



# Development of a next generation SNP genotyping array for wheat

Amanda J. Burridge<sup>1,\*</sup> , Mark Winfield<sup>1</sup> , Alexandra Przewieslik-Allen<sup>1</sup>, Keith J. Edwards<sup>1</sup>, Imteaz Siddique<sup>2</sup>, Ruth Barral-Arca<sup>2</sup>, Simon Griffiths<sup>3</sup>, Shifeng Cheng<sup>4</sup>, Zejian Huang<sup>4</sup>, Cong Feng<sup>4</sup>, Susanne Dreisigacker<sup>5</sup>, Alison R. Bentley<sup>6</sup>, Gina Brown-Guedira<sup>7</sup> and Gary L. Barker<sup>1</sup>

<sup>1</sup>School of Biological Sciences, University of Bristol, Bristol, UK

<sup>2</sup>Thermo Fisher Scientific, 3450 Central Expressway, Santa Clara, CA, USA

<sup>3</sup>John Innes Centre, Norwich Research Park, Norwich, UK

<sup>4</sup>Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China

<sup>5</sup>International Maize and Wheat Improvement Center (CIMMYT), Texcoco, Mexico

<sup>6</sup>NIAB, Cambridge, UK

<sup>7</sup>Plant Science Research Unit, USDA Agricultural Research Service, Raleigh, NC, USA

Received 7 December 2023;

revised 5 March 2024;

accepted 6 March 2024.

\*Correspondence (email [amanda.burridge@bristol.ac.uk](mailto:amanda.burridge@bristol.ac.uk))

## Summary

High-throughput genotyping arrays have provided a cost-effective, reliable and interoperable system for genotyping hexaploid wheat and its relatives. Existing, highly cited arrays including our 35K Wheat Breeder's array and the Illumina 90K array were designed based on a limited amount of varietal sequence diversity and with imperfect knowledge of SNP positions. Recent progress in wheat sequencing has given us access to a vast pool of SNP diversity, whilst technological improvements have allowed us to fit significantly more probes onto a 384-well format Axiom array than previously possible. Here we describe a novel Axiom genotyping array, the '*Triticum aestivum* Next Generation' array (TaNG), largely derived from whole genome skim sequencing of 204 elite wheat lines and 111 wheat landraces taken from the Watkins 'Core Collection'. We used a novel haplotype optimization approach to select SNPs with the highest combined varietal discrimination and a design iteration step to test and replace SNPs which failed to convert to reliable markers. The final design with 43 372 SNPs contains a combination of haplotype-optimized novel SNPs and legacy cross-platform markers. We show that this design has an improved distribution of SNPs compared to previous arrays and can be used to generate genetic maps with a significantly higher number of distinct bins than our previous array. We also demonstrate the improved performance of TaNGv1.1 for Genome-wide association studies (GWAS) and its utility for Copy Number Variation (CNV) analysis. The array is commercially available with supporting marker annotations and initial genotyping results freely available.

**Keywords:** genotyping, wheat, *Triticum aestivum*, breeding, Axiom array, single nucleotide polymorphism.

## Introduction

Single Nucleotide Polymorphism genotyping arrays (SNP arrays) play an important role in advancing studies of genetic variation in both animal (Chen *et al.*, 2014a) and plant populations (Bassil *et al.*, 2015; Koning-Boucoiran *et al.*, 2015; van Geest *et al.*, 2017). They allow the identification, and analysis of up to hundreds of thousands of SNPs in a single assay providing a high-throughput and cost-effective way to analyse genetic diversity. As such, they have been widely used to generate genetic linkage maps, study evolutionary relationships, unravel functional genomics and support conservation efforts. They have also proved to be highly valuable in breeding programmes, being used for Genomic Selection (Gebremedhin *et al.*, 2024; Kang *et al.*, 2023) to enable prediction and evaluation of quantitative traits, for marker assisted selection (MAS) (Arruda *et al.*, 2016; Thomson, 2014), genome-wide association studies (GWAS) (Balagué-Dobón *et al.*, 2022; McCouch *et al.*, 2016; Negro *et al.*, 2019; Yu *et al.*, 2023) and the mapping of QTL (Stadlmeir *et al.*, 2018; Xu *et al.*, 2017). Their power and utility are

evidenced by the large number of arrays available for crop species (strawberry – Verma *et al.*, 2017; rose – Koning-Boucoiran *et al.*, 2015; chrysanthemum – van Geest *et al.*, 2017; potato – Vos *et al.*, 2015; rice – Chen *et al.*, 2014b; Daware *et al.*, 2023; Kim *et al.*, 2022; maize – Unterseer *et al.*, 2014).

Genotyping arrays play a critical role in the genotyping of hexaploid bread wheat allowing researchers to rapidly screen wheat varieties, identify genetic variants associated with important traits and develop markers for use in breeding programmes. For wheat, several SNP arrays have been developed (Allen *et al.*, 2017; Rimbart *et al.*, 2018; Soleimani *et al.*, 2020; Sun *et al.*, 2020; Wang *et al.*, 2014; Winfield *et al.*, 2016). These arrays contain a large number of SNPs and have been demonstrated to be effective tools for linkage analysis, QTL mapping of important traits and genome-wide association analysis (Allen *et al.*, 2017; Bourke *et al.*, 2018; Vukosavljev *et al.*, 2016).

However, previously developed genotyping arrays for wheat suffer from uneven marker distribution and marker redundancy due to linkage disequilibrium (LD). Uneven marker distribution

means that regions of the genome being over- or underrepresented. This can lead to bias in the results and limit the ability to accurately detect genetic variants in certain regions of the genome. This is particularly problematic for bread wheat, which has a large and complex hexaploid genome with significant structural variation. In addition to these technical limitations, older genotyping arrays are also limited by the genetic diversity of the populations used to develop them. Bread wheat is a highly diverse crop with significant genetic variation both between and within different populations. Therefore, genotyping arrays developed using a limited set of wheat lines may not capture the full range of genetic diversity present in the crop. As a consequence of these limitations, scientists and breeders have called for a new generation of wheat genotyping arrays that overcome these technical and biological challenges, provide a more comprehensive view of the genome and capture the full range of genetic diversity present in bread wheat's pangenome. Here, we describe the development of a new SNP genotyping array for wheat, the TaNG Array, that has been designed to overcome several of these issues and, thus, provide a more comprehensive coverage of the genome than previous versions.

## Results

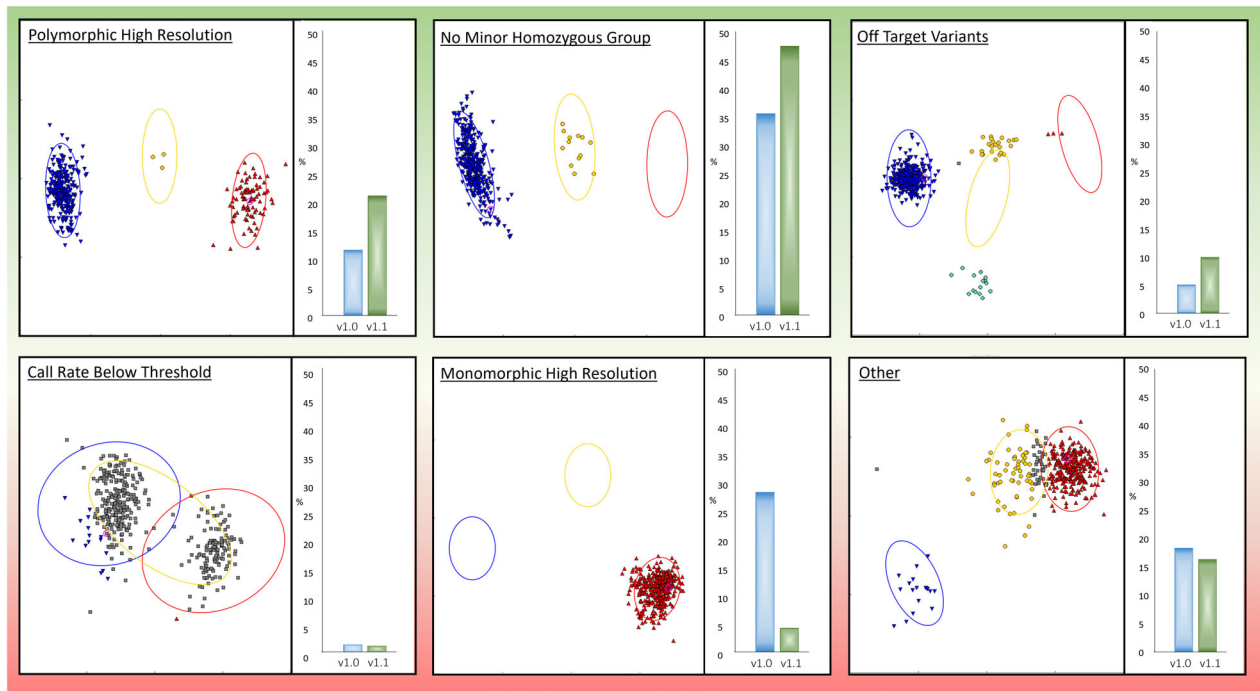
### Marker testing

The SNP markers selected for inclusion on the array were from skim sequence data from 315 wheat accessions (204 elites and 111 landraces – Data S1) and from existing genotyping arrays (see Methods). Markers taken from the Axiom™ Wheat HD Genotyping Array (Winfield *et al.*, 2016) – hereafter referred to as the 820K Axiom™ Wheat HD Genotyping Array – and the Axiom™

35K Wheat Breeder's Genotyping Array (Allen *et al.*, 2017) – hereafter referred to as the 35K Array – were entirely exonic, whilst those derived from sequence were intronic, exonic and intergenic. As a panel, the markers were evenly distributed throughout the genome based on positions relative to IWGSC RefSeq v1.0 (International Wheat Genome Sequencing Consortium (IWGSC), 2018).

The initial array design (v1.0) was screened using a standard collection of 182 elite cultivars and landraces (Data S4). The sample call rate ranged from 94.5% to 98.8%. Based on their cluster patterns, probes were classified into the following six categories: Poly High Resolution; No Minor Homozygous; Off-Target Variant; Call Rate Below Threshold; Monomorphic High Resolution and Other (Figure 1; Table 1). The first three categories are considered most useful as they generate accurate polymorphic genotype calls. Of the 44 258 probes on the initial array, 23 068 (52%) fell into the three useful categories (Table 1). Approximately 28% were monomorphic; that is, they generated a strong signal to indicate the target sequence was present, but no polymorphism was detected. Of these monomorphic markers, 93% were derived from skim sequence data rather than sourced from existing genotyping arrays, indicating that these SNPs failed to convert into useful Axiom markers.

Of the accessions used for screening TaNG Array v1.0, 144 were also present in the original skim sequencing panel (Data S1) thus allowing direct comparison to be made between genotype calls on the two platforms. Given the sequencing-derived genotypes for the 144 varieties, only four of the 12 490 markers reporting monomorphisms were predicted to be monomorphic, and only 112 of these apparently monomorphic markers were expected to have less than 10 instances of the minor allele.



**Figure 1** The percentage of probes in each probe quality category by array type. TaNG v1.1 has an increased ratio of the 'high quality' categories (Polymorphic High Resolution, No Minor Homozygous Group and Off-Target Variants) and a decreased ratio of probes in the 'low quality' categories (Call Rate below Threshold, Monomorphic High Resolution and Other).

**Table 1** Marker quality categories as designated by Axiom Analysis Suite for TaNG v1.0 and TaNG v1.1

Category	TaNG v1.0		TaNG v1.1	
	Count	Percentage	Count	Percentage
Poly high resolution	5075	11.5	9117	21.0
No minor homozygous	15 746	35.6	20 580	47.4
Off-target variant	2247	5.1	4300	9.9
Call rate below threshold	573	1.3	459	1.1
Monomorphic high resolution	12 490	28.2	1827	4.2
Other	8127	18.4	7090	16.3
Sum	44 258		43 373	

The categories considered 'high quality' are: 'Polymorphic High Resolution', 'No Minor Homozygous' and 'Off-Target Variant'. The remaining categories: 'Call Rate Below Threshold', 'Monomorphic High Resolution' and 'Other' are considered to be a lower quality genotype call.

### Design optimization

Due to the large number of monomorphic probes on the first iteration of the TaNG array (version 1.0), the array was redesigned. Monomorphic probes were replaced with probes from the Axiom™ Wheat HD Genotyping Array proven to be polymorphic; these replacement probes were selected by re-running the marker optimization algorithm with monomorphic markers excluded from the input file, whilst other markers were retained as they had performed well in screening (Data S3). Additional markers were integrated into the optimized design by analysing combining genotyping data from our existing 820K Wheat HD Genotyping Array with that derived from TaNG v1.0, where the same samples had been run on both platforms. This improved array, designated TaNG v1.1, was screened against an extended collection of elite cultivars, landraces and other *Triticum* accessions (Data S4). The sample call rate ranged from 84% to 99.8%. Compared to the initial implementation of the array, TaNG v1.1 showed an increased number of markers in each of the useful probe quality categories and a decreased number in each of the less useful categories (Table 1). Therefore, all further study was based on TaNG v1.1 and, thus, from this point forward, all results and discussion refer to comparisons between the 35K Breeders Array version v1.1 of the new array.

### Marker distribution across chromosomes

The TaNG v1.1 Array has more markers in total than the 35K Array (43 373 vs. 35 143) and, for all chromosomes except 1D and 2D, there are more markers assigned to each chromosome (Table 2). Furthermore, markers are more evenly distributed between the 21 wheat chromosomes and the number of markers per chromosome better reflects chromosome size (Data S5).

For broad scale distribution of markers, the chromosomes, regardless of their reported lengths (Table 2) were divided into 20 equally sized bins and the number of markers in each bin totalled and plotted (Figure 2a); markers on TaNG v1.1 are more evenly distributed across the chromosomes than those on the 35K Array. That is, there is neither a bias in marker number towards the telomeres nor a relative paucity of markers across the centromeres. At a smaller scale, marker distribution was determined by dividing chromosomes into 10 Mb bins and counting the number of markers in each (Figure 2b). The number of markers in the

**Table 2** Chromosome lengths in variety Chinese Spring (based on IWGSC v1.0 assembly) versus number of markers present on the 35K Wheat Breeders and TaNG v 1.1 Arrays

Chromosome	Length (bp)	Number of markers	
		35K Array	TaNG Array v1.1
1A	594 102 056	1566	2087
2A	780 798 557	1718	2295
3A	750 843 639	1486	2054
4A	744 588 157	1130	2223
5A	709 773 743	1604	2016
6A	618 079 260	1196	1897
7A	736 706 236	1678	2302
1B	689 851 870	2178	2343
2B	801 256 715	2132	2425
3B	830 829 764	1751	2429
4B	673 617 499	1135	1883
5B	713 149 757	1804	2230
6B	720 988 478	1706	2027
7B	750 620 385	1652	1890
1D	495 453 186	2045	2025
2D	651 852 609	2278	2238
3D	615 552 423	1796	2004
4D	509 857 067	872	1540
5D	566 080 677	1621	1857
6D	473 592 718	1222	1593
7D	638 686 055	1607	2015
Sum	14 066 280 851	34 177	43 373

10 Mb bins is much less variable and there are no extreme outliers with very low or very high numbers of markers. For example, on the 35K Array, the region 240–290 Mb on chromosome 4A is represented by only 6 markers; on TaNG v1.1 the same region is represented by 125 markers (Data S5). Indeed, on chromosomes 3A and 4A the 35K Array has no markers assigned at all to a small number of 10 Mb bins (Data S5).

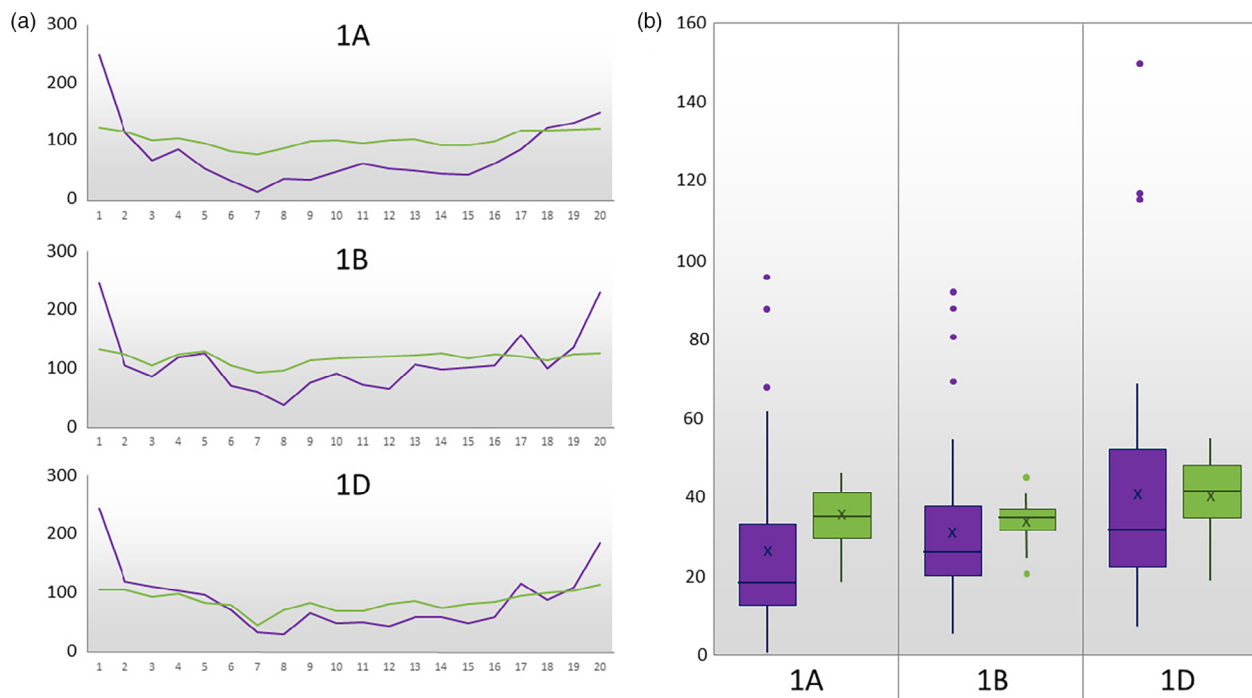
Although there is a relatively strong relationship between chromosome length and the number of markers assigned to that chromosome, the density of markers is relatively higher on the smaller chromosomes than on the larger chromosomes. The mean number of markers per 1 Mb on the A, B and D genome is 3.0, 2.9 and 3.4, respectively, highlighting the efforts made to improve D genome marker coverage compared to previous array designs.

### Source of SNP variation

The older 35K Array SNPs were derived from exome-capture sequencing, based on a set of genes de novo assembled genes from cDNA sequencing of Chinese Spring (Winfield *et al.*, 2012). Unsurprisingly, a variant effect prediction analysis annotated over 86% of these SNPs to be within or immediately adjacent to coding regions (Data S5). In contrast, more than 27% of the SNPs on the new TaNG1.1 array were annotated as having an intergenic origin (Data S5).

### Replicate testing for reproducibility

As a measure of reproducibility between technical replicates, the accessions 'Paragon' and 'Cadenza' were genotyped as four



**Figure 2** The physical distribution of chromosome 1 markers on the TaNG v1.1 (green) and 35K Wheat Breeder's Arrays (purple). (a) Number of markers in each of 20 bins spanning the chromosome. (b) Box and whisker plots of the number of markers per 10 Mb bin across the chromosome. There is a greater number of markers on the TaNG v1.1 Array (green boxes) and these are more evenly distributed than on the 35K Array (purple boxes). See Data S5 for plots of all chromosomes.

aliquots from the same DNA extraction (Data S4). The genotype correlation was extremely high with 99.86% and 99.49% correlation for 'Paragon' and 'Cadenza', respectively. This represents a less than 1% technical error rate. The genotyping errors were predominantly a mis-call between the homozygous (AA, BB) and heterozygous (AB) states with only two and four probes presenting a change in homozygous call ('hom-hom mis-call') in 'Paragon' and 'Cadenza', respectively. Only six markers presented a genotyping error between both accessions, each on a different chromosome suggesting that the 1% error rate was random in nature (Data S4).

### Genetic map construction

The genetic location of markers and performance of the haplotype optimization marker selection method was tested by generating genetic maps using three mapping populations. These were the Avalon × Cadenza (AxC) and Oakley × Gatsby (OxG) double haploid populations, and Apogee × Paragon (AxP) produced by single seed descent to the F5 generation. For the three maps, AxC, OxG and AxP, 10 113, 7734 and 4673 markers were assigned to linkage groups, respectively (Data S6).

The AxC and AxP populations had previously been genotyped using the Axiom™ Wheat Breeder's Genotyping Array, making it possible to compare the position of markers on the arrays. The TaNG AxC genetic map consisted of 10 113 markers with 1652 unique locations, whilst the Wheat Breeder's AxC genetic map consisted of 7237 markers with 1082 unique locations (Figure 3, Data S6). The TaNG AxP genetic map consisted of 4673 markers with 1984 unique locations, whilst the Wheat Breeder's AxC genetic map consisted of only 2997 markers with 1519 unique locations. The TaNG array both increased the number of markers

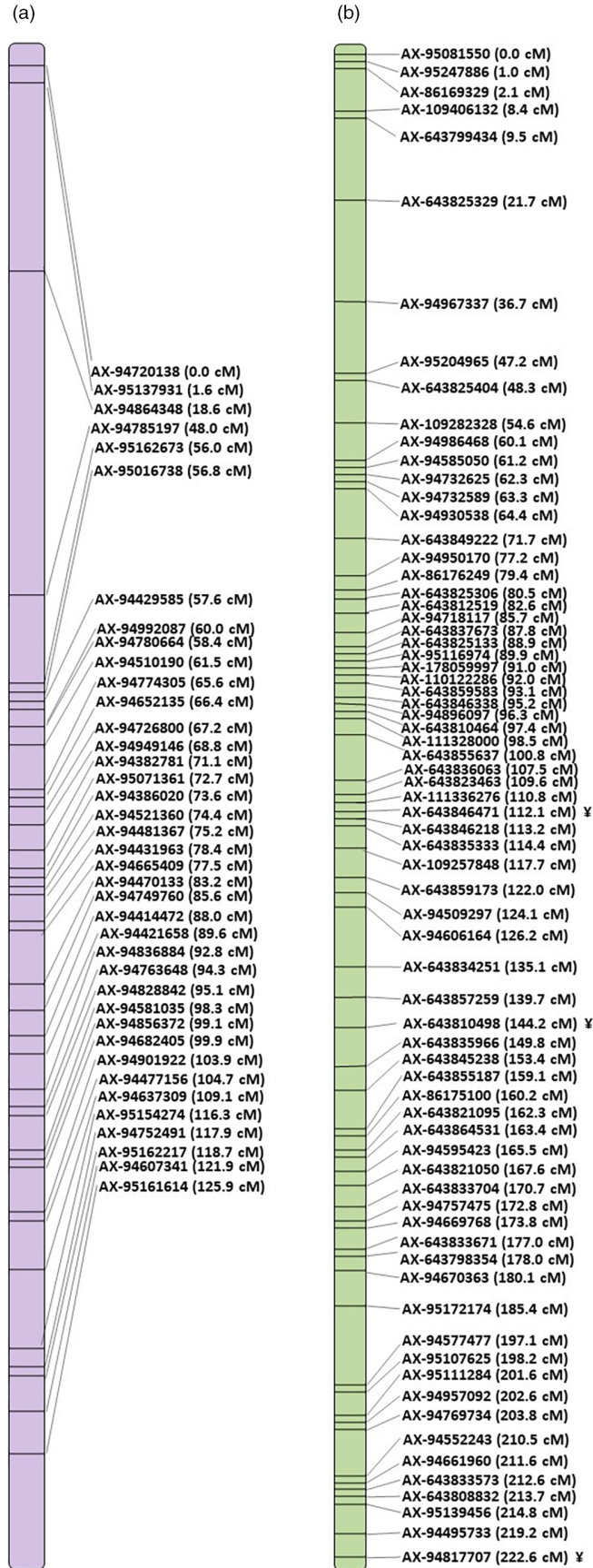
and the number of unique positions for both the AxC and AxP genetic maps. For all chromosomes the number of SNPs and the chromosome length (cM) was greater using the TaNG v1.1 array than the maps previously constructed using data from the Axiom™ Wheat Breeder's Genotyping Array. Although restricted by the limits of recombination of the population, the new array gives a greater number of more evenly spaced markers for all three populations.

The genetic map positions of markers from all three genetic maps were compared to the physical assignment based upon alignment of sequences to the Chinese Spring, IWGSC v1.0 reference assembly: physical assignment based on alignment to IWGSC assembly v1.0; Avalon × Cadenza map; Apogee × Paragon map; Oakley × Gatsby map. A marker was assigned a consensus chromosome only when at least two of the assignments were the same. In total, 12 981 markers were assigned a consensus chromosome (Data S3).

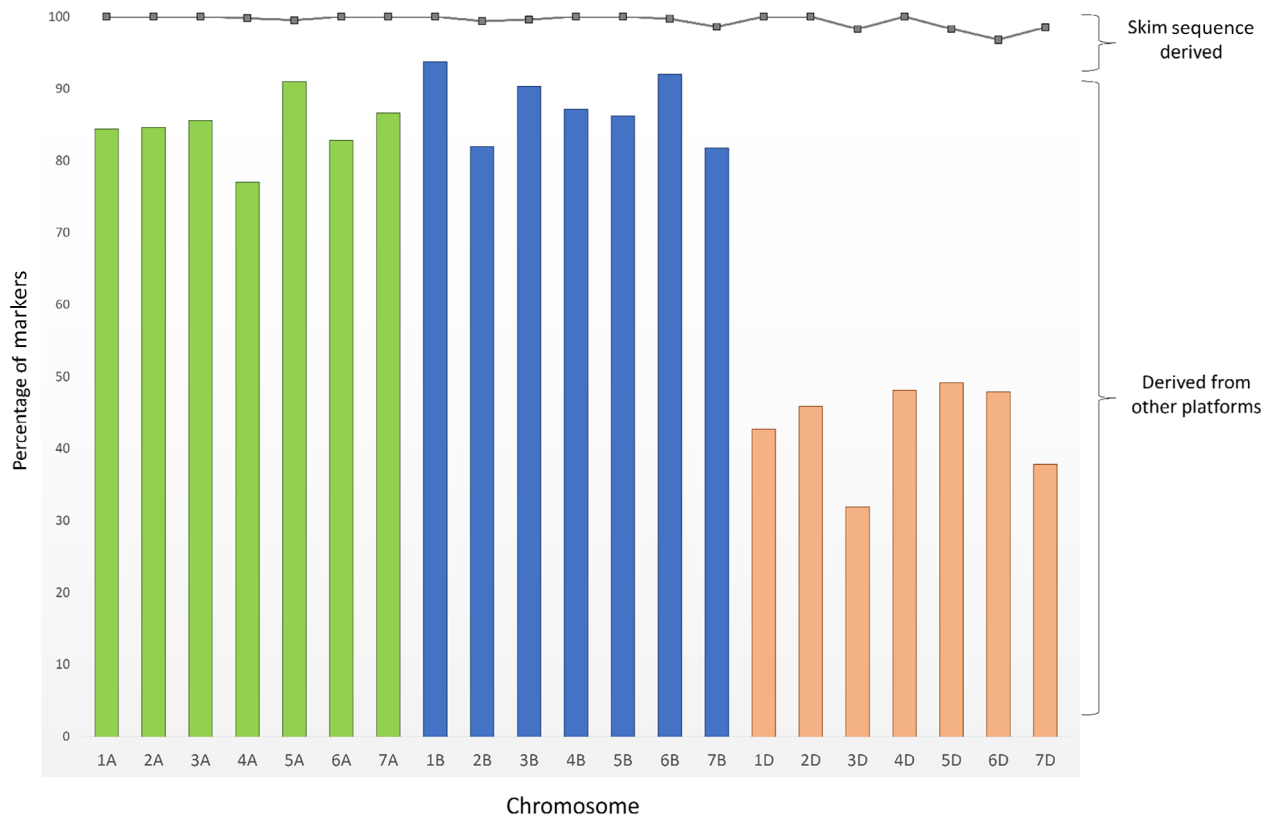
The comparison showed that the physical assignment based on skim sequence data was close to 100% accurate whilst that derived from previous platforms was less reliable, especially for D genome markers (Figure 4). There was good correlation between physical position of the markers and the cM position of the bin to which they were assigned. However, physical assignment for markers taken from previous platforms (820K Array, 35K Array and DArT marker) were more variable with an average concordance of 85% for A and B genome markers but as low as only 40% for D genome markers.

### CNV analysis

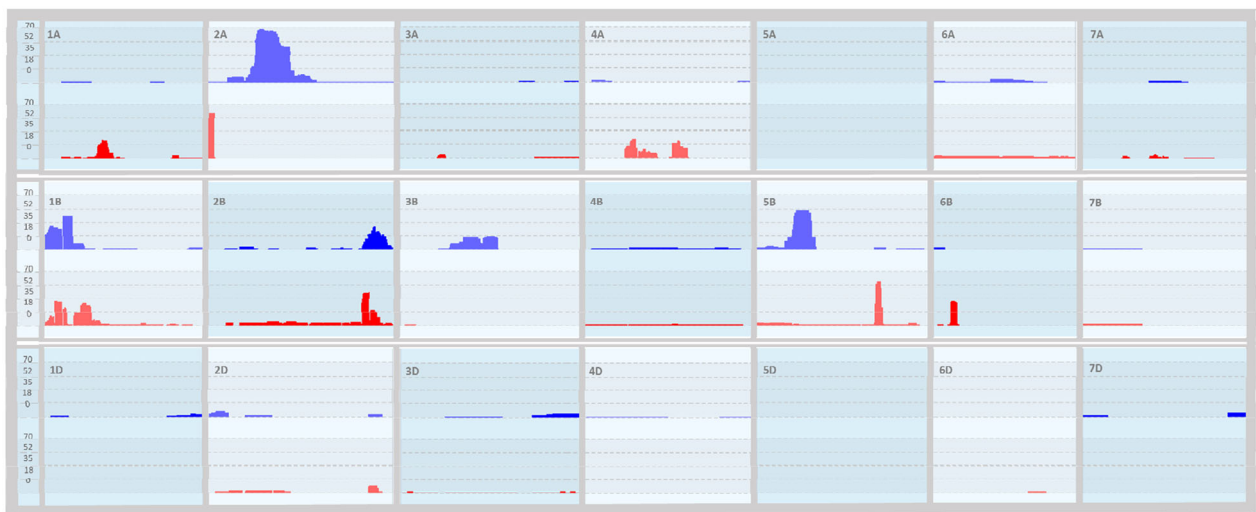
Copy number events were observed across each chromosome with the exception of 4D, 5A and 5D (Figure 5; Data S7). The



**Figure 3** Comparison of Avalon x Cadenza genetic maps for chromosome 1A using (a) 35K Array map data and (b) TaNG v1.1 map data, showing the distribution and density of markers.



**Figure 4** Comparison of chromosome location for markers based on physical and consensus assignment. Only markers with a consensus chromosome assignment are shown. Markers with physical positions derived from previous genotyping platforms (820K Array, 35K Array and DARt markers) are represented by bars. Markers derived from skim sequence data are represented by a line plot.



**Figure 5** Copy Number Variance (CNV) frequency histogram for all samples genotyped with the TaNG v1.1 array in the initial screening (Data S4) across all chromosomes. Regions of copy number gain are displayed in the top track (blue) and regions of copy number loss are displayed on the bottom track (red) for each chromosome. Start and stop positions of each event are listed in Data S7.

number of CNV varied per accession, some reported none whilst 12 were detected in the *T. aestivum* accession 'Captor' and 23 in the wild relative '*T. kiharae*'. The length of detected CNV regions ranged from 8.7 to 780 Mb with more reduced CNV losses than gains. Multiple regions are present with common variations across the screened

accessions. Large regions of variance are present on 2A (Loss: 479 492 – 24 414 560 in 50 accessions; Gain: 202 646 955 – 358 094 640 in 36 accessions), 5B (Gain: 149 551 898 – 249 788 187 in 32 accessions; Loss: 490 000 000 – 520 000 000 in 50 accessions) and 1B (Gain and Loss: 1 203 929 – 123 969 113) (Figure 5; Data S7).

## Diverse material screening

As the initial array screening was performed with a mostly European collection of elite cultivars and landraces (Data S4), additional genotyping was also performed using different sources of material. A collection of USA material grouped by geographic origin was genotyped as the initial screening panel (Data S4). The marker call rate was consistently high with 42 476 markers (98%) generating a call in 90% of samples. The performance category of markers was also high with only 3746 probes (8.6%) found to be monomorphic across the dataset. The genotype data clearly distinguished the USA material by region (Figure 6) with a clear separation of North region and East region germplasm.

A collection of 81 wheat wild relatives including *Aegilops*, *Amblyopyrum*, *Secale*, *Thinopyrum* and other *Triticum* species were also genotyped alongside 50 Durum (*T. turgidum* ssp. *durum*) landrace accessions (Data S4) to examine the suitability of the array for genotyping pre-breeding wild relative material alongside *T. aestivum*. Whilst not all samples could hybridize, 34 588 markers (80%) generated a genotype call across at least 90% of the samples. The majority of markers clustered unclearly with the 'Other' performance category (18 784; 43%) but very few markers were monomorphic (4008; 9.2%).

## GWAS analysis

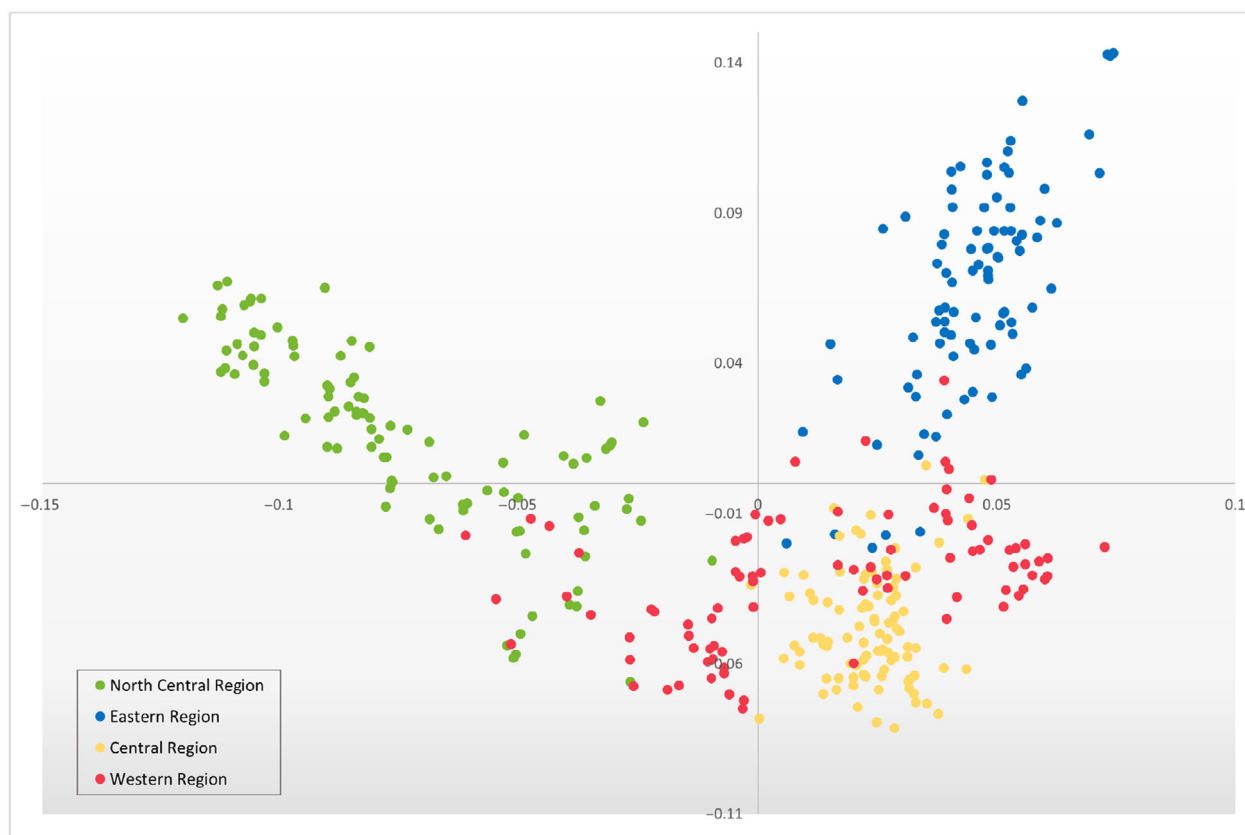
As a test of performance for the optimized SNP selection, three traits were selected for GWAS analysis: Heading Date; Response to Leaf Rust; Response to Stem Rust. Whilst the previous 35k

Breeders array was unable to identify a significant QTL for any of these traits, the 43K TaNG v1.1 array was able to identify QTL that favourably compared to the entire 10 million SNP panel generated from whole genome sequencing (Figure 7).

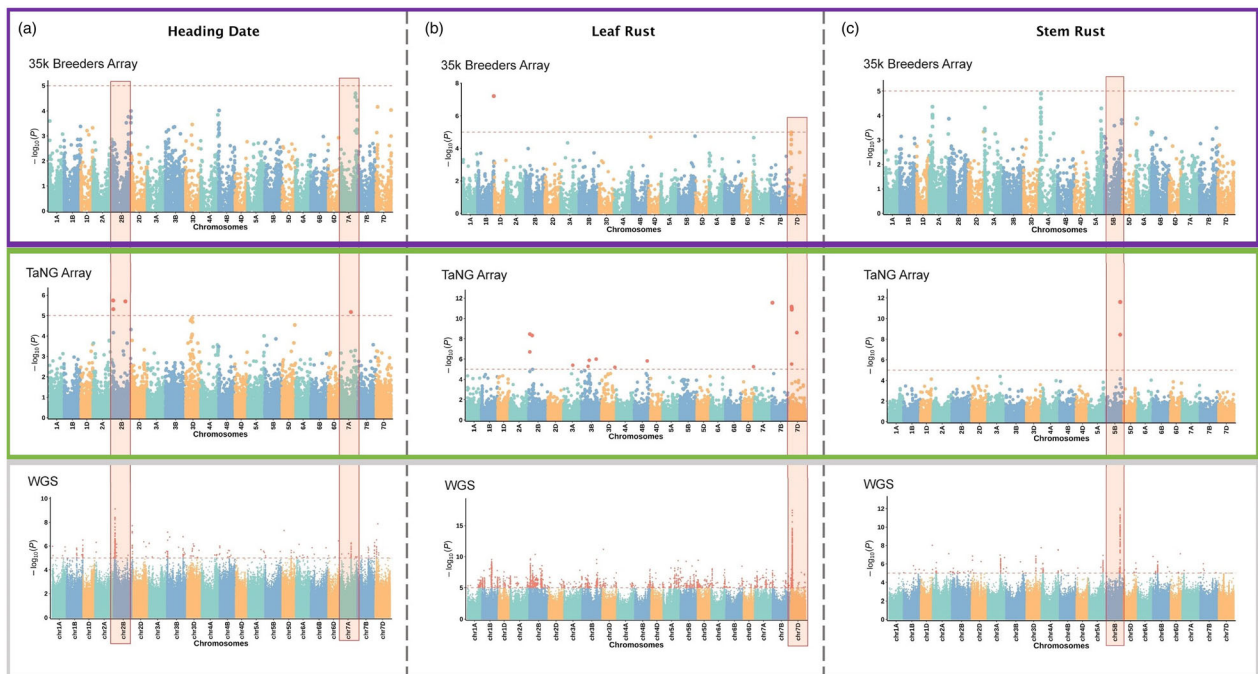
## Discussion

### Array justification

The existing 35K Axiom™ wheat breeder's genotyping array designed in 2011 has been widely used by academics and breeding companies with 288 citations in research areas recently spanning pathogen resistance (Grover *et al.*, 2022; Nannuru *et al.*, 2022), yield (Sheoran *et al.*, 2022), grain nutrient quality (Rathan *et al.*, 2022) and grain architecture (Kumari *et al.*, 2023). Whilst a valuable tool, improvements to technology and our understanding has made it possible to improve upon the design in several ways. The most noticeable difference is the source of SNPs. In the 820K Axiom™ Wheat HD Genotyping Array (Winfield *et al.*, 2016), the 35K Axiom™ wheat breeder's genotyping array (Allen *et al.*, 2017) and the 90k wheat iSelect array (Wang *et al.*, 2014) used exome-capture sequences as the source of putative SNPs. This resulted in all SNPs being within or very close to genes, resulting in uneven chromosomal distribution, the potential exclusion of rare alleles and a strong ascertainment bias (You *et al.*, 2018). Recent large scale skim sequencing (Cheng *et al.*, 2023) of a highly diverse set of globally significant breeding and landrace accessions made available a new source of SNPs free of ascertainment bias and suitable for global users.



**Figure 6** Principle Component Analysis plot based on the USA material collections coloured by region of origin. Variation in PC1 and PC2 is 5.05 and 3.93, respectively. Samples used in breeding but of non-USA origin were omitted from figure. Genotype data is available in Data S4.



**Figure 7** Genome-wide association study using the previous 35K Breeders Array, the new TaNG v1.1 array and the 10 million SNPs detected in sequenced data for three traits. (a) Heading date, (b) Leaf rust, (c) Stem rust. Identified QTL are highlighted in red.

Prior the release of the IWGSC Chinese Spring reference (IWGSC, 2018), genetic maps were used to obtain marker location information. In the 35k Array this led to significant ascertainment bias due to prioritizing SNPs that could be placed on the Avalon × Cadenza or Rialto × Savanna maps (Allen *et al.*, 2017). For SNPs without polymorphisms between these parents, the locations (physical or genetic) were initially unknown. After the release of the CS reference, the distribution was found to be variable, chromosomes had markers clustered together in some regions with large gaps in between with no marker coverage. In the areas of marker clusters, many would be in linkage disequilibrium (LD) meaning that the SNPs would be inherited together more often than expected by chance. LD represent a duplication of effort, as additional markers are no more informative. With a gold standard genome sequence now available and the physical positions of all SNPs now known, more careful consideration can be made with regards to the marker distribution.

### Two-Step design process for improved performance

The original draft TaNG v1.0 array design showed promising performance for many of the markers but with 12 490 (28.2%) markers that were found to be monomorphic. Whilst the 'Call Rate Below Threshold' and 'Other' categories may be considered less useful categories, the designation of these categories depends on the samples used. Probes with an 'Other' category in one sample set may have a clear genotype clustering in another sample set. However, as probes were designed to be polymorphic within the test set, monomorphic calls were an indication of marker design failure. This high-level of marker failure is not unexpected when converting sequence-derived SNPs to markers in a polyploid and was observed with our first high wheat axiom array (Winfield *et al.*, 2016) Because consistently monomorphic or failed probes are of no value, a two-step approach of screening

followed by redesign was used. The final design (TaNG v1.1) was produced by combining genotyping results from our original 820K Axiom™ Wheat HD Genotyping Array (Winfield *et al.*, 2016) with those from TaNG v1.0. The SNP optimization algorithm used in the initial design was re-applied to this combined dataset for marker selection to ensure that the replacement markers were fully integrated with the new design. The resulting TaNG v1.1 array had a decreased ratio of probes in all of the 'low quality' categories (Figure 1) and an improvement in D genome coverage (Data S5).

The use of a two-step design method has been employed in the design of other genotyping arrays such as maize (Unterseer *et al.*, 2014) and pear (Montanari *et al.*, 2019) to produce a high quality and reproducible array design. The number of monomorphic probes on the initial testing of the TaNG v1.1 array was 4.2%, lower than other commercial Axiom arrays such as Pine (10.1%; Perry *et al.*, 2020), Groundnut (23.7% monomorphic; Pandey *et al.*, 2017), Chickpea Array (45.7%; Roorkiwal *et al.*, 2018) and the original 35K Array (4.5%; Allen *et al.*, 2017). As not all of the accessions used for marker generation were used in the array testing (Data S1 and S4), this value may be lower when additional diverse lines are used.

### Haplotype optimization to combine new and existing probe designs

The bias towards SNPs in genic regions in the previous 820K HD Array and 35K Array resulted in markers being negatively correlated with chromosome length; that is, relative to length there were more markers on the shorter chromosomes than on the longer ones (Allen *et al.*, 2017; Winfield *et al.*, 2016). On the TaNG v 1.1 Array there is a strong positive correlation which in addition to physical distribution, has been optimized for haplotype grouping. We employed a novel selection algorithm to select the optimal combination of SNPs in each 1.5 Mb bin of



each wheat chromosome. Rather than allocating the same number of SNPs to each bin, the SNPs within a bin are minimally correlated with each other to avoid effective duplication. In this way, each SNP has considerably more diagnostic power than those identified at random. The result was a more even physical distribution than the 35K Array and greater diagnostic power even when fewer markers are present per bin (Figure 2). The power of this method was illustrated in use of the array in GWAS (Figure 7). Previously, genotyping arrays have been limited for GWAS applications due to the limited marker density. Even in regions that appear to have good coverage, SNPs in LD provide redundant information about the same genetic variation and bias kinship estimations.

To compliment the sequence-derived SNPs, markers selected from existing arrays by haplotype optimization or public nomination were included. The incorporation of a subset of markers from existing genotyping arrays can maintain continuity and consistency when comparing genetic data across different studies and populations. Combined datasets can enhance statistical power and increase the ability to detect genetic associations without the regeneration of data. In the case of wheat breeding, multi-year studies are not unusual and benefit from the use of consistent marker sets even as genotyping technologies evolve. This approach is already established in medical genotyping arrays such as the Transplant array (Li *et al.*, 2015), Axiom Asia Precision Medicine Research Array and the Axiom Human Genotyping SARs-COV-2 Array (Thermo Fisher Scientific, Santa Clara, CA, USA) which all contain cross-platform markers. More recently, agricultural genotyping arrays such as the Axiom 50K 4Tree array, Axiom 44K Rice and Infinium Apple arrays (Guilbaud *et al.*, 2020; Affymetrix Datasheet P/N GGNO05960 Rev. 1; Howard *et al.*, 2021) have been designed to include markers compatible with previous edition genotyping arrays.

As the previous 35K Array used exome-capture derived sequences for SNP discovery, there were far fewer intergenic SNPs included than has been possible with the TaNG v1.1 array (Data S5). For SNPs contained within genes it has been possible to use existing information to identify 157 which are associated with important traits (Data S3).

### Array features

The TaNG v1.1 array has a stable 1% technical variation which is in line with other Axiom arrays (GenomeWide 6.0 Human array; Hong *et al.*, 2012) and genotyping technologies such as the 1% variation reported using SNP DArTSeq (Alam *et al.*, 2018; Nantongo *et al.*, 2022), 0.5% reported using SeqSNP (Harper *et al.*, 2020) and the 0%–1% variation reported using Infinium (Cai *et al.*, 2017; Pavy *et al.*, 2016; Senthilvel *et al.*, 2019). The nature of the genotyping errors between technical replicates were predominantly hom-het mis-calls. This may be due to the probe partially binding to a secondary homoeologous site, sample contamination or due to difficulties in the genotype calling software in identifying a clear cluster. As all assays have been test-screened with a diverse set of accessions (Data S4) any probes which presented difficulties in genotype calling have been assigned SNP specific priors as described in the methods to ensure consistent calling by the software (Data S3) making the calling of SNPs on the array as accurate as possible.

Genetic maps were constructed using the TaNG v1.1 array and compared to the 35K Array. On average markers were more evenly distributed across the chromosome with a higher number

of unique locations represented and less markers clustered together at the same location (Figure 3, Data S6). Although limited by recombination in the populations this represents a significant improvement in the resolution of the maps and utility of the markers for accurately mapping QTLs and marker assisted selection. In addition, a comparison of genetic map position and physical position on the Chinese Spring v1.0 reference sequence allowed an analysis of the accuracy of the physical position assignment. This revealed that the physical assignment based on skim sequence data is close to 100% accurate. However, the physical assignment of markers derived from existing platforms still require information from mapping populations to help identify the correct homeolog (Shorinola *et al.*, 2022). To aid correct placement of markers, the mapping locations are included in Data S3.

The ability of a genotyping array to perform Copy Number evaluations is limited compared to sequencing methods, but the ease of use and the high-throughput nature allows for insight into sample panels and populations. We used copy number variation (CNV) analysis to characterize the accessions screened. Several common regions of increased (CNV gain) or reduced (CNV loss) signal were observed which could potentially represent deletions, introgressions or repeat regions. Some of these regions are already well documented such as the 1RS introgression from rye on 1BS which is reported to result in variable copy number (Xiong *et al.*, 2023) and the *Ae. ventricosa* introgression on 2A (Gao *et al.*, 2021) which is commonly found in wheat due to the addition of the Lr27 resistance gene. Deletions were also clearly represented by a copy number loss in the genotyping data such as the ph1 deletion on 5B (Figure 5: 5B).

### Supporting genotype and marker data

The bread wheat genome already contains significant genetic variation, and much work is being done to enhance the germplasm with novel alleles from wide crosses. The previous 35k Breeders array had previously been used with wheat wild relative material in a pre-breeding context (Horsnell *et al.*, 2023; Kumar *et al.*, 2020; Wright *et al.*, 2023) and for elite durum wheat cultivars (Ganugi *et al.*, 2021; Kabbaj *et al.*, 2017; Shewry *et al.*, 2023). The genotype calls generated on the TaNG v1.1 array across a diverse set of wheat relative material here (Data S4) illustrate that secondary and tertiary gene pool material may also be genotyped alongside *T. aestivum* accessions. As the primary purpose of the array was the genotyping of *T. aestivum*, we suggest a DQC cut-off of 0.6 to be used for wheat relative material to account for the absence of some reference sequences used to generate the DQC metric. The grouping of diploid, tetraploid and hexaploid material with a wide range of ancestral genomes created a valuable insight the relatives for which the TaNG array may successfully hybridize but for more accurate genotyping study, we suggest that samples be grouped by project before genotype calling.

To support cross-platform projects and better support data sharing, the TaNG array has incorporated SNP probes from other public arrays such as the CIMMYT Wheat 3.9K DArTAG array and the previous 820K HD Wheat array and 35k Breeders array. Further to this, other commercial genotyping platforms have included our probes in the same way. We have compiled these including the 90k iSelect array (Wang *et al.*, 2014) and 660k array (Cui *et al.*, 2017) synonyms with full sequences and known trait associations from literature (Data S3). We believe that together

with the TaNG array, this will be a valuable resource for all researchers working across genotyping platforms.

In essence, SNP genotyping arrays have revolutionized the way researchers and breeders study plant genetics and manipulate traits. They provide a high-throughput and cost-effective way to analyse the genetic makeup of plant populations, enabling more targeted and efficient research and breeding efforts. As technology continues to advance, SNP genotyping arrays will continue to be a cornerstone of plant science, contributing to sustainable agriculture, crop security and our understanding of plant biology.

## Experimental procedures

### Marker selection: Skim sequence sourced probes

The SNP calls generated from skim sequence data from 315 wheat accessions (204 elite wheat lines and 111 wheat landraces taken from the Watkins 'Core Collection' – Data S1) were used as the source of SNPs for haplotype optimization (Cheng *et al.*, 2023). Varieties with  $\geq 1\%$  heterozygous loci were excluded. SNPs were initially filtered to have a maximum of 0.5% heterozygous calls among all varieties, a minimum minor allele frequency of 0.01, a minimum call rate of 0.95 and a minimum mapping quality score of 5000. SNPs with a flanking sequence mapping to more than one genome location in the IWGSC v1.0 Chinese Spring genome assembly using the BWA version 0.7.12-r1039 were removed. Additionally, SNPs were checked by BLAST (blastn v2.6.0+) against the IWGSC v1.0 Chinese Spring genome assembly and those matching multiple locations were excluded. Each chromosome was then divided into 1.5 Mb intervals and up to six SNPs representing the highest combined discriminatory power were selected for each interval (Winfield *et al.*, 2020). The haplotype optimization pipeline is available at <https://github.com/pr0kary0te/GenomeWideSNP-development>.

### Existing marker designs

For cross compatibility, SNPs for which there are existing markers from various platforms were also included in the design. That is, 2528 markers selected for trait association or physical locations were taken from the CIMMYT Wheat 3.9K DARTAG array (<https://excellenceinbreeding.org/toolbox/services/mid-density-genotyping-service>) as were 4220 of the best performing markers from the existing Axiom™ Wheat Breeder's Genotyping Array (Allen *et al.*, 2017). In addition, a public call was made to researchers and wheat breeders to nominate markers from existing arrays which they would like to see included on the TaNG array; this call resulted in 1223 marker nominations (Data S3). The final design also has 936 cross-platform probes with the now discontinued Illumina 90k iSelect array (Wang *et al.*, 2014) and 8232 cross-platform probes with the Wheat 660k Axiom array (Sun *et al.*, 2020).

On an initial screening of an early version of the array (designated TaNG v1.0; Data S2) against a diverse set of 119 elite and 60 landrace accessions, 16 507 SNPs failed to convert to polymorphic SNP assays. These markers were replaced with 14 774 selected from the Axiom™ Wheat HD Genotyping Array (Winfield *et al.*, 2016), to maximize the differentiation of varieties described above. This final, optimized array design, designated TaNG v1.1 (Thermo Fisher catalogue number 551498), contains 43 373 markers (Table 1; Data S3). Some markers may be present on multiple arrays under different names, when this is the case, pseudonyms are given in each column.

### Modification of priors

For consistency, the same Dish-QC probes were used as the 35K Axiom™ wheat breeder's genotyping array for generation of the non-genotype producing DQC sample quality metrics. For probe quality, of the markers on TaNG v1.1, 299 generated clusters that were not correctly identified during allele calling using Applied Biosystems' (Waltham, MA, USA) software package Axiom™ Analysis Suite v5.2.0.65. To ensure the correct genotype call for these alleles, the analysis file was modified with sequence specific priors for the affected markers ('SSP' in Data S3). The modified analysis file designated 'Axiom\_TaNG1\_1.r4' is available from the Thermo Fisher website.

### Genotyping

Genomic DNA from wheat leaf tissue 14 days after germination was prepared as described in BurrIDGE *et al.* (2017) for samples listed in Data S4. Genotyping was performed using 11  $\mu\text{L}$  of 25 ng/ $\mu\text{L}$  DNA in water. Array processing was performed using the GeneTitan system according to the procedure outlined in Axiom™ 2.0 Assay 384HT Array Format Automated Workflow User Guide (Applied Biosystems). Allele calling was performed using Applied Biosystems' software package Axiom™ Analysis Suite v5.2.0.65 using prior file Axiom\_TaNG\_SNP.r1 for the first array design (v1.0) and Axiom\_TaNG1\_1.r4 for the final array design (v1.1). In all cases a Dish QC of 0.8 for *T. aestivum* (Initial Testing, USA Material) datasets and 0.6 for the wild relative genotyping. A sample QC call rate of 80% and 75% was used for *T. aestivum* and wild relative sets, respectively. The SNP QC cut-off for 'Call Rate Below Threshold' was 95%. Comparisons of technical replication was made using markers across all probe quality categories with 'No-call' genotypes omitted from comparison. The TaNG v1.1 Array is available from Thermo Fisher Scientific with catalogue number 551498.

### Copy number variation (CNV)

The CNV analysis and Manhattan plots were generated for all accessions screened on TaNG v1.1 using Axiom™ Analysis Suite v5.2.0.65, with prior file Axiom\_TaNG1\_1.r4 and the annotation file Axiom\_TaNG1\_1.r4.annot.db. No samples were excluded from reference creation. The recommended minimum base lengths and probe numbers for each CNV state were followed from Axiom™ Copy Number Data Analysis Guide (r3 May 2022, MAN0026736).

### SNP effect predictions

SNP effect predictions were made using the Variant Effect Predictor (VEP) hosted on the EnsemblPlants website [http://plants.ensembl.org/Triticum\\_aestivum/Tools/VEP](http://plants.ensembl.org/Triticum_aestivum/Tools/VEP) Release 110 (Martin *et al.*, 2023). The genome selected was that of *Triticum aestivum*. Variant call format (vcf) files were uploaded to the website and the web tool run using default settings.

### Genetic map construction

For the three mapping populations, markers with more than 10% missing data were removed. The remaining markers were tested for significant segregation distortion using a chi-square test. The software program MapDisto v. 1.7 (Lorieux, 2012) was used to assemble the loci into linkage groups using likelihood odds (LOD) ratios with a LOD threshold of 6.0 and a maximum recombination frequency threshold of 0.4. 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. The Kosambi mapping function (Kosambi, 1944) was used to calculate map distances (cM) from recombination frequency. Maps were drawn in MapDisto with bins represented by a single marker.

### Consensus chromosome assignment

Where possible, markers were assigned to a chromosome based on consensus of calls from four different data sets: (i) physical position from BLASTing sequence to IWGSC assembly v1.0; (ii) Avalon × Cadenza genetic map (10 113 markers); (iii) Apogee × Paragon map (4673 markers); (iv) Oakley × Gatsby map (7733 markers). To be assigned a consensus chromosome, the calls from at least two of the data sets had to agree; if a marker had only a physical position or only conflicting calls, it was not assigned a consensus. A comparison was made between the consensus calls and initial physical call to estimate agreement (Data S3); this analysis was performed taking into account the origin of the markers, skim sequence-derived versus acquired from earlier genotyping platforms (820K Array, 35K Array, DArT).

### Genome-wide association study

The GWAS analysis was performed using the Watkins collection accessions and associated phenotype data as described in (Cheng *et al.*, 2023) to compare the core 10M SNPs from the sequenced dataset (Cheng *et al.*, 2023); SNPs from the Axiom™ Wheat Breeder's Genotyping Array (CerealsDB) and those of the TaNG Array v1.1. Extreme outlier values of phenotypic data were removed. Kinship matrix was calculated as the covariate using GEMMA-kin. Based on these, GWAS was performed using GEMMA (v0.98.1) with parameters (gemma-0.98.1-linux-static -miss 0.9 -gk kinship.txt and gemma-0.98.1-linux-static -miss 0.9 -lmm -k kinship.txt). In-house R scripts were used to visualize the results.

### Acknowledgements

The array processing was performed by the Bristol Genomics Facility. We are grateful to the Wheat Genetic Improvement Network for making public the mapping data relating to the Avalon × Cadenza population. This population of doubled-haploid (DH) individuals was developed by Clare Ellerbrook, Liz Sayers and the late Tony Worland (John Innes Centre), as part of a Defra funded project led by ADAS. The parents, having contrasting canopy architectures, were originally chosen by Steve Parker (CSL), Tony Worland and Darren Lovell (Rothamsted Research). We also express gratitude to the Germplasm Resources Unit (GRU) at the John Innes Centre (JIC) for much of the elite, landrace and wheat relative germplasm used to screen the array, to the USDA Germplasm Resource Information Network (GRIN) for much of the USA germplasm and to the NBRP-Wheat gene bank for additional wild relative germplasm. With thanks to Colblindor at [color-blindness.com](http://color-blindness.com) for the Colbis colour blindness tool used to check all figures for accessibility.

### Conflict of interests

IS and RBA work for Thermo Fisher Scientific. All other authors declare no conflict of interests.

### Author contributions

SC, ZH and CF generated the original skim sequence data and SNP identification. GLB designed the haplotype optimization

method. AJB, GLB, MW, SD, APA, SG and KJE analysed data and selected markers. IS and RBA converted sequence designs to array probes. AJB performed laboratory protocols. MW analysed marker distribution data. APA generated mapping data with material provided by ARB and KJE. AJB generated the CNV data. GBG curated the USA accessions and geographical data. CF generated the GWAS data with phenotype data provided by SG. AJB, MW, APA, GLB wrote the manuscript. All authors have read and approved the final text.

### Funding

This study was funded by the Biotechnology and Biological Sciences Research Council through the Designing Future Wheat ISP (BBS/E/C/00010280), Delivering Sustainable Wheat (BB/Y003004/1) and Low Cost Identification of Crop Varieties (BB/T017031/1) awards. The improved (v1.1) design and testing was funded by the Bristol Centre for Agricultural Innovation (BCAI).

### Data availability

The raw skim sequence data was generated by Cheng *et al.* (2023). The supporting genotype calls for all screening is available in Data S4. For non-commercial accessions the germplasm source and accession numbers are listed in Data S3.

### References

- Alam, M., Neal, J., O'Connor, K., Kilian, A. and Topp, B. (2018) Ultra-high-throughput DArTseq-based silicoDArT and SNP markers for genomic studies in macadamia. *PLoS One*, **13**, e0203465.
- Allen, A.M., Winfield, M.O., Burridge, A.J., Downie, R.C., Benbow, H.R., Barker, G.L., Wilkinson, P.A. *et al.* (2017) Characterization of a Wheat Breeders' Array suitable for high-throughput SNP genotyping of global accessions of hexaploid bread wheat (*Triticum aestivum*). *Plant Biotechnol. J.* **15**, 390–401.
- Arruda, M.P., Lipka, A.E., Brown, P.J., Krill, A.M., Thurber, C., Brown-Guedira, G., Dong, Y. *et al.* (2016) Comparing genomic selection and marker-assisted selection for *Fusarium* head blight resistance in wheat (*Triticum aestivum* L.). *Mol. Breed.* **36**, 84.
- Balagué-Dobón, L., Cáceres, A. and González, J.R. (2022) Fully exploiting SNP arrays: a systematic review on the tools to extract underlying genomic structure. *Brief. Bioinform.* **23**, bbac043.
- Bassil, N.V., Davis, T.M., Zhang, H., Ficklin, S., Mittmann, M., Webster, T., Mahoney, L. *et al.* (2015) Development and preliminary evaluation of a 90 K Axiom® SNP array for the allo-octoploid cultivated strawberry *Fragaria × ananassa*. *BMC Genomics*, **16**, 155.
- Bourke, P.M., van Geest, G., Voorrips, R.E., Jansen, J., Kranenburg, T., Shahin, A., Visser, R.G.F. *et al.* (2018) polymapR -linkage analysis and genetic map construction from F1 populations of outcrossing polyploids. *Bioinformatics*, **34**, 3496–3502.
- Burridge, A.J., Winfield, M.O., Allen, A.M., Wilkinson, P.A., Barker, G.L., Coghill, J., Waterfall, C. *et al.* (2017) High-density SNP genotyping array for hexaploid wheat and its relatives. In *Wheat Biotechnology: Methods and Protocols* (Bhalla, P. and Singh, M., eds), pp. 293–306. Totowa, NJ: Humana Press.
- Cai, C., Zhu, G., Zhang, T. and Guo, W. (2017) High-density 80 K SNP array is a powerful tool for genotyping *G. hirsutum* accessions and genome analysis. *BMC Genomics*, **18**, 654.
- Chen, G.B., Lee, S.H., Brion, M.J., Montgomery, G.W., Wray, N.R., Radford-Smith, G.L., Visscher, P.M. *et al.* (2014a) Estimation and partitioning of (co) heritability of inflammatory bowel disease from GWAS and immunochip data. *Hum. Mol. Genet.* **23**, 4710–4720.
- Chen, H., Xie, W., He, H., Yu, H., Chen, W., Li, J., Yu, R. *et al.* (2014b) A high-density SNP genotyping array for rice biology and molecular breeding. *Mol. Plant*, **7**, 541–553.

- Cheng, S., Feng, C., Wingen, L., Cheng, H., Riche, A.B., Jiang, M., Leverington-Waite, M. et al. (2023) *Harnessing landrace diversity empowers wheat breeding for climate resilience*. *bioRxiv*, 2023.10.04.560903. <https://doi.org/10.1101/2023.10.04.560903>
- Cui, F., Zhang, N., Fan, X.L., Zhang, W., Zhao, C.H., Yang, L.J., Pan, R.Q. et al. (2017) Utilization of a Wheat660K SNP array-derived high-density genetic map for high-resolution mapping of a major QTL for kernel number. *Sci. Rep.* **7**, 3788.
- Daware, A., Malik, A., Srivastava, R., Das, D., Ellur, R.K., Singh, A.K., Tyagi, A.K. et al. (2023) Rice Pangenome Genotyping Array: an efficient genotyping solution for pangenome-based accelerated genetic improvement in rice. *Plant J.* **113**, 26–46.
- Ganugi, P., Palchetti, E., Gori, M., Calamai, A., Burridge, A., Biricolti, S., Benedettelli, S. et al. (2021) Molecular diversity within a Mediterranean and European Panel of Tetraploid Wheat (*T. turgidum* subsp.) landraces and modern germplasm inferred using a high-density SNP Array. *Agronomy*, **11**, 414.
- Gao, L., Koo, D.H., Juliana, P., Rife, T., Singh, D., Lemes Da Silva, C., Lux, T. et al. (2021) The *Aegilops ventricosa* 2NvS segment in bread wheat, cytology, genomics and breeding. *Theor. Appl. Genet.* **134**, 529–542.
- Gebremedhin, A., Li, Y., Shunmugam, A.S.K., Sudheesh, S., Valipour-Kahrood, H., Hayden, M.J., Rosewarne, G.M. et al. (2024) Genomic selection for target traits in the Australian lentil breeding program. *Front. Plant Sci.* **14**, 1284781.
- van Geest, G., Voorrips, R.E., Esselink, D., Post, A., Visser, R.G. and Arens, P. (2017) Conclusive evidence for hexasomic inheritance in chrysanthemum based on analysis of a 183 k SNP array. *BMC Genomics*, **18**, 585.
- Grover, G., Sharma, A., Mackay, I., Srivastava, P., Kaur, S., Kaur, J., Burridge, A. et al. (2022) Identification of a novel stripe rust resistance gene from the European winter wheat cultivar 'Acienda': a step towards rust proofing wheat cultivation. *PLoS One*, **17**, e0264027.
- Guilbaud, R., Biselli, C., Buiteveld, J., Cattivelli, L., Copini, P., Dowkiw, A., Esselink, D. et al. (2020) *Development of a new tool (4TREE) for adapted genome selection in European tree species*, GenTree, Avignon, France. GenTree, Jan 2020, Avignon, France (hal-02928391).
- Harper, H., Winfield, M.O., Copas, L., Przewieslik-Allen, S.A., Barker, G.L.A., Burridge, A., Hughes, B.R. et al. (2020) The Long Ashton Legacy: characterising United Kingdom West Country cider apples using a genotyping by targeted sequencing approach. *Plants People Planet*, **2**, 167–175.
- Hong, H., Xu, L., Liu, J., Jones, W.D., Su, Z., Ning, B., Perkins, R. et al. (2012) Technical reproducibility of genotyping SNP arrays used in genome-wide association studies. *PLoS One*, **7**, e44483.
- Horsnell, R., Leigh, F.J., Wright, T.I.C., Burridge, A.J., Ligeza, A., Przewieslik-Allen, A.M., Howell, P. et al. (2023) A wheat chromosome segment substitution line series supports characterization and use of progenitor genetic variation. *Plant Genome*, **17**, e20288.
- Howard, N.P., Troggio, M., Durel, C.E., Muranty, H., Denancé, C., Bianco, L., Tillman, J. et al. (2021) Integration of Infinium and Axiom SNP array data in the outcrossing species *Malus × domestica* and causes for seemingly incompatible calls. *BMC Genomics*, **22**, 246.
- International Wheat Genome Sequencing Consortium (IWGSC) (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, **361**, eaar7191.
- Kabbaj, H., Sall, A.T., Al-Abdallat, A., Geleta, M., Amri, A., Filali-Maltouf, A., Belkadi, B. et al. (2017) Genetic diversity within a Global Panel of Durum Wheat (*Triticum durum*) landraces and modern germplasm reveals the history of alleles exchange. *Front. Plant Sci.* **8**, 01277.
- Kang, Y., Choi, C., Kim, J.Y., Min, K.D. and Kim, C. (2023) Optimizing genomic selection of agricultural traits using K-wheat core collection. *Front. Plant Sci.* **14**, 1112297.
- Kim, K.-W., Nawade, B., Nam, J., Chu, S.-H., Ha, J. and Park, Y.-J. (2022) Development of an inclusive 580K SNP array and its application for genomic selection and genome-wide association studies in rice. *Front. Plant Sci.* **13**, 1036177.
- Koning-Boucoiran, C.F., Esselink, G.D., Vukosavljev, M., van't Westende, W.P., Gitonga, V.W., Krens, F.A., Voorrips, R.E. et al. (2015) Using RNA-Seq to assemble a rose transcriptome with more than 13,000 full-length expressed genes and to develop the WagRhSNP 68k Axiom SNP array for rose (*Rosa L.*). *Front. Plant Sci.* **21**, 249.
- Kosambi, D.D. (1944) The estimation of map distances from recombination values. *Ann. Eugenics*, **12**, 172–175.
- Kumar, D., Chhokar, V., Sheoran, S., Singh, R., Sharma, P., Jaiswal, S., Iqbal, M.A. et al. (2020) Characterization of genetic diversity and population structure in wheat using array based SNP markers. *Mol. Biol. Rep.* **47**, 293–306.
- Kumari, J., Lakhwani, D., Jakhar, P., Sharma, S., Tiwari, S., Mittal, S., Avasthi, H. et al. (2023) Association mapping reveals novel genes and genomic regions controlling grain size architecture in mini core accessions of Indian National Genebank wheat germplasm collection. *Front. Plant Sci.* **28**, 1148658.
- Li, Y.R., van Setten, J., Verma, S.S., Lu, Y., Holmes, M.V., Gao, H., Lek, M. et al. (2015) Concept and design of a genome-wide association genotyping array tailored for transplantation-specific studies. *Genome Med.* **7**, 90.
- Lorieux, M. (2012) MapDisto: fast and efficient computation of genetic linkage maps. *Mol. Breed.* **30**, 1231–1235.
- Martin, F.J., Amode, M.R., Aneja, A., Austine-Orimoloye, O., Azov, A.G., Barnes, I., Becker, A. et al. (2023) Ensembl 2023. *Nucleic Acids Res.* **51**, D933–D941.
- McCouch, S.R., Wright, M.H., Tung, C.W., Maron, L.G., McNally, K.L., Fitzgerald, M., Singh, N. et al. (2016) Open access resources for genome-wide association mapping in rice. *Nat. Commun.* **7**, 10532.
- Montanari, S., Bianco, L., Allen, B.J., Martínez-García, P.J., Bassil, N.V., Postman, J., Knäbel, M. et al. (2019) Development of a highly efficient Axiom™ 70 K SNP array for Pyrus and evaluation for high-density mapping and germplasm characterization. *BMC Genomics*, **20**, 331.
- Nannuru, V.K.R., Windju, S.S., Belova, T., Dieseth, J.A., Alsheikh, M., Dong, Y., McCartney, C.A. et al. (2022) Genetic architecture of *Fusarium* head blight disease resistance and associated traits in Nordic spring wheat. *Theor. Appl. Genet.* **135**, 2247–2263.
- Nantongo, J.S., Odoi, J.B., Agaba, H. and Gwali, S. (2022) SilicoDArt and SNP markers for genetic diversity and population structure analysis of *Trema orientalis*; a fodder species. *PLoS One*, **17**, e0267464.
- Negro, S.S., Millet, E.J., Madur, D., Bauland, C., Combes, V., Welcker, C., Tardieu, F. et al. (2019) Genotyping-by-sequencing and SNP-arrays are complementary for detecting quantitative trait loci by tagging different haplotypes in association studies. *BMC Plant Biol.* **19**, 318.
- Pandey, M.K., Agarwal, G., Kale, S.M., Clevenger, J., Nayak, S.N., Sriswathi, M., Chitkineeni, A. et al. (2017) Development and evaluation of a high density genotyping 'Axiom\_Arachis' Array with 58K SNPs for accelerating genetics and breeding in groundnut. *Sci. Rep.* **7**, 40577.
- Pavy, N., Gagnon, F., Deschênes, A., Boyle, B., Beaulieu, J. and Bousquet, J. (2016) Development of highly reliable in silico SNP resource and genotyping assay from exome capture and sequencing: an example from black spruce (*Picea mariana*). *Mol. Ecol. Resour.* **16**, 588–598.
- Perry, A., Wachowiak, W., Downing, A., Talbot, R. and Cavers, S. (2020) Development of a single nucleotide polymorphism array for population genomic studies in four European pine species. *Mol. Ecol. Resour.* **20**, 1697–1705.
- Rathan, N.D., Krishna, H., Ellur, R.K., Sehgal, D., Govindan, V., Ahlawat, A.K., Krishnappa, G. et al. (2022) Genome-wide association study identifies loci and candidate genes for grain micronutrients and quality traits in wheat (*Triticum aestivum* L.). *Sci. Rep.* **12**, 7037.
- Rimbert, H., Darrier, B., Navarro, J., Kitt, J., Choulet, F., Leveugle, M., Duarte, J. et al. (2018) High throughput SNP discovery and genotyping in hexaploid wheat. *PLoS One*, **13**, e0186329.
- Roorkiwal, M., Jain, A., Kale, S.M., Doddamani, D., Chitkineeni, A., Thudi, M. and Varshney, R.K. (2018) Development and evaluation of high-density Axiom®CicerSNP Array for high-resolution genetic mapping and breeding applications in chickpea. *Plant Biotechnol. J.* **16**, 890–901.
- Senthilvel, S., Ghosh, A., Shaik, M., Shaw, R.K. and Bagali, P.G. (2019) Development and validation of an SNP genotyping array and construction of a high-density linkage map in castor. *Sci. Rep.* **9**, 3003.
- Sheoran, S., Jaiswal, S., Raghav, N., Sharma, R., Sabhyata, Gaur, A., Jaisri, J. et al. (2022) Genome-Wide Association study and Post-Genome-Wide Association study analysis for spike fertility and yield related traits in bread wheat. *Front. Plant Sci.* **12**, 820761. <https://doi.org/10.3389/fpls.2021.820761>

- Shewry, P.R., Brouns, F., Dunn, J., Hood, J., Burrridge, A.J., America, A.H.P., Gilissen, L. *et al.* (2023) Comparative compositions of grain of *Triticum durum* wheat and bread wheat grown in multi-environment trials. *Food Chem.* **423**, 136312.
- Shorinola, O., Simmonds, J., Wingen, L.U. and Uauy, C. (2022) Trend, population structure, and trait mapping from 15 years of national varietal trials of UK winter wheat. *G3 (Bethesda)*, **12**(2), jkab415.
- Soleimani, B., Lehnert, H., Keilwagen, J., Plieske, J., Ordon, F., Naseri Rad, S., Ganal, M. *et al.* (2020) Comparison between core set selection methods using different Illumina marker platforms: a case study of assessment of diversity in wheat. *Front. Plant Sci.* **11**, 1040.
- Stadlmeir, M., Hartl, L. and Mohler, V. (2018) Usefulness of a multiple advanced generation intercross population with greatly reduced mating design for genetic studies in winter wheat. *Front. Plant Sci.* **9**, 1825.
- Sun, C., Dong, Z., Zhao, L., Ren, Y., Zhang, N. and Chen, F. (2020) The wheat 660K SNP array demonstrates great potential for marker-assisted selection in polyploid wheat. *Plant Biotechnol. J.* **18**, 1354–1360.
- Thomson, M.J. (2014) High-throughput SNP genotyping to accelerate crop improvement. *Plant Breed. Biotechnol.* **2**, 195–212.
- Unterseer, S., Bauer, E., Haberer, G., Seidel, M., Knaak, C., Ouzunova, M., Meitinger, T. *et al.* (2014) A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. *BMC Genomics*, **15**, 823.
- Verma, S., Bassil, N.V., van de Weg, E., Harrison, R.J., Monfort, A., Hidalgo, J.M., Amaya, I. *et al.* (2017) Development and evaluation of the Axiom® IStraw35 384HT array for the allo-octoploid cultivated strawberry *Fragaria × ananassa*. *Acta Hort.* **1156**, 75–82.
- Vos, P.G., Uitdewilligen, J.G.A.M.L., Voorrips, R.E., Visser, R.G. and van Eck, H.J. (2015) Development and analysis of a 20K SNP array for potato (*Solanum tuberosum*): an insight into the breeding history. *Theor. Appl. Genet.* **128**, 2387–2401.
- Vukosavljev, M., Arens, P., Voorrips, R., van't Westende, W.P., Esselink, G.D., Bourke, P.M., Cox, P. *et al.* (2016) High-density SNP-based genetic maps for the parents of an outcrossed and a selfed tetraploid garden rose cross, inferred from admixed progeny using the 68k rose SNP array. *Hortic. Res.* **3**, 16052.
- Wang, S., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B.E., Maccaferri, M. *et al.* (2014) Characterization of polyploid wheat genomic diversity using a high-density 90,000 single nucleotide polymorphism array. *Plant Biotechnol. J.* **12**, 787–796.
- Winfield, M.O., Wilkinson, P.A., Allen, A.M., Barker, G.L.A., Coghill, J.A., Burrridge, A., Hall, A. *et al.* (2012) Targeted re-sequencing of the allohexaploid wheat exome. *Plant Biotechnol. J.* **10**, 733–742.
- Winfield, M.O., Allen, A.M., Burrridge, A.J., Barker, G.L.A., Benbow, H.R., Wilkinson, P.A., Coghill, J. *et al.* (2016) High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool. *Plant Biotechnol. J.* **14**, 1195–1206.
- Winfield, M., Burrridge, A., Ordidge, M., Harper, H., Wilkinson, P., Thorogood, D., Copas, L. *et al.* (2020) Development of a minimal KASP marker panel for distinguishing genotypes in apple collections. *PLoS One*, **15**, e0242940.
- Wright, T.I.C., Horsnell, R., Love, B., Burrridge, A.J., Gardner, K.A., Jackson, R., Leigh, F.J. *et al.* (2023) A new winter wheat genetic resource harbours untapped diversity from synthetic hexaploid wheat. *Theor. Appl. Genet.* **137**, 73.
- Xiong, Z., Luo, J., Zou, Y., Tang, Q., Fu, S. and Tang, Z. (2023) The different subtelomeric structure among 1RS arms in wheat-rye 1BL.1RS translocations affecting their meiotic recombination and inducing their structural variation. *BMC Genomics*, **24**, 455.
- Xu, Y., Li, P., Yang, Z. and Xu, C. (2017) Genetic mapping of quantitative trait loci in crops. *Crop J.* **5**, 175–184.
- You, Q., Yang, X., Peng, Z., Xu, L. and Wang, J. (2018) Development and applications of a high throughput genotyping tool for polyploid crops: single nucleotide polymorphism (SNP) array. *Front. Plant Sci.* **9**, 104.
- Yu, G., Cui, Y., Jiao, Y., Zhou, K., Wang, X., Yang, W., Xu, Y. *et al.* (2023) Comparison of sequencing-based and array-based genotyping platforms for genomic prediction of maize hybrid performance. *Crop J.* **11**, 490–498.

## Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**Data S1** Skim sequence varieties.

**Data S2** TaNG v1.0 probe details.

**Data S3** TaNG v1.1 probe details.

**Data S4** Sample details and genotyping.

**Data S5** TaNG v1.1 marker distribution.

**Data S6** Genetic map data.

**Data S7** CNV summary table.