G3: Genes|Genomes|Genetics Early Online, published on May 19, 2017 as doi:10.1534/g3.117.042036

Phased genotyping-by-sequencing enhances analysis of genetic diversity and reveals divergent copy number variants in maize

Heather Manching[*], **Subhajit Sengupta**[†], **Keith R. Hopper**[‡], **Shawn W. Polson**[§], **Yuan Ji**[†,**] **and Randall J. Wisser**[*,1]

[*]Department of Plant and Soil Sciences, University of Delaware, Newark, DE 19716, [†]Program of Computational Genomics & Medicine, NorthShore University HealthSystem, Evanston, IL 60201, [‡]Beneficial Insect Introductions Research Unit, United States Department of Agriculture, Agricultural Research Service, Newark, DE 19713, [§]Center for Bioinformatics and Computational Biology, Delaware Biotechnology Institute, University of Delaware, Newark, DE 19711, [**]Department of Health Studies, University of Chicago, Chicago, IL 60637

ABSTRACT High-throughput sequencing of reduced representation genomic libraries has ushered in an era of genotyping-by-sequencing (GBS), where genome-wide genotype data can be obtained for nearly any species. However, there remains a need for imputation-free GBS methods for genotyping large samples taken from heterogeneous populations of heterozygous individuals. This requires a number of issues encountered with GBS be considered, including the sequencing of non-overlapping sets of loci across multiple GBS libraries, a common missing data problem that results in low call rates for markers per individual, and a tendency for applicability only in inbred line samples with sufficient linkage disequilibrium for accurate imputation. We addressed these issues while developing and validating a new, comprehensive platform for GBS. This study supports the notion that GBS can be tailored to particular aims, and using *Zea mays* our results indicate that large samples of unknown pedigree can be genotyped to obtain complete and accurate GBS data. Optimizing size selection to sequence a high proportion of shared loci among individuals in different libraries and using simple *in silico* filters, a GBS procedure was established that produces high call rates per marker (>85%) with accuracy exceeding 99.4%. Furthermore, by capitalizing on the sequence-read structure of GBS data (stacks of reads), a new tool for resolving local haplotypes and scoring phased genotypes was developed, a feature that is not available in many GBS pipelines. Using local haplotypes reduces the marker dimensionality of the genotype matrix while increasing the informativeness of the data. Phased GBS in maize also revealed the existence of reproducibly inaccurate (apparent accuracy) genotypes that were due to divergent copy number variants unobservable in the underlying single nucleotide polymorphism data.

KEYWORDS
GBS; haplotype phasing; copy number variant; imputation; maize

INTRODUCTION

Genome-wide genotyping of population samples is fundamental to a range of studies in genetics and genomics, and genotyping-by-sequencing (GBS) of multiplexed, high-throughput sequencing (HTS) libraries has emerged as a cost-effective strategy for obtaining this information. Genotyping-by-sequencing of multiple samples relies on a reduced representation sequencing strategy that restricts sequencing across what is hoped to be a common subspace of the genome in different samples. A prevalent technique used for this involves the ligation of barcoded adapters to DNA digested with restriction enzymes, followed by sequencing of fragments within a restricted size range. In principle, this leads to stacks of sequences anchored at the restriction cut sites across the genomes of different individuals. Following initial publications on restriction enzyme-mediated GBS (e.g. Baird *et al.* 2008; Andolfatto *et al.* 2011; Elshire *et al.* 2011), the approach has continued to be extended and optimized, such as fine-tuning the number of loci sequenced by using different restriction enzymes and size selection

Manuscript compiled: May 2017
[1]Randall J. Wisser; Department of Plant and Soil Sciences, University of Delaware, 152 Townsend Hall, 531 South College Avenue, Newark, DE, 19716; rjw@udel.edu

windows (Peterson *et al.* 2012) or selective primers (Sonah *et al.* 2013), minimizing the front-end cost by using enzymes that create blunt-end fragments for universal adapters (Heffelfinger *et al.* 2014), and increasing the scalability for large samples by including a sequence capture step (Ali *et al.* 2016). Methods and applications of GBS continue to develop at a rapid rate.

With the mass of data produced by HTS, even low sequencing error rates can lead to an abundance of inaccurate genotype calls. Moreover, errors accumulate along steps in the GBS pipeline from library preparation to data processing. As a reflection of the amount of noise in GBS data, it is not unusual to find publications (as well as our own experience) where as much as ~50% of the raw HTS data is filtered after mapping (e.g. Beissinger *et al.* 2013) and as much as ~95% of the discovered variants are filtered after variant calling (e.g. Hyma *et al.* 2015). Typically, statistically or biologically informed criteria for filtering are used to enrich for more accurate GBS data, or a model is fitted to the data to remove genotypes that do not conform to expectations (e.g. tests for independent assortment and co-segregation in genetic mapping populations, Hardy-Weinberg equilibrium in other population samples). However, estimates of genotyping accuracy for a given protocol, pipeline and application of GBS (under all of its flavors: e.g. RAD, GBS, ddRAD, Rapture, etc.) are rarely reported. One recent study compared different bioinformatic pipelines and found that genotyping accuracy for one GBS dataset ranged from 76% to 99%, depending on the bioinformatic pipeline used to score genotypes (Torkamaneh *et al.* 2016). Accuracy estimates based on resequencing of GBS loci have ranged from 51% (Rocher *et al.* 2015) to 99% (Torkamaneh *et al.* 2016). Given the wide range in accuracy and method- and study-specific nature of GBS, establishing controls or approaches that allow accuracy to be estimated is important when embarking on GBS studies.

Genotyping-by-sequencing tends to have a missing-data problem due to two primary issues: (i) for libraries with low sequence coverage per sample, the amount of missing genotype data can be high, resulting in a low call rate per marker (as low as 10%; Fu and Peterson 2011); (ii) for large population samples where many libraries need to be sequenced, the number of shared loci in different libraries may be low, resulting in low call rates per sample. This latter issue has not been fully addressed in the literature on GBS. We are aware of one study in which the consistency of genotyping the same loci in different libraries with the same samples was assessed, and in this study, overlap of loci among libraries ranged from 12% to 98% (DaCosta and Sorenson 2014). The results of that study suggested an impractical (costly) solution for genotyping large samples, where greater consistency would be achieved by pooling samples later in the protocol. This would require individual samples to be processed through size selection and PCR amplification independently, which increases the cost for reagents and time for library preparation. A solution to the GBS missing-data problem has been imputation, whereby the genotype for a missing SNP is inferred from the state of nearby SNPs. However, imputation is not always possible and may not be sufficiently accurate, in which case both of the above issues compromise the effectiveness of GBS in studies where sample sizes exceed the multiplex depth of a single sequencing library. While the former issue may be addressed by reducing the sequence space (e.g. Peterson *et al.* 2012; Sonah *et al.* 2013) or deeper sequencing, the latter issue requires techniques that provide reproducible enrichment of shared loci (e.g Ali *et al.* 2016).

Despite the shortcomings mentioned above, GBS has been effective in several settings. There is now a growing interest in using GBS for a wider range of studies (Andrews *et al.* 2016), many of which could benefit from or be advanced with phased genotype data. The standard approach when performing GBS involves scoring single nucleotide polymorphisms (SNPs). However, in GBS data, locally phased haplotypes exist as stacks of reads for each locus. If these contain more than one SNP, then multi-nucleotide polymorphisms (MNPs) can be genotyped. The use of MNPs has not been widely exploited in GBS due to the lack of tools for scoring phased genotypes. The software STACKS (Catchen *et al.* 2013) and Haplotag (Tinker *et al.* 2016) do have functions embedded within their pipelines for extracting MNP haplotypes from GBS data, but the pipeline-dependency of these functions limits their use more generally, and the algorithms rely on population-specific filtering criteria rather than statistical evaluation of the likelihood that each haplotype in an individual is real. Because HTS data is imperfect and the quality of sequenced bases and read mappings are quantitatively encoded, scoring MNPs at a locus within an individual is not straightforward if one considers this information relevant. A local haplotyping tool, LocHap, which uses community-standard file formats, offers a more generalized solution for phased genotyping. LocHap was developed under a probabilistic framework that uses the quality metrics of base calls and mapping results as well as sampling effects to phase MNPs in HTS data (Sengupta *et al.* 2015). LocHap was designed to identify distinct haplotypes in heterogeneous populations of cells (somatic mosaicism) irrespective of homology, such that the outputted haplotypes at a single locus can vary in length and are not necessarily alignable within and across individuals. Consequently, LocHap does not produce data in a format that is compatible with population or quantitative genetic studies. In our study, LocHap was extended to perform phased genotyping with GBS data on individual samples, which we call LocHap-GBS.

Phased genotyping is useful for various applications in genetics and genomics. Typing MNPs can help distinguish more than two alleles at a locus, providing greater information content for studying genetic diversity (Lu *et al.* 2011). Phased genotypes can facilitate imputation in multi-parental populations where SNPs would otherwise conflate ancestral alleles (Davies *et al.* 2016). Haplotype data can also increase the accuracy of estimated breeding values, thereby increasing the efficiency of plant and animal breeding (Ferdosi *et al.* 2016). In our study, while validating LocHap-GBS, we found that phased genotyping can also uncover copy number polymorphisms.

The aims of this study were to: (i) establish a standardized and empirically optimized flex-plex GBS protocol with an accompanying informatics pipeline for genotyping; (ii) evaluate the accuracy and potential applicability of this procedure for genotyping large samples from heterogeneous populations of heterozygous individuals; and (iii) extend the use of GBS for phased genotyping. We used maize, which is an agriculturally relevant species with an ~2.3 Gb completed reference genome sequence, but where genomic analysis is challenged by large amounts of repetitive sequence. Genetic trios of pairs of inbred parental lines and their F_1 progeny were used in assessing accuracy of GBS. We tested whether simple *in silico* filters could facilitate highly accurate scoring of genotypes irrespective of zygosity and knowledge of parentage. In addition to developing and validating a method for phased-GBS, we also examined how SNP and MNP data affect inference on the relatedness among a set of inbred lines used by the maize genetics research community.

MATERIALS AND METHODS

Study Samples

Detailed information on the samples used in this study is in Supplementary Material (File S1). Briefly, the samples included a synthetic population created from seven tropical inbred lines (used to evaluate the consistency of sequenced loci in different GBS libraries), sets of trios or pairs of inbred lines and their corresponding F_1 hybrids (used to evaluate the repeatability and accuracy of GBS), an F_2 population (used to examine the transmission of phased MNPs, along with the corresponding parental trio), and parental inbred lines of a maize nested association mapping (NAM) population (Mcmullen *et al.* 2009; used to examine phased genotyping for analysis of genetic diversity).

Genotyping-by-Sequencing

We describe the design, protocol, and associated software for GBS based on a double digestion technique similar to protocols by Poland *et al.* (2012) and Peterson *et al.* (2012). A specific interest of ours was to develop a GBS design for genotyping heterozygous samples of potentially unknown parentage with little missing data for large sample studies. The method was optimized, validated, and tested for various applications.

Detailed protocols and other relevant information are provided in Supplemental Material (File S2, S3). Briefly, two separate but related restriction-associated sequence polymorphism (RASP) adapter designs were developed: (i) RASP-1.0 was based on Illumina's "genomic DNA" adapters (Illumina, Inc., San Diego, CA) modified with appropriate overhang sequences for ligation and a 6-nucleotide inline barcode for 48-plex genotyping (RASP-1.0 and RASP-1.1 adapter oligonucleotides and primers; File S3); (ii) RASP-2.0 was based on Illumina's TruSeq adapters modified with appropriate sequence overhangs for ligation. RASP-2.0 uses variable length inline barcodes (5-10 nucleotides) for 48-plex genotyping (barcode sequences were designed using http://www.deenabio.com/gbs-adapters) along with standard TruSeq indices. The inline barcoded adapters and TruSeq barcoded adapters can be combined to construct plexes in multiples of 48 (RASP-2.0 adapter oligos and primers: File S3).

The protocol used for GBS was improved over the course of our study, including changes to the adapter sets, the size selection method and number of reactions used for PCR (Table S1). The general protocol was as follows. Purified and normalized DNA (200 ng) was digested using two restriction endonucleases, *Ngo*MIV and *Csp*6I, for 30 minutes at 37 °C. The *Ngo*MIV/*Csp*6I enzyme pair was chosen because it provided the highest read output and the lowest variation between samples based on comparisons between four different enzyme pairs (*Ngo*MIV/*Csp*6I; *Ngo*MIV/*Mse*I; *Pst*I/*Csp*6I; *Pst*I/*Mse*I; data not shown). Adapters were ligated with T4 DNA ligase using temperature-cycle ligation (Lund *et al.* 1996; 300 cycles of 30s at 10 °C and 30s at 30 °C; this was determined to improve ligation efficiency by qPCR [data not shown]), followed by heat inactivation of the ligase at 65 °C for 30 minutes. An equal volume of each ligate was pooled and then purified using either AMPure (A63880; Beckman Coulter, Inc., CA, USA) or SPRIselect (B23317; Beckman Coulter, Inc., CA, USA) beads following the manufacturer's recommended protocol for standard clean-up (exclusion of fragments <100 bp). For all libraries, prior to PCR amplification, size selection was performed on the pooled ligate using a BluePippin (Sage Science, MA, USA). Size selected samples were PCR amplified with Phusion High Fidelity Master Mix (M0531S; NEB, Inc., MA, USA) using the universal primer sequences from Illumina's genomic DNA or TruSeq sample kit. To reduce PCR bias, a minimum of eight separate PCR reactions were performed on the size selected template and the products were pooled. AMPure or SPRI beads were used to eliminate excess primers and bi-products (exclusion of fragments <100 bp). The range and peak of Bioanalyzer size fragment profiles were analyzed for each library before and after PCR to determine the consistency of the size profiles. Due to variation in fragment length profiles found following pre-PCR size selection, a post-PCR size selection was performed on some of the libraries, followed by Bioanalyzer analyses to confirm size profiles. This turned out to be critical for genotyping separate libraries, and we have since included this change in our standard protocol. Multiplex libraries were quantified using Quant-iT PicoGreen dsDNA assay kit (P7589; Thermo Fisher Scientific, Inc., MA, USA) and sequenced (1 x 101 cycles) on an Illumina HiSeq 2500 at the Delaware Biotechnology Institute .

GBS data processing

Computational pipeline: Sequences were processed using a custom *red*uced *rep*resentation "RedRep" computational pipeline with the following basic steps: 1) sequence splitting by barcode; 2) raw sequence quality control; 3) reference genome mapping; and 4) variant calling (SNPs). Briefly, sequences are deconvoluted by barcode using custom logic and the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit). Novel scripts and the CutAdapt package (Martin 2011) are then used to remove adapters, trim low quality read ends, and filter out sequences that do not meet minimum length/quality standards or lack the expected restriction site and adapter sequence expectations. Quality reads are mapped to the reference genome using the BWA-MEM algorithm (Li 2013) and SNPs are identified using the multi-sample discovery mode of the GATK Unified Genotyper (McKenna *et al.* 2010). The scripts for the pipeline and documentation are available under the open source MIT license at https://github.com/UD-CBCB/RedRep. Version 2.0 scripts committed to the repository on December 7, 2016 were used for this manuscript.

Post-vcf filtering: All 64-384 bp loci flanked by *Ngo*MIV (G↓CCGGC) and *Csp*6I (G↓TAC) recognition sites in the B73 v2 reference genome were identified by *in silico* digestion. BLASTn was used to perform genome-wide searches for each *in silico* digested sequence starting from the *Csp*6I recognition site (starting point for sequence reads) up to a maximum of 96 nucleotides (longest quality trimmed read length of our actual data). Sequences that were more than four percent distant from all other loci in the reference genome were flagged as "4PD" loci, and SNPs that occurred within these loci were maintained while the remaining SNPs were filtered. This excludes the possibility for scoring variants at loci within the reference genome that differ from one another by fewer than four nucleotides in 100 base pairs. In addition, for each individual and at each SNP site, if the genotype was not based on a minimum depth of read coverage, it was set to missing: 3X for scoring inbred samples; 12X for scoring heterozygous samples. Samples were then removed that had genotype data for <25% of the loci for 48-plex libraries and <10% for 192-plex libraries; these included problematic samples and negative controls. Finally, for each SNP, an 85% call rate threshold was used, and SNPs with more than two nucleotide variants were removed.

In silico characterization of GBS loci: The *in silico* digested loci, along with the SNP sites scored across all samples were summarized according their distribution across the B73 v2 reference genome. Using the closest-features program of BEDOPS v2.4.15 (Neph *et al.* 2012), genic and intergenic associations of the digestion

loci and SNPs were determined using gene structures in release-5b (filtered gene set; http://ftp.maizesequence.org/release-5b/).

Assessing Accuracy of GBS

A set of parent-hybrid trios were used to assess scoring accuracy: (i) CML373, CML341, and CML373 X CML341; (ii) CML341, CML277, and CML341 X CML277; (iii) Tzi9, CML258, and Tzi9 X CML258. One trio involving Tzi8 was excluded from accuracy assessment because of excess heterogeneity (~12.8%) and relatively high residual heterozygosity (~1.3%) as determined in preliminary work using the MaizeSNP50 BeadChip (Ganal *et al.* 2011). The repeatability of genotyping was estimated from replicate DNA samples of each member of the trio processed in the same library (except for the hybrid Tzi9 X CML258 because of quality issues with the replicate). Using loci with complete and consistent calls (i.e. between replicated DNA samples), genotyping accuracy was measured as the proportion of loci with the expected genotype in the hybrid given the genotype of the parents. Because GBS was performed on only one plant for each member of a trio, loci that were heterozygous in either parental line were excluded when estimating genotyping accuracy (according to prior data of ours based on the MaizeSNP50 chip, residual heterozygosity and heterogeneity for the parental lines was <0.8%). Furthermore, because the one F_1 plant genotyped was taken from bulked seed of progeny from multiple crosses between the parental lines (such that the genotyped parents may not provide the exact expectation for the specific F_1 plant that was genotyped), estimates of accuracy are expected to be slightly downward biased.

Genotyping of local haplotypes

Because GBS can resolve multiple variants present within each position-specific stack of sequence reads, phased haplotypes can be extracted. A local haplotyping program called LocHap (Sengupta *et al.* 2015) was extended for use with GBS data, which we refer to as LocHap-GBS (http://compgenome.org/lochap/GBS/). Figure 1 depicts the workflow for LocHap-GBS, which requires four input files: the standard *bam*, *bai*, and *vcf* files, plus a user-specified *bed* file containing intervals where haplotypes should be searched. To develop LocHap-GBS, a collection of modules were created that provide the following functionalities: (i) a file parser that defines intervals in which to search for haplotypes (ii) automation for running multiple samples and merging samples into a single output file; (iii) outputting of haplotypes that include bases even if they are homozygous in an individual; and (iv) additional flexibility in the format of output (either haplotype calling format [*hcf*] or *bed*).

While developing LocHap-GBS some additional improvements were made to LocHap that were related to dynamic memory allocation. Presently, LocHap is limited to calling haplotypes across a maximum of three heterozygous sites. LocHap-GBS uses a bed file with predefined intervals of windows in which to search for haplotypes, but these windows may contain more than three heterozygous sites in some individuals. Therefore, LocHap-GBS parses windows into subwindows that include a maximum of three heterozygous SNPs by referencing the coordinates of each window and the genotypes of SNPs in the *vcf* file (Figure 1). In this study, the intervals specified in the bed file were the 4PD loci from the *in silico* digest (only SNPs at these loci had been maintained in our *vcf* outputs, so this was the genotyping space of interest in this study).

We define SNPs, MNPs, and copy number polymorphisms (CNPs) as the property of a locus (loci that have variation among individuals) and single nucleotide variants (SNVs), multi-nucleotide variants (MNVs), and copy number variants (CNVs) as the property of an individual or in reference to a specific form of the polymorphism (a diploid individual is expected to have a maximum of two SNVs or MNVs at a locus). To prepare MNP data for analysis, MNVs were split into separate columns using the df2genind function within the adegenet package (Jombart 2008) in R (R Core Team 2016). An MNV containing one missing SNV was assigned a value of "NA", and a genotype containing a missing MNV was assigned a value of "NA."

Analysis of Genetic Diversity

To examine differences in information content when using SNP versus MNP data, genetic distances of the parents of the NAM population were analyzed. When analyzing the MNP data, we encountered genotypes with more than two MNVs (implying a ploidy of greater than two, which is not possible for orthologous loci in maize). After determining these were putative CNPs within the genomes of the sequenced samples relative to the reference genome sample (see Results), we examined the impact of including or excluding CNPs in the analysis of genetic diversity. The original MNP genotype matrix was split into three datasets where: (i) all putative CNPs were included for analysis (MNP1); (ii) all loci that had more than two MNVs were masked (MNP2); and (iii) loci that had more than two MNVs or were heterozygous in an inbred line (these are also potentially CNPs) were masked (MNP3). Pairwise genetic distance matrices based on shared allele distances (Bowcock *et al.* 1994) were computed for each dataset. The distance matrices were compared based on summary statistics and the Mantel test for correlation between matrices (Mantel 1967; implemented using the ade4 package in R). To visualize the relationships between the lines, multidimensional scaling (MDS, Kruskal 1964) was performed using the cmdscale function in the R package stats. Phylogenetic trees were generated using the BioNJ algorithm (Gascuel 1997) with the ape package (Paradis *et al.* 2004) in R, with 1000 bootstrap replicates performed to obtain branch support probabilities. Distances between trees were compared using symmetric Robinson-Foulds distances (Robinson and Foulds 1981) calculated using the Phangorn package in R (Schliep 2011). Cladograms were plotted using FigTree 1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/) with branches transformed to equal lengths for displaying topology.

Data Availability

Sequence data used in this study was deposited in the NCBI Short Read Archive (SRA) under two Bioprojects: PRJNA385842 and PRJNA385849. The submitted data has been de-multiplexed and processed through quality control using RedRep.

RESULTS

Genotyping-by-sequencing of *Ngo*MIV-*Csp*6I loci in maize

In silico digestion of the B73 reference genome with *Ngo*MIV and *Csp*6I identified 321,927 sequences that were 64-384 bp long and were flanked by the recognition site of each enzyme, of which 78,903 were 4PD loci. A 4PD locus in maize ensures that in most cases the distance between loci within the reference genome is greater than the average distance within a locus across genomes (on average, SNPs occur every 28 base pairs in maize [Tenaillon *et al.* 2001]). This was expected to minimize genotyping errors from ambiguous mapping. A lower PD threshold or a fixed 1 bp difference threshold may be appropriate and result in the identification of more SNPs, but we did not consider different thresholds in this study.

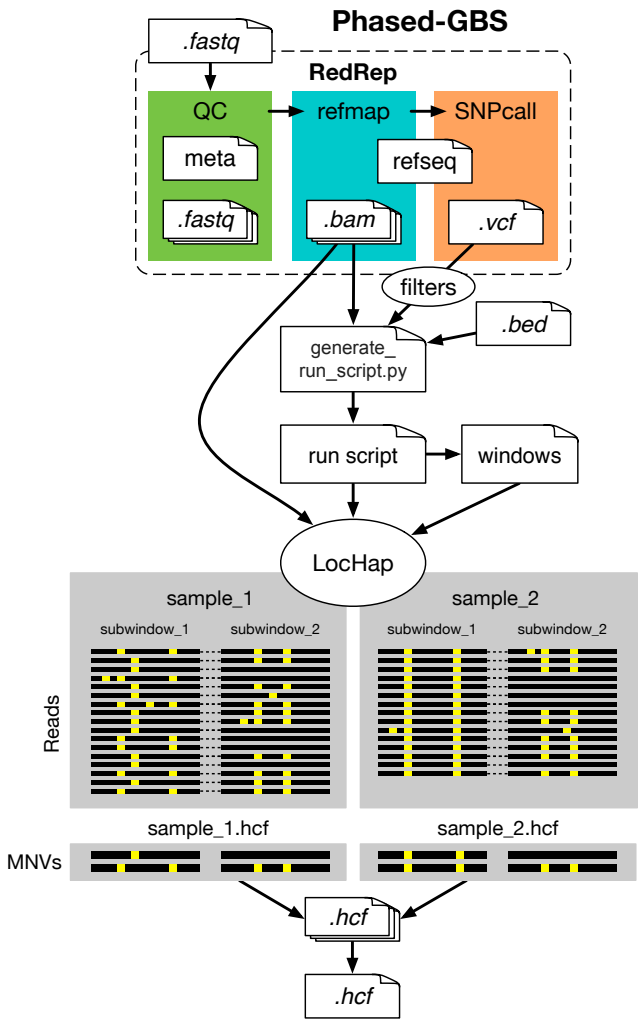

Figure 1 Phased genotyping-by-sequencing. A *red*uced *rep*resentation (RedRep) pipeline is used for SNP typing. The figure shows the basic flow of RedRep, which begins with a *fastq* data file with barcoded sequences. For quality control (QC), the "meta" file contains metadata used for demultiplexing into sample-specific *fastq* files. A reference genome sequence file is used for mapping (refmap) and variant calling (SNPcall). LocHap-GBS is run by editing a generate.py file specifying the location of the *bam* files, the filtered *vcf* file, and a *bed* file of window coordinates to search for haplotypes. A LocHap-GBS run file and windows file are automatically generated. Windows are currently split into subwindows with a maximum of three heterozygous sites within any one individual in the *vcf* file. This situation is depicted for reads across a window that has been delineated into two subwindows where phasing is performed. The dashed connecting line between reads indicates that a contiguous sequence with five SNPs was split into two subwindows. Black-filled bars represent the reference sequence and yellow squares represent SNVs. Given stacks of reads across each subwindow, LocHap-GBS uses a probabilistic model to identify haplotypes in the presence of sequence errors (depicted as one-off instances in the stacks of reads). An *hcf* file is created for each sample, which is then merged into a combined *hcf* file for downstream analysis.

For GBS, short-read sequence libraries typically comprise fragments that fall within a size range of 100 bp. To determine the use of targeted size selection to optimize the recovery of fragments from non-repetitive loci, we looked at the relative abundance of *Ngo*MIV-*Csp*6I loci classified as 4PD versus repetitive (≤4PD) in 100 bp windows (sliding every 10 bp from 64 bp to 384 bp). The median number of 4PD loci among windows was 22,854 (Figure S1), but the distribution was somewhat skewed where more 4PD loci were present in 100 bp windows of smaller fragment sizes. The ratio of 4PD:repetitive loci in 100 bp windows ranged from ~0.3-0.4, suggesting size selection might be used to maximize sequencing resources by avoiding repetitive sequences. However, many restriction endonucleases, including *Ngo*MIV, are sensitive to different types of methylation, and SNPs are not expected to be uniformly distributed across the genome, such that expectations from *in silico* analyses are only a proxy for the numbers of scorable loci for a given choice of enzyme. Moreover, the fragments enriched by size selection and PCR are not a uniform sample of the underlying distribution of digested fragments and only a fraction of those loci will have sufficient read depths for scoring genotypes across samples.

The genic space (defined here as the gene plus 5 kb flanking sequences) comprises ~27% of the B73 v2 genome (genes alone comprise ~8%). Unfiltered *Ngo*MIV-*Csp*6I loci from *in silico* digestion were distributed across the genic and intergenic spaces similar to that expected by chance alone (although significantly different, coverage of the intergenic space was greater than expected by only four percentage points). In total, these unfiltered loci encompass 3% of the genome and are associated with a majority of the gene space, including 69% (27,641) of all maize genes. Filtering *Ngo*MIV-*Csp*6I loci that were less than four percent distant from at least one alternative site in the genome removed loci associated with 7,472 genes. The 4PD loci used to score SNPs were enriched within the gene space by 32 percentage points (59% for 4PD loci versus an expected genome distribution of 27%; Figure 2A) and associated with 51% (20,169) of all genes.

A total of 159,994 post-filtered SNPs (≥3X depth at 4PD loci) were identified across all samples genotyped in this study (File S4; we have also included coordinates for SNPs identified using the B73 v3 and B73 v4 reference genomes). These were present in 30,239 loci with a median of 4 and mean of 5 SNPs per locus. As were the loci themselves, these SNPs were more abundant within genes (Figure 2B) and showed enrichment near start codons (Figure 2C). The median and mean physical distance between loci was 5.7 and 67.9 kb respectively, and the median and mean physical distance between SNPs was 14 and 12,854 bp, respectively. Previous studies, based on surveys of diverse samples of maize have catalogued 1,230 (MaizeSNP50 BeadChip; Ganal *et al.* 2011) and 64,662 (HapMap2; Chia *et al.* 2012) of the SNPs found here.

Empirical evaluation of the accuracy of GBS on heterozygous samples

The repeatability and accuracy of GBS was assessed using replicate samples and parent-hybrid trios, respectively (Table 1). At 12X coverage, the repeatability of genotyping the same DNA was 99.8% (based on 265,664 total comparisons). The accuracy of genotyping assessed in trios (i.e. observed parent genotypes serving as the expectation for predicting an F_1 genotype) was >99.6% across all monomorphic and polymorphic loci. For expected heterozygotes specifically (i.e. only sites that were polymorphic between parents of a hybrid), accuracy was >99.0%. Inspection of incorrectly predicted genotypes revealed discrepancies from either putative

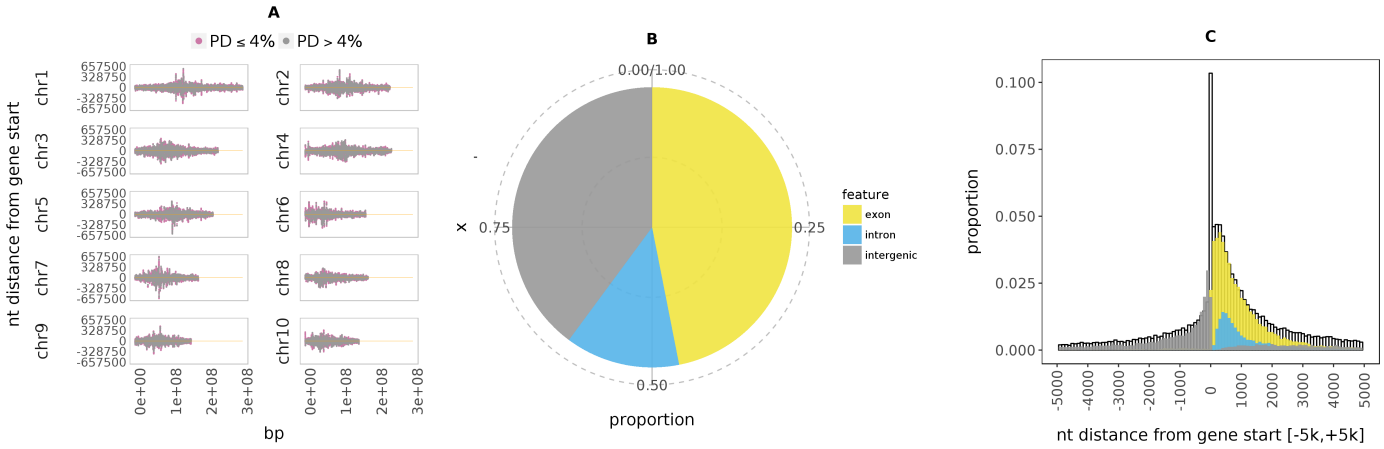

Figure 2 Genomic distribution of *Ngo*MIV-*Csp*6I loci and SNPs relative to genes in maize. (A) Nucleotide distance from each *Ngo*MIV-*Csp*6I *in silico* digested locus to its nearest start codon in the genome (y-axis), relative to the genomic location of the *Csp*6I cut site (x-axis). Two categories of percent distance (PD) are summarized: "unique" loci (greater than 4%; grey points) are overplotted on "repetitive" loci (less than or equal to 4%; red points). **(B)** Proportional distribution of all SNPs discovered in this study with respect to genome annotation categories (~60% fall within genes). **(C)** Distribution of the distance between SNPs and the nearest start site of a gene (shown: 80% of all SNPs which were located within 5 kb of a start codon). Negative values indicate SNPs upstream of the start site or 5'-end of the gene. The distribution of *in silico* digested 4PD loci is plotted as open black bars. Proportions were computed separately for SNPs and *in silico* digested loci, as a function of their respective totals.

heterogeneity of the lines or apparent mis-scoring by GATK. There were one, 12, and zero markers that were monomorphic between the parents of CML373 X CML341, CML341 X CML277, and Tzi9 X CML258, respectively, which were called as heterozygotes in the hybrid (putative heterogeneity). There were 26, 52, and 24 polymorphic SNPs between the corresponding parents that were called as homozygotes in the hybrid. Among these SNPs, there were 9, 41, and 10, respectively, that had a skewed SNV read-depth distribution in the hybrid (the minor SNV read depth proportion was no greater than ~6% of the total read count), despite a median read-depth of 45X at these loci. For the other 17, 11, and 14 SNPs, where the minor SNV read depth proportion was relatively high (minimum ~31%) and there were >10 reads for the minor SNV, the base-quality score at the variant sites in many of the reads was too low to be considered in the genotype call. Thus, considering these latter discrepancies as bonafide inaccuracies in genotyping, along with the repeatability error rate, the overall accuracy of GBS as applied in this study was estimated as 99.4%.

Surveying the same loci across GBS libraries

When working with population samples that are larger than the multiplex size of our GBS design, we found that capturing the same SNP loci across independently constructed libraries was a challenge. This is a critical issue for studies where imputation of missing data is not an option or has unacceptably low accuracy. Our initial protocol for GBS used a single size-selection step before PCR amplification and sequencing. Automated size-selection (BluePippin; Sage Science, Inc.), with the same type of gel cassette and under the same run settings, yielded differences in the peak and range of the distributions of size-selected fragments as high as 75 bp and 41 bp, respectively (Table S1). Introducing a second size-selection step after PCR minimized these differences to 38 bp and 36 bp, respectively (Table S1). Not surprisingly, libraries with similar insert size ranges showed greater correspondence in scored SNPs than those with dissimilar insert size ranges (Table S2).

Read-based haplotyping using LocHap-GBS

Variant calling was performed on 276 samples that included four test trios (each in duplicate), 23 parents of the maize NAM population, and 234 F_2 individuals derived from the CML373 x CML341 hybrid (associated with one of the trios). After filtering, there were 29,706 bi-allelic SNPs (12X read depth) across 9,667 digestion loci. In LocHap-GBS, the loci where reads stack up and searches for phased MNVs occur, are referred to as windows. Currently, phasing in LocHap-GBS is limited to two to three jointly heterozygous sites (phasing of more heterozygous sites is under development). Because individuals may be heterozygous across more than three sites within a window, LocHap-GBS automates the determination of subwindows that are delineated based on the nearest sets of three (or two) SNPs that are found to be jointly heterozygous in any one individual in the *vcf* file. However, subwindows and outputted MNVs can be derived from more than the SNP sites that define a subwindow. This happens for SNPs within a subwindow that are found only in the homozygous state across all individuals in the data set, in which case MNVs can exceed a length of three nucleotides. As an extreme example, subwindows would correspond to the original windows for a set of inbred lines where each individual is homozygous for every SNP, such that the length of MNVs would equal the number of SNP sites within a window. Also, the calling routine maintains all of the data in the *vcf* input such that some subwindows may capture individual SNPs and the *hcf* output may contain a mixture of MNPs and SNPs.

In the 9,667 digestion loci with SNPs, 10,693 subwindows were delineated that included 7,749 MNPs and 2,944 SNPs. The number of SNPs within MNP subwindows ranged from 2 to 15, with ~90% of them having five or fewer SNPs. For MNPs, three subwindows contained the observed maximum of 14 MNVsa among individuals, while ~91% of the subwindows had five or fewer MNVs. LocHap-GBS led to a 3X reduction in the dimensionality of the genotype matrix and a 1.5X increase in the median number of variants per locus (an increase from 2 to 3 variants per locus).

The trios were used to confirm whether the MNP genotypes

Table 1 Assessment of Typing Accuracy for GBS

	Description	CML373 X CML342	CML341 X CML277	Tzi9 X CML258
A	Bi-allelic SNPS	29,706	29,706	29,706
B	Not missing in trio	26,216	24,819	21,927
C	Consistent across replicates	26,114	24,729	21,878
D	Genotyping error: 1-(C/B)	0.004	0.004	0.002
E	Not heterozygous in either parent	25,811	24,420	21,555
F	Called accurately for SNPs in E	25,784	24,356	21,531
G	Genotype accuracy: F/E	0.999	0.997	0.999
H	Polymorphic between parents	5,698	5,004	5,373
I	Called accurately for loci in H	5,672	4,952	5,349
J	Heterozygote genotype accuracy: I/H	0.995	0.990	0.996
K	Not called as heterozygote for H: H-I	26	52	24
L	Low minor allele depth in K	9	41	10
M	Approximately equal allele depth in K	17	11	14
N	Adjusted genotype accuracy: G-D	0.995	0.994	0.997

scored by LocHap-GBS were valid. For each trio, subwindows were excluded if they contained missing data, had inconsistent calls between replicate samples, or contained heterozygous SNP genotypes in the parents. The MNVs expected to be observed in the hybrid were determined directly from the filtered SNP data on the parental lines (not by LocHap-GBS). These were then compared to MNVs called by LocHap-GBS. Among 25,404 MNVs recorded for the parents of all trios (ignoring whether MNVs were shared between trios), six were scored differently by LocHap-GBS in the hybrids. All of these were associated with the inaccurately scored genotypes noted previously.

When evaluating unique haplotype numbers for each subwindow we noticed some MNP loci in the hybrids that had more than two MNVs, which is not possible for maize given that it is diploid. Inspection of the raw sequence data at these loci revealed no barcode swapping errors. Examining the transmission of MNVs in the trios and F_2 population allowed us to rule out DNA cross-contamination and conclude that most, if not all of these corresponded to divergent paralogs in the sequenced samples that collapsed onto a single locus from the B73 reference genome; we refer to these as CNPs. For instance, a CML373 X CML341 F_1 sample had 21 loci with three or four MNVs that were consistent in replicate samples. For all of these loci, one or both of the inbred parents had more than one MNV (this was determined by maintaining heterozygous SNP genotypes in the inbred parents, which were filtered when estimating accuracy above) which matched each of the MNVs found in the hybrid. These same MNVs were found in the F_2 population, and tests of co-segregation indicated that 11 of them were genetically linked, but some of these loci included a mixture of linked and unlinked MNVs (data not shown).

Under the assumption that loci with greater than one MNV in an inbred line represent CNPs, among the 23 inbred parents of the NAM population there was an average of 155 (2.0%) total CNPs per individual (maximum of 305 CNPs for CML52). The cumulative number of unique CNPs was recorded with the addition of each inbred line inserted in order of their estimated genetic distance from B73 (Figure 3). There was a median increase of 30 unique CNP-associated loci with exactly two MNVs, culminating in a total of 1,195 such loci among all of the lines (Figure 3). This is an upper-bound estimate of the number of CNPs, since CNPs are conflated with heterozygous loci in inbred lines. For a lower-bound estimate, loci with more than two MNVs per line were recorded, which showed an average of 2 (0.03%) total CNPs per line (maximum of 8), a median increase of 2 CNPs with each additional line added to the dataset, and a cumulative total of 41 unique CNPs. Taken together, the estimated proportion of CNPs typed in this study among the 23 parental lines of the NAM population lies between 0.5% and 15.4% (Figure 3). Excluding the reference line B73, the correlation between genetic distance and number of unique CNPs per additional line was not significantly different from zero for both classifications of CNPs.

Genetic diversity based on SNPs versus MNPs

Results from the analysis of genetic diversity among parents of the NAM population were compared using SNP and MNP datasets. Excluding CNP-associated genotypes (MNP2 and MNP3 datasets) from the MNP1 dataset showed no significant difference in the distribution of pairwise genetic distances and all three datasets were almost perfectly correlated ($r > 0.999$; Mantel test: $p = 0.001$). Therefore, the following results are reported only for the SNP and MNP1 datasets. Although the distance matrices were significantly correlated ($r = 0.890$; Mantel test: $p = 0.001$), the average pairwise distance for SNPs was lower than that for MNPs (0.28 and 0.37 respectively) and the range in the distances was smaller for SNPs (0.12) than for MNPs (0.20). Although the mean bootstrap probabilities were effectively identical for the consensus trees based on SNPs (79.3%) and MNPs (78.9%), Robinson-Foulds distance between those trees was 12 (maximum possible distance was 40). Consequently, multidimensional scaling with MNPs produced greater separation among the lines (Figure 4A) and resulted in differences in their topology (Figure 4B; e.g., c.f. CML103, Ky21 and CML52).

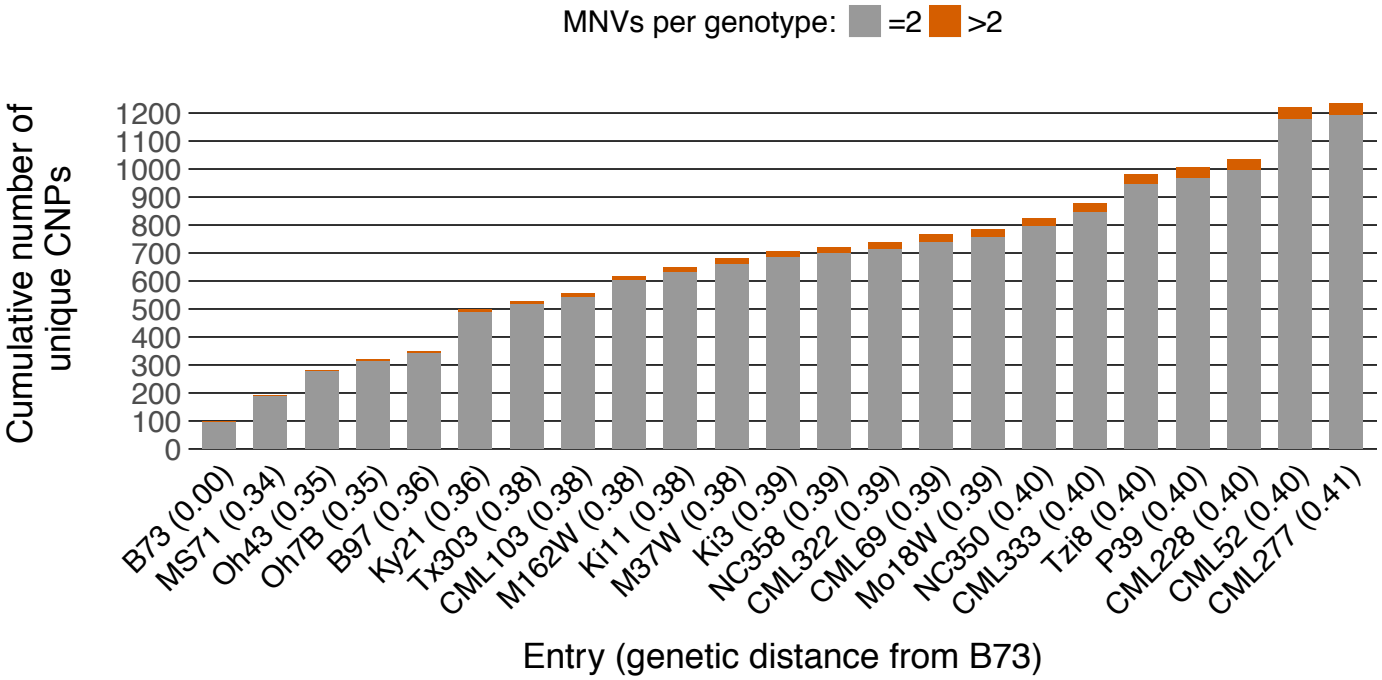

Figure 3 CNP-associated loci in the NAM inbred parents. From left-to-right, each bar indicates the cumulative number of unique CNPs (loci containing more than one MNV) in each individual compared to its preceding set of lines, with the lines ordered by their MNP1-based genetic distance from B73. The grey portion of the bar is the number of loci that had exactly two MNVs and the orange portion of the bar is the number of loci that had more than two MNVs.

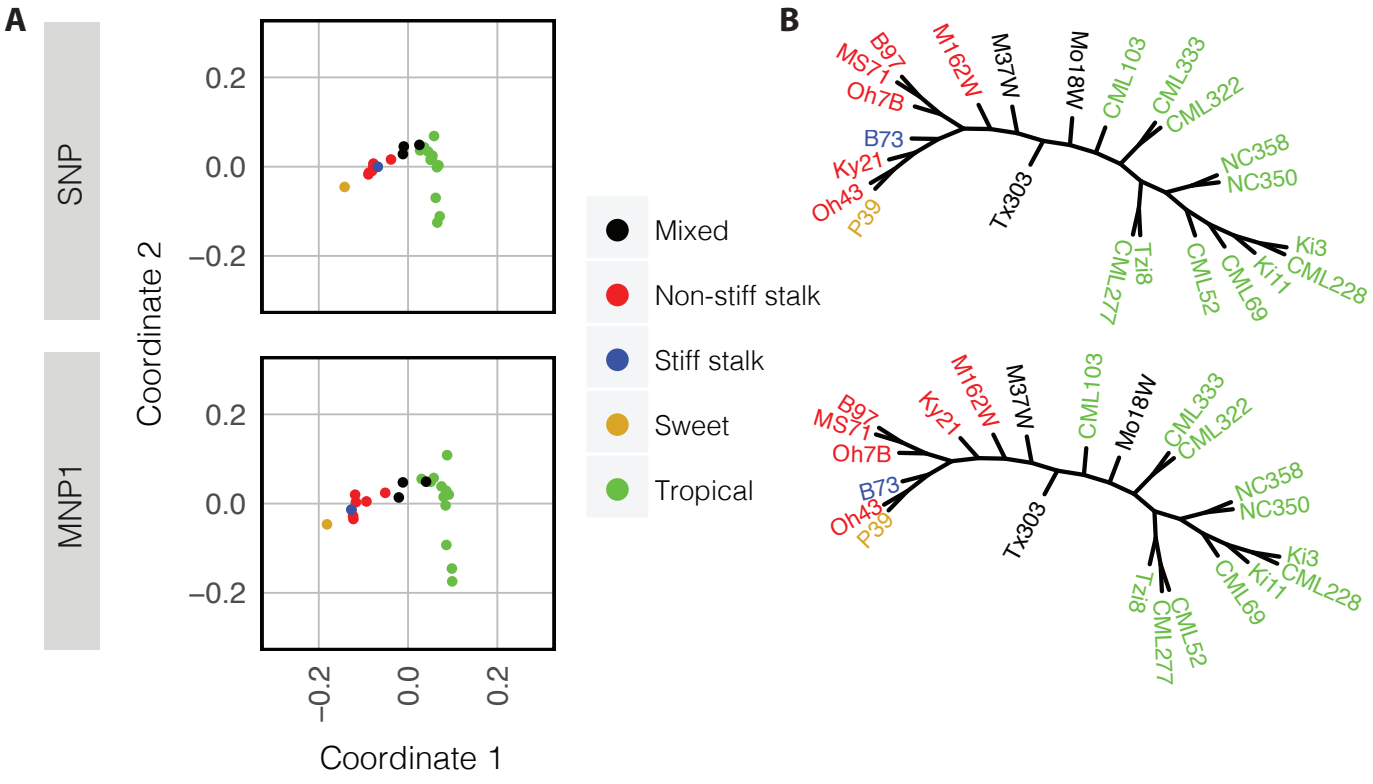

Figure 4 Analysis of genetic diversity with SNP and MNP data. (A) MDS plots based on shared allele distance for SNP (top) and MNP1 (bottom) data on 23 maize inbred lines. **(B)** Corresponding Neighbor-Joining trees. Colors represent previously assigned population structure (Liu *et al.* 2003).

DISCUSSION

Many implementations of GBS are not optimal for genotyping of heterozygous individuals or populations with unknown parentage. Imputation has proven successful on biparental families of recombinant inbred lines, but falls short when applied to populations that have low linkage disequilibrium (the median imputation accuracy was shown to be between 15-80% for markers with a $r^2 < 0.3$; He *et al.* 2015). The few approaches that have been developed for genotyping heterozygous populations use population-specific filters and depend on relatedness among individuals to score genotypes or require targeted sequencing to obtain read depths sufficient for calling heterozygotes (Uitdewilligen *et al.* 2013; Gardner *et al.* 2014; Barba *et al.* 2014; Hyma *et al.* 2015). HetMappS (Hyma *et al.* 2015) was developed specifically for dense genetic maps, and markers are filtered based on expected genotype ratios for pseudo-testcross-markers. This method was successful at increasing marker density, but was designed for use in F_1 populations only. Here, we have presented a GBS protocol and bioinformatic pipeline that, at least for maize, produces highly accurate genotype data on heterozygous individuals without requiring information on parentage or family structure nor imputation to obtain high call rates on the typed SNPs or MNPs.

As discussed by Peterson *et al.* (2012), there is a balancing act in deciding which enzymes and size-selection windows should be used when implementing GBS. In maize, which has been extensively explored for SNPs, where the enzyme *Ape*KI has been used routinely for GBS, the *Ngo*MIV-*Csp*6I enzyme combination provided an effective means for accurate genotyping that led to the discovery of many new SNPs. Assuming there is a 50% chance of sequencing each allele at a bi-allelic marker, a minimum read depth of 12X predicts that the binomial probability of sequencing each allele at least twice is >99.6%. In this study, accuracy was estimated experimentally using parent-hybrid trios to be >99.4% (Table 1), which fit closely to the expected sampling probability. Compared to an existing catalogue of 52,340,265 SNPs (Ganal *et al.* 2011; Chia *et al.* 2012) identified among diverse individuals of maize that included all but four of the ones used here (excluding: CML10, CML258, CML373, Tzi9), this study discovered approximately 100,000 new SNPs which comprised nearly two-thirds of the total SNPs found. These were enriched near genes, which was in part due to the distribution of 4PD loci, but may also be attributed to patterns in methylation: *Ngo*MIV is averse to the CpG methylation which is predominant in the intergenic repeat space of the maize genome (Antequera and Bird, 1988). Enrichment of the typed loci around the start codon of genes (Figure 2C) might be because the starts of monocot genes are GC-rich (Glémin et al. 2014) and *Ngo*MIV recognizes a GC-rich recognition site.

With the decline in cost for HTS, there is growing opportunity to apply GBS to thousands of samples for genetic studies. Sequencing the same loci within and across separate GBS libraries is required for this, but whether the same loci are sequenced has not been examined much in the literature (we are aware of one exception: DaCosta and Sorenson 2014); perhaps because GBS studies tend to rely on imputation of missing data. It has been reasoned that size selection would play an important role in sequencing shared loci (Ali *et al.* 2016). Peterson *et al.* (2012) showed that narrower size-selection windows increased the sequencing depth of shared loci in a library. As a cautionary note, some size-selection windows contain an abundance of repetitive loci (Figure S1). Here, we addressed the issue of sampling shared loci across different libraries, reaching the conclusion that two size-selection steps were needed to maximize the number of shared loci sequenced. Importantly, our results led us to the realization that the standard Y-adapters used for HTS libraries may form structures that migrate slower than fully dsDNA in dye-free agarose gels. This can lead to inconsistent size selection when using automated size-selection instrumentation (Sadaf Hoda; Sage Science, pers. comm.). Because of the inconsistency in size selection of libraries with Y-adaptors, we had to reduce the pre-PCR size-selection window and introduce a post-PCR size-selection on the dsDNA created by PCR amplification. This minimized the variation in size-selection between libraries and maximized the recovery of shared GBS loci.

This study expands the usefulness of GBS data, allowing for phased genotyping of MNPs, which enhances the information available for genetic studies. For example, founder haplotypes that are resolved for multi-parental populations can be used to impute missing data in the progeny (Gatti *et al.* 2014) or to initialize imputation (Davies *et al.* 2016). Also, the use of haplotypes may prove useful in uncovering regions associated with adaptive traits in both model and non-model species (Lorenz *et al.* 2010). The extension of LocHap for GBS was developed as a new tool for phased genotyping using standard file formats produced by most mapping and variant calling software. LocHap-GBS may be advantageous over other current tools capable of scoring haplotypes, such as STACKS and Haplotag, in that statistically vetted MNVs can be scored by LocHap-GBS on individuals without population filters only definable for specific types of samples. However, LocHap-GBS may not be readily integrated by tools that use customized data formats, though the algorithm developed for local haplotyping could (Sengupta *et al.* 2015). Finally, the software GATK used to score SNPs can also generate phased genotypes (McKenna *et al.* 2010). RedRep (Figure 1) was developed using the GATK UnifiedGenotyper, which does not have MNP phasing functionality (GATK HaplotypeCaller does). We have not experimented with the current phasing tool of GATK, but we note some unique features of LocHap-GBS. It uses a distinct algorithm for identifying MNVs, can be run independently with standard input files, reports phased genotypes in a different format (which we consider more appropriate for direct downstream analysis), and can score more than two haplotypes at a locus —this led to our finding of CNPs in maize.

Phased genotype data, while decreasing the marker dimensionality of the genotype matrix, was more informative than SNP data. Previously characterized lines of maize were more differentiated and showed some differences in relatedness when analyzed using MNPs compared to SNPs (Figure 4). In addition, phased-GBS revealed CNPs that were validated by genetic transmission yet undetectable in the underlying SNP data. These can be explained as divergent paralogs within the sample genomes that are present as only one copy in the reference genome. Because the SNP data was filtered to examine only 4PD loci, it seems unlikely that the CNPs were a result of mapping to sequences that have been collapsed in the assembly of the B73 reference genome, but this is a possibility that could result in reproducibly inaccurate CNPs. Nevertheless, having included B73 in our set of sequenced individuals, which is the same line used to sequence the reference genome, we determined that our B73 sample harbored no GBS loci with more than two MNVs, while other inbred lines did (Figure 3). There were, however, nearly 100 GBS loci that were heterozygous in the B73 sample. These may be attributed to residual heterozygosity in the line, errors in the assembly, or CNPs that had arisen during seed increases of the stock. Inspecting the genomic distribution of these loci showed no pattern of clustering as one might expect for residual heterozygosity, suggesting one or both of the latter two

explanations.

Although our study was not designed for CNP discovery, it led to the identification of at least 41 CNP-associated loci (potentially around one thousand) among 23 inbred lines of maize. The average read-depth at CNP-associated loci was 77, which was about the average read-depth of non-CNP associated loci. However, the read-depth range for CNP-associated loci was 1,480 (max: 1,492 reads) compared to 755 for non-CNP associated loci (max: 767 reads), which is expected if there are multiple paralogous loci mapping to the same place in the reference genome. Moreover, despite an average depth of sequence coverage that gives a high likelihood for detecting at least 10 CNVs, no more than four CNVs at a locus were found within an individual. There was no relationship between genome-wide estimates of genetic distance and the number of unique CNP-associated loci (Figure 3), suggesting the evolution of these putative CNPs may be different from SNPs and MNPs. This study demonstrates the potential of extending GBS for phased genotyping. As applications of GBS expand beyond bi-parental mapping populations, we foresee numerous benefits to typing MNPs by phased-GBS.

ACKNOWLEDGEMENTS

This project was supported by the Agriculture and Food Research Initiative Competitive Grant No. 2011-67003-30342 from the USDA National Institute of Food and Agriculture (Agriculture and Natural Resources Science for Climate Variability and Change Program). Other support from the University of Delaware Bioinformatics Core Facility, including use of the BIOMIX computational cluster, was made possible by Delaware INBRE (NIH P20 GM103446). We thank Dr. Karol Miaskiewicz at the Delaware Biotechnology Institute for assistance using BIOMIX. We thank Dr. Patricia Klein at TAMU for suggesting we consider *Ngo*MIV.

LITERATURE CITED

Ali, O. A., S. M. O'Rourke, S. J. Amish, M. H. Meek, G. Luikart, *et al.*, 2016 RAD Capture (Rapture): flexible and efficient sequence-based genotyping. Genetics **202**: 389–400.

Andolfatto, P., D. Davison, D. Erezyilmaz, T. T. Hu, J. Mast, *et al.*, 2011 Multiplexed shotgun genotyping for rapid and efficient genetic mapping. Genome Res **21**: 610–617.

Andrews, K. R., J. M. Good, M. R. Miller, G. Luikart, and P. A. Hohenlohe, 2016 Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet **17**: 81–92.

Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, *et al.*, 2008 Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One **3**.

Barba, P., L. Cadle-Davidson, J. Harriman, J. C. Glaubitz, S. Brooks, *et al.*, 2014 Grapevine powdery mildew resistance and susceptibility loci identified on a high-resolution SNP map. Theor Appl Genet **127**: 73–84.

Beissinger, T. M., C. N. Hirsch, R. S. Sekhon, J. M. Foerster, J. M. Johnson, *et al.*, 2013 Marker density and read depth for genotyping populations using genotyping-by-sequencing. Genetics **193**: 1073–1081.

Bowcock, A. M., A. Ruiz-Linares, J. Tomfohrde, E. Minch, J. R. Kidd, *et al.*, 1994 High resolution of human evolutionary trees with polymorphic microsatellites. Nature **368**: 455–457.

Catchen, J., P. A. Hohenlohe, S. Bassham, A. Amores, and W. A. Cresko, 2013 Stacks: an analysis tool set for population genomics. Mol Ecol **22**: 3124–3140.

Chia, J. M., C. Song, P. J. Bradbury, D. Costich, N. de Leon, *et al.*, 2012 Maize HapMap2 identifies extant variation from a genome in flux. Nat Genet **44**: 803–807.

DaCosta, J. M. and M. D. Sorenson, 2014 Amplification biases and consistent recovery of loci in a double-digest RAD-seq protocol. PLoS One **9**.

Davies, R. W., J. Flint, S. Myers, and R. Mott, 2016 Rapid genotype imputation from sequence without reference panels. Nature Genetics **48**: 965–969.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, *et al.*, 2011 A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One **6**.

Ferdosi, M. H., J. Henshall, and B. Tier, 2016 Study of the optimum haplotype length to build genomic relationship matrices. Genet Sel Evol **48**.

Fu, Y.-B. and G. W. Peterson, 2011 Genetic diversity analysis with 454 pyrosequencing and genomic reduction confirmed the eastern and western division in the cultivated barley gene pool. Plant Genome **4**: 226–237.

Ganal, M. W., G. Durstewitz, A. Polley, A. Bérard, E. S. Buckler, *et al.*, 2011 A large maize (Zea mays L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. PloS One **6**.

Gardner, K. M., P. J. Brown, T. F. Cooke, S. Cann, F. Costa, *et al.*, 2014 Fast and cost-effective genetic mapping in apple using next-generation sequencing. G3 **4**: 1681–1687.

Gascuel, O., 1997 BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol **14**: 685–695.

Gatti, D. M., K. L. Svenson, A. Shabalin, L.-Y. Wu, W. Valdar, *et al.*, 2014 Quantitative trait locus mapping methods for Diversity Outbred Mice. G3 **4**: 1623–1633.

He, S., Y. Zhao, M. F. Mette, R. Bothe, E. Ebmeyer, T. F. Sharbel, J. C. Reif, and Y. Jiang, 2015 Prospects and limits of marker imputation in quantitative genetic studies in european elite wheat (triticum aestivum l.). BMC genomics **16**: 168.

Heffelfinger, C., C. A. Fragoso, M. A. Moreno, J. D. Overton, J. P. Mottinger, *et al.*, 2014 Flexible and scalable genotyping-by-sequencing strategies for population studies. BMC Genomics **15**: 979.

Hyma, K. E., P. Barba, M. Wang, J. P. Londo, C. B. Acharya, *et al.*, 2015 Heterozygous mapping strategy (HetMappS) for high resolution genotyping-by-sequencing markers: A case study in grapevine. PLoS One **10**.

Jombart, T., 2008 adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics **24**: 1403–1405.

Kruskal, J. B., 1964 Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika **29**: 1–27.

Li, H., 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 .

Liu, K., M. Goodman, S. Muse, J. S. Smith, E. Buckler, *et al.*, 2003 Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. Genetics **165**: 2117–2128.

Lorenz, A. J., M. T. Hamblin, and J.-L. Jannink, 2010 Performance of single nucleotide polymorphisms versus haplotypes for genome-wide association analysis in barley. PlOS ONE **5**.

Lu, Y., T. Shah, Z. Hao, S. Taba, S. Zhang, *et al.*, 2011 Comparative SNP and haplotype analysis reveals a higher genetic diversity and rapider LD decay in tropical than temperate germplasm in maize. PLoS One **6**.

Lund, A. H., M. Duch, and F. S. Pedersen, 1996 Increased cloning efficiency by temperature-cycle ligation. Nucleic Acids Res **24**:

800–801.
Mantel, N., 1967 The detection of disease clustering and a generalized regression approach. Cancer research **27**: 209–220.
Martin, M., 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal **17**: 10–12.
McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, *et al.*, 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res **20**: 1297–1303.
Mcmullen, M. D., S. Kresovich, H. S. Villeda, P. Bradbury, H. Li, *et al.*, 2009 Genetic properties of the maize nested association mapping population. Science **325**: 737–740.
Neph, S., M. S. Kuehn, A. P. Reynolds, E. Haugen, R. E. Thurman, *et al.*, 2012 BEDOPS: high-performance genomic feature operations. Bioinformatics **28**: 1919–1920.
Paradis, E., J. Claude, and K. Strimmer, 2004 APE: analyses of phylogenetics and evolution in R language. Bioinformatics **20**: 289–290.
Peterson, B. K., J. N. Weber, E. H. Kay, H. S. Fisher, and H. E. Hoekstra, 2012 Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One **7**.
Poland, J. A., P. J. Brown, M. E. Sorrells, and J.-L. Jannink, 2012 Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One **7**.
Robinson, D. F. and L. R. Foulds, 1981 Comparison of phylogenetic trees. Mathematical Biosci **53**: 131–147.
Rocher, S., M. Jean, Y. Castonguay, and F. Belzile, 2015 Validation of genotyping-by-sequencing analysis in populations of tetraploid alfalfa by 454 sequencing. PLoS One **10**: 1–18.
Schliep, K., 2011 phangorn: phylogenetic analysis in r. Bioinformatics **27**: 592–593.
Sengupta, S., K. Gulukota, Y. Zhu, C. Ober, K. Naughton, *et al.*, 2015 Ultra-fast local-haplotype variant calling using paired-end DNA-sequencing data reveals somatic mosaicism in tumor and normal blood samples. Nucleic Acids Res **44**.
Sonah, H., M. Bastien, E. Iquira, A. Tardivel, G. Légaré, *et al.*, 2013 An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. PLoS One **8**.
Tenaillon, M. I., M. C. Sawkins, A. D. Long, R. L. Gaut, J. F. Doebley, *et al.*, 2001 Patterns of DNA sequence polymorphism along chromosome 1 of maize (Zea mays ssp. mays L.). P Natl Acad Sci **98**: 9161–9166.
Tinker, N. A., W. A. Bekele, and J. Hattori, 2016 Haplotag: software for haplotype-based genotyping-by-sequencing analysis. G3 **6**: 857–863.
Torkamaneh, D., J. Laroche, and F. Belzile, 2016 Genome-wide SNP calling from genotyping by sequencing (GBS) data: A comparison of seven pipelines and two sequencing technologies. PLoS One **11**.
Uitdewilligen, J. G. A. M. L., A. M. A. Wolters, B. B. D'hoop, T. J. A. Borm, R. G. F. Visser, *et al.*, 2013 A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. PLoS One **8**.