Published July, 2009

Large-Scale Discovery of Gene-Enriched SNPs

Michael A. Gore,* Mark H. Wright,* Elhan S. Ersoz, Pascal Bouffard, Edward S. Szekeres, Thomas P. Jarvie, Bonnie L. Hurwitz, Apurva Narechania, Timothy T. Harkins, George S. Grills, Doreen H. Ware, and Edward S. Buckler

Abstract

Whole-genome association studies of complex traits in higher eukaryotes require a high density of single nucleotide polymorphism (SNP) markers at genome-wide coverage. To design high-throughput, multiplexed SNP genotyping assays, researchers must first discover large numbers of SNPs by extensively resequencing multiple individuals or lines. For SNP discovery approaches using short read-lengths that next-generation DNA sequencing technologies offer, the highly repetitive and duplicated nature of large plant genomes presents additional challenges. Here, we describe a genomic library construction procedure that facilitates pyrosequencing of genic and low-copy regions in plant genomes, and a customized computational pipeline to analyze and assemble short reads (100–200 bp), identify allelic reference sequence comparisons, and call SNPs with a high degree of accuracy. With maize (*Zea mays* L.) as the test organism in a pilot experiment, the implementation of these methods resulted in the identification of 126,683 putative SNPs between two maize inbred lines at an estimated false discovery rate (FDR) of 15.1%. We estimated rates of false SNP discovery using an internal control, and we validated these FDR rates with an external SNP dataset that was generated using locus-specific PCR amplification and Sanger sequencing. These results show that this approach has wide applicability for efficiently and accurately detecting gene-enriched SNPs in large, complex plant genomes.

Published in The Plant Genome 2:121–133. Published 10 July 2009.
doi: 10.3835/plantgenome2009.01.0002
© Crop Science Society of America
677 S. Segoe Rd., Madison, WI 53711 USA
An open-access publication

THE AVERAGE NUCLEOTIDE diversity of coding regions between any two maize (*Zea mays* L.) lines (π = 1–1.4%) is two- to fivefold higher than other domesticated grass crops (Buckler et al., 2001; Tenaillon et al., 2001; Wright et al., 2005). Moreover, it is not uncommon to find maize haplotypes more than 2% diverged from one another (Tenaillon et al., 2001; Wright et al., 2005) and even as high as 5% (Henry and Damerval, 1997). Intragenic linkage disequilibrium (LD) rates rapidly decline to nominal levels within 2 kb in a population of diverse maize inbred lines (Remington et al., 2001). Of the ~2500 Mb that constitutes the maize genome, less than 25% is genic or low-copy-number sequence, with large blocks of highly repetitive DNA such as retrotransposons

M.A. Gore, Dep. of Plant Breeding and Genetics, Cornell Univ., 175 Biotechnology Bldg., Ithaca, NY 14853; M.H. Wright, Dep. of Genetics and Development, Cornell Univ., 102 Weill Hall, Ithaca, NY 14853; E.S. Ersoz, Institute for Genomic Diversity, Cornell Univ., 175 Biotechnology Bldg., Ithaca, NY 14853; P. Bouffard, E.S. Szekeres, and T.P. Jarvie, 454 Life Sciences, 20 Commercial St., Branford, CT 06405; B.L. Hurwitz and A. Narechania, Cold Spring Harbor Lab., 1 Bungtown Rd., Cold Spring Harbor, NY 11724; T.T. Harkins, Roche Applied Science Corp., 9115 Hague Rd., Indianapolis, IN 46250; G.S. Grills, Life Sciences Core Labs. Center, Cornell Univ., 139 Biotechnology Bldg., Ithaca, NY 14853; D.H. Ware, USDA-ARS, Cold Spring Harbor Lab., 1 Bungtown Rd., Cold Spring Harbor, NY 11724; E.S. Buckler, USDA-ARS, Dep. of Plant Breeding and Genetics, Institute for Genomic Diversity, Cornell Univ., 159 Biotechnology Bldg., Ithaca, NY 14853. All custom code and scripts used in this study are available upon request from M. H. Wright (mhw6@cornell.edu). M.A. Gore and M.H. Wright contributed equally to this work. Received 14 Jan. 2009. *Corresponding authors (mag87@cornell.edu) and (mhw6@cornell.edu).

Abbreviations: DFCI, Dana-Farber Cancer Institute; EST, expressed sequence tag; FDR, false discovery rate; HMPR, hypomethylated partial restriction; LD, linkage disequilibrium; MAGIv4.0 C&G, Maize Assembled Genome Island Version 4.0 Contigs and Singletons; MCS, 5-methylcytosine-sensitive; NCBI, National Center for Biotechnology Information; PDL, paralog distinguishing list; SNP, single nucleotide polymorphism; UF, unfiltered.

intermixed throughout (Hake and Walbot, 1980; Meyers et al., 2001; SanMiguel et al., 1996). Retrotransposons are generally recombinationally inert, and most meiotic recombination in the maize genome is restricted to gene-rich regions (Fu et al., 2002; Fu et al., 2001; Yao et al., 2002). Association mapping strategies, which rely on ancient recombination for dissecting complex traits, require that SNPs within these recombinationally active gene regions be identified and genotyped in phenotypically diverse populations (Reviewed by Zhu et al., 2008). Because of the rapid decay of intragenic LD in a highly diverse genome with an estimated 59,000 genes (Messing et al., 2004), several million gene-enriched SNP markers may be necessary for whole-genome association studies in diverse maize (E. Buckler, unpublished).

Retrotransposons contain a higher density of methylation in the form of 5-methylcytosine relative to genic sequences—a property unique to plant genomes (Rabinowicz et al., 2003; Rabinowicz et al., 2005). HypoMethylated Partial Restriction (HMPR) is a library construction method that exploits this property to facilitate the efficient sequencing of gene rich regions in large, highly repetitive plant genomes (Emberton et al., 2005). The principle underlying HMPR is that the complete digestion of plant genomic DNA with a 5-methylcytosine-sensitive (MCS) restriction enzyme that has a 4 bp recognition sequence permits the fractionation of genic and repetitive DNA by gel electrophoresis. Large restriction fragments (20–150 kb) contain blocks of highly methylated retrotransposons, while much smaller fragments (<1000 bp) comprise a fraction that is gene-enriched (Bennetzen et al., 1994; Yuan et al., 2002). Emberton et al. (2005) used a partial digestion of maize genomic DNA with a MCS 4 bp cutter, followed by gel-purification and cloning procedures to construct maize HMPR libraries that contained larger (1–4 kb), overlapping gene fragments more suitable for Sanger sequencing read-lengths (800–1200 bases). These maize HMPR libraries showed more than sixfold enrichment for genes compared to control libraries. This level of gene enrichment was comparable to that achieved by other non-transcriptome-based gene-enrichment sequencing technologies tested on maize (Gore et al., 2007; Palmer et al., 2003; Rabinowicz et al., 1999; Whitelaw et al., 2003; Yuan et al., 2003), but maize HMPR libraries were superior for repeat elimination and enrichment of low-copy, non-coding sequences.

With the recent emergence of ‘next-generation’ DNA sequencing technologies it is technically feasible to economically and rapidly resequence hundreds of millions of bases (Reviewed by Mardis, 2008). Using these high-throughput sequencing-by-synthesis (Bennett, 2004; Margulies et al., 2005) or sequencing-by-ligation (Shendure et al., 2005) technologies in a read-to-reference based SNP discovery approach presents computational challenges because the length and quality of obtained individual reads are shorter and potentially of lower fidelity than single-pass Sanger sequencing reads. Furthermore, the maize genome is the product of ancient and perhaps more recent tetraploidization and rearrangement events (Gaut and Doebley, 1997; Swigoňová et al., 2004; Wei et al., 2007), and as a result contains a high proportion of duplicated genes (Blanc and Wolfe, 2004; Emrich et al., 2007; Messing et al., 2004). This confounds the unique mapping of short reads if duplicated genes (i.e., paralogs) are recently diverged and thus nearly identical in nucleotide sequence. Recently, a computational SNP calling pipeline built on the POLYBAYES polymorphism detection software (Marth et al., 1999) and “monoallelism” rules was developed and used to analyze expressed sequence tags (ESTs) that were obtained by 454 pyrosequencing of cDNAs prepared from two maize inbred lines (Barbazuk et al., 2007). This pipeline reduced the number of false-positive SNPs that resulted from sequencing errors and alignment of paralogous sequences, which facilitated the identification of more than 7000 putative SNPs in expressed genes.

Nonetheless, if the discovery of maize SNP markers on the order of millions is to be economically viable, the use of low cost, next-generation DNA sequencing technologies is clearly required. These high-throughput DNA sequencing technologies can be more efficiently used in the large-scale discovery of SNPs for maize association mapping studies if resequencing is concentrated within the recombinationally active gene regions of the vastly repetitive maize genome. The objectives of this study were (i) to adapt HMPR gene-enrichment sequencing to a massively parallel pyrosequencing platform and (ii) to develop a read-to-reference based SNP calling pipeline for short reads (100–200 bp) that maximizes SNP detection power, while controlling the number of detected false-positive SNPs resulting from sequencing errors and the alignment of paralogous sequences.

MATERIALS AND METHODS

DNA Isolation from Maize

We extracted nuclear DNA from nuclei prepared from etiolated (pale green), inner husk leaves (100 g) of field-grown maize inbred line B73 as previously described by Rabinowicz (2003).

A more specialized cultivation technique was required to obtain genomic DNA from maize root tissue. Kernels from maize inbred lines B73 and Mo17 were surface sterilized in a 10% (vol/vol) bleach solution (5.25% Sodium Hypochlorite) by gently rocking for 30 min, followed by 3× 10-min rinses with sterile water. The kernels were left to imbibe overnight in sterile water at room temperature with gentle rocking. Ten kernels were placed in a vertically orientated seed germination pouch (Mega International, West St. Paul, MN) and germinated in a dark growth chamber held at 28 °C. Roots of 1-wk-old maize seedlings were bulk harvested and immediately frozen in liquid N_2 prior to storage at −80 °C. Total genomic DNA was isolated from homogenized frozen 1-week-old root tissue using the DNeasy Plant Maxi Kit

(QIAGEN, Valencia, CA) according to the manufacturer's protocol.

Modified HMPR Library Construction

Complete digestions of 5 µg of maize husk nuclear DNA (B73) and seedling root total genomic DNA (B73 and Mo17) were individually performed in 100 µL volumes with 50 U of *Hpa*II (New England Biolabs, Ipswich, MA) at 37 °C for 16 h, followed by heat inactivation of the enzyme at 65 °C for 20 min. *Hpa*II fragments ranging in size from >10 kb to less than 100 bp (data not shown) were separated on a low melting 0.8% SeaPlaque agarose gel (Cambrex Bio Science Rockland, Inc., Rockland, ME). Restriction fragments ranging in size from 100 to 600 bp were excised from the gel and purified using the QIAquick Gel Extraction kit, according to the manufacturer's protocol. Gel-isolated *Hpa*II fragments were randomly ligated to each other with 1 µL of highly concentrated T4 DNA ligase (20 U/µL) (New England Biolabs, Ipswich, MA) in a total reaction volume of 20 µl at 16 °C for 16 h, followed by heat inactivation of the enzyme at 65 °C for 20 min.

Several micrograms of concatenated *Hpa*II fragments were needed for the downstream nebulization procedure (see 454 sequencing and data processing section). However, this would typically require low-throughput, large-scale DNA extractions and gel isolations, because an estimated 95% of the maize genome was intentionally discarded. Alternatively, we found it more efficient to generate microgram quantities of concatenated *Hpa*II fragments using Phi29-based isothermal amplification of long concatemer templates in a nanogram-scale reaction. Briefly, the GenomiPhi V2 DNA Amplification Kit (GE Healthcare, Piscataway, NJ) was used to amplify 1 µL of the 10 ng/µL ligation reaction per the manufacturer's instruction. This kit uses the high fidelity Phi29 (ϕ29) DNA polymerase, dNTPs, and random hexamers to replicate linear genomic DNA by multiple displacement amplification. Several independent GenomiPhi amplification reactions were performed and pooled for each library to ensure a low level of amplification-induced bias. The GenomiPhi reaction was separated on a low melting 0.8% SeaPlaque Agarose gel, and amplification products ranging in size from 3 to 10 kb were isolated from the gel with the QIAquick Gel Extraction kit and used in the downstream 454 sample preparation procedure.

454 Sequencing and Data Processing

Sequence sample preparation and data generation were performed with the Phi29 amplified *Hpa*II concatemer DNA of two B73 HMPR libraries (husk and root) and one Mo17 HMPR library (root) using the 454 GS FLX platform at 454 Life Sciences (Branford, CT). In addition, total genomic DNA isolated from the same seedling root tissue of B73 was sequenced on the same 454 platform, which served as an unfiltered (UF) genomic control to assess the level of gene-enrichment in modified HMPR libraries. Approximately 5 µg of high molecular weight DNA was fragmented by nebulization to a size range of 300 to 500 bp. Preparation of 454 libraries, emulsion-based clonal amplification, library sequencing on the Genome Sequencer FLX System as well as signal processing and data analysis were performed as previously described by Margulies et al. (2005). Also, the 454 base-calling software (version 1.1.03.24) provided error estimates (*Q* values) for each base, none of which exceeded a value of 40.

The expected yield per run of the 454 GS FLX is approximately 100 Mb, potentially more under ideal conditions. However, sequencing the B73 husk library with a single instrument run produced only 65.6 Mb of sequence because a less than optimal DNA copy per bead ratio was used for emulsion PCR. A more optimal DNA copy per bead ratio was used for the B73 root library, improving sequence yield to 101.3 Mb in a single run. The Mo17 root library was sequenced with four runs that in total yielded 236.7 Mb of sequence. This total sequence yield for the Mo17 root library was 41% lower than expected, indicating that further optimization was still needed. In addition, we sequenced (1 run; 130.9 Mb) randomly sheared B73 total genomic DNA, which served as the UF library.

The raw 454 sequencing data are available in the NCBI Short Read Archive (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi) with accession number SRA008616.

Screening and Filtering of 454 Sequences

Because modified HMPR libraries contained *Hpa*II concatemers, 454 reads generated from sequencing these libraries were digested in silico at *Hpa*II recognition sites (5′–C/CGG–3′). This was done to produce independent, non-chimeric *Hpa*II fragment sequences. All 454 reads from the UF control library and *Hpa*II fragment sequences less than 40 bp in length were discarded (Table 1). *Hpa*II fragment sequences and UF sequences (≥40 bp) were searched using BLAT (Kent, 2002) against The Institute for Genomic Research (TIGR) maize repeat database Version 4.0 (http://maize.tigr.org/repeat_db.shtml) to identify repetitive sequences. Also, sequences were searched against mitochondrial (GenBank accession no. NC_007982.1) and chloroplast (GenBank accession no. NC_001666.2) genome sequences of maize. We performed BLAT searches with default parameters, except for a tile size of 16. We considered BLAT similarities significant if the expectation value was less than 10^{-5} and the local alignment length was 40 bp or longer. Sequences that had a significant match to a repeat sequence or an organellar genome were discarded. Remaining sequences were similarly searched with BLAT against the Maize Assembled Genome Island Version 4.0 Contigs and Singletons (MAGIv4.0 C&G) database (http://magi.plantgenomics.iastate.edu/). Because a large number of sequences did not match any sequences in the MAGIv4.0 C&G database, these unmatched *Hpa*II fragment and UF sequences were also searched against the complete genome sequences of *japonica* rice (*Oryza sativa* L.) (http://rice.plantbiology.msu.edu/) and sorghum (*Sorghum bicolor* L.)

Table 1. Sequence composition of modified HMPR and UF libraries.

					Modified HMPR							
Libraries	B73 Husk			B73 Root			Mo17 Root			UF		
	No.	Mb	%[†]	No.	Mb	%	No.	Mb	%	No.	Mb	%
454 reads[‡]	391,778	65.6	–	470,918	101.2	–	1,284,692	236.7	–	543,385	130.9	–
Total[§]	479,565	63.6	100	771,557	97.6	100	1,937,032	225.5	100	543,350	130.9	100
Chloroplast	3,771	0.6	0.8	5,567	0.9	0.7	30,835	4.1	1.6	3,118	0.8	0.6
Mitochondrial	1,319	0.2	0.3	20,332	3.0	2.6	224,593	29.7	11.6	5,493	1.4	1.0
Non-maize[¶]	6,829	0.9	1.4	530,876	67.4	68.8	454,413	49.1	23.5	41,149	9.8	7.6
Repeats[#]	150,786	21.7	31.4	34,378	5.2	4.5	75,225	9.3	3.9	343,072	83.8	63.1
Non-repeats[††]	316,860	40.2	66.1	180,404	21.1	23.4	1,151,966	133.3	59.5	150,518	35.1	27.7

[†] The number of sequences in each category expressed as a percentage of the total number of sequences.

[‡] Sequencing reads generated on the 454 GS FLX.

[§] 454 reads from modified HMPR libraries were in silico digested with *Hpa*II, and only sequences ≥40 bp were kept and BLAT searched against nucleotide databases. 454 reads from the UF library were not in silico digested with *Hpa*II, and only sequences ≥40 bp were kept and BLAT searched against nucleotide databases.

[¶] Sequences that did not significantly match any of the screened plant nucleotide, organellar, or repeat databases. All of these sequences were classified as putatively non-maize with the majority of unknown or bacterial origin.

[#] Sequences from the maize nuclear genome that significantly matched to The Institute for Genomic Research (TIGR) Maize Repeat version 4 database, which consists of characterized, uncharacterized, and predicted repeats.

[††] Sequences from putatively non-repetitive regions of the maize genome with significant matches to the Maize Assembled Gene Islands Version 4.0 Contigs and Singletons (MAGIv4.0 C&S) database, sorghum or rice genome sequences, or the Dana-Farber Cancer Institute (DFCI) maize gene index.

(http://www.phytozome.net/) as well as maize expressed sequence tag (EST) sequences within the Dana-Farber Cancer Institute (DFCI) maize gene index release 17.0 (http://compbio.dfci.harvard.edu/tgi/). Sequences that did not have a significant match in any of these additionally searched databases were considered contaminant (non-maize) sequences and discarded. Summary statistics and source information for all databases are found in Supplementary Table 3.

Assembly of 454 Sequences

We assembled the retained non-repeat *Hpa*II fragment sequences into multiple sequence alignments using the CAP3 sequence assembly program (Huang and Madan, 1999). The following CAP3 assembly options were used: –p 99 (overlaps must be >99% identity), –s 401 (alignment score must be >400, minimum value allowed), –h 3 (maximum overhang of 3%), and alignment scoring options (–m 20, –n 40, and –g 21) that allowed a perfect match overlap of 40 bp to satisfy the minimum alignment score for assembly. Additionally, CAP3 computed a *Q* value for each base of the consensus sequence. Assemblies were performed separately for B73 (husk and root) and Mo17 (root) non-repeat *Hpa*II fragment sequences. We did not assemble UF sequences, as they were only used to measure the level of gene-enrichment and repeat depletion in modified HMPR libraries.

Because CAP3 could not execute with all sequences input at once, we performed a preliminary clustering of sequences into a collection of disjoint groups with no inter-group homology. Clustering was performed by a custom program in a manner equivalent to NCBI BLASTClust (available at http://www.ncbi.nlm.nih.gov/BLAST/docs/blastclust.html). We did not use BLASTClust because it could not run on our systems with the amount of input data supplied. CAP3 was then executed on each cluster separately. The preliminary clustering revealed that about 5% of sequences were still chimeric because of an *Hpa*II site that was eliminated by a sequencing error or erroneous end-joining ligation. A simple modification to the clustering algorithm allowed almost all chimeras to be detected and split before CAP3 assembly.

We developed a custom program to analyze the CAP3 assembly output and extract a consensus sequence and associated CAP3-based *Q* values from each multiple sequence assembly as well as the number of sequences concordant with each consensus base (coverage depth). Because of partial overlaps and potential disagreements among assembled reads, coverage depth as defined here is not the same as the total number of reads aligned in the multiple sequence assembly but as the number of reads with an aligned base that supports the consensus base call. *Hpa*II fragment sequences that did not assemble into multiple sequence alignments (i.e., singletons) were used directly as consensus sequences as well as the *Q* values calculated by Roche-454's base-calling software.

Construction of the Paralog Distinguishing List (PDL)

To facilitate the identification of paralogous regions, the MAGIv4.0 C&G database of B73 reference sequences was searched and aligned against itself using BLAT, as described above. All match pairs (not the alignment) with at least 90% identity and a length of 50 bp or longer were used as input for a custom polymorphism detection program. The custom polymorphism detection program performed a Smith-Waterman (Smith and Waterman, 1981) local alignment between match pairs identified by BLAT to obtain a full representation of the alignment "in memory." This allowed alignments to be quickly scanned for single base mismatches and single base insertions/deletions (in/dels). Single base mismatches and single base in/dels were identified in the Smith-Waterman local alignments and "context sequences" were extracted: the 16 bp 5′ and 16 bp 3′ flanking the mismatch or in/del. All such putative non-allelic differences were extracted as context sequences from all pairwise matches satisfying the 90% identity minimum and 50 bp minimum. These context sequences form the PDL and represent the putative fixed differences that distinguish paralogs. The PDL was used in further analysis to search for paralogous regions, as described below.

Polymorphism Detection

Consensus sequences of B73 and Mo17 *Hpa*II fragments were searched against B73 reference sequences (MAGIv4.0 C&G database) using BLAT. Match pairs (not the alignments) were used as input for the custom polymorphism detection program, as described above. Similarly, the polymorphism detection program performed a Smith-Waterman local alignment between the *Hpa*II consensus sequence and the MAGIv4.0 C&G reference sequence (i.e., match pairs) identified by BLAT to obtain a full representation of the alignment "in memory." For each single base mismatch or in/del identified by the program, context sequences for B73 and Mo17 *Hpa*II fragment sequences were extracted: the 16 bp 5′ and 16 bp 3′ flanking the mismatch or in/del. Single base mismatches or in/dels within 16 bp of either end of the local alignment were not considered.

Implementation of the PDL and SNP Calling

With the same custom polymorphism detection program, all context sequences for B73 or Mo17 *Hpa*II fragment sequences were searched against the PDL. Any match to the PDL was considered a paralogous alignment and the entire alignment and all potential SNPs within it were discarded. Otherwise, if no PDL matches were found, all in/del contexts were discarded (not called as SNPs) and the remaining single-base mismatch contexts were scanned against a list of SNPs already called. If a single duplicate context was identified in an alignment, only that context was discarded, but if two or more duplicates were identified, the entire alignment was discarded, along with all potential SNPs, even if these SNPs were novel. Provided neither the PDL nor the duplicate alignment check resulted in discarding all potential SNPs, the remaining single-base mismatches were called SNPs and no further alignments for the current *Hpa*II consensus sequence were considered. Otherwise, if the alignment was discarded, the next strongest BLAT match was considered, continuing until an alignment was accepted, or until the next strongest BLAT match was less than 95% identity. This preset 5% maximum was not restrictive for identifying allelic variation, as it is well above the average nucleotide diversity of coding regions between any two maize lines (π = 1–1.4%) (Tenaillon et al., 2001; Wright et al., 2005), but still allows the evaluation of haplotypes that are 5% diverged from one another (Henry and Damerval, 1997). Moreover, the 5% maximum allowed us to use a smaller PDL by avoiding paralogous alignments that were more diverged and easily distinguished from previously reported allelic variation levels. Identified B73/Mo17 putative SNPs and the PDL are available for download from Panzea (http://www.panzea.org).

Panzea SNP Comparison

We extracted 6094 B73 and 6200 Mo17 sequences from the Panzea database (Zhao et al., 2006) that were generated by PCR-directed Sanger sequencing of candidate gene loci. Overlapping sequences that were amplified from the same candidate gene locus were assembled using the procedure described above, except that sequences were clustered on the basis of a common Panzea locus ID. For many of the candidate gene loci, there were two independent amplifications and sequencings of B73 and Mo17 for quality control. This resulted in 3683 (1.57 Mb) and 3696 (1.57 Mb) assemblies for B73 and Mo17, respectively. We called SNPs from these sequences using the program already described, except allelic B73 and Mo17 consensus sequences were paired on the basis of common Panzea locus ID. The PDL was not used to call SNPs with Panzea sequences, because it was assumed that all Mo17/B73 pairings were allelic on the basis of single locus PCR amplification. Identified Panzea SNPs were mapped to Mo17 454 consensus sequences on the basis of the 16 bp 5′ and 16 bp 3′ context sequences, and vice versa, to identify which SNPs from each dataset were called from sequence in common to both datasets. We separately looked at the intersection of Panzea SNPs and B73/Mo17 *Hpa*II SNPs called with (126,683 SNPs; no thresholds) and without (174,476 SNPs; no thresholds) the PDL. We then compared SNPs that mapped to both datasets to estimate the rate of false SNP discovery and power, assuming that all true Mo17/B73 SNPs were discovered in the Panzea dataset and no false SNPs were discovered.

RESULTS

Construction of Modified HMPR Libraries

We modified the previously described HMPR library construction method (Emberton et al., 2005) to allow high-throughput gene-enrichment sequencing of the maize genome using the 454 Genome Sequencer FLX (GS FLX) pyrosequencing instrument (see "Materials and Methods"). *Hpa*II, a MCS 4 bp cutter (5′-C/CGG-3′), was selected to construct modified HMPR libraries, because of its strong bias for cleaving within unmethylated genic and low-copy regions of the maize genome (Antequera and Bird, 1988; Emberton et al., 2005; Yuan et al., 2002). The first of the two major modifications to the HMPR method was to allow maize genomic DNA to be completely digested with *Hpa*II rather than partially digested. This was done to produce a more repeatable *Hpa*II restriction pattern and, as a result, consistently enrich for gene fragments mostly smaller than 600 bp. Second, *Hpa*II fragments between the sizes of 100 to

600 bp were gel-isolated and converted via random ligation into concatemers of longer lengths more suitable for nebulization (i.e., fragmentation). At the time of this experiment, it was not possible for us to execute paired-end read sequencing and to routinely obtain read-lengths longer than 250 bases on the 454 GS FLX instrument; thus, we used ligation and nebulization in combination to construct and randomly break *Hpa*II concatemers in order to completely sequence larger *Hpa*II fragments.

To test and optimize our library construction method, we constructed modified HMPR libraries for maize inbred lines B73 (husk and root) and Mo17 (root). One concern with modified HMPR and its predecessor is the potential enrichment of organellar genome fragments in constructed libraries (Emberton et al., 2005), as these genomes are unmethylated (Palmer et al., 2003) and, depending on the tissue type, may be present at a very high copy number (Li et al., 2006). Thus, we evaluated as sources of genomic DNA two etiolated tissue types that were expected to have a relatively low abundance of chloroplasts: inner husk leaves (pale green) and dark-grown seedling roots (white). For inner husk leaves, purification of nuclei prior to genomic DNA extraction was used to further limit the amount of co-isolated chloroplast DNA. For dark-grown seedling roots, we used a higher yielding and less laborious total genomic DNA extraction procedure that lacked a nuclei purification step, because dark-grown seedling roots were expected to be highly deficient in chloroplasts and other types of plastids (Reviewed by Possingham, 1980).

Compositional Analysis of Modified HMPR Libraries

Modified HMPR libraries and an unfiltered (UF) B73 library were sequenced on the 454 GS FLX instrument (see “Materials and Methods”). Because the modified HMPR libraries were comprised of randomly concatenated *Hpa*II fragments (see previous section), prior to analysis 454 reads pertaining to these libraries were in silico digested with *Hpa*II to produce independent, non-chimeric sequences. To examine the sequence composition of modified HMPR and UF libraries, *Hpa*II fragment and UF sequences were searched against several plant nucleotide databases and genome sequences (see “Materials and Methods”). The distribution of sequence among these categories is shown in Table 1. A higher level of organellar contamination was found in root libraries, but this was offset by their lower level of repeats. B73 and Mo17 root libraries were seven- to eightfold lower in repeats relative to the B73 husk library, and 14- to 16-fold lower in repeats relative to the UF library. The very low repeat content of root libraries is comparable to that previously reported in maize HMPR libraries (Emberton et al., 2005) and superior to other non-transcriptome-based gene-enrichment sequencing technologies tested on maize (Gore et al., 2007; Palmer et al., 2003; Rabinowicz et al., 1999; Whitelaw et al., 2003; Yuan et al., 2003). Even though the amount of repeat sequences within modified HMPR libraries varied substantially between tissue types (e.g., B73 husk vs. B73 root), additional biological and technical replications are needed to determine if these differences are attributed to tissue-specific differential methylation of genes and repeats.

The desired enrichment for the genic fraction of the maize genome in root libraries was compromised by an abundance of sequences that did not significantly match any of the screened plant nucleotide databases or genome sequences. These unknown contaminant sequences were most prevalent in the B73 root library, comprising 68.8% of the *Hpa*II fragment sequences. We randomly sampled 1000 of these putative non-maize sequences from each root library and searched them with BLAST (Altschul et al., 1997) against NCBI’s non-redundant nucleotide database. On average, 65% of these sampled sequences had no significant similarity (cutoff E-value of 10^{-5}) to any sequence with another 30% showing different degrees of similarity to bacterial sequences (results not shown). We suspect that bacterial endo- or exo-symbionts of maize roots were living beneath the seed pericarp layer and subsequently proliferated on seedling roots. Neither the seed surface sterilization procedure nor the sterile seedling growth conditions used in this study would have eliminated any type of bacterial symbiont from seedling roots, thus allowing the co-isolation of bacterial genomic DNA and its enrichment in modified HMPR root libraries. Regardless of the source or identity of these sequences, these putatively non-maize sequences as well as the maize repeat and organellar sequences were excluded from further analyses.

To assess the degree to which modified HMPR libraries were enriched with genic sequences, we searched non-repetitive, maize *Hpa*II sequences against the MAGIv4.0 C&S database (http://magi.plantgenomics.iastate.edu/). The MAGIv4.0 C&S database is a partial genome assembly of Sanger-based BAC end and shotgun sequences, gene-enriched genome survey sequences as well as whole-genome shotgun sequences from maize inbred line B73 (Kalyanaraman et al., 2007). In addition, the MAGIv4.0 C&S database represents the most comprehensive maize genomic database in advance of the pending draft maize genome sequence (The unassembled, draft maize B73 genome sequence is a superior reference sequence, but its use in this study is restricted by the Ft. Lauderdale agreement governing the pre-publication use of large genomic datasets). The search results revealed an intermediate to high intersection (52.2–67.0%) between the MAGIv4.0 C&S database and non-repetitive *Hpa*II fragment sequences contained within modified HMPR libraries (Supplementary Table 1). Moreover, alignment to computationally predicted genes from MAGIv4.0 Contig sequences and the DFCI maize gene index (http://compbio.dfci.harvard.edu/tgi/) showed that modified HMPR libraries were four- to fivefold enriched for genes relative to the UF library (Supplementary Table 1). This level of gene-enrichment in modified HMPR libraries was similar to that obtained with the original HMPR method (Emberton et al., 2005) and other non-EST-based gene-enrichment sequencing technologies tested on maize (Gore et al., 2007;

Palmer et al., 2003; Rabinowicz et al., 1999; Whitelaw et al., 2003; Yuan et al., 2003).

Sequence Assembly and Construction of a PDL

Why is it challenging to identify SNPs in maize using next-generation sequencing technologies? Maize is hypothesized to be an ancient tetraploid (Gaut and Doebley, 1997; Swigoňová et al., 2004; Wei et al., 2007), but its genome has lost a substantial number of unlinked duplicated genes (Lai et al., 2004). However, nearly one-third of all maize genes still have a paralog (Blanc and Wolfe, 2004), and many of these paralogs are tandemly arrayed (Messing et al., 2004). It is estimated, based on ESTs, that maize paralogs resulting from an ancient tetraploid event have diverged a minimum of 10% over time (Blanc and Wolfe, 2004), but recent evidence conservatively suggests that nearly identical paralogs (≥98% identity) are almost 13-fold more frequent in the maize genome than that of *Arabidopsis* (Emrich et al., 2007). With long enough sequencing reads, unique flanking sequences can be found to distinguish recently diverged paralogs. However, it is unlikely that *Hpa*II fragment sequences, with an average length of 120 bases after in silico digestion and a higher single-read error rate than that of Sanger sequencing, will contain sufficient and accurate information to distinguish between highly similar paralogs in the maize genome. In addition, if recently duplicated genes have diverged within the range of previously reported maize nucleotide diversity levels (π = 1–5%) (Henry and Damerval, 1997; Tenaillon et al., 2001; Wright et al., 2005), it will be difficult, if not impossible, to reliably distinguish paralogs based on the best reference match, reciprocal best match, or a conservative maximum allelic diversity threshold. Finally, the MAGIv4.0 C&S reference database used for SNP calling in this study is a partial genome assembly, thus the true allelic copy for an *Hpa*II fragment sequence may not even be present in this reference database.

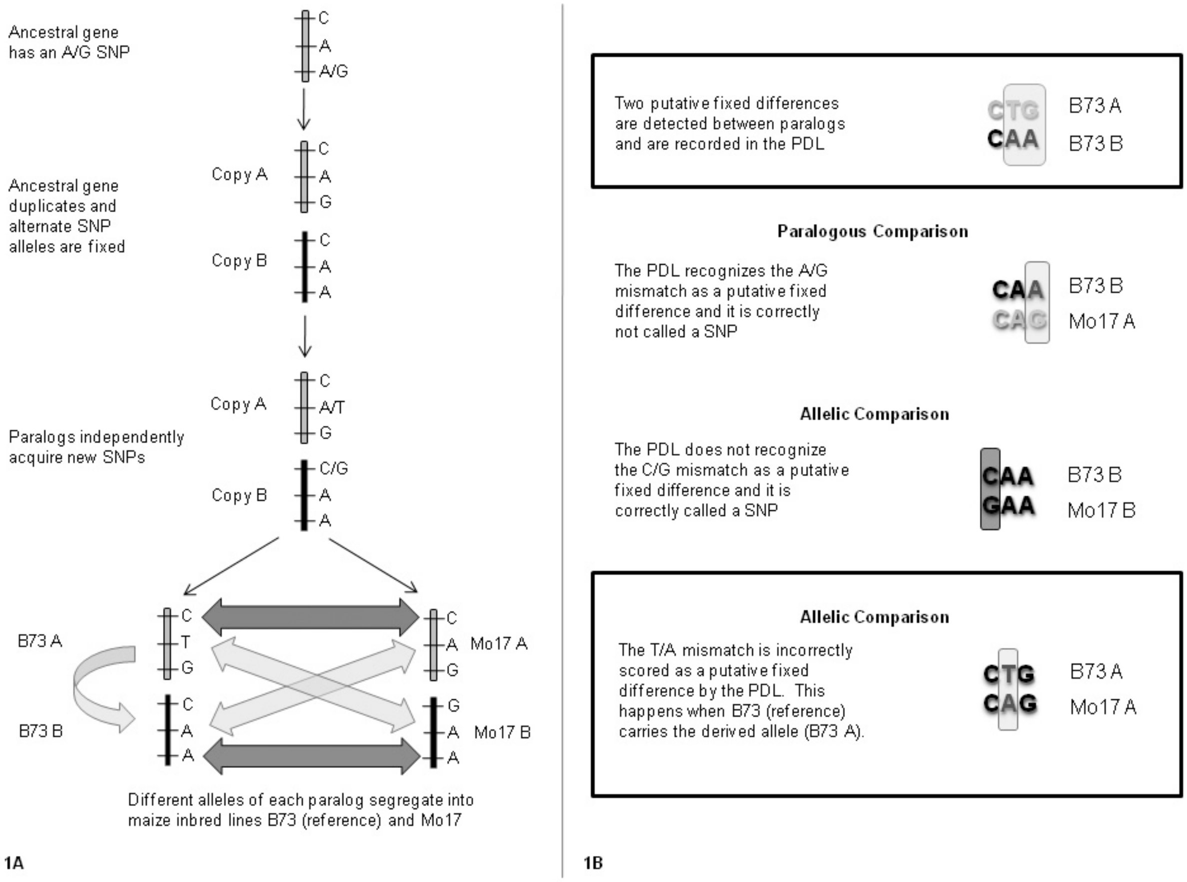
A two-pronged strategy was developed to deal with some of these challenges. First, the redundant and overlapping non-repeat B73 (husk and root: 61.3 Mb) and Mo17 (root: 133.3 Mb) *Hpa*II fragment sequences (Table 1) were assembled into multiple sequence alignments and a consensus sequence representing each alignment was derived. Assembly of these sequences resulted in the derivation of 339,730 (42.6 Mb) and 586,237 (70.7 Mb) non-redundant *Hpa*II consensus sequences from B73 and Mo17, respectively (Supplementary Table 2). In addition to providing a longer assembled sequence to help accurately align *Hpa*II fragments to allelic B73 reference sequences contained within the MAGIv4.0 C&S database (i.e., distinguish between highly similar paralogs), the assembly permitted a calculation of the per-base coverage depth, or the frequency with which any consensus base was observed in the raw data. Importantly, this metric can serve as a measure of confidence in the accuracy of consensus bases, as putative SNPs with a high coverage depth are more likely to be valid (Barbazuk et al., 2007). In addition, the assembly of cognate *Hpa*II fragment sequences reduced the computational requirements for the alignment and SNP calling process, as only unique sequences were used.

Second, we developed a computational approach to minimize the number of SNPs called from alignments of paralogous sequences, which is similar in objective to the paralog identification method used by the SNP calling software POLYBAYES (Marth et al., 1999) and to the "monoallelism" rules used by Barbazuk et al. (2007). Our approach assumes that it is possible to discover fixed differences among paralogs by comparing a reference sequence database or genome against itself, where almost all sequence differences observed in non-self paralogous alignments are non-allelic (Figure 1 A and B). Although some non-allelic differences may actually be polymorphisms at one or both of the loci, it is assumed that the majority of these identified differences are expected to be fixed differences that distinguish paralogs. Following this argument, a search of the MAGIv4.0 C&S database against itself was performed to identify all such single nucleotide differences that distinguish paralogs in the maize B73 genome. Putative non-allelic fixed differences that were identified from unique paralogous alignments were catalogued into a PDL as "context sequences" (i.e., the 16 bp 5′ and 16 bp 3′ flanking the single nucleotide difference).

SNP Identification

With the implementation of the PDL, *Hpa*II consensus sequences from Mo17 were aligned against the best reference match B73 sequence (MAGIv4.0 C&S; 675.2 Mb) and all single nucleotide differences were identified and extracted as context sequences (see "Materials and Methods"). If the context sequence of *any* of these single nucleotide differences (Mo17 *Hpa*II vs. B73 MAGIv4.0 C&S) matched a context sequence contained within the PDL, it was treated as an indication of a paralogous alignment and *all* SNP calls from such alignments were suppressed. In this case, the next strongest alignment for the same *Hpa*II consensus sequence was considered, continuing in this fashion until an alignment with no match to a PDL context sequence was found, or the rate of mismatches in the successive alignments exceeded a preset maximum of 5%. Essentially, the PDL selected which alignments to use for SNP calling but not which single nucleotide differences to call as SNPs. The same procedure was performed with B73 *Hpa*II consensus sequences, which served as an internal control to estimate the rate of false SNP discovery with and without implementation of the PDL.

Use of the PDL proved to be highly effective at preventing false SNP calls because of paralogous alignments. The estimated false discovery rate (FDR) obtained by comparing the SNP call rate for B73 (control, all SNPs considered false) and Mo17 *Hpa*II consensus sequences at various coverage depths and base quality values (*Q* values) thresholds is shown in Table 2. If SNP calls were made using the PDL and not restricted to a



Different alleles of each paralog segregate into maize inbred lines B73 (reference) and Mo17

Two putative fixed differences are detected between paralogs and are recorded in the PDL

The PDL recognizes the A/G mismatch as a putative fixed difference and it is correctly not called a SNP

The PDL does not recognize the C/G mismatch as a putative fixed difference and it is correctly called a SNP

The T/A mismatch is incorrectly scored as a putative fixed difference by the PDL. This happens when B73 (reference) carries the derived allele (B73 A).

Figure 1. Illustration of a recent single-gene duplication event that results in highly similar paralogs, and how the paralog distinguishing list (PDL) distinguishes alleles from paralogs when calling SNPs. (*A*) The PDL method is based on the assumption that a pair of duplicated genes that are fixed in the extant maize population likely originated from a single duplication event, which in many cases was the ancient tetraploidization event. If the duplication event is sufficiently old, virtually all differences among paralogs are because of mutations that have occurred since the genome duplication event, and distinguishing paralogs is easy. However, if the duplication was recent and the ancestral gene was polymorphic, alternative alleles at the paralogous loci may become fixed in the population, and the number of fixed differences between the donor and derived loci may be similar to the average allelic pairwise difference observed in maize. It is these cases for which it is very difficult to distinguish alleles from paralogs on the basis of alignment scores only. (*B*) An intra-reference alignment of B73 reference sequences discovers putative fixed differences (T/A and G/A) that differentiate paralogs (B73 A and B73 B), which are recorded as context sequences in the paralog distinguishing list (PDL). Next, *Hpa*II consensus sequences of Mo17 are aligned to B73 references sequences. Both the correct allelic (B73 B vs. Mo17 B) and erroneous paralogous (B73 B vs. Mo17 A) alignments detect a single nucleotide mismatch, and thus, cannot be distinguished from each other based solely on alignment scores. The context sequences of both single nucleotide mismatches (A/G and C/G) are searched against the PDL. The context sequence of the A/G mismatch matches a context sequence in the PDL; thus, the mismatch is correctly recognized as a putative fixed difference and not called a SNP. However, the context sequence of the C/G mismatch does not match any context sequence in the PDL and is therefore correctly called a SNP. When B73 carries a derived allele (B73A), the context sequence of the T/A mismatch in the allelic B73 A vs. Mo17 A comparison is also detected in the PDL. Thus, this true SNP is not called because it is incorrectly scored as a putative fixed difference, which ultimately leads to a reduction in SNP detection power.

specific coverage depth or *Q*-value threshold, 126,683 putative SNPs between Mo17 and B73 (1 SNP/248 bp) were discovered at an estimated 15.1% FDR. If SNP calls were made using only the most parsimonious alignment (i.e., without PDL), 174,476 putative B73/Mo17 SNPs (1 SNP/199 bp) were called at a dramatically increased FDR of 46.8%. Overall, use of the PDL effectively provided a threefold reduction in the rate of false SNP discovery at every evaluated coverage depth and *Q*-value threshold relative to rates determined without use of the PDL.

As shown in Table 2, we observed a polymorphism rate of 1 SNP every 216 bp (86,830 SNPs/18,794,000 bp) at an estimated 11% FDR (Coverage Depth: ≥1X; *Q*-score: ≥35). If we restricted SNP calling to a coverage depth of ≥2X (*Q*-score: all), then we observed a polymorphism rate of 1 SNP every 204 bp at a false SNP discovery rate of 8.4%. The SNP discovery rate for Mo17 *Hpa*II consensus sequences at only 1X coverage (i.e., singletons) and all *Q*-scores was 1 SNP every 290 bp (calculated from Table 2) at an estimated 19.7% FDR, which suggests that at higher coverage depths and with higher quality sequence data more SNPs/kb were captured (i.e., higher SNP detection power). Although the FDR was reduced nearly twofold (15.1 to 8.4%) when using the PDL and

Table 2. Summary of putative SNPs and call rates at various coverage depths and quality value thresholds with and without implementation of the paralog distinguishing list (PDL).

		With PDL							Without PDL						
CD[†]	*Q*[‡]	B73			Mo17			FDR[#]	B73			Mo17			FDR
		SNPs	Kb[§]	Rate[¶]	SNPs	Kb	Rate		SNPs	Kb	Rate	SNPs	Kb	Rate	
≥1X	All[††]	11,904	19,515	0.61	126,683	31,435	4.03	15.1%	50,936	21,675	2.35	174,476	34,756	5.02	46.8%
	≥20	10,701	18,450	0.58	119,294	29,675	4.02	14.4%	47,343	20,495	2.31	164,904	32,719	5.04	45.8%
	≥30	8,955	16,282	0.55	106,475	25,843	4.12	13.3%	39,910	17,897	2.23	147,335	28,553	5.16	43.2%
	≥35	5,703	11,182	0.51	86,830	18,794	4.62	11.0%	23,149	12,057	1.92	119,465	20,813	5.74	33.4%
	≥40	2,352	5,470	0.43	62,966	13,036	4.83	8.9%	10,378	5,830	1.78	85,547	14,451	5.92	30.1%
	≥50	1,609	4,349	0.37	57,205	11,603	4.93	7.5%	6,832	4,679	1.46	77,688	12,884	6.03	24.2%
	≥60	879	2,747	0.32	45,610	9,346	4.88	6.6%	3,724	2,956	1.26	61,991	10,384	5.97	21.1%
	≥70	634	2,113	0.30	39,787	8,153	4.88	6.1%	2,651	2,266	1.17	54,279	9,062	5.99	19.5%
≥2X	All	2,072	5,054	0.41	61,584	12,543	4.91	8.4%	9,048	5,451	1.66	83,547	13,925	6.00	27.7%
	≥20	2,057	5,017	0.41	61,527	12,531	4.91	8.4%	9,017	5,465	1.65	83,475	13,913	6.00	27.5%
	≥30	2,031	5,078	0.40	61,300	12,485	4.91	8.1%	8,910	5,433	1.64	83,173	13,862	6.00	27.3%
	≥40	1,953	4,883	0.40	60,573	12,337	4.91	8.1%	8,529	5,298	1.61	82,169	13,695	6.00	26.8%
	≥50	1,609	4,349	0.37	57,205	11,603	4.93	7.5%	6,832	4,679	1.46	77,688	12,884	6.03	24.2%
	≥60	879	2,747	0.32	45,610	9,346	4.88	6.6%	3,724	2,956	1.26	61,991	10,384	5.97	21.1%
	≥70	634	2,113	0.30	39,787	8,153	4.88	6.1%	2,651	2,266	1.17	54,279	9,062	5.99	19.5%
≥3X	All	702	2,127	0.33	37,980	7,783	4.88	6.8%	3,127	2,282	1.37	51,769	8,657	5.98	22.9%
	≥20	699	2,118	0.33	37,975	7,782	4.88	6.8%	3,124	2,280	1.37	51,763	8,656	5.98	22.9%
	≥30	697	2,112	0.33	37,966	7,780	4.88	6.8%	3,114	2,273	1.37	51,751	8,654	5.98	22.9%
	≥40	689	2,153	0.32	37,912	7,769	4.88	6.6%	3,088	2,271	1.36	51,681	8,642	5.98	22.7%
	≥50	679	2,122	0.32	37,833	7,753	4.88	6.6%	3,047	2,257	1.35	51,572	8,624	5.98	22.6%
	≥60	649	2,028	0.32	37,448	7,690	4.87	6.6%	2,899	2,196	1.32	51,044	8,550	5.97	22.1%
	≥70	529	1,763	0.30	35,417	7,272	4.87	6.2%	2,299	1,900	1.21	48,339	8,097	5.97	20.3%
≥4X	All	322	1,039	0.31	24,454	5,084	4.81	6.4%	1,452	1,108	1.31	33,403	5,662	5.90	22.2%
	≥20	319	1,029	0.31	24,454	5,084	4.81	6.4%	1,449	1,115	1.30	33,402	5,661	5.90	22.0%
	≥30	318	1,026	0.31	24,454	5,084	4.81	6.4%	1,445	1,112	1.30	33,402	5,661	5.90	22.0%
	≥40	317	1,057	0.30	24,451	5,083	4.81	6.2%	1,443	1,110	1.30	33,399	5,661	5.90	22.0%
	≥50	316	1,053	0.30	24,443	5,082	4.81	6.2%	1,437	1,105	1.30	33,391	5,659	5.90	22.0%
	≥60	313	1,043	0.30	24,430	5,079	4.81	6.2%	1,426	1,105	1.29	33,368	5,656	5.90	21.9%
	≥70	311	1,037	0.30	24,356	5,064	4.81	6.2%	1,405	1,098	1.28	33,272	5,639	5.90	21.7%

[†] CD, coverage depth. The number of reads with an aligned base that supported the consensus base call.

[‡] *Q*, quality values. Quality values were computed using the 454 base-calling software (single reads) or the CAP3 assembly program (multiple sequence alignments).

[§] The number of kilobases (Kb) of HpaII consensus sequence that aligned to MAGIv4.0 C&S database.

[¶] The number of SNPs called per Kb of *Hpa*II consensus sequence (SNPs/Kb).

[#] The percent false discovery rate (FDR) at each coverage depth was calculated by dividing the B73 call rate by the Mo17 call rate and multiplying by 100.

[††] No filtering on *Q* values.

additionally restricting SNP calls to a coverage depth of ≥2X, the FDR remained relatively unchanged at progressively higher coverage depth thresholds. This suggests that deeper sequencing would provide limited improvement in the calling accuracy of SNPs already at a coverage depth of 2X or higher, but this might not have been the case if the sequenced maize lines were highly heterozygous. The ability to reduce the number of false positive SNPs by restricting SNP calls to higher cover depths was also a key finding by Barbazuk et al. (2007), the first study that used pyrosequencing to identify SNPs within expressed maize genes. Additionally, it seems that *Q* values calculated by the 454 base calling software (single reads) or CAP3 program (multiple sequence alignments) are of minimal value for eliminating false-positive SNPs that result from sequencing errors when SNP calls are restricted to a coverage depth of 2X or higher.

SNP Validation

To independently cross-validate a subset of B73/Mo17 *Hpa*II SNPs that were identified via 454 pyrosequencing, we extracted a collection of B73 and Mo17 amplicon sequences from the Panzea database (http://www.panzea.org/) (Zhao et al., 2006) that were generated with traditional Sanger sequencing chemistry. The extracted sequences were assembled and aligned according to unique Panzea locus identifiers, which permitted the identification of SNPs. It was assumed that all paired sequences were allelic and all true SNPs were identified (i.e., 0% FDR; 100% power). To estimate an FDR for *Hpa*II SNPs, Panzea SNPs were mapped onto Mo17 *Hpa*II consensus sequences, and vice versa. The mapping resulted in the identification of a subset of SNPs in each dataset that was derived from sequence common to both datasets (Table 3).

With the constructed SNP validation dataset, we found that 85.9% (449/523) of the PDL-based *Hpa*II SNPs were concordant with Panzea SNPs. This resulted in an estimated FDR of 14.1%, which strongly agreed with the 15.1% (no thresholds; with PDL) that was estimated using the B73/Mo17 call rate comparison (Table 2). However, only 62.0% of SNPs identified in Panzea were also identified in the dataset of PDL identified B73/Mo17 *Hpa*II SNPs, whereas it was 80.9% without the PDL. This signifies a weakness of the MAGIv4.0 C&S-based PDL, as true SNPs were incorrectly considered non-allelic by the PDL.

DISCUSSION

Next-generation DNA sequencing technologies have made high-throughput resequencing efficient and affordable. However, the use of these technologies in a read-to-reference based SNP discovery approach at the level of a whole-genome has not come to fruition for agronomically important plant species. The primary reason is that many of these plant species have large, complex genomes and as a result do not have an available, accurate, or complete genome sequence. In addition, the short read-lengths produced by these high-throughput sequencing technologies are limited in ability to differentiate the large numbers of paralogs that are common to the genome of many angiosperm species (Blanc and Wolfe, 2004). Maize was chosen as the test organism for this pilot study because of three qualities of its nuclear genome: it is ~2500 Mb in size; it consists of more than 75% highly repetitive DNA (Meyers et al., 2001; SanMiguel et al., 1996); and at least one-third of its estimated 59,000 genes are duplicated (Blanc and Wolfe, 2004; Messing et al., 2004). Here, we tested a gene-enrichment sequencing approach that is applicable to virtually any plant species and a computational pipeline that enables the efficient and accurate discovery of a large number of SNPs using an incomplete and low-coverage reference sequence.

Table 3. Summary of B73/Mo17 454 SNP validation.

	With PDL	WithoutPDL
Panzea SNPs[†]	724	724
*Hpa*II SNPs	523[‡]	720[§]
Shared SNPs[¶]	449	586
*Hpa*II FDR[#]	14.1%	18.6%
*Hpa*II Power[††]	62.0%	80.9%

[†] The number of identified Panzea SNPs that mapped to Mo17 *Hpa*II consensus sequences.

[‡] The number of B73/Mo17 *Hpa*II SNPs identified via 454 pyrosequencing that mapped to Panzea sequences. These B73/Mo17 *Hpa*II SNPs that mapped are a subset of the 126,683 putative SNPs (≥1X coverage depth; All *Q* values) that were called using the paralog distinguishing list (PDL).

[§] The number of B73/Mo17 *Hpa*II SNPs identified via 454 pyrosequencing that mapped to Panzea sequences. These B73/Mo17 *Hpa*II SNPs that mapped are a subset of the 174,476 putative SNPs (≥1X coverage depth; All *Q* values) that were called without using the paralog distinguishing list (PDL).

[¶] SNPs that were identified in both the B73/Mo17 *Hpa*II SNP and Panzea SNP datasets.

[#] We assumed that all SNPs called from the Panzea sequence dataset were true SNPs. The percent false discovery rate (FDR) was calculated as [1–(449/523)*100] and [1–(586/720)*100].

[††] We assumed that all SNPs in the Panzea sequence dataset were identified. Power was calculated as [(449/724)*100] and [(586/724)*100].

We modified the previously described HMPR technique (Emberton et al., 2005) to enable shotgun sequencing of 100 to 600 bp *Hpa*II fragments in a manner that fully used the read length (potential of 200–300 bases) ability of the 454 GS FLX instrument. Of the two tissue types that were tested as sources of genomic DNA, seedling roots have a greater potential to enable the rapid construction of gene-enriched, modified HMPR libraries that have low levels of repeats and organellar DNA contamination. However, improved seed sterilization procedures and/or sterile, antibiotic-treated growing conditions are necessary to prevent the proliferation of bacterial symbionts in seedling roots, and the cytosine methylation pattern of genes and repeats in seedling root tissue needs to be more fully investigated. Since performing this experiment, we have identified unfertilized, immature ear shoots as an excellent tissue for isolating total maize genomic DNA. B73 and Mo17 immature ear *Hpa*II libraries constructed with modified HMPR technology were highly enriched (four- to fivefold) for genic sequences, while extremely depleted in repeat, organellar, and bacterial sequences (total: <10%) (M. Gore, R. Elshire, and E. Buckler, unpublished data).

Although our modified HMPR technique facilitated high throughput gene-enrichment sequencing of a large, complex plant genome, in general, the yield per run of modified HMPR libraries on the 454 GS FLX was lower than the expected 100 Mb. If the DNA copy per bead ratio is carefully optimized for modified HMPR libraries, it should be possible to routinely obtain 100 Mb of sequence data. In addition, the low sequencing yield may be because of less than optimal lengths (3–10 kb) of *Hpa*II concatemers. If so, a 6 bp MCS restriction enzyme (Fellers, 2008) may help to produce much larger concatemers that are better suited for the downstream 454 sample preparation, which is optimized for undigested total genomic DNA. Also, assembly of the larger restriction fragment sizes would produce larger consensus sequences for more accurate mapping. Alternatively, with the increased average read length (400 bases) and paired-end read capability of the new GS FLX Titanium (http://www.454.com), it might be more efficient, and as comprehensive, to directly sequence restriction fragments instead of concatemers.

We identified 126,683 putative B73/Mo17 SNPs, primarily in genic regions of the maize genome, using a computational pipeline for short read-lengths that is applicable to any plant species with at least a large collection of genome survey sequences. A computational approach was developed to distinguish between allelic and paralogous *Hpa*II consensus-MAGIv4.0 C&S

reference alignments by searching identified putative single nucleotide differences against a PDL of putative fixed differences that distinguish paralogs from each other. The false-SNP discovery rate with implementation of the PDL was estimated by two different approaches, and both were found to be at an acceptable level and highly concordant (15.1 vs. 14.1%). Detection of SNPs using the PDL was threefold more effective in controlling the FDR than a most parsimonious alignment strategy, and the FDR could be further reduced by filtering SNPs based on coverage depth and/or *Q*-value thresholds (Table 2). The most likely sources of false-positive SNPs are cloning artifacts (i.e., base substitution errors) contained within MAGIv4.0 C&S sequences (Fu et al., 2004) and paralogous alignments not identified by the PDL. Although very stringent parameters were used to assemble redundant, overlapping *Hpa*II fragment sequences, it is possible that collapsed paralogs also contributed to the identification of false-positive SNPs. The number of false-positive SNPs that result from the FLX system are expected to be low (presumably less frequent at coverage depths of 2X and higher), as other studies have shown the GS FLX single-read error rate to be ~0.5% (Droege and Hill, 2008) and substantially lower at higher coverage depths (Lynch et al., 2008; Smith et al., 2008). In addition, the rate of paralog collapse in the MAGI assemblies was estimated to be ~1% (Emrich et al., 2007); therefore, their contribution to the calling of false-positive SNPs and inaccuracies in the PDL should be very minimal.

The difference in FDR estimates between SNPs called with and without the PDL method is much less striking for the Panzea validation dataset (Table 3) than that observed for the B73/Mo17 call rate comparison (Table 2). This is most likely because Panzea sequences resulted from the preferential sequencing of putatively single-locus PCR products, as PCR reactions that appeared to amplify multiple loci were discarded prior to sequencing (E. Buckler, unpublished). Essentially, the amplicon-Sanger sequencing strategy acted as a PDL. Thus, the Panzea dataset is poorly suited to assess the ability of the PDL to detect paralogous alignments, because the Panzea database was constructed with a bias against paralogous sequences. All amplicon-Sanger sequencing strategies will have this same bias; therefore, the best external validation of the PDL is to sequence modified HMPR libraries of Mo17 on a different next-generation sequencing platform (e.g., Illumina sequencing). Currently, the B73 (internal control)/Mo17 call rate comparison is the best available method to estimate the ability of the PDL to reduce the number of false positive SNP calls from paralogous alignments (Table 2). Nevertheless, minor improvements in the FDR are still observed when the PDL is used on the Panzea dataset (Table 3).

Transcriptome sequencing is useful when the aim is enrichment of tissue and developmental-stage specific genes; however, for high coverage of the gene space it is not very cost effective. Essentially, numerous cDNA libraries capturing multiple developmental stages and environmental stresses are needed to even approach high coverage of the gene space. Therefore, we sequenced modified HMPR genomic libraries because it is expected to result in a more comprehensive sampling of genes than that of transcriptome sequencing (Emberton et al., 2005; Palmer et al., 2003), and it is also expected to provide access to the nucleotide diversity in introns, regulatory regions, and non-expressed genes. We used the Lander-Waterman model (Lander and Waterman, 1988) and the rate of contig formation as described in Whitelaw et al. (2003) to estimate the effective gene-space size sampled by the modified HMPR method, which was 136.4 Mb (~27% of the ~500 Mb maize gene-space; Palmer et al., 2003) for the Mo17 root library. This estimate of the effective gene-space size might be slightly overestimated due to the very stringent CAP3 assembly parameters that were used. Given that 70.7 Mb of *Hpa*II consensus sequence data exists for Mo17 (Supplementary Table 2), it is estimated that the library was sequenced to only 0.52X coverage. If we were to sequence the Mo17 root library to 1X coverage, then the maximum number of putative SNPs called with the PDL would be ~200,000 at a rate of 4.03 SNPs/kb. If several million SNPs are to be discovered, we will need to sequence additional maize inbred lines, possibly construct other modified HMPR libraries using different 4 bp cutter MCS restriction enzymes, and/or use the draft maize genome sequence to call SNPs.

The PDL is only as high-quality as the completeness and accuracy of the reference sequence used to construct it, but despite the shortcomings of the MAGI assemblies (e.g., 1% collapsed paralogs, cloning artifacts, and partial genome assembly), a significant reduction (threefold) in the number of false positive SNPs that resulted from paralogous alignments was still observed (Table 2). Moreover, these issues will be mostly resolved when the draft maize B73 genome sequence is available for constructing a PDL and calling SNPs.

A more important limitation of the PDL, however, is that it reduced the power to detect true SNPs. Based on the observed SNP call rate (4.91 SNPs/kb; 1 SNP/204 bp) with the PDL at a coverage depth of ≥2X, we are underestimating the expected SNP call rate (1 SNP/153 bp based on 1095 genes) between any randomly chosen diverse, temperate maize inbred lines by ~25% (Yamasaki et al., 2005). If SNPs were called without the PDL at a coverage depth of ≥2X, the observed (6.00 SNPs/kb; 1 SNP/167 bp) and expected (1 SNP/153 bp) SNP call rates are nearly identical. As shown in Table 3, based on the comparison of B73/Mo17 *Hpa*II SNPs (no threshold) with the Panzea SNP dataset, there was an 18.9% loss in SNP detection power with implementation of the PDL. The reduction in power is attributed to true SNPs being incorrectly considered non-allelic by the PDL. We hypothesize that these true SNPs could not be distinguished from actual fixed differences among paralogs on the basis of the intra-reference sequence comparison alone, which would occur if the reference line (B73) used to construct the PDL carries a derived allele (Fig. 1 A and B). This is a systematic bias that may affect both population genetics and association studies when the reference line alone carries an allele of interest. This problem is most severe when a single line is compared to the reference, but the expected rate of false negatives because of this effect decreases to $1/(n + 1)$ when n lines are compared to the reference. Further reduction may be possible if multiple non-reference lines are also compared to each other.

Although the results obtained in this pilot study are very encouraging, there are several drawbacks to this approach that should be considered. First, the method of gene enrichment used here restricts SNP discovery to sites near *Hpa*II restriction sites in unmethylated regions, which can be remedied by constructing additional modified HMPR libraries with different 4 bp cutter MCS restriction enzymes. We do not presume that *all* nucleotide variation in methylated regions of the maize genome is phenotypically irrelevant, so different methods are needed to discover SNPs from these regions. Additionally, genome-wide methylation patterns and locus specific methylation levels may vary across genetic backgrounds, tissue types, developmental stages, and even environmental conditions (Cervera et al., 2002; Finnegan et al., 2000; Lister et al., 2008; Rabinowicz et al., 1999; Vaughn et al., 2007). Thus, performing this technique across a panel of inbred lines may not result in representation of all lines at all loci. For marker discovery, this line-specific or locus-specific censoring effect may not be important overall, but population genetic studies may be adversely affected by non-random missing data.

Regardless of these limitations, a considerable number of SNPs were discovered at an acceptably low FDR for the purpose of constructing high-density multiplexed genotyping products, but sequencing of additional maize inbred lines is needed to construct an SNP dataset with low ascertainment bias that is appropriate for phylogenetics or population genetics studies. However, the SNPs identified in this study are immediately applicable for fine mapping of complex traits in the Intermated B73 × Mo17 (IBM) population, which is a widely used community resource for QTL mapping studies in maize (Lee et al., 2002). Most importantly, we estimate the cost of SNP discovery in this study at $0.38/SNP, yet note that several aspects of the molecular methods used here can be optimized for much higher sequencing yield and broader genome coverage. Such optimization, combined with further advances in high throughput sequencing yield, longer read-lengths, lower error rates, and cheaper run costs, can further reduce the cost of SNP discovery in diverse maize, such that several million gene-enriched SNPs needed for comprehensive association studies is an immediate economic possibility.

Acknowledgments

We thank D. Costich, S. Myles, J. Peiffer, and especially M. Hamblin for their insights and critical review of the manuscript, L. Rigamer Lirette for technical editing of the manuscript, and three anonymous reviewers for helpful comments. This work was supported in part by U.S. National Science Foundation grants DBI-0321467 and DBI-0638566 and United States Department of Agriculture-Agriculture Research Service. Mark H. Wright was supported by U.S. National Science Foundation grants DBI-0701382 and DBI-0606461. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the United States Department of Agriculture.

References

Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Antequera, F., and A.P. Bird. 1988. Unmethylated CpG islands associated with genes in higher plant DNA. EMBO J. 7:2295–2299.

Barbazuk, W.B., S.J. Emrich, H.D. Chen, L. Li, and P.S. Schnable. 2007. SNP discovery via 454 transcriptome sequencing. Plant J. 51:910–918.

Bennett, S. 2004. Solexa Ltd. Pharmacogenomics 5:433–438.

Bennetzen, J.L., K. Schrick, P.S. Springer, W.E. Brown, and P. SanMiguel. 1994. Active maize genes are unmodified and flanked by diverse classes of modified, highly repetitive DNA. Genome 37:565–576.

Blanc, G., and K.H. Wolfe. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell 16:1667–1678.

Buckler, E.S., J.M. Thornsberry, and S. Kresovich. 2001. Molecular diversity, structure and domestication of grasses. Genet. Res. 77:213–218.

Cervera, M.T., L. Ruiz-Garcia, and J.M. Martinez-Zapater. 2002. Analysis of DNA methylation in *Arabidopsis thaliana* based on methylation-sensitive AFLP markers. Mol. Genet. Genomics 268:543–552.

Droege, M., and B. Hill. 2008. The genome sequencer FLX(TM) system—longer reads, more applications, straight forward bioinformatics and more complete data sets. J. Biotechnol. 136:3–10.

Emberton, J., J. Ma, Y. Yuan, P. SanMiguel, and J.L. Bennetzen. 2005. Gene enrichment in maize with hypomethylated partial restriction (HMPR) libraries. Genome Res. 15:1441–1446.

Emrich, S.J., L. Li, T.-J. Wen, M.D. Yandeau-Nelson, Y. Fu, L. Guo, H.-H. Chou, S. Aluru, D.A. Ashlock, and P.S. Schnable. 2007. Nearly identical paralogs: Implications for maize (*Zea mays* L.) genome evolution. Genetics 175:429–439.

Fellers, J.P. 2008. Genome filtering using methylation—sensitive restriction enzymes with six base pair recognition sites. The Plant Genome 1:146–152.

Finnegan, E.J., W.J. Peacock, and E.S. Dennis. 2000. DNA methylation, a key regulator of plant development and other processes. Curr. Opin. Genet. Dev. 10:217–223.

Fu Y., A.-P. Hsia, L. Guo, and P.S. Schnable. 2004. Types and frequencies of sequencing errors in methyl-filtered and high C_0t maize genome survey sequences. Plant Phys. 135:2040–2045.

Fu, H., Z. Zheng, and H.K. Dooner. 2002. Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. Proc. Natl. Acad. Sci. USA 99:1082–1087.

Fu, H., W. Park, X. Yan, Z. Zheng, B. Shen, and H.K. Dooner. 2001. The highly recombinogenic bz locus lies in an unusually gene-rich region of the maize genome. Proc. Natl. Acad. Sci. USA 98:8903–8908.

Gaut, B.S., and J.F. Doebley. 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. Proc. Natl. Acad. Sci. USA 94:6809–6814.

Gore, M., P. Bradbury, R. Hogers, M. Kirst, E. Verstege, J. van Oeveren, J. Peleman, E. Buckler, and M. van Eijk. 2007. Evaluation of target preparation methods for single-feature polymorphism detection in large complex plant genomes. Crop Sci. 47:S135–S148.

Hake, S., and V. Walbot. 1980. The genome of *Zea-mays*, its organization and homology to related grasses. Chromosoma 79:251–270.

Henry, A.M., and C. Damerval. 1997. High rates of polymorphism and recombination at the Opaque-2 locus in cultivated maize. Mol. Gen. Genet. 256:147–157.

Huang, X., and A. Madan. 1999. CAP3: A DNA sequence assembly program. Genome Res. 9:868–877.

Kalyanaraman, A., S.J. Emrich, P.S. Schnable, and S. Aluru. 2007. Assembling genomes on large-scale parallel computers. J. Parallel Distrib. Comput. 67:1240–1255.

Kent, W.J. 2002. BLAT—The BLAST-like alignment tool. Genome Res. 12:656–664.

Lai, J., J. Ma, Z. Swigoňová, W. Ramakrishna, E. Linton, V. Llaca, B. Tanyolac, Y.-J. Park, O.Y. Jeong, J.L. Bennetzen, and J. Messing. 2004. Gene loss and movement in the maize genome. Genome Res. 14:1924–1931.

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

Lee, M., N. Sharopova, W.D. Beavis, D. Grant, M. Katt, D. Blair, and A. Hallauer. 2002. Expanding the genetic map of maize with the intermated B73 × Mo17 (IBM) population. Plant Mol. Biol. 48:453–461.

Li, W., S. Ruf, and R. Bock. 2006. Constancy of organellar genome copy numbers during leaf development and senescence in higher plants. Mol. Genet. Genomics 275:185–192.

Lister, R., R.C. O'Malley, J. Tonti-Filippini, B.D. Gregory, C.C. Berry, A.H. Millar, and J.R. Ecker. 2008. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. Cell 133:523–536.

Lynch, M., W. Sung, K. Morris, N. Coffey, C.R. Landry, E.B. Dopman, W.J. Dickinson, K. Okamoto, S. Kulkarni, D.L. Hartl, and W.K. Thomas. 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. Proc. Natl. Acad. Sci. USA 105:9272–9277.

Mardis, E.R. 2008. The impact of next-generation sequencing technology on genetics. Trends Genet. 24:133–141.

Margulies, M., M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen, S.B. Dewell, L. Du, J.M. Fierro, X.V. Gomes, B.C. Godwin, W. He, S. Helgesen, C.H. Ho, G.P. Irzyk, S.C. Jando, M.L. Alenquer, T.P. Jarvie, K.B. Jirage, J.B. Kim, J.R. Knight, J.R. Lanza, J.H. Leamon, S.M. Lefkowitz, M. Lei, J. Li, K.L. Lohman, H. Lu, V.B. Makhijani, K.E. McDade, M.P. McKenna, E.W. Myers, E. Nickerson, J.R. Nobile, R. Plant, B.P. Puc, M.T. Ronan, G.T. Roth, G.J. Sarkis, J.F. Simons, J.W. Simpson, M. Srinivasan, K.R. Tartaro, A. Tomasz, K.A. Vogt, G.A. Volkmer, S.H. Wang, Y. Wang, M.P. Weiner, P. Yu, R.F. Begley, and J.M. Rothberg. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380.

Marth, G.T., I. Korf, M.D. Yandell, R.T. Yeh, Z. Gu, H. Zakeri, N.O. Stitziel, L. Hillier, P.-Y. Kwok, and W.R. Gish. 1999. A general approach to single-nucleotide polymorphism discovery. Nat. Genet. 23:452–456.

Messing, J., A.K. Bharti, W.M. Karlowski, H. Gundlach, H.R. Kim, Y. Yu, F. Wei, G. Fuks, C.A. Soderlund, K.F.X. Mayer, and R.A. Wing. 2004. Sequence composition and genome organization of maize. Proc. Natl. Acad. Sci. USA 101:14349–14354.

Meyers, B.C., S.V. Tingey, and M. Morgante. 2001. Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res. 11:1660–1676.

Palmer, L.E., P.D. Rabinowicz, A.L. O'Shaughnessy, V.S. Balija, L.U. Nascimento, S. Dike, M. de la Bastide, R.A. Martienssen, and W.R. McCombie. 2003. Maize genome sequencing by methylation filtration. Science 302:2115–2117.

Possingham, J.V. 1980. Plastid replication and development in the life cycle of higher plants. Annu. Rev. Plant Physiol. 31:113–129.

Rabinowicz, P.D. 2003. Constructing gene-enriched plant genomic libraries using methylation filtration technology. Methods Mol. Biol. 236:21–36.

Rabinowicz, P.D., L.E. Palmer, B.P. May, M.T. Hemann, S.W. Lowe, W.R. McCombie, and R.A. Martienssen. 2003. Genes and transposons are differentially methylated in plants, but not in mammals. Genome Res. 13:2658–2664.

Rabinowicz, P.D., K. Schutz, N. Dedhia, C. Yordan, L.D. Parnell, L. Stein, W.R. McCombie, and R.A. Martienssen. 1999. Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. Nat. Genet. 23:305–308.

Rabinowicz, P.D., R. Citek, M.A. Budiman, A. Nunberg, J.A. Bedell, N. Lakey, A.L. O'Shaughnessy, L.U. Nascimento, W.R. McCombie, and R.A. Martienssen. 2005. Differential methylation of genes and repeats in land plants. Genome Res. 15:1431–1440.

Remington, D.L., J.M. Thornsberry, Y. Matsuoka, L.M. Wilson, S.R. Whitt, J. Doebley, S. Kresovich, M.M. Goodman, and E.S. Buckler. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc. Natl. Acad. Sci. USA 98:11479–11484.

SanMiguel, P., A. Tikhonov, Y.K. Jin, N. Motchoulskaia, D. Zakharov, A. Melake-Berhan, P.S. Springer, K.J. Edwards, M. Lee, Z. Avramova, and J.L. Bennetzen. 1996. Nested retrotransposons in the intergenic regions of the maize genome. Science 274:765–768.

Shendure, J., G.J. Porreca, N.B. Reppas, X. Lin, J.P. McCutcheon, A.M. Rosenbaum, M.D. Wang, K. Zhang, R.D. Mitra, and G.M. Church. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309:1728–1732.

Smith, D.R., A.R. Quinlan, H.E. Peckham, K. Makowsky, W. Tao, B. Woolf, L. Shen, W.F. Donahue, N. Tusneem, M.P. Stromberg, D.A. Stewart, L. Zhang, S.S. Ranade, J.B. Warner, C.C. Lee, B.E. Coleman, Z. Zhang, S.F. McLaughlin, J.A. Malek, J.M. Sorenson, A.P. Blanchard, J. Chapman, D. Hillman, F. Chen, D.S. Rokhsar, K.J. McKernan, T.W. Jeffries, G.T. Marth, and P.M. Richardson. 2008. Rapid whole-genome mutational profiling using next-generation sequencing technologies. Genome Res. 18:1638–1642.

Smith, T.F., and M.S. Waterman. 1981. Identification of common molecular subsequences. J. Mol. Biol. 147:195–197.

Swigoňová, Z., J. Lai, J. Ma, W. Ramakrishna, V. Llaca, J.L. Bennetzen, and J. Messing. 2004. Close split of sorghum and maize genome progenitors. Genome Res. 14:1916–1923.

Tenaillon, M.I., M.C. Sawkins, A.D. Long, R.L. Gaut, J.F. Doebley, and B.S. Gaut. 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Proc. Natl. Acad. Sci. USA 98:9161–9166.

Vaughn, M.W., M. Tanurd, Z. Lippman, H. Jiang, R. Carrasquillo, P.D. Rabinowicz, N. Dedhia, W.R. McCombie, N. Agier, A. Bulski, V. Colot, R.W. Doerge, and R.A. Martienssen. 2007. Epigenetic natural variation in *Arabidopsis thaliana*. PLoS Biol. 5:e174.

Wei, F., E. Coe, W. Nelson, A.K. Bharti, F. Engler, E. Butler, H. Kim, J.L. Goicoechea, M. Chen, S. Lee, G. Fuks, H. Sanchez-Villeda, S. Schroeder, Z. Fang, M. McMullen, G. Davis, J.E. Bowers, A.H. Paterson, M. Schaeffer, J. Gardiner, K. Cone, J. Messing, C. Soderlund, and R.A. Wing. 2007. Physical and genetic structure of the maize genome reflects its complex evolutionary history. PLoS Genet. 3:e123.

Whitelaw, C.A., W.B. Barbazuk, G. Pertea, A.P. Chan, F. Cheung, Y. Lee, L. Zheng, S. van Heeringen, S. Karamycheva, J.L. Bennetzen, P. SanMiguel, N. Lakey, J. Bedell, Y. Yuan, M.A. Budiman, A. Resnick, S. Van Aken, T. Utterback, S. Riedmuller, M. Williams, T. Feldblyum, K. Schubert, R. Beachy, C.M. Fraser, and J. Quackenbush. 2003. Enrichment of gene-coding sequences in maize by genome filtration. Science 302:2118–2120.

Wright, S.I., I.V. Bi, S.G. Schroeder, M. Yamasaki, J.F. Doebley, M.D. McMullen, and B.S. Gaut. 2005. The effects of artificial selection on the maize genome. Science 308:1310–1314.

Yamasaki, M., M.I. Tenaillon, I. Vroh Bi, S.G. Schroeder, H. Sanchez-Villeda, J.F. Doebley, B.S. Gaut, and M.D. McMullen. 2005. A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. Plant Cell 17:2859–2872.

Yao, H., Q. Zhou, J. Li, H. Smith, M. Yandeau, B.J. Nikolau, and P.S. Schnable. 2002. Molecular characterization of meiotic recombination across the 140-kb multigenic a1-sh2 interval of maize. Proc. Natl. Acad. Sci. USA 99:6157–6162.

Yuan, Y., P.J. SanMiguel, and J.L. Bennetzen. 2002. Methylation-spanning linker libraries link gene-rich regions and identify epigenetic boundaries in *Zea mays*. Genome Res. 12:1345–1349.

Yuan, Y., P.J. SanMiguel, and J.L. Bennetzen. 2003. High-cot sequence analysis of the maize genome. Plant J. 34:249–255.

Zhao, W., P. Canaran, R. Jurkuta, T. Fulton, J. Glaubitz, E. Buckler, J. Doebley, B. Gaut, M. Goodman, J. Holland, S. Kresovich, M. McMullen, L. Stein, and D. Ware. 2006. Panzea: A database and resource for molecular and functional diversity in the maize genome. Nucleic Acids Res. 34:D752–D757.

Zhu, C., M. Gore, E.S. Buckler, and J. Yu. 2008. Status and prospects of association mapping in plants. The Plant Genome 1:5–20.