Peer Review File

Manuscript Title: Structural variation in the pangenome of wild and domesticated barley

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referee #1 (Remarks to the Author):

The goal of this research is to generate and computationally evaluate pan-genome resources for barley. The authors focus their analyses on structural variation that are not visible with short-read sequencing with the goal of demonstrating the value of chromosome-scale pan-genomes. The analyses, hypotheses, and conclusions seem sound. Several additional data such as heterozygosity, false duplications, pangenome gene table, and graph pangenome alignment will help to further strengthen the results and usability of genomic resources. The manuscript will also benefit from more clarifications in text and figures, and more details in methods for reproducibility. Statistical results should be shown to make conclusions beside examples. More detailed comments are as follows.

Major concerns:

Many methods are overly simplified and lack sufficient details to reproduce the study.

It is critical to purge heterozygous regions in haploid assemblies before any meaningful structural variant studies. Some levels of heterozygosity are expected in these assemblies because not all of them are inbred lines. Extended Date Figure 2a shows false duplication rates in the assemblies, but it's not equivalent to heterozygosity. I would like to know how heterozygous regions were identified and removed in these assemblies. The method mentioned manual curations using Hi-C data, but no criteria were provided, making the procedure non-reproducible. Please show the heterozygosity level for all long-read assemblies in this study and how many of the unanchored contigs (~2% of assembly size) are heterozygous regions.

Figure 1b shows synteny between multiple accessions with aberrant cytotypes (HOR 14273 and HID055) which makes interpretation of any particular accession challenging. It might be useful if the accessions you wish to highlight were displayed next to the normal cytotypes. Extended Data Figure 3c shows Chr4H of HID055 was fused with Chr2H and placed between the long arm and short arm of Chr2H (or chromosome 4H fused to the short arm 2H and the long arm of 2H is a stand-alone chromosome. It's a little blurry, so hard to tell), a scenario different from what is described in the main text L141 - L143. Further, Figure 1c shows chromatin interactions from HiC data. The axes are not labeled but show chromatin interactions between part of Chr4H with part of Chr2H, suggesting it was not the entire Chr4H fused with Chr2H, conflicting with the data shown in Extended Data Figure 3c, possibly due to crossover in segregating offspring. Authors should compare the relative lengths of HID055 pseudomolecules to the length of the chromosomes in the HiC plot to see if their expected karyotype explains these size differences. Please indicate the percent lethality in offspring.

Another issue for Figure 3c - HiC captures chromatin interactions that are not equivalent to linkage disequilibrium. The latter quantifies the non-random association of alleles of different loci in a given

population. Authors should present the actual LD data from segregating offspring of HID055 x Barke.

Pangenome gene table should be constructed and provided. This would be a great resource for the barley community for functional studies but will also facilitate their own research on the Mla locus, for example, which is a proof-of-concept to demonstrate the power of pangenome. They identify CNV levels of genes (Figure 2d) and should also report orthologous genes for single-copy genes across the 76 barley accessions.

Read alignments using the graph pangenome were reduced compared to aligning to the linear genome MorexV3 (Extended Data Fig 4b), which is counter-expected since the pangenome is supposed to capture more genetic content, diversity, and variation than linear genomes do. Also, the overall read alignment rate is abnormally high, given this is a highly repetitive plant genome, and the data were short reads. Please break down read alignments into different categories such as unique, multiple, partial, unaligned, etc. Please also use long reads for this alignment test.

The genomes were estimated to have ~1% of false duplications (Extended Date Figure 2a), and their scans for long-duplication-prone-regions (l-DPR) produced ~36Mb regions in the MorexV3 genome (Suppl. Table 7) and equivalent to ~0.9% of the genome. Please show data to verify if the l-DPRs are the result of false duplications or biologically true.

Figure 4b shows the GWAS result of rachilla hair in the core1000 dataset, which is surprisingly similar to their previous result of the same phenotype in the same population (PMID: 33239781, Extended Data Fig. 8b). The GWAS method is too short to determine what was done differently. Nevertheless, the Srh1 story seems to be a follow-up of their previous publication.

Suppl. Table 25 shows many Cas9 mutant lines of the SMR-like gene, but the phenotype of only several lines is shown in Figure 4d. The quantification of the rachilla phenotype for mutant lines seems missing. Fig 4d has five panels, but its legend only shows three panels. They may change to “three independent mutant segregants showing the short-hair phenotype”. Surrounding the panels are some very tiny words that are impossible to read due to their small size and poor resolution.

Different from their previous pangenome (PMID: 33239781), they did not perform any de-novo TE annotation in these genomes. Still, several examples in this manuscript suggest the important role of TEs in the function and duplication of genes, such as Figure 2b, Figure 3a, and Figure 4c. Pangenome studies in maize and rice, two other important grass crops, showed that TEs are overrepresented in structural variations. Annotation of TEs novel to the PGSB library in the diversity panel could be valuable to discern the cause of functional variations.

The text can benefit from editorial polishing:
Line 76-77: Sentence, “For example, barley (Hordeum….” reads incorrectly and should be revised.

Line 87-89: Sentence, “In addition to these examples, traits..” reads incorrectly and should be revised.

Line 174: Sentence “…will be a desirable in agricultural genetics…”. Grammatical error here.

Line 185: Point out that RGT Planet is an accession. Like “head-to-tail in the accession RGT Planet”.

Line 250: I think the use of “respectively” could be omitted here.

Lines 257-259: Optional edit. I think info on specific clusters would be more valuable in the figure legend. In its current form, readers will need to toggle back and forth between Extended figures and text to interpret this cluster info.

Figure 2a is confusing. The cladograms on the top and left are not explained, and the coloring on the x-axis is not explained. The legend on the right is too big and merged into the background. This figure delivers unclear information and is difficult to interpret.

Figure 3: legend and figure include a different # of panels.

Extended Data Figure 1: Are the 412 non-highlighted accessions (panels a-d) from the project that sequenced ~1k genomes using short-reads? Point this out in the legend.

Extended Data Figure 2: legend and figure include a different # of panels. Panel f is nice.

Extended Data Figure 3: Please clarify the legend of Panel c, are alignment groups from top to bottom? The color scheme in panel c is confusing between the three subpanels.

The number of SNPs says 164.5M in L166 but shows 155.6M in Extended Data Figure 5b.

Extended Data Figure 6: Pane l b – Please clarify the colors used in the panel. It would be very helpful if the barley genomes were organized in different groups.

Extended Data Figure 7: too blurry to evaluate. Many figures need improved resolution. This is probably an artifact of the initial submission.

Extended data figure 8 and 9: I think these figures should switch spots. The bulk of fig 9 text comes before figure 8 in the manuscript body.

Extended Data Figure 8: panel b is difficult to follow.

Extended Data Figure 9: missing in-figure panel lettering (a-b).

Extended Data Figure 10: panel b is squeezed. panel f- the first boxplot in the figure is cut.

Please also show BUSCO results on the final gene annotation set of each genome. Indicate if the BUSCO in Ext. Data Fig2a is based on genome/transcriptome/ or gene annotations.

For the single-copy pangenome construction, the method BBDuk cited in this section was designed for read trimming and filtering, and the authors did not explain how BBDuk was utilized to “identify and filter 31-mers occurring more than once in genomic regions”.

Figure 3b: use percentage rather than accession counts since you have more domesticated than wild.

Line 915: “GWAS for was done with GEMMA”. Error here. Also needs to be expanded.

Referee #2 (Remarks to the Author):

This manuscript describes a fantastic resource for the barley genetics community, with 76 long read genomes, but the analyses provided seem cursory and disclose little new biology. Apart from the fact that this is a different species, the advance for the broader community does not go beyond similar types of papers published elsewhere for tomato, rice and soybean two or three years ago, and it does not go nearly as far as the (graph) pangenome papers for tomato and potato published in Nature last year.

The strengths of the work are a good, representative choice of accessions for the long-read genomes and a good strategy for genome annotation, using short read RNA-seq and PacBio IsoSeq to produce high-quality annotations of a small number of accessions, and then projecting these annotations onto the remaining accessions. The major weakness is that a consideration of evolution as a process is largely absent from the manuscript, and that the demonstration of the usefulness of the work is restricted to few vignettes that focus on known loci or loci that encode homologs of genes known from other species to participate in relevant processes. Notably, while “adaptation” is prominently mentioned in the title, abstract and introduction, it is completely absent from the results, as no analyses that pertain to adaptation or selection are presented.

The enormous variation in copy number of tandem repeated loci is interesting but it was disappointing that there was little systematic investigation of the evolutionary processes responsible for expansions and contractions, along the lines of the scheme presented in Extended Data Figures 6a/7b. E.g., How often are individual gene copies duplicated, how often pairs, how often trios etc.? What evidence is there for subsequent contraction? And how often are there independent expansions? Along the lines of this last question, I was intrigued by the Rabanus-Wallace et al. preprint cited in this manuscript, which suggests that expansions are driven by selfish elements – the number of independent expansions of a cluster across barley accessions would clearly speak to that question. It would also begin to answer questions about selection on entire clusters versus individual genes in such clusters.

I also have a series of technical concerns:

Given the comparatively large size of barley genomes, it is understandable that PacBio HiFi coverage was only about 20x. This may, however, lead to collapsing of tandem repeats, which is particularly relevant for long tandem repeats of identical copies. Here, I would have appreciated reading something about the limits of the assemblies: What are the longest tandem repeats of identical copies that the authors could have expected to discover? This concern is compounded by the fact the authors used an older version of Hifiasm. It might be useful to use the flagger tool (https://github.com/mobinasri/flagger/) to survey the assemblies for potentially collapsed regions. The Hi-C data might also be helpful.

Only 17k out of 96k orthology groups were present in all 76 accessions, which suggests a very strict (too strict) definition of orthology groups. Technical details are unfortunately missing. More relevant would be information on syntenic orthologs.

I understand the difficulty of using PGGB or minigraph cactus to build a pangenome for barley-sized genomes but one could use PGGB or PGR-TK for specific loci. PGR-TK shines in the comparison of tandem repeat regions, and it is a great to answer the sort of questions I raised above.
Regarding the alpha-amylase cluster: Were the results from 21-mer genotyping consistent with the assembly-based results? When did this cluster expand relative to domestication and subsequent improvement?
Finally, the font in many figures is too small. Also, both the figures as well as the figure legends are often far too terse (as are the methods), which makes it often difficult to understand how they support the interpretations in the main text.

Minor comments:

Line 130: A reference for gene flow between wild and cultivated barley is needed.

Line 140: Can short reads be used to test whether the two reciprocal translocations discovered among the 76 accessions are present in other accessions?

Line 147: The point of growth of graph structures will not be obvious to those not familiar with pangenome graphs. Needs more explanation.

Line 185: Unclear to me how anyone could deduce from Extended Fig 6b the absence of complete Mla copies. Also, what is meant with this? Truncations, premature stops, complete absence?

Line 197: “Until the advent of long-read sequencing, it was virtually impossible to resolve the structure of the Mla locus in multiple genomes at once, but now it is a corollary of pangenomics.” What is really meant here? Also, it is a bit insulting to colleagues who have been around for a while – there are examples of reconstruction of complex loci in multiple accessions using YACs, BACs and fosmids combined with Sanger or short read sequencing. There was serious genomics before PacBio HiFi sequencing, but it required more effort than Illumina whole-genome shot gun sequencing.

Line 322: Is the CATCGGATCCTT motif present in Morex at all? If yes, are there ATAC-seq data that suggest it is an active cis-regulatory element?

Line 329: The statement “Gene edits of the enhancer region, guided by the pangenome sequences, will further elucidate the transcriptional regulation of HvSRH1” is superfluous.

Figure 1c: A comparison of Hi-C linkage with genetic linkage would be useful.

Figure 1d: What is meant with “single-copy” here? Is this gene space only?

Figure 3: This would benefit from a phylogeny of non-identical ORFs. Also, does identical mean no synonymous SNPs at all? This should be spelled out.

My expertise is in structural, comparative, evolutionary and functional genomics.

Referee #3 (Remarks to the Author):

The article from Jayakodi et al. describes the construction and use of an extended barley pangenome through sequencing and comparative genomic analyses of 76 reference-quality assemblies complemented by short read-based variant calls for 1315 genotypes. The paper does not extensively describe the pangenome on its own –not enough to my opinion – but rather focus on classical comparative genomic analysis of 4 regions taken as examples in order to demonstrate the utility of having access to a wide diversity of chromosome-level assembled genomes for research. It even includes a case study with the identification of a potential structural polymorphism that may be causal of a phenotype (related to grain development).

The paper is a typical high-impact journal-oriented paper of an international consortium (30 author affiliations) describing a new milestone on the road toward characterizing the full genomic diversity of the species. It mixes the description of a massive sequence data production effort, large-scale sequence analyses, unrelated examples of comparative genomic analyses focused on four different regions (Mla, Tb1, Amy1-1, Srh1), results from wet lab experiments (alpha amylase activity), and even the first steps of a gene cloning project (with mutants for the candidate gene, isogenic lines, expression data…). People in the field of structural genomics may regret not having a more focused paper on these aspects (I do) but others may appreciate the focus on the applications. That being said, the paper is well written, the resource is of significant importance for the community, and the methods are sound, although I have remarks about methodologies to build the pangenome (see below). Since my specialty is genome sequence analysis, I will focus my review on these aspects rather than the functional part.
Although well written, the paper is not an easy read. It is often very complex, with many references to extended data, suppl materials, which makes it sometimes a pain to read. This could be improved.

My main concerns:
- I found the description of the pangenome, and the level of genomics variability, were not commented at the level I would have expected for such a large effort of data production. But maybe this paper is only the main general paper and there is another companion paper dedicated to pangenomic aspects? Here there are only a short initial paragraph (likely a quality assessment) and a second ("atlas of SV") which does not provide a deep characterization of the variability. There is nothing about TE space variability (maybe this is for another paper), and regarding genes, the only value provided in main text is 17% of core genes. A bit frustrating. What about core/shell/cloud genes? what about wild versus domesticated? What are the conclusions? Comparisons with other species?
- I also found there was not enough link made with what was known before. I say that especially because there was already a Nature article in 2020 on the same topic and by the same leader authors (Jayakody et al. 2020) and I would have appreciated to read more about it in the introduction and about what is new here.
- The Figures in the Extended data are often too small and too blurry and are not readable.

SPECIFIC REMARKS AND QUESTIONS

++ ABSTRACT

- Well written. I regret there is no summary of the main values expected for a pangenome analysis (number or % of core/dispensable genes etc.)

++ INTRODUCTION

- I suggest to add some information/results regarding the previous initiatives, especially Jayakodi et al. 2020 with 20 genomes (and Russel 2016 maybe). Size of the pangenome, etc.

++ RESULTS

+ An expanded annotated pangenome of barley

- This part is poor in terms of results, and there is no conclusion. The only result I see is the average 2% of the ~5000 single-copy genes of Poales that are absent. And there is no associated Extended Data to better describe this. It raises more questions than it brings information. Where are the results about the description of core/shell/cloud genes? I understand these results are more provided in the next paragraph (atlas of SVs)… this is not clear to me what is the purpose of this first paragraph. If it is related to genome quality assessment, I suggest to give a title accordingly and to conclude the results.
- The method employed here to build a pangenome is questionable. At least it makes appear some weaknesses, obviously due to the difficulty of analyzing these genomes correctly, but that limit our ability to reach the goal of building a high-quality pangenome. I understand that genes were predicted in 20 genomes while only projected in the 56 others. This may have strong consequences in deciphering the pangenome. I know and understand how difficult it is to get an accurate, well predicted, gene set for these genomes. But I would have appreciated to read conclusions/discussions about that. The most important question is: what is the impact of the strategy (gene projection instead of 76 independent gene predictions) on the results? Obviously if specific genes are not predicted de novo, they cannot be present in the pangenome! Please, could the authors try to discuss this?
- Why focusing on Poales single copy genes? And if these genes are the super-conserved core-genes of all Poales, how do the authors conclude on these 2%? Was this a way to estimate the error rate due to incompleteness of genome assemblies, annotation problems, etc. (I think yes)? How many of them are missing in all accessions? If there is 1-2% of genes missing in the assemblies, it could however makes the % of genes present in 76 accessions very low, this is why the authors maybe could have computed a "soft-core" gene set?

+ An atlas of SV

- This paragraph is a mix between SV detection at the whole genome sequence level (i.e., the title) and a gene clustering-based pangenome analysis… which are two different approaches to achieve similar things. I found this part is interesting but is hard to follow in the way it is written here. I suggest to try to make the text easier to read. The only value I found is the 17% core-genes. It appears very low, probably because calculated relatively to all pan-genes and not to genes present per individual. So, what is the percentage of core-genes per accession (16k/~50k)? This looks also low. I suggest to comment more on these values and to compare with knowledge from other species, which would give much more interest in reading this part.
- "At the level of individual gene models, a third were considered conserved because they belong to an orthologous group with representatives from each accession". This is an example of sentence one

could read several times but one could not make sense of it.
- What about genes subject to CNVs? (cf Extended Data where I realized there were several rounds of gene projections, probably to identify CNVs?)
- "The functional annotations […] pointed to an involvement in biotic and abiotic stress responses (Supplementary Table 4)". I read quickly the Supp table and saw many terms not related to stress response.
- "Pangenome graph". Hard to see the added value of this part, except to comment on the fact that building a high-resolution pangenome graph is still too complicated in barley and related complex genomes. However, if I understood correctly, the graph was used for mapping resequencing read of a large diversity panel in order to estimate the representativeness of haplotypes captured in this pangenome. This part is interesting.
- SVs based on genome sequence alignments + SVs based on graph-based mapping… any differences observed? (Maybe hard to answer that).
- "will be a desirable" -> should be "as desirable"

+ An inventory of complex loci
- Mla locus: very descriptive, although I understand that pangenomics is often very descriptive. The results here could be summarized as: there were 29 known Mla genes, 7 of them are present in the pangenome, but 149 homologs were clustered here? and a landrace contains 11 genes, with 2 being present in 5 copies. The conclusion is interesting, saying that resolving complex loci is now feasible and a corollary of pangenomics. But I found we miss a conclusion on Mla gene SVs.
- The thionin gene CNV example made me wonder about the strategy employed to annotate genes in the 76 genomes. I though, as explained, that genes were predicted denovo in 20 genomes and projected in the others. How does this impacted the ability to identify CNVs here? (redundant with my previous remarks on CNVs).
- Paragraph on dating duplicated loci is really interesting. But the associated Extended data 7 is not readable (blurry picture).
- "enriched in distal chromosomal regions (Fig. 2d)" -> should be Fig. 2c

+ Amplification of α-amylases in malting barley
- First paragraph is very descriptive, mentioning many details but without conclusion. Worth example is the fact that "12 had insertions of TEs". In different copies? Is it a shared polymorphism or do the authors mean that they were several independent insertions of different TEs? in the same copy?
It is hard to get what is new here compared to what was known before in term of diversity.

+ A regulatory variation controls trichome development
- The cloning about HvSRH1 is quite interesting indeed. And it is a remarkable example of the utility of the pangenome, well, at least having access to multiple well assembled genomes in order to detect structural variations that may cause phenotypic variations.
- It is mentioned a 4kb segment absent in all (14) short-haired genotypes while "exceptionally conserved" in the others i.e., with 95% identity. To me, it sounds contradictory because 95% nucleotide identity is low between two genotypes of the same species!?
- "CATCGGATCCTT, matching the sequence [ATC]T[ATC]GGATNC[CT][ATC]".
The first "C" is not part of the motif, so I guess there is something to correct here. In addition, when searching for the presence of this motif across the interval, which one was searched for? the strict

one (first) or the permissive one (second)? I scanned the Morex genome with the permissive one in order to realize that such motif is present every 8kb on average. So, the probability to find it by chance in a 4kb segment is actually high. When I search for this motif HTHGGATNCYH in the 120kb interval of Fig4, I find it 18 times which is different from the 3 motifs of the figure. I can only suggest to double check this analysis.
- The expression data illustrated at panel f of Extended Figure 10 is quite interesting. It brings a serious argument and I suggest to try to include this in the main paper.

++ Discussion
- "true to the hypothesis-generating remit of genomics", guess you mean "merit" right?
- First part is more a discussion about the interest of long-reads in complex genomes than the interest of pangenome, although I understand the two go together for complex genomes.

++ EXTENDED DATA
- Probably just a problem with the PDF conversion, but many panels of Figures are not readable. I suggest to improve the quality of the figures here. Effort has been made to detail each legend, a good point for Supplements.
- Ext Fig 2: Panel e is missing in the legend
- Ext Fig 3: Typo "pagenome". Text is too small to be readable.
- Ext Fig. 6: Alignment not readable. And I do not see how to get a message from Panel b.
- Ext Fig. 7: Cannot read that
- Ext Fig 10: I wonder if panel f should not be better part of the main Figures
- SNP and SV calling: first sentence is not finished, there is a verb missing. Same for the last sentence. I suggest to proofread the whole paragraph.
- Gene projections: this part is really hard to follow while it is critical to assess the work done and to be able to estimate the limits of the method.
"For the two top quality categories, we performed two rounds of projections, firstly inserting each source maximally only once followed by rounds allowing one source inserted multiple times into the projected annotation". I may understand that genes under CNVs were annotated here because there were several rounds of projections? If yes, I think it should be clearly mentioned, probably not only in the Extended text. And if yes, why not providing results about that?
- line 1153: "to detection the different Mla…" to be corrected.
- Code availability: "Scripts for calculation of core/shell and could genes". If such things were done, I would really appreciate to see these results.
- I quickly visited the github page https://github.com/PGSB-HMGU but cannot find any repository related to core/shell/cloud gene calculation.
- Regarding gene projection, scoring, iterations, I suggest to also make the codes available for reproducibility.

Author Rebuttals to Initial Comments:

Referee #1 (Remarks to the Author):

The goal of this research is to generate and computationally evaluate pan-genome resources for barley. The authors focus their analyses on structural variation that are not visible with short-read sequencing with the goal of demonstrating the value of chromosome-scale pan-genomes. The analyses, hypotheses, and conclusions seem sound. Several additional data such as heterozygosity, false duplications, pangenome gene table, and graph pangenome alignment will help to further strengthen the results and usability of genomic resources. The manuscript will also benefit from more clarifications in text and figures, and more details in methods for reproducibility. Statistical results should be shown to make conclusions beside examples. More detailed comments are as follows.

Major concerns:

Many methods are overly simplified and lack sufficient details to reproduce the study.
Answer: We have addressed your and the other referees' specific requests for a more detailed description of methods. Code used in our analyses is hosted in Github or Bitbucket repositories, links to which are given in the Code Availability section of the revised manuscript. We also would like to point the reviewer to the paper of Marone et al. 2022 Genome Biology (https://doi.org/10.1186/s13007-022-00964-1), which describes the genome assembly pipeline used in the present study. A detailed workflow is provided at https://tritexassembly.bitbucket.io.

It is critical to purge heterozygous regions in haploid assemblies before any meaningful structural variant studies. Some levels of heterozygosity are expected in these assemblies because not all of them are inbred lines. Extended Date Figure 2a shows false duplication rates in the assemblies, but it's not equivalent to heterozygosity. I would like to know how heterozygous regions were identified and removed in these assemblies. The method mentioned manual curations using Hi-C data, but no criteria were provided, making the procedure non-reproducible. Please show the heterozygosity level for all long-read assemblies in this study and how many of the unanchored contigs (~2% of assembly size) are heterozygous regions.
Answer: Our procedures for manual curation of contig placements in pseudomolecules is based on visual inspection of Hi-C contact matrices. A detailed description is given in a technical paper on the TRITEX assembly pipeline (Marone et al. 2022 Genome Biology (https://doi.org/10.1186/s13007-022-00964-1).

We estimated the heterozygosity rate using a k-mer based approach (see ll. 958-67 of the Methods) and report the results in column N of the revised Supplementary Table 1. Heterozygosity was low, at an average of 0.06 %, meaning that the assembled genotypes can be considered inbred lines.

Low heterozygosity was a criterion of quality control in the assembly process. Indeed, we had to exclude two accessions (Kristina and WBDC 291) that were not homozygous. **These are not part of**

the 76 genotypes we report on in the final manuscript. Our k-mer based heterozygosity estimates for these genotypes were 2.61 % and 4.75 %, i.e. more than 40-fold higher than the average. This result was supported by genome-wide k-mer profiles:

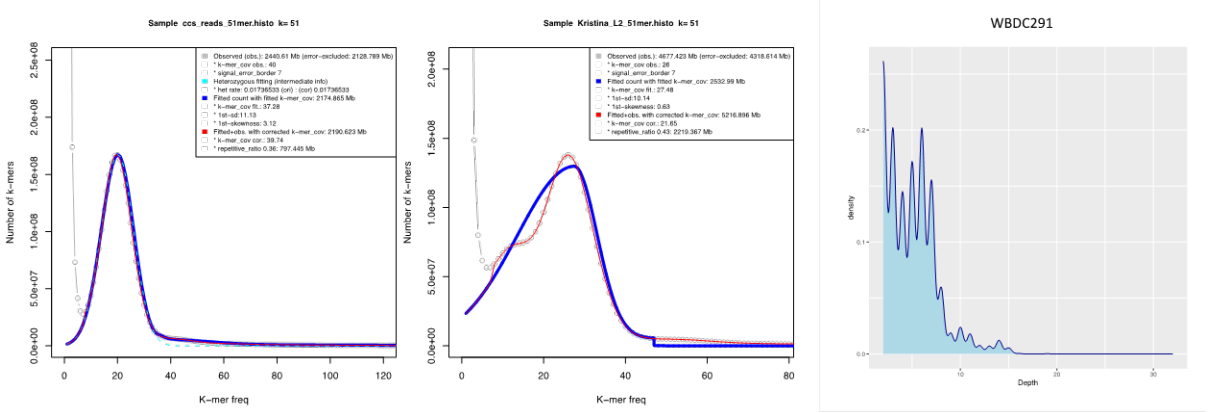

Figure: Homozygosity check with k-mer profiles: 51-mer frequency histograms were computed from HiFi data with findGSE. Left: Unimodal k-mer profile computed from HiFi data on an inbred genotype included in the pangenome. Middle / right: Multi-modal k-mer profiles of impure seed lot.

To assess how many unanchored contigs might possibly be allelic to sequences in the chromosomal pseudomolecules (and hence artificially duplicated), we aligned the former to the latter and kept only primary alignments longer than 10 kb (i.e. those spanning at least about half the length of a HiFi read). We found that an average of 53 kb of unanchored sequence matched these criteria. We consider this an acceptable amount of potentially false duplications.

Figure 1b shows synteny between multiple accessions with aberrant cytotypes (HOR 14273 and HID055) which makes interpretation of any particular accession challenging. It might be useful if the accessions you wish to highlight were displayed next to the normal cytotypes. Extended Data Figure 3c shows Chr4H of HID055 was fused with Chr2H and placed between the long arm and short arm of Chr2H (or chromosome 4H fused to the short arm 2H and the long arm of 2H is a stand-alone chromosome. It's a little blurry, so hard to tell), a scenario different from what is described in the main text L141 - L143. Further, Figure 1c shows chromatin interactions from HiC data. The axes are not labeled but show chromatin interactions between part of Chr4H with part of Chr2H, suggesting it was not the entire Chr4H fused with Chr2H, conflicting with the data shown in Extended Data Figure 3c, possibly due to crossover in segregating offspring. Authors should compare the relative lengths of HID055 pseudomolecules to the length of the chromosomes in the HiC plot to see if their expected karyotype explains these size differences. Please indicate the percent lethality in offspring.

Another issue for Figure 3c - HiC captures chromatin interactions that are not equivalent to linkage disequilibrium. The latter quantifies the non-random association of alleles of different loci in a given population. Authors should present the actual LD data from segregating offspring of HID055 x Barke.

Answer: We have changed the order of genotypes in Fig. 1b. Now the genotypes HOR 14273 and HID055, whose chromosomes 2H are involved in translocation, are not shown next to each other but are compared to non-translocated genotypes Morex and RGT Planet. We have also added Supplementary Figure 4, which shows the Hi-C contact matrices for the translocated genotypes and alignments of their pseudomolecules to non-translocated Morex. A clear inter-chromosomal Hi-C signal is seen if the Morex karyotype is used as a reference for aligning Hi-C reads but absent if the respective translocated karyotype is used.

We have removed the statement about incomplete seed set we have only anecdotal evidence (observations during population development) because a detailed quantitative assessment of lethality would require new crosses and hence more time.

Pangenome gene table should be constructed and provided. This would be a great resource for the barley community for functional studies but will also facilitate their own research on the Mla locus, for example, which is a proof-of-concept to demonstrate the power of pangenome. They identify CNV levels of genes (Figure 2d) and should also report orthologous genes for single-copy genes across the 76 barley accessions.
Answer: In the revised manuscript, we describe an orthologous genes framework/pangenome table consisting of the gene annotations of all 76 barley genotypes. This framework includes the identification of orthologous single-copy genes. We describe analysis of this orthologous framework in form of core/shell/cloud genes as well as CNV and PAV (also with respect to domestication and improvement status) in the revised manuscript (ll. 130-51). We have set a up web portal that provides access to the orthologous framework (https://panbarlex.ipk-gatersleben.de). Entry is possible via gene identifier and sequence alignment. We have added this link to the Data availability section of the revised manuscript.

Read alignments using the graph pangenome were reduced compared to aligning to the linear genome MorexV3 (Extended Data Fig 4b), which is counter-expected since the pangenome is supposed to capture more genetic content, diversity, and variation than linear genomes do. Also, the overall read alignment rate is abnormally high, given this is a highly repetitive plant genome, and the data were short reads. Please break down read alignments into different categories such as unique, multiple, partial, unaligned, etc. Please also use long reads for this alignment test.
Answer: The mapping tool used for mapping to the linear reference, BWA MEM, only aligns reads to a single location each, and partial mappings are not supported, so unfortunately the statistics requested here cannot be provided. We have double-checked the percentage of aligned/unaligned reads and the statistics we reported are correct. BWA attempts to map all reads present in the input if at all possible, and achieves this by allowing a large number of mismatches in the alignments (default = 14%). In the paper we have therefore focused on highlighting the difference in mismatch

rates and properly paired read numbers between the linear reference and the graph, both of which are generally more favorable in the graph. We have amended the text to make this clearer.

As per the request, we also conducted an additional mapping of PacBio reads. We reasoned that in order to have an unbiased comparison with the Illumina mappings, accessions used for this should not have been used in the construction of our graph, and the data should be CCS reads to avoid the inflated error rates of PacBio CLR reads and to keep things comparable with the low error rates in the Illumina mappings. There are no samples in the public read archives that satisfy these criteria, and our only recourse was to use reads from a single barley landrace accession we have recently sequenced and assembled and which is currently awaiting publication.

The mapping of these PacBio reads to the graph has essentially failed, with less than 50% of reads mapping to the graph (by comparison, 98.8% of reads mapped to the linear Morex V3 reference). We can only surmise that this is due to the apparently untested combination of vg giraffe, PacBio reads and Minigraph-based graphs which do not support small variants. This must have led to an abnormally large proportion of reads not being mappable due to excessive numbers of mismatches, as small variants are not encoded as bubbles in the graphs.

Additionally, we were unable to compute graph mapping error rates because the vg stats tool does not compute the total number of bases aligned, which is required for this statistic. For the Illumina mappings we were able to work around this by multiplying the number of mapped reads by the read length (which is the same for all reads with Illumina), but the variable read length of the PacBio reads has precluded this. Inclusion of the (incomplete) PacBio mapping results would thus be both inappropriate and misleading, and we decided against it on those grounds.

The genomes were estimated to have ~1% of false duplications (Extended Date Figure 2a), and their scans for long-duplication-prone-regions (l-DPR) produced ~36Mb regions in the MorexV3 genome (Suppl. Table 7) and equivalent to ~0.9% of the genome. Please show data to verify if the l-DPRs are the result of false duplications or biologically true.

Answer: Please note that the value 0.012 in Extended Data Fig. 2a is a percentage, not a decimal fraction. This means the false duplication rate reported by Merqury is not 1.2 %, but 0.012 % and the size of the regions potentially affected by false duplications is not 36 Mb, but 360 kb, which we deem an acceptably low number.

To further verify the integrity of our sequence assemblies in the structurally complex loci (l-DPR regions), we plotted the average HiFi read depth in those regions and did not observe elevated coverage levels (Supplementary Figure 8).

Figure 4b shows the GWAS result of rachilla hair in the core1000 dataset, which is surprisingly similar to their previous result of the same phenotype in the same population (PMID: 33239781, Extended Data Fig. 8b). The GWAS method is too short to determine what was done differently. Nevertheless, the Srh1 story seems to be a follow-up of their previous publication.

Answer: Classical barley genetics and several independent studies with different GWAS panels have arrived at the same conclusion, namely that the length of the rachilla hairs is under the control of a single major locus, *srh1*. In Jayakodi et al. 2020, we did GWAS for rachilla hair length in a panel of 200 accessions, which are a subset of the core1000 panel that we used in the present study. Apart from the larger panel size, the analytical methods employed were essentially the same. That the resultant Manhattan plot is very similar to the earlier one (although not based on the same data) is not surprising in light of our understanding of the genetic architecture of the trait in question.

Suppl. Table 25 shows many Cas9 mutant lines of the SMR-like gene, but the phenotype of only several lines is shown in Figure 4d. The quantification of the rachilla phenotype for mutant lines seems missing. Fig 4d has five panels, but its legend only shows three panels. They may change to "three independent mutant segregants showing the short-hair phenotype". Surrounding the panels are some very tiny words that are impossible to read due to their small size and poor resolution.
Answer: Supplementary Table 25 (now #26) summarizes the molecular characterization of T1/M2 progeny of individual T0/M1 plants, which were selected on the basis of their having either wildtype or mutant phenotype in M1. We have improved the presentation of this table and added more explanation.

We applied the following procedure: 31 T0/M1 regenerants were obtained for the employed CRISPR knockout construct targeting TM1a and TM1b. The plants were screened for presence of the hygromycin resistance selection marker and amplicons of the TM targets were Sanger sequenced indicating either WT sequence or heterozygous / chimeric, hence mutated sequence at the target locus. Spikes/seeds could be harvested for 23 of these M1 plants and these were inspected at several spikelets for the rachilla hair phenotype. We selected one M1 that was phenotypically and genotypically WT and three M1 plants that showed pure or mixed rachilla hair phenotype in the inspected seeds of the same M1 spike. The respective M2 families were grown for further genotypic and phenotypic validation. Supplementary Table 25 lists the genotyping results of these M2 progenies. Individual plants that turned out to be homozygous mutant were then selected for microscopic documentation of the mutant phenotype and these are the ones shown in Figure 4. The phenotypic and genotypic screening of the M1 population is shown now in addition to explain better the strategy of screening and presenting the results. Overall, we found perfect correlation between mutant phenotype and genotype supported by a minimum of three independent mutagenic events.

Different from their previous pangenome (PMID: 33239781), they did not perform any de-novo TE annotation in these genomes. Still, several examples in this manuscript suggest the important role of TEs in the function and duplication of genes, such as Figure 2b, Figure 3a, and Figure 4c. Pangenome studies in maize and rice, two other important grass crops, showed that TEs are overrepresented in structural variations. Annotation of TEs novel to the PGSB library in the diversity panel could be valuable to discern the cause of functional variations.

Answer: We agree that TEs are an important source of genetic variability and that pangenomes are an excellent tool to study TE-related diversity. Following the referee's suggestion, we have used the PGSB library to obtain full-length LTR retrotransposon (fl-LTR TEs) annotations for all 76 genomes. We report the number of fl-LTRs in Supplementary Table 2 and show the distribution of insertion ages in Supplementary Table 1. We did not observe striking differences between the genomes. We agree that TE annotations can provide valuable clues as to the potential causes of functional variation. We do report links between structural variation (*HvTB1*, *srh1*, *amy1_1*) and phenotypes, but TEs happen not to play a role in these cases. We are aware that these examples are not representative and that our fl-LTR annotation is only an initial foray into the pangenome's TE diversity. Deeper analysis including the relationship between SVs, TEs and epigenomic features will be the subject of a separate manuscript.

The text can benefit from editorial polishing:

Line 76-77: Sentence, "For example, barley (Hordeum…." reads incorrectly and should be revised.

Answer: We have corrected this.

Line 87-89: Sentence, "In addition to these examples, traits.." reads incorrectly and should be revised.

Answer: We have removed the word "both" from this sentence.

Line 174: Sentence "…will be a desirable in agricultural genetics…". Grammatical error here.

Answer: This has been corrected.

Line 185: Point out that RGT Planet is an accession. Like "head-to-tail in the accession RGT Planet".

Answer: We have followed this suggestion.

Line 250: I think the use of "respectively" could be omitted here.

Answer: We have removed "respectively" in this sentence.

Lines 257-259: Optional edit. I think info on specific clusters would be more valuable in the figure legend. In its current form, readers will need to toggle back and forth between Extended figures and text to interpret this cluster info.

Answer: We added two new panels (c and d) to Extended Figure 9 to show information on the clusters. We have shortened the main text accordingly.

Figure 2a is confusing. The cladograms on the top and left are not explained, and the coloring on the x-axis is not explained. The legend on the right is too big and merged into the background. This figure delivers unclear information and is difficult to interpret.
Answer: We have removed the cladograms and revised the legend.

Figure 3: legend and figure include a different # of panels.

Answer: This has been corrected in the revised version.

Extended Data Figure 1: Are the 412 non-highlighted accessions (panels a-d) from the project that sequenced ~1k genomes using short-reads? Point this out in the legend.

Answer: In the revised legend to Extended Data Fig. 1., we have clarified the provenance of the data shown.

Extended Data Figure 2: legend and figure include a different # of panels. Panel f is nice.

Answer: Thanks for commending panel f! We have corrected the legend.

Extended Data Figure 3: Please clarify the legend of Panel c, are alignment groups from top to bottom? The color scheme in panel c is confusing between the three subpanels.

Answer: We have split this figure into three: the new Extended Data Fig. 3 and two Supplementary Figures (# 1 and 2). We have also revised the legend to assign labels to single alignment groups.

The number of SNPs says 164.5M in L166 but shows 155.6M in Extended Data Figure 5b.

Answer: We have corrected these numbers.

Extended Data Figure 6: Panel b – Please clarify the colors used in the panel. It would be very helpful if the barley genomes were organized in different groups.

Answer: We have stated in the legend our rationale for sorting the accessions in the way we did. Note also that a more detailed version of this display item is provided in Supplementary Figure 3.

Extended Data Figure 7: too blurry to evaluate. Many figures need improved resolution. This is probably an artifact of the initial submission.

Answer: We apologize for this oversight. We have now submitted the figures as separate PDF files. Figure legends have been moved to the main manuscript file.

Extended data figure 8 and 9: I think these figures should switch spots. The bulk of fig 9 text comes before figure 8 in the manuscript body.

Answer: We tried different arrangements of panels, but in the end decided to keep the original one owing to space constraints (maximum of 10 Extended Data item). Please note that we also added two new panels to Extended Data Fig. 9.

Extended Data Figure 8: panel b is difficult to follow.
Answer: We have simplified the panel and defined what is meant by “haplotype” in the legend.

Extended Data Figure 9: missing in-figure panel lettering (a-b).

Answer: This has been corrected.

Extended Data Figure 10: panel b is squeezed. panel f- the first boxplot in the figure is cut.
Answer: We have improved the visual quality of both panels in Extended Data Figure 10.

Please also show BUSCO results on the final gene annotation set of each genome. Indicate if the BUSCO in Ext. Data Fig2a is based on genome/transcriptome/ or gene annotations.
Answer: These data are shown in Supplementary Table 4 columns J-S.

For the single-copy pangenome construction, the method BBDuk cited in this section was designed for read trimming and filtering, and the authors did not explain how BBDuk was utilized to “identify and filter 31-mers occurring more than once in genomic regions”.
Answer: BBDuk has also functions for k-mer counting and masking repetitive regions. The methods for the construction of the single-copy pangenome are described in Jayakodi et al. 2020. The code is provided at https://bitbucket.org/ipk_dg_public/barley_pangenome/src/master/. We have added this link to the code availability section. We have also expanded the relevant section of the Methods in the present manuscript.

Figure 3b: use percentage rather than accession counts since you have more domesticated than wild.

Answer: We have implemented this suggestion.

Line 915: “GWAS for was done with GEMMA”. Error here. Also needs to be expanded.
Answer: We have corrected this sentence and provided more details about parameter settings and versions.

Referee #2 (Remarks to the Author):

This manuscript describes a fantastic resource for the barley genetics community, with 76 long read genomes, but the analyses provided seem cursory and disclose little new biology. Apart from the fact that this is a different species, the advance for the broader community does not go beyond similar types of papers published elsewhere for tomato, rice and soybean two or three years ago, and it does not go nearly as far as the (graph) pangenome papers for tomato and potato published in Nature last year.

The strengths of the work are a good, representative choice of accessions for the long-read genomes and a good strategy for genome annotation, using short read RNA-seq and PacBio IsoSeq to produce high-quality annotations of a small number of accessions, and then projecting these annotations onto the remaining accessions. The major weakness is that a consideration of evolution as a process is largely absent from the manuscript, and that the demonstration of the usefulness of the work is restricted to few vignettes that focus on known loci or loci that encode homologs of genes known from other species to participate in relevant processes. Notably, while “adaptation” is prominently mentioned in the title, abstract and introduction, it is completely absent from the results, as no analyses that pertain to adaptation or selection are presented.

Answer: We thank the reviewer for commending the value of our resource.

The reviewer is correct in pointing out that our paper does not focus so much on evolutionary questions as on translational applications in crop improvement. With the latter aim in mind, we believe our comparative analysis of four complex loci does disclose some “new biology”. These four examples emerged from a discovery-driven approach that took prior knowledge of agronomically relevant loci into account. Such in-depth inquiries into loci related to crop evolution and crop improvement are exactly how we envisage other applied researchers – geneticists, breeders, genebank managers – to work with our data. In this sense, our analyses are an accompanying “how-to” guide for this community resource.

The phrase ”adaptive diversification” in the title refers to the creation of new diversity or the imposition of new selective pressures during crop evolution. Six-rowed barleys have evolved only after domestication and have been widely adopted by farmers. Likewise, the selection pressures exerted by the demands of the malting process on seed and germination traits came into play only after domestication. In the present manuscript, we analyzed structural variation at the HvTB1 and amy1_1, loci that have long been known to be involved in inflorescence architecture and malt quality, respectively. We report copy number amplification at HvTB1 in six-rowed barley and link haplotypes prevalent in malting barleys (and differing in amylase copy number) to malting quality. We believe these results do pertain to adaptation or selection and justify the conclusion “that valuable diversity can arise after domestication. Rapid evolution at structurally complex loci may endow domesticated plants with a means of adaptive diversification that aptly fulfills the needs of farmers and breeders.”

The enormous variation in copy number of tandem repeated loci is interesting but it was disappointing that there was little systematic investigation of the evolutionary processes responsible for expansions and contractions, along the lines of the scheme presented in Extended Data Figures 6a/7b. E.g., How often are individual gene copies duplicated, how often pairs, how often trios etc.? What evidence is there for subsequent contraction? And how often are there independent expansions? Along the lines of this last question, I was intrigued by the Rabanus-Wallace et al. preprint cited in this manuscript, which suggests that expansions are driven by selfish elements – the number of independent expansions of a cluster across barley accessions would clearly speak to that question. It would also begin to answer questions about selection on entire clusters versus individual genes in such clusters.

Answer: The reviewer raises the interesting question of whether TEs ("selfish elements") play a role in complex loci evolution. To address this comment, we analyzed whether certain families of TEs are enriched in complex loci which could suggest a role of these TEs in formation and evolution of complex loci. We thus compared the TE composition of complex loci with the TE composition of neighboring regions (i.e. the region 1 Mb away from each locus. This was done to compare similar chromosomal regions because TE composition varies strongly along chromosomes in barley (see Wicker et al. 2017, https://doi.org/10.1186/s13100-017-0102-3). Analysis of the most abundant TE families showed no significant difference between complex loci and their neighboring regions. This is now stated in the main manuscript (ll. 242-3) and shown in Extended Fig. 6b. A description of the TE annotation at complex loci was also added to the Methods (ll. 1207-15).

Regarding the independent expansion of clusters around the genome, we sought to characterise the evolutionary variability of gene-bearing complex loci that has accumulated in ~12 million years of barley diversification by summarising the spread of their spans and repeat contents (Supplementary Fig. 7). The repeat units vary with most around the 1 kb mark, and we can see from the distribution of spans that the characterised complex loci tend to be short (10s to 100s of kb) in most accessions, with occasional big expansions in a few exceptions. The variance in the base 10-logged spans of most regions fell around 0.1, or more intuitively, a complex region observed in a randomly-chosen accession can be expected to fall within 10% of the mean span in about two thirds of instances.

I also have a series of technical concerns:

Given the comparatively large size of barley genomes, it is understandable that PacBio HiFi coverage was only about 20x. This may, however, lead to collapsing of tandem repeats, which is particularly relevant for long tandem repeats of identical copies. Here, I would have appreciated reading something about the limits of the assemblies: What are the longest tandem repeats of identical copies that the authors could have expected to discover? This concern is compounded by the fact the authors used an older version of Hifiasm. It might be useful to use the flagger tool (https://github.com/mobinasri/flagger/) to survey the assemblies for potentially collapsed regions.

The Hi-C data might also be helpful.

Answer: Among the factors that are expected to have an impact on the correct assembly of tandem repeats are, on the technical side, the length and accuracy of reads and, on the biological side, the unit lengths and homogeneity of tandem repeat arrays. We believe that an in-depth theoretical and empirical investigation into how these factors might interact is out of the scope of the present study. We point the referee to the papers by Navratilova et al. 2022 Plant Biotechnology Journal (https://doi.org/10.1111/pbi.13816), who undertook a detailed inquiry into the limitations of the barley MorexV3 reference genome assembly, and by Mascher et al. 2021 Plant Cell (https://doi.org/10.1093/plcell/koab077), who compared long-read genome assemblies of barley cv. Morex obtained by various sequencing and assembly methods.

To address the question about collapsed regions, we performed two analyses. First, we calculated coverage in the structurally complex loci we had discovered. The data are shown in Supplementary Fig. 8. Coverage was not elevated relative to the genome-wide average. Second, we determined unit lengths of the gene-containing tandem repeats that make up the structurally complex loci. The data are shown in Supplementary Fig. 7. The majority of loci are in between 500 bp and 2 kb in size, thus multiple copies are contained in a single HiFi read. This short unit length relative to the read length reduces the chance of collapsing copies during the assembly process.

The reviewer suggested to use the tool flagger. This tool was designed with diploid assemblies in mind. Barley is an inbreeding species. Heterozygosity is low, even in wild forms and traditional varieties of the crop (Milner et al. 2019 Nature Genetics, https://www.nature.com/articles/s41588-018-0266-x); genetic purity is a criterion for plant variety registration in many countries; and pure homozygous lines are readily obtained via single-seed descent. For these reasons, genome assembly in barley (and other inbreeding crops such as wheat or rice) works essentially as in a haploid organism and the cumulative size of the resultant contig set is close to the haploid genome size.

It is of course imperative to validate the homozygosity of the assembled genotypes. We did this using a k-mer based approach that operates on the HiFi data used for genome assembly (see ll. 958-67 of the Methods). We report heterozygosity estimates for our panel based on GBS markers ascertained in a large diversity panel (Milner et al. 2019 Nature Genetics, https://doi.org/10.1038/s41588-018-0266-x) in Supplementary Table 1, column N.

Low heterozygosity was a criterion of quality control in the assembly process. Indeed, we had to exclude two accessions (Kristina and WBDC 291) that were not homozygous. **These are not part of the 76 genotypes we report on in the final manuscript.** Our k-mer based heterozygosity estimates for these genotypes were 2.61 % and 4.75 %, i.e. more than 40-fold higher than the average. This result was supported by genome-wide k-mer profiles:

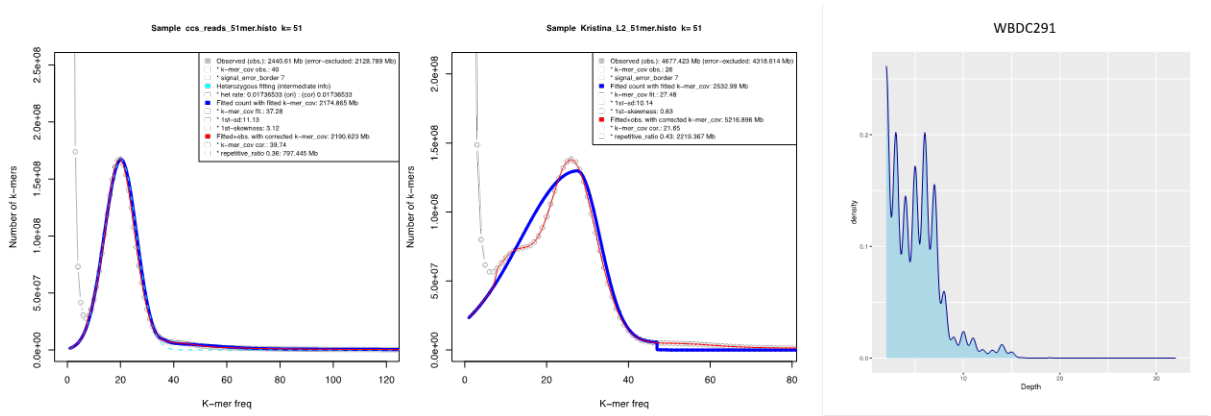

Figure: Homozygosity check with k-mer profiles: 51-mer frequency histograms were computed from HiFi data with findGSE. Left: Unimodal k-mer profile computed from HiFi data of an inbred genotype that is part of the pangenome. Middle / right: Multi-modal k-mer profiles of a HiFi dataset derived from an impure seed lot.

Only 17k out of 96k orthology groups were present in all 76 accessions, which suggests a very strict (too strict) definition of orthology groups. Technical details are unfortunately missing. More relevant would be information on syntenic orthologs.

Answer: OrthoFinder with standard parameters was used to construct Hierarchical Orthologous Groups (HOGs) from the de-novo gene annotations of 20 barley genotypes and gene projections of 56 barley genotypes. While from the perspective of total orthogroups (96,000), the number of orthogroups with genes present in all genotypes (17,000), seems small, the latter actually contain 34% of the genes. The definition of HOGs as major entities of orthologous groups (as recommended by OrthoFinder) is based on stringent criteria implemented by OrthoFinder, including cut-offs using the species tree. We believe that this definition is particularly helpful in the context of a large, but genetically diverse pan-genome. We added more details about the construction of the orthologous framework to the methods (ll. 1174-84) and made the relevant code available in a Github repository (https://github.com/PGSB-HMGU/BPGv2). Syntenic orthologs were identified by the GENESPACE software (see Figure 1b). The pangenome framework is accessible in browsable format under https://panbarlex.ipk-gatersleben.de.

I understand the difficulty of using PGGB or minigraph cactus to build a pangenome for barley-sized genomes but one could use PGGB or PGR-TK for specific loci. PGR-TK shines in the comparison of tandem repeat regions, and it is a great to answer the sort of questions I raised above.
Answer: We have constructed a local pangenome graph for the *amy1_1* locus with PGGB. Clusters defined based on structural features of the graph correlated well with copy numbers of the *amy1_1* gene in our panel of 76 genotypes. The results are shown in the new Supplementary Fig. 9.

Regarding the alpha-amylase cluster: Were the results from 21-mer genotyping consistent with the assembly-based results?

Answer: A comparison of the results from the 21-mer genotyping with the assembly-based analysis showed that the two methods are consistent (as shown in Extended Data Fig. 8c). We have now added the Pearson correlation coefficient (r=0.69) and significance level (p-value=0.004) to the legend of Extended Data Fig. 8c.

When did this cluster expand relative to domestication and subsequent improvement?

Answer: In our analysis of amy1_1 copy numbers across the 76 assemblies we see that in both wild and domesticated barley lines, more than half of the lines have 5 copies or more, but that a larger proportion of wild barley lines only have two copies. This indicates that the amy1_1 cluster expanded before domestication. Notably, the wild barley line FT880 has the highest number of amy1_1 copies. Therefore, the most straightforward interpretation is that the observed amy1_1 patterns have multiple wild origins. It remains unclear whether there was an additional wave of expansion during domestication or selection from wild ancestors. However, our data indicate that modern breeding activities select for high amy1_1 copy number with specific protein haplotype (Fig. 3). From a biological/developmental perspective we hypothesize that high amy1_1 copy number in wild barleys may improve vigor/seedling establishment and thereby be an advantage for survival in the wild. In malting and brewing the thermostability of α-amylases becomes additionally important [processing steps with increased temperature profiles such as “kilning” (drying of germinated grain) and “mashing”].

Finally, the font in many figures is too small. Also, both the figures as well as the figure legends are often far too terse (as are the methods), which makes it often difficult to understand how they support the interpretations in the main text.
Answer: *Nature*’s instructions for authors ask for short legends: “Each figure legend should begin with a brief title for the whole figure and continue with a short description of each panel and the symbols used. If the paper contains a Methods section, legends should not contain any details of methods. Legends should be fewer than 300 words each.”

Following this and the other referees' specific requests, we have expanded the Methods section and the Data and Code availability sections. We also added seven new Supplementary Figures and one new Supplementary Table.

Minor comments:

Line 130: A reference for gene flow between wild and cultivated barley is needed.
Answer: We have added two references about gene flow between wild and cultivated barley.

Line 140: Can short reads be used to test whether the two reciprocal translocations discovered among the 76 accessions are present in other accessions?
Answer: Translocations should manifest themselves in SNP matrices derived from short-read alignments as intrachromosomal LD. This was indeed what we observed in a biparental population derived from a cross between a translocated and a non-translocated genotype (Fig. 1c). However, when looking at SNP data from a diversity panel, we found that the translocated haplotypes are too rare to alter LD patterns. At the same time, we were reluctant to conclude solely from haplotypes that inversions are present. This caution is motivated by the case of an inversion on 7H reported in Jayakodi et al. 2020, whose SNP haplotype occurs in both inverted and non-inverted types.

Line 147: The point of growth of graph structures will not be obvious to those not familiar with pangenome graphs. Needs more explanation.

Answer: We have added a sentence stating that "the pan-genome growth plots illustrate the amount of presence-absence variation present in the pan-genome as a function of the number of lines added, with asymptotic curves indicating a saturation point where addition of further lines would contribute little or no further presence-absence variation" (ll. 166-170 of the revised manuscript).

Line 185: Unclear to me how anyone could deduce from Extended Fig 6b the absence of complete Mla copies. Also, what is meant with this? Truncations, premature stops, complete absence?
Answer: We have provided a more detailed version of the former Extended Data Fig. 6b as Supplementary Fig. 6. The original Extended Data Fig. 5 is now presented in larger format as Supplementary Fig. 5

Line 197: “Until the advent of long-read sequencing, it was virtually impossible to resolve the structure of the Mla locus in multiple genomes at once, but now it is a corollary of pangenomics.” What is really meant here? Also, it is a bit insulting to colleagues who have been around for a while – there are examples of reconstruction of complex loci in multiple accessions using YACs, BACs and fosmids combined with Sanger or short read sequencing. There was serious genomics before PacBio HiFi sequencing, but it required more effort than Illumina whole-genome shot gun sequencing.
Answer: We understand that the phrase “corollary of pangenomics” may sound condescending to those who have spent a lot of effort in the genomic dissection of resistance gene loci and contributed so much to our understanding of disease-resistance in crop plants. Hence we rephrased this sentence in a more forward-looking way: “We expect that pangenomes will help the genomic dissection of complex resistance gene loci in barley and other crops.”

Line 322: Is the CATCGGATCCTT motif present in Morex at all? If yes, are there ATAC-seq data that suggest it is an active cis-regulatory element?
Answer: No, this motif is not present in the Morex genome. Only the permissive C[ATC]T[ATC]GGATNC[CT][ATC] is present three times in Morex (as shown in Fig. 4c). Looking into ATAC-seq data is a good suggestion. Unfortunately, we do not have such data for the tissue in question and obtaining them is quite an intricate matter that we believe is beyond the scope of the present manuscript.

Line 329: The statement “Gene edits of the enhancer region, guided by the pangenome sequences, will further elucidate the transcriptional regulation of HvSRH1” is superfluous.
Answer: We rephrased this sentence to read “Gene edits of the putative enhancer region will be required to obtain functional proof of its involvement in the transcriptional regulation of *HvSRH1*.” We do wish to point out to our readers the remaining limitations of our current results.

Figure 1c: A comparison of Hi-C linkage with genetic linkage would be useful.

Answer: The new Supplementary Fig. 4 shows the Hi-C matrices supporting the presence of translocations in HID055 and HOR 14273.

Figure 1d: What is meant with “single-copy” here? Is this gene space only?

Answer: The single-copy pangenome is constructed by pairwise alignment and clustering of regions that are composed of single-copy k-mers (k=31) in individual genomes. We introduced this method in Jayakodi et al. 2020. The code is available from https://bitbucket.org/ipk_dg_public/barley_pangenome/src/master/.

Figure 3: This would benefit from a phylogeny of non-identical ORFs. Also, does identical mean no synonymous SNPs at all? This should be spelled out.

Answer: We thank the reviewer for the suggestion of including a phylogeny of the non-identical ORFs. Due to the high sequence identity of the amy1_1 ORFs (above 98.5%) and the accompanying paucity of informative sites we opted to construct a “median-joining haplotype network” to resolve the relatedness of the amy1_1 ORFs across all 76 BPGv2 lines. We have included this network in Figure 3 (subpanel 3c). To clearly convey the novelty and take-home message of the section “Amplification of α-amylases in malting barley” we have in large parts rewritten this paragraph. We have incorporated the “median-joining haplotype network” analysis into this rewritten text.

Regarding the second question, “identical” means no SNP differences at the ORF level, neither synonymous nor non-synonymous.

Referee #3 (Remarks to the Author):

The article from Jayakodi et al. describes the construction and use of an extended barley pangenome through sequencing and comparative genomic analyses of 76 reference-quality assemblies complemented by short read-based variant calls for 1315 genotypes. The paper does not extensively describe the pangenome on its own –not enough to my opinion – but rather focus on classical comparative genomic analysis of 4 regions taken as examples in order to demonstrate the utility of having access to a wide diversity of chromosome-level assembled genomes for research. It even includes a case study with the identification of a potential structural polymorphism that may be causal of a phenotype (related to grain development).

The paper is a typical high-impact journal-oriented paper of an international consortium (30 author affiliations) describing a new milestone on the road toward characterizing the full genomic diversity of the species. It mixes the description of a massive sequence data production effort, large-scale sequence analyses, unrelated examples of comparative genomic analyses focused on four different regions (Mla, Tb1, Amy1-1, Srh1), results from wet lab experiments (alpha amylase activity), and even the first steps of a gene cloning project (with mutants for the candidate gene, isogenic lines, expression data…). People in the field of structural genomics may regret not having a more focused paper on these aspects (I do) but others may appreciate the focus on the applications. That being said, the paper is well written, the resource is of significant importance for the community, and the methods are sound, although I have remarks about methodologies to build the pangenome (see below). Since my specialty is genome sequence analysis, I will focus my review on these aspects rather than the functional part.

Answer: We wish to point out that the comparative analysis we performed is not “classical” in the sense of adhering to past practices or established conventions. The analysis of copy number variation at *amy1_1*, *Mla*, *HvTB1* and *srh1* could not have been done without accurate long-read sequencing, which was introduced only in 2020. Structural variation at all these loci was masked by gaps and inaccuracies of previous short-read assemblies, including those described in the Jayakodi et al. 2020 paper.

The reviewer is correct in pointing out that our paper is concerned more with translational applications in crop improvement than evolutionary questions. With the former aim in mind, we believe our comparative analysis of four complex loci is not unrelated to pangenomics. These four examples emerged from a discovery-driven approach that took prior knowledge of agronomically relevant loci into account. Such in-depth inquiries into loci related to crop evolution and crop improvement are just how we envisage other applied researchers – geneticists, breeders, genebank managers – to work with our data. In this sense, our analyses are an accompanying “how-to” guide for this community resource.

Although well written, the paper is not an easy read. It is often very complex, with many references to extended data, suppl materials, which makes it sometimes a pain to read. This could be improved.
Answer: We have adhered to Nature's guidelines for authors and tried to avoid as much as possible supplementary figures, fitting as many display items as possible into the full-page Extended Data items, whose visual quality is subject to the same editorial standards as the main figures. To satisfy this and other referee requests, we had to add additional Supplementary Figures.

We also tried to limit the complexity, i.e. number of panels, of our main figures. Still, ours is a data-rich manuscript and those underlying data have to be shown in display items to support our conclusions and make our research reproducible.

My main concerns:- I found the description of the pangenome, and the level of genomics variability, were not commented at the level I would have expected for such a large effort of data production. But maybe this paper is only the main general paper and there is another companion paper dedicated to pangenomic aspects? Here there are only a short initial paragraph (likely a quality assessment) and a second ("atlas of SV") which does not provide a deep characterization of the variability. There is nothing about TE space variability (maybe this is for another paper), and regarding genes, the only value provided in main text is 17% of core genes. A bit frustrating. What about core/shell/cloud genes? what about wild versus domesticated? What are the conclusions? Comparisons with other species?

Answer: To define the core, shell and cloud genes of the barley pan-genome, we constructed an orthologous genes framework consisting of the gene annotations of all 76 barley genotypes. This framework also includes the identification of orthologous single-copy genes. We now provide results of the analysis of this orthologous framework in form of core/shell/cloud genes as a new paragraph in the main text. This analysis also contains comparisons of wild versus domesticated versus landraces gene content (CNV and PAV as well as enriched functional annotations). The reviewer is correct in surmising that there is a companion paper on deeper analysis of the transcriptome. It is currently under review at *Nature* (Guo et al., preprint available from Research Square: https://doi.org/10.21203/rs.3.rs-3787876/v1).

We did not perform comparisons to other species' pangenomes in the current manuscript. It is technically challenging to harmonize assembly and annotation workflows across different pangenome projects. But without doing so, we risk comparing apples and oranges. To give a concrete example: a very relevant comparison would be that between the wheat and barley pangenomes. However, the best available pangenome of wheat is that reported by Walkowiak et al. 2020., which was assembled from short-reads. Our comparison of short- and long-read assemblies in barley (Mascher et al. 2021 Plant Cell, https://doi.org/10.1093/plcell/koab077) has shown that the limitations of short-read assemblies hamper evolutionary analyses such as those based on gene copy number estimation. For these reasons, we believe that cross-species comparisons or "super-pangenomes" of Triticeae crops are a topic for future studies.

- I also found there was not enough link made with what was known before. I say that especially because there was already a Nature article in 2020 on the same topic and by the same leader authors (Jayakody et al. 2020) and I would have appreciated to read more about it in the introduction and about what is new here.

Answer: We have added a sentence to make mention of the effort of Jayakodi et al. 2020 (ll. 96-7). The space constraints of a Nature article do not allow for a more in-depth review of the barley genomics literature. We have laid down the rationale of barley genomics in a Perspective article in 2019 (Monat et al. Theoretical and Applied Genetics https://doi.org/10.1007/s00122-018-3234-z).

An important difference between the two Jayakodi et al. manuscripts is the expanded panel size (20 vs. 76), the inclusion of more wild barleys (1 vs. 23) and the use of a much-improved sequencing technology in our later effort. The advantages of long-read assemblies compared to short-read ones are described at length in Mascher et al. 2021 Plant Cell (https://doi.org/10.1093/plcell/koab077). Jayakodi et al. 2020 focused on the discovery of polymorphic inversions and pangenome-based GWAS. In the current manuscript, we focus on structurally complex loci and their connection to structural variation at agronomic traits.

- The Figures in the Extended data are often too small and too blurry and are not readable.

Answer: We have amended this by providing PDF instead of PPTX files during the re-submission.

SPECIFIC REMARKS AND QUESTIONS++

ABSTRACT

- Well written. I regret there is no summary of the main values expected for a pangenome analysis (number or % of core/dispensable genes etc.)

Answer: Thanks! We prefer not to report the percentages of core and dispensable genes in the abstract. Stating those numbers without providing the necessary context (which is not possible in the abstract) would mask the complexity behind those number (e.g. their dependency on panel choice, gene annotation strategies, and various thresholds).

++ INTRODUCTION

-I suggest to add some information/results regarding the previous initiatives, especially Jayakodi et al. 2020 with 20 genomes (and Russel 2016 maybe). Size of the pangenome, etc.

Answer: We have added a sentence to make mention of the effort of Jayakodi et al. 2020 (ll. 96-7). The space constraints of a Nature article do not allow for a more in-depth review of the barley genomics literature. More explanation is given in our answer one of the above comment of this referee ("I also found there was not enough link…").

++ RESULTS

+An expanded annotated pangenome of barley

- This part is poor in terms of results, and there is no conclusion. The only result I see is the average 2% of the ~5000 single-copy genes of Poales that are absent. And there is no associated Extended Data to better describe this. It raises more questions than it brings information. Where are the results about the description of core/shell/cloud genes? I understand these results are more provided in the next paragraph (atlas of SVs)… this is not clear to me what is the purpose of this first paragraph. If it is related to genome quality assessment, I suggest to give a title accordingly and to conclude the results.

Answer: We have changed the title of this section to "Annotated genome sequences of 76 barleys" to reflect the technical and expository nature of this paragraph.

- The method employed here to build a pangenome is questionable. At least it makes appear some weaknesses, obviously due to the difficulty of analyzing these genomes correctly, but that limit our ability to reach the goal of building a high-quality pangenome. I understand that genes were predicted in 20 genomes while only projected in the 56 others. This may have strong consequences in deciphering the pangenome. I know and understand how difficult it is to get an accurate, well predicted, gene set for these genomes. But I would have appreciated to read conclusions/discussions about that. The most important question is: what is the impact of the strategy (gene projection instead of 76 independent gene predictions) on the results? Obviously if specific genes are not

predicted de novo, they cannot be present in the pangenome! Please, could the authors try to discuss this?

Answer: Indeed, we have direct transcriptional evidence for only 20 of the 76 gene annotations. We agree that it is not possible to discover de novo genes without such evidence. Consequently, we did not attempt to find such genes. We added a sentence stating this limitation of our dataset in ll. 148-51 of the revised manuscript.

- Why focusing on Poales single copy genes? And if these genes are the super-conserved core-genes of all Poales, how do the authors conclude on these 2%? Was this a way to estimate the error rate due to incompleteness of genome assemblies, annotation problems, etc. (I think yes)? How many of them are missing in all accessions? If there is 1-2% of genes missing in the assemblies, it could however makes the % of genes present in 76 accessions very low, this is why the authors maybe could have computed a "soft-core" gene set?

Answer: We have clarified that we used the BUSCO Poales set only to assess the completeness of our annotations. We did not aim for a more advanced phylogenomic inquiry.

+ An atlas of SV

- This paragraph is a mix between SV detection at the whole genome sequence level (i.e., the title) and a gene clustering-based pangenome analysis… which are two different approaches to achieve similar things. I found this part is interesting but is hard to follow in the way it is written here. I suggest to try to make the text easier to read. The only value I found is the 17% core-genes. It appears very low, probably because calculated relatively to all pan-genes and not to genes present per individual. So, what is the percentage of core-genes per accession (16k/~50k)? This looks also low. I suggest to comment more on these values and to compare with knowledge from other species, which would give much more interest in reading this part.

Answer: We have expanded and revised the paragraph on genic presence/absence analysis. We believe that an interspecies comparison of pangenome complexity is out of scope of the present manuscript owing to technical challenges in harmonizing genome assembly, gene annotation and taxon sampling across species.

- "At the level of individual gene models, a third were considered conserved because they belong to an orthologous group with representatives from each accession". This is an example of sentence one could read several times but one could not make sense of it.

Answer: We rephrased the entire section (ll. 130-51 of the revised manuscript).

- What about genes subject to CNVs? (cf Extended Data where I realized there were several rounds of gene projections, probably to identify CNVs?)

Answer: We have expanded the description of the orthologous framework in form of core/shell/cloud genes (ll. 130-51) of the revised manuscript. This analysis also contains the identification of hierarchical orthologous groups subject to CNV and PAV. We compared wild barleys, cultivars and landraces and report functional enrichment in variable HOGs. As discussed above, our gene projection approach is capable of calling duplicated genes, but does have limitations in predicting true *de novo* genes.

- "The functional annotations […] pointed to an involvement in biotic and abiotic stress responses (Supplementary Table 4)". I read quickly the Supp table and saw many terms not related to stress response.

Answer: We have rephrased the text to better reflect the content of the table.

- "Pangenome graph". Hard to see the added value of this part, except to comment on the fact that building a high-resolution pangenome graph is still too complicated in barley and related complex genomes. However, if I understood correctly, the graph was used for mapping resequencing read of a large diversity panel in order to estimate the representativeness of haplotypes captured in this pangenome. This part is interesting.

Answer: Our primary objective in constructing a pan-genome graph was to create a data structure that can be used for routine bioinformatics analysis in barley. The fact that all but one of the existing graph toolkits are computationally prohibitive for a genome of this size and complexity is an important finding that needs to be reported in order to a) inform the graph tool developer community of these shortcomings and b) create awareness among colleagues attempting a similar approach in other taxa. We have, however, shown that the graph is still useful for global analyses of the kind presented here, and that it has uncovered substantially more structural variants than the linear alignment tool.

- SVs based on genome sequence alignments + SVs based on graph-based mapping… any differences observed? (Maybe hard to answer that).

Answer: We have carried out a comparative analysis of insertions and deletions (numbers, sizes and positions) called from the graph and the linear alignment and have presented the results in the manuscript.

- "will be a desirable" -> should be "as desirable"
Answer: Done.

+ An inventory of complex loci
- Mla locus: very descriptive, although I understand that pangenomics is often very descriptive. The results here could be summarized as: there were 29 known Mla genes, 7 of them are present in the pangenome, but 149 homologs were clustered here? and a landrace contains 11 genes, with 2 being present in 5 copies. The conclusion is interesting, saying that resolving complex loci is now feasible and a corollary of pangenomics. But I found we miss a conclusion on Mla gene SVs.

Answer: We have rephrased the conclusion of this paragraph: “We expect that pangenomes will help the genomic dissection of complex resistance gene loci in barley and other crops.”

- The thionin gene CNV example made me wonder about the strategy employed to annotate genes in the 76 genomes. I though, as explained, that genes were predicted denovo in 20 genomes and projected in the others. How does this impacted the ability to identify CNVs here? (redundant with my previous remarks on CNVs).

Answer: Our gene projection approach is capable of calling duplicated genes. We performed two rounds of gene projections to ensure (i) that each model has been considered at least once for a potential match and annotation, and (ii) that CNVs and gene duplications were detected for high confidence source genes. Restricting our attention to high-confidence gene models also avoids technical artifacts from potential pseudogenes and TE-derived input genes.

However, for the analysis of the 172 complex loci, pseudogenes were considered. The description of the annotation of pseudogenes at complex loci was added to the online methods (ll. 1169-72).

- Paragraph on dating duplicated loci is really interesting. But the associated Extended data 7 is not readable (blurry picture).

Answer: This has been corrected.

- "enriched in distal chromosomal regions (Fig. 2d)" -> should be Fig. 2c
Answer: Corrected.

+ Amplification of α-amylases in malting barley
- First paragraph is very descriptive, mentioning many details but without conclusion. Worth example is the fact that "12 had insertions of TEs". In different copies? Is it a shared polymorphism or do the authors mean that they were several independent insertions of different TEs? in the same copy?

It is hard to get what is new here compared to what was known before in term of diversity.

Answer: We have rewritten the section “Amplification of α-amylases in malting barley” to better convey the key take-home messages:

1) The genomic structure and linear order of the *amy1_1* cluster across diverse barley lines are for the first time resolved. This has great importance for targeted breeding approaches towards improved malt quality.
2) Ours is the first report of a correlation between *amy1_1* copy number, protein haplotypes and malt quality.
3) We determine *amy1_1* CNV across BPGv2 (and a large barley diversity panel). This identified a wild barley with 8 copies which has the potential to greatly increase the α-amylases activity of elite malting barleys (towards improved malt quality).

+ A regulatory variation controls trichome development
- The cloning about HvSRH1 is quite interesting indeed. And it is a remarkable example of the utility of the pangenome, well, at least having access to multiple well assembled genomes in order to detect structural variations that may cause phenotypic variations.

Answer: Thanks! We are glad the reviewers liked this part.

- It is mentioned a 4kb segment absent in all (14) short-haired genotypes while "exceptionally conserved" in the others i.e., with 95% identity. To me, it sounds contradictory because 95% nucleotide identity is low between two genotypes of the same species!?

Answer: We have added that this sequence segment is non-coding. We agree that we would need to explain our prior expectations in greater detail to justify the use of the word “exceptional”. We now simply say “well conserved”.

- "CATCGGATCCTT, matching the sequence [ATC]T[ATC]GGATNC[CT][ATC]".
The first "C" is not part of the motif, so I guess there is something to correct here. In addition, when searching for the presence of this motif across the interval, which one was searched for? the strict one (first) or the permissive one (second)? I scanned the Morex genome with the permissive one in order to realize that such motif is present every 8kb on average. So, the probability to find it by chance in a 4kb segment is actually high. When I search for this motif HTHGGATNCYH in the 120kb interval of Fig4, I find it 18 times which is different from the 3 motifs of the figure. I can only suggest

to double check this analysis.
Answer: We thank the reviewer for checking this carefully. We made indeed a mistake by omitting "C" as the first letter of the sequence motif. In the revised manuscript, this sentence reads now ". Within this sequence, we found the motif CATCGGATCCTT, matching the sequence C[ATC]T[ATC]GGATNC[CT][ATC], …". The correct motif with the initial "C" is present only three times.

- The expression data illustrated at panel f of Extended Figure 10 is quite interesting. It brings a serious argument and I suggest to try to include this in the main paper.
Answer: Thanks for commending this panel! Unfortunately, we cannot add it as a main figure because of constraints on the size and number of display items.

++ Discussion
- "true to the hypothesis-generating remit of genomics", guess you mean "merit" right?

Answer: We replaced "remit" with "purview".

- First part is more a discussion about the interest of long-reads in complex genomes than the interest of pangenome, although I understand the two go together for complex genomes.
Answer: Indeed, the high contiguity afforded by long reads was crucial for many of the analyses we did. The use of PacBio HiFi reads as opposed to Illumina short reads is also an important difference between our current efforts and that of Jayakodi et al. 2020. This is why we chose to give it such a prominent place in the discussion.

++ EXTENDED DATA
- Probably just a problem with the PDF conversion, but many panels of Figures are not readable. I suggest to improve the quality of the figures here. Effort has been made to detail each legend, a good point for Supplements.
- Ext Fig 2: Panel e is missing in the legend
- Ext Fig 3: Typo "pagenome". Text is too small to be readable.
- Ext Fig. 6: Alignment not readable. And I do not see how to get a message from Panel b.
- Ext Fig. 7: Cannot read that
- Ext Fig 10: I wonder if panel f should not be better part of the main Figures

Answer: We have corrected these things. We have decided to keep Extended Data Figure 10 panel f in its original place so as not to make main Fig. 4 larger than it already is.

- SNP and SV calling: first sentence is not finished, there is a verb missing. Same for the last sentence. I suggest to proofread the whole paragraph.

Answer: We have corrected this.

- Gene projections: this part is really hard to follow while it is critical to assess the work done and to be able to estimate the limits of the method.
"For the two top quality categories, we performed two rounds of projections, firstly inserting each source maximally only once followed by rounds allowing one source inserted multiple times into the projected annotation". I may understand that genes under CNVs were annotated here because there were several rounds of projections? If yes, I think it should be clearly mentioned, probably not only in the Extended text. And if yes, why not providing results about that?

We have expanded the description of the orthologous framework in form of core/shell/cloud genes (ll. 130-51) of the revised manuscript. This analysis also contains the identification of hierarchical orthologous groups subject to CNV and PAV. We compared wild barleys, cultivars and landraces and report functional enrichment in variable HOGs. We have also set up a Github repository at https://github.com/GeorgHaberer/gene_projection/tree/main/panhordeum, which includes a more detailed description of the different iterations of the method, including the annotation and identification of CNV.

- line 1153: "to detection the different Mla…" to be corrected.

Answer: Corrected.

- Code availability: "Scripts for calculation of core/shell and could genes". If such things were done, I would really appreciate to see these results.
- I quickly visited the github page https://github.com/PGSB-HMGU but cannot find any repository related to core/shell/cloud gene calculation.

Answer: We now describe core, cloud and shell genes in the main text (ll. 130-51). We also uploaded all relevant scripts for core/shell/cloud gene calculation to the Github repository at https://github.com/PGSB-HMGU/BPGv2.

- Regarding gene projection, scoring, iterations, I suggest to also make the codes available for reproducibility.

Answer: We have extended the description of the gene projection method in the Online Methods and provide the code at the Github repository https://github.com/GeorgHaberer/gene_projection/tree/main/panhordeum. This directory now also includes a detailed description of the individual iterations of our gene projection procedure.

Reviewer Reports on the First Revision:

Referee #1 (Remarks to the Author):

The revised manuscript has improved clarity in the methods section. The added pan-genome website and the pangene table should be very helpful for the barley community. A brief browsing of the website has provided a smooth experience and very helpful information in connecting genes within the pangenome. The reviewer identified several previous comments that were not fully addressed, including a suspicious use of HiC data on the LD plot, the mappability of the graph pangenome, the non-random use of mutant lines for genotype verification, and the lack of power for TE and haplotype comparisons between complex loci. Also, using the repeatedly reported srh1 gene to showcase the power of a pangenome is weak and not novel. Properly addressing these issues will further improve the robustness of this work. More detailed comments:

(1) The issue of Figure 1c is not addressed - Figure 1c looks like a HiC heatmap which shows chromatin interactions. Chromatin interactions are not equivalent to linkage disequilibrium. The latter quantifies the non-random association of alleles of different loci in a given population. Authors should present the actual LD data from segregating offspring of HID055 x Barke.

(2) L186-202: It is unclear whether Morex or the pangenome was used for the map-based analysis. Also, it's unclear whether differences in mapping rate had more to do with the alignment tool or the use of reference (linear or graph). From the response, the failure of mapping PacBio HiFi reads to the graph pangenome suggests the pangenome construction is not an improvement over a single genome. This may have also been the cause of the lower mapping rate of pangenome presented in Extended Data Fig 4b. Thus, presenting these mapping data could be misleading without further discussing the technical caveat. Please either remove the mapping rate data or provide cautious reminders for readers.

(3) It's incremental and technically unchallenging to reidentify the srh1 gene, as suggested by the authors' response. Even using the old school linkage mapping can identify the locus, which was also demonstrated in this study. Using it as one of the highlights in this pangenome barley paper does not demonstrate the power of pangenomics and the novelty of the work. Maybe they can use the pangenome graph – a decentralized approach - to analyze the rachilla hair in the core1000 dataset. However, the power of the pangenome graph may have been undermined by the mapping rate issue discussed in the previous comment. Figure 4c shows the causal variant of the rachilla hair phenotype is a structural variant. The authors may also try to identify the locus using SV data of the pangenome, which is a more novel approach compared to the traditional SNP-based GWAS.

(4) The response to comments on Suppl. Table 25 (now #26) and Fig 4d is not satisfactory. They selected four M1 mutants, including one homozygous WT and three heterozygous MTs. These selections were biased because they selected the three mutant seeds based on them having both the mutant genotype and phenotype. However, 11 out of 15 heterozygous mutants carrying the wild-type phenotype were excluded. These heterozygous M1 lines with the WT phenotype may suggest alternative genetic mechanisms for the rachilla hair phenotype. Further, they grow the three M1 mutant lines for further phenotyping and genotyping, but only select one M2 plant from each M1 mutant for presentation. These were cherry-picked on top of the cherry-picked. The "perfect

correlation between mutant phenotype and genotype" they claimed was likely due to the heavy scrutinization of mutant lines. They should have quantitatively phenotype all M2 lines and compare between homozygous and heterozygous mutant phenotypes and also contrast to all M2 lines from the homozygous M1 WT-phenotyped plants.

(5) Initial review: Different from their previous pangenome (PMID: 33239781), they did not perform any de-novo TE annotation in these genomes. Still, several examples in this manuscript suggest the important role of TEs in the function and duplication of genes, such as Figure 2b, Figure 3a, and Figure 4c. Pangenome studies in maize and rice, two other important grass crops, showed that TEs are overrepresented in structural variations. Annotation of TEs novel to the PGSB library in the diversity panel could be valuable to discern the cause of functional variations.

Response: We agree that TEs are an important source of genetic variability and that pangenomes are an excellent tool to study TE-related diversity. Following the referee's suggestion, we have used the PGSB library to obtain full-length LTR retrotransposon (fl-LTR TEs) annotations for all 76 genomes. We report the number of fl-LTRs in Supplementary Table 2 and show the distribution of insertion ages in Supplementary Table 1. We did not observe striking differences between the genomes. We agree that TE annotations can provide valuable clues as to the potential causes of functional variation. We do report links between structural variation (HvTB1, srh1, amy1_1) and phenotypes, but TEs happen not to play a role in these cases. We are aware that these examples are not representative and that our fl-LTR annotation is only an initial foray into the pangenome's TE diversity. Deeper analysis including the relationship between SVs, TEs and epigenomic features will be the subject of a separate manuscript.

Second review:
1. Fl-LTR TEs were not obtained using the PGSB library but using LTRharvest based on their methods.
2. Distribution of LTR insertions could not be found in Suppl. Table 1. Maybe they meant Suppl. Fig. 1?
3. The similar number of fl-LTR TEs in Suppl. Table 2 does not support the lack of novel LTR families in the barley pan-genome. Novel TE families are defined based on their sequences being novel compared to other genomes, not the number of sequences.
4. The claim "TEs happen not to play a role" in complex loci was based on analyses reported in L1207 – 1215, in which they split complex loci into 200bp sequences and blast against the TE library to count for accumulated TE length, then compared to that of a distant region for TE length differences. They made a similar mistake here, that the similar TE length over a large region (1Mb) does not support the lack of effect from TEs. Usually, only one TE fragment is involved in the regulatory effect of the target gene, which would not make a difference in cumulated TE length over 1Mb. They should present haplotypes of complex loci with both gene and TE annotations in the barley pangenome. In fact, they presented one such haplotype comparison in Figure 4c and showed many structural variants, suggesting TEs may play a role in cis-regulatory functions.

(6) It might help to indicate the sizes of your four selected loci, the # of tandem genes, repeat composition, and HiFi sequence coverage and length over those loci. The article takes a deep look at those four regions, which seems to necessitate a detailed assessment that your sequencing platform and coverage (20x HiFi) could resolve them. Suppl. Figures S7 and S8 try to address this but not entirely. S8 could be improved by including read depth at non-complex loci for comparison.

(7) Does projecting gene models influence your analysis of cloud/core/shell and HOGs?

Minor:
(1) cvs (cultivars) is an unspecified acronym.

(2) L86: 'in of' - typo

(3) L121: of the 20 de novo annotated accession, please state how many were wild and domesticated. Please also state what tissues were used.

(4) It's confusing to have both Supplementary Figures and Supplementary Data Figures.

(5) In the rebuttal, "We have also added Supplementary Figure 4, which shows the Hi-C contact matrices for the translocated genotypes and alignments of their pseudomolecules to non-translocated Morex. A clear inter-chromosomal Hi-C signal is seen if the Morex karyotype is used as a reference for aligning Hi-C reads but absent if the respective translocated karyotype is used." But the legend of that fig says "Hi-C contact matrices before correction (i.e. using the chromosomal configuration of Morex). (c) Hi-C contact matrices after correction" It's unsure whether the karyotype was translocated or misassembled based on the description discrepancy. It looks like the comment in the rebuttal better describes the figure.

Referee #1 (Remarks on code availability):

Links to the code are provided and could be open. However, I did not tested the code myself.

Referee #2 (Remarks to the Author):

I was reviewer #2 in the initial round of reviews.

I find the study to be mostly unchanged. As an example, I was disappointed that the authors did not take seriously my criticism that the work did not speak to the processes of adaptation. Furthermore, there is still nothing in the manuscript that supports claims of “rapid evolution”, as prominently mentioned in the abstract. The authors state that they investigate “structurally complex loci that have become hot spots of gene copy number variation in evolutionarily recent times”. This seems to imply that mutational patterns at these loci have recently changed. If this were indeed the case, this would be an amazing finding, but sadly this hypothesis is not tested in the manuscript.

There is a long section on copy number variation at the alpha-amylase locus. The authors state “Copy numbers of amy1_1 (had) on average more copies in domesticated than in wild forms” but there is no statistical test, and the differences in any case seem small. More importantly, how does copy number variation in the material examined correlate with amylase activity? A cursory perusal of the literature reveals that in vitro alpha-amylase activity varies considerably, but in the current work we learn nothing about the underlying reasons, specifically whether the known variation is due to differences in protein amounts as a consequence of gene copy number variation, due to differences in amino acid sequences and hence specific protein activity, or due to differences in promoter sequences. Also, why is there such enormous variation at all levels, rather than just expansion/contraction of very similar copies? I asked specifically what we learn about common versus independent copy number expansions/retractions, but this was also ignored by the authors in their answer.

I had a few technical concerns, and the answers to these did not convince either. I asked about the surprisingly small number of universally shared orthologs. The authors essentially state “well, that’s how the program we used, OrthoFinder, works”. They also did not bother to check whether this may have something to do with potentially problematic annotations.

In summary, this resource paper is in several aspects technically impressive, but at the same time intellectually disappointing, as it breaks little new conceptual ground. As I had stated in my confidential comments to the editor, a manuscript of this ilk needs a firm editorial stance whether or not this is the type of paper they want to publish. The good news is that there are many competent colleagues who will likely use this fantastic resource to give us more serious answers to questions about the evolutionary processes of mutation, selection, and adaptation that the (non) answers provided by the authors. I note that the authors did not contest my assertion that “the advance for the broader community does not go beyond similar types of papers published elsewhere for tomato, rice and soybean two or three years ago, and it does not go nearly as far as the (graph) pangenome papers for tomato and potato published in Nature last year”.

Despite my gripes, the work would not look out of place in the pages of Nature. Reviewer comments can only help so much in making editorial decisions.

Technical comments/suggestions for the authors to consider:

1. For the alpha-amylase cluster, the Pearson correlation coefficient is considerably lower than what was reported for humans (R=0.69 [R^2=0.4761] vs R=0.99, Vollger et al., Nature 2023). Did this perhaps come from the repetitive 21-mer introducing a bias or from partial assembly collapse?

2. You seem to build individual graphs for each chromosome. One may lose much of the variation inside the common large translocation using this approach. Have you considered “correcting” the translocation for alignment purposes?

3. BUSCO cannot distinguish between high copy number variation genes (such as alpha-amylase genes) and rapidly evolving genes (Mla). One could use the available RNA sequencing reads mapped back to the annotated gene models to support high completeness of pangenome annotation.

4. Why is embryophyta_odb9 used for assembly BUSCO and Poales_odb10 for annotation?

5. Fig. S8: How does read depth compare to non-complex regions?

Referee #3 (Remarks to the Author):

I thank the authors for the clear and detailed point-by-point answer to my remarks. I now understand the reasons behind things that could have been appeared questionable in the first version. I recognize the authors made great efforts to provide more results on core/shell/cloud genes and I found the paper has been significantly improved in quality both in term of results and writing.
I appreciated to read the novel section "Atlas of structural variations" with now interesting results on pan/core-genome and specificities of wild vs. cultivated etc., which was my main concern previously. In addition, the introduction now makes the link with previous initiatives and it looks clearer for people not fully aware of the recent papers on barley genomics. The Figures of the supplementary material are all readable in the revised version. The Github repository now contains the expected codes. Finally, I found this revised version much easier to read and follow.

Author Rebuttals to First Revision:

Referee #1

The revised manuscript has improved clarity in the methods section. The added pan-genome

website and the pangene table should be very helpful for the barley community. A brief browsing of the website has provided a smooth experience and very helpful information in connecting genes within the pangenome. The reviewer identified several previous comments that were not fully addressed, including a suspicious use of HiC data on the LD plot, the mappability of the graph pangenome, the non-random use of mutant lines for genotype verification, and the lack of power for TE and haplotype comparisons between complex loci. Also, using the repeatedly reported srh1 gene to showcase the power of a pangenome is weak and not novel. Properly addressing these issues will further improve the robustness of this work. More detailed comments:

(1) The issue of Figure 1c is not addressed - Figure 1c looks like a HiC heatmap which shows

chromatin interactions. Chromatin interactions are not equivalent to linkage disequilibrium. The latter quantifies the non-random association of alleles of different loci in a given population. Authors should present the actual LD data from segregating offspring of HID055 x Barke.

Answer: We agree that chromatin interactions are not equivalent to LD, but also reassert that Fig. 1c does show LD data. To bolster that claim, we have added the underlying data table to the supplementary dataset that is to be released under a DOI (temporary DOI link for reviewers: https://doi.ipk-gatersleben.de/DOI/d8544ae7-97e2-4723-9837-2e3407adc3a6/b3b427a2-755c-4488-abda-2b3847d7902b/2/1847940088). The relevant file is named

"gen_pos_m1m2_03_2no.csv" and located in the folder "LD".

The close ressemblance between Hi-C and LD matrices is explicable by the facts (i) that they both show "interactions" between pairs of genomic loci and (ii) that in the presence of a reciprocal translocation between chromosomes both show an interchromosomal signal.

(2) L186-202: It is unclear whether Morex or the pangenome was used for the map-based analysis. Also, it's unclear whether differences in mapping rate had more to do with the alignment tool or the use of reference (linear or graph). From the response, the failure of mapping PacBio HiFi reads to the graph pangenome suggests the pangenome construction is not an improvement over a single genome. This may have also been the cause of the lower mapping rate of pangenome presented in Extended Data Fig 4b. Thus, presenting these mapping data could be misleading without further discussing the technical caveat. Please either remove the mapping rate data or provide cautious reminders for readers.

Answer: In order to eliminate tool bias as a confounding factor, we have produced a linearised version of the pan-genome graph, and mapped reads to this following exactly the same approach used for mapping to the linear Morex V3 reference sequence. We have also restricted the analysis to perfectly matching reads so we can eliminate any issues with mismapped reads. The results are shown in Extended Data Fig. 4b. As expected, more reads were mapped to the linear pan-genome

than to the linear single cultivar sequence, while mapping rates to the graph are consistently lower, presumably due to algorithmic differences in the mapping tools.

In response to the reviewer's comments we have also included an additional analysis of the pan-genome graph with regards to the *srh1* gene. We have added a graph-based haplotype plot (Figure 4d) that clearly contrasts accessions with short and long rachilla hair, and we have also mapped Illumina WGS reads onto the graph and used the approach recommended by Hickey et al. for structural variant genotyping of graphs, which reliably identified a variant corresponding to the deleted region in all of the accessions tested. We believe these additional analyses have improved the paper by showcasing further applications of the graph-based pan-genome and would like to thank the reviewer for their suggestions.

(3) It's incremental and technically unchallenging to reidentify the srh1 gene, as suggested by the authors' response. Even using the old school linkage mapping can identify the locus, which was also demonstrated in this study. Using it as one of the highlights in this pangenome barley paper does not demonstrate the power of pangenomics and the novelty of the work. Maybe they can use the pangenome graph – a decentralized approach - to analyze the rachilla hair in the core1000 dataset. However, the power of the pangenome graph may have been undermined by the mapping rate issue discussed in the previous comment. Figure 4c shows the causal variant of the rachilla hair phenotype is a structural variant. The authors may also try to identify the locus using SV data of the pangenome, which is a more novel approach compared to the traditional SNP-based GWAS.

Answer: We agree with the reviewer that re-identification of a known gene would be technically unchallenging. But this is not case here. There is a crucial difference between *mapping* a mutant, i.e. finding a genomic region of possibly large size that is statistically associated with the occurrence of the mutant phenotype (as in earlier GWAS studies), and *cloning* genes, i.e. identifying a gene and validating its function (the present study). After this general introduction, we concede that our language in the manuscript and the response letter might have been ambiguous.

Genetic mapping of the *Sr/i1* locus has been reported several times in the literature, but the *Sr/i1* gene has eluded efforts at cloning. In this study we have identified the locus at highest genetic resolution using biparental and GWAS mapping as complementing strategies, employing unpublished data, revealing only partially overlapping results. The SMR-like gene, which could be confirmed in the present study by independent chemical mutagenesis *and* gene editing as causal to the *sr/i1* phenotype, was covered in the large interval delineated by GWAS but lay just outside the biparental mapping interval. The power of the pangenome allowed us to identify a structural variant, carrying a highly conserved enhancer motif known to be involved in the regulation of the SMR gene in Arabidopsis. This gene is perfectly linked to the rachilla hair phenotype and resides in the overlap between the biparental and the GWAS mapping intervals. We believe that our *Sr/i1* results will be of high interest of other researchers who have observed seemingly inexplicable inconsistencies between high-resolution mapping data and very plausible candidate genes located just outside of the mapping interval because the causal variants reside in a regulatory region.

To diminish the focus on mapping (as opposed to cloning) and to take on the advice from the reviewer to aim for stronger demonstration of the power of the pangenome, we have moved the former Figure 4b to the supplements (now Supplementary Figure 11) and added instead a new panel as Figure 4d to show the local pangenome graph of the *Sr/i1* locus in all 76 pangenome genotypes. This graph clearly shows the diagnostic structural variant.

(4) The response to comments on Suppl. Table 25 (now #26) and Fig 4d is not satisfactory. They selected four M1 mutants, including one homozygous WT and three heterozygous MTs. These selections were biased because they selected the three mutant seeds based on them having both the mutant genotype and phenotype. However, 11 out of 15 heterozygous mutants carrying the wild-type phenotype were excluded. These heterozygous M1 lines with the WT phenotype may suggest alternative genetic mechanisms for the rachilla hair phenotype. Further, they grow the three M1 mutant lines for further phenotyping and genotyping, but only select one M2 plant from each M1 mutant for presentation. These were cherry-picked on top of the cherry-picked. The "perfect correlation between mutant phenotype and genotype" they claimed was likely due to the heavy scrutinization of mutant lines. They should have quantitatively phenotype all M2 lines and compare between homozygous and heterozygous mutant phenotypes and also contrast to all M2 lines from the homozygous M1 WT-phenotyped plants.

Answer: We take the opportunity to explain our approach for mutant analysis in greater detail than is possible in the manuscript. This should make it clear that we were not cherry-picking and that the observed patterns of segregation are in line with our experimental design and do not contract the hypothesis that knockout of the SMR-like homolog causes the *sr/i* phenotype.

The SMR-like gene was analysed by two independent mutant analysis strategies. First, we screened by FindIt™ an ethylmethanesulfonate (EMS) mutagenized population derived from the long rachilla hair cultivar 'Etincel' and we could identify a single mutant that was affected by a predicted functionally relevant amino-acid change (non-synonymous) mutation (Extended Data Figure 10). The mutant clearly showed the short rachilla hair phenotype and thus provided us with a strong proxy about the status of the gene, while by itself, we agree, this would not have been sufficient to claim success in cloning *Sr/i1*.

Second, in a fully independent attempt we used Cas9 gene editing in the long rachilla hair genotype 'Golden Promise'. Gene editing in barley is highly efficient but time-consuming. In the primary transformant generation T0/M1, often no phenotypic effects are seen if the character in question is recessive. Nevertheless, it is still possible – and does happen quite frequently – that copies of the causal gene on both chromosomes are mutated (homozygous or hemizygous/chimeric for independent disruptive mutations). We opted for the use of multiple guide RNAs in the same transformation experiment so as to maximize the chances of mutagenesis. We have now provided phenotypic data for (i) all analysed T0/M1 primary transformants, that showed at least in individual grains the short rachilla hair phenotype, and (ii) for the progeny of four primary transformants (T1/M2 families). This was not "cherry-picking" because we had, on the basis of the FindIt experiment, initial evidence that the targeted gene is the right candidate. It was an important measure to deal with generation time in barley and effort. We did not see a reason to independently study segregation of mutants in a wildtype/mutant heterozygous family to demonstrate again linkage of the phenotype and the mutant, which anyway would have required a relatively large population. This approach is furthermore complicated by the fact that the transgene including the Cas9 enzyme is still present in the transgenic family – which was the case in all the analysed families of Cas9-derived mutants. As the reviewer can see from Supplementary Table 26, the four studied T1/M2 progenies segregated for a number of independent disruptive mutations, hence the T0/M1 plants were indeed chimeric/hemizygous for independent mutations. In family brhE19 a single plant P20 showed a wild type phenotype. This observation is consistent with the fact that this plant carried a homozygous in-frame, non-disruptive 9-bp deletion, which was expected not to have a phenotypic effect. We emphasize that this was the only case we found of a homozygous mutant with a wild-type phenotype. In all other progeny that produced viable plants, the individual guide RNA locus

mutations or the combination of mutations at both loci produced frame-shift mutations or larger deletions and all plants showed the mutant phenotype. The wild-type non-transgenic family brhE10, selected from the same transformation/gene editing experiment, produced exclusively wild-type progeny (Supplementary Table 26).

Considering the results of both independent mutagenesis strategies (FindIT and Cas9), we have confirmed by (i) at least 10 independent mutational/editing events and (ii) a perfect correlation of phenotype and genotype that our candidate gene *SMR-like* is indeed the *Sr/i1* gene.

(5) Initial review: Different from their previous pangenome (PMID: 33239781), they did not perform any de-novo TE annotation in these genomes. Still, several examples in this manuscript suggest the important role of TEs in the function and duplication of genes, such as Figure 2b, Figure 3a, and Figure 4c. Pangenome studies in maize and rice, two other important grass crops, showed that TEs are overrepresented in structural variations. Annotation of TEs novel to the PGSB library in the diversity panel could be valuable to discern the cause of functional variations.

Answer: We agree that TEs generally influence genome evolution in many different ways, and the initial gene duplications at complex loci may have indeed been caused by recombination over repeated sequences. However, in the particular case of the complex loci, our data indicate that CNV is mainly driven by unequal homologous recombination once initial tandem repeats (containing the genes) were established. For example, at the *HvTB1* locus (Fig. 2b), the duplication/CNV seems purely driven by unequal crossing-over between the ~20kb tandem repeat units that contain the genes. Unequal recombination between neighbouring tandem repeats is a well described phenomenon that is essential e.g. in the concerted evolution/homogenization of ribosomal genes.

In Figure 3a, the TE insertion in Morex is an example of a recent TE insertion that occurred after the duplication of the genes, leading to a disruption of that particular gene copy. This is to be expected from time to time, since TEs constantly (and mostly) randomly insert into the genome.

To investigate the potential impact of TEs on functional diversity, we did a more detailed analysis of TE polymorphisms in gene promoters in complex loci (see our answer to this reviewer's comment below).

Figure 4c shows the deletion of regulatory elements at the *Srh1* locus. Here, we found no indication that this was TE related. Instead, it is likely a deletion that was the result of DNA repair, for example through single-strand annealing (see e.g. J Biol Chem. 2018;293:10536-10546. doi: 10.1074/jbc.TM117.000375.). This DNA repair pathway results in deletions of mostly random segments.

We have added statements emphasizing that the main driving force for CNV in complex loci is unequal homologous recombination, but that TE insertions and random deletions contribute to diversification of individual gene copies:

> *Complex loci were enriched in distal chromosomal regions (**Fig. 2d**). In this regard, they follow the same distal-to-proximal gradient as genetic diversity and recombination frequency in barley. The latter process might play a role in their amplification and contraction owing to unequal homologous recombination between neighboring repeat units[29] (**Extended Data Fig. 6a**). We found no association of complex loci with specific TE types (**Extended Data Fig. 6b-d**). Instead, molecular dating of the tandem duplications in Morex is consistent with recent and recurring duplication/contraction*

*cycles, leading to complex pattens of higher and lower order tandem repeats (**Extended Data Fig. 7**). Indeed, many gene copies appear to have been gained within the last three million years (**Extended Data Fig. 7c**), after the H. vulgare lineage split from that of its closest relative H. bulbosum[30]. In addition, 63 loci (36.4%) underwent at least one duplication in the last 10,000 years, that is, after domestication (**Extended Data Fig. 7d**). Forty-five loci expanded so recently that the genes they harboured were identical duplicates of each other. Despite high similarity of duplicated segments, TE insertions (or excisions), random deletions and mutations contribute to diversification or pseudogenization of individual gene copies over time (**Fig. 3a**, **Supplementary Fig. 9a**)*

In addition, the new Supplementary Figure 9 focuses on the diversification of gene copies through TE insertions, deletions and mutations.

> Response: We agree that TEs are an important source of genetic variability and that pangenomes are an excellent tool to study TE-related diversity. Following the referee's suggestion, we have used the PGSB library to obtain full-length LTR retrotransposon (fl-LTR
>
> TEs) annotations for all 76 genomes. We report the number of fl-LTRs in Supplementary Table 2 and show the distribution of insertion ages in Supplementary Table 1. We did not observe striking differences between the genomes. We agree that TE annotations can provide valuable clues as to the potential causes of functional variation. We do report links between structural variation (HvTB1, srh1, amy1_1) and phenotypes, but TEs happen not to play a role in these cases. We are aware that these examples are not representative and that our fl-LTR annotation is only an initial foray into the pangenome's TE diversity. Deeper analysis including the relationship between SVs, TEs and epigenomic features will be the subject of a separate manuscript.

Second review:

1. Fl-LTR TEs were not obtained using the PGSB library but using LTRharvest based on their methods.

Answer: This was a mistake in our response letter. Full length LTRs were of course detected *de novo* and not via the PGSB TE-library, as was correctly stated in the Methods section.

2. Distribution of LTR insertions could not be found in Suppl. Table 1. Maybe they meant Suppl. Fig. 1?

Answer: We apologize for the wrong reference. The results of our fl-LTR annotation are shown in Supplementary Fig. 1 and Supplementary Fig. 2.

3. The similar number of fl-LTR TEs in Suppl. Table 2 does not support the lack of novel LTR families in the barley pan-genome. Novel TE families are defined based on their sequences being novel compared to other genomes, not the number of sequences.

Answer: The reviewer raises the question as to the detection of novel TE families in barley in general and in the 76 sequenced genomes in particular.

There are currently 247 TE families from barley in TREP, similar to 278 from wheat. The top 20 TE families contribute >60% of the genome, the rest is contributed by the remaining ~230 families. We

acknowledge that ~10% of the barley genome remains un-annotated and previous studies from Triticeae indicate that they are probably derived from low-copy, yet un-characterized TE families.

However, it was shown in previous studies that within a species, a given TE family may vary in copy numbers between haplotypes (e.g. Walkowiak et al. 2020 doi: 10.1038/s41586-020-2961-x), but it has never been found that a given haplotype contains completely "private" TE families. Neither is this expected because there is constant gene flow between haplotypes within a species, which ensures that TE families readily spread across populations.

We cannot exclude the possibility that there are low-copy TE families that have not yet been discovered in the barley pangenome. Comparing the extensive TE libraries from wheat and barley shows that there are probably genus-specific TE families (for example, a given TE family may go extinct in one genus). However, we found no evidence that the genomes sequenced here differ in the presence or complete absence of any given TE family. Supplementary Fig. 1 and Supplementary Fig. 2 are meant to illustrate the overall TE composition of the 76 genomes is very similar. They support our point that the novel TEs play only a minor role.

It would go beyond the scope of this study to analyze all un-annotated portions of the 76 genomes in search of novel TE families. We are also believe that it would not advance our main objective to characterise important differences between the genomes.

4. The claim "TEs happen not to play a role" in complex loci was based on analyses reported in L1207 – 1215, in which they split complex loci into 200bp sequences and blast against the TE library to count for accumulated TE length, then compared to that of a distant region for TE length differences. They made a similar mistake here, that the similar TE length over a large region (1Mb) does not support the lack of effect from TEs. Usually, only one TE fragment is involved in the regulatory effect of the target gene, which would not make a difference in cumulated TE length over 1Mb. They should present haplotypes of complex loci with both gene and TE annotations in the barley pangenome. In fact, they presented one such haplotype comparison in Figure 4c and showed many structural variants, suggesting TEs may play a role in cis-regulatory functions.

Answer: We apologize for our clumsy phrasing when we wrote in the responses to the first round of reviews that we "report links between structural variation (*HvTB1*, *srh1*, *amy1_1*) and phenotypes, but TEs happen not to play a role in these cases." "These cases" referred only to the three

mentioned loci, *HvTB1*, *amy1_1* and *srh1*. Our assertions may have been too rash as indeed we cannot rule out that the action of TEs might be the ultimate cause for genic copy number variation at *HvTB1* and *amy1_1* and the deletion of a regulatory motif at *srh1*. However, the more likely mechanism in these cases is unequal recombination between neighbouring tandem repeats.

We did not wish to claim that TE do not play a role in the evolution of the 172 complex loci we discovered, and indeed we are convinced that TEs are a major driving force of barley evolution. Nevertheless, the reviewer raises two important points: (i) specific TE insertion may drive functional innovation, and (ii) purely quantitative differences in TE composition may not reflect functional diversity.

As to the first point: TE insertions into (or excisions out of) promoters can indeed modify expression states of genes. Particularly, in gene clusters such as the ones described in complex

loci, for example, a new TE insertion in the promoter of a single gene can indeed give rise to an important new expression variant. This particular gene/promoter copy may then be preferentially maintained or duplicated. To address this comment, we aligned promoter sequences of genes in complex loci across the 76 genomes. Despite the loci containing very young duplications, we found 6 instances where some gene promoters contain TE presence/absence polymorphisms. However, we did not find evidence that these were preferentially amplified or selected against in the 76 genomes. But because we are looking here at products of very recent evolution, a much broader dataset would be required to arrive at conclusions with good statistical support. We added a short section in the main text describing that TE polymorphisms may be a source of sequence diversity of individual gene copies:

> *Despite high similarity of duplicated segments, TE insertions (or excisions), random deletions and mutations contribute to diversification or pseudogenization of individual gene copies over time (**Fig. 3a**, **Supplementary Fig. 9a**)*

We also added a new Supplementary Figure 9 where we show examples of polymorphic TEs and deletions in promoters of individual gene copies.

As to the second point. We replaced the quantitative annotation using only 200 bp fragments. Instead, we used a novel complete genome-wide TE annotation to compare abundance of TE superfamilies in complex loci with that of neighbouring regions and the genome overall. As in the previous (purely quantitative) analysis we found that TE content in complex loci varies broadly, with some loci having lower and some having higher TE contents than the surrounding chromosome

segments (or the genome overall). We again did not find a pattern that distinguishes complex loci from their chromosomal neighborhood or from the rest of the geome. We have added this information in the revised Extended Data Fig. 6. We also more clearly emphasize our conclusion that the main driver of CNV in complex loci is unequal recombination between tandem repeat units:

> *The latter process might play a role in their amplification and contraction owing to unequal homologous recombination between neighboring repeat units[29] (**Extended Data Fig. 6a**). We found no association of complex loci with specific TE types (**Extended Data Fig. 6b-d**). Instead, molecular dating of the tandem duplications in Morex is consistent with recent and recurring duplication/contraction cycles, leading to complex pattens of higher and lower order tandem repeats (**Extended Data Fig. 7**). Indeed, many gene copies appear to have been gained within the last three million years (**Extended Data Fig. 7c**), after the* H. vulgare *lineage split from that of its closest relative* H. bulbosum*.*

(6) It might help to indicate the sizes of your four selected loci, the # of tandem genes, repeat composition, and HiFi sequence coverage and length over those loci. The article takes a deep look at those four regions, which seems to necessitate a detailed assessment that your sequencing platform and coverage (20x HiFi) could resolve them. Suppl. Figures S7 and S8 try to address this but not entirely. S8 could be improved by including read depth at non-complex loci for comparison.

Answer: The read depth at non-complex loci was used for normalization in Supplementary Fig. 8 so it is implicitly shown. We have modified the legend of this figure for clarification:

> **Read depth at complex loci.** Each cell in the heatmap shows the average per-bp read depth *at one complex locus and one pangenome accession. The distribution of read depth is*

centered around the genomic median, i.e. the per-bp median coverage across the entire genome.

(7) Does projecting gene models influence your analysis of cloud/core/shell and HOGs?

Answer: Gene projections are a powerful way to comparatively assess the gene content of different genotypes in a pan-genome, especially for those genotypes where native expression data is not available. The inherent limitation of our gene projection approach is it would not be able to pick up completely new genes and gene structures, not present in any of the 20 de-novo annotated genotypes. While these cases are very rare, our gene projections are able to call and resolve copy number variations (such as tandem copied genes) as well as diverged gene models due to the mapping strategy applied (different rounds of mapping including multiple and low stringency mappings considered). For the analysis of cloud/core/shell and HOGs this translates into the following conclusions:

- Our study potentially underestimates the cloud gene portion, in case a significant number of true genotype-specific genes without homology to genes in other genotypes would be present in the projected barley genotypes. We expect these cases to be rare. Van Oss and Carvunis 2019 (doi: 10.1371/journal.pgen.1008160) estimated the rate of de novo gene birth at 11.6 genes per 1 million years. This means that, although *de novo* genes are by and in themselves interesting, their undercalling would not greatly influence our estimates of the relative sizes of the cloud, shell or core compartments.
- Our gene projection approach is useful even when native transcriptome data is available. It was applied to consolidate gene models across the 76 pangenome genotypes via an all-against-all mapping. Thereby, we re-annotated and rectified gene models that were missed or mis-annotated in the initial de novo annotations (a common problem found in all available gene prediction approaches and pipelines). The consolidation strategy also helps to call presence-absence and copy-number differences (and core-/shell-genome portions consequently) between the pan-genome genotypes more reliably as would be possible by full de-novo annotations without further consolidation.

Minor:

(1) cvs (cultivars) is an unspecified acronym.
(2) L86: 'in of' - typo
(3) L121: of the 20 de novo annotated accession, please state how many were wild and domesticated. Please also state what tissues were used.

Answer: We have implemented these corrections and suggestions.

(4) It's confusing to have both Supplementary Figures and Supplementary Data Figures.

Answer: It is Nature editorial policy to have both Extended Data items and Supplementary Figures and Tables. Both types of display items have different formatting requirements and there is limit to the number of Extended Data items (10).

(5) In the rebuttal, "We have also added Supplementary Figure 4, which shows the Hi-C contact matrices for the translocated genotypes and alignments of their pseudomolecules to non-translocated Morex. A clear inter-chromosomal Hi-C signal is seen if the Morex karyotype is used

as a reference for aligning Hi-C reads but absent if the respective translocated karyotype is used." But the legend of that fig says "Hi-C contact matrices before correction (i.e. using the chromosomal configuration of Morex). (c) Hi-C contact matrices after correction" It's unsure whether the karyotype was translocated or misassembled based on the description discrepancy. It looks like the comment in the rebuttal better describes the figure.

Answer: We have revised the legend of Supplementary Figure 4 to read:

> ***Translocations in HID055 and HOR 14273. (a)*** *Alignments of the final pseudomolecules to Morex.* ***(b)*** *Hi-C contact matrices using the chromosomal configuration of Morex.* ***(c)*** *Hi-C contact matrices using the chromosomal configuration of the respective native pseudomolecules. A clear interchromosomal signal is seen when the Morex reference is used, indicative of translocation relative to that genotype. If the native pseudomolecules are used no such off-diagonal signal is seen, supporting their structural integrity. Left column – HOR 14273; right column – HID055.*

Our use of the word "corrections" may have caused some confusion. The "corrections" were done as part of the usual manual curation during Hi-C based pseudomolecule construction as described in our technical guide to our TRITEX assembly pipeline (Marone et al. 2022, doi: 10.1186/s13007-022-00964-1). The corrections were not done during the first revision stage in response to the referee's comments, but during the original work that resulted in the pseudomolecules of HID055 and HOR 14273. The new phrasing provides greater clarity in this respect.

Referee #1 (Remarks on code availability):

Links to the code are provided and could be open. However, I did not tested the code myself.

Referee #2 (Remarks to the Author):

I was reviewer #2 in the initial round of reviews.

I find the study to be mostly unchanged. As an example, I was disappointed that the authors did not take seriously my criticism that the work did not speak to the processes of adaptation. Furthermore, there is still nothing in the manuscript that supports claims of "rapid evolution", as prominently mentioned in the abstract. The authors state that they investigate "structurally complex loci that have become hot spots of gene copy number variation in evolutionarily recent times". This seems to imply that mutational patterns at these loci have recently changed. If this were indeed the case, this would be an amazing finding, but sadly this hypothesis is not tested in the manuscript.

Answer: We have removed references to "rapid evolution" and "evolutionary recent times" from the abstract and discussion. We agree with the reviewer that further research is required to study the mutational patterns at complex loci. However, this is going beyond the scope of the present study.

There is a long section on copy number variation at the alpha-amylase locus. The authors state "Copy numbers of amy1_1 (had) on average more copies in domesticated than in wild forms" but there is no statistical test, and the differences in any case seem small. More importantly, how does copy number variation in the material examined correlate with amylase activity?

Answer: We thank the reviewer for pointing us to this inconsistency. We have rephrased the sentence "Copy numbers of *amy1_1* had on average more copies in domesticated than in wild forms" to "We found between two and eight copies of amy1_1 in the 76 complete genomes, with substantial variation in both wild and domesticated forms".

We have also clarified our rationale for studying alpha amylases and emphasize its industrial applications:

> *In both wild and cultivated forms, the speed and efficiency of that process determines the energy supply to and hence the vigor and survival of the young seedling when competing for sunlight and nutrient. In grains of domesticated barley, the enzymatic conversion of starch into fermentable sugars by α-amylases is a crucial step in the malting and brewing processes.*

A cursory perusal of the literature reveals that in vitro alpha-amylase activity varies considerably, but in the current work we learn nothing about the underlying reasons, specifically whether the known variation is due to differences in protein amounts as a consequence of gene copy number variation, due to differences in amino acid sequences and hence specific protein activity, or due to differences in promoter sequences.

Answer: We agree that a full-fledged proteomics study of alpha-amylase activity in light of our new findings on structural genome diversity would be valuable, but this is beyond the scope of the current study.

We have analysed transcriptomics data from micro-malted samples as a first, rough approximation to protein abundance and activity. A description of the results has been added to the manuscript:

> We confirmed lower *amy1_1* transcript abundance in micro-malted RGT Planet compared to a near-isogenic line (NIL) that carried the Barke *amy1_1* haplotype in the genomic

background of RGT Planet (**Supplementary Fig. 11**). The final end-use relevant α-amylase activity of a malted barley grain is the combination of its *amy1_1* copy number, transcription and individual protein haplotype activity.

Also, why is there such enormous variation at all levels, rather than just expansion/contraction of very similar copies? I asked specifically what we learn about common versus independent copy number expansions/retractions, but this was also ignored by the authors in their answer.

Answer: The reviewer raises an interesting point as to why we do not see simple

expansion/contraction of very similar copies. Our analysis shows that the complex loci experience duplications over long periods of time and in different parts of the complex loci. This can lead to the evolution of "higher" and "lower order" repeats (e.g. a part of a complex locus may be duplicated, giving rise to a sub-cluster of tandem repeats inside a locus). This is an important mechanism how sequence diversity can evolve even within a single locus. We now emphasize this point in the main text:

> *We found no association of complex loci with specific TE types (**Extended Data Fig. 6b-d**). Instead, molecular dating of the tandem duplications in Morex is consistent with recent and recurring duplication/contraction cycles, leading to complex pattens of higher and lower order tandem repeats (**Extended Data Fig. 7**). Indeed, many gene copies appear to have been gained within the last three million years (**Extended Data Fig. 7c**), after the* H. vulgare *lineage split from that of its closest relative* H. bulbosum.

I had a few technical concerns, and the answers to these did not convince either. I asked about the surprisingly small number of universally shared orthologs. The authors essentially state "well, that's how the program we used, OrthoFinder, works". They also did not bother to check whether this may have something to do with potentially problematic annotations.

Answer: We apologise for the incomplete response to the reviewers' initial concerns. The reasons for what the reviewer considers a paucity of core genes (universally shared orthologs) are as follows.

Our pan-genome consists of a relatively large number of genotypes from distinct gene pools including wild barleys, landraces and elite cultivars. To move an orthologous group into the shell-genome category one gene missing (or not correctly annotated) from one of the 76 genotypes is sufficient. In fact, we find a large number of genes in those groups of the shell genome that contain almost complete groups (present in 75 genotypes: 378,468 genes; present in 74 genotypes: 172,584 genes). This is visualized in Supplementary Fig. 3 (the two right-most green (shell genome) bars in panel a). In contrast to some other pan-genome studies we did not introduce an additional category sometimes referred to as "soft-core" (or similar terms) as we feel this was not required for any of the downstream analyses performed in this study.

We also wish to refer this reviewer to our response to reviewer #1, where we discuss the potential impacts of projecting genes (as done in our study for some of the genotypes) on the definition of the core-, shell- and cloud-genome compartments.

In summary, this resource paper is in several aspects technically impressive, but at the same time intellectually disappointing, as it breaks little new conceptual ground. As I had stated in my confidential comments to the editor, a manuscript of this ilk needs a firm editorial stance whether or not this is the type of paper they want to publish. The good news is that there are many competent

colleagues who will likely use this fantastic resource to give us more serious answers to questions about the evolutionary processes of mutation, selection, and adaptation that the (non) answers

provided by the authors. I note that the authors did not contest my assertion that "the advance for the broader community does not go beyond similar types of papers published elsewhere for tomato, rice and soybean two or three years ago, and it does not go nearly as far as the (graph) pangenome papers for tomato and potato published in Nature last year".

Despite my gripes, the work would not look out of place in the pages of Nature. Reviewer comments can only help so much in making editorial decisions.

Answer: We thank the reviewer for their candid appreciation of our manuscript. To assuage the reviewer's concern about a paucity of evolutionary insights, we have changed the title of the manuscript to "Structural variation in the pangenome of wild and domesticated barley."

Technical comments/suggestions for the authors to consider:

1. For the alpha-amylase cluster, the Pearson correlation coefficient is considerably lower than what was reported for humans (R=0.69 [R^2=0.4761] vs R=0.99, Vollger et al., Nature 2023). Did this perhaps come from the repetitive 21-mer introducing a bias or from partial assembly collapse?

Answer: We believe that the lower correlation mostly comes from varying sequencing depth rather than partial assembly collapse or repetitive 21-mers. We estimated copy number as the median k-mer count in the target gene region (normalized per Morex genome). Owing to varying sequencing depth along the genome, estimation of the median k-mer count in a small region will be less accurate than in a larger region. Due to high sequence similarity in alpha-amylase genes, we only selected a small conserved region (<100bp) specific to amy1_1 genes for k-mer generation. In contrast, Vollger et al. focused on larger regions (>1kb segmental duplications), which accordingly leads to better estimation. We did observe as a high correlation as Vollger et al. when estimating copy number for a larger gene (> 8000 bp) using the same dataset.

2. You seem to build individual graphs for each chromosome. One may lose much of the variation inside the common large translocation using this approach. Have you considered "correcting" the translocation for alignment purposes?

Answer: It is true that we found two large interchromosomal translocations but neither of them is common. Both were seen in only one individual each: the translocation between chromosomes 1H and 2H in the wild barley HID055 and that between 2H and 4H in the Iranian landrace HOR 14273. The reviewer is correct that our chromosome-wise graphs would not include sequences private to either of them on the respective chromosomes involved. Our two main uses of pangenome graphs were (i) to compare short-read alignments between graphs and a single linear reference and (ii) to study structural variation at the *Srh1* locus on chromosome 5H, which is not involved in either translocation. Neither use was affected by the chromosome-wise graph construction, so we decided to avoid either the substantial computational costs that whole-genome alignments between 76 would incur or the inelegant kludge of "correcting" true translocations.

We hope that the computational efficiency of graph genome tools will improve in the coming years so as to allow whole-genome instead of chromosome-by-chromosome alignments.

However, our optimism in this regard was tempered by a recent preprint on this topic by Igolkina et al. (doi: 10.1101/2024.05.30.596703), who concluded that "[i]n species [...], where [the] intergenic space is essentially unalignable even between relatively closely related agricultural varieties, the idea of representing a whole-genome alignment as a graph that captures all variation may be neither practicable nor useful."

3. BUSCO cannot distinguish between high copy number variation genes (such as alpha-amylase genes) and rapidly evolving genes (Mla). One could use the available RNA sequencing reads mapped back to the annotated gene models to support high completeness of pangenome annotation.

Answer: This is a good suggestion! We performed the proposed analysis for all genotypes with native transcriptome data. The percentage of reads mapped back to annotated genes was between 75% and 90% (average: 86 %).

4. Why is embryophyta_odb9 used for assembly BUSCO and Poales_odb10 for annotation?

Answer: We used embryophyta_odb9 to check for gene space completeness after completing the pseudomolecules and prior to annotation. These results served only for internal QC during an intermeditate step. The Poales_odb10 database was used for the annotated assemblies, which are the final results on gene space completeness. We have removed from Supplementary Table 1 and the Methods the reference to the initial BUSCO runs on the annotated assemblies to avoid confusion.

5. Fig. S8: How does read depth compare to non-complex regions?

Answer: These results are shown in Supplementary Fig. 8. The coverage values shown in each cell were normalized by dividing by the average (median) per-bp coverage (HiFi reads) of each genotype. We have revised the legend of Supplementary Fig. 8 to read:

> ***Read depth at complex loci.*** *Each cell in the heatmap shows the average per-bp read depth at one complex locus and one pangenome accession. The distribution of read depth is centered around the genomic median, i.e. the per-bp median coverage across the entire genome.*

Referee #3:

I thank the authors for the clear and detailed point-by-point answer to my remarks. I now understand the reasons behind things that could have been appeared questionable in the first version. I recognize the authors made great efforts to provide more results on core/shell/cloud genes and I found the paper has been significantly improved in quality both in term of results and writing.

I appreciated to read the novel section "Atlas of structural variations" with now interesting results on pan/core-genome and specificities of wild vs. cultivated etc., which was my main concern previously. In addition, the introduction now makes the link with previous initiatives and it looks clearer for people not fully aware of the recent papers on barley genomics. The Figures of the supplementary material are all readable in the revised version. The Github repository now contains the expected codes. Finally, I found this revised version much easier to read and follow.

Answer: We thank the reviewer for their encouraging words.

Reviewer Reports on the Second Revision:

Referee #1 (Remarks to the Author):

These revisions represent a great improvement in experimental rigor and methodological clarity. The authors have worked diligently to elevate this work to a higher level. The scale of the work is impressive and will be a great addition to the barley and plant pangenome communities. The authors have addressed most of my concerns, but a few points remain.

Detailed comments:

Mapping comparisons to linearized pangenome and the graph pangenome appear to be an improvement. It's a little unexpected that the linearized pangenome would have noticeably better mapping than a single-cultivar reference. By my understanding, a linearized pangenome ought not function as a pangenome, because all duplicate alleles have been collapsed into a single representative sequence, so its fundamental pangenome properties have been removed. Functionally, a linearized pangenome is a mosaic version of many single-cultivar references. I don't see methods describing how the linearized pangenome was prepared, but it's possible that the linearization processes retained shell and cloud seqs, which may explain better mapping. Please clarify.

The added FindIT data helps to improve clarity of this section of the paper. You chose E10, E14, E19 and E21. E10 was selected as a control and the others were selected because they have the mutant phenotype and are heterozygous. There are quite a few heterozygotes that do not show the mutant phenotype. You've dropped these and made the case that the missing phenotype is not due to an alternate genetic mechanism, but rather a methodological peculiarity. It seems risky to validate your mutation using the preferred mutant phenotype, and then subsequently making a connection between the phenotype and the genotype. Although there's some risk in your approach, the convergence of EMS and Cas9 methods seems to make a stronger case.

The genetic composition of the 4.3kb deletion is not resolved. Are they repetitive? Could they be found in sister species?

Codes for generating data and analyses are not shared. People studying pangenomics would appreciate having access to the codes of this study. Methods have very brief descriptions of what programs and parameters are used, but they are often not complete.

Minor comments:

The methods section of the manuscript still mentions the 200bp approach for TE quantification. Please

replace it with the latest method.

Suppl. Fig. S8 shows read depth is similar at complex loci as it is at non-complex loci, but I'm not sure what's the read depth at non-complex loci. Please add average read depth, total length, repeat composition at complex loci in Suppl. Table S8.

Please change the color scheme of Fig. 1c to deviate from the common color scheme of Hi-C maps. Methods for LD were described for 'Morex' x 'Barke', not HID055 x Barke.

L135: "at least one orthologous gene" By definition, one genome can contain only one orthologous gene for each gene. They may change "orthologous" to "homologous" to be accurate.

L251, they probably want to cite Extended Data Fig. 6c-e.

Extended Data Figure 10ce is mislabeled.

L396-398: Please mark the location of the putative enhancer on Fig 4d.

L744-45, the x-axis should represent the subfamily 1 or 2, but the meaning of subfamily 1 and 2 is probably convoluted for non-Mla experts. More figure captions should be added.

L751 – 766 The Figure 4 caption is not matching the figure.

Referee #1 (Remarks on code availability):

No code is shared in this study, which should not be the case.

Referee #2 (Remarks to the Author):

I was previously reviewer #2. Apart from a few technical questions, my main beef was that I felt the work was sold as something that it was not. It should stand on its merit as a resource for the community, and the resource aspect is now much clearer, and previous, not well supported claims regarding adaptation and evolutionary processes have been toned down.

In summary, I am satisfied with the authors' responses.

Author Rebuttals to Second Revision:

Referee comments:

Referee #1 (Remarks to the Author):

These revisions represent a great improvement in experimental rigor and methodological clarity. The authors have worked diligently to elevate this work to a higher level. The scale of the work is impressive and will be a great addition to the barley and plant pangenome communities. The authors have addressed most of my concerns, but a few points remain.

Reply: We are glad that our revisions improved the manuscript. We thank the reviewer for their encouragement.

Detailed comments:

1) Mapping comparisons to linearized pangenome and the graph pangenome appear to be an improvement. It's a little unexpected that the linearized pangenome would have noticeably better mapping than a single-cultivar reference. By my understanding, a linearized pangenome ought not function as a pangenome, because all duplicate alleles have been collapsed into a single representative sequence, so its fundamental pangenome properties have been removed. Functionally, a linearized pangenome is a mosaic version of many single-cultivar references. I don't see methods describing how the linearized pangenome was prepared, but it's possible that the linearization processes retained shell and cloud seqs, which may explain better mapping. Please clarify.

Reply: The minigraph algorithm used for graph construction is described on its github page at https://github.com/lh3/minigraph and is useful for explaining the concept of graph construction and subsequent linearisation. During graph construction, sequences of the input accessions are aligned to the initial reference sequence (here Morex) iteratively, and every novel sequence not currently present in the graph is added to the graph during this process until all accessions have been processed. Bubbles form in the graph where flanking sequences are identical but a stretch of sequence between these is novel.

As stated in the Methods section, the linearised pan-genome was derived from the graph using the gfatools gfa2fa command (https://github.com/lh3/gfatools). This tool essentially reverses the process of the graph construction and outputs in a FASTA file the original reference sequence (here Morex), plus all the additional sequences added to the graph from other accessions. "Duplicate alleles", i.e. graph nodes making up a bubble in the graph, are all extracted during this process. Therefore, the sequences in both the graph and FASTA file represent the entire pan-genome, including core, shell and cloud portions.

In terms of set theory, the graph/FASTA file represents the union of all the input genomes, with every sequence found in the input genomes represented in the graph/FASTA file non-redundantly. Thus, the graph/FASTA file offers a data structure for mapping of reads where the target space is maximised (within the limits of the germplasm chosen as input), and this leads to relatively greater mapping rates against the graph since a proportion of mapping targets will be unavailable in the single-accession linear reference sequence due to presence-absence variation.

2) The added FindIT data helps to improve clarity of this section of the paper. You chose E10, E14, E19 and E21. E10 was selected as a control and the others were selected because they have the mutant phenotype and are heterozygous. There are quite a few heterozygotes that do not show the mutant phenotype. You've dropped these and made the case that the missing phenotype is not due to an alternate genetic mechanism, but rather a methodological peculiarity. It seems risky to validate your mutation using the preferred mutant phenotype, and then subsequently making a connection between the phenotype and the genotype. Although there's some risk in your approach, the convergence of EMS and Cas9 methods seems to make a stronger case.

Reply: We are glad that the reviewer sees an improvement in the clarity of the descriptions of this section, however, we kindly like to point out that the FindIT information is not new and has not been newly added to this version but was already part of the first versions' contents. We appreciate that such details may not receive instant attention given the amount of data and information included in such a resources study. A tiny further detail: the FindIT population was not obtained after mutagenesis with EMS but by treating with sodium azide (Knudsen et al. 2022, DOI: 10.1126/sciadv.abq2266).

The reviewer #1's recurring criticism of the "risky" approach of "to validate your mutation using the preferred mutant phenotype" may be formally correct in a case of lack of any pre-existing knowledge. We still, however, disagree and we are convinced that our approach is not risky based on the solid cumulative evidence, which we have described in detail in our earlier reply to reviewer comments (please see our response to the comments to Version 1 of the manuscript and Supplementary Table 26). E10, E14, E19 and E21 are all regenerated plants from the same transformation experiment. It was validated that E10 is a non-transgenic WT regenerant. Therefore, E10 is serving as a proper WT control exhibiting the proper WT phenotype and WT alleles at both gRNA target sites also in 24 offspring. The other three regenerants E14, E19 and E21 were selected to be taken to the next generation based on the fact that they showed already in the T0/M1 generation consistently the mutant phenotype. This is a key observation and we want to stress here that the mutant phenotype of short rachilla hairs is the recessive trait, hence, a T0/M1 plant can only show the mutant phenotype if both chromosomal copies received a disruptive mutation, which by nature of the method is expected in most cases to be in hemizygous state (two different /independent non-functional alleles). This was consistently confirmed by the Sanger sequencing results in the offspring of all three families. The only T1/M2 plant with a WT phenotype (1/72!) was shown to carry the WT allele at the gRNA1a target sequence and a homozygous 9 bp in-frame deletion at gRNA1b target sequence, which obviously did not result in any phenotypic alteration. Together with the genetic resolution from GWAS and biparental mapping there is, therefore, no risk in our working hypothesis and our interpretation of the observed results.

3) The genetic composition of the 4.3kb deletion is not resolved. Are they repetitive? Could they be found in sister species?

Reply: We checked for repetitiveness based on k-mer frequencies and found that only the first 600 bp of the 4.3 kb interval deleted in Morex have elevated k-mer counts and are hence repetitive. The regulatory motif sits at the other end of the deletion. Apart from the first 600 bp, k-mer frequencies in the remaining sequence are consistent with it being single-copy.

We ran a BLAST search of the *Srh1* gene and its surrounding sequence against the bread wheat genome sequence (Chinese Spring RefSeq V2.1) and found hits in all three subgenomes (see the figure below). All three hits contained complete coding sequences, although only the A genome copy was properly annotated. The deleted interval was partially aligned as well with the regulatory motif present in the B and D genomes. We do not think that these results have a strong relation to the barley pangenome (i.e. main focus of this manuscript) and therefore will not include this figure to the revised manuscript.

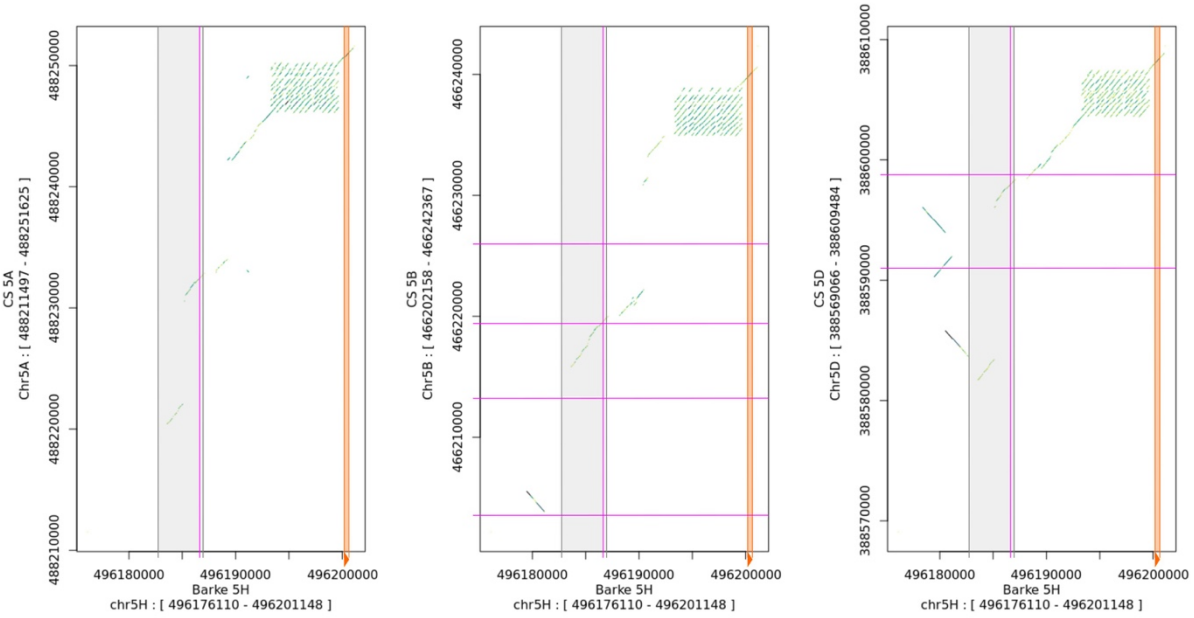

Figure: Alignment of the *Srh1* gene and its surrounding sequence to the three subgenomes of bread wheat. Orange shading marks the position of the *Srh1* gene. The deleted segment in Morex relative to Barke is shaded gray. Occurrences of the putative regulatory motif in barley and wheat are marked by purple lines.

4) Codes for generating data and analyses are not shared. People studying pangenomics would appreciate having access to the codes of this study. Methods have very brief descriptions of what programs and parameters are used, but they are often not complete.

Reply: Links to code repositories are given in the section Code Availability at the end of the Methods section:

> *Scripts for pangenome graph analyses are available at https://github.com/mb47/minigraph-barley. The scripts for the definition of core/shell and cloud genes are deposited to the repository https://github.com/PGSB-HMGU/BPGv2. Scripts used for gene projection are available from https://github.com/GeorgHaberer/gene_projection/tree/main/panhordeum. The pipeline for identifying structurally complex loci is available at https://github.com/mtrw/DGS. The pipeline for the construction of the single-copy pangenome is available from*

https://bitbucket.org/ipk_dg_public/barley_pangenome, that for heterozygosity estimation from https://bitbucket.org/ipkdg/het_estimation.

We confirm that all links are working.

Minor comments:

5) The methods section of the manuscript still mentions the 200bp approach for TE quantification. Please replace it with the latest method.

Reply: We have removed this section.

6) Suppl. Fig. S8 shows read depth is similar at complex loci as it is at non-complex loci, but I'm not sure what's the read depth at non-complex loci. Please add average read depth, total length, repeat composition at complex loci in Suppl. Table S8.

Reply: We added a column with TE content as well as start and end positions to Supplementary Table 8. Information about coverage in all 76 assemblies (after normalization for average read depth along the genome) is shown in Supplementary Fig. 8. Coverage information cannot easily be condensed into a single "average read depth" owing to the differences in average genome-wide coverage (i.e. the amount of HiFi raw data) per accession.

When reviewing this table, we noticed that in two genomic regions (on chromosomes 1H and 6H), three complex loci were redundantly mapped (i.e. covered by other complex loci). This was due to multiple types of repeated sequences which are nested within one another.

These three redundant loci were removed from the analysis. Supplementary Table 8, Figures 2c, d and Extended Data Figures 7c, d were revised accordingly. The minor changes induced thereby had no effect on our biological conclusions.

7) Please change the color scheme of Fig. 1c to deviate from the common color scheme of Hi-C maps. Methods for LD were described for 'Morex' x 'Barke', not HID055 x Barke.

Reply: We have changed the color scheme of the LD plot in Fig. 1c to yellow-red. We have added this paragraph to the Online Methods:

LD in the Barke x HID055 population

LD between each pair of SNPs (both intrachromosomal and interchromosomal) was calculated as the squared Pearson product-moment correlation between the quantitative IBD matrix scores presented in Additional File 1 of Maurer, et al.[21] *(https://datadryad.org/stash/dataset/doi:10.5061/dryad.36rm1). The LD plot was created with SAS PROC TEMPLATE and SGRENDER (SAS Institute Inc., Cary, NC, USA) on the genetic map of Maurer, et al.*[20]

8) L135: “at least one orthologous gene“ By definition, one genome can contain only one orthologous gene for each gene. They may change “orthologous” to “homologous” to be accurate.

Reply: We agree with the reviewer and have corrected “orthologous” to “homologous”.

L251, they probably want to cite Extended Data Fig. 6c-e.

Reply: We have corrected this.

Extended Data Figure 10ce is mislabeled.

Reply: We have corrected this.

L396-398: Please mark the location of the putative enhancer on Fig 4d.

Reply: Done.

L744-45, the x-axis should represent the subfamily 1 or 2, but the meaning of subfamily 1 and 2 is probably convoluted for non-Mla experts. More figure captions should be added.

Reply: We have expanded the legend of Figure 2a to explain the colour code for the *Mla* alleles and give more information on subfamily 2.

> *Presence/absence of known Mla alleles in the barley pangenome. Black and white squares denote presence and absence, respectively. The names of Mla alleles (y-axis) and genotypes (x-axis) are coloured according to, respectively, subfamily (red – 1 or black – 2, ref. [25]) and domestication status (green – domesticated, orange – wild). Only the genomes containing known alleles are displayed. Owing to higher SNP numbers and truncations[25], members of subfamily 2 are expected to be inactive forms.*

L751 – 766 The Figure 4 caption is not matching the figure.

Reply: The specified line numbers are not related to Figure 4 or its legend (they are part of the legend of Figure 2). We checked Figure 4 and its legend. They appear fine to us.

Referee #1 (Remarks on code availability):

No code is shared in this study, which should not be the case.

Reply: Links to code repositories are given in the section Code Availability at the ends of the methods section:

Scripts for pangenome graph analyses are available at https://github.com/mb47/minigraph-barley. The scripts for the definition of core/shell and cloud genes are deposited to the repository https://github.com/PGSB-HMGU/BPGv2. Scripts used for gene projection are available from https://github.com/GeorgHaberer/gene_projection/tree/main/panhordeum. The pipeline for identifying structurally complex loci is available at https://github.com/mtrw/DGS. The pipeline for the construction of the single-copy pangenome is available from https://bitbucket.org/ipk_dg_public/barley_pangenome, that for heterozygosity estimation from https://bitbucket.org/ipkdg/het_estimation.

We confirm that all links are working.

Referee #2 (Remarks to the Author):

I was previously reviewer #2. Apart from a few technical questions, my main beef was that I felt the work was sold as something that it was not. It should stand on its merit as a resource for the community, and the resource aspect is now much clearer, and previous, not well supported claims regarding adaptation and evolutionary processes have been toned down.

In summary, I am satisfied with the authors' responses.

Reply: We thank the reviewer for their positive assessment of the revised manuscript.

Reviewer Reports on the Third Revision:

Referee #1 (Remarks to the Author):

My comments have been addressed.