**Peer Review File**

**Manuscript Title:** Origin and evolution of the bread wheat D genome

## Reviewer Comments & Author Rebuttals

### Reviewer Reports on the Initial Version:

Referee #1 (Remarks to the Author):

The study by Krattinger and colleagues presents a valuable contribution to our understanding of Aegilops tauschii, the donor of the D genome to bread wheat (Triticum aestivum). By establishing a detailed pangenome resource, this work sheds light on the genetic composition and evolutionary dynamics of the bread wheat D genome, revealing previously unrecognized complexities. This study is particularly timely and useful, addressing key questions about the D-genome's evolution that have been of interest to the scientific community. Their work uncovers complexities not previously recognized, marking a significant advancement in the field of agricultural science and plant genomics.

The authors compiled an extensive k-mer matrix from 920 Ae. tauschii accessions, effectively identifying 493 non-redundant accessions. This panel covered a broad geographical range and highlighted three main lineages (L1, L2, L3), with further subdivisions in L2. The creation of chromosome-scale assemblies for 46 selected accessions from these lineages provided a detailed view of genetic diversity, capturing 99.3% of the genetic diversity present in the Ae. tauschii panel.

The researchers went on to identify two disease resistance genes and conducted a comprehensive haplotype analysis across a complex disease resistance locus. They specifically targeted rust resistance genes, which are critical due to the prevalence and impact of rust diseases on wheat crops.

The study revealed that the bread wheat D genome is derived not only from the Ae. tauschii L2E-1 subpopulation but also from contributions from all L2 subpopulations, as well as L1 and L3. This insight challenges the traditional view of a simple linear progression in crop domestication, suggesting a more intricate pattern of genetic mixing and hybridization.

Utilizing the innovative 'Missing Link Finder' pipeline, the study highlighted significant L3 introgressions in the bread wheat D genome. These introgressions are much larger than previously known, indicating a potential for greater genetic diversity and breeding potential within bread wheat than currently recognized.

The study underscores the importance of wild relatives in crop breeding and conservation genetics. It emphasizes the significance of genetic diversity in crop improvement and provides practical insights for wheat breeding. The research's findings pave the way for innovative approaches in crop improvement, offering a deeper understanding of plant genetics and evolution.

In conclusion, this study is a commendable contribution to wheat research, bridging gaps in our understanding of wheat evolution and contributing to future advancements in wheat improvement.

My detailed comments:

1. Pangenome Establishment and Analysis:
The authors compiled a comprehensive k-mer matrix from 920 Ae. tauschii accessions, identifying 493 non-redundant accessions spanning a broad geographical range. These were categorized into three main lineages: L1, L2, and L3, with further subdivisions in L2. Chromosome-scale assemblies were created for 46 selected accessions, using pseudomolecule assemblies for one representative each of L1, L2, and L3 lineages. This approach effectively bypasses the need for generating extensive HiC data for each assembly. It's important to acknowledge, however, that this method inherently limits the ability to study large-scale structural variations. While this limitation does not impact the current study's objectives, it is a notable constraint of the resource for potential future research endeavors. I think this limitation should be discussed or at least mentioned briefly.

2. Gene Discovery and Disease Resistance:
The study's focus on rust resistance genes is critical given the substantial impact of rust diseases on wheat crops.
• For stem rust resistance, a new gene (Sr66) was identified. As stem rust poses a significant threat to wheat globally, finding new resistance genes is crucial. However, I find this part of the manuscript is lacking some clarity.
◦ The number of Pgt isolates tested is three in the manuscript and 5 in supplementary table 11.
◦ Does Sr66 offers resistance similar to Sr33 against Ug99 races. I assume not?
◦ The result of the testing should be elaborated. Is Sr66 giving additional or more resistance or just to other races compared to Sr33 and other Sr genes? Are there other known resistance genes to these races? Why did you chose those races?
◦ Is Sr66 a gene that will potentially offer new resistance to stem rust/ offer additional resistance when stacked other stem rust resistance genes?
◦ It would also be beneficial to mention the scoring scale used in supplementary table 11. Stackman scale? And also a reference to the five letter code.

• For leaf rust resistance, the identification of the WTK gene is well-presented. However, quantitative results of the increased susceptibility following VIGS should be provided for completeness. And it might just be me but it took me quite a while to understand what the assemblies AUS 18913 and AUS 18911 are. It might enhance the reader’s understanding if those assemblies could be referenced in the figure caption where they are first mentioned.

3. Significance of Ae. tauschii L3 Introgressions:
The 'Missing Link Finder' pipeline's revelation of significant L3 introgressions in the D genome suggests a greater genetic diversity and breeding potential in bread wheat than previously recognized. However, the rationale behind focusing specifically on the L3 lineage over L1 or L2 is not entirely clear. It would be helpful to provide more context or introduction as to why the L3 lineage was selected for this investigation. Is it due to its relative under-exploration, or are there specific characteristics or historical bottlenecks that make L3 particularly interesting for genetic studies and potential breeding improvements? There appears to be a gap in the introduction that might leave readers seeking more context for a full understanding.

4. The data is well-presented, with clear figures and supplementary materials enhancing the study's comprehensiveness. The study appears to use statistical methods appropriately. However regarding data availability:
• The DRYAD-links are not working including the one under which the genome annotations should be available.
• I can’t find PRJNA956839 on NCBI. The genomics data has to be made available please.
• The link https://wheat.pw.usda.gov/GG3pangenome/wheat/D/taus_home.php) requires a user name and password….?

Referee #2 (Remarks to the Author):

This article describes an impressive resource for the wheat research community and an insight into the diversity of the Ae. tauschii and evolution of the D genome. These include 46 lower coverage HiFi PaCBio assemblies from across the three main lineages of Ae. tauschii; Three very high-quality assemblies and annotations from each of the lineage; WGS resequence data from 350 Ae. tauschii accessions extending existing resources.

The paper compiles diversity collections from previous work and the additional accession resequenced here. Using Structure they can split the previous L2 lineage from 3 subpopulations to 4 which can be mapped onto specific geographical regions.

They go on to describe the assembles and structural variation. This is poor, the authors fail to even include the simplest whole genome comparisons. I'm sure even the simplest structural analysis would have provided insight into the evolution of these different genomes. Moreover, it would have been useful to add the wheat D genome. This would have provided a robust link to the later parts of the paper.

The gene annotation is a big piece of work and not having an analysis of gene content is a large omission, they could have easily extended the Core/shell gene sets identified in Zhou et al. Even a simple presence/absence would have added insight with an analysis of structural variation around an example gene family eg the NLR proteins. Again contrasts with the D genome of bread wheat. The authors need to include the four high-quality genomes that already exist (zhou et al 2021) in all these analyses or explain why they have been excluded. All of this could have added value to the haplotype study in the final section.

The next section of the paper describes the utility of multiple high-quality genomes in identifying important resistance loci and how having a pangenome reveals the complexity of these loci. While this is of value, it probably could have been achieved with existing pangenome resources. It also steers away from the major thrust of the paper and would be more appropriate as a separate publication.

The authors go on to investigate L3 introgression into the D genome of wheat and demonstrate foci landraces with the L3 introgression around Georgia. The only site to date where the L3 lineages have been located. The degree of introgression decreases with distance from these foci, this is a novel and interesting finding. It would be great to discuss based on the map where maybe other L3 linage members could be collected. Eg. eastern Iran? Is the yellow dot here a new introgression? The bread quality loci is also really interesting and it would have been interesting to map this loci across the map.

The L3 lineage study is really good, however, the attempt to extend this to other lineages through a haplotype approach needs to be rethought I think there is something really powerful here. The authors might think to try matching accessions D genome haplotype and size of haplotype block with the geographic position of both the landrace and Ae. tauschii might provide the insight missing here.

Finally, what is a real missed opportunity is the fact that these lineages now have structural variants and annotated genomes. This new information isn't used in the final parts of the paper.

The discussion starts by stating that: "The comprehensive genomic resources generated in this study enabled haplotype analysis and cloning of rust resistance genes and they offered a detailed insight into the composition and origin of the bread wheat D genome." I am not entirely sure this is true. I think that the later parts of this paper could have been conducted with existing resources and that in fact, the lack of utilisation of these new data sets is a real criticism of the paper.

There are several minor corrections such as "In Figure 3c replace Chinese Spring with position (Mb)" but these would be picked up during production so I have not commented.

Referee #3 (Remarks to the Author):

Cavalet-Giorsa and colleagues have built and analyzed a tremendous resource for wheat improvement and the study of polyploid genomes in general. The article was a pleasure to read - this is quite an achievement for a large-scale and largely descriptive genomics project. I appreciate the time and effort that went into both the resource generation and the thorough analysis. Despite my overall very positive view of the manuscript, I have some comments and questions. I hope that these will help improve future drafts.

General comments:

At several points, the authors state that the A. tauschii 'pangenome' facilitated the discovery of things. I have some thoughts here.
- First off, I think it's an open question as to whether a collection of genomes and resequencing data constitutes a 'pangenome'. Typically you'd need some resource that integrates these (e.g. graph, pan-gene set, etc.) to use this term. I wonder if it wouldn't be more accurate to state that the collection of new genomes facilitated the discoveries. Certainly the kmer analysis presented here does a nice job of integrating, but my guess is that the majority of the readers will expect a graph or similar if you use the word 'pangenome'. Personally, I don't think 'pangenome' adds anything and would just drop this term in favor of something that more accurately describes the resources presented here.
- Second, it's common in this type of paper to state that the new genomes facilitated discovery, but rarely do the analyses back up this claim (i.e. the same discovery could be made from short reads mapped to a single reference). The work here clearly required the multiple genomes … I wonder if stating exactly how the multiple references were used in the abstract and summary might add some strength to the statement about the utility of the multiple references.

Throughout, it is often not clear which types of data were used to do each analysis. The first result is a good example of this. The title is "pangenome of At", but the very first result is a low-coverage reference free pop gen analysis of short reads. This is not really a pangenome in the modern definition, and just a resequencing analysis (albeit a clever one). Even though I don't really care about the wording, I think many readers would find this problematic.

The short-read reference-free pop gen work using de novo kmers is clever, but potentially biased. As the authors state in the supplement, getting high confidence kmers out is computationally tricky, but it is also biologically noisy. The resource is a collection of un-polished low-coverage short reads. These have sequencing error, which can likely be controlled through downstream kmer count distribution pruning. However, they will also miss a lot of sequence. A low-coverage 8x library (median coverage = 8.3X) will completely miss both meiotic homologous haplotypes ~0.4% of the time, and if heterozygous, will miss one of the haplotypes ~6% of the time. I understand that this could be a small effect since A. tauschii is so homozygous, but it is not necessarily. I'd be more convinced if you saw very high agreement rates between the kmers generated from the reference genomes and those from a downsampled (to 8X) short read library from the same genotype (you have this for 46 libraries). At least some context for the potential sources of error and an estimate of the % of variants that have gone un-called due to shallow coverage should be presented. Also, I

wonder if removing low-frequency kmers or low coverage libraries could have caused you to miss some subtle population genetic structure, as observed [336-337]. Commenting on this might help the reader understand the possible causes. *(this does not need to be done, just an idea for the future ... I'll also note that we have had some luck pruning to kmers that are only ever found once (or never) in all of the reference genomes; this gets rid of quite a lot of noise due to genotyping at multiple positions)*.

The reference genomes seem to be of very high contiguity relative to their CCS coverage (for the ~20X libs). However, (and maybe this is just missing in the methods) it doesn't appear that these genomes were illumina polished, despite the availability of short reads. Depending on the CCS coverage (both low and high coverage can be problematic), many thousands of short INDELs will be present in a CCS-only genome. These are typically corrected with Illumina data. If these were not corrected in your assemblies, please comment on which reported SVs could potentially be due to the systematic sequencing errors in CCS data. Since you are using a minimum of 50bp for INDELs, this probably gets rid of most, but not all error-derived SVs.

Rag-Tag assemblies can produce false and overlapping joins. It looks like visual inspection was accomplished, but no systematic treatment of potential duplicated contig ends was accomplished. I looked at the reference-level genomes and didn't see anything problematic. However, I couldn't check the shallower CCS synteny guided assemblies, since these were not made available to the reviewers (as far as I can tell). Please ensure that links to these assemblies are available for the next round.

Personal preference, but circos plots typically obscure information. I have trouble seeing what's happening in Fig. 1C. Perhaps a linear representation would emphasize the points more clearly.

I wonder if there's any way to distinguish between the "admixed Aegilops" vs "recurrent hybridization" hypotheses. Recurrent hybridization looks more likely given the variation in landraces to me.

**Author Rebuttals to Initial Comments:**

Referee #1 (Remarks to the Author):

The study by Krattinger and colleagues presents a valuable contribution to our understanding of Aegilops tauschii, the donor of the D genome to bread wheat (Triticum aestivum). By establishing a detailed pangenome resource, this work sheds light on the genetic composition and evolutionary dynamics of the bread wheat D genome, revealing previously unrecognized complexities. This study is particularly timely and useful, addressing key questions about the D-genome's evolution that have been of interest to the scientific community. Their work uncovers complexities not previously recognized, marking a significant advancement in the field of agricultural science and plant genomics.

The authors compiled an extensive k-mer matrix from 920 Ae. tauschii accessions, effectively identifying 493 non-redundant accessions. This panel covered a broad geographical range and highlighted three main lineages (L1, L2, L3), with further subdivisions in L2. The creation of chromosome-scale assemblies for 46 selected accessions from these lineages provided a detailed view of genetic diversity, capturing 99.3% of the genetic diversity present in the Ae. tauschii panel.

**Author reply:** We realized during the revision that we provided a wrong number in the initial manuscript version. 99.3% represents the percentage of *k*-mers captured by the 493 non-redundant *Ae. tauschii* accessions. The percentage of *k*-mers present in the 46 selected *Ae. tauschii* accessions is 72.5%. We investigated the *k*-mer distribution in more detail and found that most of the *k*-mers that are not captured by the 46 selected accessions are rare. We have summarized this result in the new Extended Data Fig. 1c and refer to it in the main text as follows (lines 145-149): 'The majority of the *k*-mers that were not captured in the 46 high-quality assemblies are rare and were found in fewer than 5% of the accessions that make up the *Ae. tauschii* diversity panel (Extended Data Fig. 1b, c, Supplementary Table 5).'

The researchers went on to identify two disease resistance genes and conducted a comprehensive haplotype analysis across a complex disease resistance locus. They specifically targeted rust resistance genes, which are critical due to the prevalence and impact of rust diseases on wheat crops.

The study revealed that the bread wheat D genome is derived not only from the Ae. tauschii L2E-1 subpopulation but also from contributions from all L2 subpopulations, as well as L1 and L3. This insight challenges the traditional view of a simple linear progression in crop domestication, suggesting a more intricate pattern of genetic mixing and hybridization.

Utilizing the innovative 'Missing Link Finder' pipeline, the study highlighted significant L3 introgressions in the bread wheat D genome. These introgressions are much larger than previously known, indicating a potential for greater genetic diversity and breeding potential within bread wheat than currently recognized.

The study underscores the importance of wild relatives in crop breeding and conservation genetics. It emphasizes the significance of genetic diversity in crop improvement and provides practical insights for wheat breeding. The research's findings pave the way for innovative approaches in crop improvement, offering a deeper understanding of plant genetics and evolution.

In conclusion, this study is a commendable contribution to wheat research, bridging gaps in our understanding of wheat evolution and contributing to future advancements in wheat improvement.

My detailed comments:

**Comment 1.1.** Pangenome Establishment and Analysis: The authors compiled a comprehensive k-mer matrix from 920 Ae. tauschii accessions, identifying 493 non-redundant accessions spanning a broad geographical range. These were categorized into three main lineages: L1, L2, and L3, with further subdivisions in L2. Chromosome-scale assemblies were created for 46 selected accessions, using pseudomolecule assemblies for one representative each of L1, L2, and L3 lineages. This approach effectively bypasses the need for generating extensive HiC data for each assembly. It's important to acknowledge, however, that this method inherently limits the ability to study large-scale structural variations. While this limitation does not impact the current study's objectives, it is a notable constraint of the resource for potential future research endeavors. I think this limitation should be discussed or at least mentioned briefly.

**Author reply:** We fully agree with this comment. We have added the following statement in the methods section (paragraph 'RagTag assembly of the 43 pangenome accessions') to highlight the limitations of reference-guided assemblies for large structural variant detection: 'While being great resources for gene discovery and comparative analyses, reference-guided assemblies are limited in their ability to study large structural rearrangements.' As indicated in the manuscript, our structural variation analysis only considered variants from 50 bp - 100 kb to account for this limitation. The median contig N50 value of 45 Mb in our assemblies should accommodate the discovery of most structural variants up to 100 kb.

**Comment 1.2.** Gene Discovery and Disease Resistance: The study's focus on rust resistance genes is critical given the substantial impact of rust diseases on wheat crops.

• For stem rust resistance, a new gene (Sr66) was identified. As stem rust poses a significant threat to wheat globally, finding new resistance genes is crucial. However, I find this part of the manuscript is lacking some clarity.

◦The number of Pgt isolates tested is three in the manuscript and 5 in supplementary table 11.

**Author reply:** Apologies for the discrepancy. We have updated the manuscript to mention five isolates to tally with the supplementary table.

◦Does Sr66 offers resistance similar to Sr33 against Ug99 races. I assume not?

**Author reply:** *Sr66* confers resistance only against a subset (n=2) of the isolates that were found to be avirulent against *Sr33* (n=5), based on the analysis of transgenics (Supplementary Table 12).

◦The result of the testing should be elaborated. Is Sr66 giving additional or more resistance or just to other races compared to Sr33 and other Sr genes? Are there other known resistance genes to these races? Why did you chose those races?

**Author reply:** The isolates used in Supplementary Table 11 were previously published (Arora et al., 2019 Nature Biotechnology 37:139-143; Gaurav et al., 2022 Nature Biotechnology 40:422-431). These isolates were chosen for being diagnostic for the genes being analysed by GWAS in these studies, e.g. 75ND717C (race QTHJC) is avirulent on *Sr33*, but virulent on *Sr45* and *Sr46*. The isolates in

Supplementary Table 12 were chosen as they represent three distinct phylogenetic clades of *Puccinia graminis* f. sp. *tritici* (Patpour et al., 2022 Frontiers in Plant Science 13: 882440), as well as being virulent on cv. Fielder (which is believed to carry *Sr6*), a prerequisite for the transgenic assays.

Based on the stem rust phenotypes of the 126 *Ae. tauschii* Illumina-sequenced genotypes (Arora et al., 2019; Gaurav et al., 2022), and identifying by BLAST those lines that appear to contain only *Sr33* or only *Sr66*, we were able to postulate that *Sr66* is also effective against isolate 75ND717C (race QTHJC) which is also known to be effective against *Sr33* (Arora et al., 2019).

◦Is Sr66 a gene that will potentially offer new resistance to stem rust/ offer additional resistance when stacked other stem rust resistance genes?

**Author reply:** We have updated the manuscript to provide more explicit information pertaining to this question as follows (lines 212-213): 'Our analysis thus far showed that *SrTA1662* confers resistance to a subset of the *Pgt* isolates avirulent on *Sr33*.' It would be difficult to answer the question of whether *Sr66* offers new resistance because so few isolates, if any, are known to be virulent on *Sr33*. Thus, even if we expanded the repertoire of isolates in the study we would be unlikely to find an isolate that is virulent on *Sr33* and avirulent on *Sr66*. An alternative way of answering the question, albeit beyond the scope of the present study, would be to identify the effector(s) recognized by *Sr66* and *Sr33* to see if they are different or the same.

◦It would also be beneficial to mention the scoring scale used in supplementary table 11. Stackman scale? And also a reference to the five letter code.

**Author reply:** In the updated manuscript we now make the following statement in Supplementary Table 12 (=Supplementary Table 11 in the first version): 'Infection types were assessed according to the USDA scoring system developed by Stackman et al., 1962' and 'We used the five letter code for the *Pgt* race nomenclature developed for the BGRI International Differential Core Set (https://www.fao.org/agriculture/crops/rust/stem/stem-pathotypetracker/stem-racenomenclature/en/).'

**Comment 1.3.** For leaf rust resistance, the identification of the WTK gene is well-presented. However, quantitative results of the increased susceptibility following VIGS should be provided for completeness. And it might just be me but it took me quite a while to understand what the assemblies AUS 18913 and AUS 18911 are. It might enhance the reader's understanding if those assemblies could be referenced in the figure caption where they are first mentioned.

**Author reply:** We have performed an additional *Lr39* VIGS experiment in which we quantified the relative leaf rust biomass using qPCR. The results are summarized in the new Extended Data Fig. 5c. In summary, we observe significant differences in rust biomass in the *Lr39* VIGS plants compared to the controls. We have also updated the methods section as follows (lines 834-841): 'For leaf rust biomass quantification, DNA was extracted from *P. triticina*-inoculated leaves using the CTAB method[48]. DNA concentrations were measured using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific). A 20 µl qPCR reaction containing Power SYBR Green PCR Mix (Applied Biosystems 4367659), ~25 ng of DNA, and primers specific to the *Puccinia* 28S large subunit (LSU) or the Internal transcribed spacer (ITS) region[94] and Triticeae elongation factor-specific primers[95] was run using the ABI QuantStudio 6 Flex Real-Time PCR machine. The $2^{-\Delta\Delta CT}$ method was used to

normalise rust gene amplification values relative to the *Ae. tauschii* elongation factor endogenous control.'

The original *Sr33* cloning paper reported different haplotypes spanning the *Mla* locus in CPI110799 (original donor of *Sr33*), AUS18913 (haplotype 3) and AUS18911 (haplotype 5). Since these accessions were included in our high-quality assemblies, we used them in our *SrTA1662* haplotype analysis. In the revised manuscript, we have described this in the figure legend (Fig. 2).

**Comment 1.4.** Significance of Ae. tauschii L3 Introgressions: The 'Missing Link Finder' pipeline's revelation of significant L3 introgressions in the D genome suggests a greater genetic diversity and breeding potential in bread wheat than previously recognized. However, the rationale behind focusing specifically on the L3 lineage over L1 or L2 is not entirely clear. It would be helpful to provide more context or introduction as to why the L3 lineage was selected for this investigation. Is it due to its relative under-exploration, or are there specific characteristics or historical bottlenecks that make L3 particularly interesting for genetic studies and potential breeding improvements? There appears to be a gap in the introduction that might leave readers seeking more context for a full understanding.

**Author reply:** It is L3's genetic distinctness and geographical restriction that made this lineage particularly exciting to study the spatial dynamics of introgressions. In principle, Missing Link Finder can be applied to other *Ae. tauschii* lineages and wild wheat relatives. The geographical restriction of L3 to Georgia and the genetic distinctness, however, made L3 the most interesting example to test Missing Link Finder. We have clarified this in the main text as follows (lines 272-274): 'The genetic distinctness and geographical restriction of *Ae. tauschii* L3 makes this lineage an ideal example to study the spatial dynamics of introgressions.'

**Comment 1.5.** The data is well-presented, with clear figures and supplementary materials enhancing the study's comprehensiveness. The study appears to use statistical methods appropriately. However regarding data availability:

• The DRYAD-links are not working including the one under which the genome annotations should be available.

**Author reply:** We sincerely apologize for this blunder. The DRYAD links for (i) the genome assemblies, HiC contact maps, annotations, VCF file and IBSpy dataset, and (ii) *k*-mer matrix, are now working. These are reviewer links. The data availability statement in the main text contains the actual links that will be made public upon acceptance of the manuscript:

(i) https://datadryad.org/stash/share/WlvT773KDmYWYDT-2JgyMb0plLHygdQfXEpciCfAV7Y

(ii) https://datadryad.org/stash/share/2XEz5WHXO9AeXkdGPhhIw64acNL6IQcdi6AR4lt_TNI

(ii) https://datadryad.org/stash/share/J3MPbFzZcImPN-unGTF1cJHAXFgil70IKdxxuKQAR_s

(ii) https://datadryad.org/stash/share/Fx8ZCRb6n48vrdtHVA8rLRU70kgalXu7auv8q1Rf5dE

• I can’t find PRJNA956839 on NCBI. The genomics data has to be made available please.

**Author reply:** We have provided the following Bioproject reviewer link on NCBI: PRJNA956839 containing the raw sequencing data: https://dataview.ncbi.nlm.nih.gov/object/PRJNA956839?reviewer=u6mhk79nomjt3fafqi2l8s2c76

• The link https://wheat.pw.usda.gov/GG3pangenome/wheat/D/taus_home.php) requires a user name and password….?

**Author reply:** The user name and password for the database are “wheat_guest” and “@progenitor”, respectively. Please note that the password protection will be removed before publication of the manuscript.

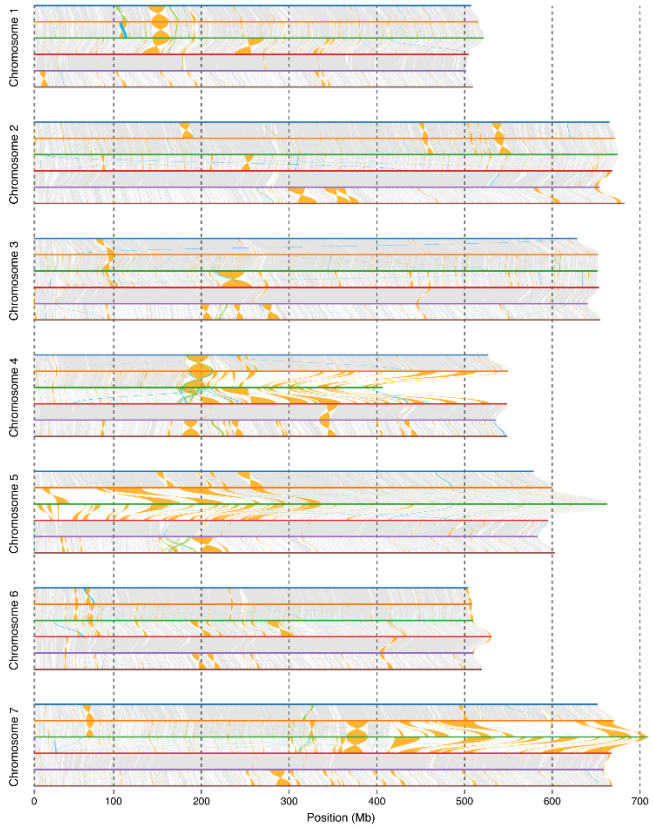
Referee #2 (Remarks to the Author):

This article describes an impressive resource for the wheat research community and an insight into the diversity of the Ae. tauschii and evolution of the D genome. These include 46 lower coverage HiFi PaCBio assemblies from across the three main lineages of Ae. tauschii; Three very high-quality assemblies and annotations from each of the lineage; WGS resequence data from 350 Ae. tauschii accessions extending existing resources.

The paper compiles diversity collections from previous work and the additional accession resequenced here. Using Structure they can split the previous L2 lineage from 3 subpopulations to 4 which can be mapped onto specific geographical regions.

**Comment 2.1.** They go on to describe the assemblies and structural variation. This is poor, the authors fail to even include the simplest whole genome comparisons. I'm sure even the simplest structural analysis would have provided insight into the evolution of these different genomes. Moreover, it would have been useful to add the wheat D genome. This would have provided a robust link to the later parts of the paper.

**Author reply:** We conducted a whole-genome synteny analysis to graphically display inversions, translocations and duplications (see figure below). In this new analysis, we also included the wheat cv. Chinese Spring D genome and two high-quality assemblies of *Ae. tauschii* Lineage 1 and 2, accessions AY61 and T093, respectively, published in the Zhou et al., 2021 study. Given the limited length of a Nature article, we prefer not to include this analysis in the manuscript. We are prepared to show this analysis as a supplementary figure in case the Reviewer and Editor think that this would be useful.




**Supplementary Fig. X**. Synteny and genomic rearrangements across chromosome-level assemblies of *Ae. tauschii* accessions TA10171 (L1; red), T093 (L1; purple), TA1675 (L2; orange), AY61 (L2; green), TA2576 (L3; brown) and the Chinese Spring D genome (blue). Syntenic regions are indicated by grey ribbons connecting the chromosomes of different accessions, inversions by orange ribbons, translocations by green ribbons and duplications by blue ribbons. Translocations and duplications greater than or equal to 40 kb are shown.

**Comment 2.2.** The gene annotation is a big piece of work and not having an analysis of gene content is a large omission, they could have easily extended the Core/shell gene sets identified in Zhou et al. Even a simple presence/absence would have added insight with an analysis of structural variation around an example gene family eg the NLR proteins. Again contrasts with the D genome of bread wheat. The authors need to include the four high-quality genomes that already exist (zhou et al 2021) in all these analyses or explain why they have been excluded. All of this could have added value to the haplotype study in the final section.

**Author reply:** In the revised manuscript, we provide an extended core/dispensable gene set analysis, including our three chromosome-scale *Ae. tauschii* assemblies (L1, L2, L3), the four high-quality assemblies produced by Zhou et al. (2021) (3 L1 and 1 L2), AL8/78, and the Chinese Spring D genome. In total, we identified 52,788 families across the nine genomes, including 18,835 core families and 33,953 shell/cloud families. These numbers are very similar to the analysis performed by Zhou et al. (2021), which found 50,323 gene families and 20,055 core families. We have summarized the gene analyses results in the new Supplementary Table 7 and included the following sentences in the main manuscript (lines 162-166): '*De novo* annotation of the three chromosome-scale assemblies revealed 43,511 to 44,275 protein-coding genes (Extended Data Table 1). We included five previously generated high-quality *Ae. tauschii* assemblies and the bread wheat D genome for a gene family analysis. In total, we identified 52,722 gene families, of which 18,835 were shared across the nine genomes (core) and 33,953 families were found in some but not all assemblies (Supplementary Table 7).'. We also updated the methods section (lines 634-635 – paragraph 'Repeat and gene annotation') as follows: 'We used OrthoFinder (v2.5.4) with default parameters to perform gene family analysis.'

Please note that the haplotype study presented in the final section of our manuscript is based on genetic groups resulting from the large diversity panel and is not based on single accessions. The haplotype analysis is thus not influenced by individual assemblies/accessions, which makes our haplotype analysis very robust (Supplementary Note 3).

**Comment 2.3.** The next section of the paper describes the utility of multiple high-quality genomes in identifying important resistance loci and how having a pangenome reveals the complexity of these loci. While this is of value, it probably could have been achieved with existing pangenome resources. It also steers away from the major thrust of the paper and would be more appropriate as a separate publication.

**Author reply:** The analyses we present would have been difficult, if not impossible, to present using the four existing assemblies in Zhou et al. because these assemblies (derived from accessions AY61, T093, AY17, XJ02) are not the original donors of *Sr33*, *SrTA1662* (*Sr66*) or *Lr39*. Indeed, BLAST analysis of these four genomes with the aforementioned resistance genes produce at best partial coding sequence hits (results not shown). The importance of using genome assemblies of *R* gene-containing reference accessions in *R* gene cloning and designation has been highlighted in several recent studies, see e.g. *Sm1* (Walkowiak et al., 2020, Nature 588:277-283), *Yr27* (Athiyannan et al., 2022, Nature Genetics 54:227-231) and *Pm69* (Li et al., 2024, Plant Communications 5:100646) to name but a few. Moreover, the previous *Ae. tauschii* assemblies do not represent an example where the *Sr33* and *SrTA1662* orthologues have naturally recombined. This natural recombinant is a critical requirement for designating *SrTA1662* as a new gene, i.e. *Sr66*. Finally, the high-quality genome assemblies in Zhou et al. (AY17, AY61, T093 and XJ02) were generated from seed sourced from Henan University. We have tried in the past to obtain seed from the Henan University germplasm bank, but without success. This complicates relating genotype to phenotype, a critical aspect in *R* gene cloning.

We essentially agree with the reviewer that with a resource manuscript, as in the present case, it is generally challenging to highlight novel uses of the resource whilst also maintaining a focus. Nevertheless, we are proud of the storyline we have presented here and expect this will enhance the appeal to a wider readership. Resistance gene cloning continues to attract high attention and the work we present here accentuates the relevance of the *Ae. tauschii* genome resources to agriculture. We are concerned that excluding this component from the manuscript would reduce the impact of our study.

**Comment 2.4.** The authors go on to investigate L3 introgression into the D genome of wheat and demonstrate foci landraces with the L3 introgression around Georgia. The only site to date where the L3 lineages have been located. The degree of introgression decreases with distance from these foci, this is a novel and interesting finding. It would be great to discuss based on the map where maybe other L3 linage members could be collected. Eg. eastern Iran? Is the yellow dot here a new introgression? The bread quality loci is also really interesting and it would have been interesting to map this loci across the map.

**Author reply:** Please note that the L3 introgression map (Fig. 3c) is based on patchy genotyping-by-sequencing data, which is insufficient to assess the bread quality locus across the 80,000 wheat accessions. The bread quality locus (Fig. 3e) could only be assessed for the 59 bread wheat landraces for which we produced additional whole-genome sequencing data. The comment about the landraces in northeastern Iran is very interesting. We investigated whole-genome sequencing data of 6 bread wheat landraces with high L3 content (yellow dots) that were collected east of the Caspian Sea. We found that the L3 genomic segments in these 6 landraces are identical by state to western bread wheat landraces from the Caucasus region. These results indicate that the L3 segments in the eastern and western bread wheat landraces share a common origin, which is most likely Georgia. We have summarized the results in the new Supplementary Table 18, which is referred to in caption of figure 3 as follows: 'Eastern bread wheat accessions from Tajikistan with high Jaccard indices carry the same L3 introgression segments compared to bread wheat landraces from Georgia, indicating a common origin of the L3 haplotype blocks (Supplementary Table 18)'.

**Comment 2.5.** The L3 lineage study is really good, however, the attempt to extend this to other lineages through a haplotype approach needs to be rethought I think there is something really powerful here. The authors might think to try matching accessions D genome haplotype and size of haplotype block with the geographic position of both the landrace and Ae. tauschii might provide the insight missing here.

**Author reply:** Please also see our response to comment 1.4 of Reviewer 1. The reason for focusing on L3 was that this lineage is genetically distinct and geographically restricted, which made *Ae. tauschii* L3 a particularly exciting case study. In principle, the pipeline can also be applied to other lineages and wild wheat relatives. It is important to note, however, that the initial analysis is based on 'noisy' genotyping data, which can result in the identification of false positives. A thorough validation of our key findings with additional whole-genome sequencing data (Extended Data Fig. 6c) was a critical step to validate the *Ae. tauschii* L3 results. The generation of whole-genome sequencing data from selected wheat lines will thus be a prerequisite for future studies on other lineages/species and the conclusions must not be based on the genotyping-by-sequencing data alone.

**Comment 2.6.** Finally, what is a real missed opportunity is the fact that these lineages now have structural variants and annotated genomes. This new information isn't used in the final parts of the paper.

**Author reply:** We appreciate the Reviewer's comment. We see the core strength of our haplotype analyses in the fact that it is based on a solid population structure that has been deduced from an extensive diversity panel. In our view, this adds robustness compared to haplotype analyses that are based on single accessions (Supplementary Note 3). Given that structural variants are often accession-specific, we do not see a straightforward way to include this information in the haplotype analysis. To expand the haplotype analysis, we now provide the haplotype information for each of the predicted genes in the high-quality bread wheat assemblies that were used in Fig. 4a. This information is summarized in the new Supplementary Table 20 and will allow researchers and breeders to determine the subpopulation origin of individual genes across 16 bread wheat lines. The new supplementary table is referred to in the main text as follows (line 361-363): 'A list of predicted genes and their corresponding subpopulation origin is provided in Supplementary Table 20'.

**Comment 2.7.** The discussion starts by stating that: "The comprehensive genomic resources generated in this study enabled haplotype analysis and cloning of rust resistance genes and they offered a detailed insight into the composition and origin of the bread wheat D genome." I am not entirely sure this is true. I think that the later parts of this paper could have been conducted with existing resources and that in fact, the lack of utilisation of these new data sets is a real criticism of the paper.

**Author reply:** We refer the reviewer back to our reply to comment 2.3 in which we argue the importance of using reference accessions and being able to link these to phenotypes. Part of the strength of our study is that seed from the vast majority of the sequence-configured panel is available from third party public good germplasm institutions with a proven capacity to export seed worldwide (Supplementary Table 1). This information is explained in the Germplasm Availability statement and summarized in detail in Supplementary Table 1. This table provides to our knowledge the most extensive and up-to-date meta-data summary of published *Ae. tauschii* accessions. The haplotype analysis presented in the later part of the manuscript was only possible because we had generated an extensive whole-genome sequencing data set from a very large diversity panel. This enabled us to define a robust population structure that is supported by a large number of *Ae. tauschii* accessions, which was a critical resource for our haplotype analysis. The haplotype analysis would not have had the same quality and robustness without the data presented in the first parts of this manuscript (Supplementary Note 3). Please also note that the inference of minimal number of hybridizations (Fig. 4b) required the 46 high-quality genome assemblies (Supplementary Note 4).

**Comment 2.8.** There are several minor corrections such as "In Figure 3c replace Chinese Spring with position (Mb)" but these would be picked up during production so I have not commented.

**Author reply:** We have carefully checked the manuscript again. Please note that we realized during the revision that we provided a wrong number regarding the *k*-mers captured by the 46 high-quality assemblies. 99.3% represents the percentage of *k*-mers captured by the 493 non-redundant *Ae. tauschii* accessions. The percentage of *k*-mers found in the 46 selected *Ae. tauschii* accessions is 72.5%. We investigated the *k*-mer distribution in more detail and found that most of the *k*-mers that are not captured by the 46 high-quality assemblies are rare. We have summarized this result in the new Extended Data Fig. 1c and refer to it in the main text as follows (lines 145-149): 'The majority of the *k*-mers that were not captured in the 46 high-quality assemblies are rare and were found in fewer than 5% of the accessions that make up the *Ae. tauschii* diversity panel (Extended Data Fig. 1b, c, Supplementary Table 5).'

Referee #3 (Remarks to the Author):

Cavalet-Giorsa and colleagues have built and analyzed a tremendous resource for wheat improvement and the study of polyploid genomes in general. The article was a pleasure to read - this is quite an achievement for a large-scale and largely descriptive genomics project. I appreciate the time and effort that went into both the resource generation and the thorough analysis. Despite my overall very positive view of the manuscript, I have some comments and questions. I hope that these will help improve future drafts.

General comments:

At several points, the authors state that the A. tauschii 'pangenome' facilitated the discovery of things. I have some thoughts here.

**Comment 3.1.** - First off, I think it's an open question as to whether a collection of genomes and resequencing data constitutes a 'pangenome'. Typically you'd need some resource that integrates these (e.g. graph, pan-gene set, etc.) to use this term. I wonder if it wouldn't be more accurate to state that the collection of new genomes facilitated the discoveries. Certainly the kmer analysis presented here does a nice job of integrating, but my guess is that the majority of the readers will expect a graph or similar if you use the word 'pangenome'. Personally, I don't think 'pangenome' adds anything and would just drop this term in favor of something that more accurately describes the resources presented here.

**Author reply:** We agree that there is no uniform and universally agreed definition of a pangenome. We have removed the word 'pangenome' from the main manuscript and replaced it with 'genome resources' or 'high-quality assemblies'.

**Comment 3.2.** - Second, it's common in this type of paper to state that the new genomes facilitated discovery, but rarely do the analyses back up this claim (i.e. the same discovery could be made from short reads mapped to a single reference). The work here clearly required the multiple genomes … I wonder if stating exactly how the multiple references were used in the abstract and summary might add some strength to the statement about the utility of the multiple references.

**Author reply:** This concern was shared by Reviewer 2 (comment 2.3). The analyses we present would have been difficult, if not impossible, to present using the four previously generated *Ae. tauschii* assemblies because none of the previously generated assemblies are original donors of *Sr33*, *SrTA1662* (*Sr66*) or *Lr39*. The importance of using genome assemblies of *R* gene-containing reference accessions in *R* gene cloning and designation has been highlighted in several recent studies, see e.g. *Sm1* (Walkowiak et al., 2020, Nature 588:277-283), *Yr27* (Athiyannan et al., 2022, Nature Genetics 54:227-231) and *Pm69* (Li et al., 2024, Plant Communications 5:100646) to name but a few. Moreover, the previous *Ae. tauschii* assemblies do not represent an example where the *Sr33* and *SrTA1662* orthologues have naturally recombined. This natural recombinant is a critical requirement for designating *SrTA1662* as a new gene, i.e. *Sr66*.

**Comment 3.3.** Throughout, it is often not clear which types of data were used to do each analysis. The first result is a good example of this. The title is "pangenome of At", but the very first result is a low-coverage reference free pop gen analysis of short reads. This is not really a pangenome in the modern definition, and just a resequencing analysis (albeit a clever one). Even though I don't really care about the wording, I think many readers would find this problematic.

**Author reply:** As indicated in comment 3.1, we have removed the word 'pangenome'. We have clarified throughout the main text, which data sets were used. In particular, we have added the new subheading '*k*-mer matrix generation, redundancy and diversity analyses' in the methods section and improved the order in which methods paragraphs are listed under this subheading.

**Comment 3.4.** The short-read reference-free pop gen work using de novo kmers is clever, but potentially biased. As the authors state in the supplement, getting high confidence kmers out is computationally tricky, but it is also biologically noisy. The resource is a collection of un-polished low-coverage short reads. These have sequencing error, which can likely be controlled through downstream kmer count distribution pruning. However, they will also miss a lot of sequence. A low-coverage 8x library (median coverage = 8.3X) will completely miss both meiotic homologous haplotypes ~0.4% of the time, and if heterozygous, will miss one of the haplotypes ~6% of the time. I understand that this could be a small effect since A. tauschii is so homozygous, but it is not necessarily. I'd be more convinced if you saw very high agreement rates between the kmers generated from the reference genomes and those from a downsampled (to 8X) short read library from the same genotype (you have this for 46 libraries). At least some context for the potential sources of error and an estimate of the % of variants that have gone un-called due to shallow coverage should be presented. Also, I wonder if removing low-frequency kmers or low coverage libraries could have caused you to miss some subtle population genetic structure, as observed [336-337]. Commenting on this might help the reader understand the possible causes. *(this does not need to be done, just an idea for the future ... I'll also note that we have had some luck pruning to kmers that are only ever found once (or never) in all of the reference genomes; this gets rid of quite a lot of noise due to genotyping at multiple positions)*.

**Author reply:** The phylogenetic analyses (Fig. 1b) and ancestry analyses (Extended Data Fig.2) were produced using SNP data, which we now clarify in the main text as follows (lines 133): 'We performed single nucleotide polymorphism (SNP)-based phylogeny (Fig. 1b, Extended Data Fig. 1a) and ancestry analyses (Extended Data Fig. 2) on the diversity panel, which defined four geographically distinct subpopulations for *Ae. tauschii* L2, referred to as L2E-1 (south-western Caspian Sea), L2E-2 (south-eastern Caspian Sea), L2W-1 (Caucasus), and L2W-2 (Turkmenistan and northern Iran), in accordance with the literature'. The SNP calling was done following an established pipeline and the quality of the SNP calling was assessed in Extended Data Fig. 8. The population structure that was used for the haplotype analysis is in agreement with previous publications (Zhou et al. 2021 Nature Plants 7: 774-786).

We observed a 97.5% agreement rate between the *k*-mers generated from the reference genomes and those from the downsampled whole-genome sequencing data. For comparison, the average agreement rate between the *k*-mers generated from the reference genomes and the full whole-genome sequencing datasets was 99.7% (new Supplementary Table 32). We have expanded Supplementary Note 3 as follows: 'Each chromosome-scale wheat assembly is divided into 50-kb windows and the origin of each window is determined based on identity-by-state to the *Ae. tauschii* subpopulations (at *K*=9) using the Identity-by-State Python pipeline (Ahmed et al. 2023, Nature 620, 830-838). We first used KMC3 to

build *k*-mer (*k* = 31) databases from the trimmed Illumina raw data of the 920 *Ae. tauschii* accessions in the diversity panel, pruning *k*-mers that only occurred once. We observed high agreement rates between the *k*-mers generated from high-quality *Ae. tauschii* reference assemblies and the *k*-mers generated from whole-genome sequencing data (Supplementary Table 32).'

The statement made in lines 392-393 'We could assign another 10.7-19.5% of the wheat D genome to L2, but without being able to infer the exact subpopulation' is not based on missing subtle population structure. These segments did not show identity-by-state to any of the *Ae. tauschii* accessions, but the IBSpy values were still in a range that allowed us to assign them to L2. Biologically, this can be explained by L2 haplotype blocks that are not present in our diversity panel because they might be rare or extinct. However, we also realized that certain technical artefacts can lead to an inflation of IBSpy values and hence, to an erroneous interpretation of 'unassigned' windows. We tried to estimate the proportion of 'unassignable' windows that are the result of technical artefacts. The results are now summarized in the new paragraph 'Unassigned subpopulations – technical artefacts vs. true unsampled haplotype blocks' in Supplementary Note 3 and the new Supplementary Table 33. Our estimates reveal that the quality of the wheat reference assembly is the biggest source of possible technical artefacts. Even when accounting for all possible technical artefacts, we still retain around 10% of the bread wheat D genome that cannot be assigned to a particular L2 subpopulation. These segments most likely represent true L2 segments that are not sampled in our diversity panel (possibly extinct L2 haplotypes).

Please note that we realized during the revision that we provided a wrong number regarding the *k*-mers captured by the 46 high-quality assemblies. 99.3% represents the percentage of *k*-mers captured by the 493 non-redundant *Ae. tauschii* accessions. The percentage of *k*-mers found in the 46 selected *Ae. tauschii* accessions is 72.5%. We investigated the *k*-mer distribution in more detail and found that most of the *k*-mers that are not captured by the 46 high-quality assemblies are rare. We have summarized this result in the new Extended Data Fig. 1c and refer to it in the main text as follows (lines 145-149): 'The majority of the *k*-mers that were not captured in the 46 high-quality assemblies are rare and were found in fewer than 5% of the accessions that make up the *Ae. tauschii* diversity panel (Extended Data Fig. 1b, c, Supplementary Table 5).'

**Comment 3.5.** The reference genomes seem to be of very high contiguity relative to their CCS coverage (for the ~20X libs). However, (and maybe this is just missing in the methods) it doesn't appear that these genomes were illumina polished, despite the availability of short reads. Depending on the CCS coverage (both low and high coverage can be problematic), many thousands of short INDELs will be present in a CCS-only genome. These are typically corrected with Illumina data. If these were not corrected in your assemblies, please comment on which reported SVs could potentially be due to the systematic sequencing errors in CCS data. Since you are using a minimum of 50bp for INDELs, this probably gets rid of most, but not all error-derived SVs.

**Author reply:** It has been found in barley (Mascher et al. 2021, Plant Cell 33:1888-1906) and human (Mc Cartney et al. 2022, Nature Methods 19:687-695) that polishing HiFi assemblies with Illumina reads reduced the assembly quality. For these reasons, we chose not to polish our genome assemblies. It is also worth mentioning that since we restricted our SV analysis to the 50 bp to 100 kb range, most homopolymer assembly errors characterized by the HiFi chemistry would not feature in our report.

**Comment 3.6.** Rag-Tag assemblies can produce false and overlapping joins. It looks like visual inspection was accomplished, but no systematic treatment of potential duplicated contig ends was accomplished. I looked at the reference-level genomes and didn't see anything problematic. However, I couldn't check the shallower CCS synteny guided assemblies, since these were not made available to the reviewers (as far as I can tell). Please ensure that links to these assemblies are available for the next round.

**Author reply:** We sincerely apologize again (as we did to Reviewer 1, see Comment 1.5) for failing to provide working links. We have now double checked that the Dryad links to the assemblies are working (see Comment 1.5 of Reviewer 1).

The DRYAD links for (i) the genome assemblies, HiC contact maps, annotations, VCF file and IBSpy dataset, and (ii) *k*-mer matrix, are now working. These are reviewer links. The data availability statement in the main text contains the actual links that will be made public upon acceptance of the manuscript:

(i) https://datadryad.org/stash/share/WlvT773KDmYWYDT-2JgyMb0plLHygdQfXEpciCfAV7Y

(ii) https://datadryad.org/stash/share/2XEz5WHXO9AeXkdGPhhIw64acNL6IQcdi6AR4lt_TNI

(ii) https://datadryad.org/stash/share/J3MPbFzZcImPN-unGTF1cJHAXFgil70IKdxxuKQAR_s

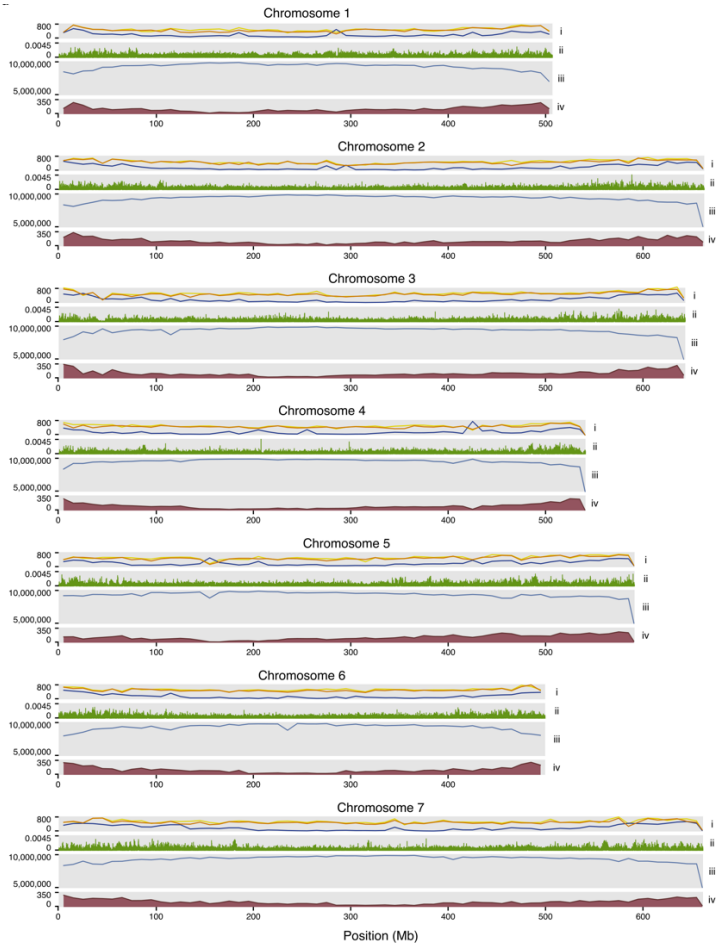
(ii) https://datadryad.org/stash/share/Fx8ZCRb6n48vrdtHVA8rLRU70kgalXu7auv8q1Rf5dE

The Bioproject PRJNA956839 containing the raw sequencing data can be accessed using the following reviewer link:

https://dataview.ncbi.nlm.nih.gov/object/PRJNA956839?reviewer=u6mhk79nomjt3fafqi2l8s2c76

**Comment 3.7.** Personal preference, but circos plots typically obscure information. I have trouble seeing what's happening in Fig. 1C. Perhaps a linear representation would emphasize the points more clearly.
**Author reply:** We have replotted the output in Figure 1c as a linear representation (see figure below). We prefer to keep the Circos plot in the main figure, but leave it up to the Reviewer and the Editor to select their preference.

**Alternative Fig. 1c**. Linear genome plot showing annotation features, nucleotide diversity, and structural variants across the *Ae. tauschii* panel and high-quality genome assemblies relative to the TA1675 L2 reference assembly. The tracks show from top to bottom: (i) mean structural variants density per 10 kb window per accession. Lineage 1 is indicated by a yellow line, Lineage 2 by a blue line, and Lineage 3 by an orange line, (ii) nucleotide diversity per 10 kb window across the panel of non-redundant accessions, (iii) repeat density for TA1675 indicated by repeat-masked bases per 10 Mb window, (iv) gene density for TA1675 indicated by high-confidence genes



**Comment 3.8.** I wonder if there's any way to distinguish between the "admixed Aegilops" vs "recurrent hybridization" hypotheses. Recurrent hybridization looks more likely given the variation in landraces to me.

**Author reply:** We have been discussing this question a lot, but have so far not come up with a satisfying approach to test the admixed *Aegilops* vs. recurrent hybridization hypotheses. Recurrent gene flow definitely occurred, but most of the alternative haplotype blocks shown in Fig. 4b are rare and geographically restricted. We therefore think that the original *Ae. tauschii* donor that contributed the D genome was already highly admixed. We would be more than happy to engage in a follow-up discussion to design an experiment or analysis that could solve this mystery in the future.

**Reviewer Reports on the First Revision:**

Referee #1 (Remarks to the Author):

Having thoroughly reviewed the revisions and the detailed responses provided by the authors to my initial comments, I am pleased to acknowledge my satisfaction with the adjustments made to the manuscript. The incorporation of additional information and the correction of previously noted discrepancies have improved the manuscript, improving the clarity of the presented study. I'm thankful for the authors' careful consideration of my feedback and their committed efforts to enhance their manuscript.

I am also pleased that all the data has been uploaded, and the links are working now. However, I noticed – and apologies for the oversight in my initial review – that the code for the missing finder pipeline is lacking a description and instruction on how to run it. This should be added including sample files in case this is meant to be used by the broader community. The same applies for the k-mer matrix generation pipeline.

I would also like to offer a further observation regarding the newly included Orthofinder analysis. The analysis conducted is not a gene family analysis in the strict sense. Therefore, it would be more accurate not to refer to Orthofinder clusters as gene families. Instead, these clusters could be described more accurately as "core and dispensable" gene content, or simply as clusters. It appears that Orthogroups were employed for clustering rather than "Phylogenetically Hierarchical Orthogroups (HOGs)," possibly to facilitate comparison with the study by Zhou et al. While this choice may have its rationale, employing HOGs would represent a more current and precise approach. However, in its current form, unfortunately I don’t see the added value of this analysis anyway. What does the information that there are 50,323 orthogroups, with 20,055 being core clusters, tell us? It seems to lack clear biological insight or relevance. For example, a functional enrichment analysis of clusters unique to certain lineages or an analysis of the clusters that are unique to Aegilops would make more sense.
Such insights could potentially illuminate pathways for the discovery of novel resistance genes, offering tangible benefits to wheat breeding efforts. I suggest that the authors consider enhancing this aspect of their study to provide a more meaningful contribution of this analysis to the understanding of genetic diversity and its implications for crop improvement.

Referee #1 (Remarks on code availability):

There is not enough instruction on how to run most of the code, but I commented on this again.

Referee #2 (Remarks to the Author):

I have now reviewed the manuscript and am very happy with how my comments have been addressed.

Referee #3 (Remarks to the Author):

Overall, I think the manuscript has been improved, especially with regard to the comments from the other reviewers. However, there are a few points that still need to be addressed. I detail these below.

Some specific points:
[Comment 3.2] I understand the value of multiple reference genomes, but my point here was that the audience would not understand this given the current presentation of the work. I feel like highlighting the points you make here and to reviewer 2 in the actual paper would be very helpful.

[Comment 3.5] "It has been found in barley (Mascher et al. 2021, Plant Cell 33:1888-1906) and human (McCartney et al. 2022, Nature Methods 19:687-695) that polishing HiFi assemblies with Illumina reads reduced the assembly quality".
This is a mis-reading of these papers. They say that in some cases it is possible that polishing doesn't help (with certain assemblers, in repeats, etc.). This does not mean that you shouldn't polish; it means you should do it carefully and don't polish the repeats. In my experience, a 20X CCS assembly will have thousands of sites that will be improved by polishing. Regardless of your opinion on this matter, please assuage my concerns and report the numbers SNPs and short INDELs that are homozygous (and would have been polished) from your Illumina reads when mapped to the genomes from the same genotype (especially the shallower sequenced ones). I normally don't ask for additional analyses, but this is a trivial analysis to do (you already have all the data) and I feel this is a crucial point since you are using kmers derived from in short read libraries - you'd want to make sure you have agreement between these and your assemblies. This analysis will also give you an empirical quality score for your assemblies.

[Comment 3.6] I was hoping the authors would provide some quantitative estimate of the number of duplicates and/or assembly errors. Since this analysis wasn't done, I did it with the data on dryad and found very few assembly issues. Seems like ragtag scaffolding worked well.

[Comment 3.7] If the goal is to show particular patterns, the linear representation is far more accessible. For example, the linear plots clearly show the differences in SV density between the two lineages, but it is impossible to tell this from the circos. I'll leave it to the authors.

**Author Rebuttals to First Revision:**

Referee #1 (Remarks to the Author):

Having thoroughly reviewed the revisions and the detailed responses provided by the authors to my initial comments, I am pleased to acknowledge my satisfaction with the adjustments made to the manuscript. The incorporation of additional information and the correction of previously noted discrepancies have improved the manuscript, improving the clarity of the presented study. I'm thankful for the authors' careful consideration of my feedback and their committed efforts to enhance their manuscript.

**Author reply:** We would like to sincerely thank the reviewer for the constructive feedback and comments that helped to improve our manuscript.

I am also pleased that all the data has been uploaded, and the links are working now. However, I noticed – and apologies for the oversight in my initial review – that the code for the missing finder pipeline is lacking a description and instruction on how to run it. This should be added including sample files in case this is meant to be used by the broader community. The same applies for the k-mer matrix generation pipeline.

**Author reply:** We have corrected this. The respective GitHub pages for the Missing Link Finder (https://github.com/emilecg/wheat_evolution) and the *k*-mer matrix pipelines (https://github.com/githubcbrc/KGWASMatrix) now include readme files with detailed instructions.

I would also like to offer a further observation regarding the newly included Orthofinder analysis. The analysis conducted is not a gene family analysis in the strict sense. Therefore, it would be more accurate not to refer to Orthofinder clusters as gene families. Instead, these clusters could be described more accurately as "core and dispensable" gene content, or simply as clusters. It appears that Orthogroups were employed for clustering rather than "Phylogenetically Hierarchical Orthogroups (HOGs)," possibly to facilitate comparison with the study by Zhou et al. While this choice may have its rationale, employing HOGs would represent a more current and precise approach. However, in its current form, unfortunately I don't see the added value of this analysis anyway. What does the information that there are 50,323 orthogroups, with 20,055 being core clusters, tell us? It seems to lack clear biological insight or relevance. For example, a functional enrichment analysis of clusters unique to certain lineages or an analysis of the clusters that are unique to Aegilops would make more sense. Such insights could potentially illuminate pathways for the discovery of novel resistance genes, offering tangible benefits to wheat breeding efforts. I suggest that the authors consider enhancing this aspect of their study to provide a more meaningful contribution of this analysis to the understanding of genetic diversity and its implications for crop improvement.

**Author reply:** We have rephrased our statement as follows (lines 162-166): 'We included five previously generated high-quality *Ae. tauschii* assemblies[15,19] and the bread wheat D genome[20] for a gene cluster analysis. In total, we identified 52,722 clusters, of which 18,835 and 33,953 were core and dispensable, respectively (Supplementary Table 7)'.

We have chosen to report orthogroups instead of phylogenetically hierarchical orthogroups (HOGs) to facilitate a direct comparison with the Zhou et al. (2021) paper, as it was requested in the original comments of reviewer 2. OrthoFinder outputs both orthogroups and HOGs in the same run. Our analysis identified 60,197 pylogenetically hierarchical orthogroups, 17,725 of which were core. The numbers are thus comparable to the orthogroup analysis. As indicated above, we prefer to keep the orthogroups to facilitate a direct comparison with the Zhou et al. study.

We fully agree with the reviewer that the gene cluster analysis is descriptive and offers limited biological insights. This analysis was specifically requested by reviewer 2 to have a comparison to a previously published paper (Zhou et al. 2021, Nature Plants 7:774-786). A thorough and biologically meaningful expansion of the gene cluster analysis would be a major new undertaking in our view that would go beyond the scope of the current study. It would also pull the paper into a direction that we did not intend in our initial submission. The novelty of our manuscript, as agreed on by all three reviewers, is the detailed haplotype and evolutionary analysis. All the data will be publically available and anyone interested in a detailed follow-up study on gene content will be able to do this.

Referee #1 (Remarks on code availability): There is not enough instruction on how to run most of the code, but I commented on this again.

**Author reply:** We have corrected this. The respective GitHub pages for the Missing Link Finder (https://github.com/emilecg/wheat_evolution) and the *k*-mer matrix (https://github.com/githubcbrc/KGWASMatrix) now include readme files with detailed instructions.

Referee #3 (Remarks to the Author): Overall, I think the manuscript has been improved, especially with regard to the comments from the other reviewers. However, there are a few points that still need to be addressed. I detail these below.

Some specific points:

[Comment 3.2] I understand the value of multiple reference genomes, but my point here was that the audience would not understand this given the current presentation of the work. I feel like highlighting the points you make here and to reviewer 2 in the actual paper would be very helpful.

**Author reply:** We have added the following statement to the end of the gene cloning paragraph (lines 219-222): **'**Several recent studies highlighted the importance of using genome assemblies of resistance gene-containing reference accessions for gene cloning and designation[11,34]. The analyses we present here would have been difficult, if not impossible, using previous *Ae. tauschii* assemblies because none of them are donors of *Sr33*, *SrTA1662* (*Sr66*) or *Lr39*.'

[Comment 3.5] "It has been found in barley (Mascher et al. 2021, Plant Cell 33:1888-1906) and human (McCartney et al. 2022, Nature Methods 19:687-695) that polishing HiFi assemblies with Illumina reads reduced the assembly quality". This is a mis-reading of these papers. They say that in some cases it is possible that polishing doesn't help (with certain assemblers, in repeats, etc.). This does not mean that you shouldn't polish; it means you should do it carefully and don't polish the repeats. In my experience, a 20X CCS assembly will have thousands of sites that will be improved by polishing. Regardless of your opinion on this matter, please assuage my concerns and report the numbers SNPs and short INDELs that are homozygous (and would have been polished) from your Illumina reads when mapped to the genomes from the same genotype (especially the shallower sequenced ones). I normally don't ask for additional analyses, but this is a trivial analysis to do (you already have all the data) and I feel this is a crucial point since you are using kmers derived from in short read libraries - you'd want to make sure you have agreement between these and your assemblies. This analysis will also give you an empirical quality score for your assemblies.

**Author reply:** We have performed the read mapping and SNP/InDel calling of the Illumina data against the respective HiFi reference assemblies. Read mapping and SNP/InDel calling have been performed with the same tools and filtering criteria already described in the manuscript. We have added the results to Supplementary Table 24 (previous Supplementary Table 32). It is important to note that we have sequenced accessions and not individual plants. For most accessions, the Illumina data were generated from a single plant and a bulk (~20 plants) representing the progeny of that plant was used for the HiFi sequencing. It is thus possible that some of the homozygous SNPs/InDels represent residual heterogeneity / heterozygosity. We have updated the methods section as follows: (lines 623-624) 'The number of homozygous SNPs and short InDels was determined comparing the HiFi assemblies against the respective WGS data (Supplementary Table 24).' and (lines 744-746) 'For the assessment of assembly quality (Supplementary Table 24), homozygous InDels were also retained (maximum number of raw reads supporting an InDel (IDV) = 3, maximum fraction of raw reads supporting an InDel (IMF) = 0.3, depth between 5 and 40, a quality higher than 30 and an allele count higher than one).'

[Comment 3.6] I was hoping the authors would provide some quantitative estimate of the number of duplicates and/or assembly errors. Since this analysis wasn't done, I did it with the data on dryad and found very few assembly issues. Seems like ragtag scaffolding worked well.

**Author reply:** We used RagTag "scaffold" with the default insertion of 100 bp gaps between contigs. To our knowledge, RagTag scaffold does not alter query sequences and should not result in overlapping joins. We have carefully checked each of the RagTag assemblies and found that the cumulative size of the placed and unplaced contigs corresponds to the total assembly size. In addition, the number of gaps corresponds to the total number of placed contigs minus seven (the number of *Ae. tauschii* chromosomes). Together, this indicates that the original HiFi contigs have not been altered while running RagTag scaffold. We have uploaded a new Excel table containing the contig information for each RagTag assembly to Dryad. Also, we added the dot-plots that were used to assess the quality of the RagTag assemblies to Dryad. Our rationale to use RagTag scaffold was to preserve the primary contig information of the HiFi assemblies. To clarify our approach, we have added the following statement in the methods section (lines 650-653): 'After running RagTag scaffold, the placed contigs had the exact same lengths as the primary contigs before running RagTag scaffold. Also, the number of gaps in each RagTag assembly corresponds to the number of placed contigs minus seven (number of chromosomes) (Supplementary Table 6). This indicates that RagTag scaffold did not introduce mis-assemblies or duplicated contig ends'.

[Comment 3.7] If the goal is to show particular patterns, the linear representation is far more accessible. For example, the linear plots clearly show the differences in SV density between the two lineages, but it is impossible to tell this from the circos. I'll leave it to the authors.

**Author reply:** We have uploaded both figure options (Circos and linear representation) with the revised manuscript version. We prefer the Circos plot, but leave the final decision to the editor.

**Reviewer Reports on the Second Revision:**

Referee #1 (Remarks to the Author):

I have reviewed the latest revisions and the authors' detailed responses to my previous comments. I am pleased to confirm that all issues have been satisfactorily addressed. The inclusion of the additional information and the corrections made have further improved the clarity and quality of the manuscript. I am now satisfied with the manuscript in its current form and have no further comments. Thank you for your thorough and considerate responses.

Referee #3 (Remarks to the Author):

All looks good.
The number of homozygous unpolished sequencing errors is about what I'd expect ... 3-40k per genome. These numbers should be mentioned in the main text. If you are concerned about heterogeneity in the sequencing pool, you could also report the number of heterozygous variants.

**Author Rebuttals to Second Revision:**

Referee #1 (Remarks to the Author):

I have reviewed the latest revisions and the authors' detailed responses to my previous comments. I am pleased to confirm that all issues have been satisfactorily addressed. The inclusion of the additional information and the corrections made have further improved the clarity and quality of the manuscript. I am now satisfied with the manuscript in its current form and have no further comments. Thank you for your thorough and considerate responses.

Our response: We want to sincerely thank reviewer 1 for the constructive and insightful comments that helped to improve our manuscript.

Referee #3 (Remarks to the Author):

All looks good.

The number of homozygous unpolished sequencing errors is about what I'd expect ... 3-40k per genome. These numbers should be mentioned in the main text. If you are concerned about heterogeneity in the sequencing pool, you could also report the number of heterozygous variants.

Our response: We now mention the number of homozygous SNPs and short InDels in the main text (methods section) as follows (lines 623-625): ‘The number of homozygous SNPs and short InDels was determined comparing the HiFi assemblies against the respective WGS data (Supplementary Table 24). They are in the range of 3,416-40,855 homozygous SNPs/InDels per accession.’.