

ORIGINAL RESEARCH

Single-Nucleotide Polymorphism Genotyping in Mapping Populations via Genomic Reduction and Next-Generation Sequencing: Proof of Concept

Peter J. Maughan,* Scott M. Yourstone, Robert L. Byers, Scott M. Smith, and Joshua A. Udall

Abstract

We report the results of a proof-of-concept experiment that validates the use of multiplex identifier (MID) barcodes and next-generation sequencing (454-pyrosequencing and Illumina Galle) to simultaneously discover and genotype single nucleotide polymorphisms (SNPs) from mapping populations using pooled genomic reduction libraries. The genomic reduction library utilized here consisted of 60 individuals from an *Arabidopsis thaliana* (L.) Heynh. mapping population. A total of 1720 and 504 SNPs were de novo identified and genotyped across the mapping population using only bioinformatic analyses of the progeny sequence fragments from the 454-pyrosequencing and Illumina datasets, respectively. The average base coverage at the SNPs was 4.5x for the 454-pyrosequencing dataset and 18.2x for the Illumina dataset. Cross validation of the genotypic scores between datasets showed 99.91% accuracy. Linkage mapping with the 454-pyrosequencing dataset produced five highly supported linkage groups that were collinear with the *Arabidopsis* physical map ($r = 0.981$), further validating the accuracy of the genotyping method. The unadjusted cost per data point ranged from US\$0.07 to \$0.147, suggesting that the technique could be broadly used for large-scale genotyping and has particular value for plant species with limited monetary resources as it circumvents the need for post-SNP discovery genotype assay development.

SINGLE NUCLEOTIDE POLYMORPHISM (SNPs) are the most abundant type of DNA polymorphism found in eukaryotic genomes (Batley et al., 2003; Garg et al., 1999) and are fast becoming the marker system of choice in marker-assisted plant breeding programs (Batley and Edwards, 2007). The high frequency of SNPs in plant genomes is well documented (Ching et al., 2002; Grivet et al., 2003; Nasu et al., 2002; Ossowski et al., 2008; Russell et al., 2004; Schneider et al., 2001; Zhu et al., 2003), with actual SNP densities ranging dramatically depending on the species type (auto- or allogamous), number and genetic diversity of the cultivars being assessed, and whether coding or noncoding regions are being considered. For example, in soybean [*Glycine max* (L.) Merr.] SNPs occur at a density of 1 per 375 bp in exon coding sequence and 1 per 490 bp in noncoding intron sequences (Choi et al., 2007), while in maize (*Zea mays* L.), a much more highly polymorphic species, 1 SNP was observed per 124 bp of coding sequence and 1 per 31 bp in noncoding regions (Ching et al., 2002). Frequency calculations are dependent on the diversity and often the number of accessions being studied. Indeed, Cronk (2005) reported that the density of SNPs in poplar [*Populus balsamifera* L. subsp. *trichocarpa* (Torr. & A. Gray) Brayshaw] doubled from 1 per 100 bp to 1 per 50 bp when geographically diverse species were included in the analysis. While the frequency of SNPs may vary across the genome due to selection pressure (e.g., coding regions), the distribution of SNPs on a genome scale appears to be fairly uniform. Indeed, in *Arabidopsis*

Published in The Plant Genome 3:166–178. Published 18 Dec. 2010.
doi: 10.3835/plantgenome2010.07.0016
© Crop Science Society of America
5585 Guilford Rd., Madison, WI 53711 USA
An open-access publication

All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Permission for printing and for reprinting the material contained herein has been obtained by the publisher.

Brigham Young Univ., Dep. of Plant & Wildlife Sciences, 285 WIDB, Provo, UT 84602. Received 20 July 2010. *Corresponding author (Jeff_Maughan@byu.edu).

Abbreviations: GRRSC, genomic reduction based on restriction-site conservation; LOD, logarithm of the odds; MAF, minor allele frequency; MID, multiplex identifier; MMD, maximum missing data; PCR, polymerase chain reaction; RIL, recombinant inbred line; RFLP, restriction fragment length polymorphism; SNP, single nucleotide polymorphism.

thaliana (L.) Heynh., the distribution of over 37,000 SNPs was found to be generally uniform across all chromosomes (Schmid et al., 2003). Compared to microsatellites (simple sequence repeat [SSR])-based markers, SNPs exhibit a lower mutation rate and thus are less problematic in population genetic analyses (Xu et al., 2005). The utility of SNPs in genetic analyses has recently been reviewed by Batley and Edwards (2007) and include association studies (Zhu et al., 2008), phylogenetic and evolutionary analysis (Cramer et al., 2008), genetic diversity analysis (Kawuki et al., 2009; Osman et al., 2003), marker-assisted selection (Hayashi et al., 2004; Tyrka et al., 2004), and genetic mapping (Rostoks et al., 2005).

Several technical methods have been reported for the initial discovery of SNPs in plant genomes, specifically: (i) expressed sequence tag (EST) sequencing (Barbazuk et al., 2007), (ii) targeted amplicon resequencing (Bundock et al., 2009), (iii) gene space resequencing with methylation-sensitive digestion (Deschamps et al., 2010; Gore et al., 2009), and (iv) genomic reduction based on restriction-site conservation (GR-RSC; Maughan et al., 2009). When combined with next-generation DNA sequencing technologies, these methods can be used to identify and validate large numbers of SNPs with limited technical expertise and at minimal cost (Deschamps and Campbell, 2010). For example, Maughan et al. (2009) utilized genome reduction to identify more than 25,000 putative SNPs in amaranth (*Amaranthus caudatus* L.) with an approximate supplies cost of less than US\$10,000. Unfortunately, before the utilization of newly discovered SNPs in marker assisted breeding programs, genotyping assay for each SNP must be developed—often with a significant conversion attrition rate (i.e., not all SNP can be converted to into a genotyping assay). While several commercial SNP genotyping assays are available (e.g., TaqMan assay [Applied Biosystems, Foster City, CA], Golden Gate system [Illumina, San Diego, CA], and Kaspar assay [KBiosciences, Hoddesdon, UK]), the conversion and validation costs associated with the development of genotyping assays is expensive, time consuming and frequently requires specific technical expertise and capital equipment.

In particular, the cost of SNP genotyping has been prohibitive for many nonmodel or nonagriculturally important plant species. Maughan et al. (2009) suggested that the use of GR-RSC, in conjunction with multiplex identifier (MID) barcodes and next-generation sequencing, could be extended to whole population studies, replacing the necessity for individual SNP assay development, since genotyping could be done directly from the DNA sequences of the population based solely on bioinformatic analysis (e.g., genotyping by sequencing). Genomic reduction, also known as reduced representation, reduces the complexity of large genomes by orders of magnitude, samples diverse but identical genomic regions from several individuals, and does not require a priori genome sequence information. Wiedmann et al. (2008) and Van Tassel et al. (2008) showed the initial successful utilization of genome reduction for SNP discovery in swine (*Sus*

scrofa) and cattle (*Bos taurus*), respectively. In both cases, pooled DNA from multiple individuals were digested with restricted endonucleases and DNA fragments of a particular size sequenced en masse. Using their protocol, SNPs were identified but alleles could not be assigned to specific individuals within the pooled sample. Maughan et al. (2009) refined the protocol by adding MID barcodes to each sample in the pool—thus allowing allele identification for each individual in the pool.

Here we investigate the extended use of MID barcodes (up to 31 MIDs per pool) and next-generation sequencing (454-pyrosequencing and Illumina Genome Analyzer) to simultaneously discover and genotype SNPs from a pooled genomic reduction library consisting of 60 individuals, including two parents and 58 recombinant inbred lines (RILs) from an *Arabidopsis thaliana* (L.) Heynh. mapping population (Lister and Dean, 1993). The incorporation of MID barcodes into specific DNA sequence fragments allowed for the initial identification of polymorphic SNPs between the parental lines, while subsequent bioinformatic analyses of the RIL subset of sequence reads was used to assign genotypic calls to each RIL individual. The use of a segregating *A. thaliana* RIL population enabled an analysis of the physical genome distribution of SNPs discovered and a comparison of the collinearity of the resultant linkage map with the physical map for each *Arabidopsis* chromosome. Moreover, we compare the efficiency of the system for genotyping using two next-generation sequencing platforms and discuss both pros and cons associated with each option. This report provides proof of concept for a technique that could be broadly used for large-scale genotyping and has particular value for plant species with limited monetary resources as it circumvents the need to develop specific SNP genotyping assays.

Materials and Methods

Plant Material and DNA Extraction

Individuals from a RIL (F_{2:9}) population, originally described by Lister and Dean (1993), were obtained from the Arabidopsis Information Resource (<http://www.arabidopsis.org/servlets/TairObject?type=germplasm&id=1005152502> [verified 19 Nov. 2010]; stock no. CS1899). The population was generated from a cross between the *Arabidopsis* ecotypes Landsberg *erecta* (Ler-0; CS20) and Columbia (Col-4; CS933) using single seed decent. For this study we utilized a random subset of individuals from the original population, including the two parental lines and 58 RILs (Table 1). All plants were grown in growth chambers using Arabidopsis Growth Medium (Lehle Seeds, Round Rock, TX) supplemented with controlled release 15–9–12 fertilizer (Lehle Seeds). Plants were maintained at 22°C under broad-spectrum halogen lamps with a 12-h photoperiod.

Total genomic DNA was extracted from 30 mg of young leaf tissue from each plant using a mini-salts protocol previously described by Sambrook et al. (1989) with slight

Table 1. Sample IDs and corresponding adapter and primer sequences. The primer sequences are designed to be complementary to the *EcoRI* and *BfaI* adapters and to carry a unique 5' multiplex identifier (MID) barcode code sequence.

Ligation Adapters					
<i>EcoRI</i> adapter1 (F) 5'/5BioTEG/-CTCGTAGACTGCGTACC					
<i>EcoRI</i> adapter2 (R) CATCTGACGATGGTTAA-/5Phos/5'					
<i>BfaI</i> adapter1 (F) 5'-GACGATGAGTCCTGAG					
<i>BfaI</i> adapter2 (R) ACTCAGGACTCAT-/5Phos/5'					
Pool-I samples	Pool-II [†] Samples	MID barcode primers pairs	Sequence (5'→3') [‡]		
			<i>EcoRI</i>		<i>BfaI</i>
Ler-0 (CS20)		MID1	ACGAGTGGTACTGCGTACCAATTC		ACGAGTGGTACTGAGTCTGAGTAG
Col-4 (CS933)		MID2	ACGCTCGACAGACTGCGTACCAATTC		ACGCTCGACAGATGAGTCTGAGTAG
CS1929	CS1963	MID3	AGACGCATCGACTGCGTACCAATTC		AGACGCATCGATGAGTCTGAGTAG
CS1930	CS1964	MID4	AGCACTGTAGGACTGCGTACCAATTC		AGCACTGTAGGATGAGTCTGAGTAG
CS1931	CS1965	MID5	ATCAGACCGGACTGCGTACCAATTC		ATCAGACCGGATGAGTCTGAGTAG
CS1932	CS1966	MID6	ATATCGCGAGGACTGCGTACCAATTC		ATATCGCGAGGATGAGTCTGAGTAG
CS1933	CS1967	MID7	CGTGTCTTAGACTGCGTACCAATTC		CGTGTCTTAGATGAGTCTGAGTAG
CS1935	CS1968	MID8	CTCGCGTGTGACTGCGTACCAATTC		CTCGCGTGTGATGAGTCTGAGTAG
CS1937	CS1969	MID9	TAGTATCAGCGACTGCGTACCAATTC		TAGTATCAGCGATGAGTCTGAGTAG
CS1940	CS1970	MID10	TCTCTATGCGGACTGCGTACCAATTC		TCTCTATGCGGATGAGTCTGAGTAG
CS1941	CS1971	MID11	TGATACGTCGACTGCGTACCAATTC		TGATACGTCGATGAGTCTGAGTAG
CS1942	CS1972	MID12	TACTGAGCTAGACTGCGTACCAATTC		TACTGAGCTAGATGAGTCTGAGTAG
CS1943	CS1973	MID13	CATAGTAGTGGACTGCGTACCAATTC		CATAGTAGTGGATGAGTCTGAGTAG
CS1944	CS1974	MID14	CGAGAGATACGACTGCGTACCAATTC		CGAGAGATACGATGAGTCTGAGTAG
CS1945	CS1975	MID15	ATACGACGTAGACTGCGTACCAATTC		ATACGACGTAGATGAGTCTGAGTAG
CS1946	CS1976	MID16	TCACGTACTAGACTGCGTACCAATTC		TCACGTACTAGATGAGTCTGAGTAG
CS1947	CS1977	MID17	CGTCTAGTACGACTGCGTACCAATTC		CGTCTAGTACGATGAGTCTGAGTAG
CS1948	CS1978	MID18	TCTACGTAGCGACTGCGTACCAATTC		TCTACGTAGCGATGAGTCTGAGTAG
CS1950	CS1980	MID20	ACGACTACAGGACTGCGTACCAATTC		ACGACTACAGGATGAGTCTGAGTAG
CS1951	CS1981	MID21	CGTAGACTAGGACTGCGTACCAATTC		CGTAGACTAGGATGAGTCTGAGTAG
CS1952	CS1982	MID22	TACGAGTATGGACTGCGTACCAATTC		TACGAGTATGGATGAGTCTGAGTAG
CS1953	CS1983	MID23	TACTCTCGTGGACTGCGTACCAATTC		TACTCTCGTGGATGAGTCTGAGTAG
CS1954	CS1985	MID24	TAGAGACGAGGACTGCGTACCAATTC		TAGAGACGAGGATGAGTCTGAGTAG
CS1955	CS1986	MID25	TCGTGCTCGGACTGCGTACCAATTC		TCGTGCTCGGATGAGTCTGAGTAG
CS1956	CS1987	MID26	ACATACGCGTACTGCGTACCAATTC		ACATACGCGTATGAGTCTGAGTAG
CS1957	CS1988	MID27	ACGCGAGTATGACTGCGTACCAATTC		ACGCGAGTATGATGAGTCTGAGTAG
CS1958	CS1989	MID28	ACTACTATGTGACTGCGTACCAATTC		ACTACTATGTGATGAGTCTGAGTAG
CS1959	CS1990	MID29	ACTGTACAGTACTGCGTACCAATTC		ACTGTACAGTATGAGTCTGAGTAG
CS1960	CS1991	MID30	AGACTATACTGACTGCGTACCAATTC		AGACTATACTGATGAGTCTGAGTAG
CS1961	CS1992	MID31	AGCGTCTGCTGACTGCGTACCAATTC		AGCGTCTGCTGATGAGTCTGAGTAG
CS1962	CS1993	MID32	AGTACGCTATGACTGCGTACCAATTC		AGTACGCTATGATGAGTCTGAGTAG

[‡]MID barcode sequences are italicized.

[†]Pool-II samples were sequenced with both 454-pyrosequencing and the Illumina 76-bp end sequencing.

modification as described by Todd and Vodkin (1996). Extracted DNA was quantified using a ND 1000 Spectrophotometer (NanoDrop Technologies Inc., Montchanin, DE) and diluted to 150 ng μl^{-1} in DNase-free water.

Genome Reduction

The genomic reduction protocol utilized here is described in detail by Maughan et al. (2009). The method is based on the conservation of restriction sites across individuals followed by genome selection and en masse fragment sequencing using next-generation sequencing technology. Specific MID barcode tags are incorporated into the DNA fragments of the individual DNA samples, which are subsequently utilized to deconvolute the sequencing

pool and search for SNPs between the parental samples and genotypes in the RIL samples.

DNA Digestion, Ligation, and Genomic Reduction

A total of 450 ng of total genomic DNA of each DNA sample was separately double digested for 1 h at 37°C with 3 U of the restriction enzymes *EcoRI* and *BfaI* (New England Biolabs, Beverly, MA) in 1x NEB4 restriction buffer. The resultant DNA fragments were immediately ligated with 1.5 μM 5'-TEG biotinylated/3'-phosphorylated *EcoRI* adapters and 15 μM 3'-phosphorylated *BfaI* adapters using 3 U of T_4 ligase (New England Biolabs, Beverly, MA; Table 1) at 16°C for 3 h. DNA fragments smaller than 100 bp were excluded from the reactions via spin chromatography

using Chroma Spin-400 columns (ClonTech Laboratories, Mountain View, CA). Nonbiotin labeled DNA fragments (*Bfa*I to *Bfa*I restriction fragments) were removed from the reaction using a biotin-streptavidin paramagnetic bead separation using M-280 streptavidin beads (Invitrogen, Carlsbad, CA) according to the manufacture's specifications. The remaining *Eco*RI-*Bfa*I and *Eco*RI-*Eco*RI fragments, with the biotin label still attached, were resuspended in 100 μ L of 10:1 TE buffer [10 mM Tris, 1 mM ethylenediaminetetraacetic acid (EDTA), and pH 8.0].

Polymerase Chain Reaction Amplification and Multiplex Identifier Barcode Addition

Thirty-one primer pairs, designed to be complementary to the *Eco*RI and *Bfa*I adaptor sequences and to carry unique 5' 10-base MID barcode sequences, were synthesized by Integrated DNA Technologies (Iowa City, IA; Table 1). A single MID barcode primer pairs was used to amplify 1 μ L of the streptavidin-cleaned DNA fragments for each DNA sample. Since two sequencing runs were utilized, only 31 MID barcode primer sets were needed (Table 1). Amplification of each sample was performed in 50 μ L polymerase chain reactions (PCRs) using 1x Advantage HF 2 PCR Master Mix (ClonTech, Mountain View, CA) and 0.2 uM of the MIDX-*Eco*RI and MIDX-*Bfa*I primer pairs, with the following thermocycling profile: 95°C for 1 min followed by 22 cycles of 95°C for 15 s, 65°C for 30 s, and 68°C for 2 min. The Advantage HF 2 PCR Master Mix was utilized as it is advertised to achieve 30-fold higher fidelity than that seen with wild-type Taq DNA Polymerase, thus minimizing errors introduced via PCR. Successful amplification appeared as a smear of fragments ranging in size from ~200 to 1500 bp.

Next-Generation Sequencing

DNA concentrations for each of the PCR reactions were determined fluorometrically using Quant-iT picogreen dye (Invitrogen, Carlsbad, CA) and were used to construct two 5 μ g pools of DNA, with each pool representing the parental DNA samples (Ler-0 and Col-4) and 29 unique RIL DNA samples (Table 1). Each pool, referred to as pool-I and pool-II, consisted of ~160 ng (~3.2% of the total) of DNA from each of the RIL PCR reactions and ~320 ng of DNA from each of the parental PCR reactions (~6.5% of the total). Each pooled sample was electrophoresized in a separate lane on a 1.5% Meaphor agarose gel (Cambrex BioScience, East Rutherford, NJ) in 0.5x Tris-acetate-EDTA (TAE) at 40V for 8 h and visualized with ethidium bromide staining. A single gel slice from each lane, representing DNA fragments ranging from ~450 to 650 bp, was removed and the DNA fragments extracted using a Qiaquick column (Qiagen, Germantown, MD). Pool-I and pool-II were separately sequenced using standard protocols for 454-pyrosequencing as a service at the Brigham Young University DNA Sequencing Center (Provo, UT) using a Roche-454 GS FLX instrument and Titanium reagents (Branford, CT) without DNA fragmentation. Pool-II was also 76-base end sequenced on a single lane using the Illumina Genome Analyzer IIe as a service from the HCI Microarray Core Facility at the University of

Utah (Salt Lake City, UT) per Illumina protocols. Utilizing sequencing run and MID barcode information, DNA fragments from all 60 individuals could be positively identified (Fig. 1).

Contig Assembly and De Novo Single Nucleotide Polymorphism Detection

DNA reads from two 454-pyrosequencing runs were bioinformatically trimmed and separated into MID barcode pools representing the different RILs and parental lines using the process-tagged sequences function in CLCBio Workbench (v. 4.0; Katrinebjerg, Aarhus N, Denmark). DNA sequence reads from the parental samples (Ler-0 and Col-4) were combined across pyrosequencing runs to make a single pool of reads for each of the parental samples. To identify SNPs between the two parental pools we de novo assembled the parental pools of sequences using the Roche Newbler assembler (v. 2.3) with the minimum overlap length set to 50 bp, the minimum overlap identity set to 95%, and the minimum contig length set to ≥ 200 bp. Putative SNPs within contigs were identified from the exported.ace file using custom perl scripts (Stajich et al., 2002; Maughan et al., 2009) when: (i) coverage depth at the SNP was ≥ 8 , (ii) the minor allele frequency (MAF) represented at least 20% of the reads, and (iii) 90% of the alleles within a parental pool were identical.

DNA reads from the 76-base Illumina GAIIe sequencing by synthesis run, representing only pool-II, were similarly processed into barcode pools and trimmed of their MID and adaptor sequences using CLCBio Workbench (v. 4.0; Katrinebjerg, Aarhus N, Denmark). The sequences from the two parental (Ler-0 and Col-4) samples were referenced mapped to the parental consensus contigs derived from 454-pyrosequencing using the Reference Assembly function with default parameters for short reads in CLCBio Workbench. Single nucleotide polymorphisms between the two parental DNA fragment pools were identified from the exported.ace file as described above.

Recombinant Inbred Line Genotype Calling

Genotype calling within the RIL fragment pools, from both the 454-pyrosequencer and the Illumina GAIIe data sets, was accomplished using a custom perl script pipeline that: (i) produced a local blast database of the 454-pyrosequence parental consensus contigs; (ii) analyzed all reads (both 454-pyrosequence and Illumina) for homology to the parental contig database using BLASTn local alignments; (iii) stored in a local database the base calls (genotype) at each putative SNP positions as well as associated E-value and percent identity of the query sequence for each homologous read; and (iv) reported genotypic calls in a SNP $\times$ RIL JoinMap4 mapping matrix (Van Ooijen 2006) based on user-specified stringency variables. In this proof-of-concept study, stringency variables for reporting of the genotypic calls included: (i) one sequence read was required at the SNP to report the genotype for a RIL, (ii) 90% of the reads within a RIL sequence pool were required to be identical at the SNP for the call to be reported, (iii) only SNPs with

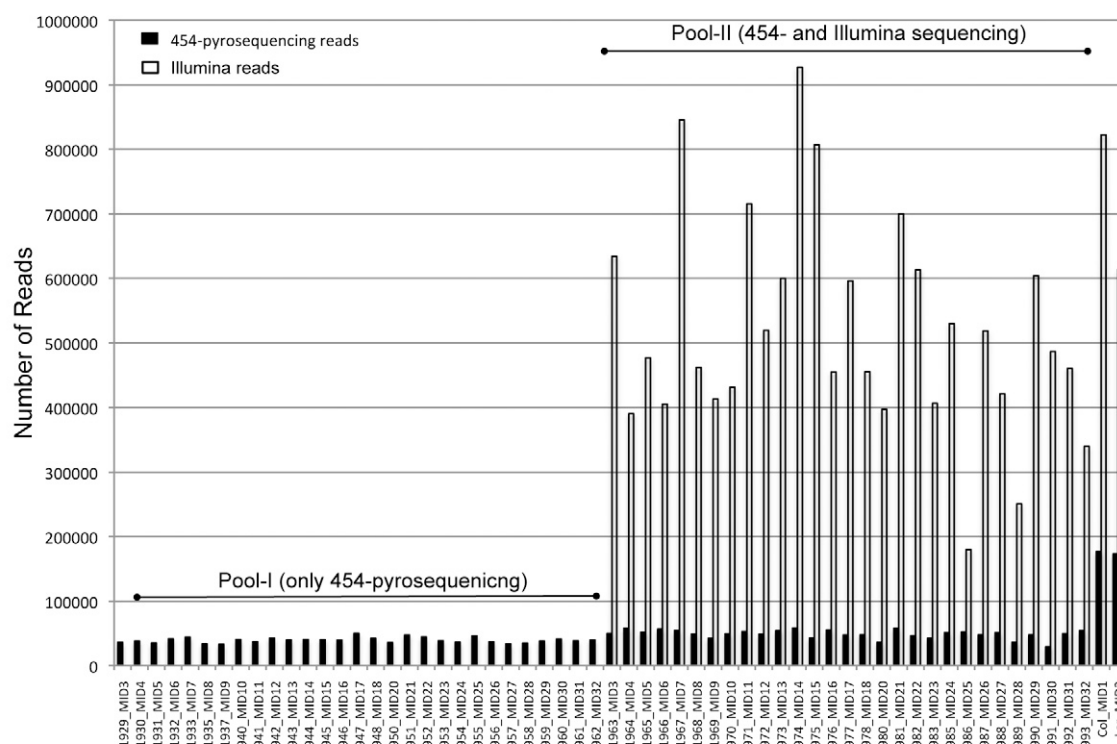


Figure 1. Number of sequence reads obtained for each individual in the pooled DNA samples. Pool-I was only sequenced using 454-pyrosequencing. Pool-II was sequenced with both 454-pyrosequencing and Illumina sequencing. Parental DNA samples were pooled at twice the concentration of the recombinant inbred line (RIL) DNA samples, which were pooled in equimolar concentrations. Parental DNA samples were present in both pools but are shown as a sum of both pools in the figure.

call rates $\geq 80\%$ were reported, and lastly (iv) only genotypic calls from query reads with E-values $\leq 1e-50$ (pyrosequencing dataset) or $\leq 1e-10$ (Illumina GAIIE dataset) and percent identities $\geq 95\%$ relative to their consensus contigs (BLASTn subject) were included in the mapping matrix.

Linkage Mapping

Marker segregation was analyzed for conformation to Mendelian ratios expected in an RIL (F_2) population using chi-squared tests. Single nucleotide polymorphism markers that showed highly significant ($p < 0.001$) were excluded from subsequent analysis. Markers were initially grouped based on independence logarithm of the odds (LOD) scores ≥ 4.0 , calculated by JoinMap 4 for recombination frequency using the G^2 statistic (Van Ooijen, 2006). Markers within groups were then ordered using maximum likelihood mapping algorithms with default parameters. This algorithm utilizes Gibbs sampling to estimate multipoint recombination frequencies, simulated annealing based on Monte Carlo optimization as a marker order search strategy and spatial sampling to escape from local optima (Cheema and Dicks, 2009).

Results

Next-Generation Sequencing and Barcode Parsing

Two full plate runs of the GS-FLX 454-pyrosequencing produced 3,098,246 total reads or 1052 Mb of total base

sequence. The first run, which was used to sequence the pool-II RIL samples, produced 1,710,561 reads with an average read length of 380 bp or ~ 652 Mb of total base sequence, while the second sequencing run, which was used to sequence the pool-I RIL sample, produced only 1,387,685 reads with an average read length of 288 bp or ~ 400 Mb of total base sequence. The difference between the runs was identified as instrument error and not an inherent problem associated with the genomic reduction of pool-I; indeed the genomic reductions for all RIL samples were performed simultaneously. The single lane of 76-base Illumina sequencing of the pool-II samples produced a total of 19,901,025 reads with an average read length of 74 bp or $\sim 1,470$ Mb of total base sequence.

The CLCBio Workbench (v.4.0) separated the reads into MID barcode pools representing each RIL and parental samples. To enter a specific pool, an exact match across the 10 bases of the MID was required. Across the 3,098,246 reads representing the two plates of 454-pyrosequencing runs, a total of 2,920,470 (94%) reads unambiguously sorted into 60 MID barcode pools (58 RILs and 2 parental lines) with an average of 44,317 reads per RIL (Fig. 1). On average each RIL reflected approximately 3.4% of the total number of reads in each run. A total of 173,151 and 176,903 reads were identified that unambiguously matched the MID barcodes of the two parental lines, Ler-0 and Col-4, respectively. The higher number of parental reads was expected and is in agreement with the fourfold increase in input

DNA of the parents into pyrosequencing runs. Increased sequencing of the parental samples was deemed necessary to ensure efficient assembly of the parental consensus contigs utilized for de novo SNP discovery.

Within the 76-base Illumina GAIIe reads from pool-II RIL samples, a total of 16,476,819 reads (83%) were unambiguously assigned to one of the 31 MID barcodes (Fig. 1). On average each RIL reflected approximately 3.1% of the total number of reads, although a fivefold level of variation existed between minimum number of reads (MID25; 180,038 reads) and the maximum number of reads (MID14; 927,263 reads) for a specific RIL, a significant observation considering that across both 454-pyrosequencing sequencing runs only a twofold difference was observed between the minimum and maximum number of reads for specific RILs. The variability in read number within the Illumina GAIIe reads may be the result of inaccurate pipetting during pooling or preferential ligation or bridge amplification of specific MID barcodes during the Illumina sequencing process.

Sequence Assembly

454-pyrosequencing reads assigned to the parental samples (Col-0 and Ler-4) were trimmed and de novo assembled into contigs (≥ 200 bp) using the Newbler assembler (v.2.3). A total of 8573 contigs were assembled, with an average size of 474 bp. Ninety-five percent of the bases within the contigs had a quality scores above 40. The average contig read depth was 27x, while the average base coverage within a contig was 16x. The ace file from the assembly was used for subsequent SNP discovery (see SNP discovery section) while the consensus sequences from the assembly were utilized as a mapping reference for the Illumina GAIIe reads. Of the 1,436,510 Illumina GAIIe reads, derived from the parental samples, 1,080,185 (75%) were successfully mapped back to the 454-parental reference contigs. Of the 25% that failed to map, most did not pass the minimum size parameter to be included in the analysis (<15 bp). Indeed the removal of both the adaptor and MID barcode sequences (25 bp), incorporated during the genomic reduction protocol, substantially reduced the overall size of these fragments. At least one Illumina read successfully mapped to 7179 (84%) of the reference contigs. The average contig read depth for the Illumina contigs was 150x, although a considerable number of reads mapped to a single reference contig homologous to the 25S-18S ribosomal DNA spacer sequence (e.g., 28% of the parental Illumina GAIIe reads). We note that the next most abundant fragment was represented by only 2.5% of the total number of reads. As expected, DNA fragments homologous to the 25S-18S ribosomal DNA spacer sequence were also the most abundant fragment found within the 454-pyrosequencing reads, although they consisted of only 5% of the total 454-reads.

Although the *Arabidopsis* genome has been fully sequenced, we chose not to use it in the assembly process since the vast majority of species, especially those with limited agricultural importance, lack genomic information (i.e., reference genome sequences). However, since

the *Arabidopsis* genome sequence, ecotype Columbia, is publicly available in GenBank (Kaul et al., 2000), we can simulate in silico the genomic reduction process. In silico genome reduction of the *Arabidopsis* nuclear chromosomes (NC_003070, NC_003071, NC_003074, NC_003075, and NC_003076) and two organellar genomes (NC_001284 and NC_000932) produced 8280 fragments. The in silico digestion produced a total of 315,494 DNA restriction fragments. After removal of the *BfaI-BfaI* fragments via paramagnetic bead and electrophoretic size selection (450–650 bp), a total of 7692 *BfaI-EcoRI* and 588 *EcoRI-EcoRI* restriction fragments remain. Thus the genomic reduction should have reduced the *Arabidopsis* genome complexity by 26-fold, leaving ~ 4.5 Mb of DNA for sequencing. Of the 8280 in silico restriction fragments, 7467 (90%) had at least one homologous read within the parental 454-pyrosequencing read dataset. As expected, greater than 75% of all parental sequences mapped to fragments within the 450 to 650 bp size range, with decreasing coverage at the range ends (Fig. 2). The overlap between the de novo assembled contigs and the in silico predicted fragments was 80%, suggesting that while the majority of in silico fragments were sequenced, many did not form contigs that passed the stringent assembly parameters utilized during the de novo assembly (minimum contig depth = 2, minimum overlap = 50 bp, and minimum overlap identity = 95%). Furthermore, we note that there are some contigs in the de novo assembly that do not show homology with fragments in the in silico-derived set, suggesting that some contigs are likely from restriction fragments from outside the size range (presumably due to trapping in the gel) or novel restriction fragments not predictable from the Columbia genome sequence.

Single Nucleotide Polymorphism Detection

Putative SNPs were identified in the ace files from the 454-pyrosequencing and Illumina GAIIe de novo and reference mapping assemblies, respectively. Similar but slightly relaxed minimum criteria were employed in the identification of the putative SNPs as compared to those reported by Maughan et al. (2009), including a minimum base coverage threshold of $\geq 8x$ and a minimum minor allele frequency (MAF) $\geq 20\%$. The parameters were relaxed in an effort to maximize the number of putative SNPs identified, with the expectation that erroneous SNPs would be subsequently identified and eliminated during the mapping phase, as they would be nonsegregating or show highly distorted segregation ratios.

A total of 6159 SNPs in 2389 contigs were identified from the 454-pyrosequencing dataset, while only 701 SNPs in 575 contigs were identified from the Illumina dataset (Table 2). The reduced number of SNPs identified in the Illumina dataset was not unexpected, considering that only the first and last 48 bases of the restriction fragments are sequenced using the Illumina sequencing technology. Indeed, considering that the average size of the restriction fragments is ~ 474 bp and assuming an even distribution of the SNPs across the fragments, the maximum SNP discovery rate within the Illumina

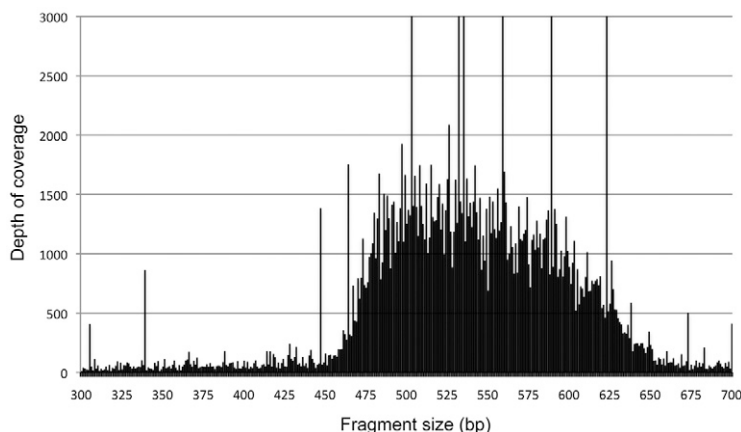


Figure 2. Depth of coverage of in silico predicted *EcoRI-EcoRI* and *EcoRI-BfaI* restriction fragments detected in the parental 454-pyrosequencing dataset. Target fragment size was 450 to 650 bp.

dataset should be only 20% of that identified using the 454-pyrosequencing technology.

The average base coverage at a SNP in the pyrosequencing dataset was 19x, while in the Illumina dataset it was 49x, reflecting the substantially deeper sequencing of the Illumina sequencing by synthesis technology (Table 2). The SNP density (SNP per bp) in the two datasets was similar at 1 per 660 bp for 454-pyrosequencing and 1 per 983 bp for Illumina sequencing. Across both sequencing platforms, transition ($A \leftrightarrow G$ or $C \leftrightarrow T$) outnumbered the transversion ($A \leftrightarrow T$, $C \leftrightarrow A$, $G \leftrightarrow C$, $G \leftrightarrow T$) mutations by an average 1.4x margin (Table 3). Transition mutations, believed to be the result of hypermutability effects of CpG dinucleotide sites and deamination of methylated cytosines, are reported as the most frequent class of SNPs in both plant and animal genomes (Morton et al., 2006; Zhang and Zhao, 2004). All SNPs, including 5' and 3' sequencing information, have been submitted to the dbSNP in GenBank under the handle MAUGHAN in batch number 2010A.

Genotype Calling and Linkage Mapping in the 454-Pyrosequencing Dataset

Genotype calling from the 454-pyrosequencing dataset for the individual RILs was accomplished using BLASTn homology searches between individual RIL sequence reads and the 454-pyrosequencing derived parental consensus contigs. The BLASTn local alignments for each read sequence were parsed relative to the SNP position. The genotype (base call), associated E-value, and percent identity (% ID) of the read (query) sequences to the homologous consensus contigs were stored in a

local database. For the genotypic call to be reported, the call had to be supported by a minimum percent identity of $\geq 95\%$ and a minimum E-value $\leq 1e-50$. Keeping the E-value and percent identity at elevated levels eliminated the reporting of reads with pseudo homologies to more than one consensus contig sequence. Since the mapping population consisted of homozygous lines ($F_2 = 99.8\%$ homozygous), if one or more reads were available for a specific SNP $\times$ RIL combination and the major frequency allele represented greater than 90% of the calls at the SNP, the genotype was reported as homozygous, otherwise the genotype was reported as heterozygous.

As reported earlier, the read depth varied across the SNPs such that some SNPs had genotypic calls for all RIL individuals, while other SNPs could only be genotyped in a subset of RIL individuals. For example, at a 10% threshold for maximum missing data (MMD) at a SNP locus (i.e., ≥ 53 RILs were scored at the SNP in question) a total of 626 SNPs were reported in the genotyping matrix, while more relaxed parameters for missing data produced significantly more data (Table 3). With conventional marker systems (restriction fragment length polymorphism [RFLP], microsatellite, amplified fragment length polymorphisms [AFLPs], etc.) it is not uncommon to utilize a MMD threshold per marker of 20%—thus, all subsequent analyses reported here were performed on the 20% MMD dataset. In this dataset, a total of 1712 SNPs were reported for the 58 RILs and 2 parental lines, totaling 89,955 genotypic data points, of which 46.3 and 53.3% were scored as homozygotes with the Col-4 and Ler-0 genotypes, respectively. The remaining 0.4% were called as heterozygotes.

Table 2. Summary of contigs assembled and single nucleotide polymorphisms (SNPs) identified for each sequencing technology.

Method	Contig metrics				Unique contigs	SNP base coverage	MAF [†] range			SNP type					
	Contigs	Bases	Avg. contig size (bp)	SNPs			20–29%	30–39%	40–50%	A/C	A/G	A/T	C/G	C/T	G/T
454-pyrosequencing	8573	4,063,602	474	6159	2389	19x	26	34	40	11%	27%	16%	7%	28%	11%
Illumina GAIIe	7179	688,753	48	701	575	49x	19	38	43	9%	30%	11%	9%	30%	11%

[†]MAF, minor allele frequency.

Table 3. The number of single nucleotide polymorphisms (SNPs), data points, and average coverage in the mapping analysis at various levels of maximum allowed missing data. Cost per data point estimates are based on in-house sequencing costs.

Method	Maximum missing data per SNP	Number of SNP genotyped	Number of data points genotyped	Avg. coverage at the genotype	Avg. missing data per RIL [†] (Min.—Max. %)	Cost per data point [‡]
454-pyrosequencing	10%	626	35,083	5.3x	6.5% (0.9–14.7%)	\$0.343
454-pyrosequencing	20% [§]	1712	89,955	4.5x	12.2% (3.6–24.3%)	\$0.134 (\$0.147) [¶]
454-pyrosequencing	30%	2825	139,571	4.0x	17.5% (6.0–33.7%)	\$0.086
Illumina GAlle	10%	461	13,993	19.1x	2.2% (0.2–12.2%)	\$0.076
Illumina GAlle	20%	504	15,119	18.2x	3.4% (0.9–14.1%)	\$0.070 (0.113) [¶]
Illumina GAlle	30%	526	15,615	17.9x	4.4% (1.5–15.6%)	\$0.068

[†]RIL, recombinant inbred line.

[‡]Pyrosequencing cost include: 2 runs at \$5,800 per 454 sequencing run and genomic reduction cost (61 samples at \$6.80 per sample) only. Illumina GAlle sequencing cost include: 1 lane at \$854 per sequencing lane and genomic reduction cost (31 samples at \$6.80 per sample) only. Labor costs are not included.

[§]Maximum missing data (MMD) describes the maximum amount of missing data allowed per SNP in order for that SNP marker to be included in the analysis. At the 20% MMD, the total amount of missing data across all markers was 12.4%.

[¶]Cost adjusted for SNPs in complete linkage disequilibrium or dropped from the mapping analysis due to distorted segregation ratios.

If there had been no bias in the selection of individuals during the development of the RILs, a 1:1 ratio of the parental alleles should be observed at all marker loci, however a 1:1.1 ratio in favor of the Ler-0 alleles was observed. The bias was not unexpected, as significantly distorted segregation ratios for RFLP markers were originally reported in the population by Lister and Dean (1993). The distortion favors the female in the cross (Ler-0), suggesting the potential that some of the SNPs may have been of extranuclear origin and showing nonmendelian maternal inheritance. Indeed, four markers in the data set showed complete distortion for the Ler-0 allele and subsequently were shown to map to the *Arabidopsis* chloroplast genome sequence (GenBank accession NC_000932.1).

454-Based Single Nucleotide Polymorphism Linkage Mapping

Of the 1712 SNPs utilized for mapping, 51 (3.0%) showed significant segregation distortion ($p < 0.0001$) and were removed before attempting linkage mapping. Another 106 (6.2%) SNP loci were also removed from the analysis

since they showed identical segregation patterns with other SNP markers (complete linkage disequilibrium). Thus, a total of 1555 SNP loci were included in the linkage analysis. At a minimum independence LOD score of 6.0, pairwise linkage analysis grouped all 1555 SNP loci into five separate linkage groups. The distribution of the markers within the linkage groups varied from 252 to 380 SNPs per linkage group (Table 4). Maximum likelihood mapping of the pairwise linkage groups successfully ordered all SNP markers within their respective linkage groups. The cM distance spanned by the SNP markers in the linkage groups ranged from a low of 102 to 174 cM (Table 4). The largest interval between two linked markers was 8.5 cM on chromosome 5, and the average distance between all loci was 0.44 cM or 174 Kb (Table 4). Most intervals (95%) were less than 2 cM apart.

The genomic location of each SNP was determined via BLASTn alignment of the parental contig sequences containing the SNPs relative to the complete *Arabidopsis* genome (including the chloroplast and mitochondrial genomes). From the comparison, we clearly identify the

Table 4. Linkage group summary based on 454-pyrosequencing derived 20% maximum missing data (MMD) dataset.

	Chromosome 1 (LG1)	Chromosome 2 (LG2)	Chromosome 3 (LG3)	Chromosome 4 (LG4)	Chromosome 5 (LG5)	Total
Total SNPs [†]	380 (24%)	271 (17%)	300 (19%)	252 (16%)	352 (23%)	1555
Chromosome size (Mb)	30.4	19.7	23.5	18.6	26.9	119.1
Distanced spanned (cM)	174	127	131	102	149	683
Largest interval (cM)	6.0	6.1	7.7	6.1	8.5	N/A [‡]
Marker density 1 per cM or Kb	0.46 cM or 174 Kb	0.47 cM or 155 Kb	0.44 cM or 179 Kb	0.40 cM or 182 Kb	0.42 cM or 181 Kb	0.44 cM or 174 Kb

[†]SNP, single nucleotide polymorphism.

[‡]NA, not applicable.

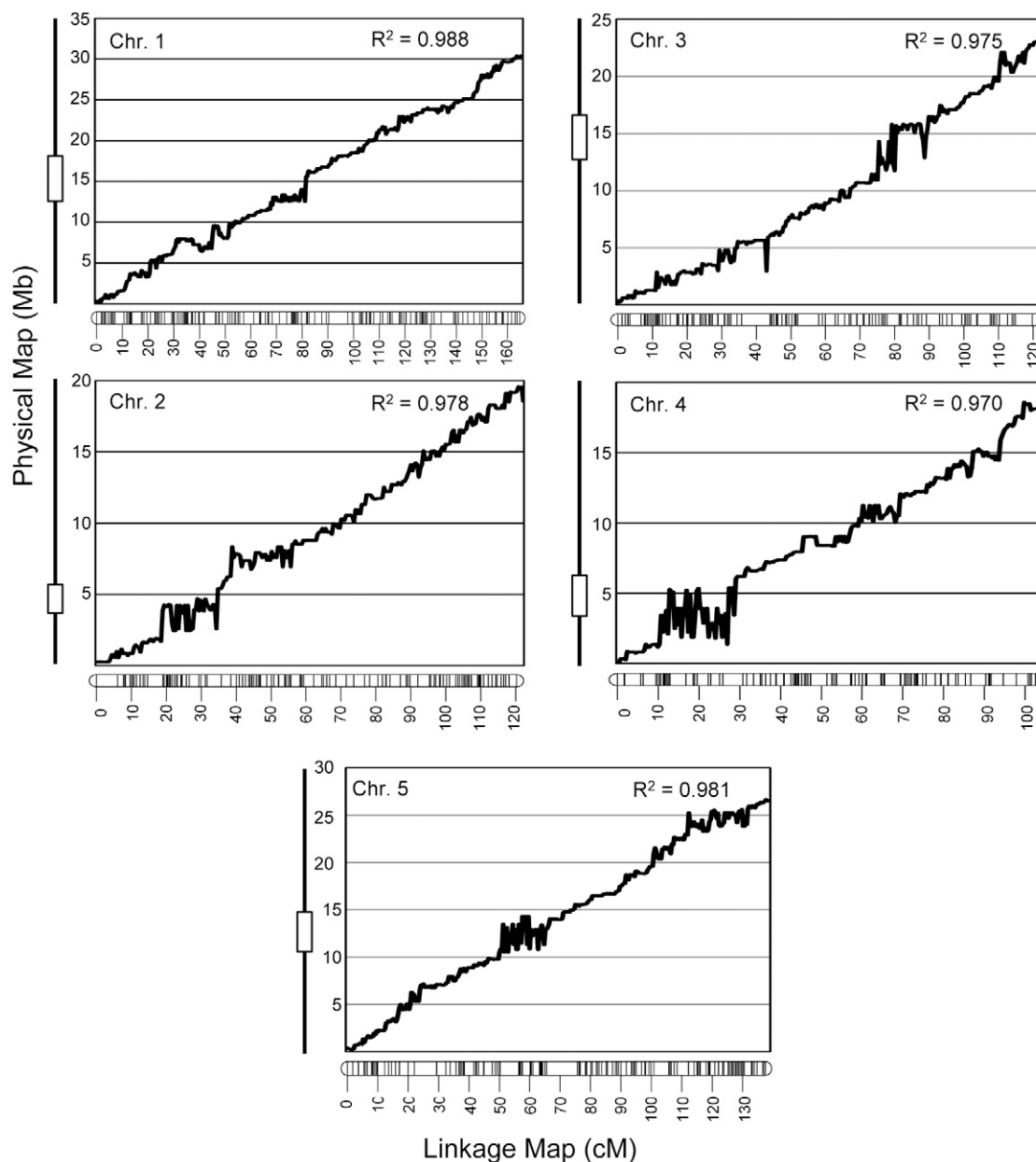


Figure 3. Collinearity analysis of the *Arabidopsis* physical maps (Mb) and linkage maps (cM) based on single nucleotide polymorphism (SNP) genotyping from 454-pyrosequencing. A depiction of the approximate physical location of the centromeric region is given to the side of the x axis, while a depiction of the linkage group provide beneath the y axis. Chr, chromosome.

association between linkage groups and physical chromosomes (Table 4; Fig. 3). Without exception, every SNP marker linked via pairwise linkage mapped to the same physical *Arabidopsis* chromosome. Furthermore, analysis of the collinearity of the linkage map and physical map showed that the linkage order of the SNP was highly correlated with physical order ($r = 0.99$). While the overall linkage and physical maps were highly collinear, several of the linkage group exhibited concentrated areas of noncollinearity (i.e., linkage group [LG] 4; Fig. 3), which corresponded, in all cases, with the physical location of the centromere (Hosouchi et al., 2002). It is well known that crossing over is suppressed at centromeres (Haupt et al., 2001; Talbert and Henikoff, 2010), with estimates of

crossover suppression ranging from fivefold to >200-fold depending on the organism (Talbert and Henikoff, 2010). Thus, these areas of noncollinearity are likely the result of insufficient numbers of crossover events to accurately order the SNP markers spanning the centromere—a situation likely exacerbated by the limited number of RIL individual utilized in this proof-of-concept experiment.

Genotype Calling and Linkage Mapping in the Illumina Dataset

Genotypic calling from the Illumina dataset at SNP loci for the individual RILs was accomplished using the same custom perl script pipeline, except that the minimum E-value was adjusted to $\leq 1e-10$ to allow for the significantly smaller

Table 5. Linkage group summary based Illumina derived 20% maximum missing data (MMD) dataset.

	Chromosome 1	Chromosome 2	Chromosome 3	Chromosome 4	Chromosome 5	Total
No. of corresponding linkage groups (LGs)	5 (LG6, 8, 10, 13, 14)	3 (LG2, 9,11)	3 (LG3, 7, 12)	2 (LG4, 5)	1 (LG1)	14
Total SNPs [†]	74 (24%)	53 (17%)	57 (18%)	51 (16%)	76 (24%)	311
Chromosome size (Mb)	30.4	19.7	23.5	18.6	26.9	119.1
Distance spanned (cM)	91	53	73	57	114	388
Largest interval (cM)	6.4	3.9	6.1	6.1	12.4	N/A [‡]
Marker density 1 per cM or Kb	1.2 cM or 410 Kb	1.0 cM or 371 Kb	1.28 cM or 412 Kb	1.1 cM or 365 Kb	1.5 cM or 354 Kb	1.2cM or 382 Kb

[†]SNP, single nucleotide polymorphism.[‡]NA, not applicable.

Illumina DNA fragment read. Using a MMD threshold of 20%, a total of 504 SNP were reported in the genotyping matrix. Unlike the 454-pyrosequencing data, relaxing the MMD threshold to 30% only marginally affected the number of SNPs (22) included in the genotyping dataset (Table 3), suggesting that most SNP in the Illumina dataset had been sequenced in excess. Indeed, the coverage at the SNPs in the 20% MMD data averaged 18.2x coverage, while coverage at the same MMD threshold in the 454-pyrosequencing dataset was only 4.5x (Table 3). In the Illumina 20% MMD dataset, a total of 15,119 (87.6%) genotypic data points were score, of which 1.0% were called as heterozygotes. The remaining 505 (3.2%) data points were reported as missing. We also observed a slight distortion (1:1.15) in favor of the Ler-0 alleles in the dataset.

Illumina-Based Single Nucleotide Polymorphism Linkage Analysis

Due to the constrained number of RILs (29) included in the Illumina sequencing run we recognized, a priori, the limited ability of the dataset to produce fully ordered linkage maps. Of the 504 SNP genotyped, 21 (4.1%) showed significant segregation distortion ($p < 0.0001$) and were removed before attempting pairwise grouping. Another 170 (34%) SNP loci were also removed from the analysis since they were in complete linkage disequilibrium with other SNP markers. The increase in the number of SNPs that were in complete linkage disequilibrium was not unexpected since fewer consensus contigs revealed SNPs—a result reflecting the fact that only SNPs in the first and last 50 bp of the consensus contigs could be detected using the Illumina sequencing technology. Thus, a total of 313 SNP loci were included in the linkage analysis. At a minimum independence LOD score of 4.0, pairwise linkage analysis grouped 311 (99.4%) of the SNP loci into 14 separate linkage groups. The distribution of the SNPs markers within the linkage groups varied from 76 to 2 per linkage group. The distance spanned by the SNP markers in the linkage groups ranged from a low of 1.2 cM (LG12; 2 loci) to 114 cM (LG1; 76 loci) (Table 4).

BLASTn analysis of these SNPs relative to *Arabidopsis* physical chromosomes clearly identifies associations between linkage groups and physical chromosomes, although several linkage groups often corresponded to a single *Arabidopsis* chromosome (Table 5). Maximum

likelihood mapping of the pairwise linkage groups ordered all SNP markers within their respective linkage groups. To analyze collinearity of the linkage and physical maps we correlated the positions of the SNPs within each linkage group to their corresponding physical chromosomal locations (Fig. 4). The average Pearson's correlation coefficient (r) of 0.98 ($R^2 = 0.96$) was slightly lower than that seen with the larger 454-pyrosequencing dataset (Table 4). The decreased correlation between the physical and linkage map of the Illumina dataset was not unexpected, since fewer RILs in the Illumina dataset meant fewer informative recombination events, inevitably leading to decreased accuracy during the marker ordering process and the inability to consolidate linkage groups. An example of this effect on ordering is seen in the linkage groups corresponding to *Arabidopsis* chromosome 2, where the linkage analysis was unable to resolve the true linear order of the markers, resulting in a 16.4 cM (4.6 Mb) inverted segment corresponding to the middle chromosome (Fig. 4). Nonetheless, considering the limited number of RIL utilized, with only a few exceptions, the linkage and physical maps are well correlated.

Cross-Platform Genotypic Score Validation

To further validate the genotyping process, genotypic calls were compared for accuracy across the two sequencing platforms. Across the two sequencing datasets, 113 common SNPs were genotyped and could be compared for accuracy. A total of 3503 data points were analyzed; 313 were missing in one or the other dataset and could not be used in the comparison. Of the remainder, 3187 were identical matches (99.91%) while three (0.09%) were mismatches. Interestingly, all three mismatches were from a single RIL individual (TAIR #CS1989) and all involved conflicts between homozygous to heterozygous calls. No opposing homozygous call conflicts were identified. Further examination of the genotypic calls for individual #CS1989 in the full 454-pyrosequencing dataset evidenced an abnormally large number (5.1%) of heterozygous calls for this RIL. Indeed, the next highest percentage of heterozygous calls was 1.2% and the average across all individuals was 0.22%. Since the population had been selfed to the $F_{2,9}$ stage, we expected an average heterozygous call rate of ~0.2%, suggesting that sample #CS1989 may represent a seed misclassification or reflect a contamination that occurred during the preparation of the sample.

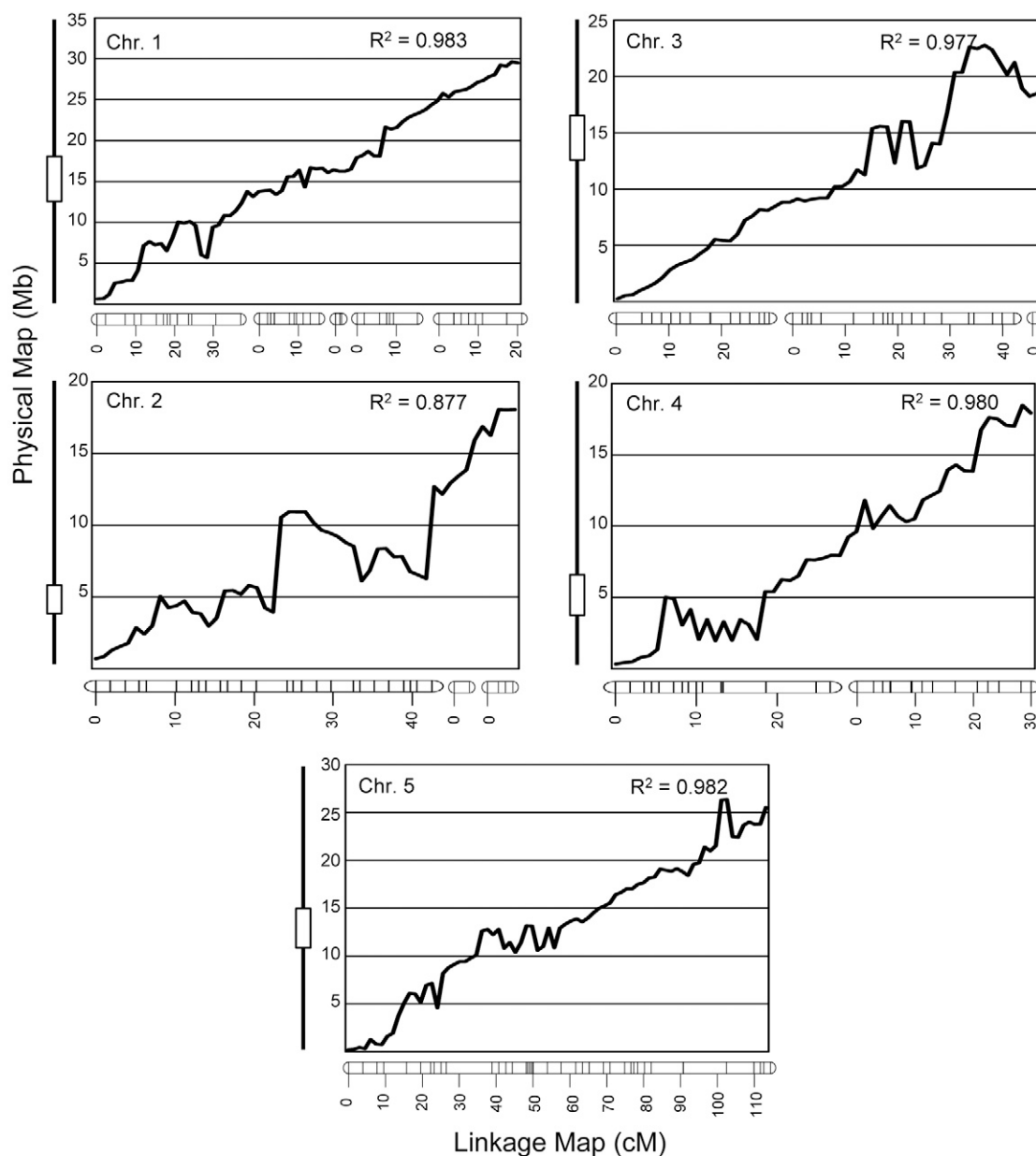


Figure 4. Collinearity analysis of the *Arabidopsis* physical maps (Mb) and linkage maps (cM) based on single nucleotide polymorphism (SNP) genotyping from Illumina sequencing. A depiction of the approximate physical location of the centromeric region is given to the side of the x axis, while a depiction of the linkage group provide beneath the y axis. Chr, chromosome.

Discussion

We have shown that next-generation sequencing (454-pyrosequencing and Illumina Genome Analyzer) coupled with genomic reduction and MID barcodes can be utilized to simultaneously identify SNPs and genotype segregating populations. The method eliminates the necessity, and often-expensive and time-consuming process, of developing independent genotype assays (e.g., TaqMan) for each newly discovered SNP.

Coverage

Of primary importance to the genomic reduction method is the necessity of achieving sufficient genomic reduction to allow adequate sequencing coverage for SNP detection and robust genotype calling. Here we utilized a F_2 RIL population,

in which the residual level of heterozygosity was less than 0.2%, which allow us to genotype based on a single read. Accurate genotyping, however, in diploid populations where heterozygous loci may be present (e.g., F_2 or backcross populations) requires a minimum of 4x and 7x sequence coverage at each SNP to achieve a genotype scoring confidence level of 95 and 99%, respectively. The massive throughput of the Illumina GAIIe sequencer should be advantageous for accurate genotyping; however, the short Illumina reads significantly reduced the number of SNPs identified, since only a small portion of each restriction fragment was sequenced.

Since coverage is a function of the total number of fragments in the pooled library and will vary relative to the genome size of the organism it is important to attempt to manage the level of coverage before any genotyping by

sequencing experiment. For example, large genomes (>500 Mb) may require the use of more infrequent cutting endonucleases (8-base recognition site) to obtain the desired magnitude of genomic reduction. The number of fragments produced from the genomic reduction can be estimated by multiplying the genome size of the organism by the predicted abundance of recognition site of the rare cutting endonuclease. For example, the predicted frequency of an *EcoRI* cut site (GAATTC) in an organism with a genome size of 466 Mb and a predicted GC content of 35% would be 0.00034 or 159,219 fragments, of which only ~20% (~31,843 fragments) are predicted to be in the sequenced fragment size range (450–650 bp). Once the number of fragments remaining has been determined, the average coverage for each fragment (contig) can be estimated based on the number of individuals in the pool and the number of reads produced by the next-generation sequencing technology. For example, a single 454-pyrosequencing run (~1.5 million reads) of a 32-sample pooled library consisting of ~31,843 unique fragments, should achieve ~1.5x sequencing coverage at each fragment for each individual, whereas the same sequencing library would have an expected 19x sequencing coverage from a single lane of Illumina GAIIe sequencing (~20,000,000 reads).

Cost per Data Point

Cost per data point is often the most critical consideration in any genotyping experiment. Current cost estimates for genotyping ranges from \$0.15 to \$0.25 per data point, depending on the genotyping volume and assay technology used. This estimate does not include the cost associated with assay development for each SNP or the labor cost associated with running individual assays (most SNP assays cannot be multiplexed). Genotyping by sequencing, utilizing genomic reduction, and pooled DNA samples as reported here had genotyping costs in the range of \$0.07 to \$0.13 per data point, depending on the sequencing technology and the stringency of the parameters employed during the genotyping process (Table 2). After adjusting for the removal of skewed and redundant SNPs, the cost per data point only increased to \$0.113 for the Illumina dataset and \$0.147 for the 454-pyrosequencing dataset. Consideration should also be given to the fact that (i) a priori information regarding SNP markers was not required, (ii) no technology or costs specific SNP assay development was necessary, (iii) since the parents and the progeny were sequenced simultaneously, no initial parental screen was needed, and lastly (iv) the genomic reduction process is simple, requiring less than 8 h to accomplish for the 61 samples utilized in this experiment (time for sequencing varied based on sequencing queues). Finally, next-generation sequencing costs have dropped significantly over the last few years, even as the number of reads per run has increased (Mardis, 2008). Any continued sequencing cost reductions or sequencing read volume increases will undoubtedly make genotyping by sequencing even more attractive.

Limitations

As with all genotyping systems, potential limitations and unknowns exist with genotyping by sequencing via genome

reduction. First, genomic reduction based SNP discovery and/or genotyping experiments have only been conducted with *Arabidopsis*, a diploid species, and amaranth (*Amaranthus caudatus* L.; Maughan et al., 2009) an ancient allo-tetraploid that shows amphidiploid inheritance. Questions regarding the use of the method in more recent allopolyploids or autopolyploid remain unclear. Undoubtedly, the potential assembly of paraologous and/or orthologous regions during the assembly process would complicate the de novo SNP discovery process. However, SNPs in such assemblies should fail the identity within-parent uniformity test and be automatically removed from subsequent genotyping analyses (Maughan et al., 2009). Second, the level of polymorphism in the population analyzed significantly influences data point cost estimates. In populations derived from narrowly diverse materials, fewer SNPs will be detected, leading to an increase in the final cost per data point. Third, linkage mapping in species that are self-incompatible is complicated by the required use of heterozygous parental lines (Hittalmani et al., 2008). In the de novo SNP identification method described here, heterozygous loci in the parental lines are eliminated by the within-parent uniformity test. Thus a different bioinformatic approach will need to be established to identify true SNPs (vs. sequencing errors) for populations derived from heterozygous parents. Fourth, de novo assembly of the short Illumina reads is highly problematic. Indeed, the Illumina assemblies reported here were based on Illumina reads mapped to consensus contigs generated from the 454-pyrosequencing reads. Thus species without readily acquired reference sequences will require at least a partial 454-pyrosequencing (or other >100 bp sequencing system) run of the parental lines to generate a reference framework for SNP discovery and SNP genotyping using the current Illumina system. Lastly, genomic reduction cannot target a specific SNPs, thus the method would be unsuitable for protocols requiring the genotyping of a single or few marker loci (e.g., marker assisted selection). However, once a trait or QTL has been linked to a specific SNP marker, sequence information from the parental consensus contigs could be used to convert the targeted SNP(s) into a traditional SNP assay format (e.g., TaqMan).

Acknowledgments

This research was supported by a Mentoring Environment Grant (MEG) from the Brigham Young University. We gratefully acknowledge Drs. Edward Wilcox (BYU) and Brian Dalley (HCI, University of Utah) for their assistance with 454- and Illumina-sequencing, respectively.

References

- 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.
- Batley, J., G. Barker, H. O'Sullivan, K.J. Edwards, and D. Edwards. 2003. Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. *Plant Physiol.* 132:84–91.
- Batley, J., and D. Edwards. 2007. SNP applications in plants. p. 95–102. *In* N. C. Oraguzie et al. (ed.) *Association mapping in plants*. Springer, New York, NY.
- Bundock, P.C., F.G. Elliott, G. Ablett, A.D. Benson, R.E. Casu, K.S. Aitken, and R.J. Henry. 2009. Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploid plant species using 454 sequencing. *Plant Biotechnol. J.* 7:347–354.

- Cheema, J., and J. Dicks. 2009. Computational approaches and software tools for genetic linkage map estimation in plants. *Brief. Bioinform.* 10:595–608.
- Ching, A., K. Caldwell, M. Jung, M. Dolan, O. Smith, S. Tingey, M. Morgante, and A. Rafalski. 2002. SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. *BMC Genet.* 3:19.
- Choi, I.Y., D.L. Hyten, L.K. Matukumalli, Q. Song, J.M. Chaky, C.V. Quigley, K. Chase, K.G. Lark, R.S. Reiter, M.S. Yoon, E.Y. Hwang, S.I. Yi, N.D. Young, R.C. Shoemaker, C.P. van Tassell, J.E. Specht, and P.B. Cregan. 2007. A soybean transcript map: Gene distribution, haplotype and single-nucleotide polymorphism analysis. *Genetics* 176:685–696.
- Cramer, E.R.A., L. Stenzler, A.L. Talaba, C.A. Makarewich, S.L. Vehrenkamp, and I.J. Lovette. 2008. Isolation and characterization of SNP variation at 90 anonymous loci in the banded wren (*Thryothorus pleurostictus*). *Conserv. Genet.* 9:1657–1660.
- Cronk, Q.C. 2005. Plant eco-devo: The potential of poplar as a model organism. *New Phytol.* 166:39–48.
- Deschamps, S., and M.A. Campbell. 2010. Utilization of next-generation sequencing platforms in plant genomics and genetic variant discovery. *Mol. Breed.* 25:553–570.
- Deschamps, S., M. la Rota, J.P. Ratashak, P. Biddle, D. Thureen, A. Farmer, S. Luck, M. Beatty, N. Nagasawa, L. Michael, V. Llaca, H. Sakai, G. May, J. Lightner, and M.A. Campbell. 2010. Rapid genome-wide single nucleotide polymorphism discovery in soybean and rice via deep resequencing of reduced representation libraries with the Illumina genome analyzer. *Plant Gen.* 3:53–68.
- Garg, K., P. Green, and D.A. Nickerson. 1999. Identification of candidate coding region single nucleotide polymorphisms in 165 human genes using assembled expressed sequence tags. *Genome Res.* 9:1087–1092.
- Gore, M.A., M.H. Wright, E.S. Ersoz, P. Bouffard, E.S. Szekeres, T.P. Jarvie, B.L. Hurwitz, A. Narechania, T.T. Harkins, G.S. Grills, D.H. Ware, and E.S. Buckler. 2009. Large-scale discovery of gene-enriched SNPs. *Plant Gen.* 2:121–133.
- Grivet, L., J.C. Glaszmann, M. Vincentz, F. da Silva, and P. Arruda. 2003. ESTs as a source for sequence polymorphism discovery in sugarcane: Example of the *Adh* genes. *Theor. Appl. Genet.* 106:190–197.
- Haupt, W., T.C. Fischer, S. Winderl, P. Frasz, and R.A. Torres-Ruiz. 2001. The centromere1 (CEN1) region of *Arabidopsis thaliana*: Architecture and functional impact of chromatin. *Plant J.* 27:285–296.
- Hayashi, K., N. Hashimoto, M. Daigen, and I. Ashikawa. 2004. Development of PCR-based SNP markers for rice blast resistance genes at the *Piz* locus. *Theor. Appl. Genet.* 108:1212–1220.
- Hittalmani, S., T.N. Girish, H. Biradar, and P.J. Maughan. 2008. Development of mapping populations: Descriptions and implications. p. 69–91. *In* C. Kole (ed.) *Principles and practices of plant molecular mapping and breeding*. Springer-Verlag, Berlin, Germany.
- Hosouchi, T., N. Kumekawa, H. Tsuruoka, and H. Kotani. 2002. Physical map-based sizes of the centromeric regions of *Arabidopsis thaliana* chromosomes 1, 2, and 3. *DNA Res.* 9:117–121.
- Kaul, S., H.L. Koo, J. Jenkins, M. Rizzo, T. Rooney, L.J. Tallon, T. Feldblyum, W. Nierman, M.I. Benito, X.Y. Lin, et al. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815.
- Kawuki, R., M. Ferguson, M. Labuschagne, L. Herselman, and D.-J. Kim. 2009. Identification, characterisation and application of single nucleotide polymorphisms for diversity assessment in cassava (*Manihot esculenta* Crantz). *Mol. Breed.* 23:669–684.
- Lister, C., and C. Dean. 1993. Recombinant inbred lines for mapping RFLP and phenotypic markers in *Arabidopsis thaliana*. *Plant J.* 4:745–750.
- Mardis, E.R. 2008. The impact of next-generation sequencing technology on genetics. *Trends Genet.* 3:133–141.
- Maughan, P.J., S.M. Yourstone, E.N. Jellen, and J.A. Udall. 2009. SNP discovery via genomic reduction, barcoding and 454-pyrosequencing in amaranth. *Plant Gen.* 2:260–270.
- Morton, B.R., I.V. Bi, M.D. McMullen, and B.S. Gaut. 2006. Variation in mutation dynamics across the maize genome as a function of regional and flanking base composition. *Genetics* 172:569–577.
- Nasu, S., J. Suzuki, R. Ohta, K. Hasegawa, R. Yui, N. Kitazawa, L. Monna, and Y. Minobe. 2002. Search for and analysis of single nucleotide polymorphisms (SNPs) in rice (*Oryza sativa*, *Oryza rufipogon*) and establishment of SNP markers. *DNA Res.* 9:163–171.
- Osman, A., B. Jordan, P.A. Lessard, N. Muhammad, M.R. Haron, N.M. Riffin, A.J. Sinskey, C. Rha, and D.E. Housman. 2003. Genetic diversity of *Eurycoma longifolia* inferred from single nucleotide polymorphisms. *Plant Physiol.* 131:1294–1301.
- Ossowski, S., K. Schneeberger, R.M. Clark, C. Lanz, N. Warthmann, and D. Weigel. 2008. Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res.* 18:2024–2033.
- Rostoks, N., S. Mudie, L. Cardle, J. Russell, L. Ramsay, A. Booth, J.T. Svensson, S.I. Wanamaker, H. Walia, E.M. Rodriguez, P.E. Hedley, H. Liu, J. Morris, T.J. Close, D.F. Marshall, and R. Waugh. 2005. Genome-wide SNP discovery and linkage analysis in barley based on genes responsive to abiotic stress. *Mol. Genet. Genomics* 274:515–527.
- Russell, J., A. Booth, J. Fuller, B. Harrower, P. Hedley, G. Machray, and W. Powell. 2004. A comparison of sequence-based polymorphism and haplotype content in transcribed and anonymous regions of the barley genome. *Genome* 47:389–398.
- Sambrook, J., E.E. Fritsch, and T. Maniatis. 1989. *Molecular cloning. A laboratory manual*. 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- Schmid, K.J., T.R. Sorensen, R. Stracke, O. Torjek, T. Altmann, T. Mitchell-Olds, and B. Weisshaar. 2003. Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. *Genome Res.* 13:1250.
- Schneider, K., B. Weisshaar, D.C. Borchardt, and F. Salamini. 2001. SNP frequency and allelic haplotype structure of *Beta vulgaris* expressed genes. *Mol. Breed.* 8:63–74.
- Stajich, J.E., D. Block, K. Boulez, S.E. Brenner, S.A. Chervitz, C. Dagdigan, G. Fuellen, J.G.R. Gilbert, I. Korf, H. Lapp, H. Lehvaslaiho, C. Matsalla, C.J. Mungall, B.I. Osborne, M.R. Pocock, P. Schattner, M. Senger, L.D. Stein, E. Stupka, M.D. Wilkinson, and E. Birney. 2002. The Bioperl toolkit: perl modules for the life sciences. *Genome Res.* 12:1611–1618.
- Talbert, P.B., and S. Henikoff. 2010. Centromeres convert but don't cross. *PLoS Biol.* 8:e1000326.
- Todd, J.J., and L.O. Vodkin. 1996. Duplications that suppress and deletions that restore expression from a chalcone synthase multigene family. *Plant Cell* 8:687–699.
- Tyrka, M., L. Blaszczyk, J. Chelkowski, V. Lind, I. Kramer, M. Weilepp, H. Wisniewska, and F. Ordon. 2004. Development of the single nucleotide polymorphism marker of the wheat *Lr1* leaf rust resistance gene. *Cell. Mol. Biol. Lett.* 9:879–889.
- Van Ooijen, J.W. 2006. JoinMap 4, Software for the calculation of genetic linkage maps in experimental populations., Kyazma B.V., Wageningen, the Netherlands.
- Van Tassell, C.P., T.P.L. Smith, L.K. Matukumalli, J.F. Taylor, R.D. Schnabel, C.T. Lawley, C.D. Haudenschild, S.S. Moore, W.C. Warren, and T.S. Sonstegard. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat. Methods* 5:247–252.
- Wiedmann, R.T., T.P. Smith, and D.J. Nonneman. 2008. SNP discovery in swine by reduced representation and high throughput pyrosequencing. *BMC Genet.* 9:81.
- Xu, H., R. Chakraborty, and Y.-X. Fu. 2005. Mutation rate variation at human dinucleotide microsatellites. *Genetics* 170:305–312.
- Zhang, F., and Z. Zhao. 2004. The influence of neighboring-nucleotide composition on single nucleotide polymorphisms (SNPs) in the mouse genome and its comparison with human SNPs. *Genomics* 84:785.
- Zhu, C., M. Gore, E.S. Buckler, and J. Yu. 2008. Status and prospects of association mapping in plants. *Plant Gen.* 1:5–20.
- Zhu, Y.L., Q.J. Song, D.L. Hyten, C.P. Van Tassell, L.K. Matukumalli, D.R. Grimm, S.M. Hyatt, E.W. Fickus, N.D. Young, and P.B. Cregan. 2003. Single-nucleotide polymorphisms in soybean. *Genetics* 163:1123–1134.