Whole genome sequencing uncovers the structural and transcriptomic landscape of hexaploid wheat/*Am. muticum* introgression lines

Benedict Coombes¹, John Fellers², Surbhi Grewal³, Rachel Rusholme-Pilcher¹, Stella Hubbard-Edwards³, Cai-yun Yang³, Ryan Joynson¹, Ian P. King³, Julie King³, Anthony Hall^{1*}

¹Earlham Institute, Norwich, Norfolk, NR4 7UZ, UK

²USDA–ARS Hard Winter Wheat Genetics Research Unit, Manhattan, KS 66506, USA

³School of Biosciences, The University of Nottingham, Sutton Bonington Campus, Loughborough, Leicestershire, LE12 5RD, UK

*Corresponding author – anthony.hall@earlham.ac.uk

Abstract

Wheat is a globally vital crop, but its limited genetic variation creates a challenge for breeders aiming to maintain or accelerate agricultural improvements over time. Introducing novel genes and alleles from wheat's wild relatives into the wheat breeding pool via introgression lines is an important component of overcoming this low variation but is limited by poor genomic resolution and limited understanding of the genomic impact of introgression breeding. By sequencing 17 hexaploid wheat/*Ambylopyrum muticum* introgression lines and the parent lines, we have precisely pinpointed the borders of introgressed segments. We report a genome assembly and annotation of *Am. muticum* that has facilitated the identification of *Am. muticum* resistance genes commonly introgressed in lines resistant to stripe rust. Our analysis has identified an abundance of structural disruption and homoeologous pairing across the introgression lines, likely caused by the suppressed Ph1 locus. mRNAseq analysis of six of these introgression lines revealed that introgressed genes tend to be downregulated, shifting the expression balance of triads towards suppression of the introgressed region, with no discernible compensation in the expression of the homoeologous copies. This analysis explores the genomic impact of introgression breeding and provides an affordable way for breeders to better characterise introgression lines and more effectively deploy wild relative variation.

Introduction

Triticum aestivum L. (bread wheat) is a vital crop, providing around 20% of calories and 25% of protein consumed globally (1). Improvements to wheat since the late 19th century have largely come from conventional breeding strategies, but these improvements rely on ample genetic variation in the primary gene pool (2). The hexaploid bread wheat grown today derives from just one or two polyploidisation events ~10,000 years ago between the tetraploid *Triticum turgidum* and the diploid *Aegilops tauschii* (3). The limited diversity stemming from this genetic bottleneck has been compounded over time by intensive breeding. Pressure on

breeders to prioritise advanced breeding material (4) for more rapid development of elite uniform, high quality varieties that perform competitively has limited the introduction of genetic variation from external sources. The genetic variation that does exist in modern wheat material is rapidly being exhausted, evident in plateauing yield improvements that left unchecked, will be insufficient to meet global demands (5). Wild relative introgression breeding will be a major component of overcoming this genetic constraint in the years to come, enabling breeders to access the secondary and tertiary gene pools of wheat (2, 4) and incorporate novel alleles or genes into modern breeding material.

There are many examples of the successful transfer of wild relative genes into wheat since first pioneered by E.R. Sears (6–10). However, challenges associated with the high-throughput production and verification of introgression lines, in addition to the linkage drag of introgressed segments, has limited the widespread adoption of introgression breeding. Utilising recombination mutants and high-throughput marker methods, introgressing entire wild relative genomes into wheat as stably inherited, homozygous segments is now possible (11, 12). These sets of lines provide the raw material required for the incorporation of alien variation into breeding programmes. Segments in these lines that confer phenotypes of interest can be identified. Lines with overlapping segments can then be crossed to break down large segments (13), resulting in genes of interest captured in short introgressed segments with reduced linkage drag, ready to be crossed into elite lines and deployed in breeding programmes.

Identifying the introgressed content of each introgression line is important for the effective utilisation of these lines. Insufficient marker density and low resolution of genomic *in situ* hybridisation (GISH) limits the resolution at which segments can be identified. Genotyping, using methods such as Kompetitive allele specific PCR (KASP), is a cheap and effective way to screen lines for introgressed segments. Determining the precise size and positions of segments and refining positions of overlap between introgression lines is important when relating to phenotypic data to narrow down regions containing genes of interest. Identifying introgression boundaries at a higher resolution will allow lines with overlapping segments to be identified; these can be crossed to break down segments and capture genes of interest in smaller segments with reduced linkage drag.

Wild relative genes have undergone selection in a different environment to the agricultural setting in which elite wheat lines are selected and thus may be deleterious, or be, at the very least imperfect replacements of their wheat orthologue when deployed in field conditions. Therefore, many genes introgressed along with a gene of interest will contribute to reduced agronomic performance of a line. This reduced performance will be driven by differences both in the encoded protein and in the pattern of expression of the introgressed gene compared to the wheat orthologue it replaced. In addition to these *cis* effects of introgression, disruptions to established regulatory networks and resulting *trans* effects on the expression of background wheat genes will likely contribute to altered performance.

Ordinarily, hexaploid wheat behaves as a diploid during meiosis. The *Pairing Homoeologous 1* (*Ph1*) locus is largely responsible for this behaviour, restricting synapsis and crossovers to homologous chromosomes (14). A

suppressed or deleted *Ph1* locus enables recombination between wheat chromosomes and non-homologous wild relative chromosomes and has been a major tool used for decades to transfer wild relative genes into wheat (15). However, this also enables homoeologous chromosomes to pair and recombine leading to transmission of chromatin between the subgenomes of wheat (16) and deletions/duplications where synteny between homoeologous chromosomes breaks.

Here, we have conducted a high-resolution genomic analysis on 17 *Am muticum*/hexaploid wheat introgression lines (11, 12), utilising whole-genome sequencing (WGS) data from the introgression lines and the parent lines and a draft genome assembly of *Amblyopyrum muticum* [(Boiss.) Eig.; *Aegilops mutica* Boiss; 2n=2X=14; genome TT], a wild relative of wheat belonging to its tertiary gene pool. Phenotypic screening of *Am. muticum* introgression lines (17) has revealed resistances to leaf, stem and stripe rust not observed in the parental wheat lines and thus likely conferred by introgressed genes. KASP genotyping to identify segments has been conducted on many of these lines (18). Through this analysis, we have pinpointed introgression segment junctions to a higher resolution than previously possible, in many cases within a single pair of reads, demonstrating segments of variable size that overlap between introgression lines, which explains some differences in resistance phenotype seen between lines. These overlaps will enable these segments to be further broken down by crossing introgression lines together. Using *in silico* karyotyping, we have shown that large-scale structural disruption is ubiquitous across the lines, including deletions and duplications up to whole-chromosome size and homoeologous recombination likely facilitated by the suppressed *Ph1* locus. A genome assembly and gene annotation of *Am. muticum* has enabled us to identify introgressed resistance genes in stripe, stem and leaf rust resistant lines that may represent novel resistance conferred by *Am. muticum* genes. Analysis of gene expression of six introgression lines compared to the wheat parent lines has revealed a tendency for introgressed genes to be downregulated or silenced compared to the wheat orthologue they replace, with minimal balancing of the homoeologous copies in the remaining subgenomes.

Results

Whole genome sequencing facilitates high resolution introgression detection

To reveal *Am. muticum* segments within introgression lines using WGS data, we developed a workflow that utilises mapping coverage and single nucleotide polymorphism (SNP) information from the introgression line and the wheat parents. If a wheat segment is replaced by an *Am. muticum* segment the mapping coverage will drop in that region due to structural variation and breaks in synteny between wheat and *Am. muticum*. Due to the homozygous nature of the lines, homozygous *muticum*-specific SNPs are indicative of the site of introgression. Reads derived from an introgressed segment that aberrantly map to a non-introgressed region will map at the same position as the wheat reads coming from that region and result in heterozygous SNP calls with *muticum*-specific and wheat-specific alleles found at the same position. Therefore, to locate introgressions, we searched for genomic blocks with reduced mapping coverage, homozygous *Am. muticum*-specific SNPs and few heterozygous *Am. muticum*-specific SNPs. We identified introgressions using 1Mbp

genomic windows and then defined the borders to a higher resolution using 100Kbp genomic windows. Figure 1A shows an example of this macro-level visualisation of introgression line DH65, which has a 51.29Mbp segment on the telomere of the short arm of chr4D, and a 139.6Mbp monosomic deletion on the short arm of chr5B. Macro-level genome plots for all lines can be seen in fig. S1.

Fig. 1. Identifying introgressed *Am. muticum* segments using whole-genome sequencing data. A Visualising introgressed segments and structural variation in introgression lines using WGS sequencing data. Each point represents the deviation in mapping coverage compared to the parent lines in 1Mbp windows across Chinese Spring RefSeq v1.0. Windows within assigned *Am. muticum* introgression blocks are coloured red. Introgression line DH65, which has a 51.29Mbp introgressed segment on chr4D and a 139.6Mbp monosomic on chr5D. **B** IGV image showing junction at the right-hand side of chr4D segment in the introgression line DH65 (fig. 1A), spanned by both illumina and nanopore data from DH65. The first four tracks show mapped illumina WGS data, the fifth track shows assembled nanopore contig for DH65 and the bottom track shows high confidence genes from the RefSeq v1.1 annotation.

Using this approach, we confirm the existence of 100% of segments previously identified with KASP genotyping (18). However, we were able to resolve the locations of segment junctions to a much higher resolution than previous methods, due to the limited marker density available for KASP genotyping and the inability of GISH to resolve segments below ~20Mbp. In addition, we were able to uncover two previously unreported segments that have been subsequently validated by KASP genotyping (18); a 17.39Mbp on the telomere of chr7D of DH195 and a 22.68Mbp segment on the telomere of chr5D in DH121. We also identified a new 3.99Mbp segment on chr6D of DH15 that we subsequently validated using 2 KASP markers, WRC1873 and WRC1890. All precise segment positions are listed in table S2.

By manually searching the boundary regions using IGV, we were able to identify segment end positions for 33/42 segment ends (78.6%). Looking solely at crossovers (excluding telomere substitutions) and excluding junctions from non-independently derived segments, we could identify the precise crossover point between wheat and *Am. muticum* in 12/17 (70.6%) cases. Of the remaining junctions, two were narrowed down to within 100kbp and three had complex structures with duplication events that prohibited precise localisation. Out of the 12 high resolution junctions, 11 (91.7%) were within 670bp upstream or downstream of a wheat gene, with 8 falling within the gene body itself, suggesting that crossover points may be localised to genes. The remaining junction was 6.75Kbp downstream from the nearest gene.

For line DH65, the pinpointed junction was validated with nanopore data (fig. 1B). 7x coverage of nanopore sequencing data for DH65 was produced and mapped to the RefSeq v1.0 along with the illumina data. Nanopore reads spanned the breakpoint between *Am. muticum* and wheat at the right-hand side of the 51.6Mbp chr4D segment, adding confidence to the illumina identification. We assembled these mapped nanopore reads using wtdbg2 (19) with relaxed parameters to include reads that were clipped due to high divergence between wheat and *Am. muticum*, producing a contig that spans the junction. The nanopore contig spans the entire junction, including regions to which neither the illumina reads from *Am. muticum* or DH65 map. These regions appear to have elevated SNP density, explaining the gaps in mapping.

For introgression lines with KASP genotyping verification, WGS data may offer an affordable tool to aid breeders to identify precise location and size of these segments. To assess the sequencing depth requirements and locate the position and size of introgressed segments using coverage deviation alone, we downsampled the paired-end Illumina sequencing data from 2 lines for which we have identified very precise positions of the junction borders; DH65, and DH92, to 1x, 0.1x, 0.01x and 0.001x to choose the lowest depth at which we could still resolve segment position. 0.01x was the lowest depth that still provided good resolution (fig. S2).

Introgression crossing induces large chromosomal aberrations and homoeologous pairing/recombination

In addition to introgression sites, we have identified large deletions and duplications, many of which were whole arm or whole chromosome in scale, based on the deviations in mapping coverage not attributable to

introgressions. Within the 17 lines examined, 12 lines (70.6%) have 1 or more very large chromosomal aberrations exceeding 140Mbp. These include 3 examples of whole chromosomes being duplicated, along with the corresponding deletion of one of its homoeologous chromosomes (fig. 2A), 2 arm or near-arm deletions (fig. 2B), 2 near arm monosomies at identical positions between 2 independently derived lines (fig. 2C), and 2 whole-chromosome monosomies (fig. 2E).

Fig. 2. Large chromosomal aberrations induced by introgression crossing. Each point shows mapping coverage deviation compared to the wheat parents in 500Kbp windows across the genome. **A** Corresponding duplication and deletion seen in both DH lines of BC3 pair, caused by pairing of a duplicated chr1A and chr1B. Mapping coverage deviation of 1 at the end of chr1A and chr1B indicates a large translocation between chr1A and chr1B has taken place in duplicated chr1A+chr1B pair and discontinuous mapping coverage deviation change towards beginning of chr1A and chr1B suggests lots of smaller translocation events. **B** Arm deletions on homoeologous chromosomes of DH pair. Possibly caused by pairing of chr4A and chr4B due to inverted synteny between these chromosomes. **C** Monosomic deletions at the same position in two independently derived lines. **D** Signal of duplicated translocations within homoeologous group 6, at similar positions in two independently derived lines. **E** Monosomic deletion of chr1A in DH195. **F** Complex chromosome structure, with a chimeric chr5A-chr5D replacing chr5B.

Homoeologous translocations resulting in the non-reciprocal transfer of genetic material can be detected through mapping coverage deviation, indicated by a duplication and deletion in corresponding homoeologous regions. We can also use differences within a DH pair (table S1) to infer what genetic events must have taken place to give rise to the segregation patterns we see from DH lines derived from the same BC3 line. We see evidence of homoeologous pairing both from duplicated/deleted pairs of chromosomes, such as in DH355 and DH124 (fig. 2A) and from signatures of duplicated translocations, such as in DH202 and BC2F420 (fig. 2D). In BC2F420, the event involves two chromosomes, and in DH202 it involves all three chromosomes in the homoeologous group, where a region on chr6B and chr6D have duplicated and translocated side by side replacing the homoeologous regions on chr6A.

Genome assembly and annotation of *Am. muticum*:

To facilitate identification of introgressed genes both for differential expression analysis and to find candidate introgressed resistance genes, we produced a draft genome assembly for *Am. muticum* comprising most of the gene space. 178Gbp of Oxford Nanopore reads were produced for *Am. muticum* accession JIC 2130012 (previously *Aegilops mutica*; GRU Store Code: T2130012). 156Gbp of these reads passed filtering and were assembled using Flye (20). After polishing the resulting assembly with long and short reads and resolving haplotigs, the assembly comprised 96,256 contigs and was 2.53Gbp in length, with an N50 of 75.5Kbp. Mapping the nanopore reads back to the assembly and computing coverage across single-copy genes allowed us to estimate the genome size at ~4.99Gbp. This is smaller than the c-value estimate for *Ae. mutica* of 6.174Gbp (21), suggesting variation in genome size between accessions. This is supported by variation in c-value seen across wild *Aegilops tauschii* accessions (21). Although heavily collapsed, spanning just 50.1% of the estimated genome size, BUSCO analysis (22) revealed that 94.2% of the expected gene space was assembled unfragmented. Gene annotation using root and shoot transcripts, proteome, and ab initio evidence resulted in 63,032 gene models, 48,606 of which received transferred functional annotations from other well annotated species. 26,597 *Am. muticum* genes were identified as orthologous to a corresponding wheat gene and assigned to an orthologue pair.

Expression of introgressed genes and trans-regulatory impact on the transcriptome

To explore how introgressed genes are expressed and to understand the impact of introgression crossing on the wheat transcriptome, we produced mRNAseq data for six of the introgression lines. *Am. muticum* genes introgressed into each line were identified using orthologue assignments and DNA read mapping evidence. RNA was mapped to a pseudo genome (ABDT) constructed by concatenating RefSeq v1.0 with the draft *Am. muticum* genome assembly; this allows us to distinguish RNA deriving from wheat from RNA deriving from *Am. muticum* genes. Differential expression was called between each introgression line and the wheat parent lines using DESeq2 (23), substituting expression counts of wheat genes in introgressed regions with those of the introgressed *Am. muticum* orthologue. Between 20.1% and 32.9% (27.2% across all lines) of introgressed genes were called as differentially expressed ($\text{abs}(\log_2\text{FC}) \geq 1$ and adj. p-value ≤ 0.05 in both parental

184 comparisons) when compared to the expression in the parent lines of the wheat orthologue they replaced (fig.
 185 3). Between 57.1% and 87.3% (73.5% across the lines) of these differentially expressed introgressed genes
 186 were downregulated or silenced in the introgression line.

Fig. 3. Differential expression within introgression lines. Differential expression in 6 introgression lines, looking at introgressed genes compared to the orthologue they replaced in the parent lines, and background wheat genes compared to the expression in the wheat parent lines. The height of the bar represented the percentage of genes differentially expressed within introgressed and background regions for each line. The number above each bar is the number of genes called as differentially expressed.

187

We hypothesised that suppression of wheat genes in an introgressed or deleted region would lead to a change in the expression of homoeologous copies of that gene in the other subgenomes to compensate for the loss of expression and partially or completely restore the overall level of expression across the triad. The results of multiple approaches support a lack of overall rebalancing of the triad following suppression of one of the copies, both in the 400Mbp introgressed region on chr5D of BC3F45 and in the deletion of chr7D in DH161 (fig. 4). In the introgression, there were 262 triads with the D copy introgressed and called as downregulated; just 1 (0.382%) of these triads had one or more homoeologues called as upregulated; this compares to 4 (0.0509%) triads out of 7863 background triads having the A or B homoeologue upregulated. For the deletion, out of 1294 triads with D copy deleted and therefore not expressed, just 5 triads had one or more homoeologues upregulated (0.386%); this compares to 13 triads (0.170%) out of 7637 background triads having the A or B homoeologue upregulated. Despite the percentage of upregulated homoeologues being lower in the background, the difference is not significant (Fisher's exact test two-tailed p-values of 0.151 and 0.167, respectively) and the number of triads with upregulated homoeologues is very small regardless.

Fig. 4. Expression profile across introgressed and a deleted region and their homoeologous regions. Exploring triads **A** Within chr5D:1-400Mbp introgression in BC3F45 where chr5D copy is replaced with *Am. muticum* orthologue and **B** within chr7D in DH161 where the chr7D copy deleted. i. DESeq2 processed $\log_2FC$ of expression compared to Paragon binned into 10Mbp windows ii. Macro level structure in 1Mbp windows. Each point represents the deviation in mapping coverage compared to the parent lines in 1Mbp windows across Chinese Spring RefSeq v1.0. Windows within assigned *Am. muticum* introgression blocks are coloured red; iii. Triad expression balance in triads where the D copy has been deleted or replaced by introgressed gene, comparing the wheat parent Paragon with the introgression line; iv. $\log_2FC$ of A, B and D in triads in which the D copy has been deleted or replaced by an introgressed gene.

To complement the above approach and consider homoeologues whose expression may have changed but not sufficiently to be called as significant by DESeq2, we looked at the $\log_2$ fold change ($\log_2FC$) in DESeq2 normalised expression counts. Plotting the $\log_2FC$ in 10Mbp windows across the chromosomes illustrates the correlation between introgression boundaries and alteration of expression as well as a lack of general expression change in the homoeologous regions. For the introgression and the deletion, we compared the $\log_2FC$ of the A and B homoeologues of triads where the D copy had been introgressed or deleted with the $\log_2FC$ of the A and B homoeologues of a control set of triads where the D copy is not differentially expressed (fig 4A iv and 4B iv). We found no statistically significant difference between the test and control sets (two-tailed t test p-values: deletion = 0.142; introgression = 0.0656). This indicates that, like the proportion of DEGs, the change in expression counts of homoeologues in which the D copy has been downregulated/silenced does not change beyond that expected by chance in triads in which the D copy is present and not differentially expressed.

In lines possessing the chr5D introgression or the chr7D deletion, the balance of expression in the triads shifted towards D suppression (fig. 4A iii), although this suppression was less extreme for the introgression due to a partial compensation from the introgressed *Am. muticum* genes. Across chr5D:1-400Mbp, the region introgressed in BC34F45, 66.4% of triads are balanced in Paragon, compared to just 1.56% of triads balanced in the introgression line, with 74.4% being D suppressed. This further supports a lack of compensation as the parental triad balance seen in the wheat parents is not retained, with triads that were balanced becoming D suppressed and triads that were D suppressed remaining D suppressed.

To explore the indirect impact of expression on the wheat background caused by trans effects and not altered copy number, we also looked at genes in genomic windows not deviating in coverage compared to the wheat parent lines (fig. 3). Between 0.182% and 2.43% (107-1268 genes; 0.890% across all lines) of these wheat genes were differentially expressed compared to the wheat parents. To assess whether any specific gene functions were enriched in the differentially expressed genes we looked for enriched GO Terms (fig. S3). We found some terms to be enriched, suggesting a non-stochastic impact on background transcription, however differences between lines suggests that the nature of trans impact depends on the genes introgressed/disrupted elsewhere in the genome. Some terms are enriched in more than 3 lines, suggesting these are commonly affected in response to the large structural changes. These are oxidoreductase activity, oxidation-reduction process, tetrapyrrole binding, catalytic activity, carbohydrate metabolic process, cofactor binding, which are enriched in downregulated genes; and ion binding, hydrolase activity, and catalytic activity, which are enriched in upregulated background genes.

Identifying candidate introgressed genes underlying *Am. muticum* derived rust resistance

Combining high resolution definition of segment positions with phenotypic information allows us to compare lines and more accurately predict the location of genes underlying the phenotype of interest. Combining this with a genome assembly of the introgressed wild relative allows us to identify introgressed genes that act as

candidate genes for the phenotype of interest. We explored this approach using previously reported rust resistance phenotypes as a case study. Two of the lines that we sequenced, DH92 and DH121 (fig. 5), have complete resistance at the seedling stage to Kansas isolates of *Puccinia striiformis tritici* (stripe/yellow rust) (17). DH92 also displays chlorotic adult resistance to leaf rust and partial resistance to stem rust that is absent in DH121. These lines have overlapping 5D segments, the positions of which were refined to 533.2-566.1Mbp (32.9Mbp) in DH92, and 544.1-566.1Mbp (22Mbp) in DH121. The source of the stripe rust resistance is likely within the overlapping 22.68Mbp region, and the source of leaf/stem rust resistance is likely within the 10.9Mbp region unique to DH92.

Fig. 5. Identifying candidate introgressed resistance genes. Introgression lines DH92 and DH121 possess a partially overlapping introgressed segment on chr5D, a common resistance phenotype to stripe rust but a differential resistance phenotype to leaf and stripe rust. A Macro-level structure of the D subgenome of DH92 and DH121 (no segments on A or B subgenomes). Each point represents the deviation in mapping coverage compared to the parent lines in 1Mbp windows across Chinese Spring RefSeq v1.0. Windows within assigned *Am. muticum* introgression blocks are coloured red. B Identifying resistance genes uniquely introgressed in DH92 and DH121 and thus candidates for the stripe rust resistance shared between the two lines.

Using a mapping-based approach to the pseudo ABDT genome and combining with functional annotation, we identified 13 complete nucleotide-binding, leucine-rich repeat (NLR) immune receptors uniquely introgressed in these two lines. 12 of these have a syntenic wheat orthologue within the overlapping region of the 5D segments. 2 of the NLRs display unique NB-ARC domain signatures (< 80.0% protein identity) when compared to 16 assembled wheat cultivars and wheat progenitor genomes; *Ae. tauschii*, *Triticum urartu*, and *T. turgidum* ssp. *dicoccoides*. 10 of the NLRs are within a 597.34Kbp cluster, which includes the 2 NLRs with novel NB-ARC domains, which are anchored within 19.06Kb from one another. We also identified 2 ABC transporters uniquely introgressed, both of which have 5D orthologues with over 97.5% protein identity, and 7 protein kinase genes uniquely introgressed. 6 of these are LRR protein kinases and 1 has a leactin-like domain and a

protein kinase domain. 3 of these protein kinases are highly diverged at the protein level compared to the closest protein in wheat (52.2%, 74.2% and 77.0%). NLRs, ABC transporters and LRR protein kinases have all been previously implicated in resistance to stripe, leaf and stem rust (24–26). Gene candidates are detailed in table S6.

We identified 3 wall-associated protein kinases (WAKs), and 3 protein kinases uniquely introgressed in DH92 with orthologues or proximal to orthologues of wheat genes in the 10.9Mbp non-overlapping region of the 5D segment. All but 1 of these were found on the same 104.7Kbp contig. 2 of the WAKs are orthologues of TaWAK388 and TaWAK390 on 5D and 1 is orthologous to TaWAK255 on chr4A. Wall-associated kinases have previously been associated with leaf rust adult plant resistance (APR) (27). Two of the protein kinases are identical at the protein level and are most similar to TaWAK387 just upstream of the TaWAK388 and TaWAK390. These may be truncated tandem duplications of this WAK. Unlike the other uniquely introgressed genes identified, the WAKs have some reads mapping to them in most of the introgression lines but only in DH92 is the coverage uniform across their lengths. This likely suggests that these are uniquely introgressed in DH92 and thus can remain as resistance candidates, but similar *Am. muticum* WAKs present in other lines are falsely mapping to these. This is supported by a lack of mapping across the rest of the contig in the other lines, unlike DH92. Gene candidates are detailed in table S7.

Discussion

Using whole-genome sequencing to pinpoint wild relative introgressions in wheat – an affordable approach to better characterise introgression lines

The current approach for studying synthetic introgression lines prior to deployment in breeding programmes relies on cytogenetic and genotyping techniques, namely GISH and KASP (12, 18). *De novo* discovery of SNPs to produce higher density KASP markers has improved the resolution and as demonstrated here, recent KASP genotyping was able to detect most of the introgressed segments identified with our approach. However, these approaches lack the resolution needed to unpick the precise size and location of segments and will likely miss small segments without the guidance of WGS data to identify areas in which additional markers should be deployed. We observe this with the new chr6D segment, the small chr7D segment in DH195 and chr5D segment in DH121, the latter two of which are novel sources of disease resistance.

We have demonstrated how whole genome sequencing data can be used to define introgressions to a very high resolution as well as resolve large-scale structural changes in these lines. Downsampling has shown that if we don't require SNP information, only 0.01x sequencing coverage is required to pinpoint the junctions of known introgressed segments to a comparable resolution. Overlaying this information with KASP genotyping will undoubtedly provide an affordable method to characterise sets of synthetic introgression lines more accurately and comprehensively.

Introgressed segments nested within complex genomic structures, such as in DH202 (fig. S1), can only be inferred in conjunction with cytogenetic data and/or segregation patterns of DH pairs. In this example, the part of the introgression signal from chr7D appears on chr7A instead, due to a 7D-7A translocation. Some introgression segment boundaries, such as the left-hand border of chr2A in DH15, can be identified but are difficult to pinpoint precisely due to structural complexities around the junction. Therefore, caution is advised when relying on introgression assignments provided by WGS data alone, particularly for complex lines with several large introgressions/deletions/duplications. However, for most lines, where genomic structure is simpler, this approach is robust, and is nevertheless an improvement on lower resolution methods, even if only to identify confounding structural complexities that would otherwise have been missed.

We have found that crossover points between wheat and *Am. muticum* mostly take place within or adjacent to genes. Previous work has shown crossovers between wheat and wild relatives are enriched in gene rich regions (28), which mirrors recombination rates along the genome (29). Here we have achieved sufficient resolution to reveal these wild relative crossovers are taking place not only in regions of open chromatin and increased recombination rate, but within the genes themselves. This may be because these are the points at which synteny is sufficiently high for recombination to take place.

Genomic instability following alien introgression crossing

We have illustrated that structural disruption is common in introgression pre-breeding material, including homoeologous pairing and recombination, and duplications and deletions up to chromosome size. This is likely caused by the *Am. muticum*-induced suppression of the *Ph1* locus which is currently used to induce recombination between wheat and wild relative chromosomes. This *Ph1* suppression enables pairing of chromosomes from different subgenomes, which can cause translocations between subgenomes and unstable chromosome pairings that can cause large deletions and duplications. Forced pairings between wheat and wild relative chromosomes in the F1 cross are likely also responsible for structural change. The DH process could be inducing structural change, but the fact we see this in lines that haven't undergone this process excludes the DH process from being the primary explanation. We can be confident that *Ph1* is contributing to the structural change due to the evidence of subgenome pairings we see tied to some of these chromosomal aberrations. Backcrossing in the development of these lines will have removed much of the initially induced structural change, so what we see here is a subset of structural change initially generated. An awareness of chromosomal aberrations is important for breeders using these lines in their breeding programmes. It will be important to identify the location of the *Ph1* suppressor in *Am. muticum* and other wild relatives that have an innate *Ph1* suppression system, such as *Ae. speltooides* (30) to prevent segments being carried forward into breeding programmes that contain a *Ph1* suppressor that could generate further genomic disruption.

Smaller scale variation in mapping coverage suggests that structural disruption is not restricted to large crossovers and provides some evidence of many potentially smaller scale structural variations in these lines that we can't accurately assess with our available data, such as transposable element mobilisation and smaller

INDELs. It will be important to assess the nature and extent of such variation in the future. To understand this type of variation, we will need a genome assembly of an introgression line and the wheat parents used in the cross, or a genome assembly of the wheat parent and long read data from the introgression line like we produced here for DH65. Unfortunately, structural variation between available chromosome-level genome assemblies and Paragon and Pavon76 is too great for structural variants arising from genome shock to be distinguished from existing structural variation between the cultivars.

Given that we see large-scale structural change as a by-product of most synthetically derived introgressions, perhaps similar structural changes have been induced from historic natural introgressions of wild relative material. We are increasingly becoming aware of the abundance of introgressions throughout wheat germplasm (31, 32); it is possible that many of the structural changes/rearrangements we see in wheat accessions has its origin in unstable chromosome pairings induced by natural introgression events and could be driving the evolution of wheat genomes beyond the addition of novel genes.

Gene expression following alien introgression, lack of homoeologous balancing, and impact on the wheat transcriptome

Many of the introgressed genes are differentially expressed compared to the wheat orthologue replaced. This makes sense biologically due to the very different genomic background the *Am. muticum* genes have been introgressed into, with many factors likely underlying these observed differences such as mismatches between transcription factor binding sites of *Am. muticum* genes and the transcription factors found in wheat. Two previous studies have explored the expression of genes in wheat introgression lines with barley (33) and *Aegilops longissima* (34) introgressed. Due to the differences in methods used for different studies, it is difficult to compare the total proportion of DEGs. However, both previous studies also show that many introgressed genes are differentially expressed with most of these being downregulated or silenced.

We expected to see some balancing of expression in the homoeologues of genes heavily downregulated or silenced due to introgression or deletion events, but we didn't observe significant alteration in the expression of homoeologues. This lack of response suggests that if large-scale balancing of triad expression does take place, it must require selection pressure, which these synthetic lines lack. Now that genome assemblies are available for wheat cultivars possessing many wild relatives introgressions (31) that have undergone extensive artificial selection in a wheat background, it will be interesting to analyse how these introgressed regions are expressed and whether balancing of triad expression arises after a period of selection.

Due to the complex nature of the introgression line genomes, it is difficult to distinguish between the effect of introgressions and deletions/monosomies on the background expression. In the future, analysing the expression of introgression lines with less complex genomes may allow us to better unpick the trans-regulatory effects of introgressed genes from those of deleted or duplicated wheat genes. We see some commonly enriched GO Terms in the genomic background that may be linked with cellular stress or loss of

cellular homeostasis; this conclusion is supported by conclusions drawn in the wheat/barley introgression line (33). The lines without these enriched GO Terms have less disruption overall, with fewer differentially expressed genes in the background and thus may either not have sufficient genomic stress to trigger these responses or lack sufficient sample size of DEGs to call significance.

A case study for uncovering candidate introgressed genes underlying phenotypes of interest

Combining high resolution detection of introgressed segment borders with phenotype information and a genic assembly of *Am. muticum* has enabled us to identify likely regions for novel resistances and produce lists of candidate genes. This will allow breeders to target regions to incorporate into elite lines and to analyse further to identify causal resistance genes. It also acts as a template that can be built upon for new methods to unpick traits of interest in sets of introgression lines.

We identified the probable region of stripe rust resistance in DH92 and DH121 as being within the 22.68Mbp overlapping region of the chr5D segment. The small size and telomeric position of this segment makes it conducive for use in breeding. Within this region, we have identified candidate resistance genes, including 3 novel NLRs and 3 novel LRR Pkinase proteins. We searched for other classes of genes that have been cloned for stripe rust resistance (35), such as hexose transporter, wheat Kinase-START (WKS), tandem kinase-pseudokinase (TKP), but found no examples of these genes uniquely introgressed in these lines. The DH92 resistance to leaf rust that is not shared with DH121, is only seen in adult plants and to a composite of isolates; this race non-specific APR tends to be more durable and, in combination with the small segment size, makes this resistance another good target for further characterisation. We identified 3 WAKs and 3 protein kinases uniquely introgressed in DH92. Wall-associated kinases have previously been shown to confer resistance to leaf rust that looks similar to APR (27) and protein kinase proteins, such as Yr36, have been implicated in APR (36). It should be noted that any resistance here could be conferred by genes that are absent from the assembly or the annotation. If only interested in either the stripe rust or leaf/stem rust resistance, DH92 and DH121 could be crossed to recover the desired resistance in a smaller segment with less linkage drag.

Accession-level differences between wild relatives used for introgression crossing might underlie differences in phenotype between introgression lines with overlapping segments. This is likely the case for the leaf rust resistance in DH195. DH195 has seedling and adult leaf rust resistance not seen in DH202, DH123, or DH121(17), all of which have 7D/7T segments. Differences in segment size and position between these lines (fig. S1) are insufficient to explain the differential resistance we see as there is no segment in DH195 with no overlaps with the other 3 lines. However, the *Am. muticum* accession used to produce DH195 and DH202 (*Am. muticum* 2130004) is different to that used to produce DH121 and DH123 (*Am. muticum* 2130012) so differences in resistance genes between these accessions in the 16.69Mbp region at the end of the chromosome may underlie resistance. Isolating this telomeric segment from the other DH195 segment and screening for resistance will be able to confirm this and provide a new segment conferring resistance for

breeders to utilise. Due to the *Am. muticum* assembly being produced from *Am. muticum* 2130012, any resistance genes specific to 2130004 would not be present in the assembly and so cannot be assessed.

These lines as well as many other sets of synthetic introgression lines are being phenotyped for a variety of agronomically important traits and genome assemblies for additional wild relatives are likely to be produced in the coming years. The analysis we have described here will work better with improved assemblies in which contiguous introgressed segments can be reconstructed and introgressed content fully assessed. This study provides a schematic that can be followed to characterise introgression lines, identify likely positions of genes of interest within segments and candidate genes underlying the phenotype. This will facilitate more effective utilisation of introgression pre-breeding material in wheat breeding programmes.

Materials and Methods

Introgression line production and selection

As described in (11, 12), *Am. muticum* accessions 2130004/2130012 were crossed with Pavon76 or Chinese Spring. The F1 interspecific hybrid was backcrossed 3 times, once with Paragon, Pavon76 or Chinese Spring and twice with Paragon. To make homozygous lines in which introgressions are stably inherited, the resulting BC3 lines were either pollinated with maize and treated with colchicine to generate doubled haploids (DH) (12) or were selfed 2 or 3 times. 13 DH lines, 3 selfed lines and 1 heterozygous BC line, along with *Am. muticum*, Paragon, Pavon76 and Chinese Spring, were selected for DNA WGS. 12 of the lines belong to a pair or a trio of lines that derive from seed from the same BC1 cross, so common segments are not independently derived. 4 DH and 2 BC lines, along with *Am. muticum*, Paragon, Pavon76 and Chinese Spring, were selected for RNA extraction and sequencing.

WGS, read processing, mapping and SNP calling:

Am. muticum introgression lines and wheat parents were grown in a growth room (16h, 21°C day/8h, 18°C night). Genomic DNA from young leaves was isolated using extraction buffer (0.1 m Tris-HCl pH 7.5, 0.05 m EDTA pH 8.0, 1.25% SDS). Samples were incubated at 65 °C for 1 h before being placed on ice and mixed with ice-cold 6 m NH₄C₂H₃O for 15 min. The samples were then spun down, the supernatant was mixed with isopropanol to pellet the DNA and the isolated DNA was treated with RNase A then purified with phenol/chloroform. DNAs were dissolved in TE (10mM Tris-HCl pH8.0, 0.1mM EDTA). PCR-free libraries were produced from this DNA with >600bp insert sizes (gel size-selection). These were sequenced on Illumina NovaSeq 6000 S4 flowcells to produce 150bp paired-end reads. Adaptors were removed and reads trimmed for quality using Trimmomatic (37) with the following parameters: ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10 SLIDINGWINDOW:4:15 MINLEN:40 CROP:150. Contaminants were screened and removed using Kontaminant (38) with *E. coli*, phage, and vector libraries.

424 The cleaned reads were mapped to the Chinese Spring reference genome RefSeqv1.0 (39) using BWA-MEM
425 (40) with the -M parameter to enable duplicates to be marked. The alignment was filtered using samtools (41):
426 supplementary alignments, improperly paired reads, and non-uniquely mapped reads ($q \leq 10$) were removed.
427 PCR duplicates were detected and removed using Picard's MarkDuplicates (42). Variants were called using
428 samtools mpileup (41) and bcftools call (43) using the multiallelic calling model -m.

429 Initial variant filtering was performed using GATK VariantFiltration (42) to remove INDELs and to retain SNPs
430 with a quality score ≥ 30 and a read depth ≥ 10 for the parental lines and ≥ 5 for the introgression lines.
431 Introgression line SNPs were further filtered as follows: homozygous SNPs were kept if 5 or more reads
432 supporting the alternative allele and the allele frequency (AF) was 1; heterozygous SNPs were kept if 3 or more
433 reads supporting each allele. Sites with more than 3 alleles were removed.

434 In silico karyotyping - calculating mapping coverage deviation compared to wheat parents

435 The number of mapped reads post filtering and duplicate removal was counted across genomic windows
436 (1Mbp and 100Kbp) in RefSeq v1.0 using bedtools makewindows (44) and hts-nim-tools (45) for the wheat
437 parents (Chinese Spring, Paragon and Pavon76) and each introgression line. Mapped read counts were
438 normalised by dividing by the total read number post duplicate removal. Normalised counts of each
439 introgression line were divided by the normalised count of each wheat parent in its crossing history (Paragon +
440 Pavon76 or Paragon + Chinese Spring) and the number closest to 1 was kept as the coverage deviation for that
441 window, under the assumption that the parent with mapping coverage closest to the introgression line is the
442 parental donor in that window. The resulting number reflects the copy number of wheat DNA in that window
443 relative to the wheat parent. A number of 1 indicates that the DNA in that window is present in the same
444 amount as in the parent line. A number approaching 0 suggests either a deletion or an introgression has
445 occurred at that region, and a number of 2 suggests a duplication event has taken place. Intermediate values
446 indicate a heterozygous copy number change. We defined windows with a coverage deviation between 0.8
447 and 1.2 as being 'normal' and not in copy number variation compared to the wheat parents.

448 Identifying *Am. muticum*-specific SNPs and assigning introgressed regions:

449 A set of custom python scripts were used to analyse the coverage deviation files and vcfs and identify the
450 introgression segments in each line. These scripts, alongside more detailed methods, are available at:
451 https://github.com/benedictcoombes/alien_detection. First, we produced 18,496,474 SNPs between *Am.*
452 *muticum* and Chinese Spring that weren't shared with either Paragon or Pavon76. *Am. muticum* SNPs were
453 retained as genome-specific if not shared with Paragon or Pavon76. At heterozygous sites where one allele
454 was specific to *Am. muticum*, the specific SNP was retained. If two alternative alleles present, both were kept if
455 both were specific to *Am. muticum*.

456 Introgression line SNPs were then assigned as *Am. muticum* if matching a muticum specific SNP in position and
457 allele. Sites exceeding 3x mean coverage level were removed; this is a signature of collapsed repeat expansion.

These SNPs were then split into homozygous and heterozygous and binned into 1Mbp windows using bedtools coverage (44). Coverage deviation blocks were defined based on contiguous blocks of 1Mbp windows with coverage deviation < 0.7, with windows within 5Mbp from the previous coverage deviation block being merged. The block was discarded if < 80.0% of constituent windows had a coverage deviation < 0.7. 1Mbp windows containing ≥ 55 homozygous *Am. muticum* specific SNPs and a ratio of homozygous to heterozygous *Am. muticum* specific SNPs ≥ 4 that fell within a coverage deviation block were classified as candidate *Am. muticum* windows. Coverage deviation blocks with $\geq 14\%$ windows assigned as *Am. muticum* using the parameters above were classed as an introgressed segment. These parameters are adjustable and were chosen because they revealed previously known introgression segments and didn't falsely class any known deletions as introgressions. The coverage deviation values in the 100Kbp windows up and downstream of either end of the introgression blocks were examined to determine where the coverage deviation starts to decrease. The first 100Kbp window either side that falls below coverage deviation of 0.7 was assigned as the border window of the segment. To locate the precise position of this junction, the BAM alignment files for *Am. muticum*, Paragon, Pavon76 and the introgression line were loaded into IGV (46). The border window ($\pm 100\text{Kbp}$) identified above was searched manually to find the position where the coverage and SNP profile switches from that of the wheat parents to that of *Am. muticum*.

KASP validation

Genomic DNA was isolated from leaf tissue of 10-day old seedlings in a 96-well plate as described by Thomson and Henry (47). All DH lines were genotyped alongside the three parental wheat genotypes (Chinese Spring, Paragon and Pavon 76) and the two *Am. muticum* accessions as controls. For each KASP™ assay, two allele-specific primers and one common primer were used (Table S1). The genotyping procedure was as described in (48). In summary, the genotyping reactions were set up using the automated PIPETMAX® 268 (Gilson, UK) and performed in a ProFlex PCR system (Applied Biosystems by Life Technology) in a final volume of 5 μl with 1 ng genomic DNA, 2.5 μl KASP reaction mix (ROX), 0.068 μl primer mix and 2.43 μl nuclease free water. PCR conditions were set as 15 min at 94°C; 10 touchdown cycles of 10 s at 94°C, 1 min at 65–57°C (dropping 0.8°C per cycle); and 35 cycles of 15 s at 94°C, 1 min at 57°C. Fluorescence detection of the reactions was performed using a QuantStudio 5 (Applied Biosystems) and the data analysed using the QuantStudio™ Design and Analysis Software V1.5.0 (Applied Biosystems).

Nanopore junction validation

DNA from introgression line DH65 extraction was prepared using ligation sequencing kit SQK-LSK109 and sequenced on a MinION using the R9.4.1_RevD flow cell. Reads were filtered using NanoFilt (49) to remove reads below a quality score of 7 or a length of 1Kbp. Filtered reads were mapped to RefSeq v1.0 using minimap2 (50) with parameters -ax map-on and --secondary=no. Mapped reads around the breakpoint (chr4D:51283000-51595000) were extracted using samtools (41), including clipped portions of mapped reads, and assembled using wtdbg2 (19). The resulting contigs were mapped to RefSeq v1.0 using minimap2 (50) with

parameters -ax map-ont and examined in IGV (46) along with the mapped illumina reads from the parent lines and DH65.

Genome assembly of *Am. muticum*:

DNA from *Aegilops mutica* (now *Am. muticum*) line 2130012 (JIC) was prepared using ligation sequencing kit SQK-LSK109 and sequenced on a MinION using the R9.4.1_RevD flow cell. 178Gbp of raw nanopore reads were filtered using NanoFilt (49), removing reads below a quality score of 7 or a length of 1Kbp. Filtered reads were assembled using the Flye assembler (20). Following nanopore read polishing integrated into Flye, we conducted 2 rounds of pilon (51) polishing using 102Gbp of illumina short reads to correct systematic errors in the nanopore data. Finally, haplotigs that were not collapsed in the assembly were detected and resolved using purge_haplotigs (52). Gene completeness was assessed using BUSCO 3.0.2 (22) with parameters -l viridiplantae_odb10 -species wheat and -m geno.

Estimating genome size:

Filtered nanopore reads were mapped back to the polished genome assembly using minimap2 (50) with parameters -x asm5 and filtered to retain uniquely mapping reads with mapQ >= 10. We computed mean coverage across single-copy genes in the assembly. These were identified by orthofinder (53) with *T. aestivum*, *T. turdigrum*, *Ae. tauschii*, *T. urartu*, *Brachypodium distachyon*, and *Arabidopsis thaliana*. *Am. muticum* genes belonging to an orthogroup with no other *Am. muticum* genes in all species represented at least once and fewer than 20 total genes in the orthogroup, were used. The estimated genome size was calculated as the total number mapped bases / mean coverage across single-copy genes, using post-filtering numbers.

Repeat annotation:

A de novo library of transposable elements was produced using EDTA v1.9.5 (54), using cds sequences from *T. aestivum* to exclude protein-coding regions from the library. The resulting library was aligned to the assembly using blastn (55) and sequences with fewer than 3 full length hits (defined as over 80.0% similarity across more than 80.0% of the length of the element) were removed and remaining sequences were clustered using cd-hit-est v4.6.7 (56) with parameters aL 0.8 aS 0.8 -c 0.8 to remove redundant sequences. The resulting library was used to mask the genome assembly using RepeatMasker v4.07 (57) with parameters -s -no_is -norna -nolow -div 40 -cutoff 225.

Gene annotation:

Ab initio, protein homology and transcriptome evidence were combined to predict protein-coding genes in the *Am. muticum* assembly. First, the *Am. muticum* RNA reads (4 root replicates and 4 shoot replicates) were trimmed using Trimmomatic (37) with the parameters ILLUMINACLIP:BBDUK_adaptor.fa:2:30:12 SLIDINGWINDOW:4:20 MINLEN:20 AVGQUAL:20. The trimmed RNA reads were mapped to the masked

genome using STAR (58). Transcripts were assembled using four independent reference-guided approaches; Trinity (59); StringTie (60), cufflinks (61) and CLASS2 (62). Transdecoder (63) was used on each set of transcripts to produce coding ORFs. PORTCULLIS (64) was used to produce splice information. Uniprot (65) reference proteomes for *T. aestivum*, *Ae. tauschii*, *T. turgidum*, *Oryza sativa*, *B. distachyon*, and *A. thaliana* were aligned to the genome using blastx (55). Mikado (66) was used to merge and refine the transcripts produced by each tool, aided by the splice site information and the protein homology evidence. The same uniprot reference proteomes as above were aligned to the masked genome using TBLASTN (55). Hits with an e-value below 1e-5 and sequence identity greater than 75% within 20Kb were merged using bedtools merge (44). These regions were searched using exonerate (67) to refine protein alignments. To produce a set of transcript annotations to train Augustus (68), Mikado protein coding genes were filtered to retain multi-exon genes with < 80.0% amino acid identity with any other protein, not have a stop codon in the ORF, and be at least 500bp away from another gene. Of the remaining transcripts, the 2000 with the highest mikado score were retained and randomly split into 1800 training genes and 200 testing genes. The AUGUSTUS HMM was trained using these transcripts using etraining followed by 10-fold cross validation using optimize_augustus.pl. To produce initial *ab initio* gene predictions, Augustus was run using this HMM, along with intron hints produced from the STAR bam file using bam2hints from Augustus. EvidenceModeler (69) was used to produce a set of HC gene models with *ab initio*, transcriptome, and protein homology evidence. This set of HC genes was used to retrain Augustus as before, as well as GlimmerHMM (70). Augustus, SNAP (71) (using the rice HMM) and GlimmerHMM were run to produce final *ab initio* gene predictions. These predictions were incorporated with the transcript and protein homology evidence using EvidenceModeler to produce final gene models with the following weightings: mikado transcriptome, weight 14; Augustus gene models - weight 4; exonerate protein evidence - weight 7; SNAP gene models - weight 1; GlimmerHMM gene models - weight 2. Predicted proteins blasted to the TREP database of transposable elements and to TREMBL proteins. Proteins were retained with no significant hit to TREP (72) but at least one significant hit to TREMBL (73). Gene functions were assigned using eggno5 5.0 (74).

Assigning orthologue pairs:

Am. muticum/wheat orthologue pairs were assigned using best reciprocal blast hits between *muticum* and each wheat subgenome independently. *Am. muticum* proteins (extracted and translated from gff) and wheat proteins (taken from IWGSC 1.1 pep.fa file for wheat proteins) were aligned reciprocally using blastp (55) with parameters -outfmt 6 -max_hsps 3 -max_target_seqs 3 -evalue 1e-6. Hits were retained if percentage identity >= 80.0% and alignment length was >= 50.0% query length. An orthologue pair was defined as a pair of proteins whose top hit were each other.

Classifying introgressed genes

RefSeq v1.0 genome and the draft *Am. muticum* assembly were concatenated to form a pseudo ABDT genome. DNA reads from the introgression lines were mapped to this genome and filtered using the same process as

mapping to RefSeq v1.0 alone. Introgressed *Am. muticum* genes in each line were defined as those with a wheat orthologue with a start position that fell within a region labelled as a *Am. muticum* introgression and $\geq 2\times$ mean mapping coverage across its length from the ABDT pseudo genome mapping above.

mRNA extraction, sequencing, alignment, and quantification

Seeds were germinated and grown to the 4th leaf stage in a growth chamber set to 22C, 16-hour days. Plants were well watered and supplemented with nutrients. The 4th leaf was extracted and snap frozen in liquid nitrogen within 20 seconds of sampling and stored at -80C. Three reps of each line were sampled individually. Each sample was ground in a pestle and mortar using liquid nitrogen and a sub-set taken for Total RNA extraction using Qiagen RNeasy plant mini kit (Qiagen, Germany). DNA was removed using a digestion step with RNase-free DNase (Qiagen, Germany). Ratios of 260:280 were measured using a Nanodrop (ThermoFisher Scientific) and samples were only used with ratios between 1.9 and 2.1. Only samples with RNA integrity number (RIN) >8 and a concentration of 50-200ng/ul (Agilent Bioanalyzer) were sequenced.

Library preparation was conducted using automated NEBNext Ultra II Directional RNA-Seq library construction with Poly-A selection and was sequenced in triplicate on an Illumina NovaSeq 6000 S2 flowcell by Genomic Pipelines at the Earlham Institute to produce 150bp paired-end reads. RNA reads were trimmed using Trimmomatic (37) with the parameters ILLUMINACLIP:BBDUK_adaptor.fa:2:30:12 SLIDINGWINDOW:4:20 MINLEN:20 AVGQUAL:20. The gff3 for the high confidence CS genes was concatenated with the gff3 for *Am. muticum* genes. Splice site hints for HISAT2 were produced using extract_splice_sites.py from HISAT-2.0.4 (75). The trimmed reads were mapped to the pseudo ABDT genome using HISAT2 with the splice hint file provided and parameters -k 101 --dta --rna-strandness RF. Non uniquely mapping reads were removed using samtools view -q 40. Stringtie (60) was used to compute gene-level abundances, outputting both raw counts and transcript-per-million (TPM) values.

Differential expression analysis:

For each wheat gene in a region identified as introgressed, they were either removed or their expression count was replaced with that of its *Am. muticum* orthologue. Differential expression analysis between each introgression line and its two wheat parents was performed using DESeq2 (23). A gene was considered to be differentially expressed if it had an adjusted p-value below 0.05 and an absolute $\log_2FC \geq 1$ in both parental comparisons. Differentially expressed genes were partitioned into those in introgressed regions, deleted regions, duplicated regions, and in the unaffected background where coverage deviation is between 0.8 and 1.2.

Testing triad expression balancing

To examine whether genes belonging to triads that have homoeologues that have been replaced by a *Am. muticum* gene or have been deleted, we took a test set of triads (76) that satisfied the following conditions:

the D copy is introgressed or deleted and called as downregulated; the A and B copy are in normal copy number regions (coverage deviation between 0.8 and 1.2); and all homoeologues have normalised expression count across samples ≥ 1 . These were compared to control sets of triads that satisfied the same conditions except the D copy was not called as differentially expressed and had $\text{abs}(\log_2\text{FC}) \leq 0.20$. These sets were used for both the comparison of number of triads with A and/or B homoeologue upregulated and for the comparisons of the mean $\log_2\text{FC}$ of the A and B homoeologues between the test and control sample of triads. The significance of these comparisons was tested using two-tailed Fisher's exact test and two-tailed t test, respectively.

GO Term analysis

We transferred functional GO Term annotation from genes in the RefSeq v1.0 annotation to genes in the RefSeq v1.1 annotation if they shared greater than 99% similarity across greater than 90.0% of their length. Statistically enriched GO Terms within the differentially expressed background gene set were computed using the R package topGO (77) with the following parameters: nodeSize=10; classicFisher test $p < 0.05$ and algorithm="parentchild". Enrichment for GO Terms involved in biological processes was tested against all background genes that fall within windows with mapping coverage deviation between 0.8 and 1.2. Bubble plots were plotted using adapted code.

Identifying introgressed resistance genes:

Am. muticum NLR genes were identified using two parallel methods. The first was to scan the gene models for typical NLR domains using hmmscan from the HMMER package (78) in the filtered protein-coding gene models produced by EVM. The second was to scan the entire genome de novo for complete, functional NLR protein-coding regions using NLRAnnotator (79). This allowed several additional NLR genes whose gene model had been filtered out or was never constructed to be recovered. Other types of genes that have been previously implicated in stripe rust resistance (35) (ABC transporters, Pkinases, hexose transporters, wheat Kinase-START genes, tandem kinase-pseudokinase proteins) were identified from the eggnoG annotation and validated manually using hmmscan and pfam domains. Resistance genes were manually checked using IGV to identify candidates with even sequencing coverage across the genes in DH92 and DH121 only, in the case of the shared stripe rust resistance, and across the genes in DH92 only, in the case of the DH92-specific leaf and stem rust resistance. To reduce the number of genes to manually check, we removed any genes with less than 2x mean mapping coverage across their length in either DH92 or DH121. The gene models were manually curated using the available evidence. For NLR predictions with no gene model but transcriptomic and ab initio evidence, gene models were manually constructed. The novelty of the uniquely introgressed NLRs was tested by extracting the NB-ARC domains using hmmscan and aligning them using blastp to the proteins of HC genes from 10 wheat cultivars, and to the proteins of the wheat progenitor species: *Ae. tauschii*; *T. urartu*; and *T. turgidum* ssp. *dicoccoides*. Hits below 85% identity were considered novel. The novelty of the other protein

627 types was tested by aligning the whole amino acid sequence to the same protein set; proteins below <80.0%
628 were considered novel.

629 References

- 630 1. M. Reynolds *et al.*, Achieving yield gains in wheat. *Plant Cell Environ.* **35**, 1799–1823 (2012).
- 631 2. M. Hao *et al.*, The resurgence of introgression breeding, as exemplified in wheat improvement. *Front.*
632 *Plant Sci.* **11**, 252 (2020).
- 633 3. G. Charmet, Wheat domestication: lessons for the future. *C R Biol.* **334**, 212–220 (2011).
- 634 4. J. J. Valkoun, Wheat pre-breeding using wild progenitors. *Euphytica.* **119**, 17-23 (2001).
- 635 5. D. K. Ray, N. D. Mueller, P. C. West, J. A. Foley, Yield trends are insufficient to double global crop
636 production by 2050. *PLoS One.* **8**, e66428 (2013).
- 637 6. G. Doussinault, A. Delibes, R. Sanchez-Monge, F. Garcia-Olmedo, Transfer of a dominant gene for
638 resistance to eyespot disease from a wild grass to hexaploid wheat. *Nature.* **303**, 698–700 (1983).
- 639 7. B. Friebe, K. S. Gill, N. A. Tuleen, B. S. Gill, Transfer of Wheat Streak Mosaic Virus Resistance from
640 *Agropyron intermedium* into Wheat. *Crop Sci.* **36**, 857–861 (1996).
- 641 8. A. M. Fatih, Analysis of the breeding potential of wheat-Agropyron and wheat-Elymus derivatives. I.
642 Agronomic and quality characteristics. *Hereditas.* **98**, 287–295 (1983).
- 643 9. E. R. Sears, *The transfer of leaf-rust resistance from Aegilops umbellulata to wheat.* (1956).
- 644 10. D. L. Klindworth *et al.*, Introgression and characterization of a goatgrass gene for a high level of
645 resistance to ug99 stem rust in tetraploid wheat. *G3 (Bethesda).* **2**, 665–673 (2012).
- 646 11. J. King *et al.*, A step change in the transfer of interspecific variation into wheat from Amblyopyrum
647 muticum. *Plant Biotechnol. J.* **15**, 217–226 (2017).
- 648 12. J. King *et al.*, Development of Stable Homozygous Wheat/Amblyopyrum muticum (Aegilops mutica)
649 Introgression Lines and Their Cytogenetic and Molecular Characterization. *Front. Plant Sci.* **10**, 34
650 (2019).
- 651 13. S. Khazan *et al.*, Reducing the size of an alien segment carrying leaf rust and stripe rust resistance in
652 wheat. *BMC Plant Biol.* **20**, 153 (2020).
- 653 14. M.-D. Rey *et al.*, Exploiting the ZIP4 homologue within the wheat Ph1 locus has identified two lines
654 exhibiting homoeologous crossover in wheat-wild relative hybrids. *Mol. Breed.* **37**, 95 (2017).

- 655 15. A. C. Martín, M.-D. Rey, P. Shaw, G. Moore, Dual effect of the wheat Ph1 locus on chromosome synapsis
656 and crossover. *Chromosoma*. **126**, 669–680 (2017).
- 657 16. D.-H. Koo, B. Friebe, B. S. Gill, Homoeologous recombination: A novel and efficient system for
658 broadening the genetic variability in wheat. *Agronomy*. **10**, 1059 (2020).
- 659 17. J. P. Fellers *et al.*, Resistance to wheat rusts identified in wheat/ *Amblyopyrum muticum* chromosome
660 introgressions. *Crop Sci* (2020), doi:10.1002/csc2.20120.
- 661 18. S. Grewal *et al.*, A novel approach to develop wheat chromosome-specific KASP markers for detecting
662 *Amblyopyrum muticum* segments in doubled haploid introgression lines. *BioRxiv* (2021),
663 doi:10.1101/2021.09.29.462370.
- 664 19. J. Ruan, H. Li, Fast and accurate long-read assembly with wtdbg2. *Nat. Methods*. **17**, 155–158 (2020).
- 665 20. M. Kolmogorov, J. Yuan, Y. Lin, P. A. Pevzner, Assembly of long, error-prone reads using repeat graphs.
666 *Nat. Biotechnol.* **37**, 540–546 (2019).
- 667 21. J. Pellicer, I. J. Leitch, The Plant DNA C-values database (release 7.1): an updated online repository of
668 plant genome size data for comparative studies. *New Phytol.* **226**, 301–305 (2020).
- 669 22. R. M. Waterhouse *et al.*, BUSCO applications from quality assessments to gene prediction and
670 phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2018).
- 671 23. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data
672 with DESeq2. *Genome Biol.* **15**, 550 (2014).
- 673 24. H. Wang, S. Zou, Y. Li, F. Lin, D. Tang, An ankyrin-repeat and WRKY-domain-containing immune receptor
674 confers stripe rust resistance in wheat. *Nat. Commun.* **11**, 1353 (2020).
- 675 25. S. G. Krattinger *et al.*, A putative ABC transporter confers durable resistance to multiple fungal
676 pathogens in wheat. *Science*. **323**, 1360–1363 (2009).
- 677 26. S. Chen *et al.*, Wheat gene Sr60 encodes a protein with two putative kinase domains that confers
678 resistance to stem rust. *New Phytol.* **225**, 948–959 (2020).
- 679 27. M. Dmochowska-Boguta *et al.*, TaWAK6 encoding wall-associated kinase is involved in wheat resistance
680 to leaf rust similar to adult plant resistance. *PLoS One*. **15**, e0227713 (2020).
- 681 28. M. Nyine *et al.*, Genomic Patterns of Introgression in Interspecific Populations Created by Crossing
682 Wheat with Its Wild Relative. *G3 (Bethesda)*. **10**, 3651–3661 (2020).
- 683 29. L.-J. Gardiner *et al.*, Analysis of the recombination landscape of hexaploid bread wheat reveals genes

684 controlling recombination and gene conversion frequency. *Genome Biol.* **20**, 69 (2019).

685 30. H. Li *et al.*, Introgression of the *Aegilops speltoides* Su1-Ph1 Suppressor into Wheat. *Front. Plant Sci.* **8**,
686 2163 (2017).

687 31. S. Walkowiak *et al.*, Multiple wheat genomes reveal global variation in modern breeding. *Nature.* **588**,
688 277–283 (2020).

689 32. A. M. Przewieslik-Allen *et al.*, The role of gene flow and chromosomal instability in shaping the bread
690 wheat genome. *Nat. Plants.* **7**, 172–183 (2021).

691 33. E. Rey *et al.*, Transcriptome reprogramming due to the introduction of a barley telosome into bread
692 wheat affects more barley genes than wheat. *Plant Biotechnol. J.* **16**, 1767–1777 (2018).

693 34. Z. Dong *et al.*, Genome-wide impacts of alien chromatin introgression on wheat gene transcriptions. *Sci.*
694 *Rep.* **10**, 4801 (2020).

695 35. S. Zheng *et al.*, Characterization and diagnostic marker development for Yr28-rga1 conferring stripe rust
696 resistance in wheat. *Eur. J. Plant Pathol.* **156**, 623–634 (2020).

697 36. J. G. Ellis, E. S. Lagudah, W. Spielmeier, P. N. Dodds, The past, present and future of breeding rust
698 resistant wheat. *Front. Plant Sci.* **5**, 641 (2014).

699 37. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data.
700 *Bioinformatics.* **30**, 2114–2120 (2014).

701 38. G. M. Daly *et al.*, Host subtraction, filtering and assembly validations for novel viral discovery using next
702 generation sequencing data. *PLoS One.* **10**, e0129059 (2015).

703 39. International Wheat Genome Sequencing Consortium (IWGSC) *et al.*, Shifting the limits in wheat
704 research and breeding using a fully annotated reference genome. *Science.* **361** (2018),
705 doi:10.1126/science.aar7191.

706 40. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. **00**, 1–3 (2013).

707 41. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* **25**, 2078–2079 (2009).

708 42. M. A. DePristo *et al.*, A framework for variation discovery and genotyping using next-generation DNA
709 sequencing data. *Nat. Genet.* **43**, 491–498 (2011).

710 43. H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population
711 genetical parameter estimation from sequencing data. *Bioinformatics.* **27**, 2987–2993 (2011).

- 712 44. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features.
713 *Bioinformatics*. **26**, 841–842 (2010).
- 714 45. B. S. Pedersen, A. R. Quinlan, hts-nim: scripting high-performance genomic analyses. *Bioinformatics*. **34**,
715 3387–3389 (2018).
- 716 46. J. T. Robinson *et al.*, Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
- 717 47. D. Thomson, R. Henry, Single-step protocol for preparation of plant tissue for analysis by PCR.
718 *BioTechniques*. **19**, 394–7, 400 (1995).
- 719 48. S. Grewal *et al.*, Development of Wheat-Aegilops caudata Introgression Lines and Their Characterization
720 Using Genome-Specific KASP Markers. *Front. Plant Sci.* **11**, 606 (2020).
- 721 49. W. De Coster, S. D’Hert, D. T. Schultz, M. Cruts, C. Van Broeckhoven, NanoPack: visualizing and
722 processing long-read sequencing data. *Bioinformatics*. **34**, 2666–2669 (2018).
- 723 50. H. Li, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. **34**, 3094–3100 (2018).
- 724 51. B. J. Walker *et al.*, Pilon: an integrated tool for comprehensive microbial variant detection and genome
725 assembly improvement. *PLoS One*. **9**, e112963 (2014).
- 726 52. M. J. Roach, S. A. Schmidt, A. R. Borneman, Purge Haplotigs: allelic contig reassignment for third-gen
727 diploid genome assemblies. *BMC Bioinformatics*. **19**, 460 (2018).
- 728 53. D. M. Emms, S. Kelly, OrthoFinder: phylogenetic orthology inference for comparative genomics.
729 *Genome Biol.* **20**, 238 (2019).
- 730 54. S. Ou *et al.*, Benchmarking transposable element annotation methods for creation of a streamlined,
731 comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
- 732 55. C. Camacho *et al.*, BLAST+: architecture and applications. *BMC Bioinformatics*. **10**, 421 (2009).
- 733 56. W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide
734 sequences. *Bioinformatics*. **22**, 1658–1659 (2006).
- 735 57. N. Chen, Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc*
736 *Bioinformatics*. **Chapter 4**, Unit 4.10 (2004), doi:10.1002/0471250953.bi0410s05.
- 737 58. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. **29**, 15–21 (2013).
- 738 59. M. G. Grabherr *et al.*, Full-length transcriptome assembly from RNA-Seq data without a reference
739 genome. *Nat. Biotechnol.* **29**, 644–652 (2011).

- 740 60. M. Pertea *et al.*, StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.
741 *Nat. Biotechnol.* **33**, 290–295 (2015).
- 742 61. C. Trapnell *et al.*, Differential gene and transcript expression analysis of RNA-seq experiments with
743 TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
- 744 62. L. Song, S. Sabunciyan, L. Florea, CLASS2: accurate and efficient splice variant annotation from RNA-seq
745 reads. *Nucleic Acids Res.* **44**, e98 (2016).
- 746 63. B. J. Haas *et al.*, De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for
747 reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
- 748 64. D. Mapleson, L. Venturini, G. Kaithakottil, D. Swarbreck, Efficient and accurate detection of splice
749 junctions from RNA-seq with Portcullis. *Gigascience.* **7** (2018), doi:10.1093/gigascience/giy131.
- 750 65. UniProt Consortium, UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515
751 (2019).
- 752 66. L. Venturini, S. Caim, G. G. Kaithakottil, D. L. Mapleson, D. Swarbreck, Leveraging multiple transcriptome
753 assembly methods for improved gene structure annotation. *Gigascience.* **7** (2018),
754 doi:10.1093/gigascience/giy093.
- 755 67. G. S. C. Slater, E. Birney, Automated generation of heuristics for biological sequence comparison. *BMC*
756 *Bioinformatics.* **6**, 31 (2005).
- 757 68. K. J. Hoff, M. Stanke, Predicting Genes in Single Genomes with AUGUSTUS. *Curr Protoc Bioinformatics.*
758 **65**, e57 (2019).
- 759 69. B. J. Haas *et al.*, Automated eukaryotic gene structure annotation using EVIDENCEModeler and the
760 Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
- 761 70. W. H. Majoros, M. Pertea, S. L. Salzberg, TigrScan and GlimmerHMM: two open source ab initio
762 eukaryotic gene-finders. *Bioinformatics.* **20**, 2878–2879 (2004).
- 763 71. I. Korf, Gene finding in novel genomes. *BMC Bioinformatics.* **5**, 59 (2004).
- 764 72. T. Wicker, D. E. Matthews, B. Keller, TREP: a database for Triticeae repetitive elements. *Trends Plant Sci.*
765 **7**, 561–562 (2002).
- 766 73. A. Bairoch, R. Apweiler, The SWISS-PROT protein sequence data bank and its supplement TrEMBL.
767 *Nucleic Acids Res.* **25**, 31–36 (1997).
- 768 74. J. Huerta-Cepas *et al.*, eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated

- 769 orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314
770 (2019).
- 771 75. D. Kim, J. M. Paggi, C. Park, C. Bennett, S. L. Salzberg, Graph-based genome alignment and genotyping
772 with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
- 773 76. R. H. Ramírez-González *et al.*, The transcriptional landscape of polyploid wheat. *Science*. **361** (2018),
774 doi:10.1126/science.aar6089.
- 775 77. A. Alexa, J. Rahnenfuhrer, Gene set enrichment analysis with topGO (2020).
- 776 78. R. D. Finn, J. Clements, S. R. Eddy, HMMER web server: interactive sequence similarity searching.
777 *Nucleic Acids Res.* **39**, W29–37 (2011).
- 778 79. B. Steuernagel *et al.*, The NLR-Annotator Tool Enables Annotation of the Intracellular Immune Receptor
779 Repertoire. *Plant Physiol.* **183**, 468–482 (2020).

780 Acknowledgements

781 We would like to acknowledge BBS/E/T/000PR9816 (NC1 - Supporting EI's ISPs and the UK Community with
782 Genomics and Single Cell Analysis) for data generation and BB/CCG1720/1 for the physical HPC infrastructure
783 and data centre delivered via the NBI Computing infrastructure for Science (CiS) group.

784 Funding:

785 BBSRC Core Capability Grant BB/CCG1720/1 (AH)

786 BBSRC funded Norwich Research Park Biosciences Doctoral Training Partnership grant BB/M011216/1 (BC)

787 BBSRC Designing Future Wheat grant BB/P016855/1 and its constituent work packages DFW WP4 Data Access
788 and Analysis

789 Author contributions:

790 Conceptualization: AH, JL, IPK, BC

791 Methodology: BC, RJ, RR-P

792 Formal analysis: BC

793 Investigation: JF, SG, CY, SE

794 Writing - Original Draft: BC

795 Writing - Review & Editing: BC, AH, RR-P, JK, JF

796 Visualization: BC

797 Supervision: AH

798 Funding acquisition: AH, JK, IPK

799 **Competing interests:**

800 Authors declare that they have no competing interests

801 **Data and materials availability:**

802 Sequencing data produced as part of this study, along with the *Am. muticum* assembly is available at:

803 https://opendata.earlham.ac.uk/wheat/under_license/toronto/Hall_2021-10-08_wheatxmuticum. [This data

804 will be uploaded to the European Nucleotide Archive (ENA) while the paper is under review]. *Am. muticum*

805 illumina sequencing data available at:

806 [https://opendata.earlham.ac.uk/wheat/under_license/toronto/Grewal_et_al_2021-09-](https://opendata.earlham.ac.uk/wheat/under_license/toronto/Grewal_et_al_2021-09-13_Amybylopyrum_muticum/)

807 [13_Amybylopyrum_muticum/](https://opendata.earlham.ac.uk/wheat/under_license/toronto/Grewal_et_al_2021-09-13_Amybylopyrum_muticum/). The Chinese Spring sequencing data used is available from ENA (study

808 PRJNA393343; runs SRR5893651 and SRR5893652). The Paragon sequencing data used is available from ENA

809 (study PRJEB35709; runs ERR3728451, ERR3760033, ERR3760405 and ERR3728448). Custom scripts used for

810 introgression detection are available at: https://github.com/benedictcoombes/alien_detection.