Single-Feature Polymorphism Mapping in Bread Wheat

Travis W. Banks,* Mark C. Jordan, and Daryl J. Somers

Abstract
Probe hybridization data from the Affymetrix GeneChip platform can be used to identify genetic polymorphisms. Developed from gene expression data, these single-feature polymorphism (SFP) markers are located in or are tightly associated with gene sequences. Using two different methods to identify SFPs, 1035 and 875 SFPs were mapped in a segregating population of 64 doubled haploid lines from the Triticum aestivum L. cross RL4452 × ‘AC Domain’. Statistical associations between the SFP maps and the rice (Oryza sativa L. ssp. japonica) genome were in agreement with known cereal syntenic blocks, and the mapping data were corroborated by previously established physical locations for wheat expressed sequence tags. This approach allowed the rapid identification of markers from the genic regions of the complex hexaploid wheat genome.

DNA-based molecular markers are an integral part of modern wheat-breeding programs. There is a long history of co-mapping phenotypes with markers (quantitative trait loci analysis), resulting in important agronomic traits being positioned on genetic maps. For efficient use in marker-assisted selection (MAS) or for map-based cloning of genes corresponding to the phenotype in large complex genomes such as wheat (Triticum aestivum L.), there must be high marker density in the region of interest so that markers closely associated with the trait can be obtained. Although a current wheat consensus map based on simple sequence repeat (SSR) markers contains 1235 markers (Somers et al., 2004), the marker density is not suitable for map-based cloning, MAS of complex traits, or anchoring physical and genetic maps without finding additional markers for fine mapping. Paux et al. (2008) developed a physical map of the 1-gigabase wheat chromosome 3B by anchoring contigs with 1443 markers. A method of rapidly adding many more markers to the wheat maps would accelerate the adoption of MAS for economically important traits in wheat and facilitate map-based cloning and genome sequencing efforts.

High-throughput methodologies are revolutionizing plant genomics and molecular breeding. The Affymetrix GeneChip Wheat Genome Array (Affymetrix Inc., Santa Clara, CA) is a microarray platform used to quantify the expression of thousands of genes simultaneously. The “chip” contains 61,127 probe sets representing 55,052 consensus sequences (unigenes) derived from the...
assembly of all expressed sequence tags (ESTs) available at the time of chip design from *Triticum aestivum* and closely related species. Each probe set contains 11 perfect match (PM) and 11 mismatch (MM) oligonucleotide probes of 25 bases in length, which collectively represent a specific unigene sequence. Gene expression values are generated when complementary RNA is hybridized to the array and probe intensity values from a probe set are combined into a single number. Originally designed for gene expression analysis, the Affymetrix expression microarray was quickly adapted for genetic mapping of genes based on gene expression values (Borevitz et al., 2003; Winzeler et al., 1998). Expression values have also been regarded as phenotypes and mapped as expression quantitative trait loci, allowing the identification of regulatory “hot spots” (Doerge, 2002; Gibson and Weir, 2005; Jordan et al., 2007; West et al., 2007).

Single-feature polymorphisms (SFPs) are a type of genetic marker generated from Affymetrix GeneChip microarrays. Based on differential hybridization intensities of individual probes, as opposed to the whole probe sets, SFPs offer a method to identify polymorphism based on DNA sequence differences and not expression differences. First presented in yeast (*Saccharomyces cerevisiae*), genomic DNA was hybridized to an Affymetrix expression array and 3714 biallelic SFP markers were identified between two isolates (Winzeler et al., 1998). In plants, this technique was first applied to *Arabidopsis thaliana* L. in which genomic DNA was hybridized to an *A. thaliana* Affymetrix expression array (Borevitz et al., 2003). Despite a genome size of 120 megabases (10x that of yeast), Borevitz et al. were able to identify 3806 SFPs between the accessions Columbia and Landsberg erecta. Taking advantage of the sequenced and assembled genome, they were also able to identify 105 genomic deletions covering 111 genes between the two strains by examining how the SFP markers from the same probe set clustered. It was also possible to extend microarray genotyping to bulk segregant analysis, which enabled the mapping of developmental mutations (Borevitz et al., 2003; Hazen et al., 2005). This approach was later expanded to allow mapping of quantitative or non-Mendelian traits with SFPs in a technique named eXtreme Array Mapping (Borevitz, 2006; Werner et al., 2005; Wolyń et al., 2004). In a more recent study, Borevitz et al. (2007) identified 77,420 SFPs among 23 accessions of *A. thaliana* with a false discovery rate (FDR) <1% (Borevitz et al., 2007). Hybridization of genomic DNA to an array to identify SFPs was also undertaken in rice where tens of thousands of SFPs were identified but at an FDR approaching 10% (Kumar et al., 2007). Detection of sequence polymorphism in large genomes using microarrays requires a reduction in complexity through polymerase chain reaction or an RNA surrogate (Kennedy et al., 2003; Rostoks et al., 2005). To identify SFPs in barley (*Hordeum vulgare* L.), a genome of 5500 megabases (http://data.kew.org/cvalues/ [verified 3 May 2009]) (Bennett and Smith, 1976), RNA from cultivars Morex and Golden Promise was collected from various tissues and time points and hybridized to 36 Affymetrix Barley GeneChips (Rostoks et al., 2005). To account for variability in the hybridization signal attributed to expression and not probe affinity between the two barley genotypes, Rostoks et al. (2005) utilized the high number of chip replications to develop a linear model that accounted for tissue, genotype, probe, and genotype × tissue effects. The residuals were fitted for a genotype effect and used to identify 10,504 SFPs in 3734 probe sets. One hundred eighteen SFPs have also been identified between two near-isogenic wheat genotypes using a large number of replicates (24 per genotype) and a novel linear model approach (Coram et al., 2008).

A drawback to the above-mentioned approaches to SFP detection is the high level of replication that is required. Another method was developed that modeled probe hybridization signal as a linear combination of probe set and probe affinity effects and used robustified projection pursuit to identify genotype-dependent, expression-adjusted probes responsible for the polymorphism. This method required fewer biological replicates (as few as three) and has been used to identify 2007 barley SFPs (Cui et al., 2005), 1058 cowpea (*Vigna unguiculata* (L.) Walp) SFPs (Das et al., 2008), and to genotype two barley cultivars thought to differ at only a single locus but that were found to contain three polymorphic haplotype blocks (Walia et al., 2007).

West et al. (2006) developed a method to identify SFPs using members of a recombinant inbred line (RIL) mapping population. RNA from 148 *A. thaliana* RILs were hybridized to the ATH1 Affymetrix GeneChip and a metric of binding affinity that is independent of expression level, called the SFPDev value, was calculated (intensity value for probe X – average intensity for remaining 10 probes in the probe set/intensity value for probe X). In cases of a polymorphism between the parental lines, the SFPDev values for a probe from all of the RILs would form a bimodal distribution, each peak representing a parental genotype. They identified SFPs in 968 genes and constructed a genetic map using 599 SFPs. Five hundred ninety-five SFPs were correctly positioned relative to the known genomic sequence. The remaining four were believed to be indicative of rearrangements that have occurred between a parental line and the sequenced accession.

A doubled haploid (DH) barley population was used to identify ~4000 SFPs and evaluate two established and two novel methods of determining SFPs (Luo et al., 2007). One method they presented was unique compared with others in that it utilized the hybridization intensity values of the MM probe in addition to the PM. Luo et al. (2007) derived a z-discrimination function that was independent of gene expression level and independent of other probes in the probe set. The researchers used cluster analysis to partition the parental genotypes and used the mean and variance of those clusters to genotype their mapping population. A t statistic was calculated between
the two groups of genotyped lines to gauge the success of the genotyping.

We applied two genotyping discrimination functions from West et al. (2006) and Luo et al. (2007) to identify SFPs in a well-characterized cross of two Canadian bread wheats (McCartney et al., 2005, 2006). More than 1500 SFPs, each representing a unigene, were placed on genetic maps using 64 DH individuals.

**Materials and Methods**

**Plant Material and RNA Collection**

Sixty-four DH lines from the *Triticum aestivum* cross RL4452 × ‘AC Domain’ described in McCartney et al. (2005), along with parental lines, were grown in a replicated field trial at one location. RNA sampling for microarray hybridization was as described in Jordan et al. (2007). Briefly, as the plants approached heading they were observed daily until heads with 50% protruding anthers were visible. Every day during the flowering period, heads with visible anthers were tagged and dated. Heads tagged on the same day were collected from each row 5 d after anthesis. Developing seeds were excised in the field from collected heads, placed into a sample tube, and immediately frozen in liquid nitrogen. Each was labeled with genotype and replicate number. Tubes were stored at −80°C until used for RNA isolation.

**RNA Isolation and Microarray Hybridization**

Each of two biological replicates for every DH line, along with three replicates of each parent, was hybridized to arrays for a total of 134 chips. Isolation of RNA, labeling, and hybridization to the Affymetrix GeneChip Wheat Genome Array were as described in Jordan et al. (2007).

**Microarray Data Processing**

The 134 Affymetrix CEL files containing the raw hybridization values for 64 field-replicated DHs and 3 replicates of each parent were background corrected and normalized using robust multiarray averaging and quantile normalization from the affy Bioconductor R-package (Gentleman et al., 2004). Background correction and normalization were applied to both the PM probes and the MM probes. Correlation coefficients between the replicates were calculated using the adjusted PM probe values, and all replicate and parental pairs met the threshold value of 0.95. We identified SFPs using two different measures of discrimination: one with the SFPDev value from West et al. (2006) and the other with a z function \(|(PM_{Xij} - MM_{Xij})/PM_{Xij}|\) from Luo et al. (2007). We refer to the first method as our SFPDev method and the latter as our z-function method. Other than the value used to discriminate genotype, the two pipelines for SFP identification were as follows:

1. For each adjusted probe in each of the 134 CEL files, the discrimination function was applied.
2. The sample values were placed into two groups by two-means clustering. If the ratio of cluster sizes was greater than 1:2, the probe was rejected. This represents our threshold for segregation distortion.
3. All of the samples from parent AC Domain had to be in one cluster and all values from RL442 in the other or else the probe was rejected.
4. Assuming the two clusters were normally distributed but without assuming equal variance, the means of the two clusters were tested for equality using a t test at a 0.05 significance level. If the cluster means were found not to be different, then the probe was rejected.
5. The probability of each cluster member belonging in the other cluster was calculated using a normal deviate as described in Luo et al. (2007). We applied a z score of 1.96. Any discriminant values in a cluster with a probability <95% of not belonging to the other cluster were regarded as missing data.
6. If >15% of the values were missing data, then the probe was rejected, else alleles were assigned based on which parent a sample clustered with.
7. We compared the two replicates of each DH line across all probes. In cases in which there was disagreement between the replicates, a missing data value was assigned.

**Genetic Mapping**

We combined the SFP markers with 293 microsatellites previously mapped to this population (McCartney et al., 2005). In cases in which a probe set, which represents a unigene, had more than one SFP we took the majority genotype scores for the individual. If the allele calls from the SFPs in a probe set did not have agreement for an individual, the consensus allele score was changed to missing data. The SFP maps were constructed using JoinMap V4.0 (Kyazma B.V., Wageningen, Netherlands; www.kyazma.nl [verified 3 May 2009]). The final maps were displayed in Kosambi centimorgan units.

**Wheat–Rice Synteny**

The consensus sequence for each probe set was taken from HarvEST (version 1.54 Wheat Chip; www.harvest.ucr.edu [verified 3 May 2009]) (Close et al., 2007) and compared using BLAST software (Altschul et al., 1997) to the predicted rice peptide sequence (Release 5), downloaded from the Rice Genome Annotation project Web site (http://rice.plantbiology.msu.edu [verified 3 May 2009]). The best rice peptide hit with an E-value ≤1 × 10⁻³ was considered an ortholog to the wheat sequence and the chromosomal location of the protein noted. For each wheat chromosome from a genetic map, we applied a sliding window approach to calculate the P value for a binomial distribution of orthologs to rice sequences. Starting at the top of the linkage group, a window of 15 SFP markers that found a rice ortholog was established. The distribution of the rice orthologs among the 12 rice chromosomes was determined and a binomial distribution P value (1:11) was calculated for each chromosome, with the null hypothesis being that orthologs from each rice chromosome are proportionally represented in the
window. A window was said to have a statistical association with a rice chromosome if the $P$ value for the distribution test was ≤ 0.05. This window was kept at a width of 15 orthologs and was moved by one ortholog at a time down the length of the linkage group. A $P$ value for each rice chromosome was calculated at each increment.

**Results**

**SFPDev Method**

Using the SFPDev method, 1925 probes from 1135 probe sets were categorized as polymorphic from our analysis pipeline. The distribution of the number of polymorphic probes within probe sets varied from 64% having a single SFP to one probe set with 10 SFPs, with an average of 1.7 (Fig. 1a). Assuming that SFP-containing probes from the same probe set were detecting polymorphism(s) at the same loci, for the purpose of mapping a majority score was used to assign genotypes to the individuals, which effectively collapsed all of the SFPs from a probe set into a single representative. There were four possible cases for determining the consensus genotype: i) all of the genotype scores for an individual across the probes in a probe set were the same; ii) the majority of individuals belonged to one genotype, resulting in that genotype being assigned to the consensus; iii) all of the individuals were scored as missing data, in which case a score of “missing” was given; and iv) both genotype scores were equally represented. In the final case, because no majority could be established, a score of “missing” was assigned to that individual (Supplementary Fig. 1). We used the disagreement in genotype assignment among probes from the same probe sets to estimate an error rate for this method of SFP identification. There were 67,873 data points that were involved in a comparison and 1110 were in disagreement with the majority value, giving an error rate of 1.7% for genotype assignment.

Majority scores for 1135 probe sets and 293 SSRs from 64 individuals were used to construct a genetic map (Fig. 2a; Supplementary Fig. 2). The map contained 1066 SFP markers and spanned 2652 cM, for an average of 2.49 cM. The remaining 69 SFPs could not be assigned a location on the genetic map with our mapping software. The map contained single gaps on chromosomes 1B, 3D, 4D, 5D, 6A, and 7B and two gaps on chromosome 1D. Probe sets were mapped to each of the 21 wheat chromosomes, with the majority being placed on the B genome (Table 1). The D genome was the least well represented in our data, having <18% of the markers. Group 4 had only 8.4% of the probe sets, followed by Group 6 at 9.5%. This is in contrast to Group 2, which had >21% of the probe sets mapped to it. Chromosome 6D had the fewest number of markers, with seven SFP-containing probe sets, followed by 5A with 12.

We identified 165 mapped probe sets that were generated from Affymetrix consensus sequences derived from assembled ESTs, of which at least one had been previously physically mapped to a wheat chromosome deletion bin (Qi et al., 2004). One hundred nine of those were genetically mapped to the same chromosome as they were physically mapped, and 25 genetically mapped to a homoeologous chromosome. Of the remaining 31 probe sets, 13 were genetically mapped to the same chromosome using the $z$-function SFP identification method.

To examine macro-synteny between our map of wheat and rice, we calculated the $P$ value for a binomial distribution test in a sliding window along the length of each chromosome (Supplementary Fig. 3). Figure 3a illustrates this analysis for wheat 2B. The top of the chromosome had the first eight windows with an association to chromosome 4 of rice ($P \leq 0.05$). Starting at Window 8 up to and including Window 28, an association with rice chromosome 7 was detected. Continuing from Window 28 to the end of the chromosome, there was an association with rice chromosome 4. Each chromosome except 4D and 6D contained regions with statistical associations ($P \leq 0.05$) to at least one chromosome of rice. These associated regions were in agreement with wheat–rice chromosomal relationships found by others (La Rota and Sorrells, 2004; Salse et al., 2008; Sorrells et al., 2003).

**$z$-Function Method**

Using the $z$-function method, 1501 probes from 921 probe sets were categorized as polymorphic. The number of SFP-containing probes per probe set varied from 65% having one SFP-containing probe to a single probe set having 10 (Fig. 1b). There was an average of 1.6 SFPs per probe set. As was done for the SFPDev method, a majority genotype was determined for each probe set containing multiple SFP probes, and we calculated an error rate for genotype assignment based on differences between probes in a probe set. The 51,032 genotype assignments were compared, yielding 663 disagreements giving an error rate of 1.30%.

The 921 probe set scores along with 293 SSRs for 64 individuals were used to construct a genetic map with 21 linkage groups with a size of 2567 cM (Fig. 2b; Supplementary Fig. 4). Of the 921 probe sets, 875 were placed on the map at an average distance of one SFP per 2.93 cM. The remaining 46 could not be assigned a location on the genetic map with our mapping software. The map contained nine single gaps on chromosomes 1A, 1D, 3B, 3D, 4D, 5A, 6A, 7B, and 7D. The distribution of markers was similar to that observed with the SFPDev method linkage map: the A, B, and D genomes had 36.8, 46.7, and 16.5% of the markers, respectively (Table 1). Group 4 had the fewest markers at 10.4% and Group 2 had the most at 19.3%. Chromosome 5A had the fewest mapped markers with only three, followed by chromosome 4D with six.

We identified 119 mapped probe sets composed of deletion-bin mapped ESTs (Qi et al., 2004). Eighty-five probe sets were located on the same chromosome on both our genetic map and the deletion-bin map and 11 probe sets were mapped to homoeologous chromosomes. From the remaining 23, 14 were genetically mapped to the same chromosome using the SFPDev method.
Figure 1. Histogram showing the distribution of unigenes based on the number of probes within the probe set found to contain a single-feature polymorphism (SFP). (a) Using the SFPDev method, 1925 SFPs representing 1135 probe sets were identified. [b] Using the z-function method, 1501 SFPs representing 921 probe sets were identified.
Figure 2. Genetic maps generated from single-feature polymorphism (SFP) and simple sequence repeat (SSR) markers. The SFP markers are shown in black if they were unique to an SFP identification method or in blue if the marker occurred in both maps generated from the two methods. The SSRs are shown in red, and gaps in the linkage map are indicated in green. Twenty centimorgans was added to the map when joining the linkage group fragments within a chromosome. (a) Genetic map of the Group 1 chromosomes generated using the SFP-Dev method of SFP discovery. (b) Genetic map of the Group 1 chromosomes generated using the z-function method of SFP discovery.
High levels of wheat–rice colinearity were observed on a number of chromosomes, with only 4D and 5A having no significant associations \((P \leq 0.05)\) with at least one rice chromosome (Supplementary Fig. 5). Figure 2b illustrates this analysis for wheat 2B. The top of the chromosome had the first five windows with an association to chromosome 4 of rice \((P \leq 0.05)\). Starting at Window 5 up to and including Window 18, an association with rice chromosome 7 was detected. Continuing from Window 15 to the end of the chromosome, there was an association with rice chromosome 4. These shared evolutionary blocks were in agreement with the overall structure reported by others (La Rota and Sorrells, 2004; Salse et al., 2003).

**Differences in Allele Assignments**

The two methods of identifying SFPs found 555 polymorphic probes in common, representing 435 probe sets that were assigned a similar genetic location on both maps. Comparison of allele calls for the same probe set between the two methods found that 211 had zero differences (Fig. 4). Over 94% of the probe sets had three or fewer allele calls different across all 64 mapping individuals. Seven probe sets had 10 or more differences between the SFPDev and \(z\)-function methods. Two were mapped to the same chromosomes by both methods and five were placed on different chromosomes. Of the five mapped to different locations, three were placed on homoeologous groups (Ta.11011.1.S1_at, Ta.4256.1.S1_at, and TaAffx.128822.1.S1). The remaining two sequences, which mapped to different chromosomes using the two methods, were Ta.5138.2.S1_a_at and TaAffx.64027.1.A1_at. Ta.5138.2.S1_a_at appears to have homology to a family of peroxidase precursor proteins from rice, and TaAffx.64027.1.A1_at has no rice orthogs. A fifth probe set, Ta.1623.2.S1_x_at, was mapped to homoeologous chromosomes but had nine differences in allele assignments between the two methods.

**Discussion**

RNA from 64 progeny and parents from a bread wheat mapping population, grown under field conditions, was collected and hybridized to Affymetrix Wheat GeneChips. The SFP determinant values developed by West et al. (2006) and Luo et al. (2007) combined with customized filtration criteria were used to develop methodologies that identified 1925 SFPs representing 1135 unigenes and 1501 SFPs representing 921 unigenes, respectively. Genetic linkage maps were developed from the two sets of unigenes in which 1066 and 875 SFP markers were assigned locations. The SFPs were placed on each of the wheat chromosomes, with the B-genome having the largest number of markers and the D-genome having the fewest, which is consistent with known geno-type diversity ranking (Huang et al., 2002). The lines of the RL4452 x AC Domain population used for analysis were selected so as to deliberately fix as homozygous

<table>
<thead>
<tr>
<th>Group</th>
<th>A</th>
<th>B</th>
<th>D</th>
<th>Total</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>SFPDev method</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>51</td>
<td>79</td>
<td>33</td>
<td>163</td>
<td>15.3</td>
</tr>
<tr>
<td>2</td>
<td>41</td>
<td>105</td>
<td>84</td>
<td>230</td>
<td>21.6</td>
</tr>
<tr>
<td>3</td>
<td>65</td>
<td>81</td>
<td>31</td>
<td>177</td>
<td>16.6</td>
</tr>
<tr>
<td>4</td>
<td>49</td>
<td>31</td>
<td>10</td>
<td>90</td>
<td>8.4</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>101</td>
<td>10</td>
<td>123</td>
<td>11.5</td>
</tr>
<tr>
<td>6</td>
<td>68</td>
<td>26</td>
<td>7</td>
<td>101</td>
<td>9.5</td>
</tr>
<tr>
<td>7</td>
<td>112</td>
<td>56</td>
<td>14</td>
<td>182</td>
<td>17.1</td>
</tr>
<tr>
<td>Total</td>
<td>398</td>
<td>479</td>
<td>189</td>
<td>1066</td>
<td></td>
</tr>
<tr>
<td>%</td>
<td>37.3</td>
<td>44.9</td>
<td>17.7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| z-Function method |  |  |  |  |  |
| 1 | 40 | 72 | 12 | 124 | 14.2 |
| 2 | 31 | 75 | 63 | 169 | 19.3 |
| 3 | 67 | 61 | 28 | 156 | 17.8 |
| 4 | 55 | 30 | 6 | 91 | 10.4 |
| 5 | 3 | 79 | 12 | 94 | 10.7 |
| 6 | 43 | 45 | 11 | 99 | 11.3 |
| 7 | 83 | 47 | 12 | 142 | 16.2 |
| Total | 322 | 409 | 144 | 875 |  |
| % | 36.8 | 46.7 | 16.5 |  |  |

Rht-D1, a major plant-height gene located on chromosome 4D, and could account for the lower number of SFPs found on this chromosome (Table 1). Chromosome 5A was also not as well populated with SFP markers as other chromosomes in the two genetic maps. The reason for this lack of identified diversity is unknown, although McCartney et al. (2005) found a similar lack of diversity using SSR markers in the same cross.

Identification of SFPs in this study required the detection of variation in probe binding between genotypes in a manner that was independent of gene expression. The first method presented in this report uses the SFPDev value (West et al., 2006), which achieves expression-level independence by calculating how much the hybridization value of a probe deviates from the average of the other probes in the probe set and normalizes that difference by the probe’s own hybridization value. The expectation is that changes in gene expression will result in changes in hybridization signal to all probes in a probe set and that the ratio of signal values remains constant. West et al. (2006) demonstrated that this
calculation yielded a reproducible representative value and resulted in identifying 1257 SFPs across 148 RILs. Another strength of the SFPDev approach that we drew on was capitalizing on the high level of genetic replication per locus available by genotyping the progeny of two homozygous parental lines. Each SFPDev value that was calculated was, in essence, the sampling of one of the parental alleles. With a segregation distortion cutoff of 1:2, an allele was measured at least 42 times (64 lines × 2 replicates × 33%) and from this an SFPDev value distribution for the two parental alleles could be established. We deviated from the methodology presented by West et al. (2006) in how the two clusters of SFPDev values were identified. The “iterative slice” method used
by these researchers was not effective with our data, and application of the Perl program provided by the authors to our data resulted in no SFPs being found (data not shown). The plants used to generate the West et al. (2006) microarray data were grown in growth chambers, and it is likely the consistent environmental conditions would have resulted in uniform expression profiles and, therefore, SFPDev values that clustered tightly. Our data were generated from plants grown under field conditions where the environmental conditions would have introduced more variable expression between biological replicates and between individuals sharing the same parental allele. We found that the iterative slice approach could not properly partition our data and so we used two-means clustering to group the values into two distributions.

The second approach we applied to identify SFPs was the discriminant $z_{Xij}$ value developed by Luo et al. (2007), where $z_{Xij} = (\text{PM}_{Xij} - \text{MM}_{Xij})/\text{PM}_{Xij}$. Unlike the SFPDev value, $z_{Xij}$ is calculated without consideration to the other probes in a probe set and uses the hybridization value of the MM probe. Luo et al. (2007) used two-means clustering to group the parental observations and assigned alleles to the progeny using the mean and variance of the replicate parental scores, while we used all of the values in a cluster to determine an estimate for the true parental determinant values for the alleles and assigned genotypes.

As part of the common pipeline to identify SFPs, $t$ tests were performed on clusters of permuted discriminant values at a significance level of 0.05. Because of the large number of tests performed, 13,434 for the SFPDev method and 12,755 for the $z$-function method, we would expect 672 and 638 false positives, respectively. However, of the 13,434 $P$ values calculated for the SFPDev clusters, only 10 had $P > 0.0001$ and 11 $z$-function clusters had $P > 0.0001$. These low $P$ values suggest that few false positives were passed to the next stage of the pipeline. In addition, the pipeline stage of generating consensus genotypes resulted in very few probe sets, which deviated from the 2:1 segregation ratio established earlier in the processing. For the SFPDev method, 43 of the 1066 mapped markers had segregation ratios beyond 1:2, 39 had ratios between 1:2 and 1:3, three between 1:3 and 1:4, and a single marker above this. Similar results were observed with $z$-function–derived markers: 31 had ratios between 1:2 and 1:3, and two markers with ratios between 1:3 and 1:4. These skewed ratios are not a reflection on the mapping population but are an artifact of attempting to reduce errors in genotyping calls. The generation of consensus calls also affected the level of missing data in the markers. This is particularly true when a probe set had two polymorphic probes. Because a majority could not be determined from a sample of two differing probes, both calls were changed to missing data, potentially elevating a consensus marker beyond the earlier established threshold. The effects of this were minimal, as only six SFPDev-derived and three $z$-function–derived markers fell outside of this threshold (Supplementary Fig. 1).

The polyploid nature of the bread wheat genome means that it is possible for at least three different homoeologous loci contribute to the expression of each gene and that the hybridization signal measured can represent the binding of more than one homoeologous transcript. There was no evidence of probes from the same probe set detecting polymorphism at different loci. The two SFP discovery methods found 555 probe sets in common, 435 of which were placed on both genetic maps and six of which mapped to different chromosomes. Four of those mapped to syntenic regions of homoeologous groups, indicating that the two methods were likely detecting SFPs in two transcripts from homoeologous chromosomes. Ta.5138.2.S1_a_at, which mapped to nonhomoeologous chromosomes, appears to be a member of a large peroxidase family in rice, and it

![Figure 4. Genotyping differences between single-feature polymorphism (SFP) discovery methods. The two methods used to identify polymorphisms in the doubled haploid (DH) population identified 435 SFPs in common between the two linkage maps generated. There was not perfect agreement in genotype assignment for each DH. The histogram illustrates the distribution of differing genotype assignments in each of the 435 shared SFPs. Two hundred eleven SFPs had perfect agreement in genotypes between the two methods, and ~94% of the SFPs had fewer than four differences across all 64 DHs.](image-url)
is possible that the transcripts from those two locations were binding probes from the same probe set.

West et al. (2006) were able to use the *A. thaliana* genomic sequence to conclude that the ordering of genes present in their map was correct and they also identified potential regions of translocation and inversion between two ecotypes. Luo et al. (2007) created a genetic map from the SFPs they discovered, though they did not evaluate the quality of the map but were able to estimate the efficacy of their detection methods using EST sequences. Without a complete genomic sequence for *Triticum aestivum*, we are unable to determine to what extent the ordering of probe sets, which represents expressed genic content, is correct. To overcome this limitation, we used indirect evidence to validate the correctness of our map and, therefore, our SFP discovery methodologies. Probes from the same probe set should have the same allele calls for all of the individuals and can be viewed as replicates for a locus. Four hundred three probe sets from the SFPDev method and 319 probe sets from the z-function method contained more than one probe that was found to be an SFP. The allele calls for probes were in agreement with their probe set partners for >98.3% of all genotype assignments for the SFPDev method and 98.7% for the z-function–based assignments. This level of shared identity between “replicates” suggests that our methods of genotype assignment are highly reproducible and likely correct. The calculation of genotyping error rate is based on the assumption that recombination in our mapping population has not occurred within genes. Although intragenic crossovers do occur, given the limited number of recombination events represented in a population of 64 individuals and the tendency for recombination rates to be lower within genes (Myers et al., 2005), we believe the effect this has on error calculations to be minimal. This assumption also has the effect of introducing more missing data for consensus genotypes, potentially excluding them from the mapping stage of our analysis. Because of the rarity of crossing-over within genes, the majority of disagreements between probes within probe sets were likely genotyping errors and correctly excluded. We were also able to integrate the SFP genotyping data from both methods with previously determined genotypic data for 293 SSRs and generate two genetic maps. The Mendelian segregation of our SFPs suggests that the assignment of alleles to individuals is not random and that we are measuring heritable traits. The genetic maps generated from two different methods of SFP determination shared 435 probe sets. The allele calls for the mapping individuals was highly consistent, with 94% (409) of the shared probe sets having three or fewer allele assignment differences, which are assumed to be genotyping errors, between the two methods (Fig. 4). Ninety-nine percent (429) of the 435 probe sets mapped to the same chromosome in both methods. This level of continuity between SFP detection approaches is further evidence that the SFPs identified are valid and represent true polymorphism.

A gross physical position (bin) has been determined for 6426 wheat ESTs using 101 wheat deletion lines (Qi et al., 2004). Of the ESTs used by Affymetrix to construct their unigenes from which the probes on the array were designed, 165 SFPDev-method and 119 z-function–method mapped probe sets were developed from physically mapped ESTs. For the SFPDev-method SFPs, 81% mapped to the same chromosome or a homoeologous chromosome between the deletion and genetic maps and 8% were mapped to a different location but corresponded to the chromosome location found in the z-function method. Only 11% of the probe set locations were not verifiable by other methods. Verification of the z-function–method SFPs revealed similar results: 81% mapped to the same chromosome or a homoeologous chromosome on the genetic and deletion maps and 12% differed but were mapped to the same chromosome in the SFPDev method, leaving only 7% of probe set positions not verifiable. The corroboration of the chromosomal location of expressed sequence by SFP mapping and deletion mapping is strong evidence that our methodology and conclusions of the efficacy of our SFP genotyping is correct and robust.

Single-feature polymorphisms are associated with gene sequences and, when used to construct a genetic map, they infer an order to the genic regions along the linkage groups. West et al. (2006) were able to use the genomic sequence of *A. thaliana* to verify their SFP map. The bread wheat genome has not been sequenced and we were unable to do a direct verification of our SFP ordering, so we used rice as a surrogate genome. Rice has a realized and highly annotated genomic sequence and a well-studied syntenic relationship with wheat. The approach of using a binomial distribution test within a sliding window to examine the syntenic organization of our linkage maps relative to rice revealed blocks of genic conservation. The position of these blocks was consistent with results found by others (La Rota and Sorrells, 2004; Salse et al., 2008; Sorrells et al., 2003) and supports the SFP-based genotyping of bread wheat.

The nature of the polymorphism that is captured with the SFPDev and z-function methods is unknown. The sources of detected polymorphism may include SNPs, deletions, insertions, splice variants, or some cis-acting element affecting transcript binding as discussed by Luo et al. (2007). The type of the polymorphism(s) identified in this study is less important for mapping and marker development in that the polymorphism is detectable and heritable in a Mendelian manner. Luo et al. (2007) found that ~64% of the barley SFPs they identified were not based on sequence polymorphism at the site of probe binding but were likely expression-level polymorphisms (ELPs), and it is possible that in some cases we are detecting differences in gene expression rather than sequence polymorphisms as well. This could occur with the SFPDev method if one or more probes in a probe set had different levels of affinity for their target transcripts relative to the other probe set members or if some probes in a probe set were competitively bound by other probes on the chip. *z*-Function SFPs could actually be ELPs if the relative binding affinities of the PM and MM probes
change as the concentration of transcript changes. Our use of two methods that normalize the SFP discriminant value against gene expression levels and the large number of shared markers between the methods suggests that our methods of detection is preferentially identifying SFP.

Our strategy for SFP marker discovery used probe set–dependent and –independent methods, and its robustness is confirmed based on several lines of evidence: i) overlap in SFP content and position from linkage maps generated using two different methods; ii) Mendelian segregation of the discriminant values and integration of the SFPs with an SSR map; iii) corroborations of the genetic map chromosomal locations of SFPs using deletion-mapped ESTs; and iv) consistent alignment of wheat gene positions with their rice orthologs in conserved evolutionary blocks.

The results indicate that genotyping with the Affymetrix Wheat GeneChip offers an effective high-throughput platform to generate mapped markers from genomic regions in species with large, complex genomes, such as hexaploid bread wheat.

Acknowledgments

The authors wish to acknowledge the Canadian Crop Genomics Initiative of Agriculture and Agri-Food Canada for funding as well as Monika Eng, Brenda Terwisscha, and Kerry Ward for their technical assistance.

References


