Plant Biotechnology Journal

Plant Biotechnology Journal (2013) **11**, pp. 279–295 doi: 10.1111/pbi.12009

Discovery and development of exome-based, co-dominant single nucleotide polymorphism markers in hexaploid wheat (*Triticum aestivum* L.)

Alexandra M. Allen[1,*], Gary L. A. Barker[1], Paul Wilkinson[1], Amanda Burridge[1], Mark Winfield[1], Jane Coghill[1], Cristobal Uauy[2], Simon Griffiths[2], Peter Jack[3], Simon Berry[4], Peter Werner[5], James P. E. Melichar[6], Jane McDougall[7], Rhian Gwilliam[7], Phil Robinson[7] and Keith J. Edwards[1]

[1]*School of Biological Sciences, University of Bristol, Bristol, UK*
[2]*John Innes Centre, Norwich, UK*
[3]*RAGT, Ickleton, Essex, UK*
[4]*Limagrain, Woolpit, Suffolk, UK*
[5]*KWS, Thriplow, Hertfordshire, UK*
[6]*Syngenta Seeds Ltd, Whittlesford, Cambridge, UK*
[7]*KBioscience Unit 7, Hertfordshire, UK*

Received 20 April 2012;
revised 6 August 2012;
accepted 10 August 2012.
**Correspondence* (Tel 44 117 331 6770;
fax 44 117 925 7374;
email a.allen@bristol.ac.uk)

Keywords: wheat, next-generation sequencing, KASPar genotyping, single nucleotide polymorphism.

Summary

Globally, wheat is the most widely grown crop and one of the three most important crops for human and livestock feed. However, the complex nature of the wheat genome has, until recently, resulted in a lack of single nucleotide polymorphism (SNP)-based molecular markers of practical use to wheat breeders. Recently, large numbers of SNP-based wheat markers have been made available via the use of next-generation sequencing combined with a variety of genotyping platforms. However, many of these markers and platforms have difficulty distinguishing between heterozygote and homozygote individuals and are therefore of limited use to wheat breeders carrying out commercial-scale breeding programmes. To identify exome-based co-dominant SNP-based assays, which are capable of distinguishing between heterozygotes and homozygotes, we have used targeted re-sequencing of the wheat exome to generate large amounts of genomic sequences from eight varieties. Using a bioinformatics approach, these sequences have been used to identify 95 266 putative single nucleotide polymorphisms, of which 10 251 were classified as being putatively co-dominant. Validation of a subset of these putative co-dominant markers confirmed that 96% were true polymorphisms and 65% were co-dominant SNP assays. The new co-dominant markers described here are capable of genotypic classification of a segregating locus in polyploid wheat and can be used on a variety of genotyping platforms; as such, they represent a powerful tool for wheat breeders. These markers and related information have been made publically available on an interactive web-based database to facilitate their use on genotyping programmes worldwide.

Introduction

Bread wheat (*Triticum aestivum*) is an allohexaploid (AABBDD) crop derived from the hybridisation of the diploid genome of *Aegilops tauschii* (DD) with the AABB tetraploid genome of *Triticum turgidum* (Dubcovsky and Dvorak, 2007). These hybridisation events, the domestication process and the inbreeding nature of wheat have together resulted in a reduced level of genetic diversity between cultivated wheat varieties, when compared with their wild ancestors (Haudry *et al.*, 2007). Wheat breeders and geneticists require tools to exploit the genetic diversity available within germplasm collections and carry out breeding programmes, which utilise this diversity to maximum effect. Molecular markers enable breeders and geneticists to carry out this process; however, in allohexaploid wheat, the development of molecular markers has, until recently, been problematic due to the presence of homoeologous and paralogous copies of the various genes (Kaur *et al.*, 2012). Recent advances in genotyping platforms have built upon the wealth of data provided by next-generation sequencing (NGS) technologies to enable, for the first time, the large-scale identification, validation and application of molecular markers in wheat breeding programmes (Berkman *et al.*, 2012; Paux *et al.*, 2011). These developments have come at a critical time, where the need for a substantial increase in yields to feed a growing global population has coincided with reduced genetic gains and increasing climatic and environmental pressures (Dixon *et al.*, 2009; Reynolds *et al.*, 2009).

Many of the recently developed genotyping platforms rely on the identification of single nucleotide polymorphisms (SNPs), which are polymorphic between different wheat varieties (Paux *et al.*, 2011). To overcome the various bottlenecks and problems associated with SNP generation, characterisation and most importantly validation in wheat, we and others have previously used NGS-based technology to identify and map relatively large numbers of gene-based SNP loci (Allen *et al.*, 2011; Akhunov *et al.*, 2009; Chao *et al.*, 2010). However, these studies used cDNA and EST sequences and were therefore subject to variation

in expression of homoeologous and paralogous genes. In hexaploid wheat, this situation is further aggravated as homoeologous and paralogous genes are often silenced or can show differential spatial and/or temporal expressions (Adams and Wendel, 2005; Akhunova *et al.*, 2010; Liu *et al.*, 2009). Genomic DNA is likely to be a more reliable source of putative SNPs; however, the size of the wheat genome means that sequencing the whole genome of multiple varieties to the depths required for successful SNP identification is impractical, time consuming and costly (Biesecker *et al.*, 2011). To overcome these resource-associated problems, we have used a recently developed sequence capture targeted resequencing approach to characterise a significant proportion of the wheat exome (Winfield *et al.*, 2012). By using a reference collection of the wheat exome as the basis of our SNP collection, we have been able to sequence and compare equivalent regions of the wheat genome from several wheat varieties.

To be fully utilised in breeding programmes, putative SNPs need to be identified and converted to working assays on a high-throughput genotyping platform. Recently, several technologies have revolutionised wheat genotyping: Illumina's GoldenGate/Infinium technologies and KBioscience's KASPar (Akhunov *et al.*, 2009; Allen *et al.*, 2011). Development of these platforms has encouraged the widespread uptake of SNP-based genotyping in wheat; however, both technologies have two significant drawbacks. Firstly, they require the identification and characterisation of varietal SNPs among an excess of homoeologous and paralogous SNPs. Secondly, as both platforms were developed for diploid species, they have problems with the scoring of varietal SNPs in polyploid heterozygotes, for instance, F_2 and backcross populations. The detection of heterozygous SNPs in allohexaploid wheat is dependent on the ability of the system to accurately discriminate between different call ratios. For 'dominant' SNP assays, which amplify all three homoeologous copies, these systems are often incapable of distinguishing homozygote (having a call ratio of 4 : 2) and heterozygote lines (having a call ratio of 5 : 1) (Allen *et al.*, 2011; Paux *et al.*, 2011). In contrast, both genotyping platforms work well when the SNP is amplified from just a single homoeologous/paralogous copy. Such SNP assays are usually referred to as co-dominant SNP assays, that is, they are capable of differentiating between homozygotes (having a call ratio of 2 : 0) and heterozygotes (having a call ratio of 1 : 1). As such, co-dominant SNP assays are preferred markers compared with dominant SNP assays. Unfortunately, co-dominant SNP assays usually make up < 20% of the SNP assays generated by conventional means (Allen *et al.*, 2011). However, careful primer design can lead to the successful amplification of just one homoeolog/paralog, but this process is time consuming as the variable nature of each set of sequences demands a manual approach to primer design. To overcome this bottleneck, we have developed a SNP identification pipeline which incorporates a novel bioinformatics procedure designed to identify putative co-dominant SNP assays.

The developments described here have led to both the generation of an extensive set of putative varietal SNPs from genomic DNA and within this data set the identification of a subset of putative co-dominant SNPs. The use of an exome-based SNP discovery strategy has targeted gene discovery to genic regions. Validation of a subset of these putative co-dominant SNP assays and a comparison with dominant SNP markers has provided useful insights into their design and characteristics. Finally, the work described here has resulted in a significant increase in the number of gene-derived co-dominant SNP assays, which will be of considerable interest to wheat researchers, and in particular, the breeding community.

Results

SNP discovery

In this study, the exome of the UK varieties Alchemy, Avalon, Cadenza, Hereward, Rialto, Robigus, Savannah and Xi19 was captured using the NimbleGen capture array (NimbleGen array reference 100819_Wheat_Hall_cap_HX1) described in Winfield *et al.* (2012). This generated between 9.8 and 48.7 million reads on the Illumina GAIIx platform. Sequence data were filtered as described in the experimental procedures. Varietal SNPs were called in the filtered data where read coverage was sufficiently high that there was less than a 0.1% chance of an observed allelic difference between two varieties being due to failure to sample an allele. For example, if the varieties Avalon and Cadenza have observed calls of A(20) and A(10)G(10), respectively, we would expect half of the alleles in Avalon (10 calls) to be G under the null hypothesis that there is no real genotypic difference. Randomisation tests showed that for the data set as a whole, using a minimum expected count of 10 for null bases resulted in a false discovery rate of < 1%. Putative co-dominant SNP markers were identified as the subset of SNPs meeting the above criteria, but where every variety had only a single allele called. The SNP discovery pipeline identified 95 266 putative varietal SNPs in 26 551 distinct reference sequences (Winfield *et al.*, 2012). Examination of these SNPs suggested that as in our previous work, only 10%–20% were co-dominant (Table 1; Allen *et al.*, 2011), with 10 251 putative co-dominant SNP markers identified within 5308 contigs.

Co-dominant SNP validation

As co-dominant SNP assays are of significant interest to the wheat community, it is important that such assays have a level of polymorphism that is not significantly different to that previously shown for dominant SNP assays. In addition, it is important that the distribution of the co-dominant SNP markers across the three homoeologous genomes do not show a bias beyond those shown previously for mapped dominant SNP markers (Allen *et al.*, 2011). To assess both of these features, we selected a subset of 1337 putative co-dominant SNP assays for validation and further analysis. While the selection process used to identify this subset was essentially random, to aid further investigation, we selected those SNPs that appeared, via the sequence analysis, to be polymorphic between the parents of the UK mapping populations Avalon × Cadenza and Savannah × Rialto. Of the 1337 SNPs selected, we were able to design working KASPar assays for 1190 SNPs (89%; Data S1).

Genotyping of a panel of 47 wheat varieties using the 1190 KASPar probes resulted in 1138 probes (96%) generating data that could be scored consistently and were polymorphic in at least one of the varieties screened (Data S2). Examination of the genotypic data revealed three types of varietal SNP markers: co-dominant SNP markers (where homozygous scores were detected for all hexaploid varieties, for instance, only scores of A:A or T:T were obtained, Figure 1a); partially co-dominant SNP markers (where heterozygous and homozygous scores were detected in hexaploid varieties, that is, A:A, T:T and the mixed A:T, Figure 1b); and dominant SNP markers (where a single homozygous and heterozygous score was detected in hexaploid

Table 1 Summary of next-generation sequence data and SNPs identified for eight wheat varieties

Variety	No. of sequences (million)	No. mapped reads (million)	No. of SNPs (compared with Avalon)	No. of co-dominant SNPs (compared with Avalon)	Proportion of total SNPs that are co-dominant (%)
Avalon	27.2	10.9	N/A	N/A	N/A
Alchemy	44.4	16.9	22 092	2558	11.6
Cadenza	20.2	4.9	8909	1550	17.4
Hereward	41.0	15.2	13 379	2399	17.9
Rialto	30.4	11.8	15 141	2662	17.6
Robigus	31.4	6.3	10 823	2114	19.5
Savannah	48.7	16.8	18 648	2818	15.1
Xi19	9.8	2.9	4391	899	20.5

varieties, that is, A:A and mixed A:T only, Figure 1c). Of the 1138 validated probes, 734 (65%) were co-dominant, 194 (17%) were partially co-dominant and 210 (18%) were dominant. Dominant and co-dominant markers were used to screen an F_4 population known to contain homozygote and heterozygote individuals. Screening this population with co-dominant SNP assays resulted in three separate clusters for the various homozygote and heterozygote individuals (Figure 1d). However, screening the same population with dominant SNP assays produced a more scattered cluster where homozygous and heterozygous loci were indistinguishable (Figure 1e). Screening the F_4 population with partially co-dominant SNPs yielded the same results as described for both dominant and co-dominant SNP assays depending on the genotypes of the population parents (data not shown).

The SNP markers developed through the NimbleGen exome capture were compared with the existing database of SNP markers developed from EST and normalised cDNA sequences using the experimental procedures described in Allen *et al.*, 2011; (Table 2; Data S3). The number of co-dominant SNP assays generated was significantly higher, and dominant SNPs significantly lower, when compared with the previous data set of validated EST/cDNA-derived SNPs ($\chi^2 = 131.98$, $P < 0.001$). Comparison of the two data sets showed that the numbers of partially co-dominant SNP assays were not significantly different between the two data sets ($\chi^2 = 3.87$, $P = 0.14$). In addition, the polymorphism information content (PIC) scores and minor allele frequencies (MAF) were similar between the two data sets; they were highest for the partially co-dominant SNP assays and lowest for the dominant SNP assays.

Characterisation of the different SNP types

To characterise the different SNP marker types identified in this study, and in particular, the co-dominant and partially co-dominant SNP assays, several analyses were performed using the contig sequences containing the SNPs. The average sizes of contigs containing different SNP types were similar (co-dominant, 692 bp; dominant, 690 bp; partially co-dominant, 666 bp). We hypothesised that co-dominant SNP assays were likely to be derived from 5′ or 3′ untranslated regions (UTRs) of genic sequences. In the absence of functional coding constraints, such regions are more likely to have diverged between homoeologs and thus represent effectively unique regions of sequence. To address this hypothesis, SNP-containing contig sequences were used to screen, via BLASTX (Altschul *et al.*, 1990), the non-redundant (nr) protein database. If a match was found (E-value 1e-5), a further analysis was then performed to identify whether the SNP was located inside or outside the coding region. This analysis showed that a higher proportion of the contig sequences used to develop co-dominant SNP assays returned no hit when subjected to a BLASTX analysis against the nr database, compared with dominant SNP assay sequences. Where a hit was identified, a higher proportion of the co-dominant SNPs were found to lie outside the coding region, compared with dominant SNPs. The number of co-dominant SNPs located within known coding

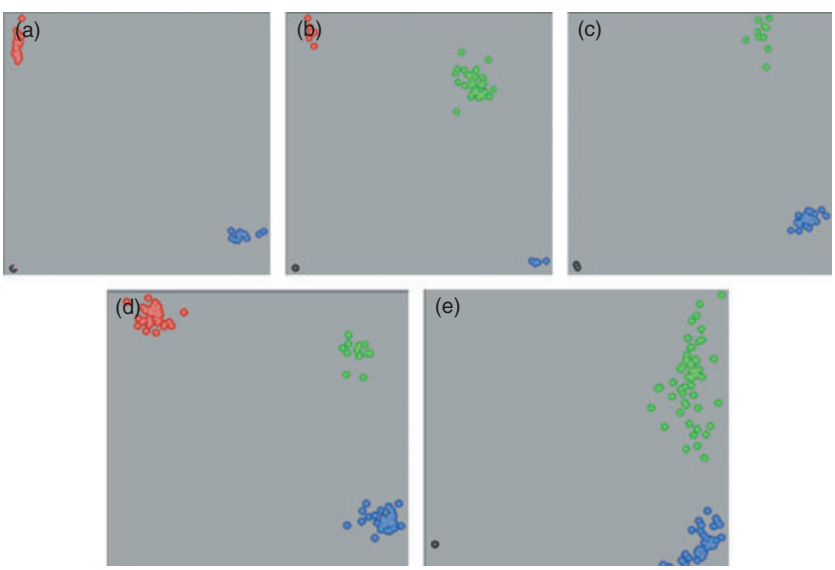

Figure 1 KASPar plots of different varietal SNP types screened against a panel of hexaploid wheat varieties with examples of (a) a co-dominant SNP assay, (b) a partially co-dominant SNP assay and (c) a dominant SNP assay. Screening an F_4 population containing heterozygotes with a co-dominant SNP assay results in a separate cluster for heterozygote individuals (d). Screening the same population with a dominant SNP assay produces a more scattered cluster where homozygote and heterozygote individuals are indistinguishable.

Table 2 Summary of validated SNPs

SNP type	Validated NimbleGen SNPs (%)	Validated EST/cDNA SNPs (%)	Total validated SNPs	Average minor allele frequency	Average PIC score
Dominant	210 (19)	1195 (58)	1407	0.249	0.273
Partially co-dominant	194 (17)	437 (21)	632	0.315	0.315
Co-dominant	734 (64)	444 (21)	1175	0.270	0.287
All SNPs	1138	2076	3214	0.270	0.286

PIC, polymorphism information content.

regions was significantly lower than would be expected to occur by chance ($\chi^2 = 7.56$, $P = 0.02$). Partially co-dominant SNPs were midway between dominant and co-dominant SNPs (Figure 2a). Across all SNP marker types, the average length of contigs returning no hit was lower (approximately 520 bp) than the average length of contigs returning BLASTX hits (approximately 770 bp), suggesting that contig length affected the likelihood of obtaining a BLASTX match. To check whether the contigs returning no hit represent genes that had not yet been annotated, the same contig sequences were subjected to a BLASTN analysis (E-value 1e-3) against the NCBI nr nucleotide database; 86% of the contigs returning no hit from the BLASTX analysis also had no match in the BLASTN nr database.

We further hypothesised that some co-dominant SNP assays may have been derived from single-copy regions of the wheat genome. Such regions may have either been unique to only one progenitor genome or alternatively one or more copies have been lost since polyploidisation. To address this second hypothesis, the same sets of sequences were screened, using BLASTN, against the 5 × Chinese Spring genomic raw reads at http://www.cerealsdb.uk.net/. BLAST hit coverage was calculated for every nucleotide position in the query sequence and averaged over the whole sequence to derive a mean contig coverage. All three SNP types peak in coverage at 15 × , indicative of three gene copies each at 5 × coverage; however, the co-dominant SNPs and to a lesser extent the partially co-dominant SNPs had a secondary peak of coverage at fivefold coverage, indicative of single-copy number sequences. This peak is absent from the dominant SNPs (Figure 2b).

These analyses were combined by comparing the coverage of sequences containing co-dominant SNPs that had different BLASTX results. The sequences with 5 × coverage are most highly represented by those returning a 'no hit' from BLASTX analysis (Figure 2c). When contig length was plotted against median coverage for co-dominant SNPs, no relationship was observed ($r = -0.0009$). A similar result ($r = -0.0009$) was obtained by performing the same analysis using only sequences returning no BLASTX hit, suggesting that the contig length does not affect the number of hits returned from the BLASTN analysis or the estimated level of coverage.

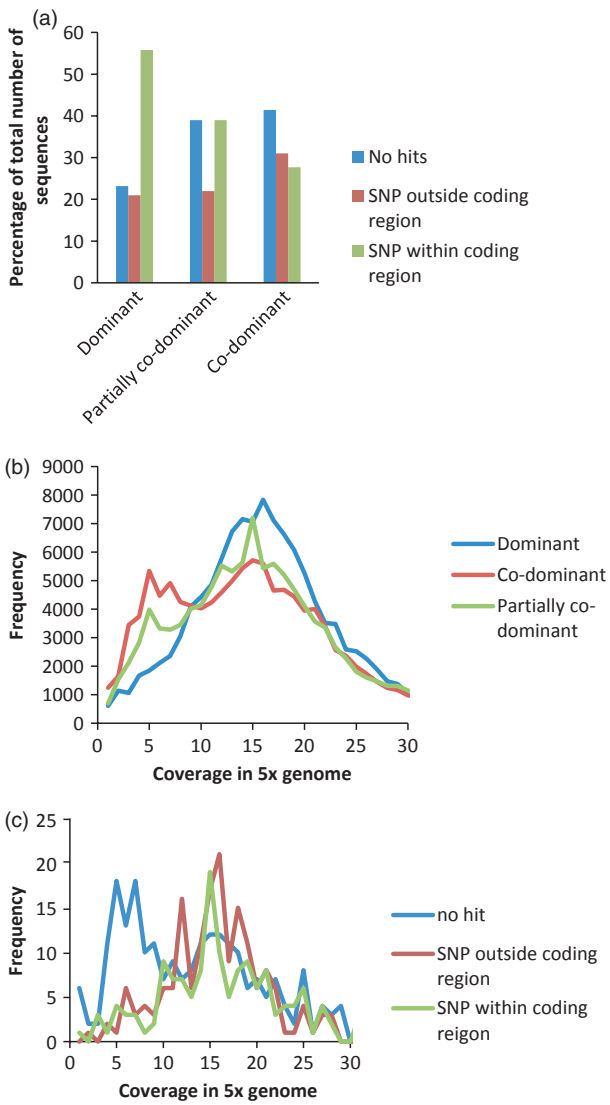

Figure 2 Characteristics of sequences containing dominant, partially dominant and co-dominant SNP types. (a) BLASTX analysis against the non-redundant (nr) protein database and (b) BLASTN against the 5 × Chinese Spring raw reads. (c) Coverage of sequences containing co-dominant SNPs against the 5 × Chinese Spring raw reads classified according to the BLASTX designation in (a).

Map location of dominant, partially co-dominant and co-dominant SNP assays

The map positions of the different SNP marker types were investigated to determine whether any bias in genetic location was introduced by using co-dominant SNP assays in two doubled-haploid mapping populations developed from UK cultivars Avalon × Cadenza (A × C) and Savannah × Rialto (S × R). Of the 3214 SNP markers developed to date (Table 2), 2109 were identified as polymorphic between Avalon and Cadenza (via screening of the 47 varieties above), of which 1807 were placed on the Avalon × Cadenza map. These consisted of 1152 EST/cDNA-derived markers and 655 NimbleGen-derived markers. Of the remaining 1105 SNP assays not polymorphic between Avalon

and Cadenza, 562 were identified as polymorphic between Savannah and Rialto and 541 of these markers were placed on the Savannah × Rialto map. These consisted of 187 EST/cDNA-derived markers and 375 NimbleGen-derived markers. To enable comparisons between the maps, 231 evenly spaced loci from the Avalon × Cadenza map were also included on the Savannah × Rialto map (Figure 3; Data S3). For the Avalon × Cadenza map, previously mapped SSR markers were used to help assign linkage groups to chromosomes (http://www.wgin.org.uk/resources/MappingPopulation/TAmapping.php; Data S4). In total, 2350 (73%) of the validated SNP markers were mapped; these comprised of 969 dominant SNP loci, 444 partially co-dominant loci and 937 co-dominant SNP loci. In the Avalon × Cadenza map, the linkage groups ranged from 54.5 to 239.0 centiMorgans (cM) in size, with 8–214 SNP markers. The total map length was 2434.4 cM with an average spacing of 1.3 cM between SNP loci. In the Savannah × Rialto map, linkage groups ranged from 1.3 to 221.1 cM, with 2–98 SNP markers. The total Savannah × Rialto map length was 2861.8 cM with an average spacing of 3.8 cM between SNP loci (Table 3). The two linkage maps aligned well with each other, showing similar arrangements of common loci within linkage groups.

In both populations, over 97.5% of the SNP markers could be mapped unequivocally to a linkage group and assigned to a unique chromosome position. The lack of markers on the short arm of chromosome 1B in the Savannah × Rialto map can be attributed to the presence of the same 1BL.1RS rye translocation in both Savannah and Rialto, where the short arm of rye chromosome 1B has replaced the short arm of wheat chromosome 1B (Figure 3). Clustering of SNP markers was observed in both linkage maps, with 55% of A × C markers and 61% of S × R markers being completely linked (0 cM distance between them). Of the remaining markers, 81% of A × C markers are separated by < 5 cM and 90% are within 10 cM of the next marker. For the S × R map, these figures are lower (55% markers separated by < 5 cM and 74% within 10 cM), probably due to a smaller number of markers on the map. Similar levels of clustering were observed for the different SNP marker types; 60% of A × C co-dominant markers and 52% of dominant markers were completely linked. Of the remaining co-dominant markers, 77% of markers are within 5 cM of each other and 88% are within a 10 cM interval. The corresponding proportions for dominant markers are similar (84% within 5 cM and 92% within 10 cM).The different SNP types showed similar patterns of distribution between the A, B and D linkage groups in both the Avalon × Cadenza and Savannah × Rialto maps, with the only difference of a higher proportion of the partially co-dominant markers mapped to the D genome (Figure 4a). Similarly, although clustering of SNP markers was observed within the linkage groups, there was no obvious bias of different marker types (Figure 3).

Summary statistics of mapped loci

The summary statistics of mapped SNP markers were compared to assess whether different marker types had varying levels of polymorphism in the 47 varieties screened and to ensure that the co-dominant SNP markers developed in this study would be useful across a wide range of material. The mean MAF and levels of genetic diversity of SNP markers were compared between the different marker types and assigned genomes of the Avalon × Cadenza and Savannah × Rialto maps (Table 4). These summary statistics were very similar for both maps; however, differences within the maps were observed. Partially co-dominant SNP assays had the highest average MAF and PIC scores, and dominant SNP assays had the lowest. Loci from the separate homoeologous genomes had consistent MAF and PIC measurements, although the A and D genome measurements were slightly higher than the B genome (Table 4). The different classes of SNP loci showed differences in the distribution of MAF scores. Co-dominant and partially co-dominant loci showed an increased proportion of medium and high frequency alleles compared with dominant loci (Figure 5a). Similarly, D genome loci showed a trend to have higher MAF compared with A and B genome loci (Figure 5b). The distribution of PIC scores showed that co-dominant and partially co-dominant loci types had a higher proportion of high PIC scores than dominant SNP assays (Figure 5c). A and B genome loci had a similar distribution of PIC scores, while D genome loci had a comparatively higher proportion of high PIC scores (Figure 5d).

Discussion

In this study, we present a SNP discovery pipeline capable of identifying large numbers of putative SNPs from genomic sequence obtained by targeted exome capture. This proved an efficient method to generate equivalent sequences from multiple varieties from which we were able to generate over 90 000 putative SNPs between eight elite UK cultivars. Given our results, this same approach is likely to prove highly effective at identifying SNPs across a wide range of cultivars, and in a wider range of germplasms, such as landraces, progenitors and alien species. By cataloguing SNPs using a reference collection of sequences derived from just the wheat exome, we provide a unique context for each SNP, thereby both reducing the chance of duplications within the SNP data set and allowing direct comparisons between different wheat lines. A key advantage to the SNP collection described here compared with other SNP markers such as insertion site–based polymorphisms (ISBPs; Paux *et al.*, 2010) is that by the nature of the targeted sequencing, all the SNPs developed are associated with genes, and as such, are likely to prove useful in gene-based marker-assisted breeding.

When examined further, the SNP database was shown to contain 10 251 putative co-dominant SNPs. Validation of over 10% of the co-dominant SNP assays on the KASPar genotyping platform resulted in a significantly improved validation rate compared with our previous study with 96% being polymorphic between the varieties screened compared with 67% as described in Allen *et al.* (2011). This increased validation rate is probably due to the use of genomic DNA, as opposed to transcriptome-derived data, in the SNP discovery phase where the problems of expression differences and presence of intron–exon splice sites hindered effective SNP identification and primer design (Trick *et al.*, 2012). Of the subset of putative co-dominant SNP assays, over 80% were validated as co-dominant or partially co-dominant, compared with < 20% in previous studies where random SNPs were validated (Allen *et al.*, 2011). The occurrence of putative co-dominant SNP assays, which were dominant when validated, is likely to be due to the presence of homeologous sequences that were not represented in the sequence data but were amplified by the KASPar primers. This may be due to a feature of individual sequences that prevent them from mapping on to the assembly or a consequence of reduced sequence coverage.

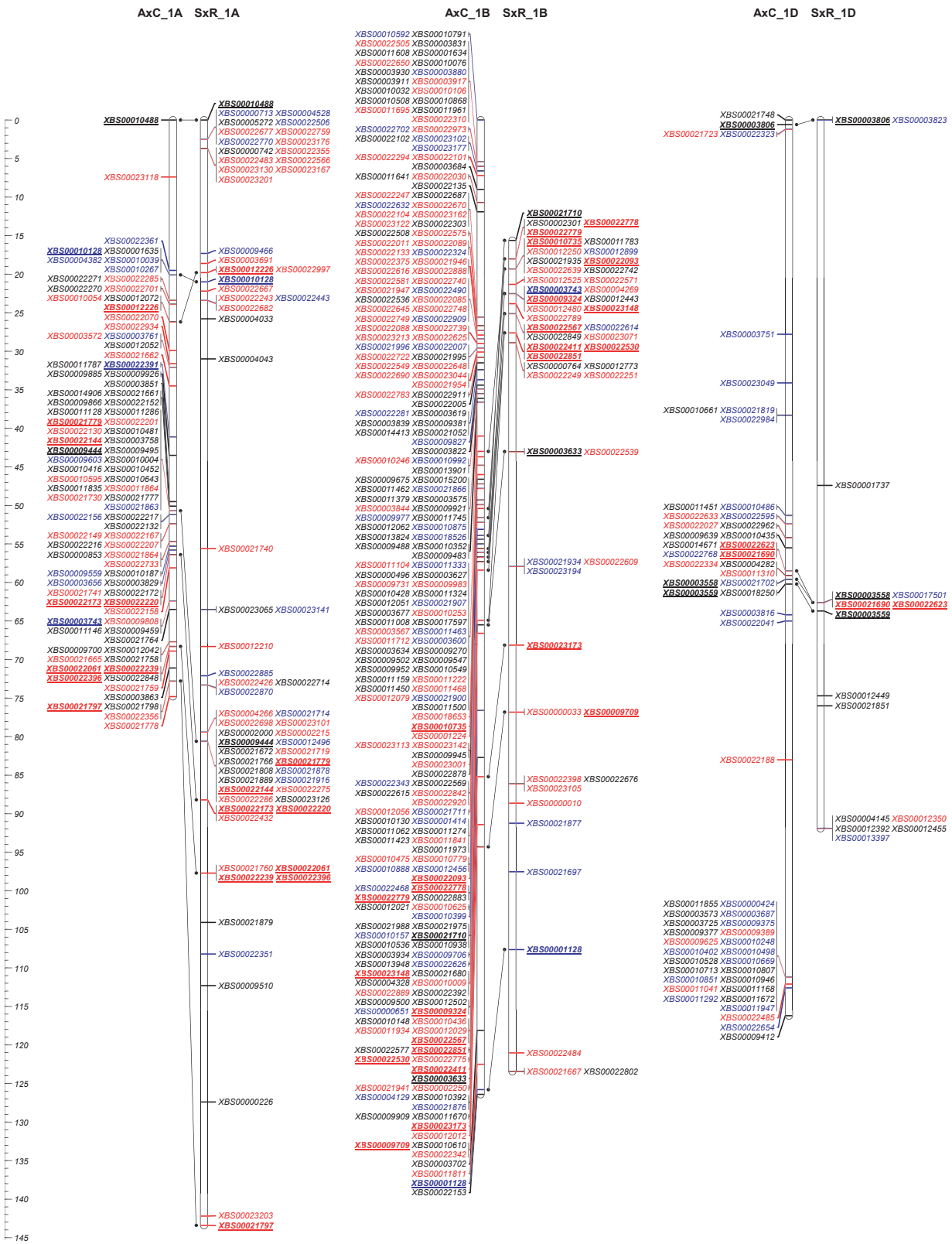

Figure 3 Genetic linkage maps of wheat derived from 190 Avalon × Cadenza doubled-haploid lines and 95 Savannah × Rialto doubled-haploid lines. Each linkage group was assigned to a chromosome indicated above the linkage group, and chromosomes are arranged with the short arm above the long arm. SNP loci mapped in this study are designated XBS and are coloured according to the SNP type: Dominant SNP loci are shown in black, co-dominant SNP loci are shown in red and partially co-dominant SNP loci are shown in blue. Common markers between the Avalon × Cadenza and Savannah × Rialto maps are underlined. Map distances, calculated using the Kosambi mapping function, are shown in centiMorgans (cM) on the ruler to the left of linkage groups.

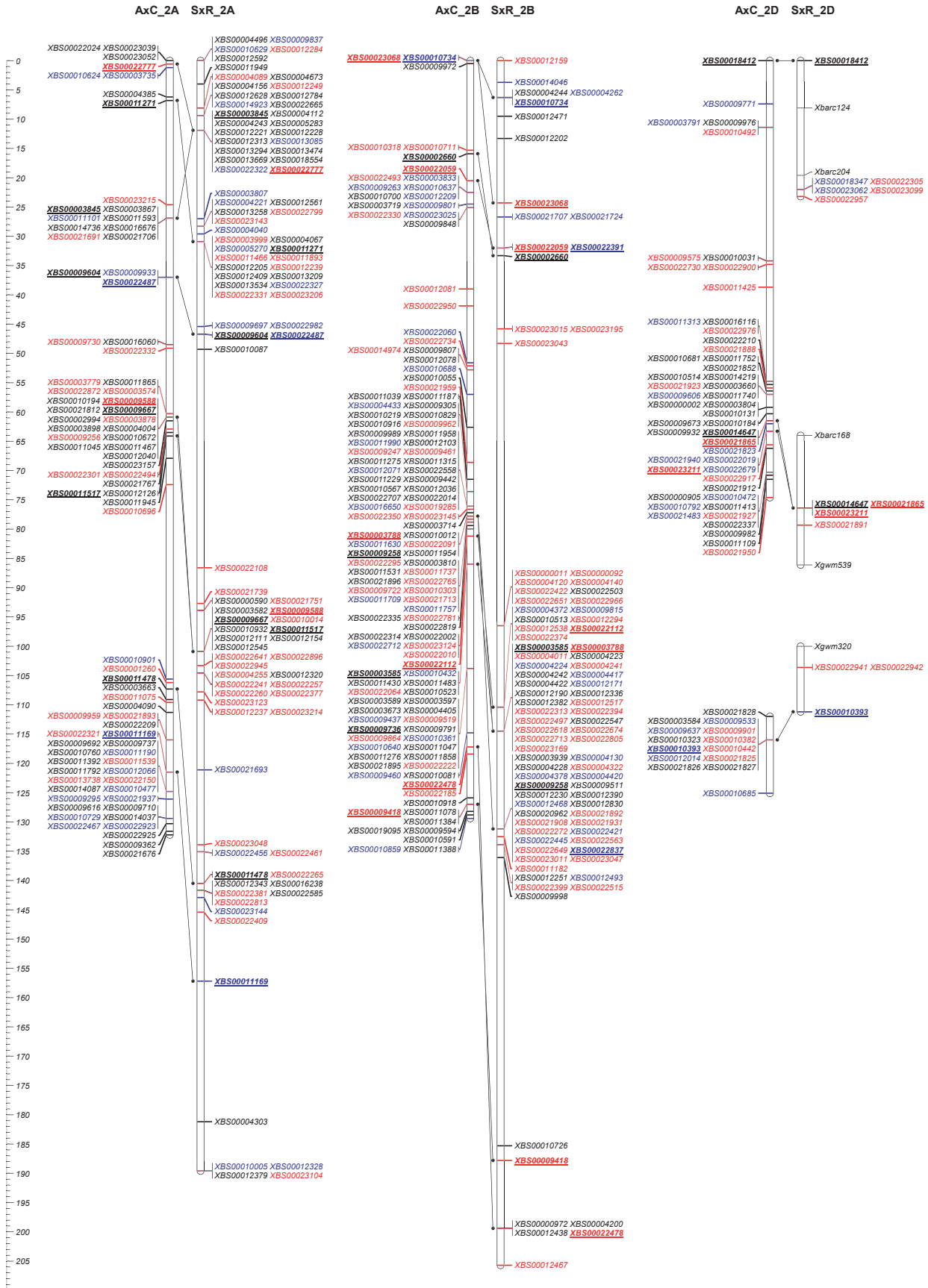

Figure 3 Continued

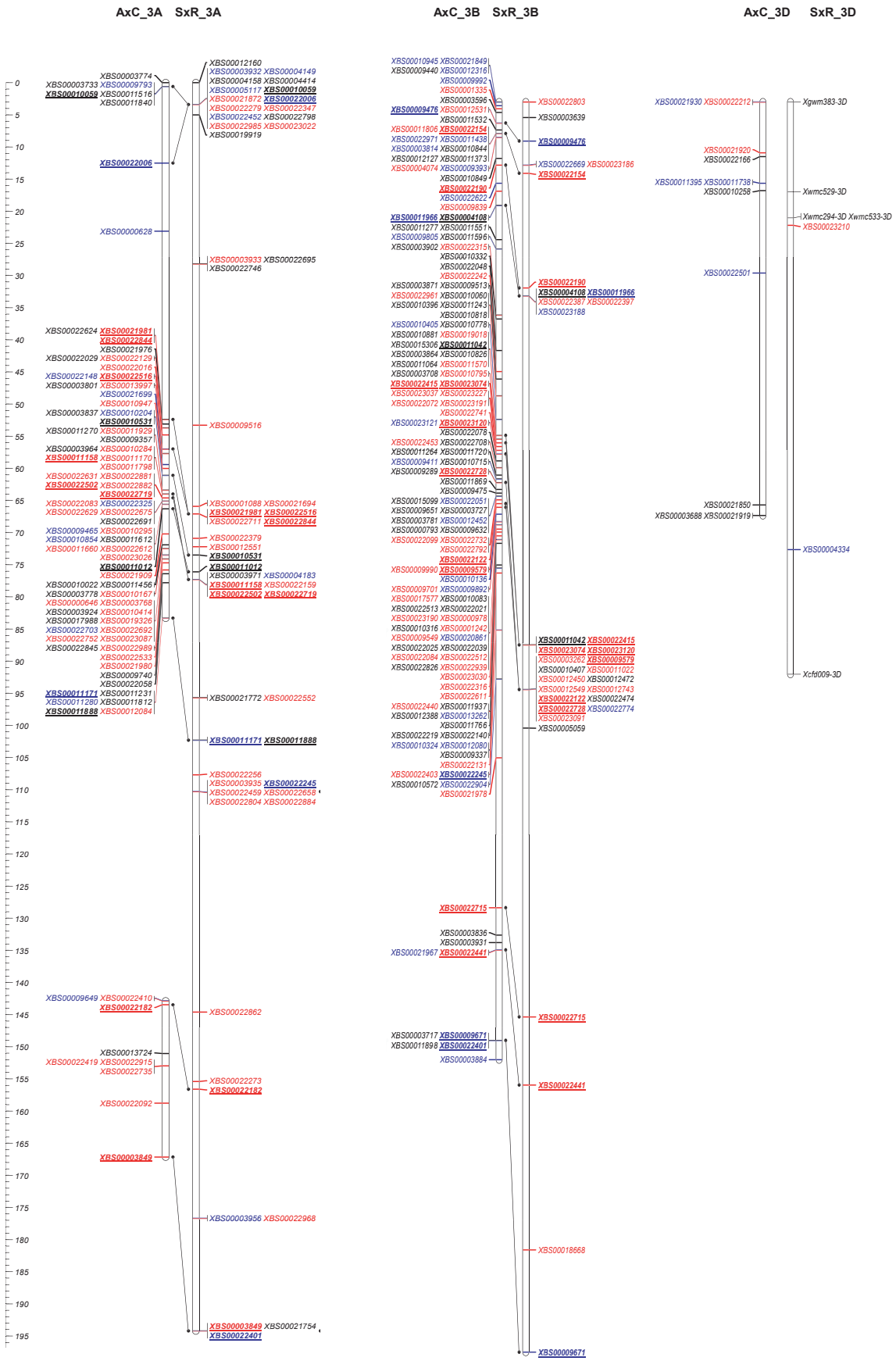

Figure 3 Continued

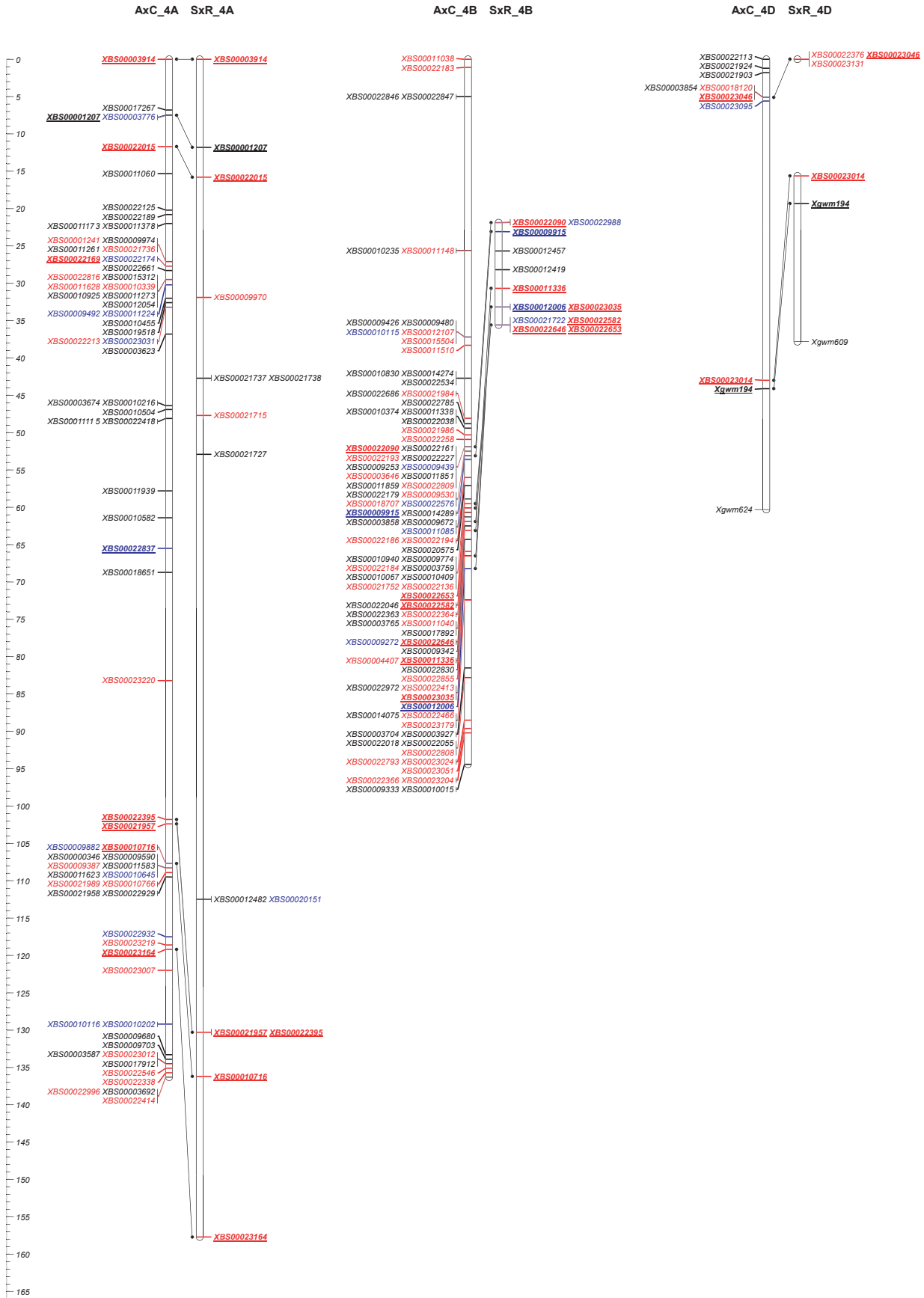

Figure 3 Continued

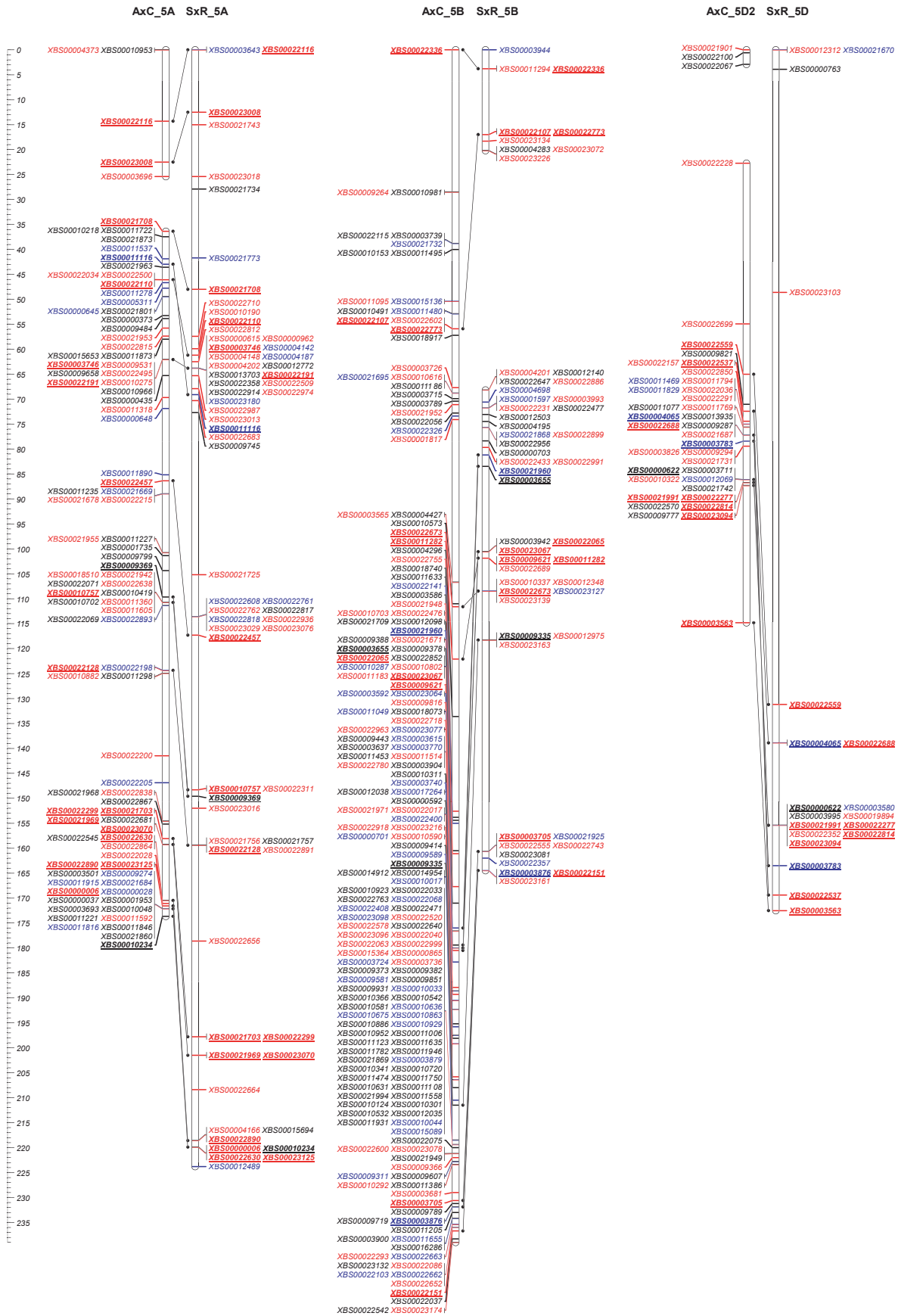

Figure 3 Continued

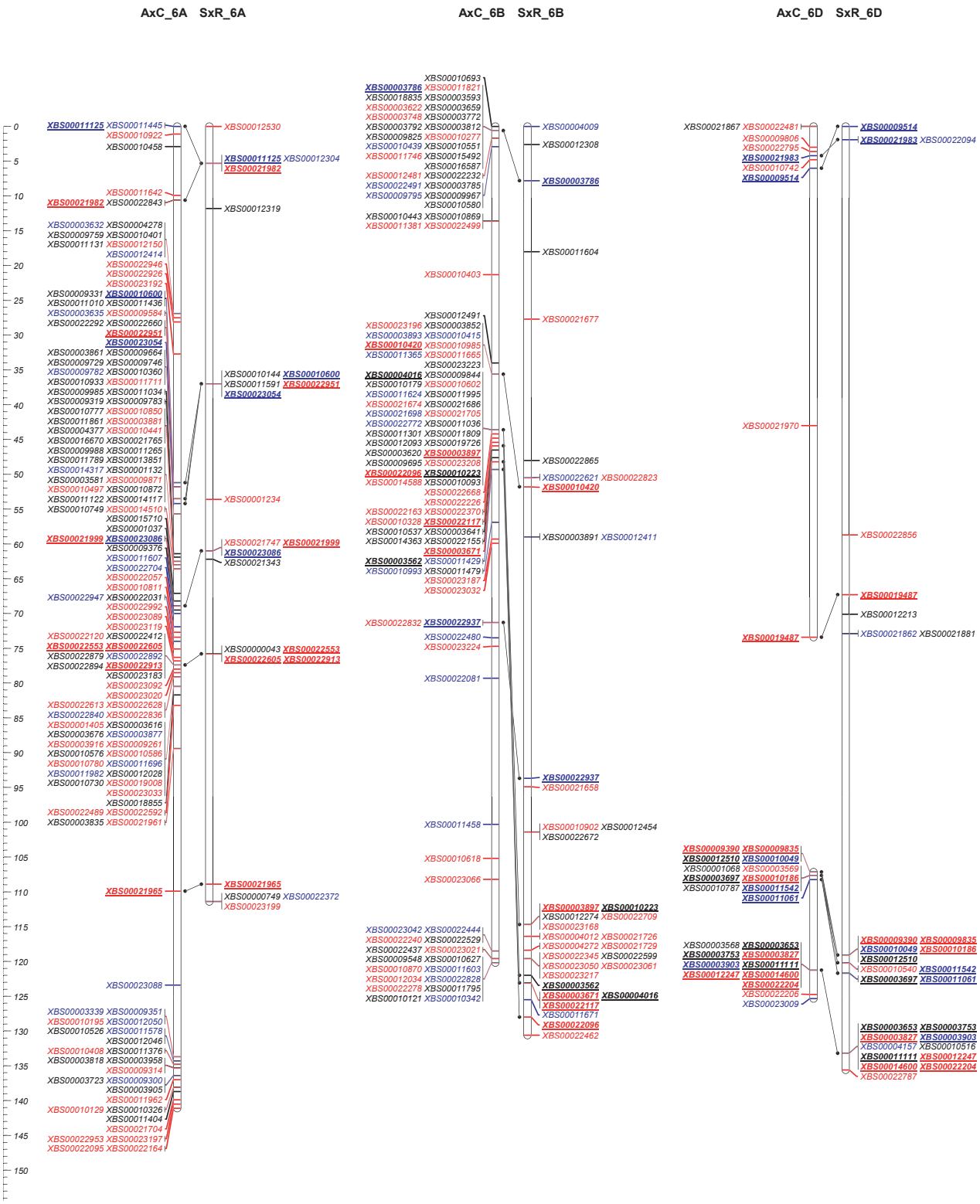

Figure 3 Continued

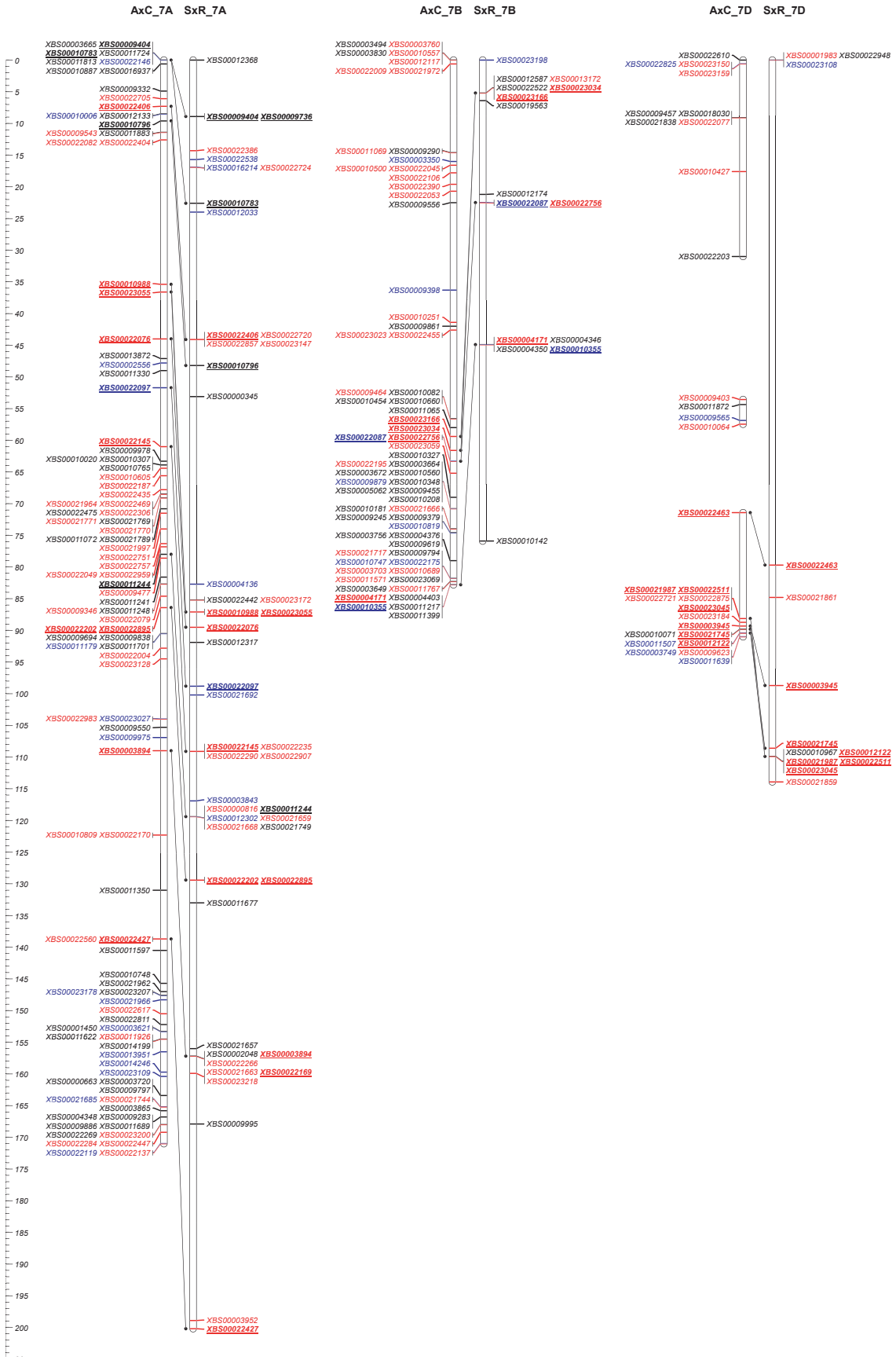

Figure 3 Continued

Table 3 Summary of linkage groups and mapped loci

	Avalon × Cadenza map			Savannah × Rialto map		
Chromosome	Number of bristol SNP loci	Size (cM)	Average spacing between loci (cM)	Number of bristol SNP loci	Size (cM)	Average spacing between loci (cM)
1A	89	74.8	0.8	68	143.4	2.1
1B	214	126.4	0.6	50	115.9	2.3
1D	57	116.2	1.6	15	104.1	6.9
2A	82	132.2	1.6	98	211.1	2.2
2B	118	129.4	1.1	91	221.7	2.4
2D	63	77.7	1.2	13	49.8	3.8
3A	86	107.6	1.3	55	194.3	3.5
3B	136	154.5	1.1	34	205.0	6.0
3D	11	66.7	6.1	2	72.3	36.2
4A	71	136.3	1.9	14	157.7	11.3
4B	87	94.4	1.1	12	23.9	2.0
4D	8	60.3	7.5	4	1.3	0.3
5A	93	162.7	1.7	64	223.8	3.5
5B	172	239.0	1.4	51	147.2	2.9
5D	38	95.0	2.5	19	194.6	10.2
6A	136	141.1	1.0	23	111.4	4.8
6B	105	120.2	1.1	37	158.6	4.3
6D	31	91.6	3.0	28	135.6	4.8
7A	103	171.0	1.7	48	200.2	4.2
7B	64	82.8	1.3	14	56.1	4.0
7D	29	54.5	1.9	13	133.8	10.3
Total	1793	2434.4	1.3	753	2861.8	3.8
A genome	660	925.7	1.4	370	1241.9	3.4
B genome	896	946.7	1.1	289	928.4	3.2
D genome	237	562	2.2	94	691.4	7.4
Group 1	360	317.4	0.8	133	363.4	2.7
Group 2	263	339.3	1.3	202	482.6	2.4
Group 3	233	328.8	1.4	91	471.6	5.2
Group 4	166	291	1.8	30	182.9	6.1
Group 5	303	496.7	1.6	134	565.6	4.2
Group 6	272	352.9	1.3	88	405.6	4.6
Group 7	193	308.3	1.6	75	390.1	5.2

Characterisation of the validated co-dominant SNP assays showed that their PIC scores and MAF were on average higher than dominant SNP assays, suggesting they are highly useful genetic markers for use on a range of materials. Analyses using the contig sequences containing the different SNP types revealed that co-dominant SNP assays were more likely to be located in contigs returning no BLAST hit to either protein or nucleotide databases, or outside coding regions in those contigs returning a BLASTX hit. Our analysis is consistent with the hypothesis that a proportion of the contigs used to develop co-dominant SNP assays represent single-copy genes. These contigs most likely represent genes that were lost before or during the domestication process as they are found as single copies in both landraces, such as Chinese Spring, and modern varieties. For those SNP contigs with 15 × Chinese Spring genomic coverage, it is quite possible that while these are represented as three homoeologs in Chinese Spring, they have undergone gene loss down to single copy in the UK germplasm we have studied. Intracultivar heterogeneity has been documented between elite inbred lines of crop species, and there are reports of intervarietal gene loss in wheat (Haun *et al.*, 2011; Swanson-Wagner *et al.*, 2010; Winfield *et al.*, 2012).

In addition to the factors outlined above, the Chinese Spring reference used to map the NimbleGen-captured sequences was based upon cDNA. If only one homoeolog was sampled in the cDNA data, and this was sufficiently divergent from the other two homoeologous copies, we may have only been able to map Illumina sequence data to that single genome. This would be the case in many 3′ UTR regions that are more divergent than protein-coding sequence and have diverged sufficiently during evolution to preclude their co-amplification in the KASPar PCR. This homoeolog-specific amplification could fortuitously lead to the development of co-dominant markers, yet BLAST analysis of such sequences against the Chinese Spring genome would show them to be present in three copies. In summary, investigations into the average copy number of sequences used to develop co-dominant SNP assays and the location of the SNP in the sequence suggests that these SNPs are likely to reside in single-copy genes of as yet unknown function, and/or three-copy genes which are sufficiently divergent that sequence data from one homoeolog does not map to other copies. The uncharacterised nature of these genes makes them an exciting and intriguing source of further co-dominant markers and scientific investigation.

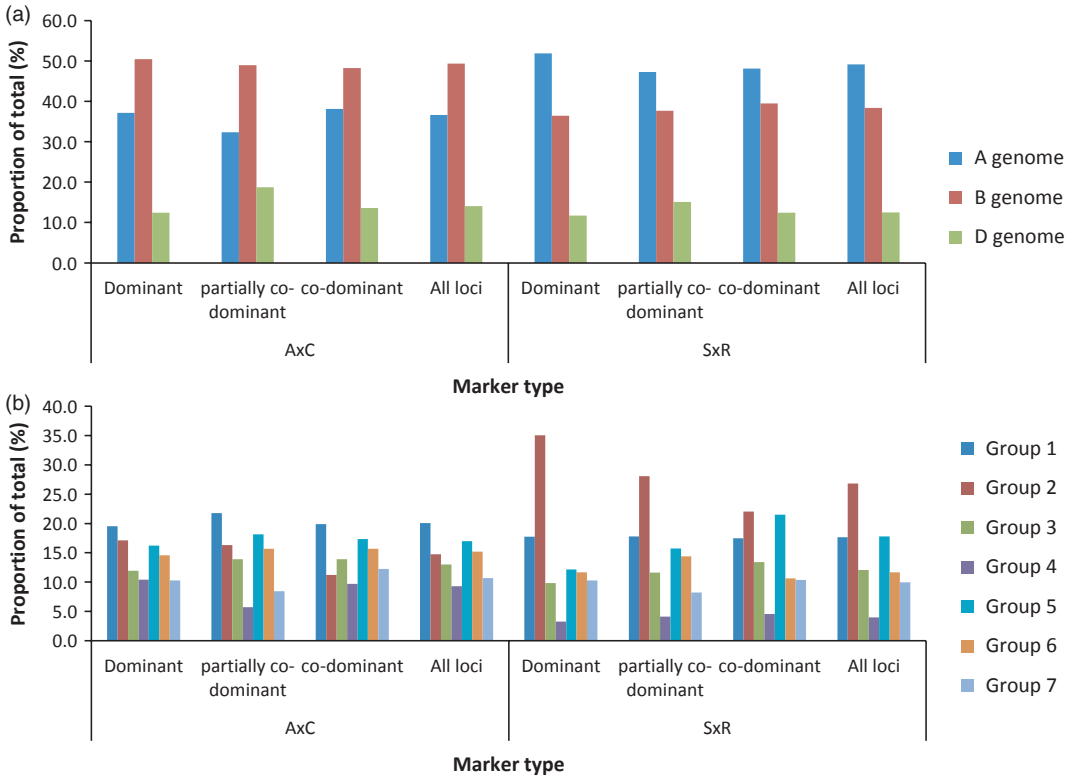

Figure 4 Distribution of the different marker types across the (a) A, B and D linkage groups and (b) homoeologous chromosome groups of the Avalon × Cadenza and Savannah × Rialto genetic maps.

Table 4 Summary statistics for mapped loci

	Number of loci	Minor allele frequency	Polymorphism information content
Avalon × Cadenza mapped loci	1793	0.264	0.284
Co-dominant loci	672	0.277	0.290
Partially co-dominant loci	332	0.308	0.313
Dominant loci	789	0.235	0.266
A genome	660	0.263	0.284
B genome	896	0.260	0.283
D genome	237	0.278	0.288
Savannah × Rialto mapped loci	753	0.284	0.299
Co-dominant loci	395	0.291	0.300
Partially co-dominant loci	213	0.319	0.323
Dominant loci	145	0.246	0.280
A genome	370	0.290	0.305
B genome	289	0.272	0.290
D genome	94	0.294	0.305

During this study, we have created two complementary genetic maps, enabling 73% of our validated SNPs to be assigned a map location. The co-dominant SNP loci had a similar pattern of distribution between linkage groups compared with dominant loci, suggesting that co-dominant SNP markers have a similar distribution to the previously used dominant markers. Analysis of the MAF and PIC scores of the different types of mapped SNPs demonstrated that the co-dominant and partially co-dominant SNP markers had higher levels of genetic diversity within the lines tested, compared with dominant SNP assays, suggesting that co-dominant SNP assays are highly suitable for use as genetic markers.

The two genetic maps aligned well with each other, with a similar assignment and order of common markers. Clustering of SNP markers was observed in both linkage maps, indicating that despite the relatively large mapping populations used, a lack of recombination events between these markers may affect map resolution. This may be overcome by mapping these markers against a larger number of individuals. Preliminary results indicate that mapping a subset of 223 evenly spaced A × C markers on 566 individuals from an extended A × C population reduced the proportion of completely linked markers from 50.4% to 44.9%, and it is likely that this figure could be further decreased by specifically targeting clustered markers. However, despite the high proportion of clustered markers, 89% of the remaining markers map to within 10 cM of the next marker, suggesting that these provide good overall coverage of the genome, with few gaps. When co-dominant and dominant markers were compared separately, similar proportions of markers were observed to map to within 10 cM of each other (88% and 92%, respectively), suggesting that both marker types are similarly distributed across the map.

Both maps had a relatively low proportion of D genome loci; this has been observed in previous studies and is likely to relate to a lower level of diversity found in the D genome due to the effects of the genetic bottleneck that accompanied the domestication of hexaploid wheat (Allen *et al.*, 2011; Caldwell *et al.*, 2004; Chao

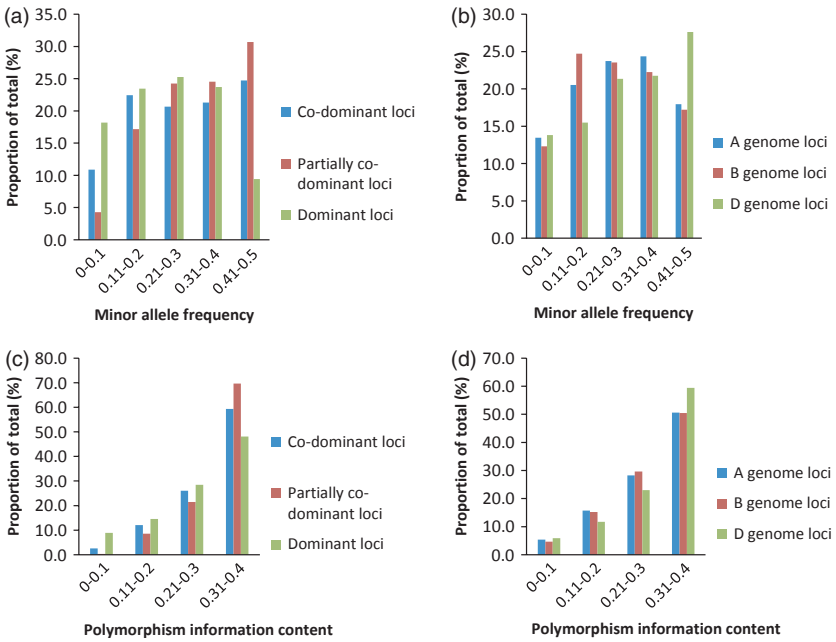

Figure 5 Distribution of minor allele frequency (MAF) and polymorphism information content (PIC) scores among the 47 wheat varieties. Loci were separated into subgroups according to (a,c) marker type and (b,d) genome.

et al., 2009). Although the mean MAF and PIC scores were similar for A, B and D genome loci, some differences were observed in the distributions of these measurements. The results for A and B genome markers were similar; however, loci assigned to the D genome had a higher proportion of high MAF and PIC scores compared with A and B genome loci. This is the opposite to what has been detected in previous studies (Akhunov *et al.*, 2010; Chao *et al.*, 2009) and suggests that, although the lower genetic diversity within the D genome hinders SNP discovery and marker development, the D genome SNPs identified by our pipeline are as informative and useful as loci from the A and B genome.

This study has described the design, implementation and validation of a pipeline designed to identify gene-based co-dominant SNP assays from genomic DNA sequence data. The validation results suggest that this approach is highly efficient and the resulting co-dominant SNP markers are evenly distributed across the genome with relatively high MAF and PIC scores. As such, these should prove a highly valuable resource for use in breeding programmes. The construction of two complementary genetic maps has maximised the amount of mapped SNP loci and allowed comparisons between UK breeding materials. The genotype data generated in this study for 47 widely used wheat lines, combined with genetic map locations for SNP markers, should enable wheat researchers to target their efforts to regions of interest and enable QTL studies and marker-assisted selection. The markers described in this study will be useful in linking the genetic map with the developing physical maps and so will enhance the possibility of efficient map-based cloning in hexaploid wheat. The entire data set presented in this study has been made publicly available via the provision of supplementary data sets and an interactive website (http://www.cerealsdb.uk.net/), to make this resource as accessible and useful as possible. These new co-dominant wheat SNP-based markers will be useful on a number of genotyping platforms and germplasm collections and hence should be a powerful new tool for wheat breeders and researchers alike. In addition, the pipeline developed here to identify co-dominant SNP markers should be applicable to other polyploid crops where SNP discovery and marker development have previously been challenging (Cordeiro *et al.*, 2006; Trick *et al.*, 2009; Yu *et al.*, 2012).

Experimental procedures

Plant material

Forty-seven wheat varieties were grown for DNA extraction (for details see Data S5). The Avalon × Cadenza doubled-haploid (DH) population was supplied by the John Innes Centre and was developed by Clare Ellerbrook, Liz Sayers and the late Tony Worland as part of a Defra-funded project led by ADAS. The parents were originally chosen (to contrast for canopy architecture traits) by Steve Parker (CSL), Tony Worland and Darren Lovell (Rothamsted Research). The Savannah × Rialto DH population was supplied by Limagrain UK Limited (Woolpit, Suffolk, UK). All plants were grown in pots in a peat-based soil and maintained in a glasshouse at 15–25 °C under a light regime of 16 h light and 8 h dark. Leaf tissues were harvested from 6-week-old plants and immediately frozen on liquid nitrogen and stored at −80 °C until nucleic acid extraction. Genomic DNA was prepared from leaf tissue using a phenol–chloroform extraction method (Sambrook *et al.*, 1989).

Preparation of NimbleGen libraries

The NimbleGen capture array was designed to capture a significant proportion of the wheat exome and was developed using a gene-rich assembly of 454 titanium sequence data from normalised and non-normalised cDNA libraries of Chinese Spring line 42, publically available EST sequences and the NCBI unigene set (Winfield *et al.*, 2012). The resulting assembly was used by NimbleGen to design an array containing 132 605 features with an average length of 426 bp (NimbleGen array reference 100819_Wheat_Hall_-cap_HX1). NimbleGen sequence libraries were prepared for eight wheat varieties (Alchemy, Avalon, Cadenza, Hereward, Rialto, Robigus, Savannah and Xi19) as described by Winfield *et al.* (2012). Post-capture-enriched sequencing libraries were subjected to 110 bp of paired end sequencing on a Illumina Genome

Analyser (GAIIx) using Illumina TruSeq v5 Cluster Generation (Illumina Inc., San Diego, CA) and sequencing reagents following the manufacturers preparation guides for paired end runs (Part 15019435 RevB, Oct2010 and Part 15013595 Rev C, Feb 2011, respectively).

SNP discovery

After pre-processing of reads, where adapter sequences were removed, the data were submitted to a custom pipeline (Winfield *et al.*, 2012). NGS sequences generated from the eight varieties were mapped to the NimbleGen array reference using BWA version 0.5.9-r16 (Li and Durbin, 2009) with a seed length of 32 bases, and the resulting SAM files were used for downstream analysis. Uniquely mapped reads were analysed using a series of custom PERL scripts designed to identify only differences between varieties as opposed to those between each variety and the reference sequence. This enabled the exclusion of homoeologous SNPs (which are not useful markers), which were removed from the SNP discovery pipeline. SNPs were called where there were at least two alternative bases predicted at a reference position. An additional constraint on SNP prediction required each SNP to be represented by two or more independent reads or 2% of all reads examined (whichever was the greater). Only bases that were located at the centre of a three-base window of PHRED quality $\geq$ 20 were included in the analysis. Sequences were discarded if they displayed more than 10% sequence variation from the reference over their length or if they mapped equally well to more than one locus, as the mapping in these situations could be regarded as uncertain. In cases where multiple reads started at the same position in the reference, all but one were ignored to guard against clonal reads being sampled more than once. All NGS data generated for this study will be available at: http://www.cerealsdb.uk.net. In addition, the Illumina fastq files and associated metadata have been uploaded to NCBI Sequence Read Archive (SRA) under the study accession SRP011067. Accession numbers of fastq files for each variety are as follows: Alchemy (SRR417586.1), Avalon (SRR417587.1), Cadenza (SRR417953.1), Hereward (SRR417954.1), Rialto (SRR417955.1), Robigus (SRR418209.1), Savannah (SRR418210.1) and Xi19 (SRR418211.1).

SNP validation

For each putative varietal SNP, two allele-specific forward primers and one common reverse primer (Data S1) were designed (KBioscience, Hoddesdon, UK). Genotyping reactions were performed in a Hydrocycler (KBioscience) in a final volume of 1 μL containing 1 × KASP 1536 Reaction Mix (KBioscience), 0.07 μL assay mix (containing 12 μM each allele-specific forward primer and 30 μM reverse primer) and 10–20 ng genomic DNA. The following cycling conditions were used: 15 min at 94 °C; 10 touchdown cycles of 20 s at 94 °C, 60 s at 65–57 °C (dropping 0.8 °C per cycle); and 26–35 cycles of 20 s at 94 °C, 60 s at 57 °C. Fluorescence detection of the reactions was performed using a Omega Pherastar scanner (BMG LABTECH GmbH, Offenburg, Germany), and the data were analysed using the KlusterCaller 1.1 software (KBioscience).

Genetic map construction

The software programme MapDisto v. 1.7 (Lorieux, 2012) was used to place the SNP markers in the previously established genetic map for Avalon × Cadenza (http://www.wgin.org.uk/resources/MappingPopulation/TAmapping.php). A chi-square test was performed on all loci to test for segregation distortion from the expected 1 : 1 ratio of each allele in a DH population, and any loci showing significant distortion were removed from the data set before constructing the linkage groups. Loci were assembled into linkage groups using likelihood odds (LOD) ratios with a LOD threshold of 6.0 and a maximum recombination frequency threshold of 0.40. The linkage groups were ordered using the likelihoods of different locus-order possibilities and the iterative error removal function (maximum threshold for error probability 0.05) in MapDisto and drawn in MapChart (Voorrips, 2002). The Kosambi mapping function (Kosambi, 1944) was used to calculate map distances (cM) from recombination frequency.

SNP data analysis

Summary statistics (MAF and PIC estimates) were calculated for loci using Powermarker 3.25 software (Liu and Muse, 2005).

Acknowledgements

We are grateful to the Biotechnology and Biological Sciences Research Council, UK, and the Crop Improvement Research Club (CIRC) for providing the funding for this work (awards BB/I003207/1, BB/I017496/1). We are grateful to the Wheat Genetic Improvement Network for making the mapping data relating to the Avalon × Cadenza population public. For further details of the Avalon × Cadenza mapping population, please refer to the Wheat Genetic Improvement Network web site at: http://www.wgin.org.uk/resources/MappingPopulation/TAmapping.php. We also thank Limagrain UK limited for supplying the Savannah × Rialto mapping population and related marker data.

References

Adams, K.L. and Wendel, J.F. (2005) Novel patterns of gene expression in polyploid plants. *Trends Genet.* **21**, 539–543.

Akhunov, E., Nicolet, C. and Dvorak, J. (2009) Single nucleotide polymorphism genotyping in polyploidy wheat with the Illumina GoldenGate assay. *Theor. Appl. Genet.* **119**, 507–517.

Akhunov, E.D., Akhunova, A.R., Anderson, O.D., Anderson, J.A., Blake, N., Clegg, M.T., Coleman-Derr, D., Conley, E.J., Crossman, C.C., Deal, K.R., Dubcovsky, J., Gill, B.S., Gu, Y.Q., Hadam, J., Heo, H., Huo, N., Lazo, G.R., Luo, M.C., Ma, Y.Q., Matthews, D.E., McGuire, P.E., Morrell, P.L., Qualset, C.O., Renfro, J., Tabanao, D., Talbert, L.E., Tian, C., Toleno, D.M., Warburton, M.L., You, F.M., Zhang, W. and Dvorak, J. (2010) Nucleotide diversity maps reveal variation in diversity among wheat genomes and chromosomes. *BMC Genomics*, **11**, 702.

Akhunova, A.R., Matniyazov, R.T., Liang, H. and Akhunov, E.D. (2010) Homoeolog-specific transcriptional bias in allopolyploid wheat. *BMC Genomics*, **11**, 505.

Allen, A.M., Barker, G.L., Berry, S.T., Coghill, J.A., Gwilliam, R., Kirby, S., Robinson, P., Brenchley, R.C., D'Amore, R., McKenzie, N., Waite, D., Hall, A., Bevan, M., Hall, N. and Edwards, K.J. (2011) Transcript-specific, single-nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (*Triticum aestivum* L.). *Plant Biotechnol. J.* **9**, 1086–1099.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Berkman, P.J., Lai, K., Lorenc, M.T. and Edwards, D. (2012) Next generation sequencing applications for wheat crop improvement. *Am. J. Bot.* **99**, 365–371.

Biesecker, L.G., Shianna, K.V. and Mullikin, J.C. (2011) Exome sequencing: the expert view. *Genome Biol.* **12**, 128.

Caldwell, K.S., Dvorak, J., Lagudah, E.S., Akhunov, E., Luo, M.C., Wolters, P. and Powell, W. (2004) Sequence polymorphism in polyploid wheat and their D-genome diploid ancestor. *Genetics*, **167**, 941–947.

Chao, S., Zhang, W., Akhunov, E., Sherman, J., Ma, Y., Luo, M.C. and Dubcovsky, J. (2009) Analysis of gene-derived SNP marker polymorphism in US wheat (*Triticum aestivum* L.) cultivars. *Mol. Breed.* **23**, 23–33.

Chao, S., Dubcovsky, J., Dvorak, J., Luo, M.C., Baenziger, S.P., Matnyazov, R., Clark, D.R., Talbert, L.E., Anderson, J.A., Dreisigacker, S., Glover, K., Chen, J., Campbell, K., Bruckner, P.L., Rudd, J.C., Haley, S., Carver, B.F., Perry, S., Sorrells, M.E. and Akhunov, E.D. (2010) Population- and genome-specific patterns of linkage disequilibrium and SNP variation in spring and winter wheat (*Triticum aestivum* L.). *BMC Genomics*, **11**, 727.

Cordeiro, G.M., Eliott, F., McIntyre, C.L., Casu, R.E. and Henry, R.J. (2006) Characterisation of single nucleotide polymorphisms in sugarcane ESTs. *Theor. Appl. Genet.* **113**, 331–343.

Dixon, J., Braun, H.J. and Crouch, J. (2009) Transitioning wheat research to serve the future needs of the developing world. In: *Wheat Facts and Futures* (Dixon, J., Braun, H.J. and Kosina, P., eds), pp. 1–19. Mexico: CIMMYT.

Dubcovsky, J. and Dvorak, J. (2007) Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science*, **316**, 1862–1866.

Haudry, A., Cenci, A., Ravel, C., Bataillon, T., Brunel, D., Poncet, C., Hochu, I., Poirier, S., Santoni, S., Glémin, S. and David, J. (2007) Grinding up wheat: a massive loss of nucleotide diversity since domestication. *Mol. Biol. Evol.* **24**, 1506–1517.

Haun, W.J., Hyten, D.L., Xu, W.W., Gerhardt, D.J., Albert, T.J., Richmond, T., Jeddeloh, J.A., Jia, G., Springer, N.M., Vance, C.P. and Stupar, R.M. (2011) The composition and origins of genomic variation among individuals of the soybean reference cultivar Williams 82. *Plant Physiol.* **155**, 645–655.

Kaur, S., Francki, M.G. and Forster, J.W. (2012) Identification, characterization and interpretation of single-nucleotide sequence variation in allopolyploid crop species. *Plant Biotechnol. J.* **10**, 125–138.

Kosambi, D.D. (1944) The estimation of map distances from recombination values. *Ann. Eugen.* **12**, 172–175.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Liu, K. and Muse, S.V. (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics*, **21**, 2128–2129.

Liu, B., Xu, C., Zhao, N., Qi, B., Kimatu, J.N., Pang, J. and Han, F. (2009) Rapid genomic changes in polyploid wheat and related species: implications for genome evolution and genetic improvement. *J. Genet. Genomics*, **36**, 519–528.

Lorieux, M. (2012) MapDisto: fast and efficient computation of genetic linkage maps. *Mol. Breeding*, **30**, 1231–1235.

Paux, E., Faure, S., Choulet, F., Roger, D., Gauthier, V., Martinant, J.P., Sourdille, P., Balfourier, F., Le Paslier, M.C., Chauveau, A., Cakir, M., Gandon, B. and Feuillet, C. (2010) Insertion site-based polymorphism markers open new perspectives for genome saturation and marker-assisted selection in wheat. *Plant Biotechnol. J.* **8**, 196–210.

Paux, E., Sourdille, P., Mackay, I. and Feuillet, C. (2011) Sequence-based marker development in wheat: advances and applications to breeding. *Biotechnol. Adv.* http://dx.doi.org/10.1016/j.bbr.2011.03.031.

Reynolds, M., Foulkes, M.J., Slafer, G.A., Berry, P., Parry, M.A.J., Snape, J.W. and Angus, W.J. (2009) Raising yield potential in wheat. *J. Exp. Bot.* **60**, 1899–1918.

Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual*, 2nd edn. Cold Spring Harbor: Cold Spring Harbor Laboratory Press.

Swanson-Wagner, R.A., Eichten, S.R., Kumari, S., Tiffin, P., Stein, J.C., Ware, D. and Springer, N.M. (2010) Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Res.* **20**, 1689–1699.

Trick, M., Long, Y., Meng, J. and Bancroft, I. (2009) Single nucleotide polymorphism (SNP) discovery in the polyploid *Brassica napus* using Solexa transcriptome sequencing. *Plant Biotechnol. J.* **7**, 334–346.

Trick, M., Adamski, N., Mugford, S.G., Jiang, C., Febrer, M. and Uauy, C. (2012) Combining SNP discovery from next-generation sequencing data with bulked segregant analysis (BSA) to fine-map genes in polyploid wheat. *BMC Plant Biol.* **12**, 1–14.

Voorrips, R.E. (2002) MapChart: software for the graphical presentation of linkage maps and QTLs. *J. Hered.* **93**, 77–78.

Winfield, M.O., Wilkinson, P.A., Allen, A.M., Barker, G.L.A., Coghill, J.A., Burridge, A., Hall, A., Brenchley, R.C., D'Amore, R., Hall, N., Bevan, M., Richmond, T., Gerhardt, D.J., Jeddeloh, J.A. and Edwards, K.J. (2012) Targeted re-sequencing of the allohexaploid wheat exome. *Plant Biotechnol. J.* **10**, 733–742.

Yu, J.Z., Kohel, R.J., Fang, D.D., Cho, J., Van Deynze, A., Ulloa, M., Hoffman, S.M., Pepper, A.E., Stelly, D.M., Jenkins, J.N., Saha, S., Kumpatla, S.P., Shah, M.R., Hugie, W.V. and Percy, R.G. (2012) A high-density simple sequence repeat and single nucleotide polymorphism genetic map of the tetraploid cotton genome. *Genes Genomes Genetics*, **2**, 43–58.

Supporting information

Additional Supporting information may be found in the online version of this article:

Data S1 KASPar assay details for 1190 SNPs designed as part of this study.

Data S2 Genotyping results of a panel of 47 wheat varieties screened with 1138 KASPar assays.

Data S3 KASPar assay details for all assays, including genetic map position.

Data S4 Mapping data for all assays mapped on the Avalon × Cadenza and Savannah × Rialto cross.

Data S5 Details of the 47 wheat varieties used in this study.