Single feature polymorphism discovery using the wheat Affymetrix gene chip

Banks TW, Somers DJ, Jordan MC
Agriculture and Agri-Food Canada, Cereal Research Centre, 195 Dafoe Rd. Winnipeg, MB, Canada

INTRODUCTION

Genetic mapping of DNA-based molecular markers in crop plants has a long history of being connected to co-mapping of phenotypic traits. The approach has been used widely to position many genes of agronomic importance onto crop genetic maps and subsequently this information is used for marker-assisted selection (MAS). Wheat has benefited enormously from genetic mapping, particularly through QTL analysis of many disease resistance and seed quality traits that are now selected in breeding programs via markers.

The Affymetrix GeneChip® Wheat Genome Array is a microarray platform used to measure the expression of thousands of genes simultaneously. The 'chip' contains 61,127 probe sets representing 55,052 genes (unigenes) from *Triticum aestivum* and closely related species. Each probe set contains 11 perfect match and 11 mismatch oligonucleotide probes of 25 bases in length which collectively represent a specific gene sequence. Gene expression values are generated from combining the probe intensity values into a single measure of expression when hybridized with cRNA.

A Single Feature Polymorphism (SFP) is a recently developed type of molecular marker that detects sequence polymorphisms associated with the hybridization of cRNA to individual probes on a microarray[1]. SFPs have been developed for a number of plant species including *Arabidopsis*[2,3], rice[4] and barley[5,6]. Here we present the identification of 750 SFP markers in hexaploid wheat and their incorporation into a genetic map for a population of 81 doubled haploid lines.

MATERIALS AND METHODS

Plant Material and RNA Collection

Ninety-five doubled haploid (DH) lines from an extensively mapped population from the cross RL4452/AC Domain[7] were grown in a replicated field trial at one location.

RNA sampling for microarray hybridization was as described in Jordan et al.[8]. 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 days 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 labelled with genotype and replicate number. Tubes were stored at –80 °C until use for RNA isolation.

Labelling and Microarry Hybridization

mRNA was used for first and second strand cDNA synthesis using the One Cycle cDNA Synthesis kit (http://www.affymetrix.com) following the manufacturer's instructions. The resulting cDNA was purified and used to make biotin-labelled cRNA which was hybridized to an Affymetrix GeneChip® Wheat Genome Array (http://www.affymetrix.com) according to the manufacturer's instructions.

SFP Identification

196 CEL files (containing the raw hybridization intensities for each probe) for 95 field replicated DHs and 3 replicates of each parent, AC Domain and RL4452, were background-corrected and normalized using the 'affy' Bioconductor R-package[9]. Files for 14 individuals were removed from our analysis because the correlation coefficient between the replicates fell below a threshold value of 0.92. Our approach to identify SFPs is a hybrid of methods described by West[3] and Luo[6]. For each perfect match probe on the Affymetrix Wheat Array the following procedure was applied; a) an SFPDev value was calculated[3] from each sample and the values were placed into two groups using k-means clustering; b) the ratio of cluster sizes had to be 1:2 or less else the probe was rejected (this represents our threshold for segregation distortion.); c) all of the SFPDev values for AC Domain had to be in one cluster and the values for RL4452 in the other; d) we assumed that the SFPDev values were normally distributed in each cluster and tested that the mean SFPDev values between them were not the same using Student's t-test ($P<0.05$); e) continuing with the assumption of normality, the deviate method as describe by Luo et al. was used to test the likelihood that a probe's SFPDev value in one cluster did not belong in the other ($P<=0.003$)[6] and any values not meeting this criterion were regarded as missing data; f) if greater than 15% of the samples had missing data then the probe was rejected else alleles were assigned based on which parent a sample clustered with. Finally we compared the two replicates of each DH line across all probes. In cases where there was disagreement between the replicates a missing data value was assigned. The results of this processing pipeline were our pool of potential SFPs and were used to construct a genetic map.

Genetic Mapping and Rice Synteny

A genetic map containing the replicated potential SFP values and 453 existing SSR markers was constructed using JoinMap V4.0 (http://www.kyazma.nl). Only elements with a LOD score >= 4 were included. The probes that were placed on the map were considered to be truly polymorphic and deemed SFPs.

The program HarvEST was used to identify rice orthologs of unigenes containing SFPs (http://harvest.ucr.edu/).

RESULTS

SFP Detection

Using a hybrid approach of West's SFPDev[3] and methods to group and eliminate probes unlikely to be SFPs by Luo[6] we identified 750 SFPs between the parents of a DH population of 81 individuals. There were 443 unigenes represented by 750 SFPs. 63% of the unigenes contained a single SFP, and 84% had 1 or 2 polymorphic probes (Figure 1).

The segregation data for the SFPs and 453 SSRs were combined to form a genetic map of 27 linkage groups from 21 chromosomes and had a length of 2096 cM. Nearly 50% of the SFPs mapped to the B genome, 30 % to the A and 20 % to the D (Table 1). SFPs did not map evenly among homoeologous chromosomes with chromosomes 4D, 5D, 6D, 7D and 5A having only 2, 4, 4, 8 and 6 SFPs respectively.

The order of the wheat unigenes along the chromosomes was compared to rice and an alignment of the wheat map to the rice genome sequence showed conservation of gene order (Figure 2). For example chromosome 2B was determined to be syntenic with rice chromosome 4 and chromosome 7B was syntenic with rice chromosomes 6 and 8. These orthologous relationships agree with the wheat-rice synteny described by La Rota et al.[10].

DISCUSSION

The calculation used by West et al[3] to identify differences in probe hybridization was used in this study. Combined with filtration criteria from Luo et al.[6] we developed a method used to identify 750 SFPs, representing 443 unigenes, from a population of 81 DHs and their parents. The process used biological replicates of our population, k-means clustering and statistically based criteria to assign an allele score to each individual based on parent hybridization signals.

The majority of unigenes (probe sets) contain only one or two SFPs. This indicated that the algorithm, replication and mapping approach was preferentially selecting polymorphisms that are independent of expression level. In situations where the expression level of the gene is the polymporphism we would expect to observe unigenes with the majority of probes in the probe set being different between parents. The West method was used because it took advantage of the large amount of genetic replication per locus from our DH chips and it preferentially identified polymorphoric elements linked to genes. The nature of the polymorphisms identified in our study is unknown. Possibilities include SNPs, deletions, insertions, splice variants or flanking sequence difference that results in altered probe binding efficiencies between individuals. The nature of the polymorphism does not need to be known, only that the difference can be reliably exploited as a genetic marker.

SFPs are associated with gene sequences and therefore when used to construct our genetic map gave an order to the coding regions along the wheat linkage groups. Wheat does not have a sequenced genome so we used rice as a surrogate to evaluate the wheat gene positions. Rice was chosen because it has a realized and highly annotated genome sequence and a known syntenic relationship with wheat. When the wheat unigenes were compared with their orthologs in rice their map position was in agreement with wheat-rice synteny described by La Rota[10].

Our strategy for SFP marker discovery in wheat is based on several lines of evidence: 1) single probe specific hybridization with large sampling, 2) Mendelian segregation of derived SFPDev values that are independent of gene expression levels, 3) integration of SFPs into an SSR map and 4) consistent alignment of wheat gene positions with their rice orthologs.

REFERENCES

1 Borevitz, J., *Methods Mol Biol* **323**, 137 (2006).

2 Borevitz, J. O. et al., *Proc Natl Acad Sci U S A* **104** (29), 12057 (2007).

3 West, M. A. et al., *Genome Res* **16** (6), 787 (2006).

4 Kumar, R. et al., *PLoS ONE* **2** (3), e284 (2007).

5 Cui, X. et al., *Bioinformatics* **21** (20), 3852 (2005); Rostoks, N. et al., *Genome Biol* **6** (6), R54 (2005); Walia, H. et al., *BMC Genomics* **8**, 87 (2007).

6 Luo, Z. W. et al., *Genetics* **176** (2), 789 (2007).

7 McCartney, C. A. et al., *Genome* **48** (5), 870 (2005); McCartney, C. A. et al., *Plant Breeding* **125**, 565 (2006).

8 Jordan, M. C., Somers, D. J., and Banks, T. W., *Plant Biotechnol J* **5** (3), 442 (2007).

9 Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A., *Bioinformatics* **20** (3), 307 (2004).

10 La Rota, M. and Sorrells, M. E., *Funct Integr Genomics* **4** (1), 34 (2004).

	Genome				
Group	A	B	D	TOTAL	%
1	28	57	17	102	13.6
2	15	94	95	204	27.2
3	48	63	24	135	18.0
4	46	10	2	58	7.7
5	6	71	4	81	10.8
6	47	34	2	83	11.1
7	46	33	8	87	11.6
TOTAL	236	362	152	750	
%	31.5	48.3	20.3		

Table 1 – Distribution of SFP markers among the 21 wheat chromosomes. Markers are mapped to each of the chromosomes with the B genome being most strongly represented and the D genome having the fewest markers.

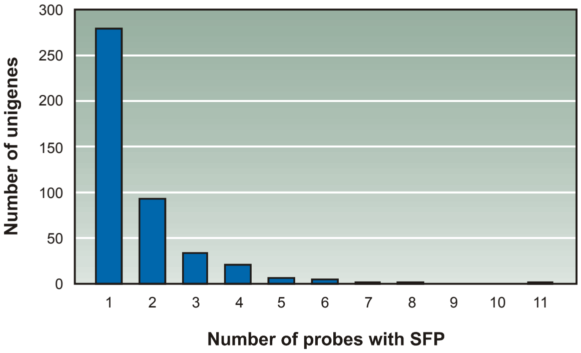

Figure 1. Histogram showing the distribution of unigenes based on the number probes identified to contain an SFP.

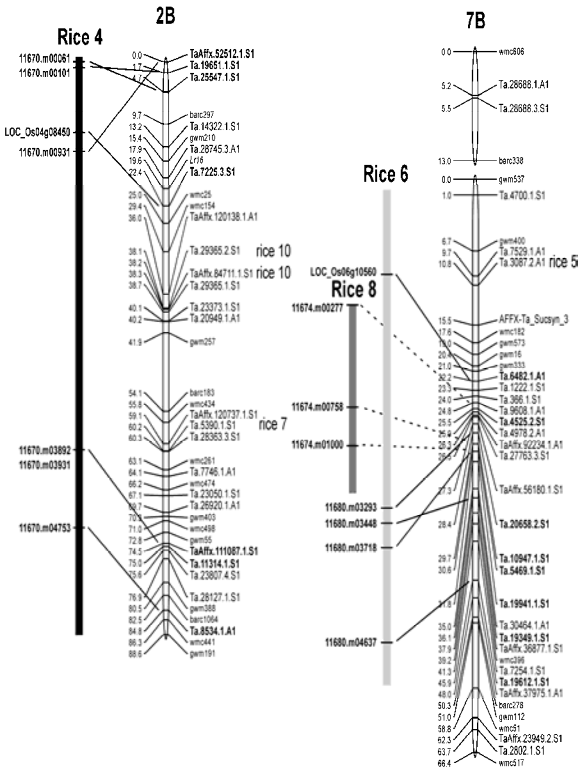

Figure 2 – Comparative mapping between the RL4452 x AC Domain chromosome 2B and 7B SSR-SFP maps and the orthologous rice chromosome physical maps. Wheat unigenes are represented by Affymetrix probe set designators beginning with Ta. Unigenes with BLAST hits to the rice genome sequence are indicated in bold. A further 4 unigenes are orthologous to rice 5, 7, and 10. Rice orthologs are presented showing the rice gene name and physical position on the rice chromosome.