

SNP Discovery via Genomic Reduction, Barcoding, and 454-Pyrosequencing in Amaranth

Peter J. Maughan,* Scott M. Yourstone, Eric N. Jellen, and Joshua A. Udall

Abstract

The grain amaranths are important pseudo-cereals native to the New World. During the last decade they have garnered increased international attention for their nutritional quality, tolerance to abiotic stress, and importance as a symbol of indigenous cultures. We report the development of a novel genomic reduction protocol based on restriction-site conservation and multiplex identifiers (MID) barcodes to discover single nucleotide polymorphisms (SNPs) from a pooled amaranth library of four mapping parents. The incorporation of MID barcodes allowed for DNA sample pooling, sequence deconvolution, and the identification of SNPs for specific mapping populations without additional genotyping. Approximately 1.3 million sequence-reads with an average read length of 440 bp were collected from a single 454-pyrosequencing run. Contigs specific to each of the four mapping populations were assembled. The assembled contigs had an average read-length of 464 bp, and an average read-depth of 16X. A total of 27,658 SNPs were observed across all populations. The average base coverage at all SNPs was 20X. Thirty-four of 35 (97%) predicted SNPs were verified via resequencing and the random genomic distribution of the SNPs generated using this approach was shown in *Arabidopsis*. The method does not require a priori sequence information and should be readily adaptable to other species.

THE GENUS *Amaranthus* (Caryophyllales: Amaranthaceae) encompasses about 60 species with world-wide distribution (Sauer, 1976). Three species produce edible seeds and collectively are known as the grain amaranths (*A. hypochondriacus* L., *A. cruentus* L., and *A. caudatus* L.). While the grain amaranths remain important food staples in several regions of South and Central America, they have been underutilized since the Spanish conquest, when they were replaced by Old World crops and their cultivation suppressed due to their deeply rooted use in indigenous religious practices (Iturbide and Gispert, 1994; Sauer, 1976; Sauer, 1993). In the last decade they have received new attention, largely due to recognition of the nutritional value of the seed for human consumption as well as their ability to thrive under extreme abiotic stress (Brenner et al., 2000). Among the nutritional characteristics are relatively high seed protein content and a favorable balance of essential dietary amino acids (Becker et al., 1981). The species thrive under warm and dry conditions, an attribute related to their C₄ photosynthesis (Kadereit et al., 2003), and have been suggested as a potential alternative crop in developing nations, especially in overpopulated and undernourished areas (Pal and Khoshoo, 1974; Sauer, 1993). Notwithstanding their renewed potential as an emerging crop species, only a few molecular investigations have been reported for the grain amaranths and even fewer molecular tools, needed for advanced genomic studies, have been developed (Mallory et al., 2008; Maughan et al., 2008).

Single nucleotide polymorphism (SNPs), defined as single-base changes, are the most abundant type of sequence variation in eukaryotic genomes (Batley

Published in The Plant Genome 2:260–270. Published 5 Nov. 2009.
doi: 10.3835/plantgenome2009.08.0022
© Crop Science Society of America
677 S. Segoe 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 University, Dep. of Plant & Wildlife Sciences, Provo, 285 WIDB, UT, 84602. Received 12 Aug. 2009. *Corresponding author (Jeff_Maughan@byu.edu).

Abbreviations: GR-RSC, genomic reduction base on restriction-site conservation; MAF, minor allele frequency; MID, multiplex identifiers; PCR, polymerase chain reaction; SNP, single-nucleotide polymorphism.

et al., 2003; Garg et al., 1999). Theoretically SNPs can have four alleles, but most show only two alleles and are regarded as bi-allelic (Krawczak, 1999). Estimates of SNP densities in plants vary dramatically depending on the species (predominantly self- or cross-pollinating) and whether coding or non-coding regions are considered. For example, in soybean [*Glycine max* (L.) Merr.], SNPs occur at a frequency of 1 per 2038 bp in coding sequence and 1 per 191 bp in non-coding sequence (Van et al., 2005); while in maize (*Zea mays* L.) 1 SNP was observed per 124 bp of coding sequence and 1 per 31 bp in non-coding regions (Ching et al., 2002). The high frequency of SNPs in plant species offers the possibility of constructing extremely dense genetic maps that are particularly valuable for map-based gene cloning efforts as well as haplotype-based association studies. Since SNPs are fundamental changes at the DNA level, their mapping could potentially identify causal mutations (Rafalski, 2002). Compared to tandem repeat (i.e., microsatellites) -based markers, SNPs exhibit a lower mutation rate and thus are less problematic in population genetic analyses (Xu et al., 2005). Indeed, SNPs have already been utilized in wide array of research areas, including association studies (Andrew et al., 2008), conservation genetics (Cramer et al., 2008), genetic diversity analysis (Kawuki et al., 2009), and genetic mapping (Rostoks et al., 2005).

Unfortunately, the initial discovery cost of SNPs in plant genomes has been high and is a limiting factor in their use in many non-model or non-agricultural plant species. 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. When linked with next-generation sequencing, genomic reduction provides a cost-effective means to identify large numbers of high-confidence SNPs. Recently, genomic reduction was successfully used with next-generation sequencing (454-pyrosequencing and Illumina Genome Analyzer) for SNP discovery and allele frequency estimation in cattle (Van Tassel et al., 2008) and pig (Wiedmann et al., 2008). In both cases, pooled restriction fragments of a specific size-range from multiple individuals were sequenced *en masse* and thousands of SNPs were identified. However, since the genomic fragments from several individuals were pooled for sequencing, alleles for specific DNA samples could not be inferred without additional genotyping. In plants, reduced representation libraries have also been used for SNP discovery. Barbazuk et al. (2007) reduced genomic complexity in maize via construction of

tissue-specific cDNA libraries and 454-based transcriptome sequencing. However, due to low sequence coverage, high stringency filtering was required to identify 7016 putative SNPs with an 85% validation rate and, as was the case with reduced representational library strategies in cattle and pig, additional independent genotyping was required to assign specific alleles to specific DNA samples in the sequenced pool.

Here we report the use of 454-pyrosequencing to discover SNPs from a pooled library of four amaranth mapping parents using a novel genomic reduction protocol based on restriction site conservation (GR-RSC) with each sample labeled independently using a unique multiplex identifier (MID) barcode. The incorporation of MID barcodes into specific DNA sequence fragments allows for the unambiguous assignment of fragments to specific samples in the sequence pool, eliminating the need for additional genotyping, while identifying SNPs that will segregate in specific mapping populations. This report provides proof-of-concept of a technique that could be broadly used to discover SNPs in other genomes.

Materials and Methods

Plant Material and DNA Extraction

Four *A. caudatus* plant introduction lines were kindly provided by D. Brenner (USDA, Iowa State University, Ames, IA; Table 1). The four *A. caudatus* plant introduction lines are the parents of four independent F₂ mapping populations that will be used for the future development of the first genetic linkage maps in this species (Table 1). Since an objective of the experiment was to identify SNPs specific to each population, we report SNP discovery for each of the F₂ populations. The four populations are designated Pop1-3, Pop1-4, Pop2-3, and Pop2-4 (Table 1).

Table 1. Description of the *Amaranth caudatus* populations used for single nucleotide discovery.

Population Designation	PI designation	MID barcode	Origin [†]	Perisperm type [‡]	Inflorescence type [‡]	Seedling color [‡]	Plant height [‡] cm	Avg. 100 seed weight [‡] g
Pop1-3								
Female parent	Ames15170	MID1	Nepal	Translucent	Drooping	Green	200	0.05
Male parent	PI 553073	MID3	USA	Translucent	Drooping	Red	200	0.05
Pop1-4								
Female parent	Ames15170	MID1	Nepal	Translucent	Drooping	Green	200	0.05
Male parent	PI 642741	MID4	Bolivia	Opaque	Erect	Red	250	0.09
Pop2-4								
Female parent	PI 481125	MID2	India	Opaque	Erect	Green	215	0.08
Male parent	PI 642741	MID4	Bolivia	Opaque	Erect	Red	250	0.09
Pop2-3								
Female parent	PI 481125	MID2	India	Opaque	Erect	Green	215	0.08
Male parent	PI 553073	MID3	USA	Translucent	Drooping	Red	200	0.05

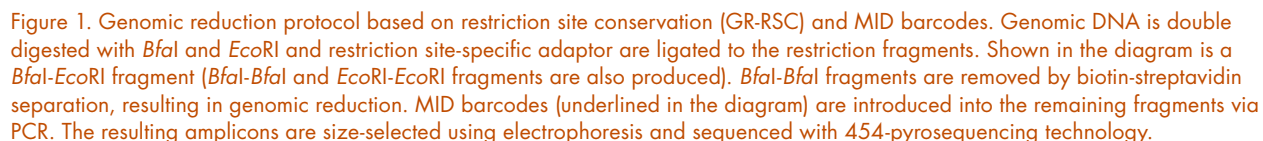
[†]*A. caudatus* is native to the New World and its spread to Asia, Europe, and Africa is a relatively recent event, occurring during post-Colonial American times (Sauer, 1976).

[‡]Data obtained by Germplasm Resources Information Network (<http://www.ars-grin.gov/npgs/>).

Total genomic DNA was extracted from 30mg of freeze-dried leaf tissue from a single plant for each of the four plant introduction lines according to procedures previously described (Sambrook et al., 1989), with modifications described by Todd and Vodkin (Todd and Vodkin, 1996). Extracted DNA was quantified using a Nanodrop (ND 1000 Spectrophotometer, NanoDrop Technologies Inc., Montchanin, DE) and diluted to 150 ng μl^{-1} in DNase-free water.

The principle of the genomic reduction protocol presented here for SNP discovery is based on restriction site conservation across individuals, removal of >90%

Similar to the first steps in amplified fragment length polymorphism analysis (Vos et al., 1995), we double-digest genomic DNA with restriction endonucleases that recognize 4-base and 6-base recognition sites, producing three types of DNA fragments on the basis of their end restriction sites, specifically: 4-base to 4-base, 6-base to 6-base, and 6-base to 4-base fragments. Subsequently, double stranded adapters are ligated to the ends of the digested DNA fragments. The adaptor ligated to the end of the 6-base recognition site is end-labeled with a 5'-biotin molecule, while the adaptor on the 4-base recognition site is unlabeled. Genomic reduction is accomplished by removing the non-labeled DNA fragments (4-base to 4-base fragments) from the reaction using a biotin-streptavidin paramagnetic bead separation. MID barcode sequences are then added to the DNA fragments using PCR primers complementary to the adaptor sequences and that carry a 5' 10-base MID barcode. During the initial PCR cycle, the primers anneal to the adaptor sequence and the MID



barcode is incorporated into the amplified DNA fragment. A high-fidelity polymerase with proofreading capabilities is used to avoid introducing amplification errors. Since the MID barcode is a 10-base sequence, several variants of the barcode can be synthesized, thus each individual used in the SNP discovery experiment can be labeled with a specific MID barcode. Equimolar amounts of each individual PCR sample are pooled together and size selected (500–650 bp) via electrophoresis in preparation for 454-pyrosequencing. The size selection further reduces the genome complexity and the number of loci sequenced. Since all the samples are sequenced together, only a single em-PCR reaction is required and no space is lost to sub-gasketing of the 454 sequencing plate, thus maximizing the number of potential reads in the pyrosequencing run.

DNA Digestion

Four hundred and fifty ng of total genomic DNA of each DNA sample was separately double-digested with 3 U of the restriction enzymes *EcoRI* and *BfaI* (New England Biolabs, Beverly, MA) in 1X NEB4 restriction buffer. The DNA fragments were then ligated with 5'-TEG biotinylated/3'-phosphorylated *EcoRI* adapters and 3'-phosphorylated *BfaI* adapters (Table 2). Following ligation, small DNA fragments were excluded from the samples using Chroma Spin-400 columns (ClonTech, Mountain View, CA). The Chroma spin column removes ~90% of fragments less than 170 bp and 99% of fragments less than 100 bp. Lastly, the DNA fragments containing the biotin labeled *EcoRI* adapters were separated from the reaction using M-280 streptavidin beads (Invitrogen, Carlsbad, CA) according to the manufacturer's specifications, thus reducing the genomic complexity by removing the more abundant *BfaI*-*BfaI* fragments from the reaction mixture. The remaining *EcoRI*-*BfaI* and *EcoRI*-*EcoRI* fragments, with the biotin label still attached, were resuspended in 100 µL of TE.

PCR Amplification and MID Barcode Addition

Primers designed to be complementary to the *EcoRI* and *BfaI* adaptor+restriction sites and to carry a unique 5' MID barcode sequence were synthesized and HPLC purified by Integrated DNA Technologies (Iowa City, IA; Table 2). In total, four barcode+adaptor specific primers pairs were synthesized (i.e., MID1-*EcoRI* and MID1-*BfaI* represent one such pair). For each DNA sample, a single adaptor+barcode primer pair was used to amplify 1 µL of streptavidin-cleaned DNA fragments (biotin labeled *EcoRI*-*BfaI* and *EcoRI*-*EcoRI* fragments). Amplification of each sample was performed in 50 µL PCR reactions using 1X Advantage HF 2 PCR Master Mix (ClonTech, Mountain View, CA) and 0.2 µM of the MIDX-*EcoRI* and MIDX-*BfaI* primers. The Advantage HF 2 PCR Master Mix was chosen 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. The thermocycling profile was: 95°C for 1 min followed by 22 cycles of 95°C for 15 s, 65°C for 30 s, 68°C for 2 min. PCR products were visualized using a 1.2% Agarose FlashGel system (Lonza, Rockland, ME). Successful amplification appeared as a smear of fragments from 250 to 2000 bp.

Pyrosequencing

Sample concentrations for each of the four PCR reactions were measured fluorometrically using Quant-iT picogreen dye (Invitrogen, Carlsbad, CA) and pooled in equimolar amounts. The pooled PCR sample was electrophoresed in a 1.5% Metaphor agarose gel (Cambrex BioScience, East Rutherford, NJ), run in 0.5X TAE at 40V for 8 h and visualized using ethidium bromide staining. A single gel slice, representing DNA fragments ranging from 500 to 650 bp, was removed and the fragments extracted using a Qiaquick column (Qiagen, Germantown, MD). A single micro-bead sequencing run was performed as a service at the Brigham Young University DNASC (Provo, UT) using a Roche-454 GS FLX instrument and Titanium reagents (Branford, CT) without DNA fragmentation.

Assembly and SNP Detection and Validation

DNA reads were bioinformatically separated into MID barcode pools representing each mapping parent using the process-tagged sequences function in CLCBio Workbench (v. 3.6.1; Katrinebjerg, Aarhus N, Denmark). To generate the contigs for each of the four mapping populations, trimmed reads specific to each of the two parents of a specific population were de novo assembled using the Roche Newbler assembler (v. 2.0.01) with the minimum overlap length set to 50 bp and the minimum overlap identity set to 95%. Separate assemblies were constructed for each of the four mapping populations. SNPs specific to each of the mapping populations were identified within large contigs (>200 bp) using custom perl-scripts (Stajich et al., 2002). Putative SNPs were tagged if the following conditions were met: (i) coverage depth at the SNP was ≥10; (ii) the minor allele represented at least

Table 2. Adapter and primer sequences designed to be complementary to the *EcoRI* and *BfaI* adaptor+restriction sites and to carry a unique 5' MID-barcode code sequence.

Ligation adapters		
<i>EcoRI</i> adapter1 (F)	5'/5BioTEG/-CTCGTAGACTGCGTACC	
<i>EcoRI</i> adapter2 (R)	CATCTGACGCATGGTTAA/-5Phos/5'	
<i>BfaI</i> adapter1 (F)	5'-GACGATGAGTCTTGAG	
<i>BfaI</i> adapter2 (R)	ACTCAGGACTCAT/-5Phos/5'	
MID-barcode primer pairs	Sequence (5'→3') [†]	
	<i>EcoRI</i>	<i>BfaI</i>
MID1	ACGAGTGCCTGACTGCGTACCAATTC	ACGAGTGCCTGATGAGTCTTGAGTA
MID2	ACGCTCGACAGACTGCGTACCAATTC	ACGCTCGACAGATGAGTCTTGAGTA
MID3	AGACGCACTCGACTGCGTACCAATTC	AGACGCACTCGATGAGTCTTGAGTA
MID4	AGCACTGTAGGACTGCGTACCAATTC	AGCACTGTAGGATGAGTCTTGAGTA

[†]MID barcode sequences are bolded.

30% of the alleles observed; and (iii) 90% of the alleles from a specific MID barcode (i.e., mapping parent) were identical. Therefore, at the minimum read depth of 10, three of the sequence reads (30%) would need to be called as the minor allele variant, of which all would have to be derived from the same MID barcode pool.

The SNP validation was accomplished via Sanger resequencing. In brief, PCR was used to amplify 35 putative SNP loci (supplemental Table 1) followed by TA-cloning into a pGEM-T vector (Invitrogen, Carlsbad, CA) following the manufacturer's recommendation. Plasmid DNA from recombinant clones were isolated and sequenced bidirectionally using M13 primers (Forward: 5'-GTA AAA CGA CGG CCA GT; Reverse: 5'-CAG GAA ACA GCT ATG AC) and standard Sanger-based ABI Prism Taq dye terminator cycle sequencing methodology at the Brigham Young University DNA sequencing center.

Results

Genomic Reduction

Of primary importance to maximizing the number of SNP identified is the selection of the appropriate restriction endonucleases that will minimize the amount of repetitive DNA content in the target sequencing size range (500–650 bp), while providing sufficient genomic reduction to provide adequate sequencing coverage (≥ 10 -fold) for SNP detection. To determine which enzymes to use in the genomic reduction, we screened several enzyme combinations, specifically *EcoRI/MseI*, *EcoRI/NdeI*, and *EcoRI/BfaI* for the presence of repetitive DNA. The combination of *EcoRI/MseI* and *EcoRI/NdeI* both produced obvious bands within our target size range, while the *EcoRI/BfaI* produced a continuous smear of DNA, suggesting the absence of repetitive DNA fragments within the digest.

Assuming a genome size of 466 Mb and a 35% GC content estimate for the amaranth genome (Bennett and Smith, 1976; Maughan et al., 2008), we calculated that the frequency of recognition sites for a *BfaI* (CTAG) and *EcoRI* (GAATTC) to be 1.5 million and 159,219, respectively. The removal of the *BfaI-BfaI* end fragments combined with the selection of ~20% of the remaining fragment in the target size range (500–650 bp) reduces the complexity of the DNA nearly 50-fold, leaving approximately 10 MB of DNA for

sequencing. Using Titanium reagents, each 454-pyrosequencing run is capable of producing 1.3 million reads with an average read length of ~400 bp or slightly more than 500 MB. Thus, we estimated that a single 454-pyrosequencing run, with four-pooled MID-barcoded individuals would allow for 13X coverage across the 10 MB of DNA of each pooled individual. The GR-RSC protocol presented here allows for flexibility in the amount of the genome sampled, either through the selection of restriction endonucleases or by the size range of fragments sequenced, and therefore allows the researcher the ability to scale the sequencing run by the number of pooled DNA samples or by the number of loci (restriction fragments) to be sequenced.

Genetic Material

Prescreening of the parents of the four populations with 32 microsatellite markers, known to have high polymorphic information content (Mallory et al., 2008), provided an initial estimate of the level of polymorphism within the populations. Of the 32 microsatellites screened, 6, 47, 81 and 84% of the markers were polymorphic in Pop1-3, Pop1-4, Pop2-4, and Pop2-3, respectively (Table 3), suggesting that Pop2-4 and Pop2-3 are the most widely divergent of the populations and that Pop1-3 is potentially unsuitable for the development of a dense genetic map.

Pyrosequencing and Barcode Parsing

A total of 1,343,976 reads were obtained in the single run of the GS-FLX pyrosequencer producing 541,261,089 total bases of sequence with an average read length of 440 bp, with 84% of the bases having a quality score greater than 20. The percentage of bases called as "N" was 0.03%.

The CLCBio Workbench (v.2.6.1) separated the reads into four MID-barcode-specific pools, representing each mapping parent (where Ames15170, PI481125, PI553073, and PI642741 were respectively tagged with MID1-, MID2-, MID3-, and MID4-barcodes). To enter a pool, an exact match to all 10 bases on the MID barcode was required. A total of 1,272,089 (95%) of the reads were unambiguously sorted into four parental-specific barcode pools. Although we attempted to mix the samples in equimolar amounts before sequencing, thus expecting an equal distribution (25%) of the reads to each of the specific MID-barcode pools, our observed distribution of reads was 25.2% (320,759; MID1), 20.6% (262,467;

Table 3. Summary of contigs assembled and SNPs identified for each population.

	% Poly microsatellite	Large contig metrics [†]			SNPs [‡]	Unique contigs [§]	SNP base coverage	Minor allele freq. range		SNP type					
		Contigs	Bases	Avg. contig size				30–39%	40–50%	A/C	A/G	A/T	C/G	G/T	C/T
Pop1-3	6%	29,278	13,848,131	472	140	75	14	53.6%	46.4%	12%	31%	9%	5%	9%	33%
Pop1-4	47%	31,862	14,930,184	468	5433	3241	23	51.1%	48.9%	9%	33%	11%	8%	9%	31%
Pop2-4	81%	35,141	16,080,961	457	11,038	4585	22	53.0%	47.0%	9%	33%	11%	7%	8%	32%
Pop2-3	84%	33,524	15,348,505	457	11,047	4622	21	50.7%	49.3%	9%	32%	11%	7%	8%	32%
Average:		32,451	15,051,945	464	6915	3131	20	52.1%	47.9%	10%	32%	11%	7%	9%	32%

[†]Only contigs larger than 200 bp were included in the SNP discovery analysis.

[‡]SNP were called only if coverage at the base was ≥ 10 X, the frequency of the minor allele was ≥ 30 %, and ≥ 90 % of the alleles called within a parental line were identical.

[§]Number of contigs with at least one SNP.

MID2), 26.6% (338,228; MID3), and 27.6% (350,637; MID4) across the barcode pools, a discrepancy likely due to difficulties associated with fluorometric quantification of the PCR samples before pooling and/or inaccurate pipetting during the pooling process. Within a MID pool we also searched for reads that spanned the length of a total amplicon. In such cases, we observed the same MID on both sides of the DNA read. Since the average read length in the library was 440 bp, and our target sequencing size range was 500 to 650 bp, we did not anticipate that many of the reads would be full-length amplicons. Indeed, on average only 13.6% of the reads evidenced MID sequences on both sides of the read and, as expected, in all of these cases the same MID was present on both sides of the DNA fragment.

Sequence Assembly

Once reads were partitioned into four MID-barcode pools, the Newbler assembler (v. 2.0.01) was used to remove the MID barcode and adaptor sequences. Contigs, specific to each mapping population, were constructed using the Newbler de novo assembler, and only large contigs (≥ 200 bp; 93% of all contigs) were used in all subsequent analyses. The average contig size, across all assemblies, was 464 bp, with 96% of the bases with quality scores above 40. The average contig read depth across the four populations was 15X, while the average base coverage was 13X. Pop2-3 and Pop2-4, presumed to be the most divergent of the populations based on our preliminary microsatellite analysis, showed the highest number of large contigs (Table 3). Pop1-3, deemed to be the least divergent of the populations, showed the fewest number of large contigs assembled (29,278), while Pop1-4 showed an intermediate number (Table 3). We speculate that the increase in the number of aligned contigs in Pop2-4 and Pop2-3, relative to the other population assemblies, is due to fewer overlapping restriction fragments in their respective parental libraries and further supports the conclusion that these populations are the most divergent.

Population-Specific SNP Detection

To identify population-specific SNPs we analyzed each population assembly independently. In the analysis, SNPs were first identified using a minimum base cutoff coverage threshold ($\geq 10X$) and a minor allele frequency threshold ($\geq 30\%$; see SNP validation). While these thresholds are already conservative, we further filtered the SNPs on the basis of a $\geq 90\%$ within-parent uniformity threshold. In this parameter a threshold is set such that $\geq 90\%$ of the SNP alleles for a specific parental library must be identical. This parameter helps eliminate (i) erroneous SNPs introduced via the preparatory PCR steps; (ii) invalid SNPs caused by faulty assemblies; and (iii) SNPs that were heterozygous in the parental lines. The outcrossing rate in the grain amaranths has been estimated to range from 3.5 to 31% (Jain et al., 1982), thus heterozygosity is expected. The heterozygous SNPs, while potentially valid polymorphisms, are problematic for

downstream segregation analysis, especially if more than a single F_1 individual is used to develop the subsequent mapping population. Moreover, since amaranth is an ancient allotetraploid, paralogous regions of the genome may incorrectly align during the assembly process, even with the increased stringency parameters (minimum length overlap = 50 bp and minimum overlap identity = 95%) utilized in the assembly process.

Using the stated parameters, a dramatic range of SNPs was observed across the populations, ranging from a high of 11,047 SNPs in 4622 contigs in Pop2-3 to a low of only 140 SNPs in 75 contigs in Pop1-3 (Table 3), with an average of 6915 SNPs observed in 3131 contigs across all populations. The extremely low SNP rate in Pop1-3, also reflected in the microsatellite data, suggests that Ames 15170 and PI 553073 are closely related and that Pop1-3 is unsuitable for dense genetic mapping. We note that the available observation data for plant height, seed size, inflorescence type, and perisperm type in the Germplasm Resources Information Network (<http://www.ars-grin.gov/npgs/>) also suggest a potentially close genetic relationship between these two parents (Table 1).

The number of contigs that contained at least one SNP varied from a high of 4622 (14%) in Pop2-3 to a low of 75 (0.3%) in Pop1-3. Excluding Pop1-3, 12.8% of all contigs contained at least 1 SNP, with the largest class of contigs containing a single SNP (47%; Fig. 2). The average base coverage at an SNP across all populations was 20X (Fig. 3; Table 3), while the SNP density (SNP/bp) in the four populations was 1/98,915 bp, 1/2748 bp, 1/1457 bp, and 1/1389 bp in Pop1-3, Pop1-4, Pop2-4, and Pop2-3, respectively. Fifty-two percent of the SNPs identified fit within the 30 to 40% range for minor allele frequency (MAF), while the remaining 48% fit within the 40 to 50% range. Across all populations, transition mutations (A/G or C/T) were the most numerous, outnumbering transversions (A/T, C/A, G/C, G/T) by 1.8X margin (Table 3). These findings weren't unexpected as transition mutations, believed to be the result of hypermutability effects of CpG dinucleotide sites and deamination of methylated cytosines, are reported to be the most frequent SNPs in both plant and animal genomes (Morton et al., 2006; Zhang and Zhao, 2004). Across all four populations, we identified a total of 27,658 SNPs, of which 87% (24,162) were identified in just one population, while the remaining 13% were identified in at least one other population. The largest overlap of SNPs was between populations Pop2-4 and Pop2-3, where 1443 SNPs were shared in common among the 22,085 SNPs identified in these two populations. All SNPs, including 5' and 3' sequence information, have been deposited in dbSNP in GenBank. The SNPs are submitted under the handle MAUGHAN in batch number 2009A (GenBank: ss161123993 to ss161151650; build B131).

SNP Validation

Thirty-five SNPs, mapping to 30 contigs in Pop2-4, were chosen for validation via resequencing (Supplemental

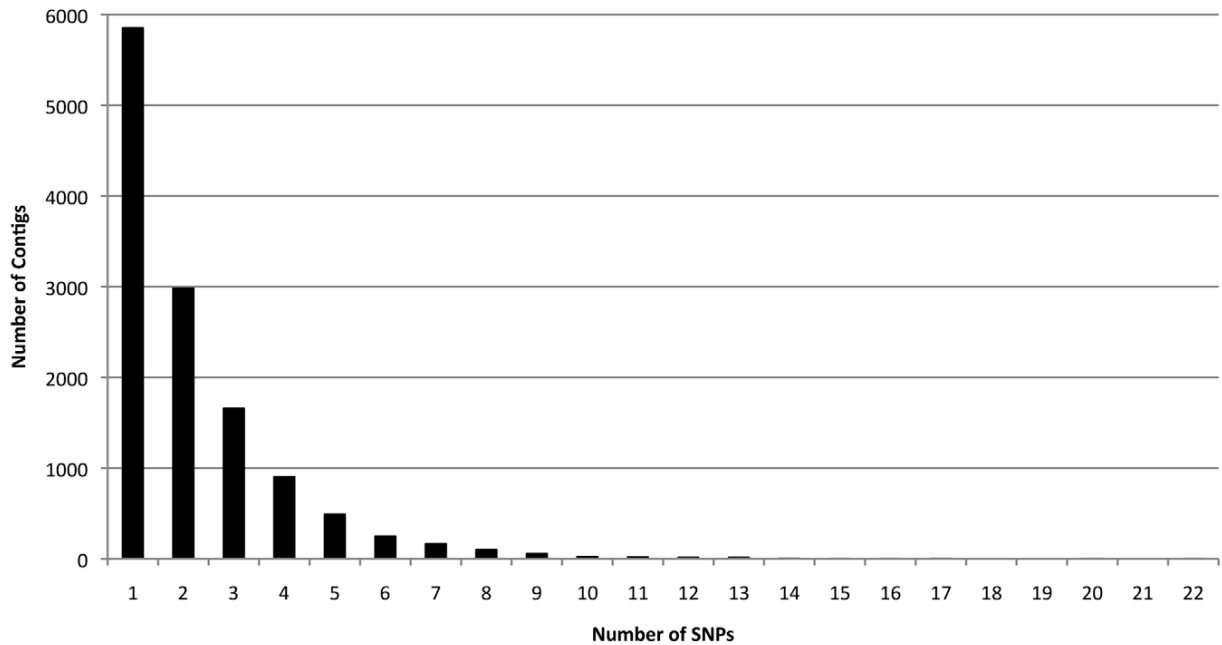


Figure 2. The distribution of contigs by number of SNPs.

Table 1). Of the targeted SNPs, 8 had MAF between 20 and 29%, 9 had frequencies between 30 and 39% and 19 had frequencies between 40 and 50%. The targeted SNPs ranged in base coverage from 9 to 57. The only SNP that failed to validate was derived from a contig consisting of nine reads and a MAF of 22% (2C to 7T). All of the remaining SNPs, 34 of 35 (97%), were validated as expected. Although seven of eight (87%) SNPs within the 20 to 29% MAF range were validated, we report here only those SNPs that were identified at a MAF $\geq 30\%$ and at a

base coverage of $\geq 10X$. Consequently, we note that we are likely under-reporting the true number of SNPs present in these populations. Indeed, because SNP detection is also a function of read depth, we anticipate that deeper sequencing of the genomic reduction libraries would also increase the number of SNPs identified.

SNP Distribution

Unfortunately, like many other under-researched species, little to no genome sequence information exists for

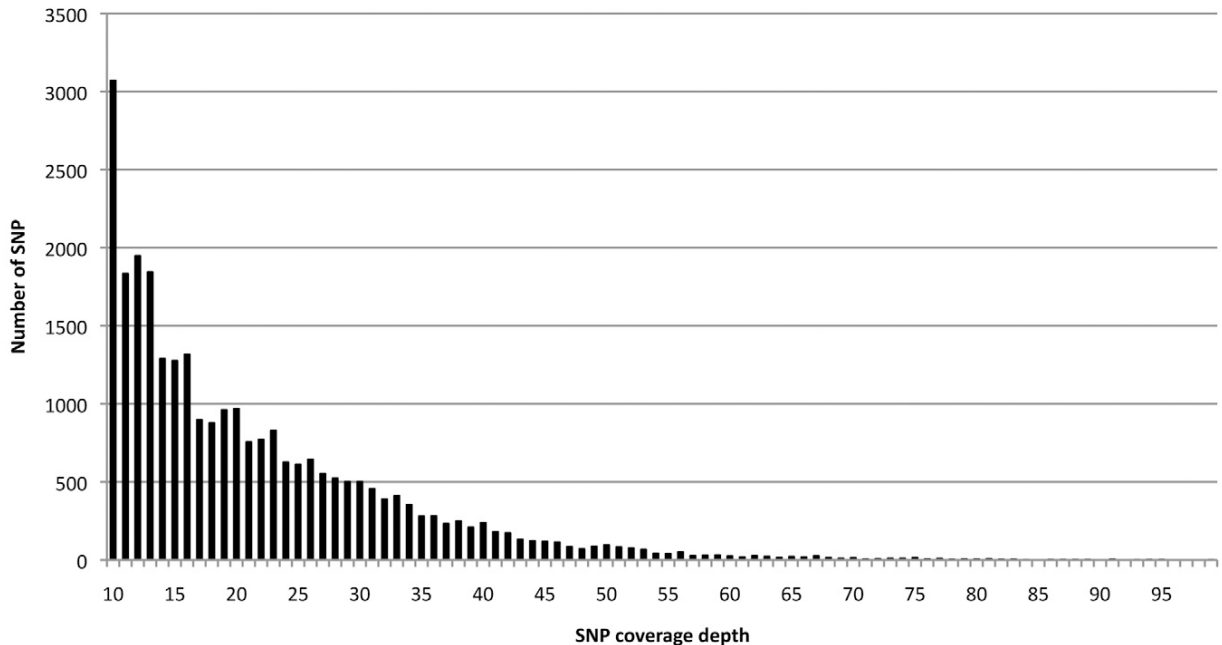


Figure 3. The distribution of the number of SNPs by coverage depth. The average depth of coverage for an SNP was 20X.

A. caudatus. The lack of a draft genome sequence prevents us from demonstrating the genomic distribution of SNPs identified with the GR-RSC method in amaranth. Thus, to demonstrate the genomic distribution of SNPs identified using genome reduction, we developed a GR-RSC library for *Arabidopsis*, ecotype Columbia. The ecotype Columbia is fully sequenced and publicly available in GenBank. To show the distribution of the GR-RSC restriction fragments across the *Arabidopsis* genome, we 454-pyrosequenced a total of 12,761 reads (363,618 bp) from the library and mapped them back to Columbia's five sequenced nuclear chromosomes (NC_003070, NC_003071, NC_003074, NC_003075, NC_003076) and two organellar genomes (NC_001284, NC000932) (Fig. 4). Eighty-five percent of the reads unambiguously mapped back to the reference chromosomes and/or organellar genomes at high stringency. Within the nuclear chromosomes, the greatest number of reads (2164) mapped to the largest chromosome (Chr1; 30.4 Mb) while the fewest (1506) mapped to the smallest chromosome (Chr4; 18.6 Mb), with an average of 1935 reads mapping to each chromosome. To determine the distribution of reads across individual chromosomes, we divided each chromosome into 100,000 bp segments and plotted a histogram of the number of reads mapping to each segment along the length of the chromosomes (Fig. 4). The distribution of reads was fairly uniform with an average of 6.3 reads per each 100,000 bp chromosomal segment. Indeed, of the 1198 chromosomal segments analyzed, only 43 did not contain at least one sequence read. Notable are two chromosomal segments that showed over-representation of read depth. The first was related to a telomeric region on chromosome 2 (reference position 3283–3596) while the second was related to a centromeric region on chromosome 3 (reference position 14,209,433–14,215,545; Fig. 4) and are thus likely related to repeat elements present at these sites.

An additional 1183 reads mapped to the organellar genomes, of which 1170 mapped to the chloroplast genome (154 Kbp), and 13 mapped to the mitochondrial genome (366 Kbp). The chloroplast reads mapped to 34 distinct chromosomal segments, while the mitochondrial reads mapped to nine segments. The read depth within the chloroplast was also significantly deeper, with an average 37.7 reads per contig, while the mitochondria averaged only 1.4 reads per contig. The reason for the discrepancy between the distribution of mapped reads and the read depth between the two organellar genomes is unclear but may, in part, be due to a difference in the number of *Bfa*I and *Eco*RI restriction sites and/or a difference in the initial abundance of each organelle in the photosynthetic leaf material used to generate the GR-RSC library. Etiolation of the plant before sampling or the use of non-photosynthetic tissue (i.e., root tissue) could potentially be used to minimize oversampling of the chloroplast genome (Hoover, 2006). The sequenced *Arabidopsis* chloroplast genome provides us with the ability to measure the sensitivity of the genomic

reduction. In silico digestion of the *Arabidopsis* chloroplast genome produces 299 *Bfa*I-*Bfa*I, 150 *Eco*RI-*Bfa*I, and 29 *Eco*RI-*Eco*RI DNA fragments, ranging in size from 7 to 2300 bp. All 35 chloroplast fragments that we sequenced were either *Eco*RI-*Bfa*I or *Eco*RI-*Eco*RI fragments, suggesting that the biotin-based genome reduction strategy was successful. Furthermore, as expected, the majority (77%) of the 35 fragments sequenced were *Eco*RI-*Bfa*I fragments. The last step in the GR-RSC protocol, before 454-pyrosequencing, is size selection of DNA fragments in the range of 500 to 650 bp via gel electrophoresis. We note that these fragments already have adapters and MID barcodes attached and thus are increased in length by 51 bp. Thus, the actual size of the in silico fragments being selected were between ~450 to 600 bp. Of the 179 *Eco*RI-*Bfa*I or *Eco*RI-*Eco*RI chloroplast DNA fragments remaining after the biotin-based reduction, 11 are in the 450 to 600 bp size range and all are accounted for in the sequencing. All of the remaining 24 chloroplast fragments that we sequenced are within the 300 to 700 bp size range, with fragments furthest from the center of the selected size range showing only minimal read-depth coverage (Fig. 5). The inclusion of the fragments outside, but close to the selected size range, is likely due to factors related to agarose gel electrophoresis resolution and imprecise cutting of the gel fragment for size selection.

Discussion

We have shown 454-based pyrosequencing coupled with genomic reduction to be a reliable method for mass SNPs discovery in amaranth. A genomic reduction protocol using deep sequencing was recently reported for the discovery of large numbers of SNP in cattle and pig (Van Tassel et al., 2008; Wiedmann et al., 2008). Unique to the protocol reported here was the use of genomic reduction in a plant genome with almost no prior sequences information and the use of the Roche-454 GS FLX instrument with Titanium reagents and MID barcodes.

The Titanium reagents produce longer sequence reads, often in excess of 400 bp, which are more likely to assemble correctly. The longer reads are also advantageous for subsequent SNP detection assays as they allow added flexibility in primer design. The incorporation of MID barcodes allowed for DNA sample pooling, read deconvolution, and the identification of SNPs for specific mapping populations. The use of MID barcodes significantly simplifies both the pyrosequencing step as well as downstream interpretation of the results. In the previous attempts at large-scale SNP discovery using genome reduction the researchers anonymously pooled genomic DNA from multiple individuals into a single sequencing reaction, thus allele calls for specific individuals within the pool could only be inferred with additional genotyping. Separating the individuals into specific regions of the pyrosequencing plate by gasketing could be used to assign reads to specific individuals (and thus avoid additional genotyping), but results in significant read-loss

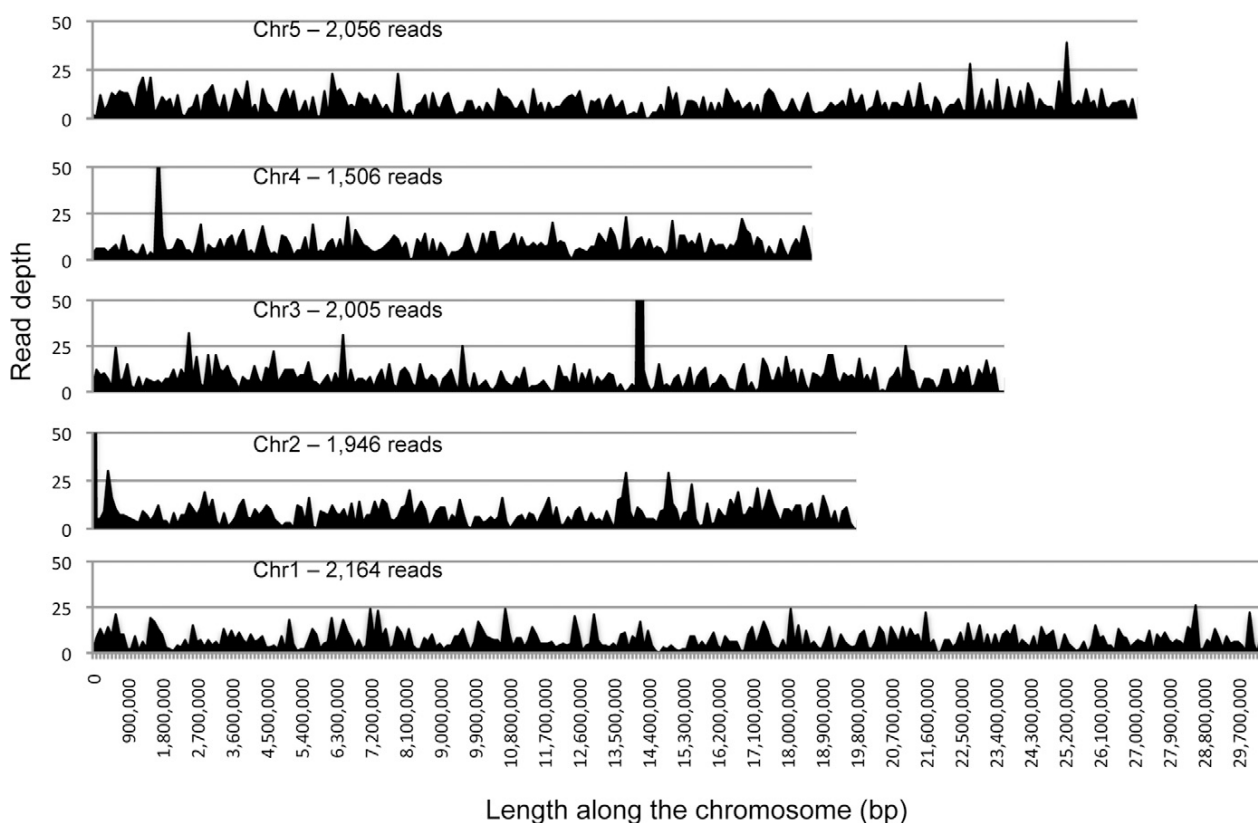


Figure 4. Read distribution across the *Arabidopsis* nuclear genome. An average of 6.3 reads per 100,000 bp chromosomal segment was observed.

due to the area of the plate occupied by the gasket. The amount of read-loss due to the gasket increases with the number of gasketed sub-regions. For example, a four sub-regions gasket reduces the overall number of plate reads by 300,000 reads (~120 Mb), while a 16 sub-region gasket reduces the number of reads by 660,000 reads (~264 Mb). Currently gaskets are only available for 2, 4, 8, and 16 sub-regions. Attaching a unique barcode to each sample before sequencing allows for pooling of the samples and negates the need for gasketing, thus maximizing the number of reads obtained in each 454-pyrosequencing run. We note that barcoding, as described here, does reduce the amount of target DNA sequenced since the barcode+adaptor (~26 bp) are part of the sequenced fragments. Assuming an average read length of 400 bp, this reduction represents a 6.5% decrease in the target sequence or approximately 16.9 Mb across a picotiter plate, regardless of the number of samples barcoded. Post sequencing, the samples are bioinformatically deconvoluted on the basis of their attached barcode sequence. The process is relatively simple and resulted in the accurate association of reads to specific DNA samples. Using MID barcoded samples, true allele frequencies can be calculated accurately for each SNP within the pool without additional genotyping and specific comparisons between individuals can be studied. Knowing the true allele frequency for a particular SNP has direct implications on

its utility, as SNPs with low allele frequencies have low polymorphic information content.

Although we only report the barcoding of four individuals, the application of GR-RSC and MID barcodes could easily be extended to larger pools of individuals, requiring only that new barcode primers be synthesized for each individual included in the pool. Besides their utility for SNP discovery, the use of MID barcodes and next-generation sequencing represent a new and potentially cost-effective alternative for large-scale genotyping (genotyping by sequencing). The current cost estimates for genotyping ranges from \$0.15 to \$0.25 per data point, depending on the genotyping volume and technology used. Genotyping by sequencing easily fits within that cost range, is scalable and does not require extensive robotics for sample preparation. The cost per data point is determined by the number of loci genotyped, the coverage depth required at each locus to assure accurate genotyping and the cost associated with the 454-pyrosequencing run. A rough estimate of genotyping by sequencing on a data point cost basis can be determined using the parameters generated in this proof-of-concept experiment. For example, Pop1-4 is segregating for 5433 SNP loci and our in-house 454-pyrosequencing costs are \$7200 per picotiter plate. The average SNP base coverage in our experiment with four individuals was nearly 20X—significantly more than what would be needed in

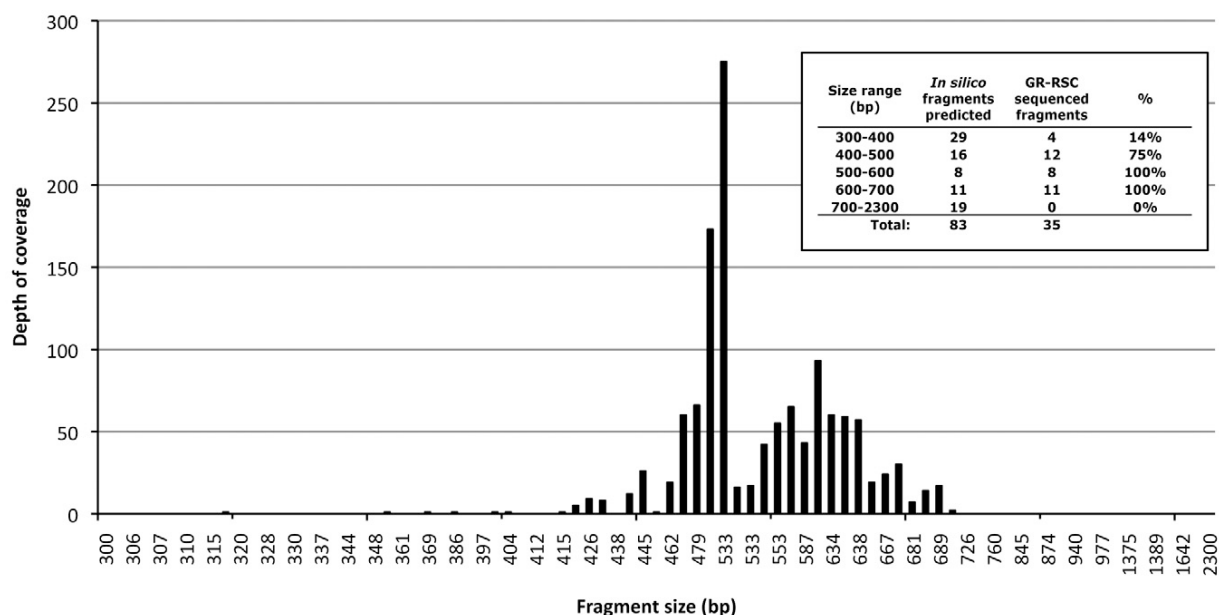


Figure 5. Read coverage and fragment size for 83 *Arabidopsis* *in silico* predicted chloroplast *EcoRI-EcoRI* and *EcoRI-Bfal* fragments. Only fragments that range in size from 300 to 2300 bp are shown.

a genotyping experiment using a recombinant inbred line or double haploid population (all individuals are homozygous). Thus, assuming that an average base coverage of 4X at all SNP loci for all individuals is sufficient for accurate genotype calling, we conservatively estimate that 20 individuals with an approximate genome size of amaranth can be genotyped simultaneously via sequencing with genomic reduction. Thus the number of data points generated is 108,660 (20 individuals times 5433 loci) and the cost per data point is \$0.07. If fewer loci are acceptable in the mapping experiment, the genomic reduction can easily be increased by decreasing the size of the gel cut or changing the restriction endonucleases. Increasing the genomic reduction will lead to fewer SNP loci being detected, but allow for increased numbers of individuals to be barcoded and pooled within the same picotiter plate without losing any base coverage at genotyped SNP loci.

Next-generation sequencers are accessible in most university DNA sequencing core facilities or from many commercial genomic service providers. Next-generation sequencing costs have dropped significantly over the last few years, even while the number of sequences produced per run has increased (Mardis, 2008). Any continued price reductions or sequencing volume increases will undoubtedly make genotyping by sequencing even more attractive. Genotyping of individuals from pooled, barcoded DNA samples has direct application for association studies, population genetics, genetic diversity analysis, as well as genetic linkage mapping and its variants (e.g., bulk segregant and near-isogenic line analysis). The genome reduction method described here is proof-of-concept for genotyping by sequencing from pooled

DNA samples, does not require a priori DNA sequence information, and should be readily adaptable to most species. What is lacking currently is the collection of the user-friendly bioinformatic tools needed for automated genotype calling for specific individuals from the voluminous data produced by next-generation sequencers.

Acknowledgments

This research was funded by the Ezra Taft Benson Agriculture and Food Institute. We gratefully acknowledge Dr. Edward Wilcox (BYU) for his assistance with 454-pyrosequencing and Alicia Barreda and Nathan Barney for their assistance with the microsatellite genotyping.

References

- Andrew, T., N. Maniatis, F. Carbonaro, S.H.M. Liew, W. Lau, T.D. Spector, and C.J. Hammond. 2008. Identification and replication of three novel myopia common susceptibility gene loci on chromosome 3q26 using linkage and linkage disequilibrium mapping. *PLoS Genet.* 4(10): e1000220. doi:10.1371/journal.pgen.1000220.
- 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.
- Becker, R., E.L. Wheeler, K. Lorenz, A.E. Stafford, O.K. Grosjean, A.A. Betchart, and R.M. Saunders. 1981. A compositional study of amaranth grain. *J. Food Sci.* 46:1175–1180.
- Bennett, M.D., and J.B. Smith. 1976. Nuclear DNA amounts in angiosperms. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 274:227–274.
- Brenner, D.M., D.D. Baltensperger, P.A. Kulakow, J.W. Lehmann, R.L. Myers, M.M. Slabbert, and B.B. Sleugh. 2000. Genetic resources and breeding of amaranthus. *Plant Breed. Rev.* 19:227–285.
- 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.
- Cramer, E.R.A., L. Stenzler, A.L. Talaba, C.A. Makarewicz, S.L. Vehrencamp, 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.

- 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.
- Hooper, J. 2006. The structure and function of plastids. Chloroplast development: Whence and whither. In R.R. Wise and J.K. Hooper (ed.) *Advances in photosynthesis and respiration*. Springer, Dordrecht.
- Iturbide, G.A., and M. Gispert. 1994. Grain amaranths (*Amaranthus* spp.). p. 93–101. In J.E. Hernandez-Bermejo and J. Leon (ed.) *Neglected crops: 1492 from a different perspective*. FAO, Rome, Italy.
- Jain, S.K., H. Hauptil, and K.R. Vaidya. 1982. Outcrossing rate in grain amaranths. *J Hered.* 73:71–72.
- Kadereit, G., T. Borsch, K. Welsing, and H. Freitag. 2003. Phylogeny of Amaranthaceae and Chenopodiaceae and the evolution of C4 photosynthesis. *Int. J. Plant Sci.* 164:959–986.
- 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.
- Krawczak, M. 1999. Informativity assessment for biallelic single nucleotide polymorphisms. *Electrophoresis* 20:1676–1681.
- Mallory, M.A., R.V. Hall, A.P. McNabb, D.B. Pratt, E.N. Jellen, and P.J. Maughan. 2008. Development and characterization of microsatellite markers for the grain amaranths. *Crop Sci.* 48:1098–1106.
- Mardis, E.R. 2008. The impact of next-generation sequencing technology on genetics. *Trends Genet.* 3:133–141.
- Maughan, P.J., N. Sisneros, M.Z. Luo, D. Kudrna, J.S.S. Ammiraju, and R.A. Wing. 2008. Construction of an *Amaranthus hypochondriacus* bacterial artificial chromosome library and genomic sequencing of herbicide target genes. *Crop Sci.* 48:S85–S94.
- 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.
- Pal, M., and T.N. Khoshoo. 1974. Grain amaranths. p. 129–137. In J.B. Hutchinson (ed.) *Evolutionary studies in world crops: Diversity and change in the Indian subcontinent*. Cambridge Univ. Press, Cambridge, UK.
- Rafalski, A. 2002. Applications of single nucleotide polymorphisms in crop genetics. *Curr. Opin. Plant Biol.* 5:94–100.
- 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.
- 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.
- Sauer, J.D. 1976. Grain amaranths. p. 4–7. In N.W. Simmonds (ed.) *Evolution of crop plants*. Longman Group, London, UK.
- Sauer, J.D. 1993. Amaranthaceae: Amaranth family. p. 9–14. In *Historical geography of crop plants: A select roster*. CRC Press, Boca Raton, FL.
- 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. Lehtvaslaiho, 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.
- 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.
- Van, K., E.Y. Hwang, M.Y. Kim, H.J. Park, S.H. Lee, and P.B. Cregan. 2005. Discovery of SNPs in soybean genotypes frequently used as the parents of mapping populations in the United States and Korea. *J. Hered.* 96:529–535.
- Van Tassel, 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.
- Vos, P., R. Hogers, M. Bleeker, M. Reijans, T.v.d. Lee, M. Hornes, A. Fritters, J. Pot, J. Paleman, M. Kuiper, and M. Zabeau. 1995. AFLP: A new technique for DNA fingerprinting. *Nucleic Acids Res.* 23:4407–4414.
- 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.