identified from each SNP set by running DNAcopy (31) using the segment function with the parameters $\alpha = 0.01$, nperm = 10000, p.method = 'perm', eta = 0.01 and min.width = 3. A segment genotype was determined by identifying which genotype was the majority in the given segment. The SNP genotyping calls from the each filtered SNP set were compared to the segmentation genotype from each method. Each putative error was examined to determine the genotypes of flanking markers. If the genotype of the putative error agreed with at least one of the flanking markers, the marker was no longer considered an error. Individual SNPs that did not match the segment genotype and had no flanking markers that would indicate the segment was generated incorrectly were considered errors.

The accuracy of tGBS calls conducted in the B73 × Mo17 $F_2$ individuals was also determined by using segmentation. tGBS was performed on 192 $F_2$ individuals at tGBS (GRL2). Because an $F_2$ population is expected to segregate ∼1:2:1 at sites that are polymorphic for the two different parental alleles, the 4032 SNP sites with a 70% minimum call rate (MCR; i.e. at least 70% of the samples were genotyped), minor allele frequencies ≥0.35 and a proportion of heterozygous genotypes between 0.35 and 0.65 were used for segmentation. Using the same parameters for DNAcopy described above, segments of similar genotypes were identified in each of the $F_2$ individuals. Within each individual, marker genotypes that did not agree with the segment genotype (reference, heterozygous or non-reference) were flagged as putative errors.

The accuracy of tGBS (GRL2) and cGBS was compared as described above for the NAM concordance where SNP genotypes obtained from three methods are expected to agree and if one genotype does not agree, that genotype call is considered an error. The HapMap2 genotypes from Panzea were used as the third comparison. SNP genotypes from the cGBS data were called in three ways: SNP calling using a method that allows for heterozygous calls (equivalent to previous SNP calling descriptions), SNP calling using a method that allows for only homozygous calls (the most common allele must be supported by at least 80% of all aligned reads instead of at least 30%), or downloaded from Panzea. tGBS SNPs were obtained from either the heterozygous or homozygous SNP calling methods.

### Construction of genetic maps

Genetic maps were constructed from 70% MCR, 50% MCR and 20% MCR tGBS (GRL2) SNP sets in the IBM RILs with the same filtering described for segmentation using ASMap (32). LinkImpute (33) was run with the default settings. SNPs imputed from LinkImpute and unimputed SNPs for each MCR were imported into ASMap (32) for map construction. RILs with high similarity were detected using the comparegeno function. Six RILs (M0122, M0173, M0177, M0187, M0209, M0252) were removed for having >90% similarity with another RIL. Markers with segregation distortion were identified and any markers with a *P*-value < 1e-10 were removed. Genetic maps were constructed using the mstmap.cross function. The *P*-value cutoff for genetic map construction (with and without imputation) was adjusted so that 10 or more distinct linkage groups were identified, and the detection of bad markers was set to 'yes'. The genotyping error of genetically mapped markers was estimated by determining the maximum likelihood from a range of potential errors using R/qtl (34).

Genetic maps were also constructed from 70% MCR tGBS (GRL2) filtered SNP set for 192 B73 × Mo17 $F_2$ individuals using ASMap. Imputation and genetic mapping were performed as described for the IBM RILs but using a more stringent *P*-value (<1e-5) for segregation distortion.

### Comparisons between tGBS and cGBS

cGBS data were downloaded from GenBank SRP021921 (32). Barcodes were removed and reads were trimmed for quality as described above for tGBS reads. However, because tGBS data were generated using Ion Proton technology and cGBS data were generated using Illumina sequencing technology and these technologies produce reads of different lengths, it was necessary to control for read length before conducting comparisons. To compare read depth per interrogated site, tGBS and cGBS were standardized by trimming all reads to 75 bp. The observed reduction in read number from raw to standardized reads is primarily due to the removal of reads <75 bp in length, which were not used in this analysis. These 'standardized reads' were then aligned to the B73 reference genome.

Similarly, to compare the number of interrogated sites at various MCR values standardized tGBS (GRL2) and cGBS reads were subsampled to obtain equal numbers of reads for each method (0.25, 0.5, 0.75 and 1 million reads). Samples that had fewer than the desired number of subsampled reads were not used in this analysis. Because MCR is affected by the number of samples included in the comparison, equal numbers of tGBS and cGBS samples were selected based on the method with the smaller number of samples having the appropriate number of subsampled reads.

## RESULTS

### tGBS for genome reduction

During tGBS, genomic DNA is subjected to double digestion with two enzymes, producing DNA fragments with a 5′ overhang on one end and a 3′ overhang on the other (Figure 1). In contrast to other methods (7–12) that employ adaptors, a unique oligo is ligated to each overhang. This strategy ensures that only double-digested fragments are sequenced, thereby increasing specificity. One of the oligos is unique to an individual sample and contains a DNA barcode (35) (barcode oligo) while the other oligo is common to all samples and contains a universal sequence (universal oligo) for subsequent construction of sequencing libraries. Following ligation, two PCR steps complete the construction of the sequencing library. For the first PCR (selective PCR), two PCR primers that partially match the ligation oligos are used. The primer matching the universal oligo (selective primer) is designed to be the reverse complement

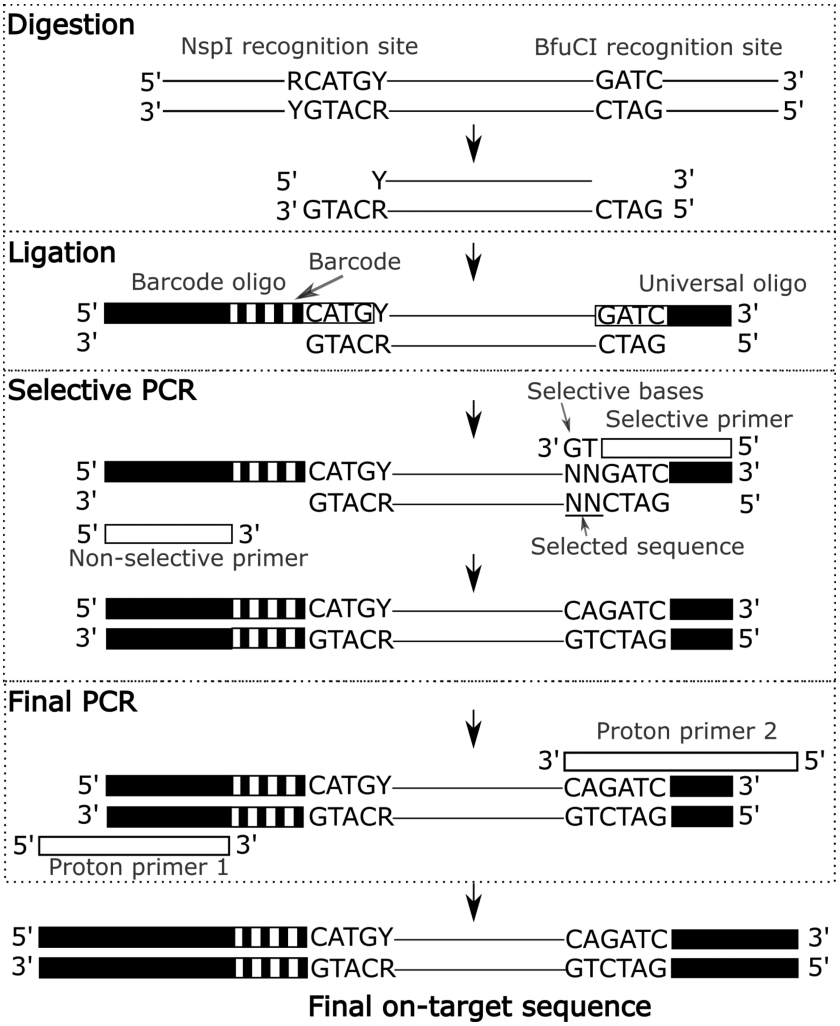


**Figure 1.** Diagram of tGBS. **Digestion**. Genomic DNA is digested with two REs: NspI leaves a 3′overhang and BfuCI leaves a 5′ overhang. **Ligation**. Two distinct oligos are ligated to the complementary 3′ and 5′ overhangs. The oligo matching the 3′ overhang contains a sample-specific internal barcode 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 selected sequences in the digested genome fragments and a non-selective primer. When properly amplified, the selected sequence is complementary to the selective bases. **Final PCR**. Primers matching the amplification primer and the selective primer which contain the full Proton adaptor sequence are used for amplification of the final library. **Final on-target sequence**. The final sequence contains the 5′ Proton adaptor sequence, an internal barcode, the NspI RE site, the target molecule, selective bases, the BfuCI RE site and the 3′ Proton adaptor sequence. It is possible to adapt the tGBS protocol for sequencing on an Illumina instrument by redesigning the ligation oligos and PCR primers.

of the universal ligation oligo; however, it extends an additional 1–3 nt (selective bases) at its 3′ end which can only perfectly anneal to a subset of the genomic fragments created by RE digestion and oligo ligation, thus reducing the number of targeted sites to be amplified. As a result, genomic fragments that include the complement of the selective bases and the universal oligo will be preferentially amplified. The non-selective primer used in selective PCR matches the 5′ end of the barcode oligo. Because this primer will anneal and amplify the sequence preceding the barcode, the primer itself does not need to be designed to match the barcode, reducing primer complexity and cost. For the second PCR (final PCR), two primers (Proton/Illumina primer 1 and 2) compatible with the appropriate sequencing platform are used to create the sequencing library.

Based on their cutting frequencies and abilities to generate appropriate overhangs (one 5′ overhang and one 3′ overhang), NspI and BfuCI were selected for tGBS. Simulation analysis of the maize genome constrained to only non-repetitive DNA-fragments with different cut sites on each end with a total size between 100 and 300 bp yielded a total of 246 124, 44 372 and 8645 non-repetitive DNA fragments for 1-, 2- or 3-bp of selective bases (T, TG and TGT) respectively.

### tGBS strongly selects for reads at target sites

The maize inbreds B73, Mo17 and the 25 parents of the NAM population (19) were genotyped via tGBS using the enzymes NspI and BfuCI and 1, 2 and 3 selective bases (Supplementary Table S3). Each level of selection is named

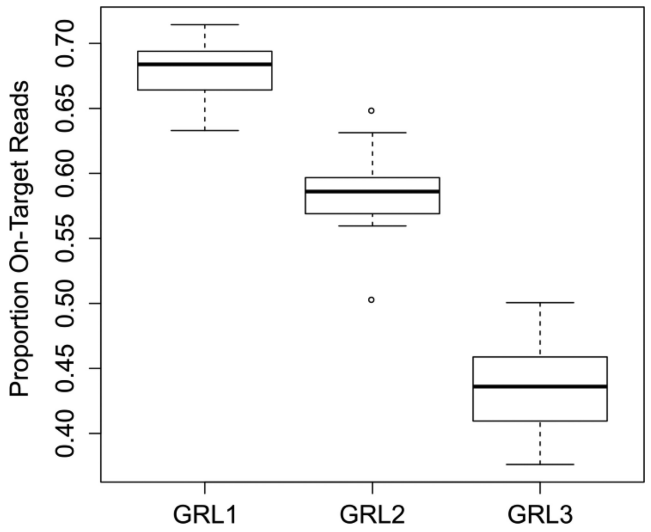


**Figure 2.** Selectivity in B73, Mo17 and the NAM founders. In the absence of selection, the proportion of random reads in the target size range from the B73 genome with 'T', 'TG' and 'TGT' selection in GRL1, GLR2 and GRL3 would be ∼25, ∼6 and ∼2%, respectively.

based on the number of selective bases: e.g. genome reduction level 1 (GRL1) involves a single selective base. An average of 6.4M (GRL1), 8.1M (GRL2) and 6.3M (GRL3) reads were generated per line. These reads were then subjected to quality trimming and aligned to the B73 reference genome.

At all tGBS GRLs, over 90% of the aligned reads contain the expected RE recognition sites. For tGBS (GRL2), the selective primer had the selective bases 'TG' at its 3′ end. Ideally, all amplified reads will be derived from restriction fragment that contain the selected 'AC' sequence. However, mis-annealing of primers during PCR can lead to off-target amplification. To measure the specificity of selection during our PCR protocol, the bases adjacent to the BfuCI restriction recognition site of sequenced reads were examined. Target sites in the genome contain the appropriate RE recognition site adjacent to the selected nucleotides ('AC' in the case of GRL2 'TG' selection); reads that align to such target sites are termed on-target reads. tGBS (GRL1) had the highest percent of on-target reads, with an average of ∼68% of the reads across all samples containing both the RE sites and the correct selective bases based on the B73 genome. For tGBS (GRL2) and tGBS (GRL3) the average percent of on-target reads were 58 and 44%, respectively, across all samples (Figure 2). Note that for each additional selective base, genome-wide the number of on-target sites decreases by ∼1/4 (Supplementary Table S4). Therefore, even though the on-target rate was somewhat lower for tGBS (GRL3) than for tGBS (GRL1) and tGBS (GRL2), the read depth of covered bases at on-target sites was highest for tGBS (GRL3) (Supplementary Table S4). As a consequence of the size selection conducted prior to Proton sequencing, 68% of all uniquely aligning reads (4248, 425/6271,577) and 88% of on-target reads (3569,220/4071,296) were from on-target sites consisting of between 100 and 300 bp.

### Genotyping the founders of the nested association mapping (NAM) population

Genotyping genetically diverse lines such as the NAM founders is important for GWAS and genomic selection (Supplementary Table S6). A MCR cutoff was implemented. At 70% MCR, each SNP must have been genotyped in ≥70% of the samples. Among the 25 NAM founders, 6665 (GRL1), 11 883 (GRL2) and 3253 (GRL3) SNPs were identified at 70% MCR (Supplementary Table S5). These SNPs are distributed relatively evenly across the genome (Figure 3 and Supplementary Figure S1), and the number of reads per SNP site per sample had a mean of 63 and a median of 31 (Supplementary Figure S2.).

The numbers of SNPs discovered in the NAM founders are not directly comparable across tGBS GRLs due to the variation in the average read number per sample (Supplementary Table S7). To overcome this limitation, a subset of NAM founders with comparable minimum numbers of reads were used in the analysis described below. To examine the trade-offs in SNP discovery associated with variation in the amount of sequencing data generated we subsampled the sequencing reads from each of the NAM founders independently. In our dataset 11 of the 25 NAM founders had a sufficient number of reads across all three tGBS GRL to perform comparable subsampling (Supplementary Table S6). From this analysis, the diminishing returns of SNP discovery with increased sequencing can be seen in tGBS (GRL3), which begins to plateau after 3 million raw reads. At tGBS (GRL2), additional sequencing exhibits diminishing returns such that the benefits of additional sequencing begin to level off around 4 million subsampled reads (Supplementary Figure S3). tGBS (GRL1) had not reached the point of diminishing returns, which is expected to be much higher than 4 million reads (Supplementary Figure S3).

The minimum accuracy rate of SNP calling in the 25 NAM founders was determined by calculating the concordance of tGBS SNPs with those derived from HapMap2 (36) and RNA-sequencing data (28) from the same lines. The HapMap2 and RNA-sequencing data were obtained via whole genome resequencing and transcriptome sequencing of five maize tissues for each of the NAM founders, respectively. For this analysis, individual SNPs were compared, therefore an MCR cut-off was not employed. Across the 25 founders, 90 902 (GRL1), 95 028 (GRL2) and 30 051 (GRL3) SNPs were genotyped in all three experiments (Supplementary Table S7). To calculate minimum accuracy rates, if only two of the three experiments yielded a concordant genotyping call at a particular site, the non-concordant call at that site was considered an error. tGBS had >99% concordant calls for all GRL, which was higher than the other two methods (Supplementary Table S7). Note that this approach probably over-estimates genotyping errors because the lack of concordance between methods may be due to biological differences among the different pedigrees of samples used in the three experiments. Hence, the minimum SNP calling accuracy of tGBS as determined in this analysis of inbred lines is >99%.

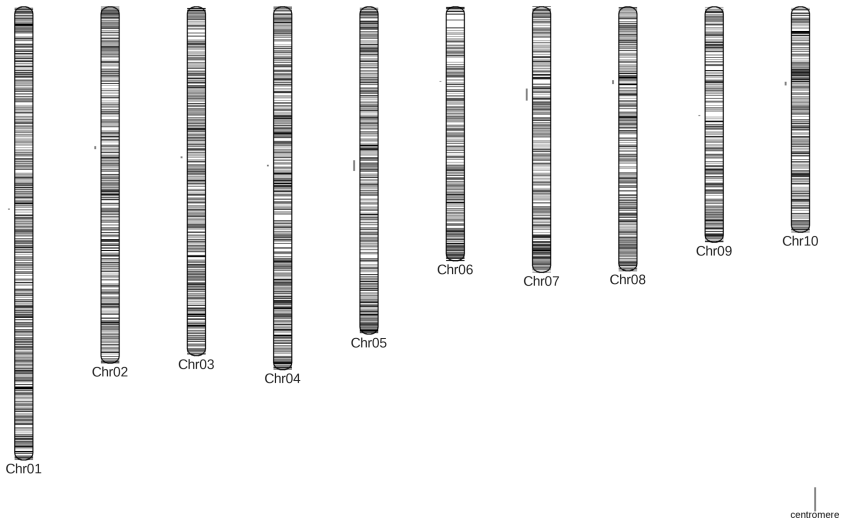


**Figure 3.** Genomic distribution of SNPs discovered in the 25 NAM founders using tGBS (GRL2) at 70% MCR. Each horizontal line represents the physical position of a SNP identified by alignment to the B73 reference genome. The circles to the left of each chromosome represent the location of the centromere.

### Genotyping recombinant inbred lines (RILs) and construction of a genetic map

The IBM RILs were developed by crossing B73 and Mo17 (37). Random mating was performed for several generations before extensive inbreeding (37). tGBS was conducted on 232 IBM RILs (Supplementary Table S8) using tGBS (GRL2). A mean of 2.1 M reads and a median of 1.8 M reads were obtained per sample which is similar to target sequencing read numbers per SNP generally employed by other GBS protocols (12).

The accuracy of the 70% MCR SNP calls was assessed by comparing tGBS SNP calls with Sequenom-based genotyping results (30) and RNA-sequencing (29) for 67 IBM RILs genotyped with all three methods, similar to the comparison performed for the NAM founders. However, unlike the NAM founders, it was possible to use SNP genotypes to subdivide the genome of each RIL into segments, each of which was derived from one of the two RIL parents: B73 or Mo17. This segmentation allowed us to compare all of the SNPs within a segment, rather than only those SNPs that had been genotyped with multiple methods. Thus, this approach enabled us to compare most SNPs (Supplementary Table S9). Another difference in this analysis as compared to the analysis of the NAM founders was that heterozygosity and minor allele frequency filters (based on expected segregation patterns in RILs) were employed to exclude errors due simply to mis-alignment of reads to the genome. Following filtering, each of the three datasets was used to generate segments, which were compared to the original SNP calls used as input data for segmentation. As expected, the agreement between the input data and the segmented data was high. In this analysis tGBS also had a minimum accuracy of 99% (Supplementary Table S9).

Genetic maps were constructed using the tGBS data from the IBM RILs, both with and without SNP imputation at various MCR cutoffs (Figure 4). Based on Spearman rank correlation, marker orders were well conserved between the genetic and physical maps (Supplementary Table S10). At 70% MCR, about 4000 (∼90%) SNPs were mapped using both imputed and non-imputed data. As expected, more SNPs were obtained using more relaxed MCR cut-offs (50 or 20% MCR). At an MCR of only 20%, imputation increased both the number and the percentage of SNPs successfully placed on the genetic map. The generation of ∼10 linkage groups corresponding to the 10 maize chromosomes, the high percentage of markers that were mapped, the extremely low proportion of markers assigned to an incorrect chromosome, the low estimated error rate of markers on the genetic map and the high Spearman correlation values demonstrate the accuracy of the tGBS genotyping calls for these homozygous RILs (Supplementary Table S10).

### Genotyping heterozygous loci

To assess the accuracy of genotyping heterozygous sites, SNPs were called in 192 $F_2$ progeny of the B73 × Mo17 cross at tGBS (GRL2). After filtering for MCR, minor allele frequency and heterozygosity, the set of 70% MCR ('Materials and Methods' section) SNPs called in the $F_2$ population were used to create a genetic map consisting of 3498 markers. Because we were able to map similar numbers of markers in this F2 population as compared to IBM RILs, and considering the low rate of genotyping errors (0.5%), and the high correlations of physical and genetic marker orders (0.99) we concluded that tGBS performs well on both inbred and heterozygous populations (Supplementary Table S10). The presence of both homozygous and heterozygous genotypes also allowed us to classify genotyping errors identified in the $F_2$ population as being false homozygous or false heterozygous calls using segmentation (see 'Materials and Methods' section). Only a small proportion (1.7%, 11 848/677 929) of genotyping calls at polymorphic sites were putative errors, and heterozygous calls were as accurate as homozygous calls (Supplementary Table S11). Additionally, the intersection of SNPs in the $F_2$ and IBM RIL populations was examined. The majority of overlap-

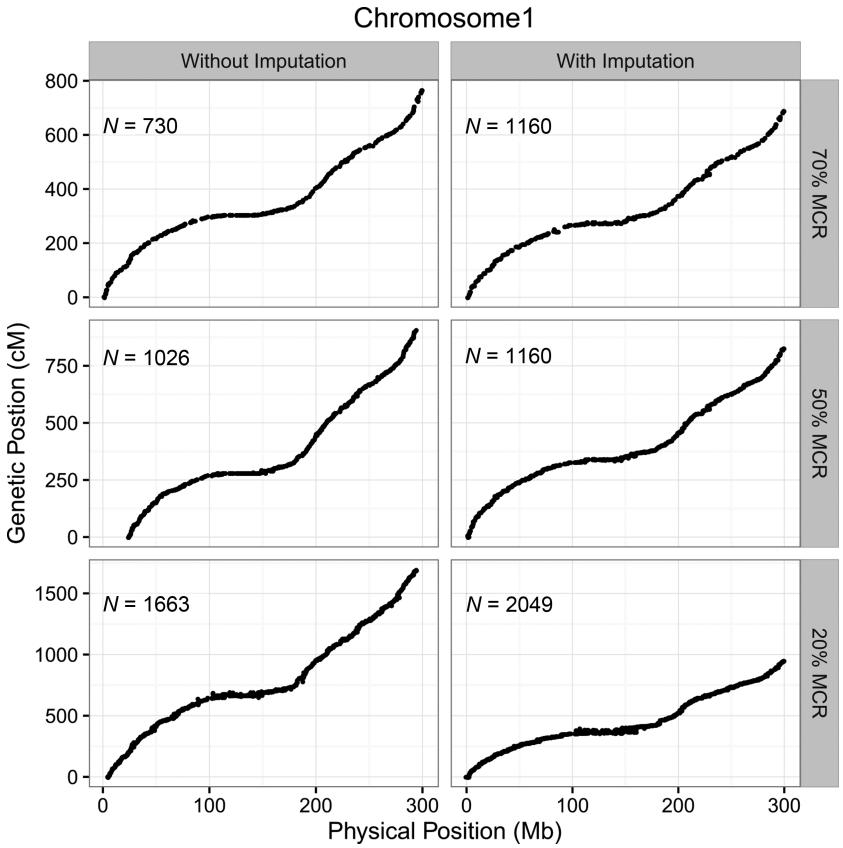


**Figure 4.** Genetic mapping in the IBM RILs. Comparisons of genetic and physical positions on chromosome 1 generated from ASMap for various MCRs, without and with LinkImpute-based imputation. Each dot represents the positon of a single SNP on a genetic and physical map.

ping high MCR SNPs are at on-target sites (Supplementary Table S12).

### Comparisons between tGBS and cGBS

We compared read depths for tGBS data from the NAM founders, IBM RILs and B73 × Mo17 $F_2$ reported in this study with read depths for cGBS data from a large maize diversity panel generated using ApeKI as the RE ($N = 3172$) by Romay *et al.* (38). For each technology we determined the median read depth at interrogated sites, i.e. those sites covered by at least five reads ('Materials and Methods' section). When comparing libraries with similar numbers of raw reads that are controlled for read length ('Materials and Methods' section; Supplementary Figures S4 and 5), the median read depths for tGBS (GRL1) and cGBS were similar, while in contrast and as expected tGBS (GRL2) and tGBS (GRL3) have greater read depth per interrogated site than cGBS (Figure 5).

We also compared the numbers of interrogated sites for tGBS (GRL2) and cGBS after controlling for read length and library size ('Materials and Methods' section). To conduct this analysis we subsampled standardized reads, i.e. both mapped and unmapped reads and both on- and off-target reads that have been truncated to equal lengths ('Materials and Methods' section) from the two datasets. At modest read depths and for a given MCR, tGBS yielded more interrogated sites than did cGBS (Figure 6). This ad-

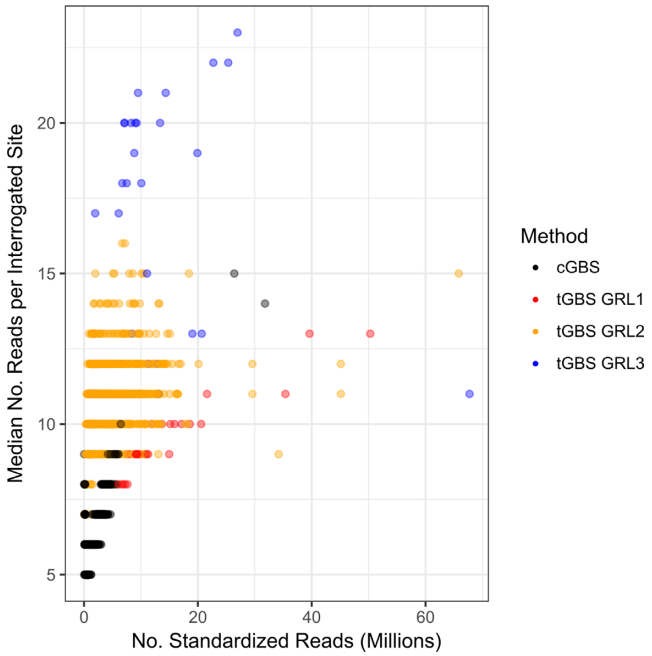


**Figure 5.** Median read depth per interrogated site for tGBS and cGBS data. Each dot represents a sample. For each GRL, tGBS data were analyzed for each of 25 NAM founders. Additionally, the IBM RIL ($N = 232$) and F2 samples ($N = 192$) were analyzed for tGBS (GRL2). The evaluation of cGBS is based on 3172 samples (38).

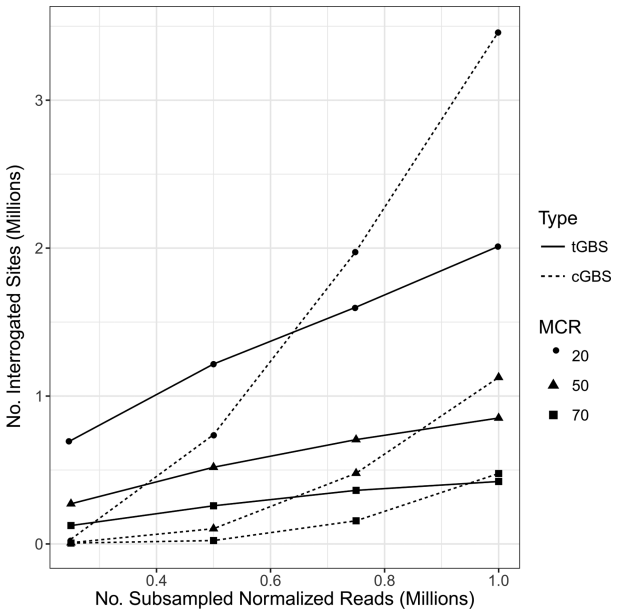


**Figure 6.** Numbers of interrogated sites from equal numbers of standardized tGBS (GRL2) and cGBS reads at various MCR cut-offs. As a consequence of data availability, data points (dots) are based on different numbers of samples: 198 samples were assessed using 1M subsampled reads and 433 samples were assessed using 0.75 M, 0.5 M and 0.25 M subsampled reads.

vantage of tGBS relative to cGBS increases at higher MCR values. In summary, although tGBS targets fewer sites than cGBS, when considering only sites that are consistently scored across many individuals within a population, tGBS yields more interrogated sites than cGBS.

The greater read depth of tGBS as compared to cGBS would be expected to result in higher SNP calling accuracy. The large maize diversity panel analyzed via cGBS includes the NAM founders which, as discussed earlier, were also genotyped via tGBS. It was therefore possible to compare the accuracies of tGBS and cGBS, using SNP calls from HapMap2 data to establish truth in cases where tGBS and cGBS SNP calls disagreed. HapMap2 was selected due to its accuracy (>98%) as demonstrated via comparison to tGBS and RNA-sequencing, and because calls from the Panzea SNP calling pipeline were available. Because tGBS and cGBS target only partially overlapping regions of the genome the number of sites that could be compared in this analysis was limited. Even so, this comparison demonstrated that tGBS provides greater accuracy than cGBS. When consistent heterozygous SNP genotyping ('Materials and Methods' section) was performed on both cGBS and tGBS sequencing reads, the accuracy of cGBS was only 90.5% as compared to 99.7% for tGBS (Table 1). The accuracy of cGBS genotyping can be improved if one can assume that most loci are homozygous, as is the case of the NAM founders. When SNP calls were generated from cGBS and tGBS sequencing reads using a more stringent homozygous SNP genotyping pipeline ('Materials and Methods' section), the accuracy of cGBS improved to 94.3% as compared to 98.6% for tGBS. However, this assumption of homozygosity is only appropriate when dealing with inbred lines or natural populations of self-pollinating species. Given the accuracy of tGBS at calling heterozygous loci we can conclude that tGBS in superior to cGBS when genotyping samples that are expected to be heterozygous such as the individuals from many types of genomic selection experiments and natural populations of outcrossing species.

It is possible to increase the accuracy of cGBS via the introduction of a minor allele frequency filter. For example, if similar comparisons are performed using Panzea's SNP calling of the cGBS reads described above, the accuracy increases to 99.3% (comparable to that of tGBS). Unfortunately, this increased accuracy achieved via the introduction of a 10% minor allele frequency filter substantially degrades the utility of cGBS for the discovery of rare novel alleles, highlighting the superiority of tGBS for analyzing diversity panels.

## DISCUSSION

Here, we present a novel approach to genotyping using sequence data, tGBS, which uses selection at the 3′ ends of a PCR primer to enhance genome reduction in an adjustable manner. tGBS employs oligos instead of adaptors, which has a number of technical advantages (20). This genotyping approach is simple and cost-efficient. We have demonstrated its high accuracy for genotyping both homozygous and heterozygous sites in diversity populations, RILs and $F_2$s.

### Technical advantages of tGBS

Our strategy of selecting only a subset of restriction digestion fragments for amplification and sequencing provides for flexible genome reduction. Adjusting GRLs provides different numbers of target sites for sequencing. While fewer SNPs are obtained at higher GRL levels, the number of reads per sample necessary to saturate the genotyping of on-target SNPs is also reduced (Supplementary Figure S3 and Table S4). Importantly, this results in more of the same sites across panels of samples having genotyping calls, resulting in lower levels of missing data per marker (Supplementary Figure S3). Additionally, the increased read depth at target sites allows for accurate genotyping of both homozygous and heterozygous sites (Figure 5 and Table 1).

The fact that higher GRL sites are a subset of lower GRL sites (i.e. 'TG' sites from GRL2 are a subset of 'T' sites from GRL1) offers advantages both within and across experiments. For example, in a given population, it is possible to use a lower GRL level to obtain more markers for higher resolution mapping subsequent to conducting a pilot study with a higher GRL. Perhaps more significantly, haplotypes can easily be tracked across populations even if these populations were analyzed using differing GRLs.

We have been unable to find published studies reporting RAD-Seq based genotyping of maize lines and thus it was not possible to directly compare the number of SNP sites identified by this method and tGBS. However, a report has recently been published using double digestion RAD-Seq and the REs PstI and Alw1 to genotype Sitka spruce (39). This study relied on pilot sequencing studies

**Table 1.** Concordant SNP calls summed across the NAM founders

| Genotyping pipeline used for cGBS | No. (%) concordant | | | Total comparisons |
|---|---|---|---|---|
| | tGBS (GRL2) | cGBS | HapMap2 | |
| Heterozygous | 30 309 (99.7) | 27 525 (90.5) | 29 831 (98.1) | 30 412 |
| Homozygous | 26 073 (98.6) | 24 926 (94.3) | 26 245 (99.3) | 26 440 |
| Panzea[1] | 24 772 (98.9) | 25 125 (99.3) | 25 098 (99.1) | 25 307 |

[1]tGBS SNPs were genotyped using the heterozygous method.
cGBS SNPs were genotyped using three different methods. The heterozygous genotyping pipeline allows for heterozygous calls ('Materials and Methods' section). The homozygous SNP calling pipeline discards genotyping calls that appear to be heterozygous ('Materials and Methods' section). The cGBS and HapMap2 SNPs downloaded from the Panzea website were genotyped using a SNP calling pipeline that is similar to our homozygous genotyping pipeline, though using a different software and with the addition of a minor allele frequency filter (37). tGBS and cGBS SNPs reported in the same row of this table were genotyped using the same SNP calling pipelines unless indicated otherwise.

to select the REs. If different enzymes are selected for different experiments, the resulting targeted sites will not be comparable among experiments. Considering the respective RE recognition sites and the selective bases of tGBS the estimated levels of genome reduction for this version of RAD-Seq and tGBS (GRL2) are similar; however, PstI (like the ApeKI used in cGBS) is methylation sensitive at sites which are often methylated in plants. Consequently, the number of RAD-Seq sites sequenced following the protocol of Fuentes-Utrilla *et al*. would be expected to be smaller than for GRL2. On the other hand, differential methylation among individuals has the potential to increase the amount of missing data when a GBS method relies on a RE that unlike BfuCI (see below) relies on pronounced methylation sensitivity to achieve genome reduction. There is, however no technical barrier to employing methylation sensitive REs within tGBS.

cGBS as described by Elshire *et al*., (12) relies on the restriction enyzme ApeKI which is sensitive to CpG methylation which is prevalent in plant genomes. BfuCI has also been reported exhibit sensitivity to CpG methylation. However, in our data 89% of predicted on-target BfuCI sites in the maize genome were represented by tGBS sequencing reads (data not shown), indicating that most of the associated BfuCI restriction sites had been digested by BfuCI. Given the prevalence of CpG methylation in the maize genome, this result is not consistent with BfuCI exhibiting substantial sensitivity to CpG methylation. Hence, to the extent that samples differ in their methylation patterns, the use of ApeKI may contribute to the higher missing rate of missing genotype calls from cGBS as compared to tGBS.

Conventional GBS methods (including RAD-Seq and other protocols) use adaptors. Because annealing/ligation can occur between adaptors (inter-adaptor-annealing/ligation) via overhang pairing, it is critical to control the ratio of adaptors and input genomic DNAs in the ligation reaction. In contrast, tGBS uses oligos thereby avoiding the serious problem of inter-adaptor annealing/ligation. In our tGBS experiments, satisfactory results were obtained despite the fact that we did not conduct titration experiments to optimize the ratio of oligos to template. When combined with REs that generate opposite direction overhangs, the use of oligos increases the specificity of the PCR because the PCR reaction will always begin with extension on the BfuCI end of the fragment, and extension from the NspI side will only proceed after this first extension has occurred. Consequently, fragments that do not contain a BfuI oligo will not be extended during selective PCR, thereby essentially eliminating the amplification of off-target NspI/NspI fragments. Further, because sequencing is initiated from the NspI side, any fragments that lack the BfuCI oligo would not be amplified in the final PCR and therefore would not be sequenced.

In breeding and diagnostics projects quick turn-around can be essential. Hence we used the Ion Proton sequencing platform which offers one-day turn-around at a per data point cost comparable to the low cost but slower turn-around Illumina sequencing platforms and a much lower cost than Illumina's fast-turnaround MiSeq technology. However, tGBS can also be conducted using Illumina platforms. We have tested Illumina oligos and barcodes (Supplementary Table S2) and obtained similar levels of accuracy as reported for the Ion Proton platform (data not shown). Combining tGBS oligo barcodes with barcodes on Illumina adaptors increases the ability to pool large numbers of samples without the need to synthesize a large number of barcoded oligos.

### Determination of selection levels and pooling size

One of the critical decisions in any GBS experiment is how much sequencing data to generate per sample to obtain the desired number of SNPs. In maize, ∼12 000 and ∼2000 consistently covered SNPs were obtained across 11 samples from 3 million raw tGBS (GRL2) reads and 1 million raw tGBS (GRL3) reads per sample, respectively (Supplementary Figure S3). In the case of the IBM RILs with tGBS (GRL2), 4293 high MCR SNPs and 10 736 low MCR SNPs were identified from an average of 2 million raw reads across all the RILs (Supplementary Table S10). SNPs with high missing data come predominantly from off-target sites and can be imputed or disregarded, while high MCR SNPs are predominantly from on-target sites and are consistently genotyped from one experiment to the next (Supplementary Table S12). The appropriate GRL and number of reads per sample will vary based on the organism and project goals; however, regardless of genome complexity and diversity among individuals, sequencing depths required to cover on-target sites at any given threshold are linearly related to genome size. tGBS has been conducted at various GRLs with the described REs and selective bases in over two dozen species with excellent results (40–43).

### Accuracy of genotyping with tGBS

Complementary methods were used to assess the accuracy of tGBS for genotyping inbreds. For the NAM founders and the IBM RILs, genotyping calls made at polymorphic sites were compared using three independent genotyping methods. Concordance was considered an indication of accuracy. Hence, if one method disagreed with the other two methods, the discordant method was assumed have been generated via a genotyping error. Even considering the potential of biological differences among samples used in the different methods to inflate estimates of errors, the genotyping accuracy estimated from the tGBS NAM concordance study was >99% (Supplementary Table S7). While concordance in the NAM founders was limited to polymorphic sites that had been genotyped by each of the three methods, segmentation of the IBM RILs could be used to identify regions in each RIL that are derived from either the B73 or Mo17 parent. By comparing each SNP call from multiple methods within a segment to the consensus genotype of that segment, it was possible to compare genotype calls at more sites. The concordance was high for all three methods, regardless of which SNP set was used to define the segments, with tGBS having a concordance >99% (Supplementary Table S9). The reported values should be considered minimum estimates of accuracy because true genotyping errors and small regions with double cross overs are confounded, resulting in a potentially inflated estimate of genotyping error rates. Further support for the accuracy of tGBS data is that the RIL genetic maps exhibited a high correlation with the physical marker order (>0.997), even in genetic maps constructed using non-imputed SNP sets that include markers with high levels of missing data (Supplementary Table S10). Because tGBS has an enhanced ability to discover and genotype rare alleles as compared to cGBS, it is also the preferred technology for genotyping diversity panels, even if these panels consist of inbred lines.

tGBS also provides accurate genotyping of heterozygous loci without the extensive filtering relied upon by other GBS methods. The accuracy of tGBS genotyping calls were between 98 and 99% in a segregating F2 population using a similar segmentation-based metric (Supplementary Table S11), and the correlation between the physical maize genome sequence and marker order on a genetic map constructed using these data was >0.999 (Supplementary Table S10). In contrast, the accuracy of cGBS genotype calls suffers even at homozygous sites when a SNP calling pipeline is allowed to make heterozygous calls (Table 1). The accuracy of cGBS genotyping calls only reaches the level of tGBS when minor allele frequency filters are employed, which prevents the discovery of rare alleles. The high genotyping accuracy of tGBS at heterozygous loci makes it suitable for genotyping $F_2$ and $F_1BC_1$ mapping populations where 50% of segregating markers are expected to be heterozygous, as well as in natural populations of outcrossing species that are expected to exhibit high levels of heterozygosity. The accuracy of tGBS heterozygous genotyping will be particularly useful for conducting genomic selection, which requires the ability to genotype populations of individuals that are heterozygous at many loci.

## ACCESSION NUMBERS

Debarcoded tGBS sequencing reads generated in this study are available in the Sequence Read Archive with the identifiers SRP095743 (RILs), SRP095751, SRP095750, SRP095749 (NAM GRL1, GRL2, and GRL3, respectively), and SRP095555 (F2s).

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

The authors thank Dr Wei Wu, Molly Parsons and Samantha Hoesel for assistance with wet lab experiments. Finally, we would like to dedicate this paper to the memory of Matt Hickenbotham.

## FUNDING

National Science Foundation [IOS1027527 to P.S.S.]; Iowa State University's Office of Biotechnology Fellowship (in part) (to A.O.); National Science Foundation Graduate Research Fellowship [DGE1247194 to A.O., in part]. Funding for open access charge: ISU Internal Funding.

*Conflict of interest statement.* The tGBS method is covered by patents pending in the USA and in other countries that are owned by Data2Bio LLC. S.L., J.C.S., C.-T.Y. and P.S.S. have equity interests in Data2Bio LLC.

## REFERENCES

1. Kwok,P.-Y. (2001) Methods for genotyping single nucleotide polymorphisms. *Annu. Rev. Genomics Hum. Genet.*, **2**, 235–258.
2. Morris,T., Robertson,B. and Gallagher,M. (1996) Rapid reverse transcription-PCR detection of hepatitis C virus RNA in serum by using the TaqMan fluorogenic detection system. *J. Clin. Microbiol.*, **34**, 2933–2936.
3. Oliphant,A., Barker,D.L., Stuelpnagel,J.R. and Chee,M.S. (2002) BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *Biotechniques*, **32**, 56–58.
4. Chee,M., Yang,R., Hubbell,E., Berno,A., Huang,X.C., Stern,D., Winkler,J., Lockhart,D.J., Morris,M.S. and Fodor,S.P.A. (1996) Accessing genetic information with high-density DNA arrays. *Science*, **274**, 610–614.
5. Egan,A.N., Schlueter,J. and Spooner,D.M. (2012) Applications of next-generation sequencing in plant biology. *Am. J. Bot.*, **99**, 175–185.
6. Mardis,E.R. (2011) A decade's perspective on DNA sequencing technology. *Nature*, **470**, 198–203.
7. Sun,X.W., Liu,D.Y., Zhang,X.F., Li,W.B., Liu,H., Hong,W.G., Jiang,C.B., Guan,N., Ma,C.X., Zeng,H.P. *et al.* (2013) SLAF-seq: an efficient method of large-scale de novo SNP discovery and genotyping using high-throughput sequencing. *PLoS One*, **8**, e58700.
8. Russello,M.A., Waterhouse,M.D., Etter,P.D. and Johnson,E.A. (2015) From promise to practice: pairing non-invasive sampling with genomics in conservation. *Peerj*, **3**, e1106.
9. Baird,N.A., Etter,P.D., Atwood,T.S., Currey,M.C., Shiver,A.L., Lewis,Z.A., Selker,E.U., Cresko,W.A. and Johnson,E.A. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*, **3**, e3376.
10. van Orsouw,N.J., Hogers,R.C.J., Janssen,A., Yalcin,F., Snoeijers,S., Verstege,E., Schneiders,H., van der Poel,H., van Oeveren,J. and Verstegen,H. (2007) Complexity reduction of polymorphic sequences (CRoPS): a novel approach for large-scale polymorphism discovery in complex genomes. *PLoS One*, **2**, e1172.

11. Andolfatto,P., Davison,D., Erezyilmaz,D., Hu,T.T., Mast,J., Sunayama-Morita,T. and Stern,D.L. (2011) Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Res.*, **21**, 610–617.
12. Elshire,R.J., Glaubitz,J.C., Sun,Q., Poland,J.A., Kawamoto,K., Buckler,E.S. and Mitchell,S.E. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, **6**, e19379.
13. Peterson,B.K., Weber,J.N., Kay,E.H., Fisher,H.S. and Hoekstra,H.E. (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One*, **7**, e37135.
14. Wang,S., Meyer,E., McKay,J.K. and Matz,M.V. (2012) 2b-RAD: a simple and flexible method for genome-wide genotyping. *Nat. Methods*, **9**, 808–810.
15. Stolle,E. and Moritz,R.F.A. (2013) RESTseq-efficient benchtop population genomics with RESTriction Fragment SEQuencing. *PLoS One*, **8**, e63960.
16. Beissinger,T.M., Hirsch,C.N., Sekhon,R.S., Foerster,J.M., Johnson,J.M., Muttoni,G., Vaillancourt,B., Buell,C.R., Kaeppler,S.M. and de Leon,N. (2013) Marker density and read depth for genotyping populations using genotyping-by-sequencing. *Genetics*, **193**, 1073–1081.
17. 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*, **11**, e0161333.
18. 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**, 1–8.
19. Yu,J., Holland,J.B., McMullen,M.D. and Buckler,E.S. (2008) Genetic design and statistical power of nested association mapping in maize. *Genetics*, **178**, 539–551.
20. Liu,S., Hsia,A.P. and Schnable,P.S. (2013) Digestion-ligation-amplification (DLA): a simple genome walking method to amplify unknown sequences flanking mutator (Mu) transposons and thereby facilitate gene cloning. *Methods Mol. Biol.*, **1057**, 167–176.
21. Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, **8**, 186–194.
22. Ewing,B., Hillier,L., Wendl,M.C. and Green,P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, **8**, 175–185.
23. Chou,H.H., Sutton,G., Glodek,A. and Scott,J. (1998) Lucy—a sequence cleanup program. In: *Proceedings of the Tenth Annual Genome Sequencing and Annotation Conference (GSAC X)*, Miami.
24. Li,S. and Chou,H.H. (2004) LUCY2: an interactive DNA sequence quality trimming and vector removal tool. *Bioinformatics*, **20**, 2865–2866.
25. Schnable,P.S., Ware,D., Fulton,R.S., Stein,J.C., Wei,F., Pasternak,S., Liang,C., Zhang,J., Fulton,L. and Graves,T.A. (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.
26. Wu,T.D. and Nacu,S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**, 873–881.
27. Glaubitz,J.C., Casstevens,T.M., Lu,F., Harriman,J., Elshire,R.J., Sun,Q. and Buckler,E.S. (2014) TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS One*, **9**, e90346.
28. Yu,J., Li,X., Zhu,C., Yeh,C.-T., Wu,W., Takacs,E., Petsch,K., Tian,F., Bai,G. and Buckler,E. (2012) Genic and non-genic contributions to natural variation of quantitative traits in maize. *Genome Res.*, **22**, 2436–2444.
29. Li,L., Petsch,K., Shimizu,R., Liu,S., Xu,W.W., Ying,K., Yu,J., Scanlon,M.J., Schnable,P.S. and Timmermans,M. (2013) Mendelian and non-Mendelian regulation of gene expression in maize. *PLoS Genet.*, **9**, e1003202.
30. Liu,S., Chen,H.D., Makarevitch,I., Shirmer,R., Emrich,S.J., Dietrich,C.R., Barbazuk,W.B., Springer,N.M. and Schnable,P.S. (2010) High-throughput genetic mapping of mutants via quantitative single nucleotide polymorphism typing. *Genetics*, **184**, 19–26.
31. Olshen,A.B., Venkatraman,E.S., Lucito,R. and Wigler,M. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.
32. Taylor,J. and Bulter,D. (2017) R Package ASMap: Efficient Genetic Linkage Map Construction and Diagnosis. *Journal of Statistical Software*, **79**, 1–29.
33. Money,D., Gardner,K., Migicovsky,Z., Schwaninger,H., Zhong,G.Y. and Myles,S. (2015) LinkImpute: fast and accurate genotype imputation for nonmodel organisms. *G3 (Bethesda)*, **5**, 2383–2390.
34. Broman,K.W. (2010) Genetic map construction with R/qtl. *Technical Report# 214*. University of Wisconsin-Madison, Department of Biostatistics & Medical Informatics.
35. Qiu,F., Guo,L., Wen,T.J., Liu,F., Ashlock,D.A. and Schnable,P.S. (2003) DNA sequence-based "Bar codes" for tracking the origins of expressed sequence tags from a maize cDNA library constructed using multiple mRNA sources. *Plant Physiol.*, **133**, 475–481.
36. Chia,J.-M., Song,C., Bradbury,P.J., Costich,D., de Leon,N., Doebley,J., Elshire,R.J., Gaut,B., Geller,L. and Glaubitz,J.C. (2012) Maize HapMap2 identifies extant variation from a genome in flux. *Nat. Genet.*, **44**, 803–807.
37. Lee,M., Sharopova,N., Beavis,W.D., Grant,D., Katt,M., Blair,D. and Hallauer,A. (2002) Expanding the genetic map of maize with the intermated B73 x Mo17 (IBM) population. *Plant Mol. Biol.*, **48**, 453–461.
38. Romay,M.C., Millard,M.J., Glaubitz,J.C., Peiffer,J.A., Swarts,K.L., Casstevens,T.M., Elshire,R.J., Acharya,C.B., Mitchell,S.E., Flint-Garcia,S.A. *et al.* (2013) Comprehensive genotyping of the USA national maize inbred seed bank. *Genome Biol.*, **14**, 1–18.
39. Fuentes-Utrilla,P., Goswami,C., Cottrell,J.E., Pong-Wong,R., Law,A., A'Hara,S.W., Lee,S.J. and Woolliams,J.A. (2017) QTL analysis and genomic selection using RADseq derived markers in Sitka spruce: the potential utility of within family data. *Tree Genet. Genomes*, **13**, 1–12.
40. Tang,H.B., Zhang,X.T., Miao,C.Y., Zhang,J.S., Ming,R., Schnable,J.C., Schnable,P.S., Lyons,E. and Lu,J.G. (2015) ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biol.*, **16**, 1–15.
41. Guo,Y., Lin,W.K., Chen,Q., Vallejo,V.A. and Warner,R.M. (2017) Genetic Determinants of Crop Timing and Quality Traits in Two Interspecific Petunia Recombinant Inbred Line Populations. *Sci. Rep.*, **7**, 1–12.
42. Pang,Y., Chen,K., Wang,X., Wang,W., Xu,J., Ali,J. and Li,Z. (2017) Simultaneous Improvement and Genetic Dissection of Salt Tolerance of Rice (Oryza sativa L.) by Designed QTL Pyramiding. *Front Plant Sci.*, **8**, 1–11.
43. Goiffon,M., Kusmec,A., Wang,L., Hu,G. and Schnable,P.S. (2017) Improving Response in Genomic Selection with a Population-Based Selection Strategy: Optimal Population Value Selection. *Genetics*, **206**, 1675–1682.