A wheat chromosome segment substitution line series supports characterisation and use of progenitor genetic variation

Richard Horsnell[1‡], Fiona J Leigh[1‡], Tally IC Wright[1‡], Amanda J Burridge[2], Aleksander Ligeza[1], Alexandra M. Przewieslik-Allen[2], Philip Howell[1], Cristobal Uauy[3], Keith J. Edwards[2], Alison R Bentley[1,4*]

[1]The John Bingham Laboratory, NIAB, 93 Lawrence Weaver Road, Cambridge, UK

[2]Life Sciences, University of Bristol, Bristol, BS8 1TQ, UK

[3]John Innes Centre, Norwich Research Park, Norwich, NR4 7UH, UK

[4]Current address: International Wheat and Maize Improvement Center (CIMMYT), El Batan, Mexico

[*]Corresponding author: a.bentley@cgiar.org

[‡]These authors contributed equally to this work

Abstract

Genome-wide introgression and substitution lines have been developed in many plant species, enhancing mapping precision, gene discovery and the identification and exploitation of variation from wild relatives. Created over multiple generations of crossing and/or backcrossing accompanied by marker-assisted selection, the resulting introgression lines are a fixed genetic resource. In this study we report the development of spring wheat chromosome segment substitution lines generated to systematically capture genetic variation from tetraploid (*Triticum turgidum* ssp *dicoccoides*) and diploid (*Aegilops tauschii*) progenitor species. Generated in a common genetic background over four generations of backcrossing, the material is a base resource for the mapping and characterisation of wheat progenitor variation. To facilitate further exploitation the final population was genetically characterised using a high-density genotyping array and a range of agronomic and grain traits assessed to demonstrate the the potential use of the populations for trait localisation in wheat.

Keywords: introgression, CSSL, *Aegilops tauschii, Triticum aestivum*

Introduction

Hexaploid wheat (*Triticum aestivum*) arose through natural hybridization between the tetraploid grass species *T. turgidum* and the diploid species *Aegilops tauschii* (Fu & Somers, 2009). The infrequency of hybridisation resulted in limited genetic diversity (Cox, 1997) and progenitor species are a valuable source of variation (Zhang *et al.*, 2015). Numerous strategies have been proposed and deployed to capture this for application in genetic studies and wheat breeding (Leigh *et al.*, 2022).

A wealth of genomic and germplasm resources is now available for functional genetic characterisation in polyploid wheat (Adamski *et al*., 2020). This includes mapping and gene discovery resources for wild relatives (e.g. stable wild wheat introgression lines developed by Grewal *et al*., 2020). These build on early cytogenetic resources including monosomic and nullisomic chromosome substitution lines series (Sears, 1954; Unrau *et al.*, 1956) which supported the foundations of wheat genetics (Kuspira and Unrau, 1957; Law, 1966). The original monosomics took over 25 years to develop (Berke *et al.*, 1992) and were used to characterise a range of traits (Zemetra *et al*., 1986) as well as to create reciprocal series of chromosome substitution lines, allowing genetic trait mapping (Berke *et al*., 1992).

Substitution line series serve many purposes. Law (1966) described their benefits as three-fold: (1) localisation of genetic effects enabling targeted use in breeding, (2) revealing genetic trait architecture and relating it to patterns of descent and (3) validating predictions of magnitude of genetic effects. Their direct use in breeding has been demonstrated as base populations for trait mapping and line selection (Eshed *et al*., 1992), identification (Basava, 2019) and transfer of wild species variation (Eshed & Zamir, 1994; Nie *et al*., 2015) and pyramiding of target regions or introgression segments (Ali *et al*., 2010). In understanding trait architecture, they support precise QTL mapping (Basava, 2019) and fine mapping (Fulop *et al*., 2016), gene discovery and cloning (Eshed & Zamir, 1994) along with marker development (Qiao *et al*., 2016). Substitution lines typically show little variation for agronomic (growth and development) traits and can be used to confirm predicted magnitude of genetic effect, thereby increasing homogeneity between experiments (Keurentjes *et al*., 2007) as well as eliminating the interference of genetic background (Zhai *et al*., 2016). They are permanent populations (Zhai *et al*., 2016) that can be maintained as immortal seed stocks allowing assessment across environments and traits (Keurentjes *et al.*, 2007), thereby increasing statistical power to detect small effects QTLs and accurately determine the magnitude of genotype x environment interactions (Rae *et al*, 1999; Keurentjes *et al.*, 2007).

Chromosome segment substitution lines (CSSLs) are a form of substitution series and capture genome-wide diversity from a donor species in a fixed genetic background. First described in tomato as a series of interspecific introgression lines capturing small, overlapping chromosome segments of the wild species *Lycopersicum pennellii* into cultivated tomato (*L. esculentum*; Eshed *et al.*, 1992; Eshed & Zamir, 1994) the resulting lines were used to map quantitative traits including yield (Eshed & Zamir, 1995; Fridman *et al.*, 2004) and leaf characters (Holtan & Hake, 2003). The introgression lines have been subsequently used to detect thousands of QTLs for adaptation, morphological characters, yield, metabolism and fruit quality traits (reviewed in Lippman *et al.*, 2007) and cellular features underlying leaf traits (Chitwood *et al.*, 2013).

Further application has been demonstrated in the model species *Arabidopsis thaliana* (Keurentjes *et al.*, 2007; Fletcher *et al.*, 2013) as well as in several cultivated crops including rice (reviewed by Ali *et al.*, 2010), brassica (Howell *et al*., 1996; Ramsay *et al.*, 1996; Rae *et al.*, 1999; Li *et al.*, 2015), peanut (Fonceka *et al.*, 2012), durum wheat (Blanco *et al*., 2006) and pearl millet (Kumari *et al.*, 2014; Basava *et al.*, 2019). A recent review of resources by Balakrishnan *et al.* (2019) detailed populations available over sixteen cultivated crop species.

Despite the importance of wheat in global food security and its limited diversity due to evolutionary history (Stebbins, 1950; Gross & Olsen, 2010) there are few substitution resources available beyond the foundation monosomic and nullisomic series. In hexaploid wheat Zemetra *et al.* (1986) used the Nebraska reciprocal CSSL series to map heading date. More recently Gu *et al.* (2015) developed four sets of hexaploid introgression lines capturing the endemic Chinese wheat subspecies *T. aestivum yunnanense*, *tibetanum* and *petropavlovskyi* along with a synthetic hexaploid wheat. The lines were used to map QTLs for height, spike length and grain number per spike and several lines were identified for use in breeding to incorporate favourable introgression segments. In tetraploid wheat, Joppa and Cantrell (1990) created a 'Langdon'(durum)-*T. dicoccoides* substitution series via crossing between a high grain protein content (GPC) *T. dicoccoides* donor and a series of 'Langdon' D-genome disomic substitution lines. This allowed the identification of lines with superior GPC, subsequently leading to its mapping on chromosome 6B (*QGpc.ndsu-6Bb*; Joppa *et al.*, 1997). Further refinement of mapping to the 6B short arm (Olmos *et al.*, 2003) provided the basis for the cloning of the *GRAIN PROTEIN CONTENT-B1* (*GPC-B1*) gene (Uauy *et al.*, 2006a; Uauy *et al.*, 2006b). This demonstrates the utility of precision introgression resources for trait mapping and gene cloning. In durum wheat, Blanco *et al*. (2006) also developed a set of 92 backcross introgression lines using *T. turgidum* ssp. *dicoccoides*. Developed to capture high GPC from

the wild species, several generations of backcrossing allowed extraction of a single favourable line. This was used to derive a population with the original recurrent parent and to extract high and low protein bulks from the derived F_3 population for marker discovery and QTL detection.

Selection of CSSLs within a series aims to capture segments supporting genetic dissection and that optimise donor coverage. Balakrishnan *et al.* (2019) reviewed the breadth and characteristics of existing CSSL resources reporting that sets were composed of 35-200 lines, typically capturing 90% of donor genome. Tian *et al.* (2006) reported 67% coverage for wild to cultivated rice compared to near complete (99%) coverage in other rice population (Jie *et al*., 2006). In supporting genetic dissection, substitution line phenotypic effects that differ between an introgression line and the recurrent parent are ascribed to the substituted region (Rae *et al*., 1999). To optimise this Wu *et al*. (2006) proposed an additive-dominance model for accurately determining and ascribing substituted segment effects.

In this study we describe the development and characterization of a CSSL series capturing the genomes of hexaploid wheat's primary progenitor species: tetraploid wild emmer (*T. turgidum ssp. dicoccoides;* AB genomes) and diploid goat grass (*Ae. tauschii;* D genome). We genetically characterize the resource using a high-density marker platform to provide accurate resolution of introgression boundaries. We demonstrate that the series can be used to associate trait variation from progenitor species to specific genetic regions. The series is publicly available to provide characterised introgression lines (and associated markers) to support further trait research and progression into breeding.

Materials and Methods

Selection of parental material

Two independent backcross-derived CSSL populations were generated to capture the genomes of the progenitor species *T. turgidum ssp. dicoccoides* (denoted NIAB_AB) and *Ae. tauschii* (denoted NIAB_D). The hexaploid spring wheat 'Paragon' was used as the recurrent parent for both populations as it has been widely used for genetic studies and the creation of genetic stocks. For the NIAB_AB population, the Israeli wild emmer wheat accession TTD-140 was selected as the donor parent. For the NIAB_D population, the Armenian *Ae. tauschii* accession ENT-336 was selected and the synthetic hexaploid line NIAB-SHW041 (created via resynthesis from a Hoh-501 (AB) x ENT-336 (D) cross) used as the introgression donor.

Comparative genetic analysis of 'Paragon', TTD-140 and NIAB-SHW041 used single nucleotide polymorphism (SNP) marker data generated using the Axiom® Wheat Breeder's Genotyping Array (Allen *et al.*, 2017). Two datasets were used, the first with A and B genome

mapped markers and 'Paragon', 24 *T. dicoccoides* (including TTD-140), 12 *T. dicoccum*, 12 *T. durum* accessions and the second using only D genome markers and 'Paragon' and 51 primary synthetic hexaploid wheats (including NIAB-SHW041). Markers were thinned using pairwise comparisons between SNPs using a Pearson correlation test and a single marker was removed in comparisons that yielded an absolute value of the correlation coefficient (*r*) greater than 0.75. The final marker number in the A/B and D genome datasets 2,748 and 217, respectively. The relationships between individuals in each set were compared using principal coordinate analysis (PCoA) of Euclidean genetic distance matrices computed from the marker data, implemented in the R package ape (v5.3, Paradis & Schliep 2018).

Construction of populations

For NIAB_AB, pollen from the wild emmer wheat donor TTD-140 was used to fertilise 'Paragon' to create a pentaploid F_1 generation. Six F_1 plants were used as the maternal parents in the creation of the BC_1 generation, where 'Paragon' was used to pollinate F_1 plants, continuing to BC_4. For NIAB_D the donor ENT-336 was first crossed to the durum wheat Hoh-501 to produce a fertile hexaploid synthetic wheat (NIAB-SHW041). The synthetic wheat was crossed to 'Paragon' and the F_1 progeny then backcrossed four times to 'Paragon' to derive BC_4 plants. At each BC generation genomic DNA was extracted from 2-week-old seedlings of all progenies using a modified Tanksley extraction protocol (Fulton *et al.*, 1995). Marker-assisted selection (MAS) was used to select individuals carrying heterozygous introgressions in the F_1 to BC_4 generations and homozygous introgressions in the BC_4F_2 generation. A schematic of the crossing scheme is given in **Supplementary Figure S1**. The number of lines generated, genotyped, and advanced at each generation is summarised in **Supplementary Table S1**.

MAS at each BC generation used co-dominant KASP™ markers developed to cover the TTD-140 and ENT-336 genomes and to allow for differentiation from 'Paragon'. In addition to foreground selection, evenly spaced markers on the non-target genomes were used to allow background selection against the donor. At BC_4F_4 additional genome-wide SNP genotyping (using the Axiom® Wheat Breeder's Genotyping Array; Allen *et al.*, 2017) was used to improve the accuracy of final selection.

For the NIAB_AB population, 191 KASP™ markers (98 A genome; 93 B genome) distributed at approximately 20-30 cM intervals were selected to screen the BC generations (**Supplementary Table S2**). An additonal set of 50 co-dominant D-genome mapped markers (1-2 per chromosome arm) was used to facilitate selection against tetraploids and nullisomic

aneuploids and amplified a product in D-genome diploids and hexaploids, but not in AB-genome tetraploids, and gave expected results across nulli-tetrasomic series (**Supplementary Table S3**).

Based on graphical genotypes derived from KASP™ marker positions, subsets of BC_1 lines were manually created. The BC_1 lines were selected to carry multiple but complementary heterozygous introgressions across the TTD-140 genome. Lines that failed to amplify at loci targeted by D genome specific KASP™ markers were discarded. For each subsequent cycle of backcrossing, the same selection criteria were applied. A summary of the number of individuals created, genotyped, and selected at each crossing generation is given in **Supplementary Table S1**. From each of the 138 selected BC_4F_1 individuals, 8 BC_4F_2 seeds were grown and genotyped (using the above KASP™ marker set; **Supplementary Table S2**) and one individual was selected based on homozygosity for the marker at the target introgression but minimising background introgressions (using markers in **Supplementary Table S3**). Eight BC_4F_3 seeds from each selected homozygous BC_4F_2 line were subsequently sown and genotyped with the KASP™ marker set to identify the target introgression and to ensure progression of homozygous lines.

At the BC_4F_4 generation, the lines were genotyped using the Axiom® Wheat Breeder's Genotyping Array (Allen *et al.*, 2017). To derive the final NIAB_AB population we used 1,444 A-genome and 1,707 B-genome SNPs which were verified to share the same genetic and physical genome chromosome. Genetic positions were taken from the existing consensus map (Allen *et al.*, 2017) and physical positions determined using the Basic Local Alignment Search Tool (BLAST+) command-line applications on the RefSeq v1.0 genome assembly with default parameters (Camacho *et al.*, 2009, International Wheat Genome Sequencing Consortium (IWGSC) 2018). Only markers with a single A or B genome physical hit were used. Introgression selections were made using physical positions and graphical visualisation of allele variation different from 'Paragon' using Flapjack (v-1.18.06.29, Milne *et al.*, 2010). Introgressions were visually selected on: (a) coverage across the target genome chromosome, (b) homozygosity within each introgression and (c) minimal off-target introgressions.

For the NIAB_D population, 60 D-genome KASP™ markers were selected based on distribution across the D-genome, co-dominance and polymorphism between the recurrent parent 'Paragon', the tetraploid component of NIAB-SHW041, Hoh-501 and the *Ae. tauschii* donor ENT-336 (**Supplementary Table S4**). In addition, 66 A- (33) and B-genome (33) markers were used for background selection against Hoh-501 (**Supplementary Table S5**). The

CSSLs were developed using MAS as described for the NIAB_AB population but with 'Paragon' used as the maternal parent, as summarized in **Supplementary Table S1**.

At the BC_4F_4 generation, lines were genotyped using the Axiom® Wheat Breeder's Genotyping Array. Using the same approach as for the NIAB_AB population, 644 SNPs which shared a genetic and physical D genome chromosome (where markers had only a single physical D genome hit) were used for the final line selection. Introgression line selections were based on coverage across each chromosome, maximising homozygosity and minimising off-target introgressions. Seed of both the NIAB_AB and NIAB_D populations is available from the Germplasm Resources Unit at the John Innes Centre (entries WCSSL0001 to WCSSL0112; available via https://www.seedstor.ac.uk/) and all genotypic data is available from www.cerealsdb.uk.net/cerealgenomics/CerealsDB/array_info.php.

Phenotypic characterisation

The NIAB_AB population was grown under glasshouse conditions to assess grain yield per plant and grain characteristics. Two or three replicate plants of each line were grown in an incomplete block design, surrounded by a 'Paragon' border to minimise edge effects. Each plant was grown in a 1L pot with 16-hour daylength at 22°C and 15°C overnight. At maturity, all plants plus one ear per plant were photographed, and all plant images are available from: https://opendata.earlham.ac.uk/wheat/under_license/toronto/Horsnell,Leigh,Bentley,Wright_2020_NIAB_CSSL_D_genome_yield_trial_H2020/. Seed from each plant was hand threshed and weighed to estimate grain yield per plant. Grains were analysed using a MARVIN seed analyser (MARViTECH GmbH) to obtain grain dimensions and thousand grain weight (TGW).

To assess phenotypic variance in the NIAB_D CSSLs, three replicates of each line along with six replicates of the recurrent parent 'Paragon' were sown in a randomised yield (2x6 m harvested area) trial in Hinxton, Cambridge in 2020. The trial was laid out in an incomplete block design, with 6 plots per block and three complete replicates. Traits assessed in the field included: grain yield (kg/plot at 85% dry matter content), early ground coverage using normalised difference vegetation index ($NDVI^{early}$; an average of two measurements taken between 11:00 and 14:00 GMT on the 15th and 21st of April 2020 at 2m above the canopy using a handheld RapidScan CS-45 (Holland Scientific)), plant height (cm from base to tip (including scurs or awns) of three random plants per plot at maturity), flowering time (days to growth stage (GS) 61 (start of anthesis) recorded when 50% of the plot was at GS61; Zadoks *et al.*, 1974). All field trial data is available from Grassroots (Bian *et al.*, 2017): https://grassroots.tools/fieldtrial/study/61faaf25c68884365e7bcc34. Images of each plot were

taken to assess glaucosity using a RGB camera (Olympus TG4) at a height of 320cm from the ground. Images were cropped to eliminate edge distortion with the final 8 Mpix image covering 85% of the plot. Post-processed images for each plot are available here: https://opendata.earlham.ac.uk/wheat/under_license/toronto/Horsnell,Leigh,Bentley,Wright_2020_NIAB_CSSL_D_genome_yield_trial_H2020/. Post-harvest yield component traits were assessed on ten replicate ears per plot (collected at random) including ear length (cm from base to tip of the ear, not including scurs or awns) and spikelet number. Each of the 10 ears was weighed to determine ear weight, threshed individually and grains analysed using a MARVIN seed analyser (MARViTECH GmbH) to determine seed number, thousand grain weight (TGW) and grain dimensions (length, width).

Phenotypic analysis

For both the NIAB_AB glasshouse and NIAB_D field trial, analysis was conducted in R v4.0.3 (R Core Team, 2020). Each trial included the final CSSLs selected to form the graphical genotype as well as a proportion of CSSLs that were not selected in the final population but carried introgressions. The NIAB_AB glasshouse trial analysis included 146 replicated CSSL individuals. In the NIAB_D field trial analysis there were 33 replicated CSSLs. For each trait, variance between replicates and population distributions were plotted and visually inspected to remove outliers. In the NIAB_D field trial, 10 ears were harvested per plot as technical replicates for measurements on ear and seed characteristics. The ears were measured separately and variance within a plot was then inspected to aid in outlier removal. The technical replicates were then averaged to form plot means for the analysis.

Mixed linear effect models were constructed with restricted maximum likelihood (REML) using the R packages lme4 (Bates *et al.*, 2015) and lmerTest (Kuznetsova *et al.*, 2017). Genotype was treated as a fixed factor and the trial design components including ‘replicate’ and ‘block’ were tested within each model as random factors. Due to visual patterns in model residuals, in addition to the direction of phenotypic assessment and application of agronomic inputs, the trial layout components ‘column’ in the NIAB_D and ‘row’ in the NIAB_AB glasshouse trial were included as random factors. Final model selection was based on Akaike Information Criterion (AIC) and (exact) restricted likelihood ratio tests implemented through the R package RLRsim (Scheipl *et al.*, 2008). In each model, the significance of the genotypic effect was estimated using ANOVA with Satterthwaite’s degrees of freedom method, implemented through lmerTest. Generalised heritability (H^2) was estimated using genotypic best linear unbiased predictors (BLUPs) and the Cullis *et al.* (2006) model, using the R script

developed by Schmidt *et al.* (2019). Ranked bar plots, including only the final selected CSSLs, were created for grain yield (NIAB_D trial), seed width (NIAB_D trial) and seed length (NIAB_AB trial). The package emmeans (Lenth, 2020) was used to complete pairwise comparisons of all BLUEs for these traits and a false discovery rate (FDR) correction was used to adjust *P* values for multiple testing. Finally, the package predictmeans (Luo *et al.*, 2020) was used to inspect the residuals from final models and to extract the BLUEs for examining variation across all traits.

Introgression-trait association

Categorical traits were assessed in each CSSL population to demonstrate the use of the resource for detecting introgression-trait association. In NIAB_AB this used a major known locus linked to awn presence and in NIAB_D to a putative marker association with flag leaf glaucousness. Awn presence was scored in the NIAB_AB glasshouse trial and graphical genotypes were inspected to detect introgressions from the awned donor TTD-140 overlapping with regions where major awn inhibitors have previously been mapped. Huang *et al.* (2020) characterised and identified the *B1* locus as a C2H2 zinc finger encoding gene that corresponds to a coding region on chromosome 5A (698,528,636 to 698,529,001bp; IWGSC 2018). This region was taken as the physical location of the gene for the introgression-trait association.

To demonstrate that CSSLs can also be used to dissect the genetic basis of a less well characterised association, ear, stem, and leaf glaucousness was scored using visual assessment and RGB images in the NIAB_D field trial. This identified a subset of individuals that shared the non-glaucous phenotype, and these were assessed for common introgression regions. A previous study (Würschum *et al.*, 2020a) detected a putative marker-trait association linked to flag leaf glaucousness mapped to 2D (~2.9 Mb) and the ability of the series to validate this interval was assessed via introgression-trait association. Off target regions were also inspected to confirm that the subset shared no other common introgressions. Once each introgression-trait association was identified, schematics using graphical genotype selections were formed for each association and a SNP cluster plot was extracted from Axiom Analysis Suite (Thermo Fisher Scientific) that confirmed the CSSL region segregating within the introgressions of interest. For both populations, markers that were excluded or dropped during the QC were inspected visually to improve resolution of introgression boundaries.

Results

Genetic relationships between parents

The genetic relationships between 48 tetraploid wheat accessions and 51 D genome donors (as novel SHWs) were compared based on Euclidean genetic distance and plotted through PCoA (**Supplementary Figure S2).** TTD-140, the tetraploid donor of the NIAB_AB population clustered with most of the *T. dicoccoides* lines, though the *T. dicoccoides* accessions were shown to contain significant diverse (**Supplementary Figure S2A)**. The main cluster of *T. dicoccoides* was separate from the cultivated wheats by PCoA1, which explained 12% of the variation. The domesticated wheats, including accessions of *T. durum*, *T. dicoccum* and 'Paragon', all clustered above zero on PCoA1. These cultivated groups were separated by the second axis which explained 9% of the variation.

The synthetic donor used for the NIAB_D population (NIAB-SHW041) clustered closely with recurrent parent 'Paragon' on the first PCoA axis (**Supplementary Figure S2B**), where there was a distinct group of primary synthetics that clustered separately from the rest of the material towards the right of the plot (slightly above 5 on PCoA1). PCoA1 accounted for 20% of the variation compared to 12% PCoA2. However, within the lineage shared by NIAB-SHW041 and 'Paragon', the two lines do not appear closely related, which is supported by their separation on PCoA2 in **Supplementary Figure S2B**.

Selection of introgression lines based on graphical genotypes

Final lines were selected based on graphical genotypes from the BC_4F_4 (**Supplementary Table S1**). Two different sets of SNP markers were used for line selection in NIAB_AB: 1,444 markers with physical locations on the A genome and 1,707 markers with physical positions on the B genome. From these data sets, 44 genotypes were identified that had A genome introgressions from TTD-140 (**Figure 1A**). Using the B genome data set, 33 genotypes were identified with introgressions covering the B genome (**Figure 1B**). These 77 selections originated from 62 unique NIAB_AB individuals with some genotypes selected for introgressions at multiple loci. For example, NIAB_AB 1A.04 and 5A.05 represent the same unique individual but are included twice to represent distinct target introgressions. Using the NIAB_D population and a D genome marker set of 644 SNPs, 32 genotypes were selected with introgressions from NIAB-SHW041 (**Figure 1D**). These selections originated from 25 unique NIAB_D lines.

Individuals were typically selected for a single introgression and genotypes with multiple introgressions were selected against. However, the majority of BC_4F_4 individuals carried small off-target introgressions (**Figure 1**). Within each genome, each selected genotype contained on average two introgressions (**Figure 1; Supplementary Table S6**). The average

length of the introgression ranged from an average of 102 Mb for the B genome to 129 Mb for the D genome. Off-target introgressions were also present across the different genomes for each selected line (**Supplementary Figures S3-S5)**. Therefore, as an estimate of recurrent parent background recovery, the percentage of markers with a 'Paragon' allele was quantified for each genotype, using markers with mapped genetic positions on the consensus map (Allen *et al.*, 2017) and genotypes had 96% average recurrent parent recovery (**Supplementary Table S6**). Several NIAB_AB lines appeared to have large chromosomal introgressions from TTD-140 across the D genome (**Supplementary Figures S3, S4)**. As the tetraploid donor lacks a D genome, these lines were likely D-genome aneuploids and alternative (not 'Paragon') alleles were called for markers in that region. The extent of amplification of homeologous loci is not characterised for the markers described herein, although D genome amplification in tetraploid accessions indicates that for some markers, priming sites are sufficiently conserved across genomes to allow amplification.

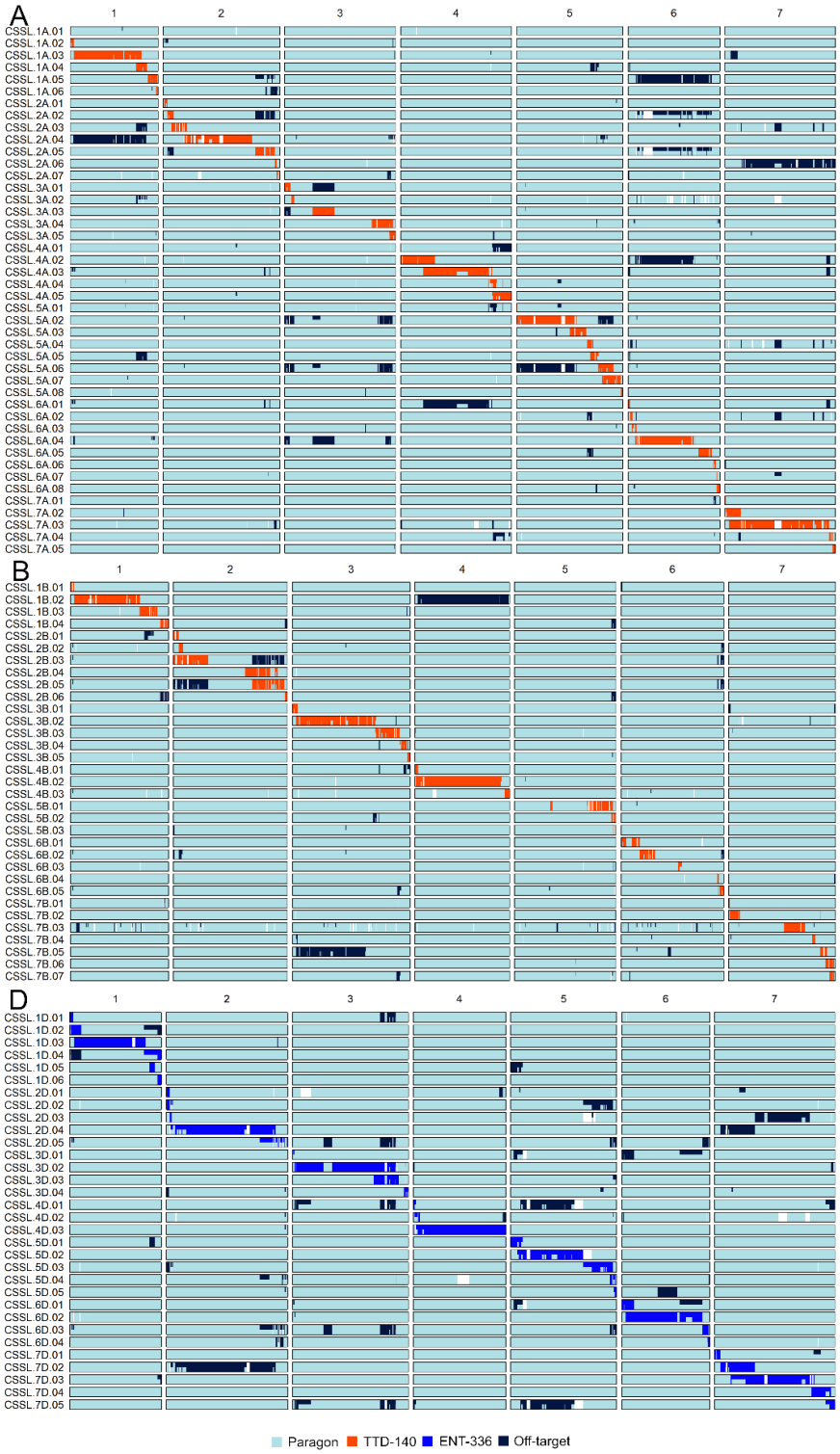

Figure 1. Graphical genotype selections for the A, B and D genome, using the NIAB_AB and NIAB_D populations. A selection of 77 lines was used to form the A and B graphical genotype, originating from 62 unique NIAB_AB CSSLs. The light blue colour represents a 'Paragon' like genotype, whereas orange represents a TTD-140 introgression. Off-target introgressions are shown by navy blue. For the D graphical genotype, 32 lines were selected using 25 unique NIAB_D CSSLs. The medium blue represents an introgression from the NIAB-SHW041 donor, which was formed using the *Ae. tauschii* ENT-336. The chromosomes are scaled on physical position (IWGSC 2018) and the length of each chromosome is determined by marker coverage. White spaces show missing

data and heterozygous regions are represented by divided colours. The figure was created using the R package SelectionTools (v-19.1, www.uni-giessen.de).

Phenotypic variation

Variation was observed for every trait measured in comparison to the recurrent parent 'Paragon'. Trait summary statistics from the NIAB_AB glasshouse and NIAB_D field trial are shown in **Table 1**. For grain length in both trials, and plant height in the NIAB_D field trial, the mean of the CSSL lines was on average higher than 'Paragon' (**Table 1**). However, most trait means for the CSSLs were lower than the recurrent parent, although individual CSSL lines had higher means than 'Paragon' for all traits. In the NIAB_AB glasshouse trial there was a 43.4% increase in grain yield per plant in the highest yielding CSSL line (compared to 'Paragon') and in NIAB_D field trial the highest TGW for a CSSL line gave a 32.5% increase over 'Paragon'. For other traits, such as spikelet number and flowering time in NIAB_D, there were more moderate increases in the maximum CSSL mean compared to 'Paragon' (**Table 1).**

A significant genotypic effect ($P < 0.001$) was observed for all traits (**Table 1**). In NIAB_D, moderate or high generalised heritability (H^2) was observed. The highest H^2 was found for grain length ($H^2 = 0.91$) and flowering time, ear length and seed area also showed high H^2 (**Table 1**). Estimates of H^2 were lowest for grain yield in the NIAB_AB glasshouse assessment ($H^2 = 0.39$), likely due to measurement on a per plant basis.

Table 1. Summary statistics and results from two replicated trials: NIAB_AB glasshouse and NIAB_D field trial. Results are taken from all introgressed lines included in both trials and the recurrent parent 'Paragon'. The best linear unbiased estimates (BLUEs) are shown for 'Paragon' and the CSSLs that had the maximum and minimum BLUE for each trait. The genotypic effect for each trait is shown through ANOVA, using the Satterthwaite's degrees of freedom method. Generalized heritability (H^2) is shown for every trait. For the NIAB_D field trial, ear and seed characteristics (including seed weight) are shown on a per ear basis, averaged from 10 technical replicates per plot.

Trait	Paragon	CSSL $_{min}$	CSSL $_{mean}$	CSSL $_{max}$	Num *df*	Den *df*	*F*	*P*	H^2
NIAB_AB glasshouse trial									
Grain yield (g/plant)	4.38	1.06	3.74	6.28	143	161.36	1.67	<0.001	0.39
Grain length (mm)	6.32	5.83	6.53	7.14	144	204.02	3.56	<0.001	0.70
NIAB_D field trial									
Grain yield (85% DMC kg/plot)	6.13	4.90	5.89	6.73	33	55.85	3.07	<0.001	0.68

Flowering time (Days from drilling to GS65)	189.13	183.99	189.05	191.75	33	54.33	8.34	<0.001	0.88
Plant height (cm)	97.65	92.31	97.84	104.68	33	57.62	4.17	<0.001	0.76
NDVI ^{early} (index)	0.64	0.52	0.59	0.67	33	51.97	2.94	<0.001	0.65
Spikelet no. (count)	22.29	20.22	21.57	22.93	33	59.68	6.51	<0.001	0.83
Ear length (cm)	10.57	9.51	10.33	11.83	33	59.10	8.76	<0.001	0.88
Ear weight (g)	2.95	2.38	2.77	3.51	33	56.52	2.81	<0.001	0.64
Seed no. (no. ear $^{-1}$)	58.62	42.81	56.43	63.64	33	57.46	5.88	<0.001	0.82
Seed weight (g ear $^{-1}$)	2.39	1.91	2.21	2.89	33	56.65	3.20	<0.001	0.68
TGW (g)	40.56	31.69	39.10	53.74	33	59.21	7.66	<0.001	0.87
Grain length (mm)	6.63	6.35	6.70	7.33	33	64.46	11.22	<0.001	0.91
Grain width (mm)	3.43	3.13	3.35	3.65	33	59.68	5.09	<0.001	0.80
Grain area (mm^2)	16.80	14.66	16.55	20.12	33	60.15	8.65	<0.001	0.88

In the NIAB_AB glasshouse trial, significant variation was observed for grain length (**Table 1**, **Figure 2A**). The majority of the CSSLs had a higher mean than 'Paragon'; compared to the recurrent parent there was a 13.0% increase in the CSSL with the maximum grain length (NIAB_AB 5A.02/5A.06). For the NIAB_AB selections, post hoc tests indicated several lines with grain length significantly different to 'Paragon' (**Figure 2A**) and variation for a subset of lines is shown in **Figure 2B**, clearly demonstrating visual variation. Grain length showed a positive weak correlation with grain yield per plant (r = 0.3, P = 0.02; **Figure 2C**). Overall, there was a significant genotypic effect observed for grain yield per plant (**Table 1**), although the post hoc tests did not identify any CSSLs that differed significantly to 'Paragon' and measurement on a per plant basis in the glasshouse is the likely cause of the large standard error associated with the BLUEs (**Figure 2D**).

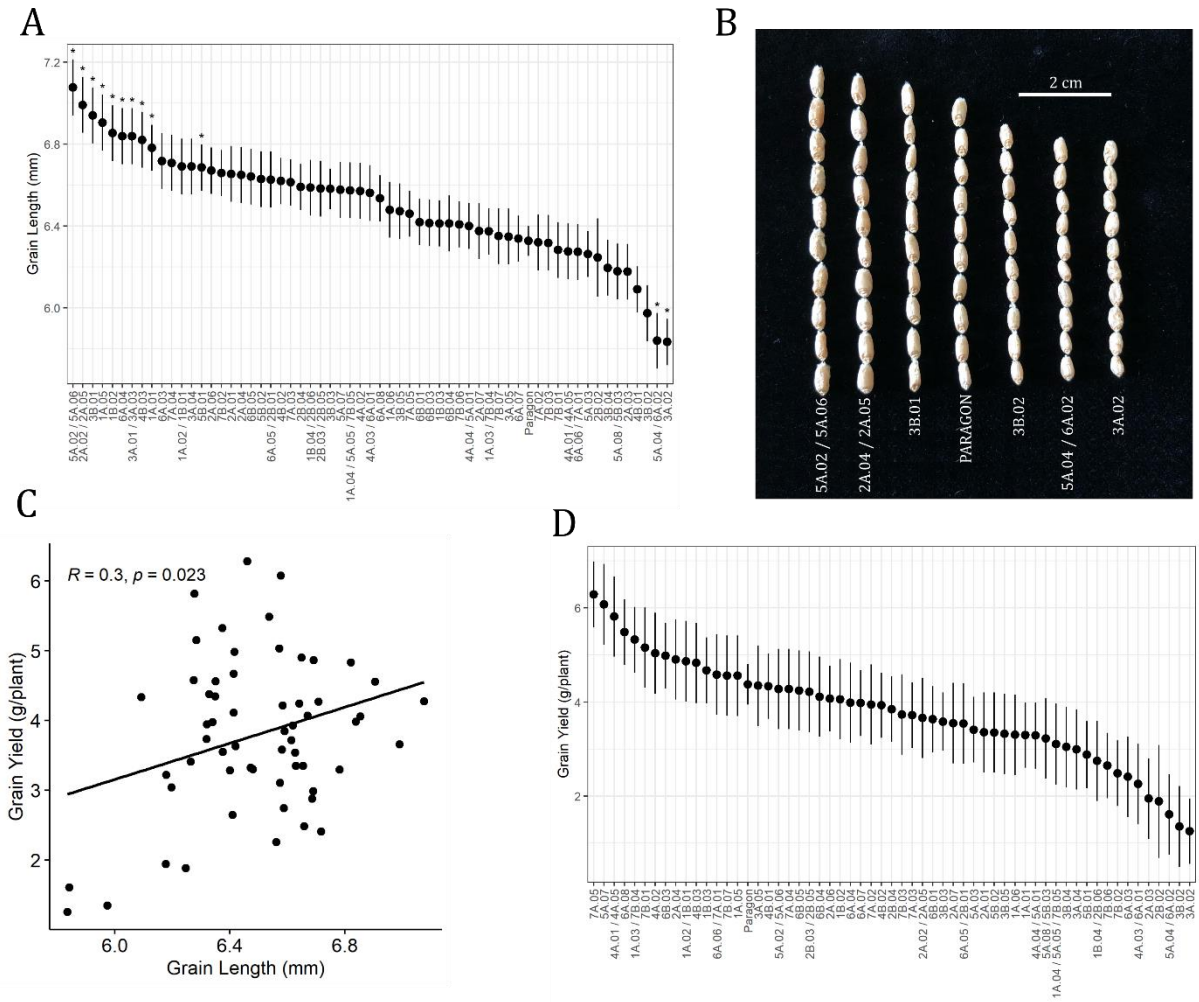

Figure 2. Variation for grain length and yield per plant in the NIAB_AB glasshouse trial: A) Ranked best linear unbiased estimates (BLUEs) for seed length, with the NIAB_AB lines selected for the graphical genotype and the recurrent parent 'Paragon' shown. The asterisks above genotypes indicate where CSSLs were significantly different to 'Paragon' ($P < 0.05$). B) Variation in seed length observed with CSSLs with the longest and shortest seeds included, and 'Paragon' shown for comparison. C) The relationship between grain length and yield, with the Pearson correlation coefficient test statistics overlayed on the plot. D) NIAB_AB CSSLs and 'Paragon' ranked BLUEs for grain yield per plant. Error bars in A and D represent standard error of the BLUEs.

Several traits showed a positive correlation with grain yield in the NIAB_D trial (**Figure 3A)**, of which the positive correlation with grain width was the most significant (r = 0.61, $P < 0.001$; **Figure 3A**). Several significant negative trade-offs were also observed with flowering time having the most negative correlations with other traits (**Figure 3A)**. Earlier flowering was negatively associated with ear length, TGW and grain length, width, and area. A significant negative relationship was also observed between seed number (per ear) and other grain size related traits (including TGW, seed length and area).

Grain yield was reduced in the majority of selected CSSLs compared to 'Paragon', except in five CSSLs (**Figure 3B**). Lines selected for multiple introgressions across the D genome gave some of the lowest yield performance. Post-hoc pairwise comparisons showed that three of these were significantly lower than 'Paragon' (4D.01/7D.05, 2D.02/5D.03 and 1D.05/5D.01). None of the CSSLs had significantly higher grain yield than 'Paragon'. Several NIAB_D CSSLs showed significantly different means for grain width compared to 'Paragon': five CSSLs had a lower predicted mean and a single CSSL had a higher mean (2D.04/7D.02, **Figure 3C**). There were clear visual differences in grain width observed across the NIAB_D CSSLs (**Figure 3D**). However, despite wide grain width CSSL.2D.04/7D.02 did not have a higher grain yield than 'Paragon' (**Figure 3B**).

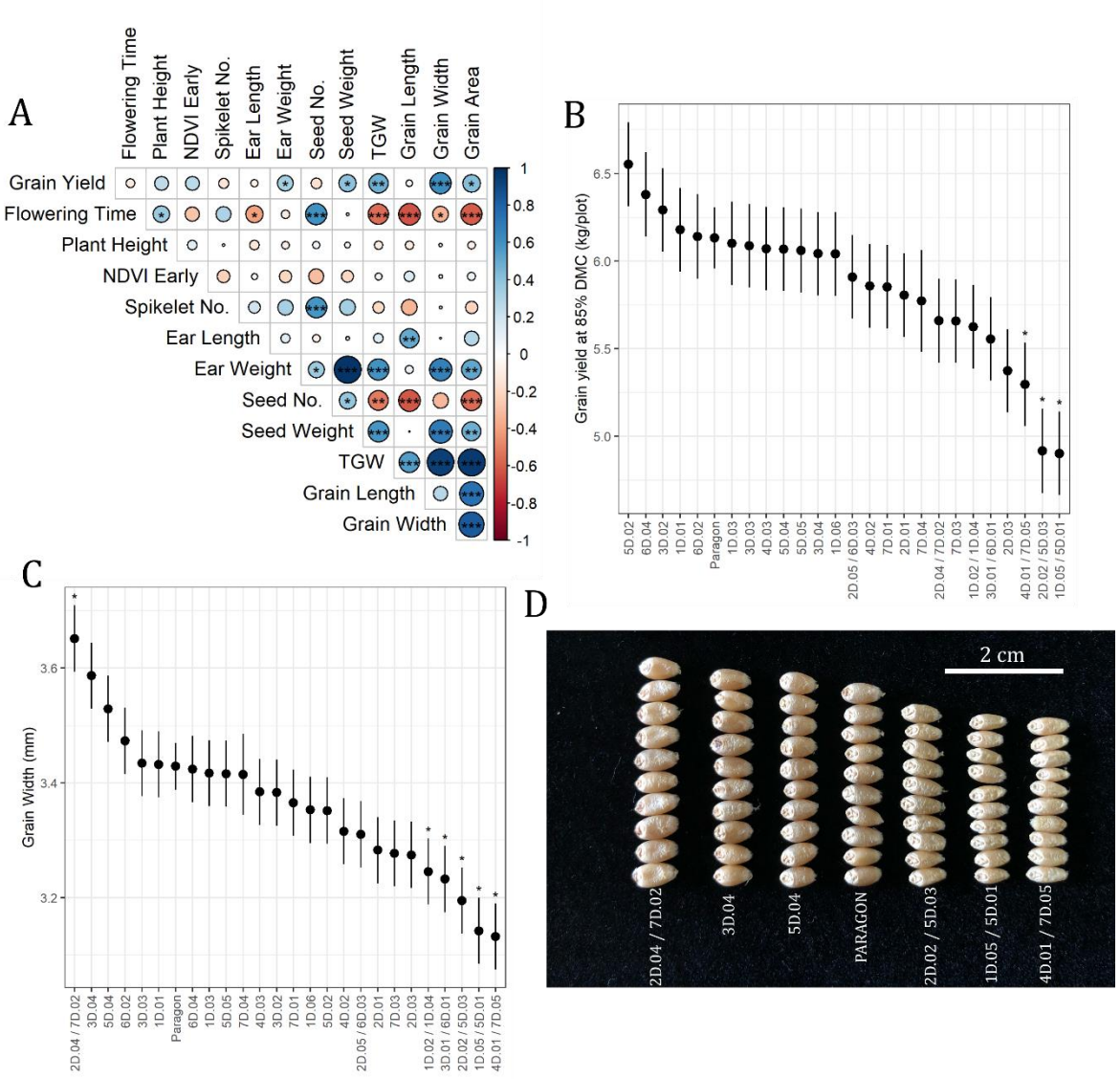

Figure 3. Trait correlations and variation for grain yield and grain width in the NIAB_D field trial: A) a Pearson's correlation matrix calculated using the best linear unbiased estimates (BLUEs) from models fitted for each trait.

412 The heat scale bar shows the Pearson correlation coefficient (*r*). The asterisks within the matrix show the
413 significance threshold taken from the *P*-value for each comparison: *** = $P < 0.001$, ** = $P < 0.01$ and * = $P < 0.05$
414 created using R package corrplot (Wei and Simko, 2017). B) Ranked BLUEs for grain yield and C) grain width
415 from a mixed linear model fitted via lme4 (Bates et al., 2015). Post-hoc pairwise comparisons of the BLUEs used
416 the package emmeans (Lenth 2020). The asterisks indicate significant difference to 'Paragon' ($P < 0.05$). All error
417 bars shown represent standard error of the predicted means. D) A photograph showing diversity in grain width
418 using samples taken from plots in the NIAB_D field trial. The CSSLs with the three highest and lowest grain
419 widths are shown, with 'Paragon' included for comparison.
420
421 *Introgression-trait association and validation*
422 For the introgression-trait analysis, markers that were excluded during the earlier stages of QC
423 were included to provide further resolution on introgression boundaries. The presence of awns
424 was detected in a single NIAB_AB line (5A.07; **Figure 4A**). The CSSLs with neighbouring
425 introgressions (5A.06 and 5A.08) were non-awned (**Figure 4A**). The awn inhibitor *B1* has been
426 previously identified (TraesCS5A02G542800, 5A; 698,528,636 to 698,529,001bp; Huang *et*
427 *al.*, 2020). The introgression from TTD-140 in 5A.07 overlapped this gene (570 Mb to 700
428 Mb), whereas in 5A.06 and 5A.08 the introgressions were in adjacent regions (**Figure 4B**). The
429 SNP marker AX-94504627 had a physical location (698,003,176 bp) close to the *B1* locus. The
430 SNP cluster plot extracted from the Axiom Analysis Suite (Thermo Fisher Scientific) showed
431 that the only CSSL to carry the TTD-140 allele at this location was 5A.07. As shown in **Figure**
432 **1**, NIAB_AB_5A.07 exhibits few off target introgressions on other chromosomes. No
433 introgressions were observed in the regions harbouring the other reported awn inhibitors
434 (Hooded locus (*Hd*)) on 4A and the awn inhibitor (*Tipped 2*) B2 on 6B (Mcintosh *et al.*, 2013).
435

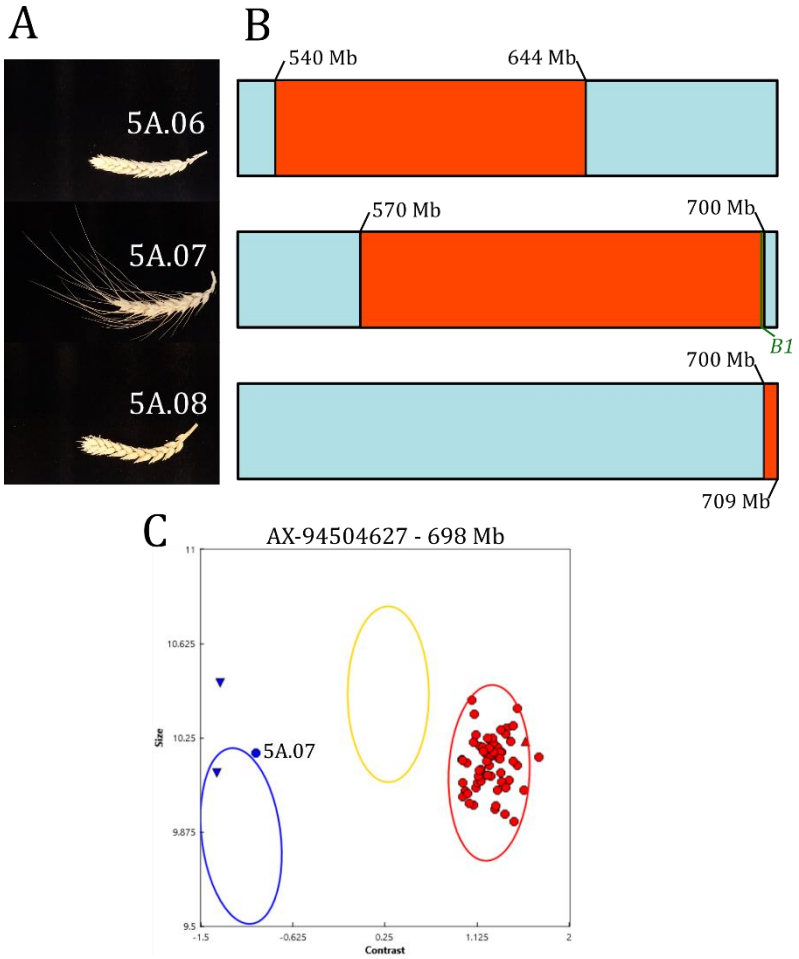

Figure 4. Introgression-trait association using the NIAB_AB population. A) Ears from the NIAB_AB glasshouse trial, showing the single awned CSSL (5A.07). B) The awned CSSL had an introgression from TTD-140 on 5A that extended over the *B1* awn inhibitor locus (Huang *et al.*, 2020) and awnless CSSLs 5A.06 and 5A.08 did not have an introgression in this region. A TTD-140 introgression is represented by the colour orange, light blue represents a 'Paragon' genotype. C) A genotyping plot taken from Axiom Analysis Suite (Thermo Fisher Scientific) for a SNP close to the gene (AX-94504627; 698,003,176 bp) showing CSSLs from the AB graphical genotype (circular points), two TTD-140 replicates (inverted triangle points) and 'Paragon' (a single triangle point). The only line to share the TTD-140 allele (blue) is CSSL.5A.07. Physical positions are in mega base pairs (Mb) and were taken from IWGSC (2018).

Three NIAB_D CSSLs displayed non-glaucous ear, stem and leaf phenotypes based on visual and RGB image assessment: 2D.01, 2D.02 and 3D.04. The CSSLs 2D.01 and 2D.02 were selected for NIAB-SHW041 introgressions at the start of 2D (**Figure 5A**), while 3D.04 was selected for a 3D introgression. However, 3D.04 also carried an off-target introgression at the start of 2D, from 1.6 to 13.7 Mb (**Figure 1**). The CSSL lines 2D.01 and 2D.02 carried an introgression on 2D from 2.7 to 17.6 Mb and 1.6 to 17.6 Mb, respectively. Additionally, 2D.02 had a downstream heterozygous introgression starting at around 19.7 Mb (**Figure 5A**). The

CSSL 2D.03 did not share the non-glaucous phenotype, which is visible in the RGB canopy photographs of field plots for these genotypes (**Figure 5B**). The line 2D.03 had a homozygous introgression from 19.6 to 29.2 Mb. The previously identified putative marker-trait association with flag leaf glaucousness was based on a marker located at 2.9 Mb on 2D and hypothesized to be *Iw2* (Würschum *et al.*, 2020a), the D-genome paralog of *INHIBITOR of WAX1* (*Iw1*; identified by Huang *et al.*, 2017). All three NIAB_D lines that were non-glaucous in the graphical genotype carried an introgression overlapping this region, which can be observed for 2D.01 and 2D.02 in **Figure 5A**. A SNP marker located at 5.1 Mb also confirmed that the three non-glaucous CSSLs did not possess the 'Paragon' allele close to the marker-trait association (**Figure 5C)**. The introgressions in these lines at the start of 2D were typically homozygous (**Figure 1**). However, the genotype calls at AX-94815994 are heterozygous (**Figure 5C)**. It is likely that the calls for the lines were homozygous for the ENT-336 (D-genome) allele but differences in ploidy impacting SNP calling between the CSSLs and the *Ae. tauschii* donor has separated the cluster. Whole genome graphical genotypes were inspected visually and the CSSLs 2D.01, 2D.02 and 3D.04 did not share any other overlapping introgressions.

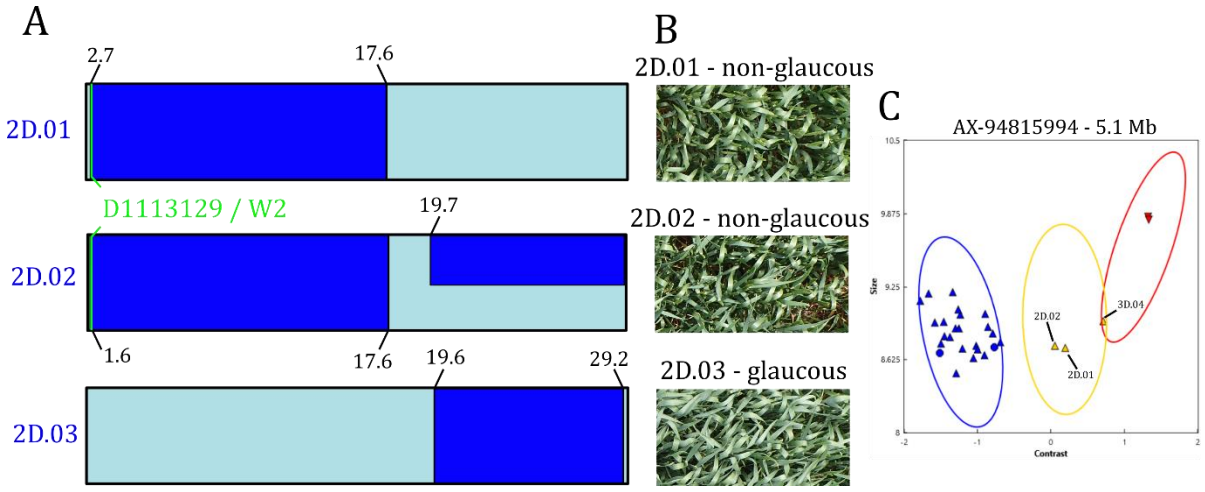

Figure 5. A) Schematic showing regions of chromosome 2D on three NIAB_D genome lines; light blue indicates a 'Paragon' genotype and dark blue indicates an introgression from NIAB-SHW041. The location of a putative QTL (linked to the marker D1113129) proposed to correspond to a known locus for wax production *Iw2* (Würschum *et al,*. 2020a) is overlayed on the plot. B). Canopy photos from the NIAB_D field trial show 2D.01 and 2D.02 were non-glaucous. C) A plot extracted from the genotyping software Axiom Analysis Suite (Thermo Fisher Scientific) for a SNP (AX-94815994; 5.1 Mb on 2D) that shows CSSLs genotypes from the D graphical genotype (triangle points), two of the D genome donor ENT-336 replicates (inverted triangle points) and 'Paragon circular points). The three CSSLs that were non-glaucous in the NIAB_D field trial are labelled. Physical positions are shown in mega base pairs (Mb) taken from (IWGSC 2018).

Discussion

We report the creation of a genetic resource for detecting and exploiting wheat progenitor variation. Publicly available pure seed, phenotyping and genotyping data support use as introgression donors and/or near-isogenic lines. The introgressions are genetically tractable, of reasonable size for forward use and have low off-target segments. This builds on early work to develop wheat resources for gene localisation and characterisation (Sears 1953 & 1954) and genetic mapping (Zemetra *et al*., 1986). Since this time numerous forward and reverse wheat genetic resources have been developed and the reported wheat CSSL series adds a further precision resource for exploiting progenitor diversity.

The NIAB_AB and NIAB_D series were created using backcrossing and MAS followed by high-density genotyping to facilitate progenitor segment introgression with low off-target effects. This approach is relatively time and cost intensive. Different methods of population construction have been proposed, including those with a lower marker requirement (e.g. Bian *et al*. (2010) used 8 informative markers at BC_3F_4 followed by pooled genotyping to extract individuals with small donor segments). This reduced genotyping approach is an alternative, low-cost approach to support CSSL development although larger numbers of individuals need to be progressed through backcrossing. Future development of CSSL series is likely to benefit from the reducing costs of genotyping, allowing more intensive screening and enhancing both accuracy of donor introgression and background recovery. The reducing cost of whole genome skim sequencing is also likely to allow higher resolution of introgression segments.

In both the NIAB_AB and NIAB_D series, we detected significant trait variation and transgressive segregation compared to ‘Paragon’. Grain size and shape are largely independent traits and we observed strong variation for grain length and width. Grain length of ancestral wheat has been shown to be generally higher than in elite bread wheat (Gegas *et al.*, 2010) and selection of alleles that confer increased grain length could increase grain size. Brinton *et al.* (2017) identified a significant QTL for TGW on chromosome 5A (*Qtgw-cb.5A*) demonstrating that the increase in grain area (driving TGW) was predominantly conferred by an increase in grain length. Yang *et al.* (2019) also identified a novel allele associated with increased grain length on 5AL (cloning *TaGL3-5A-G*). This allele, found in ancestral wheats, represents a breeding target, particularly if it could be pyramided in combination with favourable alleles for grain width (e.g. *TaGW2*; Simmonds *et al.*, 2016) and grain size (*TaGS5-3A*; Ma *et al.*, 2015). In the NIAB_AB population, there was a significant positive effect of grain yield per plant correlated with grain length. This identified individual CSSL lines with introgressions

associated with increased grain length which are effectively near-isogenic lines for the TTD-140 segment. The introgression in the lines with the greatest increase in grain length is in the same region as *TaGL3-5A* (Yang *et al.,* 2019) and future work could establish whether it contains the favourable allele described by Yang *et al* (2019). These lines could be field tested to validate the effect, and the segments transferred to additional elite backgrounds to test stability. In addition, the extreme positive and negative (compared to 'Paragon') lines could be used in a bulk-segregant genotyping approach (e.g. based on exome sequencing; Martinez *et al*., 2020 or RNA-seq; Zhu *et al.,* 2020) to determine the specific grain length controllers.

For the NIAB_D population, significant trait variation and high trait heritabilities were observed. The significant genotypic effects, despite the high recovery of 'Paragon' background, indicate effective capture (and contribution) of functional variation from the *Ae. tauschii* donor ENT-336. Variation was observed for grain yield although was only significant for CSSLs with yields lower than 'Paragon'. Although not statistically significant, five CSSLs yielded higher than 'Paragon'. Further testing through multi-site trials could establish if there is a significant yield effect associated with these CSSLs. Grain width was highly correlated with yield. This allowed identification of CSSLs with positive and negative effects for grain width which are available as near-isogenic lines, introgression segment donors and for further genetic characterisation, as above. Grain width has been well characterised in hexaploid wheat (Simmonds *et al.,* 2016) and alternative alleles for *TaGW2* can increase grain width and therefore grain weight, though phenotypes may be subtle due to allele dosage effects. Previous GWAS analysis of *Ae. tauschii* collections have identified numerous QTL for grain characteristics conferred by genetic effects on multiple chromosomes (Arora *et al*., 2017; Zhao *et al.,* 2021). The further development of markers for the regions underlying these QTLs will facilitate the validation of allelic effects and potential use for breeding.

In addition to identification of lines with positive and negative introgression effects, specific trait-introgression effects could be assessed based on physical locations of the segments. Although the series can't be used directly for mapping *per se* due to the low power of detection it can be used to associate introgressions with phenotypes of interest. This is useful to validate previously observed effects as well as being a starting point for fine mapping. Here, our definition of introgression segment boundaries is limited to the coverage of Axiom® Wheat Breeder's Genotyping Array (Allen *et al.,* 2017) but further genotyping would allow higher resolution and definition of boundaries and allow full automation of the process (array data requires manual curation below software quality thresholds). However, the current resolution is likely to be sufficient for transfer of segments via MAS.

549 The recurrent parent 'Paragon' carries the *B1* awn inhibition locus. Using the
550 NIAB_AB population, and through replacement of the region with the wild type (recessive)
551 progenitor allele via introgression we confirmed the location of the locus. This was near the
552 edge of the introgression boundary but could be reliably detected and linked to the phenotype.
553 The *Tipped 1* (*B1*) mutant is one of the three major loci linked to awn presence in wheat
554 (Watkins and Ellerton, 1940; Sourdille *et al*., 2002; Yoshioka *et al.*, 2017) and has been
555 identified as an important determinant of awn development (Mackay *et al*., 2014; Würschum
556 *et al.*, 2020b). The *B1* locus has been fine mapped in multiple studies to 5AL (DeWitt *et al.*,
557 2020; Huang *et al.*, 2020; Niu *et al*., 2020; Wang *et al*., 2020; Würschum *et al.*, 2020b). Awn
558 presence and the *B1* locus have been recently linked to an offset in the negative association
559 between grain yield and protein content (Scott *et al.*, 2021). The detection of the awn inhibitor
560 segment demonstrates the use of the resource for categorical traits and could be used to further
561 understand awn and linked-awn trait haplotype contributions from *T. diccocoides* and expedite
562 their transfer to breeding. CSSLs with and without awns could also be used to further
563 understand the mechanisms of the recently reported trade-offs between awns, quality, and
564 nutritional traits (Scott *et al.*, 2021).

565 In the NIAB_D population we demonstrate the use of trait-introgression association to
566 validate a putative association with a categorical trait contributed by the progenitor *Ae. tauschii*.
567 This used a step-by-step statistical approach enabling validation of a marker-trait association
568 for glaucousness which fell slightly below the significance threshold in Würschum *et al.*
569 (2020a). Our approach aligned the glaucous phenotype of the CSSL introgression segment with
570 the physical position of the peak marker reported by Würschum *et al.* (2020a). In addition to
571 confirming the genetic effect, this provides a 35K Axiom marker to facilitate MAS and tracking
572 in breeding, adding to the previously identified DArTseq peak marker from Würschum *et al.*
573 (2020a). This is likely to be advantageous given the relative ease of conversion of array probes
574 compared to sequence-based probe conversion (Burridge *et al.*, 2018).

575 Here we report the development, characterization, and initial testing of a CSSL wheat
576 series as a tractable genetic resource adding to the existing portfolio of functional genomics
577 resources for wheat improvement. The use of 'Paragon' as a genetic background complements
578 other available wheat resources including the A.E. Watkins hexaploid wheat nested association
579 mapping panel, and EMS and γ mutant populations (Wingen *et al.*, 2017), QTL datasets
580 (Przewieslik-Allen *et al.*, 2021), and publicly available synthetic hexaploid wheat populations
581 (Adamski *et al.*, 2020). Further work is required to reduce off-target introgressions and to split
582 multiple introgression lines into discreet single-introgression lines. The application of higher

583 density markers and/or sequencing will support further precision in defining introgression
584 boundaries. The extreme phenotypic lines can be further validated as near isogenic lines and
585 used as donor lines for transfer of introgressions to additional elite parents for use in breeding.
586 The NIAB_AB and NIAB_D lines can also be intercrossed to create favourable introgression
587 complements (e.g. grain length and grain width) for further testing and development of targeted
588 multi-introgression lines for specific traits of interest. Overall, the resource reported here
589 increases access to wheat genomic resources for research and breeding improvements.
590
591 **Acknowledgements**
592 Development of the CSSL resource was supported by the UK's Biotechnology and Biological
593 Sciences Research Council via the Wheat Improvement Strategic Program (BB/I002561/1) and
594 Designing Future Wheat (BB/P016855/1). We thank Professor Andy Greenland, Dr Nicolas
595 Gosman and Dr Keith Gardner for their contribution and insights into the development of the
596 resource. We thank Michael Scott for supplying the BLAST+ output for the SNP markers and
597 Dr Rob Davey for facilitating trial and image uploads to Grassroots.
598
599 **Conflict of interest**
600 The authors declare no conflict of interest.
601
602 **Author contributions**
603 P.H., C.U., K.J.E. and A.R.B. conceived the project; R.H., F.J.L, T.I.C.W., A.R.B designed
604 experiments; R.H., F.J.L, A.J.B., A.L., A.M.P.-A. performed the experiments; F.J.L, T.I.C.W.,
605 P.H. and A.R.B analysed experiments, R.H., F.J.L, T.I.C.W. and A.R.B wrote the manuscript
606 with inputs from all other authors.
607
608 **Data availability**
609 Seed of the germplasm reported here is directly available from the Germplasm Resources Unit
610 at the John Innes Centre (via https://www.seedstor.ac.uk/), including the parental line
611 'Paragon' (WCSSL0001), donor lines TTD-140 (WCSSL0002) and NIAB-SHW041
612 (WCSSL0003) and for the NIAB_AB (WCSSL0004 to WCSSL0080) and NIAB_D
613 (WCSSL0081 to WCSSL0112) populations.
614 The Axiom® Wheat Breeder's Genotyping Array data is available for direct download for each
615 of the three genomes here:

https://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/genotyping_data/NIAB_CSSL_Agenome.csv
https://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/genotyping_data/NIAB_CSSL_Bgenome.csv
https://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/genotyping_data/NIAB_CSSL_Dgenome.csv

Plant images from the NIAB_AB glasshouse trial are available from: https://opendata.earlham.ac.uk/wheat/under_license/toronto/Horsnell,Leigh,Bentley,Wright_2020_NIAB_CSSL_D_genome_yield_trial_H2020/

Plot images from the NIAB_D field trial are available from: https://opendata.earlham.ac.uk/wheat/under_license/toronto/Horsnell,Leigh,Bentley,Wright_2020_NIAB_CSSL_D_genome_yield_trial_H2020/

Data from the NIAB_D field trial are available from: https://grassroots.tools/fieldtrial/study/61faaf25c68884365e7bcc34

References

Adamski NM, Borrill P, Brinton J, *et al,* (2020) A roadmap for gene functional characterisation in crops with large genomes: Lessons from polyploid wheat. eLife 9:e55646 doi:10.7554/eLife.55646

Ali ML, Sanchez PL, Yu S *et al.* (2010) Chromosome Segment Substitution Lines: a powerful tool for the introgression of valuable genes from *Oryza* wild species into cultivated rice (*O. sativa*). *Rice* 3: 218–234. https://doi.org/10.1007/s12284-010-9058-3

Allen AM, Winfield MO, Burridge AJ, 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 Biotechnology Journal* 15: 390–401. https://doi.org/10.1111/pbi.12635

Arora S, Singh N, Kaur S, Bains NS, Uauy C, Poland J, Chhuneja P (2017) Genome-wide association study of grain architecture in wild wheat *Aegilops tauschii*. *Frontiers in Plant Science* 8: 886.

649 Balakrishnan D, Surapaneni M, Mesapogu S, Neelamraju S (2019) Development and use of
650 chromosome segment substitution lines as a genetic resource for crop improvement.
651 Theoretical and Applied Genetics 132: 1-25. doi:10.1007/s00122-018-3219-y
652
653 Basava RK, Hash CT, Mahendrakar MD, Kishor KPB, Satyavathi CT, Kumar S, Singh RB,
654 Yadav RS, Gupta R, Srivastava RK (2019) Discerning combining ability loci for divergent
655 environments using chromosome segment substitution lines (CSSLs) in pearl millet. *PLoS*
656 *ONE* 14: e0218916. https://doi.org/ 10.1371/journal.pone.0218916
657
658 Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using
659 lme4. *J Stat Softw* 67 https://doi.org/10.18637/jss.v067.i01
660
661 Berke TG, Baenziger PS, Morris WR (1992) Chromosomal location of wheat quantitative trait
662 loci affecting agronomic performance, using reciprocal chromosome substitutions. *Crop*
663 *Science* 32: 621–627.
664
665 Bian JM, Jiang L, Liu LL, Wei XJ, Xiao YH, Zhang LJ, et al. (2010) Construction of a new set
666 of rice chromosome segment substitution lines and identification of grain weight and related
667 traits QTLs. *Breeding Science* 60: 305–13.
668
669 Bian X, Tyrrell S, Davey RP (2017) The Grassroots life science data infrastructure.
670 https://grassroots.tools
671
672 Blanco A, Simeone R, Gadaleta A (2006) Detection of QTLs for grain protein content in durum
673 wheat. *Theoretical and Applied Genetics* 112: 1195–1204.
674
675 Brinton J, Simmonds J, Minter F, Leverington-Waite M, Snape J, Uauy C (2017) Increased
676 pericarp cell length underlies a major quantitative trait locus for grain weight in hexaploid
677 wheat. *New Phytologist* 215: 1026-1038.
678
679 Burridge A, Wilkinson P, Winfield M, Barker G, Przewieslik-Allen S, Coghill J, Waterfall C,
680 Edwards K (2018) Conversion of array-based single nucleotide polymorphic markers for use
681 in targeted genotyping by sequencing in hexaploid wheat (*Triticum aestivum*). *Plant*
682 *Biotechnology Journal* 16: 867-876.

Camacho C, Coulouris G, Avagyan V, Ma N, Papdopoloulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. *BMC Bioinformatics* 10, 421. https://doi.org/10.1186/1471-2105-10-421

Chitwood DH, Kumar R, Headland LR, Ranjan A, Covington MF et al. (2013) A quantitative genetic basis for leaf morphology in a set of precisely defined tomato introgression lines. *Plant Cell* 25: 2465–2481.

Cox TS (1997) Deepening the wheat gene pool. *Journal of Crop Production* 1: 1-25. doi: 10.1300/J144v01n01_01.

Cullis BR, Smith AB, Coombes NE (2006) On the design of early generation variety trials with correlated data. *Journal of Agricultural Biology and Environmental Statistics* 11: 381-393. https://doi.org/10.1198/108571106X15443.

DeWitt, N., Guedira, M., Lauer, E., Sarinelli, M., Tyagi, P., Fu, D., et al. (2020) Sequence-based mapping identifies a candidate transcription repressor underlying awn suppression at the *B1* locus in wheat. *New Phytologist*, 225, 326–339.

Eshed Y, Abu-Abied M, Saranga Y Zamir D (1992) *Lycopersicon esculentum* lines containing small overlapping introgressions from *L. pennellii*. *Theoretical and Applied Genetics* 83: 1027–1034.

Eshed Y, Zamir D (1994) A genomic library of *Lycopersicon pennellii* in *L. esculentum*: a tool for fine mapping of genes. *Euphytica* 79: 175–79.

Eshed Y Zamir D (1995) An introgression line population of *Lycopersicon pennellii* in the cultivated tomato enables the identification and fine mapping of yield-associated QTL. *Genetics* 141: 1147–1162.

Evans LT, Bingham J, Jackson P, Sutherland J (1972) Effect of awns and drought on the supply of photosynthate and its distribution within wheat ears. *Annals of Applied Biology*, 70, 67–76.

717 Falke KC, Miedaner T, Frisch M (2009) Selection strategies for the development of rye
718 introgression libraries. *Theoretial and Applied Genetics* 119: 595–603.
719
720 Feldman M, Millet E (1995) Methodologies for identification, allocation and transfer of
721 quantitative genes from wild emmer into cultivated wheat. In: Li ZS, Xin ZY (eds) Proceedings
722 of the 8th international wheat genetics symposium, Beijing, China, pp 19–27.
723
724 Fletcher RS, Mullen JL, Yoder S *et al.* (2013) Development of a next-generation NIL library
725 in *Arabidopsis thaliana* for dissecting complex traits. *BMC Genomics*. 14: 655.
726 doi:10.1186/1471-2164-14-655.
727
728 Fonceka D, Tossim HA, Rivallan R, et al. (2012) Construction of chromosome segment
729 substitution lines in peanut (*Arachis hypogaea* L.) using a wild synthetic and QTL mapping for
730 plant morphology. *PLoS One* 7: e48642. doi:10.1371/journal.pone.0048642
731
732 Fridman E, Carrari F, Liu Y-S, Fernie AR, Zamir D (2004) Zooming in on a quantitative trait
733 for tomato yield using interspecific introgressions. *Science* 17: 1786-9.
734
735 Fu YB, Somers DJ (2009) Genome-wide reduction of genetic diversity in wheat breeding. *Crop*
736 *Science* 49: 61–168.
737
738 Fulop D, Ranjan A, Ofner I, Covington MF, Chitwood DH, West D, Ichihashi Y, Headland L,
739 Zamir D, Maloof JN, Sinha NR (2016) A new advanced backcross tomato population enables
740 high resolution leaf QTL mapping and gene identification. *G3: Genes, Genomes, Genetics* 6:
741 3169-3184.
742
743 Fulton TM, Chunwongse J, Tanksley SD (1995) Microprep protocol for extraction of DNA
744 from tomato and other herbaceous plants. *Plant Molecular Biology Reporter* 13: 207–209.
745
746 Gaurav K, Arora S, Silva P *et al.* (2021) Population genomic analysis of *Aegilops tauschii*
747 identifies targets for bread wheat improvement. *Nature Biotechnology*
748 https://doi.org/10.1038/s41587-021-01058-4
749

Gegas VC, Nazari A, Griffiths S, Simmonds J, Fish L, Orford S, Sayers L, Doonan JH, Snape JW (2010) A genetic framework for grain size and shape variation in wheat. *The Plant Cell* 22: 1046–1056,

Grewal S, Hubbart-Edwards S, Yang C, et al. (2020) Rapid identification of homozygosity and site of wild relative introgressions in wheat through chromosome-specific KASP genotyping assays. *Plant Biotechnology Journal* 18: 743-755. doi:10.1111/pbi.13241

Gross BL, Olsen KM (2010) Genetic perspectives on crop domestication. *Trends in Plant Science* 15: 529–537.

Gu L, Wei B, Fan R, Jia X, Wang X, Zhang X (2015) Development, identification and utilization of introgression lines using Chinese endemic and synthetic wheat as donors. *Journal of Integrative Plant Biology* 8: 688–697

Holtan HEE, Hake S (2003) Quantitative trait locus analysis of leaf dissection in tomato using *Lycopersicon pennellii* segmental introgression lines. *Genetics* 165: 1541-1550.

Howell PM, Lydiate DJ, Marshall DF (1996). Towards developing intervarietal substitution lines in *Brassica napus* using marker-assisted selection. *Genome* 39: 348-358.

Huang D, Feurtado A, Smith MA, Flatman LK, Koh C, Cutler AJ (2017) Long noncoding miRNA gene represses wheat ß-diketone waxes. *Proceedings of the National Academy of Sciences* 114: E3149-E3158.

Huang D, Zheng Q, Melchkart T, Bekkaoui Y, Konkin DJF, Kagale S et al. (2020) Dominant inhibition of awn development by a putative zinc-finger transcriptional repressor expressed at the *B1* locus in wheat. *New Phytologist*, 225, 340–355.

International Wheat Genome Sequencing Consortium (IWGSC) (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*. 17: 361 (6403): eaar7191. doi: 10.1126/science.aar7191.

783 Jie C, Bughio H, Chen DZ, Liu G, Zheng K, Zhuang J (2006) Development of chromosomal
784 segment substitution lines from a backcross recombinant inbred population of interspecific rice
785 cross. *Rice Science* 13: 15–21.
786
787 Joppa LR, Cantrell RG (1990) Chromosomal location of genes for grain protein content of wild
788 tetraploid wheat. *Crop Science* 30: 1059-1064.
789
790 Joppa LR, Du C, Hart GE, Hareland GA (1997) Mapping gene(s) for grain protein in tetraploid
791 wheat (*Triticum turgidum* L.) using a population of recombinant inbred chromosome lines.
792 *Crop Science* 37: 1586-1589.
793
794 Keurentjes JJB, Bentsink L, Alonso-Blanco C, Hanhart CJ, Vries HB-D, et al. (2007)
795 Development of a near-isogenic line population of *Arabidopsis thaliana* and comparison of
796 mapping power with a recombinant inbred line population. *Genetics* 175: 891–905.
797
798 Kishii M (2019) An update of recent use of Aegilops species in wheat breeding. *Frontiers in*
799 *Plant Science* 10:585. doi: 10.3389/fpls.2019.00585
800
801 Kumari BR, Kolesnikova-Allen MA, Hash CT, Senthilvel S, Nepolean T, KaviKishor PB,
802 Riera-Lizarazu O, Witcombe JR, Srivastava RK (2014) Development of a set of chromosome
803 segment substitution lines in pearl millet [*Pennisetum glaucum* (L.) R. Br.]. *Crop Science* 54:
804 2175–2182.
805
806 Kuspira J, Unrau J (1957) Genetic analysis of certain characters in common wheat using whole
807 chromosome substitution lines. *Canadian Journal of Plant Sciences* 37: 300-326.
808
809 Kuznetsova A, Brockhoff PB, Christensen RHB (2017) lmerTest package: tests in linear mixed
810 effects models. *Journal of Statistics Software* 82: https://doi.org/10.18637/jss.v082.i13
811
812 Law CN (1966) Biometrical analysis using chromosome substitutions within a species. P59-85
813 In R. Riley and KR Lewis (ed.) Chromosome manipulations in plant genetics. Plenum Press,
814 New York.
815

816 Leigh FJ, Wright TIC, Horsnell RA, Dyer S, Bentley AR (2022) Progenitor species hold
817 untapped diversity for potential climate-responsive traits for use in wheat breeding and crop
818 improvement. *Heredity*
819
820 Lenth RV (2020) emmeans: estimated marginal means, aka least-squares means. R package
821 version 1.5.3. https://CRAN.R-project.org/package=emmeans.
822
823 Martinez SA, Shorinola O, Conselman S. *et al.* (2020). Exome sequencing of bulked segregants
824 identified a novel TaMKK3-A allele linked to the wheat ERA8 ABA-hypersensitive
825 germination phenotype. *Theoretical and Applied Genetics* 133: 719–736.
826
827 Li X, Wang W, Wang Z, Li K, Lim YP, Piao Z (2015). Construction of chromosome segment
828 substitution lines enables QTL mapping for flowering and morphological traits in *Brassica*
829 *rapa*. *Frontiers in Plant Science* 6:432. doi:10.3389/fpls.2015.00432
830
831 Luo D, Ganesh S, Koolaard J (2020). Predictmeans: calculate predicted means for linear
832 models. R package version 1.0.4. https://CRAN.R-project.org/package=predictmeans.
833
834 Lippman ZB, Semel Y, Zamir D (2007) An integrated view of quantitative trait variation using
835 tomato interspecific introgression lines. *Current Opinions in Genetics and Development* 17:
836 545-552.
837
838 Mcintosh R, Dubcovsky J, Rogers WJ, Morris *et al.* (2013) Catalogue of gene symbols for
839 wheat. In 12th International Wheat Genetics Symposium, Yokohama, Japan.
840
841 Ma L, Li T, Hao C, Wang Y, Chen X, Zhang X (2016) *TaGS5-3A*, a grain size gene selected
842 during wheat improvement for larger kernel and yield. *Plant Biotechnology Journal*. 14: 1269-
843 1280. https://doi.org/10.1111/pbi.12492
844
845 Mackay IJ, Bansept-Basler P, Barber T, Bentley AR, Cockram J, Gosman N, Greenland AJ,
846 Horsnell R, Howells R, O'Sullivan D, Rose GA, Howell P (2014) An eight-parent Multiparent
847 Advanced Generation Inter-Cross population for winter-sown wheat: creation, properties and
848 validation. *G3: Genes, Genomes, Genetics* 4:1603-1610.
849

Millet E, Rong J-K, Qualset C, Mcguire PE, Bernard A, Sourdille P, Feldman M (2014) Grain yield and grain protein percentage of common wheat lines with wild emmer chromosome-arm substitutions. *Euphytica*. 195. 10.1007/s10681-013-0975-2.

Milne I, Shaw P, Stephen G, *et al.* (2010) Flapjack-graphical genotype visualization. *Bioinformatics* 26: 3133–3134. https://doi.org/10.1093/bioinformatics/btq580

Nevo E (2014) Evolution of wild emmer wheat and crop improvement. *Journal of Systematics and Evolution* 52: 673-696.

Nie X, Tu J, Wang B, Zhou X, Lin Z (2015) A BIL population derived from *G. hirsutum* and *G. barbadense* provides a resource for cotton genetics and breeding. *PLoS ONE* 10: e0141064. https://doi.org/10.1371/journal.pone.0141064

Olmos S, Distelfeld A, Chicaiza O, Schlatter AR, Fahima T, Echenique V, Dubcovsky J (2003) Precise mapping of a locus affecting grain protein content in durum wheat. *Theoretical and Applied Genetics* 107: 1243-1251.

Paradis E, Schliep K (2018) Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35: 526-528.

Przewieslik-Allen A, Wilkinson PA, Burridge A, Winfield M, Dai X, Beaumont M, King J, Yang C, Griffiths S, Wingen L, Horsnell R, Bentley AR, Shewry P, Barker GLA, Edwards KJ (2021) The role of gene flow and chromosomal instability in shaping the bread wheat genome. *Nature Plants* 7: 172-183.

Qiao W, Qi L, Cheng Z, Su L, Li J, Yan S, Ren J, Zheng Z, Yang Q (2016) Development and characterization of chromosome segment substitution lines derived from *Oryza rufipogon* in the genetic background of *O. sativa* spp. *indica* cultivar 9311. *BMC Genomics* 17: 580. https://doi.org/10.1186/s12864-016-2987-5.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

884 Rae AM, Howell EC, Kearsey MJ (1999) More QTL for flowering time revealed by
885 substitution lines in *Brassica oleracea*. Heredity 83: 586-596. doi:10.1038/sj.hdy.6886050
886
887 Ramsay LD, Jennings DE, Bohuon EJR, Arthur AE, Lydiate DJ *et al.* (1996) The construction
888 of a substitution library of recombinant backcross lines in *Brassica oleracea* for the precision
889 mapping of quantitative trait loci. *Genome*. 39: 558–567.
890
891 Rodriguez-Alvarez MX, Boer MP, van Eeuwijk FA, Eilers PHC (2018) Correcting for spatial
892 heterogeneity in plant breeding experiments with P-splines. *Spatial Statistics* 23:52-71
893 https://doi.org/10.1016/j.spasta.2017.10.003
894
895 Scheipl F, Greven S, Kuechenhoff H (2008) Size and power of tests for a zero random effect
896 variance or polynomial regression in additive and linear mixed models. *Computational*
897 *Statistics and Data Analysis*, 52:3283-3299.
898
899 Scott MF, Ladejobi O, Amer S *et al.* (2020) Multi-parent populations in crops: a toolbox
900 integrating genomics and genetic mapping with breeding. *Heredity* 125: 396-416.
901
902 Scott MF, Fradgley N, Bentley AR, Brabbs T, Corke F, Gardner KA *et al*. (2021) Limited
903 haplotype diversity underlies polygenic trait architecture across 70 years of wheat breeding.
904 *Genome Biology*, 22, 137.
905
906 Sears ER (1953) Nullisomic analysis in common wheat. *American Naturalist* 87: 245–252.
907
908 Sears ER (1954) The aneuploids of common wheat. *Research Bulletin Missouri Agricultural*
909 *Experiment Station* 572: 1-59.
910
911 Simmonds J, Scott P, Brinton J, Mestre TC, Bush M, Del Blanco A, Dubcovsky J, Uauy C
912 (2016) A splice acceptor site mutation in *TaGW2-A1* increases thousand grain weight in
913 tetraploid and hexaploid wheat through wider and longer grains. *Theoretical and Applied*
914 *Genetics* 129:1099–1112.
915
916 Sourdille P, Cadalen T, Gay G, Gill B, Bernard M. (2002) Molecular and physical mapping of
917 genes affecting awning in wheat. *Plant Breeding* 121, 320–324.

Stebbins GL (1950) Variation and Evolution in Plants. Columbia University Press, New York, USA.

Tian F, Li DJ, Fu Q, Zhu ZF, Fu YC, Wang XK, Sun CQ (2006) Construction of introgression lines carrying wild rice (*Oryza rufipogon* Griff.) segments in cultivated rice (*Oryza sativa* L.) background and characterization of introgressed segments associated with yield-related traits. *Theoretical and Applied Genetics* 112: 570–580.

Uauy C, Brevis JC, Dubcovsky J (2006a) The high grain protein content gene *Gpc-B1* accelerates senescence and has pleiotropic effects on protein content in wheat. *Journal of Experimental Botany* 57: 2785-94.

Uauy C, Distelfeld A, Fahima T, Blechl A, Dubcovsky J (2006b) A NAC gene regulating senescence improves grain protein, zinc, and iron content in wheat. *Science* 314:1298-301.

Unrau J, Person C, Kuspira J (1956) Chromosome substitution in hexaploid wheat. *Canadian Journal of Botany* 34: 629–640.

Urrutia M, Bonet J, Arús P, Monfort A (2015) A near-isogenic line (NIL) collection in diploid strawberry and its use in the genetic analysis of morphologic, phenotypic and nutritional characters. *Theoretical and Applied Genetics* 128:1261-1275. doi:10.1007/s00122-015-2503-3.

Walkowiak S, Gao L, Monat C *et al*. (2020) Multiple wheat genomes reveal global variation in modern breeding. *Nature* 588 : 277–283 https://doi.org/10.1038/s41586-020-2961-x

Wang D, Yu K, Jin D, Sun L, Chu J, Wu W *et al.* (2020) Natural variations in the promoter of *Awn Length Inhibitor 1* (*ALI-1*) are associated with awn elongation and grain length in common wheat. *The Plant Journal* 101; 1075–1090.

Watkins AE, Ellerton S (1940) Variation and genetics of the awn in *Triticum*. *Journal of Genetics* 40: 243–270.

952 Wei T, Simko, V (2017) R Package "Corrplot": visualization of a correlation matrix. Available
953 online: https://github.com/taiyun/corrplot
954
955 Wilkinson PA, Winfield MO, Barker GLA, Tyrell S, Bian X *et al.* (2016) CerealsDB 3.0:
956 expansion of resources and data integration. *BMC Bioinformatics doi*.org/10.1186/s12859-
957 0161139-x
958
959 Wingen LU, West C, Leverington-Waite M, Collier S, Orford S, Goram R, Yang CY, King J,
960 Allen AM, Burridge A, Edwards KJ, Griffiths S (2017) Wheat landrace genome diversity.
961 *Genetics*. 205: 1657-1676.
962
963 Wu J, Jenkins JN, McCarty JC, Saha S, Stelly DM (2006) An additive-dominance model to
964 determine chromosomal effects in chromosome substitution lines and other gemplasms.
965 *Theoretical and Applied Genetics* 112: 391–399.
966
967 Würschum T, Langer SM, Longin CFH, Tucker MR, Leiser WL (2020a) Refining the genetic
968 architecture of flag leaf glaucousness in wheat. *Theoretical and Applied Genetics*, 133, 981–
969 991.
970
971 Würschum T, Jähne F, Phillips AL, Langer SM, Longin CFH, Tucker MR, Leiser WL (2020b)
972 Misexpression of a transcriptional repressor candidate provides a molecular mechanism for the
973 suppression of awns by *Tipped 1* in wheat. *Journal of Experimental Botany*, 71, 3428–3436.
974
975 Yu J, Holland JB, McMullen MD, Buckler ES (2008) Genetic design and statistical power of
976 nested association mapping in maize. *Genetics* 178: 539-551.
977 doi:10.1534/genetics.107.074245
978
979 Yang J, Zhou Y, Wu Q. *et al.* (2019) Molecular characterization of a novel *TaGL3-5A* allele
980 and its association with grain length in wheat (*Triticum aestivum* L.). *Theoretical and Applied*
981 *Genetics* 132:1799–1814. https://doi.org/10.1007/s00122-019-03316-1
982
983 Yoshioka M, Iehisa JCM, Ohno R, Kimura T, Enoki H, Nishimura S *et al.* (2017) Three
984 dominant awnless genes in common wheat: Fine mapping, interaction and contribution to
985 diversity in awn shape and length. *PLoS ONE*, 12: 1–21.

Zemetra RS, Morris WR, Schmidt JW (1986) Gene locations for heading date using reciprocal chromosome substitutions in winter wheat. *Crop Science* 26:531–533.

Zhai H, Gong W, Tan Y, Liu A, Song W, Li J, Deng Z, Kong L, Gong J, Shang H, Chen T, Ge Q, Shi Y, Yuan Y (2016) Identification of chromosome segment substitution lines of *Gossypium barbadense* introgressed in *G. hirsutum* and quantitative trait locus mapping for fiber quality and yield traits. *PLoS One* 11: e0159101. doi: 10.1371/journal.pone.0159101.

Zhang P, Dundas IS, McIntosh RA, Xu SS, Park RF, Gill BS, et al. (2015) Wheat–Aegilops Introgressions. In: *Alien Introgression in Wheat: Cytogenetics, Molecular Biology, and Genomics* (Molnár-Láng M, Ceoloni C, Doležel J, eds) pp 221-243. New York: Springer. doi: 10.1007/978-3-319-23494-6

Zhao X, Lv L, Li J, Ma F, Bai S, Zhou Y, Zhang D, Li S, Song CP (2021) Genome-wide association study of grain shapes in *Aegilops tauschii*. *Euphytica*, 217: 1-14.