Marker Density and Read Depth for Genotyping Populations Using Genotyping-by-Sequencing

Timothy M. Beissinger,*,† Candice N. Hirsch,‡,§ Rajandeep S. Sekhon,*, Jillian M. Foerster,* James M. Johnson,* German Muttoni,* Brieanne Vaillancourt,‡,§ C. Robin Buell,‡,§ Shawn M. Kaeppler,*,** and Natalia de Leon*,**,[1]**

*Department of Agronomy, †Department of Animal Sciences, and **Department of Energy Great Lakes Bioenergy Research Center, University of Wisconsin, Madison, Wisconsin 53706, and ‡Department of Plant Biology and §Department of Energy Great Lakes Bioenergy Research Center, Michigan State University, East Lansing, Michigan 48824

ABSTRACT Genotyping-by-sequencing (GBS) approaches provide low-cost, high-density genotype information. However, GBS has unique technical considerations, including a substantial amount of missing data and a nonuniform distribution of sequence reads. The goal of this study was to characterize technical variation using this method and to develop methods to optimize read depth to obtain desired marker coverage. To empirically assess the distribution of fragments produced using GBS, ~8.69 Gb of GBS data were generated on the *Zea mays* reference inbred B73, utilizing *Ape*KI for genome reduction and single-end reads between 75 and 81 bp in length. We observed wide variation in sequence coverage across sites. Approximately 76% of potentially observable cut site-adjacent sequence fragments had no sequencing reads whereas a portion had substantially greater read depth than expected, up to 2369 times the expected mean. The methods described in this article facilitate determination of sequencing depth in the context of empirically defined read depth to achieve desired marker density for genetic mapping studies.

HIGH-density genotypic information on large numbers of individuals is crucial for quantitative trait locus (QTL) mapping and association analysis. Cost efficiency is an important component in generating genotypic data. Previous genotyping methods include markers such as microsatellites, amplified fragment length polymorphisms (AFLPs), and restriction fragment length polymorphisms, among others. The relatively high cost and limited marker density of these methods led to the use of single-nucleotide polymorphisms (SNPs) as the current preferred genotyping system. As such, a large number of array-based SNP genotyping platforms are currently available (reviewed by Fan *et al.* 2006) as well as targeted or whole-genome sequencing-based technologies.

Sequencing-based approaches to SNP allele calling include whole-genome sequencing (Hillier *et al.* 2008), exome capture (Ng *et al.* 2009), RNA sequencing (Hansey *et al.* 2012), methylated DNA sequencing (Brunner *et al.* 2009), and restriction enzyme (RE) digestion (reviewed by Davey *et al.* 2011). RE-based approaches include restriction-site associated DNA sequencing (RAD-seq) (Baird *et al.* 2008), complexity reduction of polymorphic sequences (CRoPS) (Van Orsouw *et al.* 2007), and genotyping-by-sequencing (GBS) (Elshire *et al.* 2011). All of these methods represent efficient and cost-effective approaches to produce genetic information, but differ in their implementation. For instance, RAD-seq and GBS both involve sequencing DNA fragments adjacent to RE cut sites, yet while RAD-seq involves sequencing these fragments to high coverage, the focus of GBS is to sequence with low target coverage. Alternatively, CRoPS is based on sequencing DNA fragments that were originally generated as AFLP markers. In this study, we utilized the GBS protocol of Elshire *et al.* (2011) with minor modifications.

In brief, GBS utilizes RE digestion to preferentially target sites in low-copy genomic regions, minimizing reads in repetitive sequences that are abundant in maize (Schnable *et al.* 2009) and that produce ambiguous SNP information. *Ape*KI is a RE frequently used for GBS in maize because it cuts retrotransposons infrequently and is partially methylation sensitive, thereby preferentially generating fragments from

doi: 10.1534/genetics.112.147710
Manuscript received November 25, 2012; accepted for publication February 6, 2013
Available freely online through the author-supported open access option.
Sequence data have been deposited in the NCBI Short Read Archive (BioProject accession no. PRJNA169551).
[1]Corresponding author: Department of Agronomy, University of Wisconsin, 1575 Linden Dr., Moore Hall Room 459, Madison, WI 53706. E-mail: ndeleongatti@wisc.edu

low-copy genic regions (Elshire *et al.* 2011), but additional enzymes can also be used for GBS. Cost efficiency is achieved by multiplexing barcoded individuals (Baird *et al.* 2008; Elshire *et al.* 2011). Analysis pipelines rely on mapping the resulting fragments to a reference genome, if available. Otherwise, linkage relationships can be used to genetically map sequenced DNA in organisms without a sequenced reference genome. Recently, numerous researchers have successfully employed RE-based genotyping protocols to develop maps and/or map QTL in such species (Amores *et al.* 2011; Chutimanitsakun *et al.* 2011; Pfender *et al.* 2011; Baxter *et al.* 2011; Poland *et al.* 2012).

Although all of the sequencing-based genotyping strategies are capable of generating substantial amounts of data in a cost-efficient manner, peculiarities of the genome structure can limit the utility of the data. For instance, features such as the presence of repetitive DNA make the unique alignment of sequence reads difficult or impossible. This is particularly problematic in plant species because of the high proportion of repeats frequently present (Treangen and Salzberg 2011). Similarly, repetitive DNA that is not accounted for in reference genomes may allow repetitive sequences to be aligned uniquely and therefore cause false polymorphisms to be identified. Finally, differences in guanine–cytosine (GC) content and other potential sources of sequencing biases can leave important genomic regions under- or overrepresented (Minoche *et al.* 2011).

Addressing this, theoretical work has been done to determine the expected sequence coverage obtained from these technologies (Lander and Waterman 1988; Wendl 2006). Also, studies investigating desirable marker coverage for genotype–phenotype associations in the context of classical genotyping technologies have been performed (Piepho 2000). Herein, we describe theoretical and empirical considerations of using GBS (Elshire *et al.* 2011) for genetic analysis, with the goal of determining reasonable marker expectations and corresponding resource investments. GBS was conducted on the maize reference inbred, B73, using replicated DNA samples, barcodes, and independent sequencing lanes to gather empirical information. We then compared the theoretical coverage distribution to the actual distribution that was obtained through GBS. Next, we developed a theoretical tool to determine the appropriate marker number for QTL mapping in biparental populations, as well as assessed the marker number required for association mapping in diverse inbred populations. Finally, we provided recommendations for the target number of raw sequence reads that should be generated to attain an effective density of markers. Although our results pertain to GBS in maize, simple adjustments make our techniques potentially applicable to a wide variety of protocols and species.

Materials and Methods

Library construction, sequencing, and read mapping

DNA was isolated from pooled leaf tissue from 5–10 seedlings of the reference inbred line B73. Genomic DNA was extracted using a modified CTAB method (Saghai-Maroof *et al.* 1984). Multiple DNA extractions from B73 tissue were performed. Next, extracted DNA was barcoded and pooled following the procedure described by Elshire *et al.* (2011), with an additional gel-based size selection step to enrich for fragments of intermediate size. The size selection was incorporated because Illumina reports that to optimize cluster formation the ideal fragment size range for single-end libraries is 150–300 bp (http://www.illumina.com/support/faqs.ilmn). Additionally, the size selection step allows for further reduction of the effective genome size. Sequencing was conducted using Illumina TruSeq SBS 36-bp kits, versions 3–5, on eight lanes of the Illumina Genome analyzer II (GAII) (Illumina, San Diego) at the University of Wisconsin Biotechnology Center (Madison, WI). For each library, 48 barcoded samples were pooled. The eight lanes of sequences were generated over multiple sequencing runs that were run to variable read lengths. For each lane of sequence, read quality was evaluated based on the Illumina purity filter, percentage of low-quality reads, and distribution of Phred-like scores at each cycle. Lanes that had a lower quartile Phred-like score <20 prior to base 40 were not included in this analysis. Individual reads from sequencing lanes that passed this quality control (four lanes of 75-bp reads, one lane of 76-bp reads, and three lanes of 81-bp reads) were then cleaned using the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html) with the fastx_clipper program, requiring a minimum length of 20 bp after clipping. After running fastx_clipper with both adapter sequences across the eight lanes, 84% of the reads were retained with a minimum of 80% retention from one of the lanes. Sequences from each lane that passed this filtering were parsed to remove the barcode sequences, using a custom Perl script requiring a perfect match to the barcode and the *Ape*KI cut site (*i.e.*, GC[A/T]GC). Cleaned reads were mapped to the maize B73 version 2 pseudomolecules (AGPv2; http://ftp.maizesequence.org/) (Schnable *et al.* 2009), using Bowtie version 0.12.7 (Langmead *et al.* 2009), allowing up to two mismatches and requiring a single best alignment to nuclear DNA.

Computational digestion and analysis

A custom Perl script was used to identify all *Ape*KI cut sites, irrespective of methylation state, in the maize B73 version 2 pseudomolecules (AGPv2; http://ftp.maizesequence.org/) (Schnable *et al.* 2009) and the expected fragment sizes including the GC[A/T]G overhangs were determined. GC content for a 50-bp window up- and downstream of the cut sites excluding the GC[A/T]GC sequence was determined using a custom Perl script. All subsequent analyses were performed using standard functions implemented within R version 2.13 (R Development Core Team 2011).

QTL mapping

QTL mapping was performed for two data sets. In the first, previously published data from 283 individuals of the maize

intermated B73 × Mo17 (IBM) population was analyzed (Eichten *et al.* 2011). The phenotype under study was plant height, and the total set of markers included 1340 simple sequence repeats (SSRs). These markers were not generated using GBS, but the data set was chosen because the total number of SSRs and relative spacing of the markers approximate what could be expected per individual from a low-coverage deployment of GBS. First, QTL mapping was conducted with the full data set, using composite-interval mapping with the software program R/qtl (Broman *et al.* 2003). The analysis included five covariates selected using forward selection, and the LOD threshold was determined according to a Bonferroni-corrected 0.05 significance level. The total set of identified QTL was recorded. Next, randomly chosen marker subsets were used for QTL mapping. Subset sizes ranged from 100 markers to 1300 markers, in increments of 100. For each marker subset size, QTL mapping was repeated 1000 times with randomly selected marker subsets of the specified size. The proportion of QTL in the total set that were identified in each run was recorded. Finally, the power of mapping with additional markers was evaluated based on the mean proportion of expected QTL that were identified at each marker subset size.

The second experiment involved a simulation study representing a nonintermated maize recombinant inbred (RI) population with 250 individuals. Ten QTL of equal effect were simulated, one on each chromosome, with an overall heritability of 0.5. A set of 11,917 markers was simulated, corresponding to what could be expected from a higher-coverage deployment of GBS. QTL mapping was conducted in the same manner as for the plant height data described above, except in this case the true underlying QTL were known based on the simulation. Mapping was conducted at marker subset sizes ranging from 100 to 1000, by 100, and then from 2000 to 11,917, by 1000. Again, the power of additional markers was evaluated based on the mean proportion of known QTL that were identified at each marker subset size.

Determining appropriate target coverage for mapping purposes

A bootstrapping scheme was developed to determine the genotyping resources required to obtain reads from a specified number of distinct RE fragments for an individual DNA sample. First, the set of all B73 fragments that were successfully aligned to the B73 reference genome was considered representative of the comprehensive set of all sites with the potential of being both sequenced and aligned. Next, increasing numbers of fragments were sampled from this set, with replacement. The probability of sampling each fragment was made proportional to the number of times it was actually observed. Next, the number of additional unique fragment reads that were obtained at each round of sampling was counted to estimate how many total reads are required to obtain a desired number of distinct fragments per individual.

Results

Repeated sequencing of the maize inbred line B73

In total, we generated >118 million GBS sequence reads from the reference inbred B73 to determine the distribution of reads throughout the genome. These reads corresponded to ~8.69 Gb of B73 sequence data before adapter and barcode clipping. There are ~3.9 million *Ape*KI sites in the B73 reference genome and our sequencing approach had the potential to capture up to 77 bp on either side of each cut site (up to 81 bp per read minus the 4- to 8-bp barcode). It is expected, consequently, that this sequencing should provide ~14.3× coverage of the *Ape*KI target space, assuming no size selection $[8,690,000,000/(3,936,260\times77\times2) = 14.3]$. However, due to our additional step of size selection and technical bias for smaller fragments by the Illumina procedure, the expected coverage per observable position was substantially greater. Because gel-based size selection cannot perfectly isolate fragments of a particular size, we empirically estimated the experimentally optimal fragment size by observing that 95% of the observed sequencing reads resulted from *Ape*KI fragments between 70 and 318 bp in length (here, we define *Ape*KI fragments as DNA segments between *Ape*KI cut sites). Since there are ~1.4 million optimally sized *Ape*KI fragments predicted from the B73 reference genome, our sequencing provided an expected coverage of ~40.1× over the *Ape*KI reduced and size selected space $[8,690,000,000/(1,406,269\times77\times2) = 40.1]$.

Using an informatics pipeline that allowed up to two mismatches in a read to map to the reference, we found that 43.9% of the B73 fragment reads had a single best alignment to the reference, 46.7% could be aligned to multiple positions, and 9.3% could not be aligned to the reference at all. The B73 fragments that did not align to the reference likely resulted from the requirement imposed of two or fewer mismatches for a read to be mapped, in the context of the relatively high error rate of the Illumina techonology used (Luo *et al.* 2012). Recent advances in Illumina sequencers, such as HiSeq, have reduced error rates relative to previous technologies which will improve the proportion of reads mapped. More permissive alignment algorithms may also reduce the proportion of fragments that cannot be aligned, but could increase spurious alignments in complex genomes such as maize. From the 3,936,260 potential *Ape*KI cut sites identified in the reference genome, only 35.7% (1,406,269) were expected to generate at least one fragment in the optimal size range of 70–318 bp. It was found that 27.4% (384,887) of these cut sites had at least one sequence read on at least one side of the cut site with a unique alignment, although some were sequenced many more times. Additionally, we obtained 174,954 uniquely aligned reads from *Ape*KI fragments that were larger or smaller than the 70- to 318-bp range (Figure 1). Finally, we were able to uniquely align reads from 52,123 sites that were not predicted to be *Ape*KI-site adjacent based on requiring a perfect cut site sequence in the reference genome. These unpredicted cut sites could be the result of errors

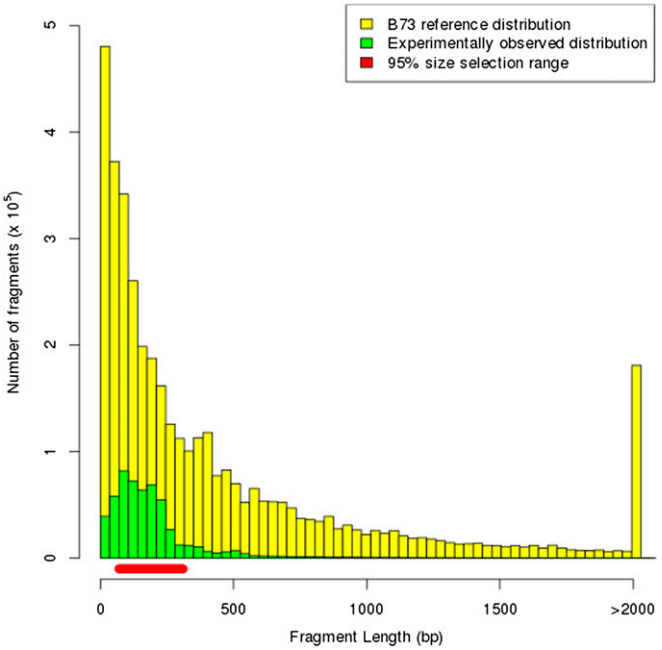

Figure 1 Distribution of the length of B73 *Ape*KI fragments expected based on an analysis of the reference genome and experimentally observed from ~8.69 Gb of B73 DNA sequence reads.

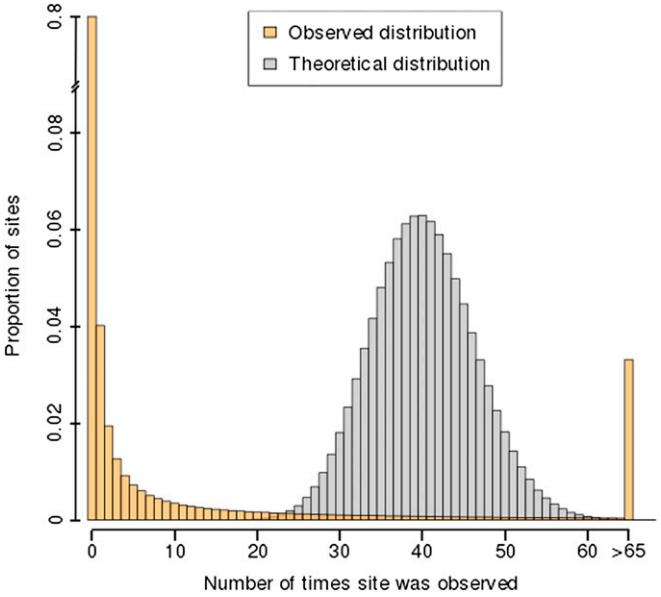

Figure 2 Observed and theoretical frequency distributions of the number of times that optimally sized B73 *Ape*KI fragments were sequenced. Note the break in the vertical axis. "Sites" refers to DNA segments from either end of an *Ape*KI fragment. The number of reads per site is expected to follow a Poisson distribution with mean equal to the average coverage.

in the reference, <100% restriction accuracy of *Ape*KI, or differences between our B73 source and that used to produce the reference sequence.

The number of reads from each fragment is expected to follow a Poisson distribution (Lander and Waterman 1988); however, the empirical data did not follow such expectation (Figure 2). Additionally, the median number of reads per sequence fragment was zero, which was substantially less than expected (~40). The likely cause for this deviation is that many of the sequence fragments were observed thousands of times more than expected (up to 95,014 reads from a single end of an *Ape*KI fragment). At the same time, a disproportionately large number (1,021,382 or 72.6%) of the predicted *Ape*KI fragments of size 70–318 bp had no observed sequencing reads from either end (Figure 2).

Analysis of the overrepresented fragments relative to the B73 organelle reference genome revealed that a subset of these fragments were also present in DNA from organelles. Hence, they correspond to historical insertions of organellar DNA into the nuclear genome, an occurrence that has been documented previously (Lough *et al.* 2008). Other overrepresented fragments are likely due to repeats in the maize genome that include *Ape*KI cut sites that were collapsed into a single, nonrepetitive segment in the reference sequence. In these instances, the repeated fragments were mapped uniquely because the reference genome does not capture their repetitive nature. Highly overrepresented sites (>500 reads per site, corresponding to nearly a 0% probability based on the expected distribution) represented 0.5% of the sequenced *Ape*KI sites, but accounted for 41.7% of the total reads.

To further investigate the reason for the highly overrepresented sites, potential biases due to fragment GC content were explored. High or low content of GC within a fragment can affect read depth using Illumina sequencing technologies (Minoche *et al.* 2011). Technological advances have reduced this bias over time, but nevertheless, fragments extremely high or low in GC content are likely to be underrepresented. We assessed the GC content in windows of 50 bp adjacent to cut sites. The mean level of GC for these windows across B73 was 53.9% with a range of 0–100% (Figure 3A). It was observed that sites with GC content between 40 and 50% were more frequently sequenced using Illumina GAII. When GC content was outside of a seemingly ideal 10–70% range, the mean number of times that sites were sequenced decreased from 12.8 to 0.46 (Figure 3B). But, regions with such high or low GC content accounted for <7% of all optimally sized sequencing fragments in B73. Therefore, GC content bias could explain some of the rarely observed sites but such bias does not account for the highly overrepresented sites.

The skewed distribution of sequencing reads we observed in B73 is not unique to this inbred line. Across three RI populations and an association panel (totaling ~1500 diverse inbred lines), for which we performed a comparable GBS protocol, many of the same positions that were overrepresented through sequencing B73 also had disproportionately high coverage across the set of diverse lines (data not shown).

Marker number for QTL mapping

It is desirable to optimize marker density to maximize the efficient application of genotyping technologies. For the

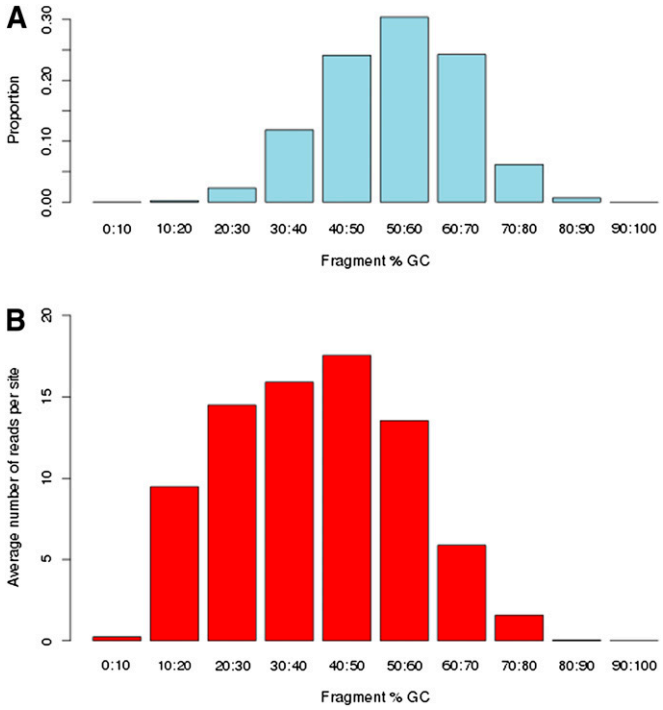

Figure 3 Distribution of GC content and coverage of optimally sized (70–318 bp) sites. (A) The proportion of optimally sized sequencing fragments with the specified GC content (computationally determined by analysis of the reference genome). (B) Mean number of reads for optimally sized B73 sequencing fragments with given GC content. Extremely high or low GC content negatively affected read number per site, but the majority of fragments are in the intermediate GC range.

purpose of QTL mapping in biparental populations, marker density can be optimized by first recognizing that adding markers to stretches of the genome that correspond to the same parental genotype does not provide additional information. On the other hand, when markers flank a recombination event, additional markers in the region will increase resolution of the recombination position. Still, if the recombination event is not in the region where a true QTL lies, there is no practical use for increased marker density in this region. Thus, the probability of having both a QTL and a recombination event occur between two markers should be minimized (Figure 4).

Consider an individual with c chromosomes, n evenly spaced markers, r recombination events, and q QTL. A lower limit, p_0, on the probability of both a recombination event and a QTL not occurring in a region with unknown parental genotype (*i.e.*, between two markers) is given by

$$p_0 = \frac{(n-r+c)^q}{(n+c)^q}. \quad (1)$$

To derive (1), consider the available markers as dividing the target DNA into $n + c$ bins, which are the spaces between two markers or between a marker and a chromosome end. The numerator of (1) is the number of ways the QTL can be placed into these bins such that they are not in a bin where a recombination event occurred, while the denominator of (1) is the total number of ways that QTL can be placed into bins. An implicit assumption is that the probability of a QTL being located in any of the bins is equal, which is met when markers are evenly spaced. Although GBS markers are not perfectly evenly spaced, there is no large-scale clustering of *Ape*KI cut sites throughout the genome. More specifically, no one chromosome or chromosomal region has an abnormally high or low concentration of cut sites. This means that even spacing is a reasonable approximation on a genome-wide basis. Also, this estimation provides a lower limit because (1) assumes that every recombination event occurs in a unique bin—if there happen to be multiple recombination events in any single bin, the numerator will be slightly increased. To obtain the number of markers (n) that will provide an expected proportion of p_0 individuals without any particular QTL flanked by markers of alternate parent genotypes, one may solve the above formula for n, with $q = 1$, which is given by

$$n = \frac{r - c(1-p_0)}{1-p_0}. \quad (2)$$

Equation 2 was verified based on two QTL mapping experiments. The first experiment consisted of previously published data on plant height in the IBM population (Eichten *et al.* 2011), while the second was based on simulations for a nonintermated RI population of 250 individuals. In both cases, the optimal number of markers suggested by (2) was within 200 of the observed marker number that provided a maximally powerful test (Figure 5). Furthermore, this confirms that the assumption of even spacing is robust against minor violations, as the IBM markers used were not evenly spaced. Interestingly, the simulated QTL mapping experiment data depicted a slight decrease in power with an unnecessarily dense set of markers. This is likely attributable to the Bonferroni correction for LOD threshold that was implemented.

Marker number for mapping with recombinant inbred and association populations

RI populations are commonly used for QTL mapping. Application of Equation 2 requires knowledge of the genome-wide recombination rate (r) and the number of chromosomes (c). In the maize IBM RI population, for example, the average number of effective recombination events per individual is 57 (Fu *et al.* 2006). Since the IBM population was intermated four times before selfing (Lee *et al.* 2002), this value can be scaled to nonintermated standard RI populations by first multiplying by the reciprocal of the genetic map expansion factor incurred during the development of the IBM population and then by the expansion factor incurred during the development of a standard RI population. Respectively using the expansion factors $j/2 + (2^i - 1)/2^i$ and $(2^{i+1} - 1)/2^i$, for lines that have

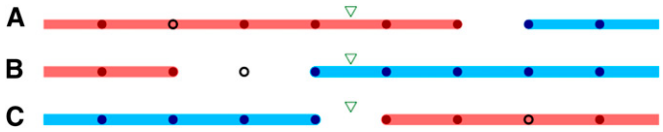

Figure 4 An example of genotypes for three hypothetical RI lines, A, B, and C. Red circles correspond to observed marker genotypes from one of the parental lines, blue circles correspond to observed maker genotypes from the other parent, and open circles correspond to missing marker information. Red and blue shading illustrates that between two markers of the same parental genotype, genotypes can be inferred with great accuracy, even in the case of a missing marker genotype. However, genotypes between markers of alternate parental types remain unknown. The green arrowheads show the location of a "true" quantitative trait locus (QTL). Note that line C has unknown genotype at the QTL and therefore does not add power to a statistical test for QTL identification (although this individual would be particularly useful for downstream fine mapping). Equations 1 and 2 provide the number of markers needed for the probability of occurrence of case C to be minimized.

been inbred for i generations after being intermated for j generations (Teuscher *et al.* 2006), on average, there are ~38 recombination events per individual in a nonintermated maize RI population.

Application of (2) to the IBM RI population given that $r = 57$ and $c = 10$, and allowing an expected $p_0 = 95\%$ of individuals without uncertain genotype at each QTL, we determined that the marker goal across the genome is $n = 1130$. For a standard RI population, applying (2) with the same parameters except that $r = 38$ results in the genome-wide marker goal being 750. We emphasize that these marker number requirements represent the minimum number of markers needed to produce an expected 95% of individuals with known genotypes at any particular QTL.

It is important to note that in RI populations, the genotyped markers need not be the same for each individual. In the case where several of the markers typed on each individual are distinct, as occurs in GBS, imputing markers that were not observed is the appropriate action. In the case of hundreds of individuals, each with mostly different markers typed, this will lead to a large proportion of imputed markers for the population. However, as long as the number of markers that were observed for each individual meets the values described above, mapping will have near the maximum power possible for the particular population under study. Moreover, although imputing between observed markers allows comparisons to be made between individuals with observed genotypes at different markers, it cannot increase mapping resolution (Figure 4). Therefore, a distinction must be made between the total number of GBS markers generated and a value that we deem the "effective number of markers." The effective number of markers for an individual is the number of markers with observed genotypes. Thus, a RI line typed at n markers but with unobserved genotypes at p markers before imputation has an effective marker number with respect to resolution of n, not of $n + p$.

Also, the marker number suggested by (2) should be viewed as a minimum number of effective markers for mapping purposes. Additional markers will provide only minimal increased power to detect QTL, but they will reduce the proportion of individuals with uncertain parental genotypes due to recombination events near the QTL to fewer than the expected value of $1 - p_0$. It is important to highlight that based on the estimation that maize contains ~39,500 genes (AGPv2; http://ftp.maizesequence.org/) (Schnable *et al.* 2009), the marker numbers suggested here will generate maps that have, on average, ~35 genes between markers in the IBM population and ~53 genes between adjacent markers in nonintermated RI populations. However, improving QTL mapping resolution requires not just additional markers but also an increase in recombination events, which can be achieved only with an altered population structure or size.

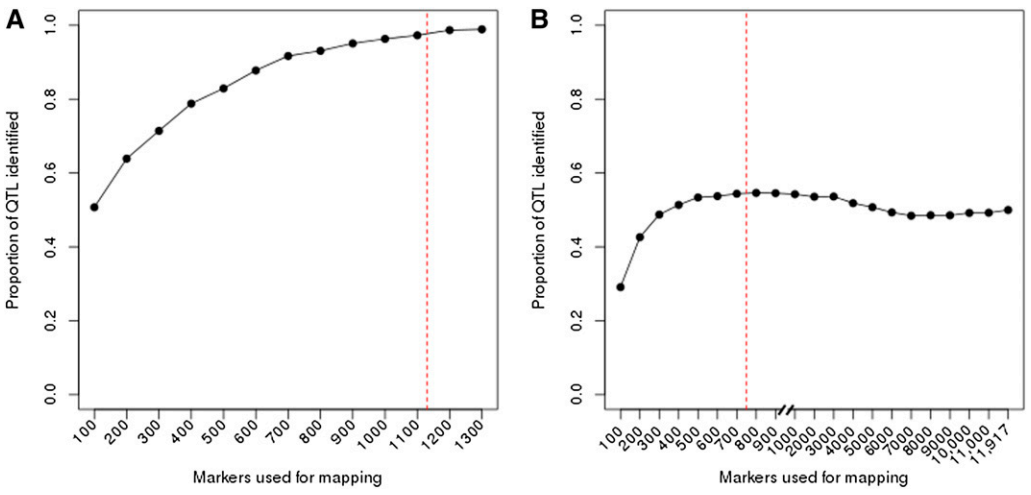

Figure 5 Validation of marker number estimate. Two quantitative trait loci (QTL) mapping studies were performed to validate Equation 2, which estimates the number of markers required to maximize the power of a biparental QTL mapping study based on the number of chromosomes and level of recombination in a population. Depicted in both A and B is the mean proportion of QTL identified from 1000 replicated mapping experiments at each marker subset level. (A) For the intermated B73 × Mo17 (IBM) RI population, the maximum number of QTL that could be identified was three, which was the number identified from mapping with the full data set. (B) For the simulated RI population, which was not intermated before inbred development, the maximum number of QTL that could be identified was all 10 QTL simulated. In each plot, the red line depicts the number of markers suggested by Equation 2. For experimental data from the IBM RI population, as well as data from a simulated nonintermated RI population, Equation 2 closely approximates the ideal marker number for maximal QTL identification.

Similar to QTL mapping with a RI population, the goal of association mapping using a diverse set of inbred lines is to associate genotypes with phenotypes. However, these methods differ in that RI populations have a simpler structure of relatedness generated by the expected recombination of regions originated from the two parents of the population, whereas a more complex structure is present in a diverse population. Similarly, greater levels of linkage disequilibrium (LD) are expected to be present in RI populations compared to diverse populations. Therefore, it is well established that more markers are required for association mapping to capture markers within historical blocks of LD compared to structured biparental populations. In maize, for instance, it has been suggested that association mapping should be conducted with SNP markers spaced every 100–200 bp (Tenaillon *et al.* 2001). In other species, the required density of SNP markers for effective association mapping is dependent upon the level of historical LD.

Determination of read depth required to achieve desirable marker density based on an empirical distribution of read coverage per marker position

The number of unique fragment reads that should be expected for a given total number of fragment reads was quantified to determine the optimal depth at which sequencing should be performed to achieve a desirable marker coverage based on an empirically determined distribution of reads. The quantification is based on a resampling of the data we generated from the maize reference inbred B73. The approach utilized is likely to provide a slight bias toward B73 fragments. For other lines or species, the proportion of reads that can be aligned to the reference genome is likely to vary. An adjustment for the proportion of fragments that can be aligned is needed for this method to be globally applicable to other maize lines, and a repetition of the process will apply for other species.

To quantify optimal target depth in a RI population, the number of unique fragments required per individual will vary with target marker number and average SNP density between the parents. But, given a target number of unique fragments per individual, a recommendation for the total number of reads that should be obtained for each sample is provided (Figure 6). Based on resampling-observed B73 fragments, the expected number of unique fragments sequenced for a given number of total fragments sequenced incorporates the uneven coverage distribution. The focus is on unique fragments because these have the potential to contribute additional information. However, if the sequencing technology used is error prone, repetitive sequencing of sites may be required, and the approach utilized here can be modified to evaluate the expected number of fragments sequenced a specified number of times for a given total number of fragments sequenced.

As described above, diverse association panels require substantially more markers than do RI lines for effective mapping. With millions of *Ape*KI sites in maize, GBS based on that RE seemingly has the potential to generate marker densities near the target. However, based on the uneven coverage of sites that we observed, GBS would have to be performed with substantially greater depth than calculated simply by reads divided by target sites to obtain information at the majority of the desired sites. For instance, from the ~8.69 Gb of B73 data generated in our study from 118 million reads, only 559,841 unique *Ape*KI fragments of the 1.4 million expected to pass our size selection step were successfully sequenced and aligned. It appears that the additional ~840,000 fragments had an extremely low probability of being captured through sequencing. Given this, and the fact that LD decays over a span of only a few hundred base pairs in maize (Tenaillon *et al.* 2001), relying on downstream LD-based imputation for those sites that were missed is expected to be relatively ineffective. Instead, a reasonable approach is to minimize the amount of missing data by sequencing fewer sites at a higher target coverage, taking into account the variable sequencing depth that will be observed. Our resampling analysis suggests that using *Ape*KI-based GBS in maize, genotyping with a target of 23, 41, or 80 million reads is expected to result in missing data at ~30%, 20%, or 10% of sites, respectively, for a given individual (Figure 6). Determination of the appropriate target number of sequence reads in different species or by the use of alternative sequencing-based genotyping methods can be achieved by first sequencing a representative individual at high coverage and subsequently performing empirical resampling to identify the point of adequate coverage as suggested here.

Discussion

We have shown that the coverage of different sites throughout the maize genome as captured through the *Ape*KI-based GBS protocol is highly variable, although the reasons for the extreme variability are only partially understood. Therefore, sequencing approaches that succeed even when coverage is variable, or approaches that reduce the uneven coverage, are necessary. Alternative sequencing approaches for genotyping individuals are abundant, including GBS with different or multiple enzymes (Poland *et al.* 2012), RAD-seq (Baird *et al.* 2008), and CRoPS (Van Orsouw *et al.* 2007). In situations where highly variable levels of coverage are still observed, the strategy proposed here first operates on a single individual (B73 in this case) to be sequenced extensively. The variability of site coverage in this individual will approximate the variability yet to be generated from later individuals. From the full set of sequenced fragments obtained from the first individual, including repeated fragments, random computer-based subsamples are drawn, with replacement. These are evaluated for the amount of additional sites observed as subsample size increases. The subsample size that provides enough site information for the desired marker number or level of missing data dictates the coverage that should be targeted.

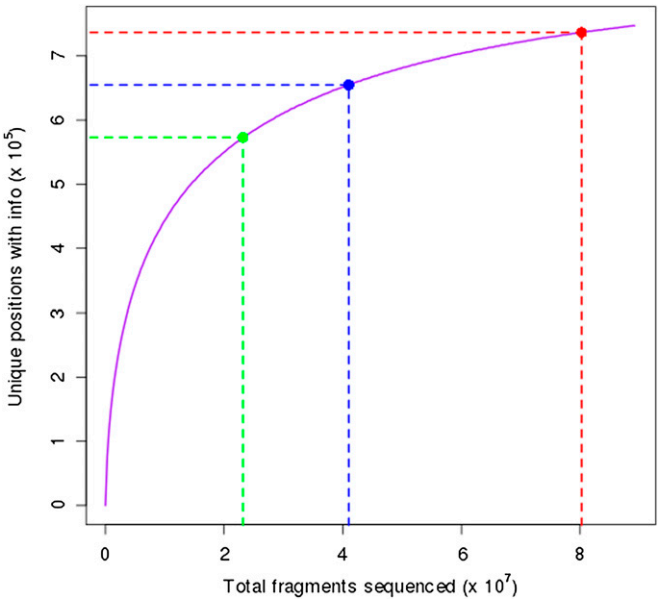

Figure 6 Resampling method to determine target sequencing depth. A resampling analysis was conducted to determine the number of total fragment reads needed to achieve desirable levels of coverage. Plotted is how the number of uniquely identifiable sequenced DNA fragments resulting from sheared *Ape*KI fragments varies with the total number of sequenced DNA fragments. Results were generated based on empirically determined frequencies of fragment reads from ~8.69 Gb of B73 DNA sequence reads. The red, blue, and green points highlight the number of total fragment reads necessary to observe 90%, 80%, and 70% of the potential fragments, respectively.

Carrying out this strategy in maize suggested, for example, that acquiring 300,000 unique fragments per individual can be obtained by sequencing ~3.6 million total fragments per individual (Figure 6). But because of the variable coverage distribution, doubling the number of acquired unique fragments to 600,000 requires a more than ninefold increase in total fragments sequenced, to ~27.9 million. Moreover, our results show that to minimize the level of missing data across individuals even more sequencing per individual is required (Figure 6). However, these target sequencing levels will vary and potentially be reduced in species with less repetitive genomes.

Although the sequencing coverage required to generate a dense marker set seems daunting, we have demonstrated that a substantial marker density is not required, or even useful, for the purpose of QTL mapping in RI populations. For this type of population structure, desirable marker coverage is given by Equation 2. Representative populations suggest that the number of markers for efficient QTL mapping in biparental populations is on the order of hundreds to thousands. Substantially larger numbers of markers would be necessary, however, when association mapping is being conducted in a diverse population.

In summary, performing GBS on the maize inbred line B73 produced a highly skewed coverage of genomic positions, which is only partially accounted for by GC bias and duplicated positions. The result of the uneven coverage distribution is that no information is available at the majority of positions for which information was initially expected. Still, our findings suggest that even at relatively low coverage, GBS can produce enough information for powerful QTL mapping in biparental populations. However, obtaining dense genotyping resolution for downstream fine mapping will require increased target coverage per individual. Using the method for association studies in maize, for example, will require genotyping at substantially greater target coverage. Therefore, researchers must be aware that in complex genomes, using simple approximations and standard distributions to determine target coverage vastly underestimates the sequence depth required to generate adequate data for complex analyses. Before large-scale sequencing commences, empirical enumerations of target coverage that account for potentially complicated genome compositions will lead to more complete and useful data sets relative to study goals.

Acknowledgments

This work was funded by the Department of Energy (DOE) Great Lakes Bioenergy Research Center (BER) (DOE BER Office of Science grant DE-FC02-07ER64494). T.M.B. and J. M.F. were supported by a gift to the University of Wisconsin, Madison, Plant Breeding and Plant Genetics program from Monsanto. G.M. was supported by a fellowship from DuPont–Pioneer Hi-Bred International. J.M.J. was supported by Hatch funds from the National Institute of Food and Agriculture, United States Department of Agriculture Project WIS01330.

Literature Cited

Amores, A., J. Catchen, A. Ferrara, Q. Fontenot, and J. H. Postlethwait, 2011 Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted gar, an outgroup for the teleost genome duplication. Genetics 188: 799–808.

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(10): e3376.

Baxter, S. W., J. W. Davey, J. S. Johnston, A. M. Shelton, and D. G. Heckel, 2011 Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. PLoS ONE 6(4): e19315.

Broman, K. W., H. Wu, S. Sen, and G. A. Churchill, 2003 R/qtl: QTL mapping in experimental crosses. Bioinformatics 19: 889–890.

Brunner, A. L., D. S. Johnson, S. W. Kim, A. Valouev, T. E. Reddy *et al.*, 2009 Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. Genome Res. 19: 1044–1056.

Chutimanitsakun, Y., R. W. Nipper, A. Cuesta-Marcos, L. Cistue, A. Corey *et al.*, 2011 Construction and application for QTL analysis of a restriction site associated DNA (RAD) linkage map in barley. BMC Genomics 12: 4.

Davey, J. W., P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen *et al.*, 2011 Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. 12: 499–510.

Eichten, S. R., J. M. Foerster, N. de Leon, Y. Kai, C. Yeh *et al.*, 2011 B73-Mo17 near-isogenic lines demonstrate dispersed structural variation in maize. Plant Physiol. 156(4): 1679–1690.

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(5): e19379.

Fan, J., M. S. Chee, and K. L. Gunderson, 2006 Highly parallel genomic assays. Nat. Rev. Genet. 7: 632–644.

Fu, Y., T. Wen, Y. I. Ronin, H. D. Chen, L. Guo *et al.*, 2006 Genetic dissection of intermated recombinant inbred lines using a new genetic map of maize. Genetics 174: 1671–1683.

Hansey, C. N., B. Vaillancourt, R. S. Sekhon, N. de Leon, S. M. Kaeppler *et al.*, 2012 Maize (*Zea mays* L.) genome diversity as revealed by RNA-sequencing. PLoS ONE 7(3): e33071.

Hillier, L. W., G. T. Marth, A. R. Quinlan, D. Dooling, G. Gewell *et al.*, 2008 Whole-genome sequencing and variant discovery in *C. elegans*. Nat. Methods 5(2): 183–188.

Lander, E. S., and M. S. Waterman, 1988 Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics 2: 231–239.

Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg, 2009 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10(3): R25.

Lee, M., N. Sharapova, W. D. Beavis, D. Grant, M. Katt *et al.*, 2002 Expanding the genetic map of maize with the intermated B73 x Mo17 (IBM) population. Plant Mol. Biol. 48: 453–461.

Lough, A. N., L. M. Roark, K. Akio, T. S. Ream, J. C. Lamb *et al.*, 2008 Mitochondrial DNA transfer to the nucleus generates extensive insertion site variation in maize. Genetics 178: 47–55.

Luo, C., D. Tsementzi, N. Kyrpides, T. Read, and K. T. Konstantinidis, 2012 Direct comparisons of Illumina *vs.* Roche 454 sequencing technologies on the same microbial community DNA sample. PLoS ONE 7(2): e30087.

Minoche, A. E., J. C. Dohm, and H. Himmelbauer, 2011 Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. Genome Biol. 12: R112.

Ng, S. B., E. H. Turner, P. D. Robertson, S. D. Flygare, A. W. Bigham *et al.*, 2009 Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461: 272–276.

Pfender, W. F., M. C. Saha, E. A. Johnson, and M. B. Slabaugh, 2011 Mapping with RAD (restriction-site associated DNA) markers to rapidly identify QTL for stem rust resistance in *Lolium perenne*. Theor. Appl. Genet. 122: 1467–1480.

Piepho, H. P., 2000 Optimal marker density for interval mapping in a backcross population. Heredity 84: 437–440.

Poland, J. A., P. J. Brown, M. E. Sorrells, and J. Jannink, 2012 Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS ONE 7(2): e32253.

R Development Core Team, 2011 R: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. Available at http://www.r-project.org/.

Saghai-Maroof, M. A., K. M. Soliman, R. A. Jorgensen, and R. W. Allard, 1984 Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. Proc. Natl. Acad. Sci. USA 81: 8014–8018.

Schnable, P. S., D. Ware, R. S. Fulton, J. C. Stein, F. Wei *et al.*, 2009 The B73 maize genome: complexity, diversity, and dynamics. Science 326(5956): 1112–1115.

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.). Proc. Natl. Acad. Sci. USA 98: 9161–9166.

Teuscher, F., V. Guiard, P. E. Rudolph, and G. A. Brockmann, 2005 The map expansion obtained with recombinant inbred strains and intermated recombinant inbred populations for finite generation designs. Genetics 170: 875–879.

Treangen, T. J., and S. L. Salzberg, 2011 Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet. 13: 36–46.

Van Orsouw, N. J., R. C. J. Hogers, A. Janssen, F. Yalcin, S. Snoeijers *et al.*, 2007 Complexity reduction of polymorphic sequences (CRoPS): a novel approach for large-scale polymorphism discovery in complex genomes. PLoS ONE 2(11): e1172.

Wendl, M. C., 2006 Occupancy modeling of coverage distribution for whole genome shotgun DNA sequencing. Bull. Math. Biol. 68: 179–196.

Communicating editor: C. D. Jones