nature portfolio

Peer Review File

Genome Evolution and Diversity of Wild and Cultivated Rice Species

Corresponding Author: Dr Weixiong Long

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
This manuscript entitled “Genome Evolution and Diversity of Wild and Cultivated Rice Species” reported thirteen representative wild rice species from the Oryza genus and constructed a super pangenome by integrating them with four previously reported genomes in this genus. It discussed extensive topics including genomic and pangenomic features, phylogenetics, structural variations and variable genes, TEs and other functional elements, which would be much valuable for rice researchers and breeders with amount new resource to improve cultivated rice and utilize wild rice germplasm. However, there are amounts of issues in the results and methods sections, which makes me very concerned about the reliability of these results.

major points:
1. For the sampling of different wild rice species, why a BBCC type of Oryza punctata was selected? Usually, a typical O. punctata sample is considered as an BB type, although there exist other types.

2. Although the authors constructed the phylogenetic tree of Oryza genus by distinguishing different genome types, the introgression and gene flow between different genomes may still cause strong interference in the phylogenetic relationships, which should be taken into account when constructing the phylogenetic tree.

3. The obtained phylogenetic relationships in this study is obviously conflicted with previous reported results. For the tree of nuclear genome in Fig 1b, O. glaberrima is clustered with O. sativa ssp. indica in the same clade while the two subspecies of O. sativa, O. sativa ssp. indica and O. sativa ssp. japonica are clustered in different clades. Moreover, the O. rufipogon that is commonly known as wild relative species of O. sativa, is placed as an outgroup of both O. glaberrima and O. sativa. This totally contradicts the consensus on the origin and evolution of Asia and African cultivated rice over the past half century.

Thus, the supporting rate for different topologies of species tree could be invalid and the authors should check about their methods related to this part. Additionally, two supporting rates (44% and 46) very close to 50% on branch 2 suggest that the phylogenetic relationship here is not resolved, and also suggest a potential risk in the calculation method.

For the tree of chloroplast genome in supplemental figure 3, it is also conflicted with previous reported results, e.g., the O. punctata should be the closest lineage to the AA-genome species. And two important species, O. rufipogon and O. glaberrima, are absent in this tree.

4. The large inversion in Chr6 is detected in both O. sativa ssp. japonica and O. rufipogon as revealed by Fig 2a. As we know, the O. rufipogon is mainly distributed in South and Southeast Asia. The japonica rice consists of tropical japonica rice, subtropical japonica rice and temperate japonica rice, etc. But the Lines 223-225 described that “The inversion occurred only among the common wild rice and Oryza sativa japonica with high latitude distribution, indicating that it has risen to expansion of geographical range.” This causal relationship lacks evidence and may even be incorrect. It also affects the significance and necessity of the subsequent analysis on the gene OsMFT1.

5. Fig 3a presented the function genes with CNV mutations, however, all genes showed more than two copies in the Nipponbare genome but present-or-absent variations in other species. If this gene dataset can be easily obtained by selecting genes with multiple copies from Nipponbare genome, what is the significance of constructing a super pangenome?

The authors should recollect a more meaningful gene dataset to conduct an in-depth analysis on the CNV of important genes across the wild and cultivated rice species, highlighting the advantage of this super pangenome.

Furthermore, the Fig 3j showed a more similar collinearity of NLR genes between O. sativa ssp. indica and O. sativa ssp. japonica than that between O. sativa ssp. indica and O. glaberrima, indicating a closer phylogenetic relationship between the two subspecies of O. sativa than that between indica and O. glaberrima. This result is conflict with the results of phylogenetic analysis in this study but is consistent with the current consensus on the origin and evolution of Asia and African cultivated rice.

Additionally, the authors clarified that “Analysis of NLR distribution showed that while R gene singletons were similar between wild and cultivated rice, cultivated rice tended to have a higher number of R genes in pairs or clusters compared to wild rice” at Lines 390-392. However, considering that most of the existing rice NLR gene data sets are based on studies on cultivated rice genome, this may also be the case that NLR in wild rice genomes has not been fully identified.

Minor points:
1. Line 64: lack of species name of Asian cultivated rice, Oryza sativa.

2. The Latin names of indica and japonica rice are not properly written, e.g. should be Oryza sativa ssp. indica and Oryza sativa ssp. japonica. Please check through the manuscript and all the figures and tables.

3. Line 97: the reference 9 (Qin et al., 2021) was used for Nipponbare genome, but actually this study used the IRGSP-1.0 (Kawahara, Y. et al., 2013) as shown in the M&M section.

4. Extended Data Fig. 3c, the second member of dispensable family showed a presence in all types of genomes, which is identical with the core family.

5. The chromosomes of different species were drawn with lengths not consistent with their genome size, e.g., in extended Data Fig. 6 and 7, supplementary 7 and 10.

6. The dotplot of pairwise alignment for Nipponbare genome and other genomes were drawn with a reverse orientation and should be corrected.

Reviewer #2

(Remarks to the Author)
Sequencing the genomes of crop wild relatives is very important for identifying novel and adaptive variation, especially in the light of combating climate change. The paper by Long et al. is an excellent example of the types of study that should be undertaken. I have some suggestions for improvements, some I would suggest are vital and others are smaller or optional, as follows:

Major changes:
1. Figure legends need significant improvement. For example, for fig1:
“structure, panicle” > “structure of the panicle”. Also the panicle panel doesn’t have the species names indicated. If they are in the same order as the whole plant photos then say this.
“TE content in each subgenome/genome of the wild and cultivated rice”. > what about it? – say it is indicated on the right of the phylogeny
“The order of species is corresponding to the phylogeny tree”. > use “phylogenetic tree” or “phylogeny” and I don’t understand why you need to say this statement, the species are clearly in phylogeneticv order, that’s the point of a phylogeny.
“Proportions of contrasting gene tree topologies for the 100 genes with regard to two major conflicting relationships”. > indicate these as panels c and d.
All figure legends need work.
2. Present the BUSCO analysis for each sample, not just the average. As allotetraploids were involved it would be very interesting and important to be able to see the BUSCOs.
3. The orthofinder analysis would give very different results if different settings were used. How were the settings chosen selected? The very high number of orthogroups (101k) from comparing many genome but each only containing 30k genes suggests a very large number are single copy and species-specific, which is what is reported. However I doubt it is true that 1/3 of genes are truly species-specific, and certainly relaxing the settings would reduce the 101k and reduce the number of species specific. So simply stating there are 101k orthogroups is premature and not validated. Only 10% of genes being core genes suggest to me the settings were too strict.
4. The section on TEs (from line 154) is very descriptive – I suggest adding some statistics to back up the statements, eg you say there is an increase or decrease in some TE family sixed but is this just random variation across the phylogeny or are there significant increases in some species/groups relative to others?
In the same section, at line 178 the authors state “they influenced a certain proportion of genes expression” but I don’t see evidence for this. Extended data Fig 5c, Supplementary Fig. 7b do not have anything to do with expression.
Also, line 181 says TEs are related to centromeres, but in the figure the centromeres are not indicated so how is there evidence for this statement?
5. The section on synteny is unusual because everything is compared to NIP. So therefore statements such as “The AA

genomic type genome displayed a significant level of chromosomal conservation." is biased –of course it will show synteny because its compared to NIP which is AA too. This statement is not backed up unless B is compared to B, C to C and D to D etc. Related, "the gene synteny findings provide support for the phylogenetic position of the rice genus (Supplementary Fig. 10)." – this figure does not tell me anything about the phylogeny. It makes sense that related species have similar genomes, but unless this is specifically being tested, then remove this statement.
6. lines 262 – 274 – without any stats this can just be random variation between species. This must be added to allow for these statements to be made. Line 404, again without stats this can be random variation. Stats must be added.
7. There are several analyses that are poorly explained or, in my opinion, irrelevant, for example:
Line 342-7 – this seems like any genome region could have the same properties – this statement simply says that the different species have different SVs. But nothing about selection during domestication. In addition, it is already known that Rc controls pericarp colour, so what is the need to focus on it here? > this analysis doesn't seem to be meaningful and is not well-explained.
L382 – "This suggests that the immune system in wild rice has a more diverse evolutionary history compared to cultivated rice." – the comparison is two cv and many wilds, it is therefore not surprising there is this difference. Also the cv rice is diploid and the wild contains polyploid therefore at least double the number of genes is expected in punctata. > this analysis is not meaningful.

Smaller comments:
Line 39 – expansion of range in which species? Osat? or both Osat and Orufi?
Line 43 – I don't agree with the term 'hidden legacy' and I think it should be removed here.
Line 52 – state these are genome groups, because as written it is discussed like 6+5 species (but 20 species). Same for lines 61-2 where you say 11 species and list six genomes.
You say there is a lack of genome for BB but you didn't sequence a BB genome species according to Supp table 1, correct?
The introduction appears to brief in my opinion. Some more about what each of the species being sequenced brings would be appropriate, eg "species X (BB genome) is tolerant of X" and so on. Just a couple of sentences would help.
Several of the figures state the species name but a lot of the text refers to genome groups. It would be good for genome group to be in parentheses on some figures, especially figure 1.
Line 89 – how does supp figure 1 relate to this statement about the quality of the genomes?
93-4 – 37000 is not higher than 41000 – please reword this sentence.
Line 98 – "contradicted" in what way? Where are the contradictions? Why were 1000 of the 3555 used in the analysis not all 3555? This refers to fig 1b but see comment above about naming panels c and d.
Line 102 – state what was found too, don't leave it to the reader to download the supp figure. Does this match previous work on the cpDNA genome?
Lines 123-8 – it is confusing to sometimes use the number and sometimes the percentage. Please clarify this.
Line 131 remove "approximately" this number is not an approximation. In this section it would be interesting to see what percentage of core, disp, and sp-sp genes are supported by transcriptome data.
136 – you mention length then Ka/Ks. This is confusing.
140 – remove "to adapt to diverse environments" this is complete speculation.
150 – what do you mean "except for"?
156 – "Scientists have been intrigued by" is not a good phrase. Remove
172 – "puffy"? this is definitely not the correct word.
224 – "it has risen to expansion of geographical range" is pure speculation and not backed up by any data, only a correlation. Genetic drift is equally plausible.
276 – how does "substantial variation in SV sizes" indicate "an enrichment of SVs in repetitive DNA regions"?
Line 288 – "This gene might be de novo birth to contribute to the ability of wild rice to adapt to poor and problem soil in Sri Lanka." – this statement is confusing but also completely speculative. What does the analysis indicate the function of this gene is? Is it supported by expression data?
Line 290 – "The phylogenetic results revealed that the gene originated from O. eichingeri and then transferred to O. punctata, providing evidence that O. eichingeri was the progenitor of tetraploid O. punctata, consistent with our chloroplast evolution results." > where is the evidence it was transferred to punctata? Isn't this simply because punc couldn't have evolved without eich? How does it back up the cpDNA results, these are completely different things. The parents are shown by the nuclear genome and the female parent by the cpDNA. The cpDNA has nothing to do with determining the parents. Surely the parents were already known too prior to this analysis. This could just be removed, it doesn't mean anything.
L296 – 301 – picking one gene out of a QTL region is again entirely speculation that it is involved in resistance. What is the evidence that this gene is the likely causative gene? How many genes were in this region? The gene structure looks very different between sat and off, is this really an orthologue?
L302-6 – I have no idea what is the relevance of this.
L311 – what are "whole-genome alleles"?
L315 – "decreased, ranging from 18,463 to 23,812," – this sounds like an increase. Please reword.
L333-7 – I don't understand this. It sounds like the genes in collinear regions were translated and their domains compared. But then 7 clusters were found. Surely there are more than 7 clusters of proteins. Do you mean 7 members per cluster? Ext Fig 11d does nothing to help me understand.
Line 342-7 – this seems like any genome region could have the same properties – this statement simply says that the different species have different SVs. But nothing about selection during domestication. In addition, it is already known that Rc controls pericarp colour, so what is the need to focus on it here?
L382 – "This suggests that the immune system in wild rice has a more diverse evolutionary history compared to cultivated rice." – the comparison is two cv and many wilds, it is therefore not surprising there is this difference. Also the cv rice is diploid and the wild contains polyploid therefore at least double the number of genes is expected in punctata.
Nip has been sequenced before. How does your Nip genome compare to the previously sequenced Nip?

Reviewer #3

(Remarks to the Author)
The paper describes the generation of reference-level genomes of 13 wild rice accessions (including BB, BBCC, CC, EE, FF, and GG) using HiFi technology. The genome assemblies are generally of high quality and will be an important resource for utilizing useful alleles in wild rice gene pools. However, there are several major concerns about the work, which are listed below:

1. RNA sampling of diverse tissues is critical for genome annotation. Currently, the data seem to not have been comprehensively collected and are not well utilized.
2. Most of the findings, including allelic and regulatory element variations, are based purely on sequences, and there is a lack of genetic evidence using intercross populations and functional validation through transgenic experiments. Without substantial results, the conclusions should be provided more carefully.
3. The investigation of genomic heterozygosity, phasing, and intraspecific variation is completely lacking.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I am very glad to see that the current manuscript has been largely improved. Most of my comments on the last version of the manuscript have been well addressed. However, there are still some small issues should be fixed.

1. The authors claimed that “The NLR identification in this study was used the combination of denovo annotation of the genomes and RGAugry software” as revealed in their response to Reviewer #1. But I don’t find such description under the subtitle of “The analysis of pan-NLRome” in the M&M section. The de novo annotation of NLR genes should be updated with details in the M&M section, accordingly.

2. The authors claimed that they used the Nipponbare genome (Qin et al., 2021) as the reference. However, Qin et al. 2021 did not assemble the genome for Nipponbare, instead, they just used the IRGSP-1.0 (Kawahara, Y. et al., 2013), which is a high-quality assembly of Nippponbare genome, for their analysis. Thus, it is not suitable to cite Qin et al., 2021 here. This also suggests that the authors may lack sufficient investigation on the rice genomics. Please check throughout the manuscript to avoid such common-sense mistakes.

3. Line 61, “A total of 20045 gene families”. The thousandth separator is missing here.

4. The Latin names of some species are inconsistently written, e.g., the abbreviation of Oryza, O., was used in Lines 678-680 before the use of its full name in Lines 683-685 (the caption of Fig. 4). Please check throughout the manuscript to revise such typesetting errors.

Reviewer #2

(Remarks to the Author)
Based on the response to reviewers I can see the authors have made the revisions I requested in the first version. A cursory look over some of the other edits however finds a few errors and as such I think a small revision is still required:

1. Oryza sataiva spelled incorrectly in Supp Figure 4. Why is diploid punctata not included in Supp Fig 4 and only in the response to reviewers? Also needs bootstrap values adding.
2. A small point but for accuracy use the same number of decimal places for a set of data, for example in table R2 (Supp Table 1) all values in a column should have the same number of dp (e.g. 93.0 not 93).
3. Is table R5 in the main ms or only in the response. It seems important to include, but it is currently very confusing. For example “syntenic regions” and “inversions” – compared to what? Similarly “SNP” – what about them? Again if its compared to NIP how useful is this, especially for the very distantly related species? See also point above about being consistent in the number of dp among individuals in a column.

Reviewer #3

(Remarks to the Author)
The manuscript has been enhanced based on the reviewer's feedback. One additional recommendation: include the accession number for each RNA-seq dataset in Supplementary Data 3, and ensure that all genomic (raw reads and genome assemblies) and RNA-seq data are publicly accessible and thoroughly described.

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The current manuscript has been further improved and all my comments on the last version of the manuscript have been well addressed.

Reviewer #2

(Remarks to the Author)
The authors have made the small edits I suggested.

Reviewer #1 (Remarks to the Author):

This manuscript entitled “Genome Evolution and Diversity of Wild and Cultivated Rice Species” reported thirteen representative wild rice species from the Oryza genus and constructed a super pangenome by integrating them with four previously reported genomes in this genus. It discussed extensive topics including genomic and pangenomic features, phylogenetics, structural variations and variable genes, TEs and other functional elements, which would be much valuable for rice researchers and breeders with amount new resource to improve cultivated rice and utilize wild rice germplasm. However, there are amounts of issues in the results and methods sections, which makes me very concerned about the reliability of these results.

RESPONSE: Thanks for your valuable comments and suggestions that helping us to improve the manuscript. We have carefully revised and improved the manuscript.

major points:

1. For the sampling of different wild rice species, why a BBCC type of Oryza punctata was selected? Usually, a typical O. punctata sample is considered as an BB type, although there exist other types.

Response: Thanks for bring up this point. We agree BB genome type is more typical for the *Oryza punctata*. However, we haven’t collected the BB genome type. To avoid

confusion, we have labeled our *O. punctata* samples as "*O. punctata* (BBCC, 4X)" throughout our manuscript.

2. Although the authors constructed the phylogenetic tree of Oryza genus by distinguishing different genome types, the introgression and gene flow between different genomes may still cause strong interference in the phylogenetic relationships, which should be taken into account when constructing the phylogenetic tree.

Response: Thanks for your comment. The incomplete lineage sorting (ILS), horizontal gene transfer, natural selection, evolutionary radiation, and introgression caused by hybridization should be taken into account when constructing the phylogenetic tree.

In this revision, we applied a multi-species concatenation-based method incorporated in ASTRAL (v.5.7.8) which takes ILS into account to generate a species tree. ASTRAL took 3,555 single-copy gene trees as input and generated a species tree estimated by searching for the species tree that was most congruent with quartets garnered from the input gene trees.

As suggested by reviewer, we have corrected the inaccurate placement of AA genome species and updated the version of *Oryza* genus phylogenetic tree in Fig. 1b in revised manuscript (Fig. R1).

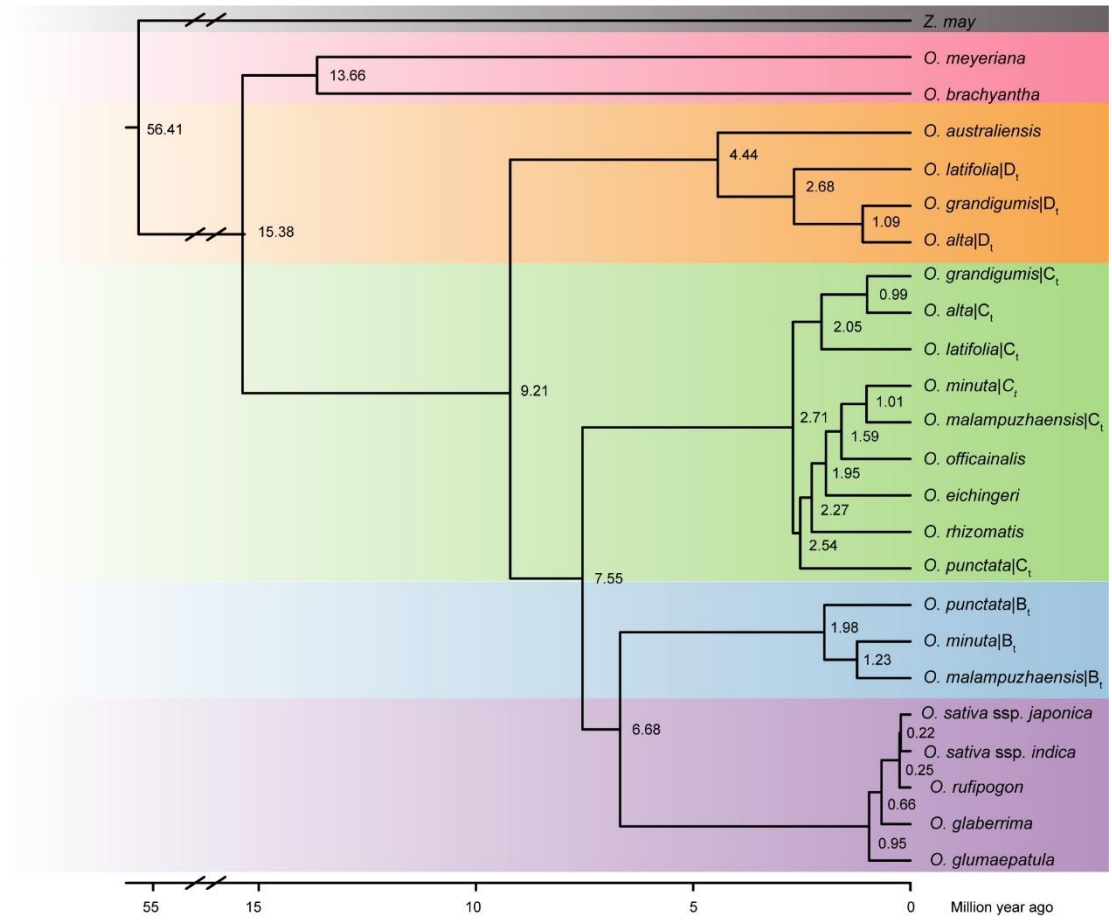


Fig. R1: Phylogenetic trees of the *Oryza* genus. allotetraploid wild rice genomes were split into sub-genomes to construction and the single copy genes were concatenated to generate the phylogenetic tree. Numbers at nodes indicate the median value for the divergence time (Mya) estimates for each clade.

3. The obtained phylogenetic relationships in this study is obviously conflicted with previous reported results. For the tree of nuclear genome in Fig 1b, O. glaberrima is clustered with O. sativa ssp. indica in the same clade while the two subspecies of O. sativa, O. sativa ssp. indica and O. sativa ssp. japonica are clustered in different clades. Moreover, the O. rufipogon that is commonly known as wild relative species of O. sativa, is placed as an outgroup of both O. glaberrima and O. sativa. This totally contradicts the consensus on the origin and evolution of Asia and African cultivated

rice over the past half century.

Response: Thank you very much for the careful review and detailed comments. We apologize for the error caused by simply selected 1000 out of 3500 genes for tree construction and performed only 100 bootstraps in previous analysis. The 22% supported rate in the right part of figure 1B (last version) was consistent with previous reported results. In this revision, we have now redone the analysis by reconstructing the phylogenetic tree based on the concatenation of 3555 single copy gene sequences and got a tree that is consistent with previously reported results, and the corrected *Oryza genus* phylogenetic tree can be found in Fig. 1b (Fig. R1). And we thank you again for improving our manuscript.

Thus, the supporting rate for different topologies of species tree could be invalid and the authors should check about their methods related to this part. Additionally, two supporting rates (44% and 46) very close to 50% on branch 2 suggest that the phylogenetic relationship here is not resolved, and also suggest a potential risk in the calculation method.

Response: Thanks for your suggestions and kind reminder. We apologize for the error caused by simply selected 1000 out of 3500 genes for tree construction and performed only 100 bootstraps in previous analysis. In this revision, we have now redone the analysis by reconstructing the phylogenetic tree based on the concatenation of 3555 single copy gene sequences from 18 species (24 subgenomes, *Z. may* as outgroup), and the corrected *Oryza genus* phylogenetic tree can be found in Fig. 1b (Fig. R1).

For the tree of chloroplast genome in supplemental figure 3, it is also conflicted with previous reported results, e.g., the O. punctata should be the closest lineage to the AA-genome species. And two important species, O. rufipogon and O. glaberrima, are absent in this tree.

Response: We apologize for any inaccuracies in our description that may have led to misunderstandings. The *O. punctata* used in this study is an allotetraploid (BBCC), with the chloroplast genome derived from the CC genome. Consequently, it clusters with CC genome species, consistent with previously reported findings (Zou et al., 2015). We have made revisions in Supplementary Fig. 4, which now includes *O. rufipogon* (KF359902) and *O. glaberrima* (KF359905) (see Fig. R2a) as your suggested. Additionally, we conducted a phylogenetic analysis of the chloroplast genome, incorporating *O. punctata* (BB genome, GenBank: KF359908). The results are presented in Fig. R2b.

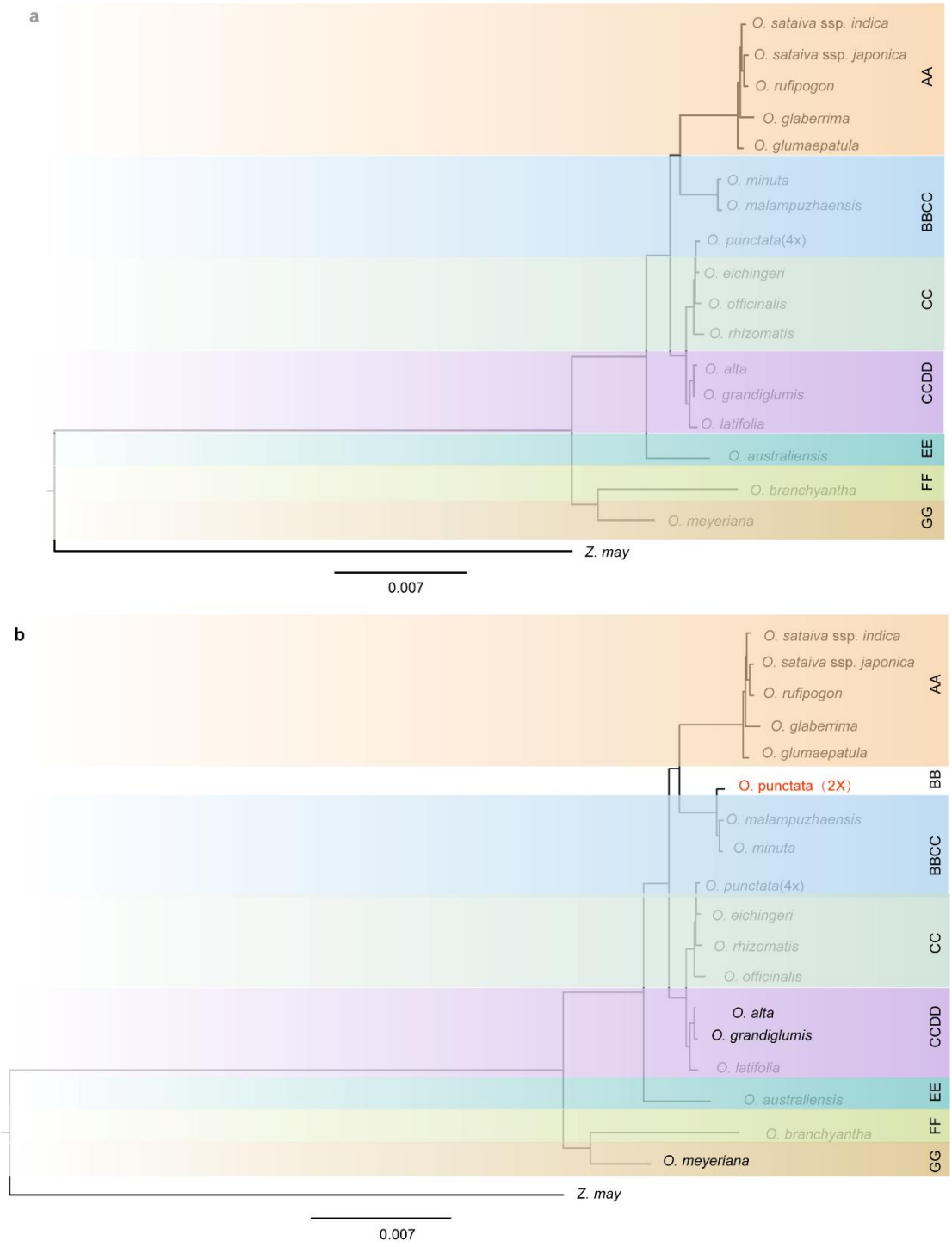


Fig R2: **Phylogenetic tree based on the chloroplast genome sequences.** a. Phylogenetic tree based on 17 wild rice species chloroplast genome sequences (Z. may as outgroup). b. Phylogenetic tree constructed based on 18 wild rice (including a BB genome type *O. punctata,* KF359908) species chloroplast genome sequences.

Zou, XH., Du, YS., Tang, L. et al. Multiple origins of BBCC allopolyploid species in

the rice genus (Oryza). Sci Rep 5, 14876 (2015).

4. The large inversion in Chr6 is detected in both O. sativa ssp. japonica and O. rufipogon as revealed by Fig 2a. As we know, the O. rufipogon is mainly distributed in South and Southeast Asia. The japonica rice consists of tropical japonica rice, subtropical japonica rice and temperate japonica rice, etc. But the Lines 223-225 described that "The inversion occurred only among the common wild rice and Oryza sativa japonica with high latitude distribution, indicating that it has risen to expansion of geographical range." This causal relationship lacks evidence and may even be incorrect. It also affects the significance and necessity of the subsequent analysis on the gene OsMFT1.

Response: Thanks for point out this problem. We agree with you. It is inappropriate to make such a conclusion "The inversion occurred only among the common wild rice and *Oryza sativa* japonica with high latitude distribution, indicating that it has risen to expansion of geographical range". We have removed these sentences in this revision.

5. Fig 3a presented the function genes with CNV mutations, however, all genes showed more than two copies in the Nipponbare genome but present-or-absent variations in other species. If this gene dataset can be easily obtained by selecting genes with multiple copies from Nipponbare genome, what is the significance of constructing a super pangenome? The authors should recollect a more meaningful gene dataset to conduct an in-depth analysis on the CNV of important genes across the wild and cultivated rice species, highlighting the advantage of this super pangenome.

Response: Thanks for your constructive suggestions. We agree that detecting function genes with CNV mutations in the *Oryza* genus super pangenome is more challenging and meaningful than that in Nipponbare. In this study, we conducted comprehensive analyses of the CNV alleles within the rice genus and examined the CNV genes associated with significant agricultural traits in rice. However, genes related to agricultural traits have been less explored in wild rice. Consequently, we divided the CNV genes into two categories: the first category comprises genes with multiple copies identified from the Nipponbare genome, while the second category was derived from comparisons across wild rice species, as these genes are absent in the Nipponbare genome. Detailed information can be found in Fig. 6 of the revised manuscript (Fig. R3).

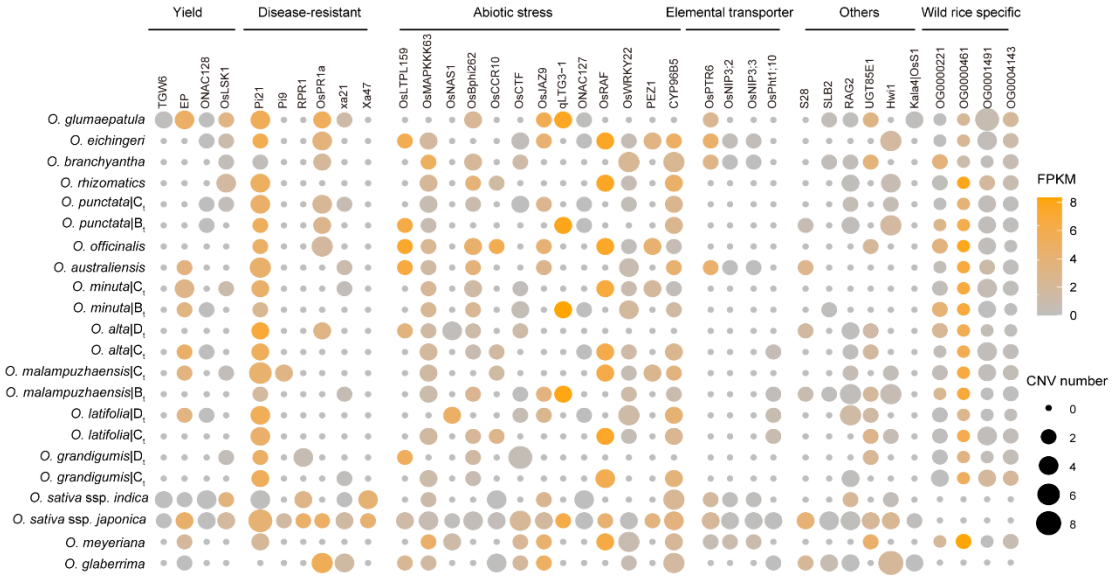


Fig. R3: Characteristics of gene CNVs related to important agronomic traits in the Oryza genus super pangenome. Circle size shows the number of gene copies, color from grey to yellow represents the gene expression.

Furthermore, the Fig 3j showed a more similar collinearity of NLR genes between O.

sativa ssp. indica and O. sativa ssp. japonica than that between O. sativa ssp. indica and O. glaberrima, indicating a closer phylogenetic relationship between the two subspecies of O. sativa than that between indica and O. glaberrima. This result is conflict with the results of phylogenetic analysis in this study but is consistent with the current consensus on the origin and evolution of Asia and African cultivated rice.

Response: Thanks for your careful and conscientious review. We agree with you that *O. sativa* ssp. *indica* and *O. sativa* ssp. *japonica* had a closer phylogenetic relationship than that between *O. sativa* ssp. *indica* and *O. glaberrima*. The error was caused by our previous phylogenetic tree construction. In this revision, we have reconstructed the *Oryza* phylogenetic tree by concatenated based methods. The revised tree now correctly places *O. sativa* ssp. *japonica* closer to *O. sativa* ssp. *indica* than to *O. glaberrima* in Fig. 6 (Fig R4). In addition, we have carefully revised all the results related to the phylogenetic tree in this article.

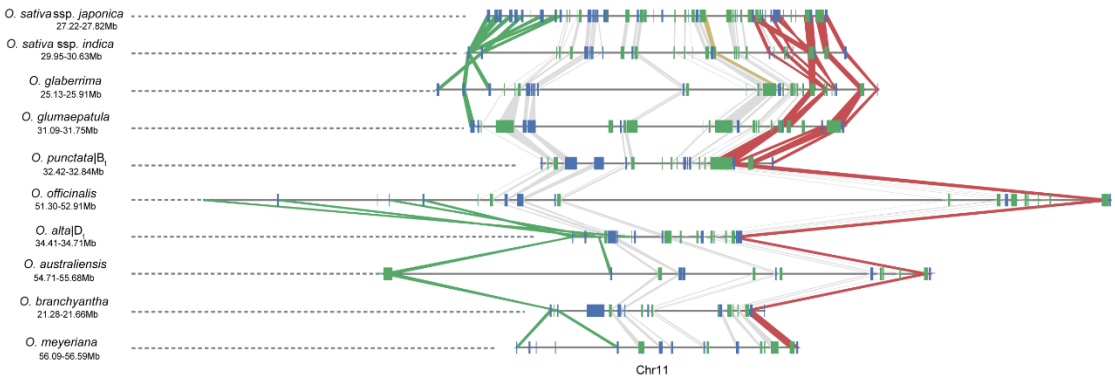


Fig. R4: Collinearity of NLRs in various *Oryza* genomes within a specific region.

Additionally, the authors clarified that "Analysis of NLR distribution showed that while R gene singletons were similar between wild and cultivated rice, cultivated rice tended to have a higher number of R genes in pairs or clusters compared to wild rice" at Lines 390-392. However, considering that most of the existing rice NLR gene data sets are

based on studies on cultivated rice genome, this may also be the case that NLR in wild rice genomes has not been fully identified.

Response: Thank you very much for the professional suggestions. We believe our analysis does provide the comprehensive NLR gene data sets in wild rice genomes. The NLR identification in this study was used the combination of denovo annotation of the genomes and RGAugry software. For genome sequence annotation, LRR domain were extracted from the denovo annotation of the genomes (Table R1). The another method was as follow: Firstly, RGAugry was widely used in R gene identification of wild rice[1], wild potatoes[2], and wild radishes[3] and obtained a comprehensive results. The total NBS gene number of *O. alta* identified in our results was content with the previous study[1]. Secondly, the NLR results was verified by using Interproscan and hmmsearch software. The both identified NLR was recorded for subsequent analysis. To obtain a comprehensive NLR in the wild rice genomes, we selected the NLR region which cultivar had more NLR than wild rice and selected one as example for validation (Figure R5).

Table R1: The denovo annotation of the genome with LRR domian

Species	LRR domain number	R gene number
O. sativa ssp. indica	789	511
O. sativa ssp. japonica	709	473
O. glaberrima	732	419
O. rufipogon	448	260
O. australiensis	432	195
O. brachyantha	472	245
O. rhizomatis	627	343
O. glumaepatula	613	354
O. eichingeri	588	288
O. meyeriana	482	238
O. alta|Ct	508	260

O. alta|Dt	458	216
O. grandiglumis|Ct	515	224
O. grandiglumis|Dt	456	183
O. latifolia|Ct	585	314
O. latifolia|Dt	460	206
O. minuta|Bt	553	258
O. minuta|Ct	593	336
O. punctata|Bt	596	320
O. punctata|Ct	623	349
O. malampuzhaensis|Bt	559	302
O. malampuzhaensis|Ct	559	277

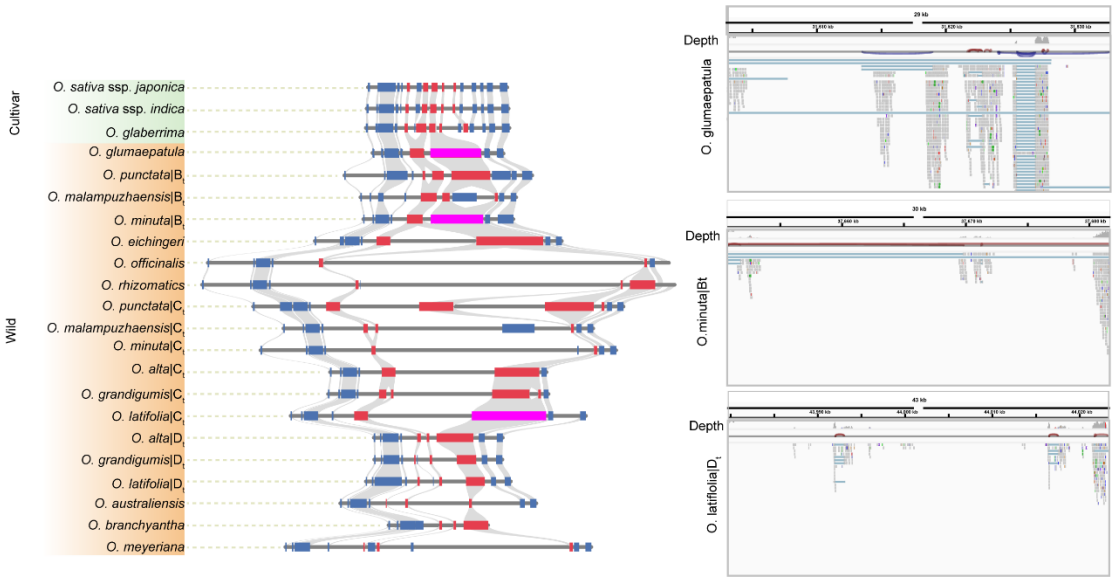


Fig. R5: A NLR synteny region and RNA evidence of the annotated NLR genes. The pink color in *O. glumaepatula, O. minuta*|B_t, *O. latifolia* NLR gene were validated by the RNA-seq in IGVtools.

Minor points:

1. Line 64: lack of species name of Asian cultivated rice, Oryza sativa.

Response: Corrected.

2. The Latin names of indica and japonica rice are not properly written, e.g. should be Oryza sativa ssp. indica and Oryza sativa ssp. japonica. Please check through the

manuscript and all the figures and tables.

Response: We have revised it through the manuscript.

3. Line 97: the reference 9 (Qin et al., 2021) was used for Nipponbare genome, but actually this study used the IRGSP-1.0 (Kawahara, Y. et al., 2013) as shown in the M&M section.

Response: Thanks for point out the questions. We used the Nipponbare genome (Qin et al., 2021) as the reference. We have corrected it in the M&M section.

Qin, P. et al. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* **184**, 3542–3558 e3516 (2021).

4. Extended Data Fig. 3c, the second member of dispensable family showed a presence in all types of genomes, which is identical with the core family.

Response: Corrected in Fig. 2c in this revision.

5. The chromosomes of different species were drawn with lengths not consistent with their genome size, e.g., in extended Data Fig. 6 and 7, supplementary 7 and 10.

Response: Thanks for your constructive suggestions. The chromosomes of different species were drawn with lengths consistent with their genome size in Supplementary Fig. 7 (Supplementary Fig. 10 in this revision). Extended Data Fig 6 was only used to show the alignment of collinear gene pairs, and does not account for chromosome lengths. It is not taken the length of chromosome into consideration. We have corrected the Extended Data 7 and Supplementary Fig 10 to Supplementary Fig 17 and Supplementary Fig 14 in line in this revised manuscript.

6. The dotplot of pairwise alignment for Nipponbare genome and other genomes were drawn with a reverse orientation and should be corrected.

Response. Corrected in Supplementary Fig. 11 in this revision.

Reviewer #2 (Remarks to the Author):

Sequencing the genomes of crop wild relatives is very important for identifying novel and adaptive variation, especially in the light of combating climate change. The paper by Long et al. is an excellent example of the types of study that should be undertaken. I have some suggestions for improvements, some I would suggest are vital and others are smaller or optional, as follows:

Thank you very much for your recognition and your valuable comments. We have responded to them point by point, and these modifications have helped us to improve the manuscript significantly.

Major changes:

1. Figure legends need significant improvement. For example, for fig1:

“structure, panicle” > “structure of the panicle”. Also the panicle panel doesn’t have the species names indicated. If they are in the same order as the whole plant photos then say this.

“TE content in each subgenome/genome of the wild and cultivated rice”. > what about it? – say it is indicated on the right of the phylogeny

“The order of species is corresponding to the phylogeny tree”. > use “phylogenetic tree” or “phylogeny” and I don’t understand why you need to say this statement, the species are clearly in phylogeneticv order, that’s the point of a phylogeny.

“Proportions of contrasting gene tree topologies for the 100 genes with regard to two major conflicting relationships”. > indicate these as panels c and d.

All figure legends need work.

Response: We agree with your suggestion. Thus, we have updated all the figure legends through the manuscript. Accordingly, we have updated the Figure 1 and modified the legend as following:

Line 626-637: “**Fig. 1: Geographical distribution and phylogeny of wild and cultivated rice.** **a** Geographic distribution of the 13 wild rice varieties and their diverse agronomic characteristics, such as plant height. The fonts colors of wild rice species corresponding to the distribution on the world map. These 13 accessions covered 13 species in the rice genus. **b** Panicle architecture of the 13 wild rice species. **c** Phylogenetic tree based on the conserved single-copy gene illustrates evolutionary history in the Oryza genus. All the allotetraploid wild rice genome was split into subgenomes, the ASTRAL concatenation-based species tree estimated by 3555 single copy gene generated by orthofinder. Numbers at nodes indicate the median value for the divergence time (Mya) estimates for each clade. **d** TE content in each subgenome/genome of the wild and cultivated rice. RNA transposons were presented as following: blue for LTR, yellow represents Helitron, pink indicates LINEs, orange shows SINEs, orange blue stand for unclassified. DNA transposons was showed as wine.”

2. Present the BUSCO analysis for each sample, not just the average. As allotetraploids were involved it would be very interesting and important to be able to see the BUSCOs.

Response: Thank for your insightful feedback. In response to your comments, we have added the BUSCO value of each subgenomes of allotetraploid in Supplementary Table

1 (Table R2). There is a higher value of BUSCO in subgenome than that in allotetraploid due to gene loss in subgenome also famous for subgenomic dominance.

we also noticed that the BUSCO value in C subgenome was significantly larger than that in D subgenome, which indicated C was more conserved than D subgenome during selection and evolution.

Table R2: Quality assessments of 6 diploid genomes.

Species	BUSCOs assessment					
	Completeness (%)	Single_Completeness (%)	Duplicated_Completeness (%)	Fragment (%)	Missing (%)	TotalNumber (%)
O. alta|Ct	94.3	85.4	8.9	0.6	5.1	1614
O. alta|Dt	89.5	86.5	3	0.9	9.6	1614
O. grandiglumis|Ct	95.1	85.9	9.2	0.7	4.2	1614
O. grandiglumis|Dt	89.7	86.7	3	0.7	9.6	1614
O. latifolia|Ct	98.1	96.1	2	0.3	1.6	1614
O. latifolia|Dt	97.9	96.1	1.8	0.6	1.5	1614
O. minuta|Bt	94.7	93	1.7	2.6	2.7	1614
O. minuta|Ct	94.1	91.9	2.2	2.8	3.1	1614
O. punctata|Bt	96.1	94.9	1.2	1.9	2	1614
O. punctata|Ct	96.8	94.4	2.4	1.4	1.8	1614
O. malampuzhaensis|Bt	95.7	93.9	1.8	2	2.3	1614
O. malampuzhaensis|Ct	97.4	94.5	2.9	0.9	1.7	1614

3. The orthofinder analysis would give very different results if different settings were used. How were the settings chosen selected? The very high number of orthogroups (101k) from comparing many genome but each only containing 30k genes suggests a very large number are single copy and species-specific, which is what is reported. However I doubt it is true that 1/3 of genes are truly species-specific, and certainly relaxing the settings would reduce the 101k and reduce the number of species specific. So simply stating there are 101k orthogroups is premature and not validated. Only 10% of genes being core genes suggest to me the settings were too strict.

Response: Thank you very much for your kind advice. The different parameters indeed influence the size of the super pangenome and its individual components. In the previous manuscript, we initially constructed the super pangenome using default parameters (e^{-3}), which generated a total of 96,962 orthogroups. This included the following counts and percentages: core gene families (10,015, 10.33%), dispensable gene families (43,270, 56.84%), and private gene families (43,667, 45.05%). Notably, the percentage of private gene families is relatively high. To achieve a more comprehensive and accurate representation of the *Oryza* genus super pangenome, we selected a stricter setting (e^{-10}) for this analysis. Although the total number of gene families in the pan-genome is slightly larger than that obtained with the default parameters, there is no significant difference in the counts of core gene families. Detailed information regarding the *Oryza* genus pangenome constructed with the default parameters is presented in Figure R6

We felt sorry for our unclear expression, we have moved the Extended Data Fig. 2 to

Fig. 2 in the main text. As the Fig. 2b shown, a total of 9,834 (9.66%) core gene families with 318,253 (33.22%) genes, 57,822(56.84%) gene families contains 599412(62.58%) genes in dispensable genome, and 34,067(33.48%) gene families includes 40115(4.18%) genes. Extended Data Fig. 2a (Fig. 2a in this revision manuscript) showed that 1/3 gene families and 4.18% gene being species-specific, we have added annotation to the Figure 2a in this revision.

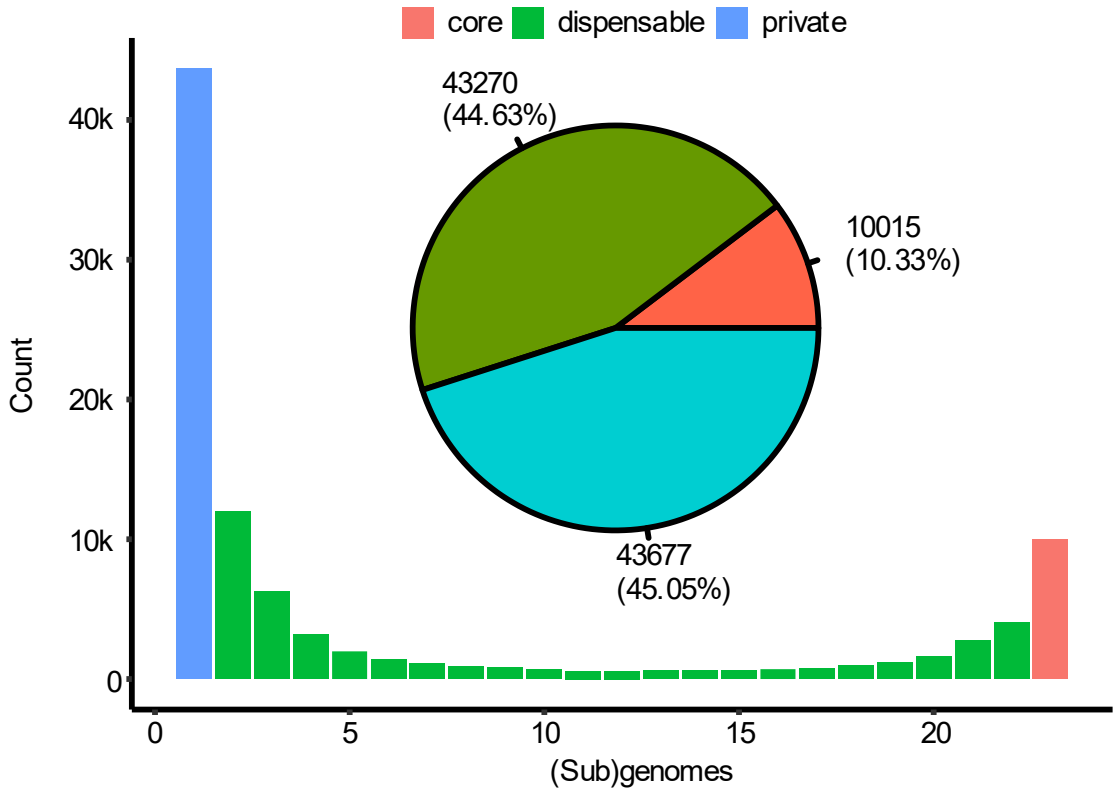


Fig. R6: The *Oryza* genus super-pangenome constructed by using default parameters.

4. The section on TEs (from line 154) is very descriptive – I suggest adding some statistics to back up the statements, eg you say there is an increase or decrease in some TE family sixed but is this just random variation across the phylogeny or are there significant increases in some species/groups relative to others?

Response: We are grateful to the reviewer for bringing up this question. We have revised

it from line 223 to line 244 in this revision.

In the same section, at line 178 the authors state “they influenced a certain proportion of genes expression” but I don’t see evidence for this. Extended data Fig 5c, Supplementary Fig. 7b do not have anything to do with expression.

A: Thank you for pointing out this concern. Previous findings indicated that gene expression levels were suppressed by the abundance and physical distance from neighboring TEs (Zhang et al., Proc Natl Acad Sci U S A, 2015; Hanada et al., Plant cell, 2009; Castanera et al., PLoS Genetic, 2016; Gagliardi et la., Proc Natl Acad Sci U S A, 2019). Therefore, in this study, TE that overlapped with gene positions were identified as influencing gene expression TE.

1.Zhang J, Liu Y, Xia EH, Yao QY, Liu XD, Gao LZ. Autotetraploid rice methylome analysis reveals methylation variation of transposable elements and their effects on gene expression. Proc Natl Acad Sci U S A. 2015 Dec 15;112(50):E7022-9.

2. Hanada K, Vallejo V, Nobuta K, Slotkin RK, Lisch D, Meyers BC, Shiu SH, Jiang N. The functional role of pack-MULEs in rice inferred from purifying selection and expression profile. Plant Cell. 2009 Jan;21(1):25-38.

3. Castanera R, López-Varas L, Borgognone A, LaButti K, Lapidus A, Schmutz J, Grimwood J, Pérez G, Pisabarro AG, Grigoriev IV, Stajich JE, Ramírez L. Transposable Elements versus the Fungal Genome: Impact on Whole-Genome Architecture and Transcriptional Profiles. PLoS Genet. 2016 Jun 13;12(6):e1006108.

4. Gagliardi D, Cambiagno DA, Arce AL, Tomassi AH, Giacomelli JI, Ariel FD, Manavella PA. Dynamic regulation of chromatin topology and transcription by inverted

repeat-derived small RNAs in sunflower. Proc Natl Acad Sci U S A. 2019 Aug 27;116(35):17578-17583.

Also, line 181 says TEs are related to centromeres, but in the figure the centromeres are not indicated so how is there evidence for this statement?

A: Corrected in Supplementary Figure 12(Figure R7) in this revision.

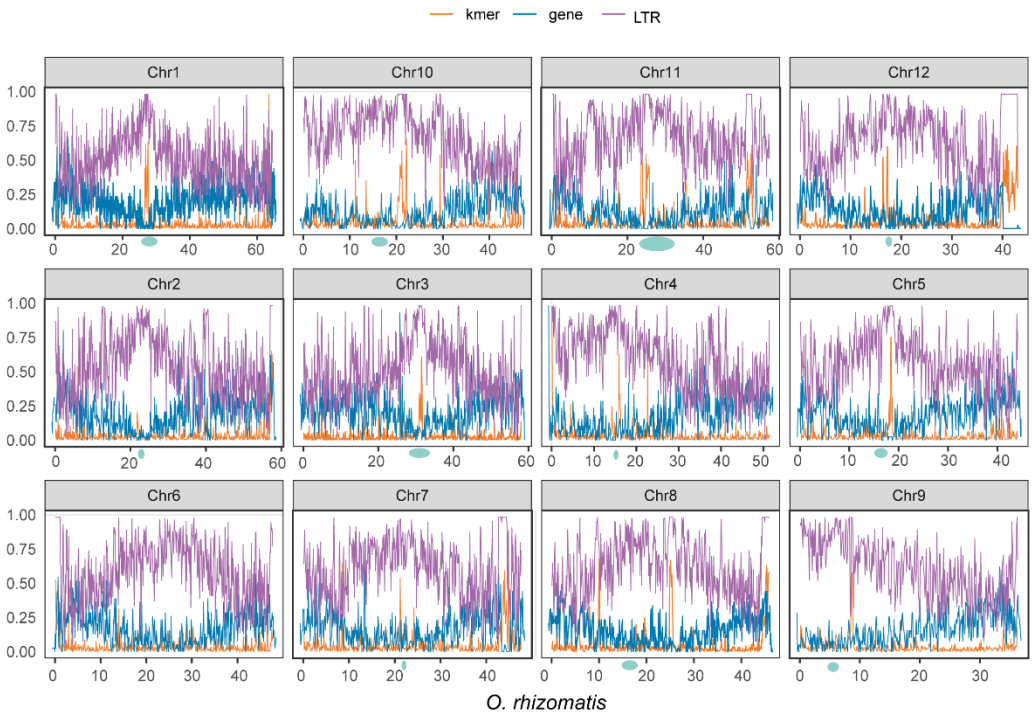


Fig. R7: Landscape of the centromere in wild rice *O. rhizomatis*.

5. The section on synteny is unusual because everything is compared to NIP. So therefore statements such as "The AA genomic type genome displayed a significant level of chromosomal conservation." is biased –of course it will show synteny because its compared to NIP which is AA too. This statement is not backed up unless B is compared to B, C to C and D to D etc. Related, "the gene synteny findings provide support for the phylogenetic position of the rice genus (Supplementary Fig. 10)." – this figure does not tell me anything about the phylogeny. It makes sense that related species

have similar genomes, but unless this is specifically being tested, then remove this statement.

A: We appreciate your suggestions. We agree with you that "The AA genomic type genome displayed a significant level of chromosomal conservation." is biased. In this study, the synteny and variation were identified compared to NIP reference genome. Hence, **we have removed these sentences in this revision**. Furthermore, **we also removed the sentence** "the gene synteny findings provide support for the phylogenetic position of the rice genus (Supplementary Fig. 10)." as you suggested. The supplementary 10 was moved in line 278 in Supplementary figure 14 in this revision to show the SV within rice diploid genomes.

6. lines 262 – 274 – without any stats this can just be random variation between species. This must be added to allow for these statements to be made. Line 404, again without stats this can be random variation. Stats must be added.

Response: We agree with your suggestion and we have rewritten the sentence in line 320-335.

7. There are several analyses that are poorly explained or, in my opinion, irrelevant, for example:

Line 342-7 – this seems like any genome region could have the same properties – this statement simply says that the different species have different SVs. But nothing about selection during domestication. In addition, it is already known that Rc controls pericarp colour, so what is the need to focus on it here? > this analysis doesn't seem to be meaningful and is not well-explained.

Response: Thank you for your comments. There is no relevant with the selection during domestication. We have removed these statements in the revised manuscript as your suggested.

L382 – "This suggests that the immune system in wild rice has a more diverse evolutionary history compared to cultivated rice." – the comparison is two cv and many wilds, it is therefore not surprising there is this difference. Also the cv rice is diploid and the wild contains polyploid therefore at least double the number of genes is expected in punctata. > this analysis is not meaningful.

Response: Thank you very much for the kind suggestion. That is a mistake of writing from our side. In our opinon, *Oryza* genus can divided into 25 species, which including two cultivated species, *O. glaberrima*, and *O. satiava*. Wherea *O. sativa* contains two subspecies: *O. sativa.* ssp. *indica* and *O. sativa.* ssp. *japonica.* and the remain species all belong to wild rice. We have rewritten the sentence "This suggests that the immune system in wild rice has a more diverse evolutionary history compared to cultivated rice." to "which revealed a significant expansion of NLRs in the cultivar rice" in lines 409-410.

We identified the R gene number in two subgenomes (diploid genomes) instead of the allotertraploid wild rice (polyploid genomes). Thus, all analysis was conducted at diploid genome level.

Smaller comments:

Line 39 – expansion of range in which species? Osat? or both Osat and Orufi?

Response: We have removed these sentences according to the suggests of Reviewer 1.

Line 43 – I don't agree with the term 'hidden legacy' and I think it should be removed here.

Response: Corrected

Line 52 – state these are genome groups, because as written it is discussed like 6+5 species (but 20 species). Same for lines 61-2 where you say 11 species and list six genomes.

Response: We feel very sorry for causing your confusing due to our imprecise expression and thank you very much to point out this issue. Wild rice includes more than 20 species, and they are currently divided into 11 genome types: AA, BB, CC, EE, FF, BBCC, CCDD, HHJJ, HHKK, and KKLL. Among the wild rice, 6 species belong to AA genome type with the cultivated rice: *O. rufipogon*, *O. nivara*, *O. barthii*, *O. glumaepatula*, *O. longistaminata*, and *O. meridionalis*. And most of AA genome type wild rice had been assembled to chromosome level. Hence, we only selected *O. glumaepatula* in AA-genome type wild rice in this study. While BBCC genome types contains three species, named *O. punctata*, *O. minitus*, and *O. malampuzhensis*. *O. latifolia*, *O.alta*, and *O. grandigumis* were belong to CCDD genome type. GG genome type include two species: *O. meryeriana* and *O. granulata*.

We have revised the sentence in lines 52-53 as "These wild rice are currently divided into 11 genome types: AA, BB, CC, EE, FF, GG, BBCC, CCDD, HHJJ, HHKK and KKLL," and line 66 as "Currently, there is a lack of high-quality reference genomes for

the most of rice wild relatives" in this revised manuscript.

You say there is a lack of genome for BB but you didn't sequence a BB genome species according to Supp table 1, correct?

Response: Thank you very much for bringing this up. We have not sequenced the diploid genome (BB genome type) of the *Oryza punctata* species because we have not yet collected the sample of this diploid species. This issue has been addressed throughout the entire manuscript

The introduction appears to brief in my opinion. Some more about what each of the species being sequenced brings would be appropriate, eg "species X (BB genome) is tolerant of X" and so on. Just a couple of sentences would help.

Response: Thank you very much for your suggestions and recommendations. We have made changes to the introduction as suggested, "A few non-A genome type *Oryza* genomes have been published with a chromosome level to date. The G genome type was firstly assembled to a chromosome level by using gradient resequencing and Pacbio SMRT data and combing the positive selection revealed photosynthesis and energy production related gene might have facilitated its tolerance to shade. Allotetraploid *O. coarctata* were released with a total length of 573.4 Mb and an N50 of 23.1 Mb by using HIFI and HiC data. Yu assembled a high-quality reference genome of *O. alta* and through gene editing for important agricultural gene, completing the de novo

domestication of allotetraploid wild rice" was added into the introduction in lines 67-75 in this revision.

Several of the figures state the species name but a lot of the text refers to genome groups. It would be good for genome group to be in parentheses on some figures, especially figure 1.

Response: Corrected.

Line 89 – how does supp figure 1 relate to this statement about the quality of the genomes?

Response: Thank you for raising this question. We acknowledge the oversight in our writing. We used Merqury to evaluate completeness of assembled genomes. And the consensus quality values were listed in the Supplementary table 4(Table R3), we have corrected the Supplementary figure 1 to Supplementary table 4 in this revision.

Table R3: The Quality assessments of 13 assemblies.

No	Species	QV	Completeness
M-6	*O. meyeriana*	49.4919	75.1455
W-100	*O. grandigumis*	53.5847	99.087
W-128	*O.latifolia*	55.0804	98.6829
W-171	*O. malampuzhaensis*	54.143	99.134
W-18	*O. alta*	53.7814	81.8891
W190	*O. minuta*	54.9547	99.2397
W-26	*O. australiensis*	57.6184	95.5999
W-323	*O. officinalis*	50.3274	90.1201

W-363	*O. punctata*	55.9447	99.5892
W-375	*O. rhizomatis*	52.1117	91.9521
W-59	*O. branchyantha*	49.5088	99.5186
W71	*O. eichingeri*	54.4785	68.6229
W-97	*O. glumaepatula*	48.3143	99.2445

93-4 – 37000 is not higher than 41000 – please reword this sentence.

Response: Corrected.

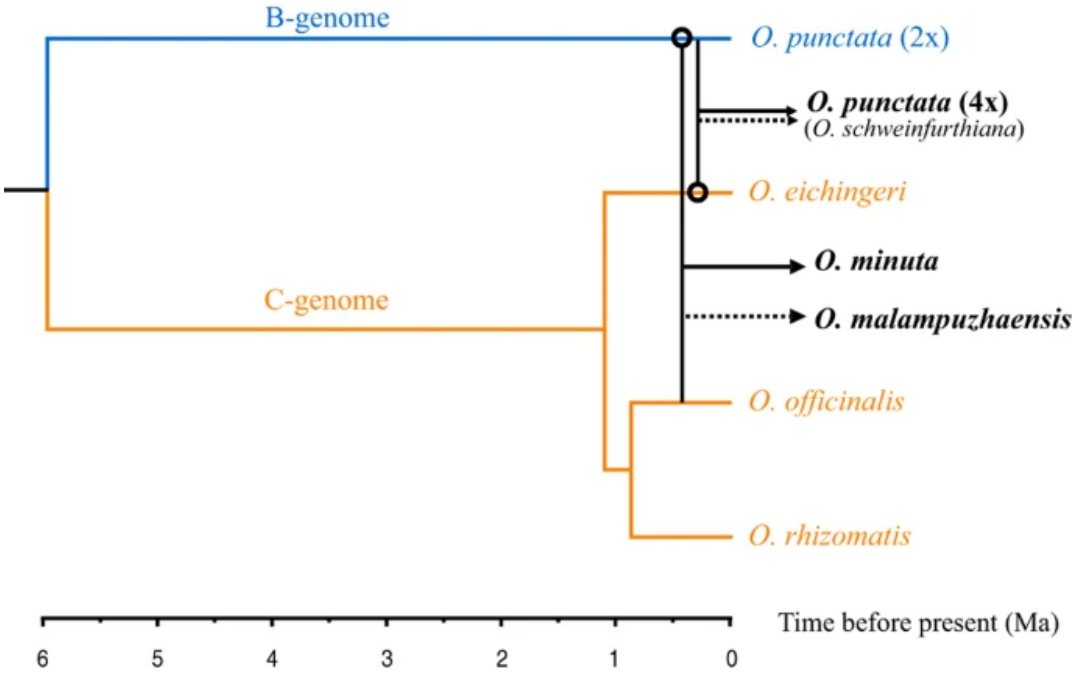
Line 98 – "contradicted" in what way? Where are the contradictions? Why were 1000 of the 3555 used in the analysis not all 3555? This refers to fig 1b but see comment above about naming panels c and d.

Response: We apologize for the error in the phylogenetic tree. We have reconducted the phylogenetic tree using all the 3555 single copy orthologs identified by Orthofinder. The revised phylogenetic tree now was consistent with previous results. In previous study, we want to save the save time and computing resources, then we simply select 1000 of the 3555 genes and bring such an error in the phylogenetic construction. We have revised it in Figure 1 in this revised manuscript.

Line 102 – state what was found too, don't leave it to the reader to download the supp figure. Does this match previous work on the cpDNA genome?

Response: Thank you very much for your nice advice. We have revised it to "revealed *O. eichingeri* is the female parent of *O. punctata* (BBCC) (Supplementary Fig. 2)." In

lines 148-149 in this revision. Our findings were consistent with the results of the previous study (Zou et al., Sci Rep, 2015). The reference can be found as followed:



Zou, XH., Du, YS., Tang, L. et al. Multiple origins of BBCC allopolyploid species in the rice genus (Oryza). Sci Rep 5, 14876 (2015).

Lines 123-8 – it is confusing to sometimes use the number and sometimes the percentage. Please clarify this.

Response: We have corrected it to "Additionally, a syntenic pangenome was constructed to analyze the differentiation among the 7 diploid genomes, revealing core gene families shared across these genomes accounted for 17.37% of total gene sets, Dispensable families, accounted for 29.73% of the total gene sets (Supplementary Fig. 6, Supplementary Table 6), with the largest proportion represented by genome type private gene sets, unique to individual genome types, making up 52.90% of the total gene sets (Supplementary Table 6)." in lines 172-178 in the revised manuscript.

Line 131 remove “approximately” this number is not an approximation. In this section it would be interesting to see what percentage of core, disp, and sp-sp genes are supported by transcriptome data.

Response: We agree with your suggestions. The word “approximately” is not suitable here. We have corrected it to “Within the *Oryza* genus, 81.14% of the core genes could be assigned to protein domains in our Pfam and InterPro databases, which is nearly twice the percentage of dispensable genes (41.82%) and more than seven times higher than that of accession-specific genes (10.83%).” In lines 179-182 in this revision.

Additionally, we calculated the percentage of genes in each component of the pangenome supported by transcriptome data. An average of 92.40% of the genes were expressed in the core genome, which is significantly higher than the 63.31% of genes expressed in the dispensable genome and the 52.15% in private genome in Fig. 2d in the revised manuscript.

136 – you mention length then Ka/Ks. This is confusing.

Response: We have corrected it to “Core genes showing significantly lower (0.15-fold on average) pairwise nonsynonymous substitution/synonymous substitution ratios (Ka/Ks) compared to the dispensable genes (Figs. 2f, g), indicating conservation of function among core genes in the *Oryza* genus, while variable genes evolved more rapidly” in lines 186-190 in this revision.

140 – remove “to adapt to diverse environments” this is complete speculation.

Response: Corrected

150 – what do you mean “except for”?

Response: Thanks for your careful reading. We are regret for our expression careless, and we have clarified this sentence as bellow.

A further GO analyze of the different genome type rice private gene was performed to understand their functional differentiation. A genome wild rice private genes were enriched in resistance to heavy metal and salt. B genome wild rice were focus on response to oxidative stress. C genome type wild rice private gene significantly enriched in regulation of salicylic acid and steroid hormone, which contributed their high resistance to various diseases. D genome type wild rice private genes were significantly associated with arginine synthase, which give raise to sense and adapt to environmental changes for rice. E genome type wild rice private genes were significantly involved in cellulose synthase, which benefit to large biomass. F genome type private genes were significantly associated with defense response. While the G genome type rice specific genes were focus on electron transfer, which contribute to shape its shade tolerance in lines 198-211 in the revised manuscript.

156 – "Scientists have been intrigued by" is not a good phrase. Remove

Response: Corrected.

172 – "puffy"? this is definitely not the correct word.

Response: Thank you for carefully review. We had corrected it to "The genome size of G genome type (*O. meyeriana)* was predominantly influenced by Retand LTR amplification, which counted as 8.66% of the genome size (Fig 3c), and the 10.17% abundance of Tekay in *O. glumaepatula* surpassed that 2.78% in the *O. sativa* resulted in the increase of genome size." in lines 232-236 in this revision.

224 – “it has risen to expansion of geographical range” is pure speculation and not backed up by any data, only a correlation. Genetic drift is equally plausible.

Response: Thank you very much for bringing this up and help us to improvement our manuscript. We acknowledge that the conclusion was not sufficiently rigorous, and we have removed it in this revision.

276 – how does “substantial variation in SV sizes” indicate “an enrichment of SVs in repetitive DNA regions”?

Response: Thank you for your comments. We have rewritten the sentence in lines 331-335as follows:

However, the size of structural variation in wild rice, particularly in *O. australiensis* and *O. meyeriana*, was found to be 614.70 Mb and 711.90 Mb, respectively, was overlapped with repeat sequences. This is significantly higher than the 274.43 Mb and 258.09 Mb associated with non-repeat sequences (Supplementary Fig. 17).

Line 288 – “This gene might be de novo birth to contribute to the ability of wild rice to adapt to poor and problem soil in Sri Lanka.” – this statement is confusing but also completely speculative. What does the analysis indicate the function of this gene is? Is it supported by expression data?

Response: We agree with your suggestion. The evolution of PSTOL1 was speculative, the more direct evidence was lacking. After our comprehensive consideration, we have removed this statement in this revised manuscript.

Line 290 – “The phylogenetic results revealed that the gene originated from O.

eichingeri and then transferred to O. punctata, providing evidence that O. eichingeri was the progenitor of tetraploid O. punctata, consistent with our chloroplast evolution results." > where is the evidence it was transferred to punctata? Isn't this simply because punc couldn't have evolved without eich? How does it back up the cpDNA results, these are completely different things. The parents are shown by the nuclear genome and the female parent by the cpDNA. The cpDNA has nothing to do with determining the parents. Surely the parents were already known too prior to this analysis. This could just be removed, it doesn't mean anything.

Response: We agree with your suggestion. We have removed these sentences.

L296 – 301 – picking one gene out of a QTL region is again entirely speculation that it is involved in resistance. What is the evidence that this gene is the likely causative gene? How many genes were in this region? The gene structure looks very different between sat and off, is this really an orthologue?

Response: We agree with your suggestion. we have removed the *Bph3* as an example in this revision. And subsequent complementary verification will be demonstrated in our following studies.

L302-6 – I have no idea what is the relevance of this.

Response: We have removed it.

L311 – what are "whole-genome alleles"?

Response: The term "whole-genome alleles" refers to genome-wide allelic in all *Oryza* species in comparison to the NIP reference genome. Our objective is to investigate the

genome-wide allelic variation in wild rice, as the functions of most genes are well-documented in the reference genome. We have corrected it in this revision.

L315 – "decreased, ranging from 18,463 to 23,812," – this sounds like an increase. Please reword.

Response: Corrected.

L333-7 – I don't understand this. It sounds like the genes in collinear regions were translated and their domains compared. But then 7 clusters were found. Surely there are more than 7 clusters of proteins. Do you mean 7 members per cluster? Ext Fig 11d does nothing to help me understand.

Response: We revised these sentence as "A genome-wide protein cluster was created based on their domain similarity, revealing that the protein in wild rice had 7.71 clusters in average, whereas only grouped to 1.78 in cultivated rice." in lines 367-369.

Line 342-7 – this seems like any genome region could have the same properties – this statement simply says that the different species have different SVs. But nothing about selection during domestication. In addition, it is already known that Rc controls pericarp colour, so what is the need to focus on it here?

Response: Thank you for your suggestions. We just selected one example to illustrated that wild rice had more variations than cultivar in protein sequence, and it is nothing about selection during domestication. We have revised it in lines 374-376.

L382 – "This suggests that the immune system in wild rice has a more diverse evolutionary history compared to cultivated rice." – the comparison is two cv and many wilds, it is therefore not surprising there is this difference. Also the cv rice is

diploid and the wild contains polyploid therefore at least double the number of genes is expected in punctata.

Nip has been sequenced before. How does your Nip genome compare to the previously sequenced Nip?

Response: We are sorry for the confusion. That is a mistake of writing from our side. In our opinon, *Oryza* genus can divided into 25 species, which including two cultivated species, *O. glaberrima*, and *O. satiava*. Wherea *O. sativa* contains two subspecies: *O. sativa.* ssp. *indica* and *O. sativa.* ssp. *japonica*. and the remain species all belong to wild rice. We have rewritten the sentence "This suggests that the immune system in wild rice has a more diverse evolutionary history compared to cultivated rice." into "which revealed a significant expansion of NLRs in the cultivar rice".

We identified the R gene number in two subgenomes (diploid genomes) instead of the allotertraploid wild rice (polyploid genomes). Thus, all analysis was conducted in diploid genome level.

We did not sequence the Nipponbare genome; instead, we used NIP (Qin et al., 2021, Cell) as the reference genome.

Qin, P. et al. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* **184**, 3542–3558 e3516 (2021).

Reviewer #3 (Remarks to the Author):

The paper describes the generation of reference-level genomes of 13 wild rice accessions (including BB, BBCC, CC, EE, FF, and GG) using HiFi technology. The genome assemblies are generally of high quality and will be an important resource for

utilizing useful alleles in wild rice gene pools. However, there are several major concerns about the work, which are listed below:

Thank you for your valuable comments and suggestions helping us to improve the manuscript. We have carefully revised and improved the manuscript.

1. RNA sampling of diverse tissues is critical for genome annotation. Currently, the data seem to not have been comprehensively collected and are not well utilized.

Response: We appreciate your comments and suggestions. All tissues, including leaf, stem, root, and spikelet, have been collected for the 13 wild rice species in this study, except for the species *O. meyeriana* (Supplementary Data 3). We collected only one plant of *O. meyeriana*, which has low biomass; therefore, the RNA data for this species has not been comprehensively collected (only the root is included). However, we have obtained public RNA data for *O. meyeriana* (SRR16192096 and SRR1685731) and the closest relatives *O. granulata* (PRJNA756291) for genome annotation. Furthermore, using the transcriptome data, we investigated gene expression in wild rice to enhance our understanding of gene expression supporting the core, dispensable, and private genome genes. In our copy number variation (CNV) analysis, we assessed whether gene expression is influenced by copy number variations. We identified 207 genes related to important agricultural traits (Supplementary Data 12). Additionally, we calculated the expression of NLR genes in both wild and cultivated species. Taking all of the above into consideration, we believe we have effectively utilized the RNA data, including both our manuscript data and publicly available data.

2. Most of the findings, including allelic and regulatory element variations, are based

purely on sequences, and there is a lack of genetic evidence using intercross populations and functional validation through transgenic experiments. Without substantial results, the conclusions should be provided more carefully.

Response: Thank you for point out these questions. Our manuscript presents the generation of reference-level genomes of 13 wild rice accessions (including B, BC, C, D, E, F, and G) using HiFi and Hic sequencing. The rice genus super pangenome is an important resource for utilizing useful alleles in wild rice gene pools.

However, we indeed lack of the genetic evidence for the gene functions. Hence, After Comprehensive consideration, the conclusions for the function gene had been removed in this revision.

3. The investigation of genomic heterozygosity, phasing, and intraspecific variation is completely lacking.

Response: Thank you for your suggestion and help us improving our manuscript. We have investigated the genomic heterozygosity in supplementary Table 2 by using GCE software (Table R4).

Additionally, we performed haplotype assembly using Hifiasm with parameters -h1 -h2 -ul, producing one primary genome and two haplotype draft contig genomes, the valid interaction pairs of HI-C data were used for further assembled to primary, Hap1, and Hap2 contigs genomes. The primary assembly chromosome ID and orientation were adjusted in alignment with the R498 rice genome by RagTAG. 3D-DNA was used to connect and order contigs forming pseudomolecules for the Hap1 and Hap2 genomes. The information regarding intraspecific variation information was listed in Table R5.

Table R4: Genome heterozygosity of the 13 wild rice.

Number	Rice species	Genome heterozygosity(%)
W-18	O. alta	1.0396
W-26	O. australiensis	0.31
W-59	O. brachyantha	0.073
W-71	O. eichingeri	0.1116
W-97	O. glumaepatula	0.1148
W-100	O. grandiglumis	0.147
W-171	O. malampuzhaensis	0.2033
W-190	O. minuta	0.1553
W-323	O. officinalis	0.5927
W-363	O. punctata	0.133
W-375	O. rhizomatis	0.5141
M-6	O. meyeriana	0.2993
W-128	O. latifolia	0.1905

Table R5. The intraspecific variation between the 13 wild rice assemblies (19 diploid genomes).

Species	Syntenic regions(Number)	Syntenic regions(Hap1/Hap2,Mb)	Inversion(Number)	Inversion(Hap1/Hap2,Mb)	Translocations (Number)	Translocations (Length,Mb)	SNP	Insertions(number/Length,Mb)	Deletion(number/Length,Mb)
O. glumaepatula	120	342.35/337.69	0	/	18	0.82/0/82	31756	4480/0.48	2477/0.50
O. punctata|B_t	36	430.81/428.71	0	/	10	0.77/0.77	18870	1169/0.06	1059/0.003
O. punctata|C_t	45	485.37/481.96	1	0.10/0.10	17	1.49/1.31	53315	2224/0.005	2149/0.005
O. malampuzhaensis|B_t	59	460.03/455.91	5	0.45/0.45	23	1.30/1.30	282151	13304/0.05	16706/0.05
O. malampuzhaensis|C_t	65	501.81/498.59	3	0.08/0.08	18	0.89/0.89	153079	5992/0.02	6214/0.02
O. minuta|B_t	50	433.21/432.03	1	0.008/0.008	18	0.48/0.48	191158	46162/0.06	18281/0.06
O. minuta|C_t	93	528.61/524.67	1	0.11/0.09	167	35.27/35.12	507967	107120/0.14	41279/0.14
O. rhizomatis	1446	568.03/563.25	55	7.90/8.13	425	3.50/3.47	2257727	138972/8.27	125527/8.62
O. officinalis	2372	526.63/527.29	68	17.64/17.91	811	7.27/7.28	2399716	159186/14.09	135082/14.16
O. eichingeri	2086	339.87/336.68	112	31.18/19.43	1140	18.11/17.03	4064028	358214/6.93	289957/7.46
O. alta|C_t	1150	370.39/376.19	88	51.28/43.00	394	2.83/2.77	2993400	303809/15.30	229606/14.59
O. alta|D_t	1657	375.11/366.26	86	37.38/36.93	654	4.50/4.49	3504563	307778/19.75	253034/19.20
O. grandigumis|C_t	81	424.63/423.63	0	/	28	2.02/2.03	46815	3298/0.02	2651/0.01
O. grandigumis|D_t	84	422.28/419.39	2	1.38/1.36	24	1.18/1.20	53563	3398/0.11	2907/0.44
O.latifolia|C_t	80	545.16/539.79	0	/	18	1.39/1.39	20050	1471/0.003	904/0.06
O.latifolia|D_t	89	504.49/500.25	1	0.02/0.02	51	43.25/42.14	14451	875/0.002	674/0.002
O. australiensis	948	827.93/841.86	37	14.97/19.13	209	22.01/21.91	1100303	103936/8.26	90954/ 3.28
O. branchyantha	17	150.29/158.92	3	8.57/8.57	2	0.003/0.003	9424	638/0.0001	792/0.85
O. meyeriana	2116	501.91/493.93	167	145.58/123.67	878	29.27	10524288	548069/29.05	413047/31.67

Point-by-point response

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

I am very glad to see that the current manuscript has been largely improved. Most of my comments on the last version of the manuscript have been well addressed. However, there are still some small issues should be fixed.

1. The authors claimed that "The NLR identification in this study was used the combination of denovo annotation of the genomes and RGAugry software" as revealed in their response to Reviewer #1. But I don't find such description under the subtitle of "The analysis of pan-NLRome" in the M&M section. The de novo annotation of NLR genes should be updated with details in the M&M section, accordingly.

Response: We thank you for pointing out this issue. We have added the descriptions of the de novo annotation of NLR genes in the M&M sections in line 275 to 278 in this revision.

2. The authors claimed that they used the Nipponbare genome (Qin et al., 2021) as the reference. However, Qin et al. 2021 did not assemble the genome for Nipponbare, instead, they just used the IRGSP-1.0 (Kawahara, Y. et al., 2013), which is a high-quality assembly of Nippponbare genome, for their analysis. Thus, it is not suitable to cite Qin et al., 2021 here. This also suggests that the authors may lack sufficient

investigation on the rice genomics. Please check throughout the manuscript to avoid such common-sense mistakes.

Response: We apologize for the mistake. We agree with the reviewer that we used the IRGSP-1.0 (Kawahara, Y. et al., 2013) as the reference. However, we used Qin's re-annotated genome file to construct a gene-based super pangenome and we described it in the "Analyses of the protein-coding-gene based *Oryza* super pangenome" section of M&M. We had missed the cite of IRGSP-1.0 (Kawahara, Y. et al., 2013) and have corrected it in line 78 and 89 in the revised manuscript and in line 202, 219, and 238 in the revised M&M. We have also revised the cite in the manuscript throughout in this revision. Finally, we sincerely appreciate your rigorous attitude towards scientific research, which helped us to further improve the manuscript.

3. Line 61, "A total of 20045 gene families". The thousandth separator is missing here.

Response: We have corrected it and carefully checked the whole manuscript to revised such mistakes in the revised manuscript.

4. The Latin names of some species are inconsistently written, e.g., the abbreviation of Oryza, O., was used in Lines 678-680 before the use of its full name in Lines 683-685 (the caption of Fig. 4). Please check throughout the manuscript to revise such typesetting errors.

Response: Thank you for your detailed and careful reviews. We have corrected it in line 743-745 and checked throughout the manuscript and supplementary files. And we thank you again for improving our manuscript.

Reviewer #2 (Remarks to the Author):

Based on the response to reviewers I can see the authors have made the revisions I requested in the first version. A cursory look over some of the other edits however finds a few errors and as such I think a small revision is still required:

1. Oryza sataiva spelled incorrectly in Supp Figure 4. Why is diploid punctata not included in Supp Fig 4 and only in the response to reviewers? Also needs bootstrap values adding.

Response: Thank you for your constructive suggestion. In our opinion, it is appropriate to correspond to the sample which used in the nuclear genome phylogenetic tree construction, so we did not add the bb genome. We have added the bootstrap values in Supplementary Fig 4 in this revision.

2. A small point but for accuracy use the same number of decimal places for a set of data, for example in table R2 (Supp Table 1) all values in a column should have the same number of dp (e.g. 93.0 not 93).

Response: Thank you for your great suggestion, very appreciated. We have corrected it in Supplementary Table 1 and checked throughout the manuscript.

3. Is table R5 in the main ms or only in the response. It seems important to include, but it is currently very confusing. For example “syntenic regions” and “inversions” – compared to what? Similarly “SNP” – what about them? Again if its compared to NIP

how useful is this, especially for the very distantly related species? See also point above about being consistent in the number of dp among individuals in a column.

Response: We apologize for the confusion. The R5 data were exclusively used to respond to Reviewer 3. The information presented in Table 5 illustrates the variation between the two haplotype assemblies of the 13 wild rice species. We have thoroughly checked and revised the number of data points among individuals in each column throughout the manuscript.

Reviewer #3 (Remarks to the Author):

The manuscript has been enhanced based on the reviewer's feedback. One additional recommendation: include the accession number for each RNA-seq dataset in Supplementary Data 3, and ensure that all genomic (raw reads and genome assemblies) and RNA-seq data are publicly accessible and thoroughly described.

Response: Thank you for your recommendation. We had updated the accession number for each RNA-seq dataset in Supplementary Data 3. And the genomic and RNA-seq data were share to public according to your suggestion.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

This manuscript entitled “Genome Evolution and Diversity of Wild and Cultivated Rice Species” reported thirteen representative wild rice species from the Oryza genus and constructed a super pangenome by integrating them with four previously reported genomes in this genus. It discussed extensive topics including genomic and pangenomic features, phylogenetics, structural variations and variable genes, TEs and other functional elements, which would be much valuable for rice researchers and breeders with amount new resource to improve cultivated rice and utilize wild rice germplasm. However, there are amounts of issues in the results and methods sections, which makes me very concerned about the reliability of these results.

major points:
1. For the sampling of different wild rice species, why a BBCC type of Oryza punctata was selected? Usually, a typical O. punctata sample is considered as an BB type, although there exist other types.

2. Although the authors constructed the phylogenetic tree of Oryza genus by distinguishing different genome types, the introgression and gene flow between different genomes may still cause strong interference in the phylogenetic relationships, which should be taken into account when constructing the phylogenetic tree.

3. The obtained phylogenetic relationships in this study is obviously conflicted with previous reported results. For the tree of nuclear genome in Fig 1b, O. glaberrima is clustered with O. sativa ssp. indica in the same clade while the two subspecies of O. sativa, O. sativa ssp. indica and O. sativa ssp. japonica are clustered in different clades. Moreover, the O. rufipogon that is commonly known as wild relative species of O. sativa, is placed as an outgroup of both O. glaberrima and O. sativa. This totally contradicts the consensus on the origin and evolution of Asia and African cultivated rice over the past half century.

Thus, the supporting rate for different topologies of species tree could be invalid and the authors should check about their methods related to this part. Additionally, two supporting rates (44% and 46) very close to 50% on branch 2 suggest that the phylogenetic relationship here is not resolved, and also suggest a potential risk in the calculation method.

For the tree of chloroplast genome in supplemental figure 3, it is also conflicted with previous reported results, e.g., the O. punctata should be the closest lineage to the AA-genome species. And two important species, O. rufipogon and O. glaberrima, are absent in this tree.

4. The large inversion in Chr6 is detected in both O. sativa ssp. japonica and O. rufipogon as revealed by Fig 2a. As we know, the O. rufipogon is mainly distributed in South and

Southeast Asia. The japonica rice consists of tropical japonica rice, subtropical japonica rice and temperate japonica rice, etc. But the Lines 223-225 described that “The inversion occurred only among the common wild rice and Oryza sativa japonica with high latitude distribution, indicating that it has risen to expansion of geographical range.” This causal relationship lacks evidence and may even be incorrect. It also affects the significance and necessity of the subsequent analysis on the gene OsMFT1.

5. Fig 3a presented the function genes with CNV mutations, however, all genes showed more than two copies in the Nipponbare genome but present-or-absent variations in other species. If this gene dataset can be easily obtained by selecting genes with multiple copies from Nipponbare genome, what is the significance of constructing a super pangenome? The authors should recollect a more meaningful gene dataset to conduct an in-depth analysis on the CNV of important genes across the wild and cultivated rice species, highlighting the advantage of this super pangenome.

Furthermore, the Fig 3j showed a more similar collinearity of NLR genes between O. sativa ssp. indica and O. sativa ssp. japonica than that between O. sativa ssp. indica and O. glaberrima, indicating a closer phylogenetic relationship between the two subspecies of O. sativa than that between indica and O. glaberrima. This result is conflict with the results of phylogenetic analysis in this study but is consistent with the current consensus on the origin and evolution of Asia and African cultivated rice.

Additionally, the authors clarified that “Analysis of NLR distribution showed that while R gene singletons were similar between wild and cultivated rice, cultivated rice tended to have a higher number of R genes in pairs or clusters compared to wild rice” at Lines 390-392. However, considering that most of the existing rice NLR gene data sets are based on studies on cultivated rice genome, this may also be the case that NLR in wild rice genomes has not been fully identified.

Minor points:
1. Line 64: lack of species name of Asian cultivated rice, Oryza sativa.

2. The Latin names of indica and japonica rice are not properly written, e.g. should be Oryza sativa ssp. indica and Oryza sativa ssp. japonica. Please check through the manuscript and all the figures and tables.

3. Line 97: the reference 9 (Qin et al., 2021) was used for Nipponbare genome, but actually this study used the IRGSP-1.0 (Kawahara, Y. et al., 2013) as shown in the M&M section.

4. Extended Data Fig. 3c, the second member of dispensable family showed a presence in all types of genomes, which is identical with the core family.

5. The chromosomes of different species were drawn with lengths not consistent with their genome size, e.g., in extended Data Fig. 6 and 7, supplementary 7 and 10.

6. The dotplot of pairwise alignment for Nipponbare genome and other genomes were drawn with a reverse orientation and should be corrected.

Reviewer #2 (Remarks to the Author):

Sequencing the genomes of crop wild relatives is very important for identifying novel and adaptive variation, especially in the light of combating climate change. The paper by Long et al. is an excellent example of the types of study that should be undertaken. I have some suggestions for improvements, some I would suggest are vital and others are smaller or optional, as follows:

Major changes:
1. Figure legends need significant improvement. For example, for fig1:
“structure, panicle” > “structure of the panicle”. Also the panicle panel doesn’t have the species names indicated. If they are in the same order as the whole plant photos then say this.
“TE content in each subgenome/genome of the wild and cultivated rice”. > what about it? – say it is indicated on the right of the phylogeny
“The order of species is corresponding to the phylogeny tree”. > use “phylogenetic tree” or “phylogeny” and I don’t understand why you need to say this statement, the species are clearly in phylogeneticv order, that’s the point of a phylogeny.
“Proportions of contrasting gene tree topologies for the 100 genes with regard to two major conflicting relationships”. > indicate these as panels c and d.
All figure legends need work.
2. Present the BUSCO analysis for each sample, not just the average. As allotetraploids were involved it would be very interesting and important to be able to see the BUSCOs.
3. The orthofinder analysis would give very different results if different settings were used. How were the settings chosen selected? The very high number of orthogroups (101k) from comparing many genome but each only containing 30k genes suggests a very large number are single copy and species-specific, which is what is reported. However I doubt it is true that 1/3 of genes are truly species-specific, and certainly relaxing the settings would reduce the 101k and reduce the number of species specific. So simply stating there are 101k orthogroups is premature and not validated. Only 10% of genes being core genes suggest to me the settings were too strict.
4. The section on TEs (from line 154) is very descriptive – I suggest adding some statistics to back up the statements, eg you say there is an increase or decrease in some TE family sixed but is this just random variation across the phylogeny or are there significant increases in some species/groups relative to others?
In the same section, at line 178 the authors state “they influenced a certain proportion of genes expression” but I don’t see evidence for this. Extended data Fig 5c, Supplementary Fig. 7b do not have anything to do with expression.

Also, line 181 says TEs are related to centromeres, but in the figure the centromeres are not indicated so how is there evidence for this statement?
5. The section on synteny is unusual because everything is compared to NIP. So therefore statements such as “The AA genomic type genome displayed a significant level of chromosomal conservation.” is biased –of course it will show synteny because its compared to NIP which is AA too. This statement is not backed up unless B is compared to B, C to C and D to D etc. Related, “the gene synteny findings provide support for the phylogenetic position of the rice genus (Supplementary Fig. 10).” – this figure does not tell me anything about the phylogeny. It makes sense that related species have similar genomes, but unless this is specifically being tested, then remove this statement.
6. lines 262 – 274 – without any stats this can just be random variation between species. This must be added to allow for these statements to be made. Line 404, again without stats this can be random variation. Stats must be added.
7. There are several analyses that are poorly explained or, in my opinion, irrelevant, for example:
Line 342-7 – this seems like any genome region could have the same properties – this statement simply says that the different species have different SVs. But nothing about selection during domestication. In addition, it is already known that Rc controls pericarp colour, so what is the need to focus on it here? > this analysis doesn’t seem to be meaningful and is not well-explained.
L382 – “This suggests that the immune system in wild rice has a more diverse evolutionary history compared to cultivated rice.” – the comparison is two cv and many wilds, it is therefore not surprising there is this difference. Also the cv rice is diploid and the wild contains polyploid therefore at least double the number of genes is expected in punctata. > this analysis is not meaningful.

Smaller comments:
Line 39 – expansion of range in which species? Osat? or both Osat and Orufi?
Line 43 – I don’t agree with the term ‘hidden legacy’ and I think it should be removed here.
Line 52 – state these are genome groups, because as written it is discussed like 6+5 species (but 20 species). Same for lines 61-2 where you say 11 species and list six genomes.
You say there is a lack of genome for BB but you didn’t sequence a BB genome species according to Supp table 1, correct?
The introduction appears to brief in my opinion. Some more about what each of the species being sequenced brings would be appropriate, eg “species X (BB genome) is tolerant of X” and so on. Just a couple of sentences would help.
Several of the figures state the species name but a lot of the text refers to genome groups. It would be good for genome group to be in parentheses on some figures, especially figure 1.
Line 89 – how does supp figure 1 relate to this statement about the quality of the genomes?
93-4 – 37000 is not higher than 41000 – please reword this sentence.
Line 98 – “contradicted” in what way? Where are the contradictions? Why were 1000 of the 3555 used in the analysis not all 3555? This refers to fig 1b but see comment above

about naming panels c and d.
Line 102 – state what was found too, don't leave it to the reader to download the supp figure. Does this match previous work on the cpDNA genome?
Lines 123-8 – it is confusing to sometimes use the number and sometimes the percentage. Please clarify this.
Line 131 remove "approximately" this number is not an approximation. In this section it would be interesting to see what percentage of core, disp, and sp-sp genes are supported by transcriptome data.
136 – you mention length then Ka/Ks. This is confusing.
140 – remove "to adapt to diverse environments" this is complete speculation.
150 – what do you mean "except for"?
156 – "Scientists have been intrigued by" is not a good phrase. Remove
172 – "puffy"? this is definitely not the correct word.
224 – "it has risen to expansion of geographical range" is pure speculation and not backed up by any data, only a correlation. Genetic drift is equally plausible.
276 – how does "substantial variation in SV sizes" indicate "an enrichment of SVs in repetitive DNA regions"?
Line 288 – "This gene might be de novo birth to contribute to the ability of wild rice to adapt to poor and problem soil in Sri Lanka." – this statement is confusing but also completely speculative. What does the analysis indicate the function of this gene is? Is it supported by expression data?
Line 290 – "The phylogenetic results revealed that the gene originated from O. eichingeri and then transferred to O. punctata, providing evidence that O. eichingeri was the progenitor of tetraploid O. punctata, consistent with our chloroplast evolution results." > where is the evidence it was transferred to punctata? Isn't this simply because punc couldn't have evolved without eich? How does it back up the cpDNA results, these are completely different things. The parents are shown by the nuclear genome and the female parent by the cpDNA. The cpDNA has nothing to do with determining the parents. Surely the parents were already known too prior to this analysis. This could just be removed, it doesn't mean anything.
L296 – 301 – picking one gene out of a QTL region is again entirely speculation that it is involved in resistance. What is the evidence that this gene is the likely causative gene? How many genes were in this region? The gene structure looks very different between sat and off, is this really an orthologue?
L302-6 – I have no idea what is the relevance of this.
L311 – what are "whole-genome alleles"?
L315 – "decreased, ranging from 18,463 to 23,812," – this sounds like an increase. Please reword.
L333-7 – I don't understand this. It sounds like the genes in collinear regions were translated and their domains compared. But then 7 clusters were found. Surely there are more than 7 clusters of proteins. Do you mean 7 members per cluster? Ext Fig 11d does nothing to help me understand.
Line 342-7 – this seems like any genome region could have the same properties – this statement simply says that the different species have different SVs. But nothing about

selection during domestication. In addition, it is already known that Rc controls pericarp colour, so what is the need to focus on it here?
L382 – “This suggests that the immune system in wild rice has a more diverse evolutionary history compared to cultivated rice.” – the comparison is two cv and many wilds, it is therefore not surprising there is this difference. Also the cv rice is diploid and the wild contains polyploid therefore at least double the number of genes is expected in punctata.
Nip has been sequenced before. How does your Nip genome compare to the previously sequenced Nip?

Reviewer #3 (Remarks to the Author):

The paper describes the generation of reference-level genomes of 13 wild rice accessions (including BB, BBCC, CC, EE, FF, and GG) using HiFi technology. The genome assemblies are generally of high quality and will be an important resource for utilizing useful alleles in wild rice gene pools. However, there are several major concerns about the work, which are listed below:

1. RNA sampling of diverse tissues is critical for genome annotation. Currently, the data seem to not have been comprehensively collected and are not well utilized.
2. Most of the findings, including allelic and regulatory element variations, are based purely on sequences, and there is a lack of genetic evidence using intercross populations and functional validation through transgenic experiments. Without substantial results, the conclusions should be provided more carefully.
3. The investigation of genomic heterozygosity, phasing, and intraspecific variation is completely lacking.

RESPONE TO REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

This manuscript entitled "Genome Evolution and Diversity of Wild and Cultivated Rice Species" reported thirteen representative wild rice species from the Oryza genus and constructed a super pangenome by integrating them with four previously reported genomes in this genus. It discussed extensive topics including genomic and pangenomic features, phylogenetics, structural variations and variable genes, TEs and other functional elements, which would be much valuable for rice researchers and breeders with amount new resource to improve cultivated rice and utilize wild rice germplasm. However, there are amounts of issues in the results and methods sections, which makes me very concerned about the reliability of these results.

RESPONSE: Thanks for your valuable comments and suggestions that helping us to improve the manuscript. We have carefully revised and improved the manuscript.

major points:

1. For the sampling of different wild rice species, why a BBCC type of Oryza punctata was selected? Usually, a typical O. punctata sample is considered as an BB type, although there exist other types.

Response: Thanks for bring up this point. We agree BB genome type is more typical for

the *Oryza punctata*. However, we haven't collected the BB genome type. To avoid confusion, we have labeled our *O. punctata* samples as "*O. punctata* (BBCC, 4X)" throughout our manuscript.

2. Although the authors constructed the phylogenetic tree of Oryza genus by distinguishing different genome types, the introgression and gene flow between different genomes may still cause strong interference in the phylogenetic relationships, which should be taken into account when constructing the phylogenetic tree.

Response: Thanks for your comment. The incomplete lineage sorting (ILS), horizontal gene transfer, natural selection, evolutionary radiation, and introgression caused by hybridization should be taken into account when constructing the phylogenetic tree.

In this revision, we applied a multi-species concatenation-based method incorporated in ASTRAL (v.5.7.8) which takes ILS into account to generate a species tree. ASTRAL took 3,555 single-copy gene trees as input and generated a species tree estimated by searching for the species tree that was most congruent with quartets garnered from the input gene trees.

As suggested by reviewer, we have corrected the inaccurate placement of AA genome species and updated the version of *Oryza* genus phylogenetic tree in Fig. 1b in revised manuscript (Fig. R1).

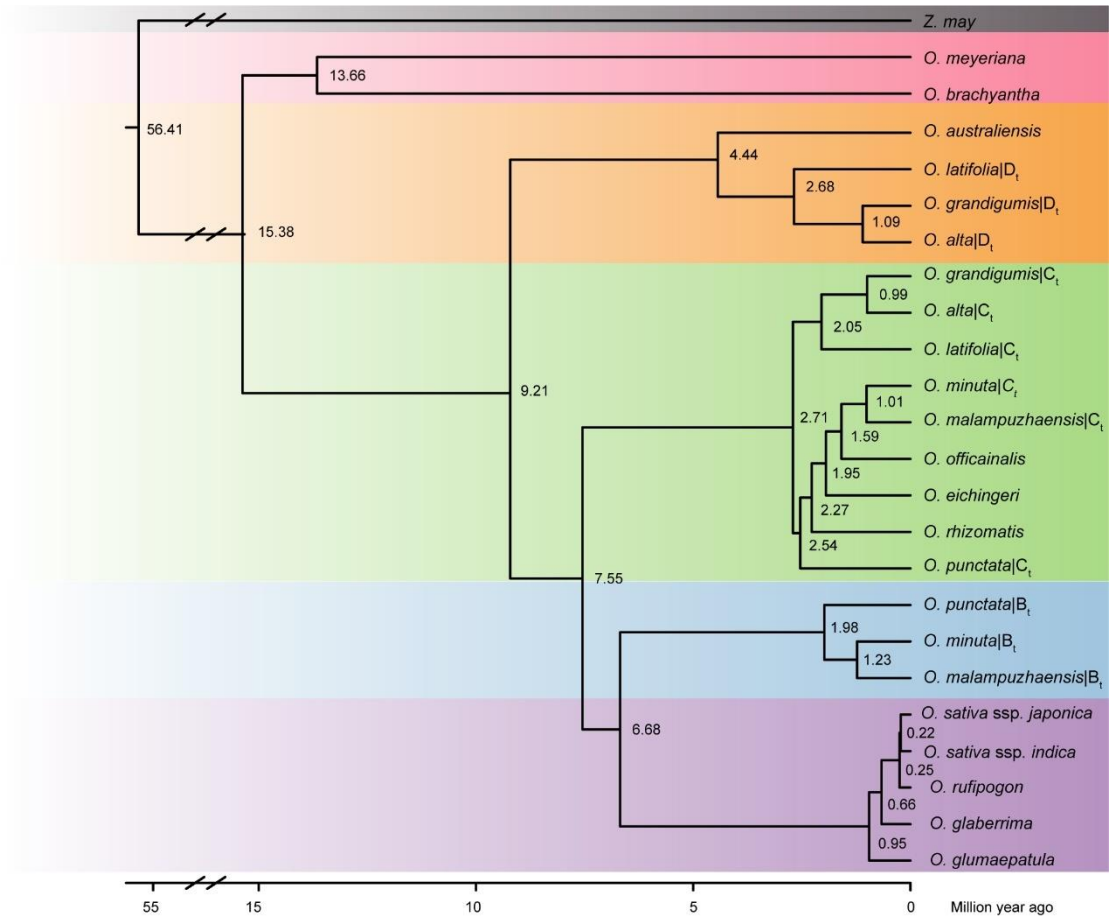


Fig. R1: Phylogenetic trees of the *Oryza* genus. allotetraploid wild rice genomes were split into sub-genomes to construction and the single copy genes were concatenated to generate the phylogenetic tree. Numbers at nodes indicate the median value for the divergence time (Mya) estimates for each clade.

3. The obtained phylogenetic relationships in this study is obviously conflicted with previous reported results. For the tree of nuclear genome in Fig 1b, O. glaberrima is clustered with O. sativa ssp. indica in the same clade while the two subspecies of O. sativa, O. sativa ssp. indica and O. sativa ssp. japonica are clustered in different clades. Moreover, the O. rufipogon that is commonly known as wild relative species of O. sativa, is placed as an outgroup of both O. glaberrima and O. sativa. This totally contradicts the consensus on the origin and evolution of Asia and African cultivated

rice over the past half century.

Response: Thank you very much for the careful review and detailed comments. We apologize for the error caused by simply selected 1000 out of 3500 genes for tree construction and performed only 100 bootstraps in previous analysis. The 22% supported rate in the right part of figure 1B (last version) was consistent with previous reported results. In this revision, we have now redone the analysis by reconstructing the phylogenetic tree based on the concatenation of 3555 single copy gene sequences and got a tree that is consistent with previously reported results, and the corrected *Oryza genus* phylogenetic tree can be found in Fig. 1b (Fig. R1). And we thank you again for improving our manuscript.

Thus, the supporting rate for different topologies of species tree could be invalid and the authors should check about their methods related to this part. Additionally, two supporting rates (44% and 46) very close to 50% on branch 2 suggest that the phylogenetic relationship here is not resolved, and also suggest a potential risk in the calculation method.

Response: Thanks for your suggestions and kind reminder. We apologize for the error caused by simply selected 1000 out of 3500 genes for tree construction and performed only 100 bootstraps in previous analysis. In this revision, we have now redone the analysis by reconstructing the phylogenetic tree based on the concatenation of 3555 single copy gene sequences from 18 species (24 subgenomes, *Z. may* as outgroup), and the corrected *Oryza genus* phylogenetic tree can be found in Fig. 1b (Fig. R1).

For the tree of chloroplast genome in supplemental figure 3, it is also conflicted with previous reported results, e.g., the O. punctata should be the closest lineage to the AA-genome species. And two important species, O. rufipogon and O. glaberrima, are absent in this tree.

Response: We apologize for any inaccuracies in our description that may have led to misunderstandings. The *O. punctata* used in this study is an allotetraploid (BBCC), with the chloroplast genome derived from the CC genome. Consequently, it clusters with CC genome species, consistent with previously reported findings (Zou et al., 2015). We have made revisions in Supplementary Fig. 4, which now includes *O. rufipogon* (KF359902) and *O. glaberrima* (KF359905) (see Fig. R2a) as your suggested. Additionally, we conducted a phylogenetic analysis of the chloroplast genome, incorporating *O. punctata* (BB genome, GenBank: KF359908). The results are presented in Fig. R2b.

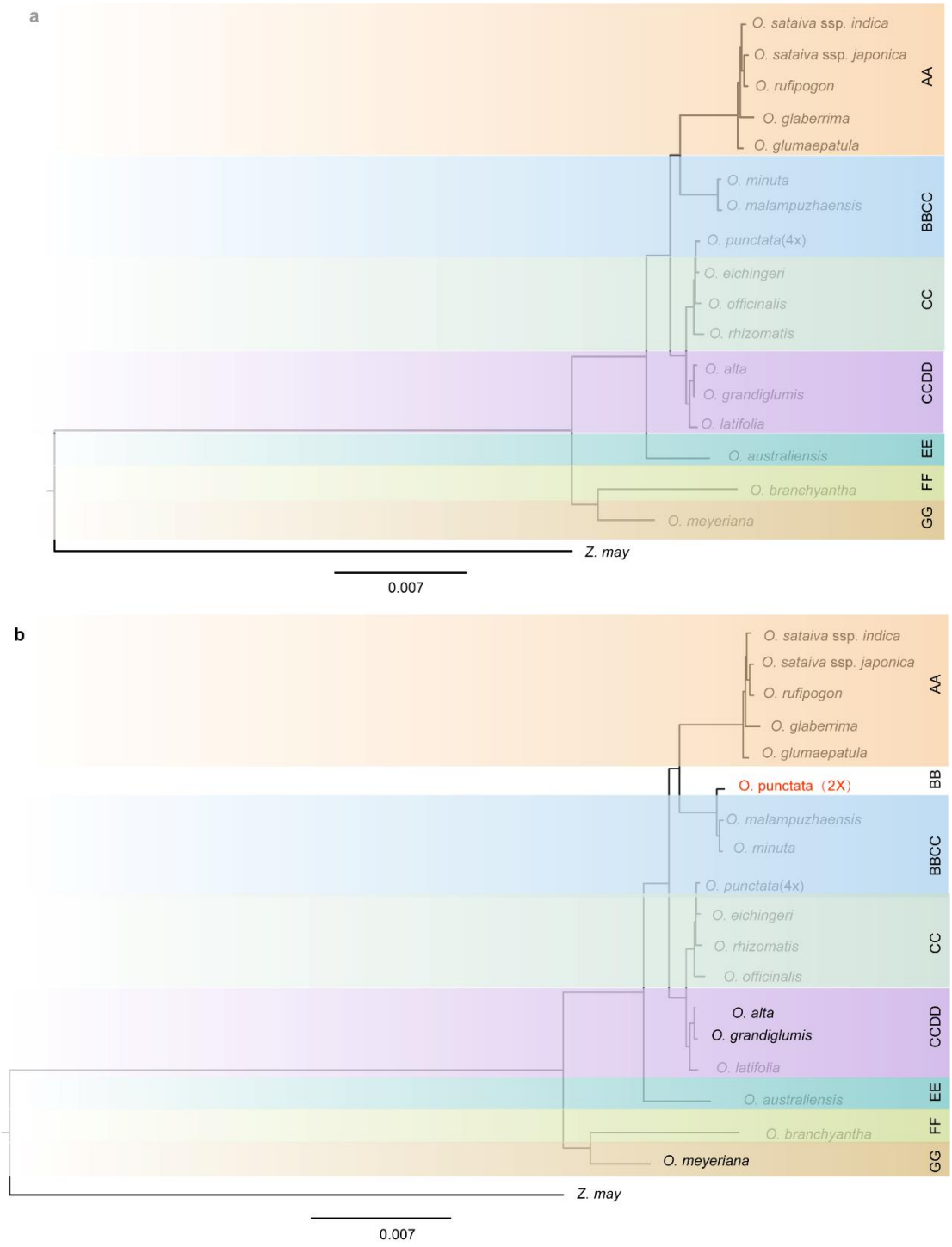


Fig R2: **Phylogenetic tree based on the chloroplast genome sequences.** a. Phylogenetic tree based on 17 wild rice species chloroplast genome sequences (Z. may as outgroup). b. Phylogenetic tree constructed based on 18 wild rice (including a BB genome type *O. punctata,* KF359908) species chloroplast genome sequences.

Zou, XH., Du, YS., Tang, L. et al. Multiple origins of BBCC allopolyploid species in

the rice genus (Oryza). Sci Rep 5, 14876 (2015).

4. The large inversion in Chr6 is detected in both O. sativa ssp. japonica and O. rufipogon as revealed by Fig 2a. As we know, the O. rufipogon is mainly distributed in South and Southeast Asia. The japonica rice consists of tropical japonica rice, subtropical japonica rice and temperate japonica rice, etc. But the Lines 223-225 described that "The inversion occurred only among the common wild rice and Oryza sativa japonica with high latitude distribution, indicating that it has risen to expansion of geographical range." This causal relationship lacks evidence and may even be incorrect. It also affects the significance and necessity of the subsequent analysis on the gene OsMFT1.

Response: Thanks for point out this problem. We agree with you. It is inappropriate to make such a conclusion "The inversion occurred only among the common wild rice and *Oryza sativa* japonica with high latitude distribution, indicating that it has risen to expansion of geographical range". We have removed these sentences in this revision.

5. Fig 3a presented the function genes with CNV mutations, however, all genes showed more than two copies in the Nipponbare genome but present-or-absent variations in other species. If this gene dataset can be easily obtained by selecting genes with multiple copies from Nipponbare genome, what is the significance of constructing a super pangenome? The authors should recollect a more meaningful gene dataset to conduct an in-depth analysis on the CNV of important genes across the wild and cultivated rice species, highlighting the advantage of this super pangenome.

Response: Thanks for your constructive suggestions. We agree that detecting function genes with CNV mutations in the *Oryza* genus super pangenome is more challenging and meaningful than that in Nipponbare. In this study, we conducted comprehensive analyses of the CNV alleles within the rice genus and examined the CNV genes associated with significant agricultural traits in rice. However, genes related to agricultural traits have been less explored in wild rice. Consequently, we divided the CNV genes into two categories: the first category comprises genes with multiple copies identified from the Nipponbare genome, while the second category was derived from comparisons across wild rice species, as these genes are absent in the Nipponbare genome. Detailed information can be found in Fig. 6 of the revised manuscript (Fig. R3).

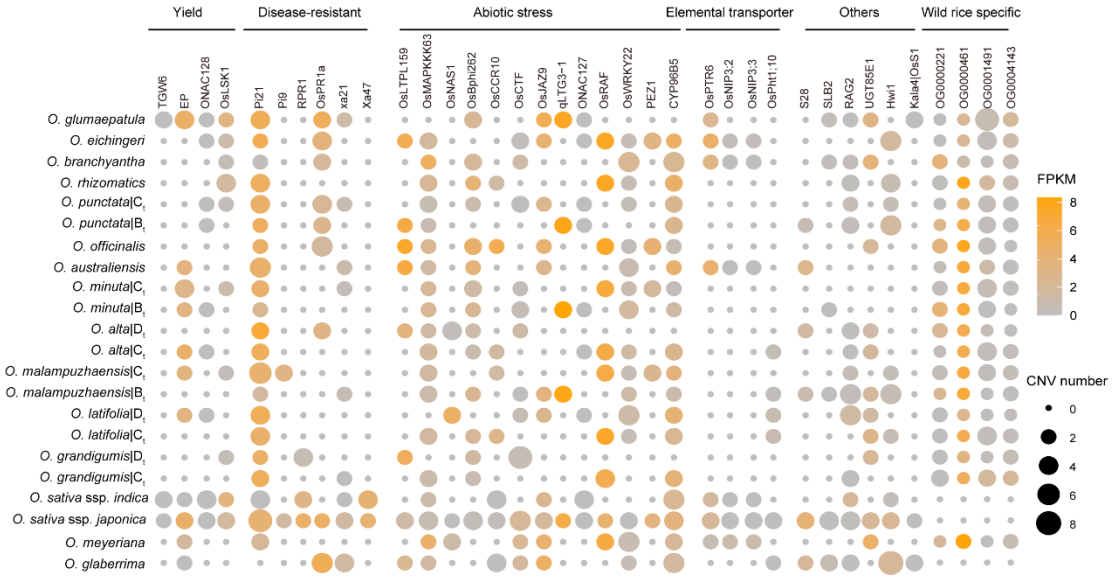


Fig. R3: Characteristics of gene CNVs related to important agronomic traits in the Oryza genus super pangenome. Circle size shows the number of gene copies, color from grey to yellow represents the gene expression.

Furthermore, the Fig 3j showed a more similar collinearity of NLR genes between O.

sativa ssp. indica and O. sativa ssp. japonica than that between O. sativa ssp. indica and O. glaberrima, indicating a closer phylogenetic relationship between the two subspecies of O. sativa than that between indica and O. glaberrima. This result is conflict with the results of phylogenetic analysis in this study but is consistent with the current consensus on the origin and evolution of Asia and African cultivated rice.

Response: Thanks for your careful and conscientious review. We agree with you that *O. sativa* ssp. *indica* and *O. sativa* ssp. *japonica* had a closer phylogenetic relationship than that between *O. sativa* ssp. *indica* and *O. glaberrima*. The error was caused by our previous phylogenetic tree construction. In this revision, we have reconstructed the *Oryza* phylogenetic tree by concatenated based methods. The revised tree now correctly places *O. sativa* ssp. *japonica* closer to *O. sativa* ssp. *indica* than to *O. glaberrima* in Fig. 6 (Fig R4). In addition, we have carefully revised all the results related to the phylogenetic tree in this article.

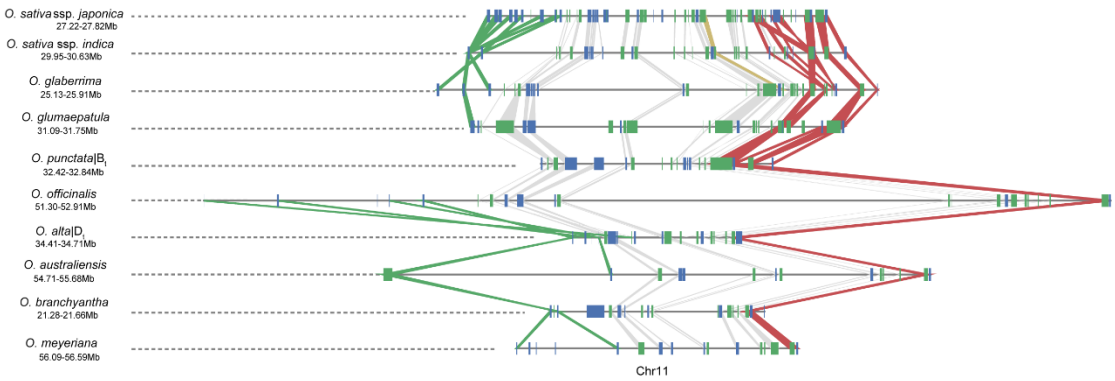


Fig. R4: Collinearity of NLRs in various *Oryza* genomes within a specific region.

Additionally, the authors clarified that "Analysis of NLR distribution showed that while R gene singletons were similar between wild and cultivated rice, cultivated rice tended to have a higher number of R genes in pairs or clusters compared to wild rice" at Lines 390-392. However, considering that most of the existing rice NLR gene data sets are

based on studies on cultivated rice genome, this may also be the case that NLR in wild rice genomes has not been fully identified.

Response: Thank you very much for the professional suggestions. We believe our analysis does provide the comprehensive NLR gene data sets in wild rice genomes. The NLR identification in this study was used the combination of denovo annotation of the genomes and RGAugry software. For genome sequence annotation, LRR domain were extracted from the denovo annotation of the genomes (Table R1). The another method was as follow: Firstly, RGAugry was widely used in R gene identification of wild rice[1], wild potatoes[2], and wild radishes[3] and obtained a comprehensive results. The total NBS gene number of *O. alta* identified in our results was content with the previous study[1]. Secondly, the NLR results was verified by using Interproscan and hmmsearch software. The both identified NLR was recorded for subsequent analysis. To obtain a comprehensive NLR in the wild rice genomes, we selected the NLR region which cultivar had more NLR than wild rice and selected one as example for validation (Figure R5).

Table R1: The denovo annotation of the genome with LRR domian

Species	LRR domain number	R gene number
O. sativa ssp. indica	789	511
O. sativa ssp. japonica	709	473
O. glaberrima	732	419
O. rufipogon	448	260
O. australiensis	432	195
O. brachyantha	472	245
O. rhizomatis	627	343
O. glumaepatula	613	354
O. eichingeri	588	288
O. meyeriana	482	238
O. alta|Ct	508	260

O. alta|Dt	458	216
O. grandiglumis|Ct	515	224
O. grandiglumis|Dt	456	183
O. latifolia|Ct	585	314
O. latifolia|Dt	460	206
O. minuta|Bt	553	258
O. minuta|Ct	593	336
O. punctata|Bt	596	320
O. punctata|Ct	623	349
O. malampuzhaensis|Bt	559	302
O. malampuzhaensis|Ct	559	277

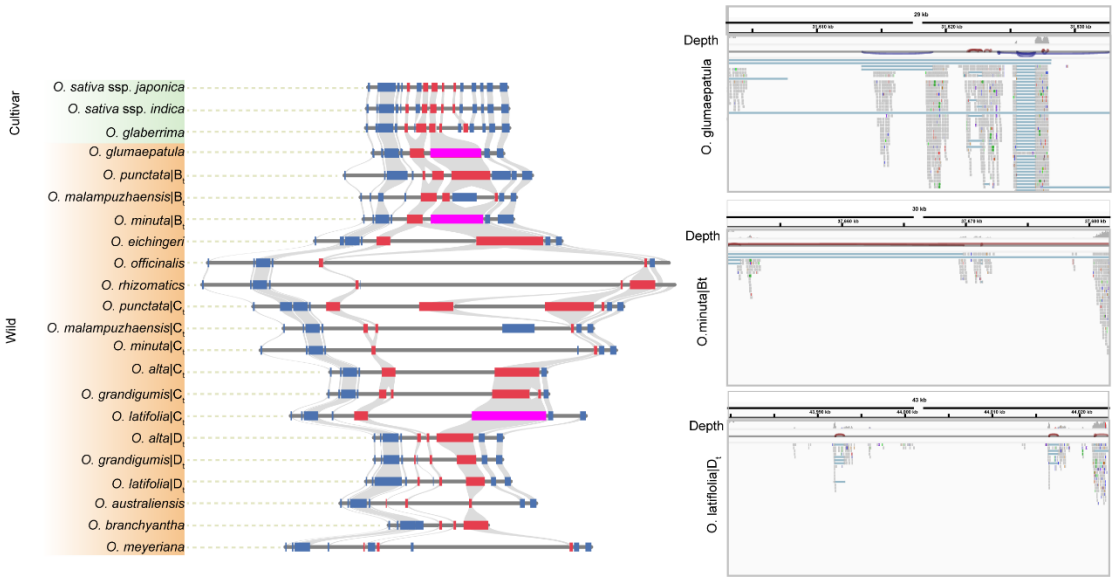


Fig. R5: A NLR synteny region and RNA evidence of the annotated NLR genes. The pink color in *O. glumaepatula, O. minuta*|B_t, *O. latifolia* NLR gene were validated by the RNA-seq in IGVtools.

Minor points:

1. Line 64: lack of species name of Asian cultivated rice, Oryza sativa.

Response: Corrected.

2. The Latin names of indica and japonica rice are not properly written, e.g. should be Oryza sativa ssp. indica and Oryza sativa ssp. japonica. Please check through the

manuscript and all the figures and tables.

Response: We have revised it through the manuscript.

3. Line 97: the reference 9 (Qin et al., 2021) was used for Nipponbare genome, but actually this study used the IRGSP-1.0 (Kawahara, Y. et al., 2013) as shown in the M&M section.

Response: Thanks for point out the questions. We used the Nipponbare genome (Qin et al., 2021) as the reference. We have corrected it in the M&M section.

Qin, P. et al. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* **184**, 3542–3558 e3516 (2021).

4. Extended Data Fig. 3c, the second member of dispensable family showed a presence in all types of genomes, which is identical with the core family.

Response: Corrected in Fig. 2c in this revision.

5. The chromosomes of different species were drawn with lengths not consistent with their genome size, e.g., in extended Data Fig. 6 and 7, supplementary 7 and 10.

Response: Thanks for your constructive suggestions. The chromosomes of different species were drawn with lengths consistent with their genome size in Supplementary Fig. 7 (Supplementary Fig. 10 in this revision). Extended Data Fig 6 was only used to show the alignment of collinear gene pairs, and does not account for chromosome lengths. It is not taken the length of chromosome into consideration. We have corrected the Extended Data 7 and Supplementary Fig 10 to Supplementary Fig 17 and Supplementary Fig 14 in line in this revised manuscript.

6. The dotplot of pairwise alignment for Nipponbare genome and other genomes were drawn with a reverse orientation and should be corrected.

Response. Corrected in Supplementary Fig. 11 in this revision.

Reviewer #2 (Remarks to the Author):

Sequencing the genomes of crop wild relatives is very important for identifying novel and adaptive variation, especially in the light of combating climate change. The paper by Long et al. is an excellent example of the types of study that should be undertaken. I have some suggestions for improvements, some I would suggest are vital and others are smaller or optional, as follows:

Thank you very much for your recognition and your valuable comments. We have responded to them point by point, and these modifications have helped us to improve the manuscript significantly.

Major changes:

1. Figure legends need significant improvement. For example, for fig1:

“structure, panicle” > “structure of the panicle”. Also the panicle panel doesn’t have the species names indicated. If they are in the same order as the whole plant photos then say this.

“TE content in each subgenome/genome of the wild and cultivated rice”. > what about it? – say it is indicated on the right of the phylogeny

“The order of species is corresponding to the phylogeny tree”. > use “phylogenetic tree” or “phylogeny” and I don’t understand why you need to say this statement, the species are clearly in phylogeneticv order, that’s the point of a phylogeny.

“Proportions of contrasting gene tree topologies for the 100 genes with regard to two major conflicting relationships”. > indicate these as panels c and d.

All figure legends need work.

Response: We agree with your suggestion. Thus, we have updated all the figure legends through the manuscript. Accordingly, we have updated the Figure 1 and modified the legend as following:

Line 626-637: “**Fig. 1: Geographical distribution and phylogeny of wild and cultivated rice.** **a** Geographic distribution of the 13 wild rice varieties and their diverse agronomic characteristics, such as plant height. The fonts colors of wild rice species corresponding to the distribution on the world map. These 13 accessions covered 13 species in the rice genus. **b** Panicle architecture of the 13 wild rice species. **c** Phylogenetic tree based on the conserved single-copy gene illustrates evolutionary history in the Oryza genus. All the allotetraploid wild rice genome was split into subgenomes, the ASTRAL concatenation-based species tree estimated by 3555 single copy gene generated by orthofinder. Numbers at nodes indicate the median value for the divergence time (Mya) estimates for each clade. **d** TE content in each subgenome/genome of the wild and cultivated rice. RNA transposons were presented as following: blue for LTR, yellow represents Helitron, pink indicates LINEs, orange shows SINEs, orange blue stand for unclassified. DNA transposons was showed as wine.”

2. Present the BUSCO analysis for each sample, not just the average. As allotetraploids were involved it would be very interesting and important to be able to see the BUSCOs.

Response: Thank for your insightful feedback. In response to your comments, we have added the BUSCO value of each subgenomes of allotetraploid in Supplementary Table

1 (Table R2). There is a higher value of BUSCO in subgenome than that in allotetraploid due to gene loss in subgenome also famous for subgenomic dominance.

we also noticed that the BUSCO value in C subgenome was significantly larger than that in D subgenome, which indicated C was more conserved than D subgenome during selection and evolution.

Table R2: Quality assessments of 6 diploid genomes.

Species	BUSCOs assessment					
	Completeness (%)	Single_Completeness (%)	Duplicated_Completeness (%)	Fragment (%)	Missing (%)	TotalNumber (%)
O. alta|Ct	94.3	85.4	8.9	0.6	5.1	1614
O. alta|Dt	89.5	86.5	3	0.9	9.6	1614
O. grandiglumis|Ct	95.1	85.9	9.2	0.7	4.2	1614
O. grandiglumis|Dt	89.7	86.7	3	0.7	9.6	1614
O. latifolia|Ct	98.1	96.1	2	0.3	1.6	1614
O. latifolia|Dt	97.9	96.1	1.8	0.6	1.5	1614
O. minuta|Bt	94.7	93	1.7	2.6	2.7	1614
O. minuta|Ct	94.1	91.9	2.2	2.8	3.1	1614
O. punctata|Bt	96.1	94.9	1.2	1.9	2	1614
O. punctata|Ct	96.8	94.4	2.4	1.4	1.8	1614
O. malampuzhaensis|Bt	95.7	93.9	1.8	2	2.3	1614
O. malampuzhaensis|Ct	97.4	94.5	2.9	0.9	1.7	1614

3. The orthofinder analysis would give very different results if different settings were used. How were the settings chosen selected? The very high number of orthogroups (101k) from comparing many genome but each only containing 30k genes suggests a very large number are single copy and species-specific, which is what is reported. However I doubt it is true that 1/3 of genes are truly species-specific, and certainly relaxing the settings would reduce the 101k and reduce the number of species specific. So simply stating there are 101k orthogroups is premature and not validated. Only 10% of genes being core genes suggest to me the settings were too strict.

Response: Thank you very much for your kind advice. The different parameters indeed influence the size of the super pangenome and its individual components. In the previous manuscript, we initially constructed the super pangenome using default parameters (e^{-3}), which generated a total of 96,962 orthogroups. This included the following counts and percentages: core gene families (10,015, 10.33%), dispensable gene families (43,270, 56.84%), and private gene families (43,667, 45.05%). Notably, the percentage of private gene families is relatively high. To achieve a more comprehensive and accurate representation of the *Oryza* genus super pangenome, we selected a stricter setting (e^{-10}) for this analysis. Although the total number of gene families in the pan-genome is slightly larger than that obtained with the default parameters, there is no significant difference in the counts of core gene families. Detailed information regarding the *Oryza* genus pangenome constructed with the default parameters is presented in Figure R6

We felt sorry for our unclear expression, we have moved the Extended Data Fig. 2 to

Fig. 2 in the main text. As the Fig. 2b shown, a total of 9,834 (9.66%) core gene families with 318,253 (33.22%) genes, 57,822(56.84%) gene families contains 599412(62.58%) genes in dispensable genome, and 34,067(33.48%) gene families includes 40115(4.18%) genes. Extended Data Fig. 2a (Fig. 2a in this revision manuscript) showed that 1/3 gene families and 4.18% gene being species-specific, we have added annotation to the Figure 2a in this revision.

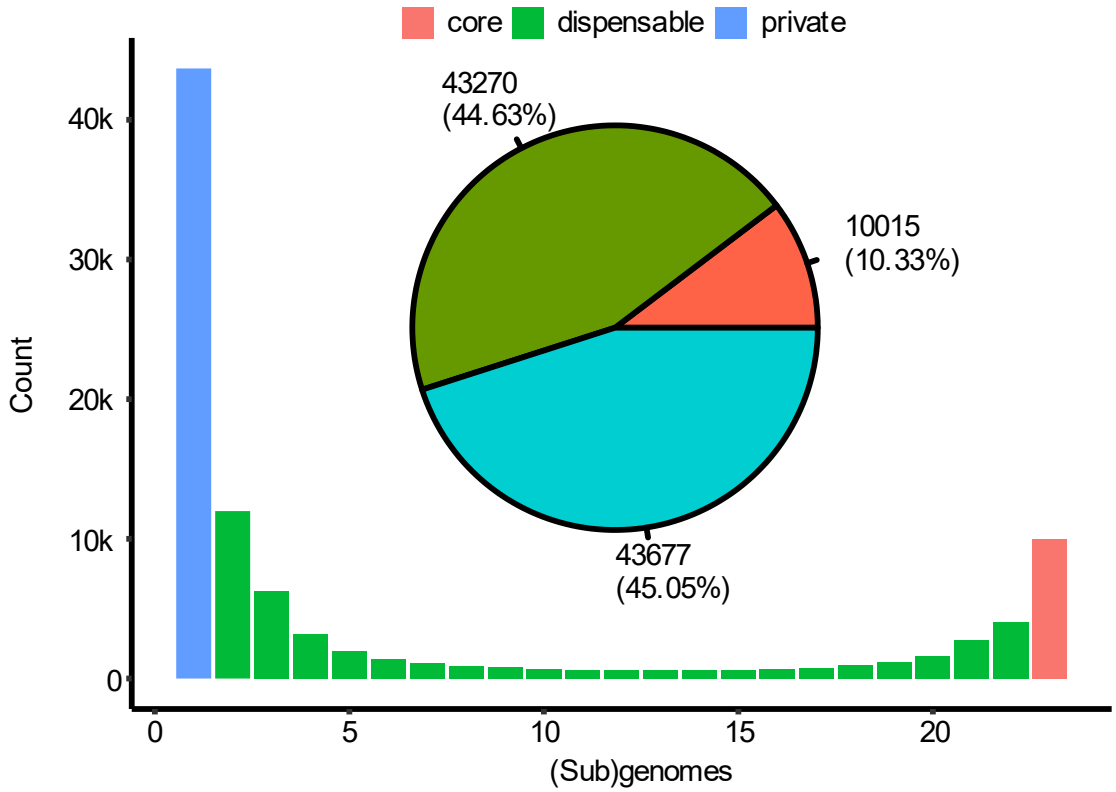


Fig. R6: The *Oryza* genus super-pangenome constructed by using default parameters.

4. The section on TEs (from line 154) is very descriptive – I suggest adding some statistics to back up the statements, eg you say there is an increase or decrease in some TE family sixed but is this just random variation across the phylogeny or are there significant increases in some species/groups relative to others?

Response: We are grateful to the reviewer for bringing up this question. We have revised

it from line 223 to line 244 in this revision.

In the same section, at line 178 the authors state “they influenced a certain proportion of genes expression” but I don’t see evidence for this. Extended data Fig 5c, Supplementary Fig. 7b do not have anything to do with expression.

A: Thank you for pointing out this concern. Previous findings indicated that gene expression levels were suppressed by the abundance and physical distance from neighboring TEs (Zhang et al., Proc Natl Acad Sci U S A, 2015; Hanada et al., Plant cell, 2009; Castanera et al., PLoS Genetic, 2016; Gagliardi et la., Proc Natl Acad Sci U S A, 2019). Therefore, in this study, TE that overlapped with gene positions were identified as influencing gene expression TE.

1.Zhang J, Liu Y, Xia EH, Yao QY, Liu XD, Gao LZ. Autotetraploid rice methylome analysis reveals methylation variation of transposable elements and their effects on gene expression. Proc Natl Acad Sci U S A. 2015 Dec 15;112(50):E7022-9.

2. Hanada K, Vallejo V, Nobuta K, Slotkin RK, Lisch D, Meyers BC, Shiu SH, Jiang N. The functional role of pack-MULEs in rice inferred from purifying selection and expression profile. Plant Cell. 2009 Jan;21(1):25-38.

3. Castanera R, López-Varas L, Borgognone A, LaButti K, Lapidus A, Schmutz J, Grimwood J, Pérez G, Pisabarro AG, Grigoriev IV, Stajich JE, Ramírez L. Transposable Elements versus the Fungal Genome: Impact on Whole-Genome Architecture and Transcriptional Profiles. PLoS Genet. 2016 Jun 13;12(6):e1006108.

4. Gagliardi D, Cambiagno DA, Arce AL, Tomassi AH, Giacomelli JI, Ariel FD, Manavella PA. Dynamic regulation of chromatin topology and transcription by inverted

repeat-derived small RNAs in sunflower. Proc Natl Acad Sci U S A. 2019 Aug 27;116(35):17578-17583.

Also, line 181 says TEs are related to centromeres, but in the figure the centromeres are not indicated so how is there evidence for this statement?

A: Corrected in Supplementary Figure 12(Figure R7) in this revision.

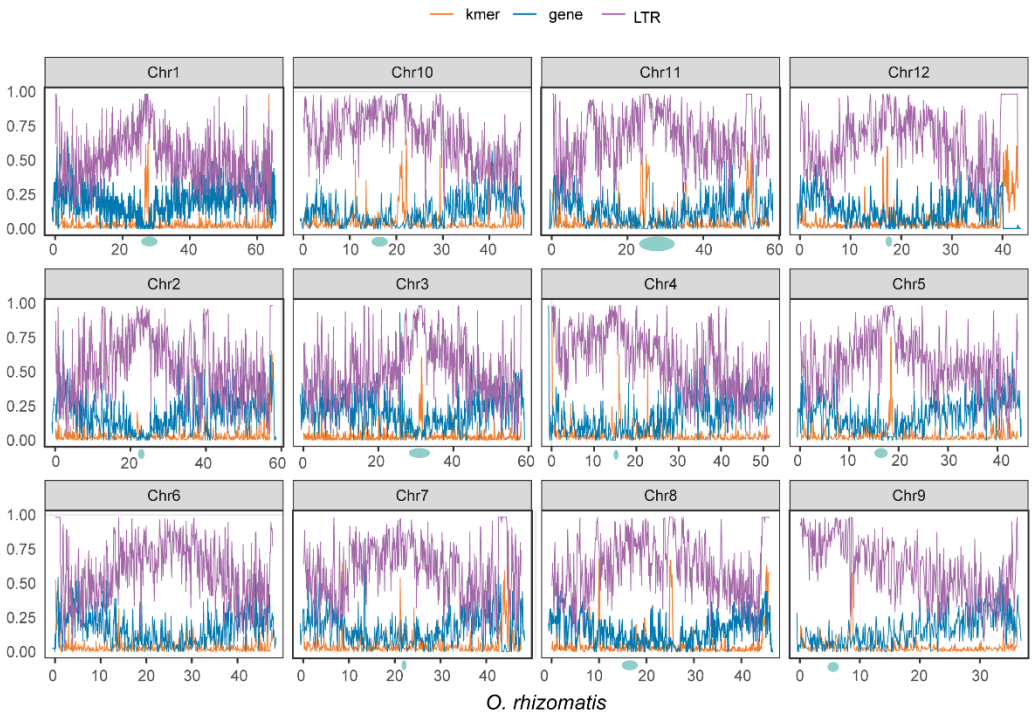


Fig. R7: Landscape of the centromere in wild rice *O. rhizomatis*.

5. The section on synteny is unusual because everything is compared to NIP. So therefore statements such as "The AA genomic type genome displayed a significant level of chromosomal conservation." is biased –of course it will show synteny because its compared to NIP which is AA too. This statement is not backed up unless B is compared to B, C to C and D to D etc. Related, "the gene synteny findings provide support for the phylogenetic position of the rice genus (Supplementary Fig. 10)." – this figure does not tell me anything about the phylogeny. It makes sense that related species

have similar genomes, but unless this is specifically being tested, then remove this statement.

A: We appreciate your suggestions. We agree with you that "The AA genomic type genome displayed a significant level of chromosomal conservation." is biased. In this study, the synteny and variation were identified compared to NIP reference genome. Hence, **we have removed these sentences in this revision**. Furthermore, **we also removed the sentence** "the gene synteny findings provide support for the phylogenetic position of the rice genus (Supplementary Fig. 10)." as you suggested. The supplementary 10 was moved in line 278 in Supplementary figure 14 in this revision to show the SV within rice diploid genomes.

6. lines 262 – 274 – without any stats this can just be random variation between species. This must be added to allow for these statements to be made. Line 404, again without stats this can be random variation. Stats must be added.

Response: We agree with your suggestion and we have rewritten the sentence in line 320-335.

7. There are several analyses that are poorly explained or, in my opinion, irrelevant, for example:

Line 342-7 – this seems like any genome region could have the same properties – this statement simply says that the different species have different SVs. But nothing about selection during domestication. In addition, it is already known that Rc controls pericarp colour, so what is the need to focus on it here? > this analysis doesn't seem to be meaningful and is not well-explained.

Response: Thank you for your comments. There is no relevant with the selection during domestication. We have removed these statements in the revised manuscript as your suggested.

L382 – "This suggests that the immune system in wild rice has a more diverse evolutionary history compared to cultivated rice." – the comparison is two cv and many wilds, it is therefore not surprising there is this difference. Also the cv rice is diploid and the wild contains polyploid therefore at least double the number of genes is expected in punctata. > this analysis is not meaningful.

Response: Thank you very much for the kind suggestion. That is a mistake of writing from our side. In our opinon, *Oryza* genus can divided into 25 species, which including two cultivated species, *O. glaberrima*, and *O. satiava*. Wherea *O. sativa* contains two subspecies: *O. sativa.* ssp. *indica* and *O. sativa.* ssp. *japonica.* and the remain species all belong to wild rice. We have rewritten the sentence "This suggests that the immune system in wild rice has a more diverse evolutionary history compared to cultivated rice." to "which revealed a significant expansion of NLRs in the cultivar rice" in lines 409-410.

We identified the R gene number in two subgenomes (diploid genomes) instead of the allotertraploid wild rice (polyploid genomes). Thus, all analysis was conducted at diploid genome level.

Smaller comments:

Line 39 – expansion of range in which species? Osat? or both Osat and Orufi?

Response: We have removed these sentences according to the suggests of Reviewer 1.

Line 43 – I don't agree with the term 'hidden legacy' and I think it should be removed here.

Response: Corrected

Line 52 – state these are genome groups, because as written it is discussed like 6+5 species (but 20 species). Same for lines 61-2 where you say 11 species and list six genomes.

Response: We feel very sorry for causing your confusing due to our imprecise expression and thank you very much to point out this issue. Wild rice includes more than 20 species, and they are currently divided into 11 genome types: AA, BB, CC, EE, FF, BBCC, CCDD, HHJJ, HHKK, and KKLL. Among the wild rice, 6 species belong to AA genome type with the cultivated rice: *O. rufipogon*, *O. nivara*, *O. barthii*, *O. glumaepatula*, *O. longistaminata*, and *O. meridionalis*. And most of AA genome type wild rice had been assembled to chromosome level. Hence, we only selected *O. glumaepatula* in AA-genome type wild rice in this study. While BBCC genome types contains three species, named *O. punctata*, *O. minitus*, and *O. malampuzhensis*. *O. latifolia*, *O.alta*, and *O. grandigumis* were belong to CCDD genome type. GG genome type include two species: *O. meryeriana* and *O. granulata*.

We have revised the sentence in lines 52-53 as "These wild rice are currently divided into 11 genome types: AA, BB, CC, EE, FF, GG, BBCC, CCDD, HHJJ, HHKK and KKLL," and line 66 as "Currently, there is a lack of high-quality reference genomes for

the most of rice wild relatives" in this revised manuscript.

You say there is a lack of genome for BB but you didn't sequence a BB genome species according to Supp table 1, correct?

Response: Thank you very much for bringing this up. We have not sequenced the diploid genome (BB genome type) of the *Oryza punctata* species because we have not yet collected the sample of this diploid species. This issue has been addressed throughout the entire manuscript

The introduction appears to brief in my opinion. Some more about what each of the species being sequenced brings would be appropriate, eg "species X (BB genome) is tolerant of X" and so on. Just a couple of sentences would help.

Response: Thank you very much for your suggestions and recommendations. We have made changes to the introduction as suggested, "A few non-A genome type *Oryza* genomes have been published with a chromosome level to date. The G genome type was firstly assembled to a chromosome level by using gradient resequencing and Pacbio SMRT data and combing the positive selection revealed photosynthesis and energy production related gene might have facilitated its tolerance to shade. Allotetraploid *O. coarctata* were released with a total length of 573.4 Mb and an N50 of 23.1 Mb by using HIFI and HiC data. Yu assembled a high-quality reference genome of *O. alta* and through gene editing for important agricultural gene, completing the de novo

domestication of allotetraploid wild rice" was added into the introduction in lines 67-75 in this revision.

Several of the figures state the species name but a lot of the text refers to genome groups. It would be good for genome group to be in parentheses on some figures, especially figure 1.

Response: Corrected.

Line 89 – how does supp figure 1 relate to this statement about the quality of the genomes?

Response: Thank you for raising this question. We acknowledge the oversight in our writing. We used Merqury to evaluate completeness of assembled genomes. And the consensus quality values were listed in the Supplementary table 4(Table R3), we have corrected the Supplementary figure 1 to Supplementary table 4 in this revision.

Table R3: The Quality assessments of 13 assemblies.

No	Species	QV	Completeness
M-6	*O. meyeriana*	49.4919	75.1455
W-100	*O. grandigumis*	53.5847	99.087
W-128	*O.latifolia*	55.0804	98.6829
W-171	*O. malampuzhaensis*	54.143	99.134
W-18	*O. alta*	53.7814	81.8891
W190	*O. minuta*	54.9547	99.2397
W-26	*O. australiensis*	57.6184	95.5999
W-323	*O. officinalis*	50.3274	90.1201

W-363	*O. punctata*	55.9447	99.5892
W-375	*O. rhizomatis*	52.1117	91.9521
W-59	*O. branchyantha*	49.5088	99.5186
W71	*O. eichingeri*	54.4785	68.6229
W-97	*O. glumaepatula*	48.3143	99.2445

93-4 – 37000 is not higher than 41000 – please reword this sentence.

Response: Corrected.

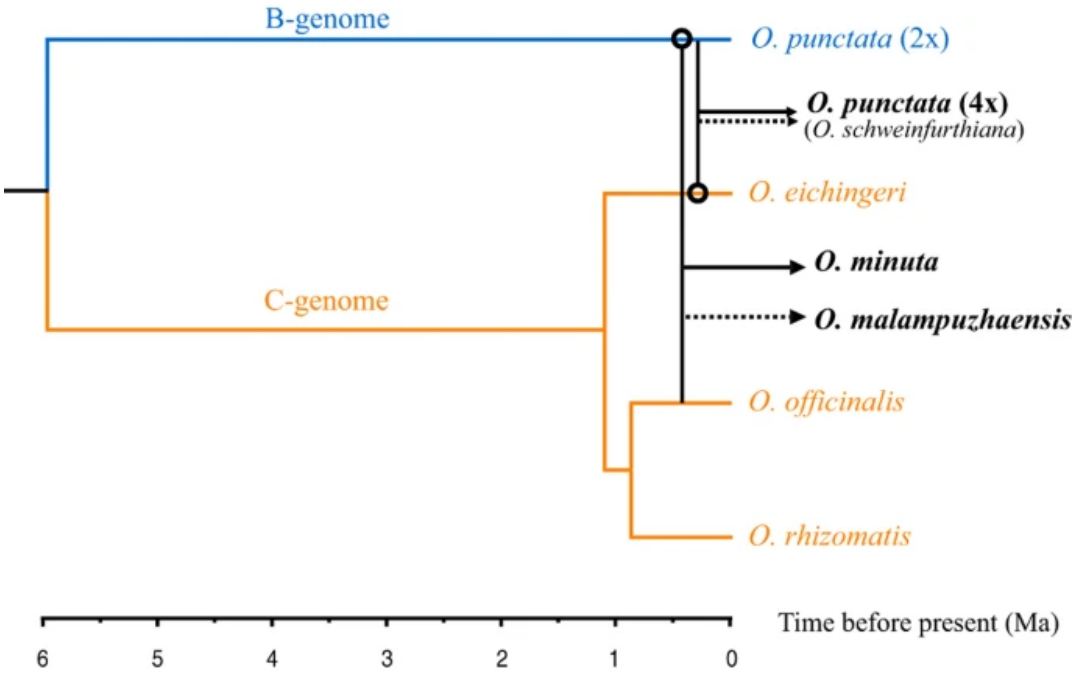
Line 98 – "contradicted" in what way? Where are the contradictions? Why were 1000 of the 3555 used in the analysis not all 3555? This refers to fig 1b but see comment above about naming panels c and d.

Response: We apologize for the error in the phylogenetic tree. We have reconducted the phylogenetic tree using all the 3555 single copy orthologs identified by Orthofinder. The revised phylogenetic tree now was consistent with previous results. In previous study, we want to save the save time and computing resources, then we simply select 1000 of the 3555 genes and bring such an error in the phylogenetic construction. We have revised it in Figure 1 in this revised manuscript.

Line 102 – state what was found too, don't leave it to the reader to download the supp figure. Does this match previous work on the cpDNA genome?

Response: Thank you very much for your nice advice. We have revised it to "revealed *O. eichingeri* is the female parent of *O. punctata* (BBCC) (Supplementary Fig. 2)." In

lines 148-149 in this revision. Our findings were consistent with the results of the previous study (Zou et al., Sci Rep, 2015). The reference can be found as followed:



Zou, XH., Du, YS., Tang, L. et al. Multiple origins of BBCC allopolyploid species in the rice genus (Oryza). Sci Rep 5, 14876 (2015).

Lines 123-8 – it is confusing to sometimes use the number and sometimes the percentage. Please clarify this.

Response: We have corrected it to “Additionally, a syntenic pangenome was constructed to analyze the differentiation among the 7 diploid genomes, revealing core gene families shared across these genomes accounted for 17.37% of total gene sets, Dispensable families, accounted for 29.73% of the total gene sets (Supplementary Fig. 6, Supplementary Table 6), with the largest proportion represented by genome type private gene sets, unique to individual genome types, making up 52.90% of the total gene sets (Supplementary Table 6).” in lines 172-178 in the revised manuscript.

Line 131 remove “approximately” this number is not an approximation. In this section it would be interesting to see what percentage of core, disp, and sp-sp genes are supported by transcriptome data.

Response: We agree with your suggestions. The word “approximately” is not suitable here. We have corrected it to “Within the *Oryza* genus, 81.14% of the core genes could be assigned to protein domains in our Pfam and InterPro databases, which is nearly twice the percentage of dispensable genes (41.82%) and more than seven times higher than that of accession-specific genes (10.83%).” In lines 179-182 in this revision.

Additionally, we calculated the percentage of genes in each component of the pangenome supported by transcriptome data. An average of 92.40% of the genes were expressed in the core genome, which is significantly higher than the 63.31% of genes expressed in the dispensable genome and the 52.15% in private genome in Fig. 2d in the revised manuscript.

136 – you mention length then Ka/Ks. This is confusing.

Response: We have corrected it to “Core genes showing significantly lower (0.15-fold on average) pairwise nonsynonymous substitution/synonymous substitution ratios (Ka/Ks) compared to the dispensable genes (Figs. 2f, g), indicating conservation of function among core genes in the *Oryza* genus, while variable genes evolved more rapidly” in lines 186-190 in this revision.

140 – remove “to adapt to diverse environments” this is complete speculation.

Response: Corrected

150 – what do you mean “except for”?

Response: Thanks for your careful reading. We are regret for our expression careless, and we have clarified this sentence as bellow.

A further GO analyze of the different genome type rice private gene was performed to understand their functional differentiation. A genome wild rice private genes were enriched in resistance to heavy metal and salt. B genome wild rice were focus on response to oxidative stress. C genome type wild rice private gene significantly enriched in regulation of salicylic acid and steroid hormone, which contributed their high resistance to various diseases. D genome type wild rice private genes were significantly associated with arginine synthase, which give raise to sense and adapt to environmental changes for rice. E genome type wild rice private genes were significantly involved in cellulose synthase, which benefit to large biomass. F genome type private genes were significantly associated with defense response. While the G genome type rice specific genes were focus on electron transfer, which contribute to shape its shade tolerance in lines 198-211 in the revised manuscript.

156 – "Scientists have been intrigued by" is not a good phrase. Remove

Response: Corrected.

172 – "puffy"? this is definitely not the correct word.

Response: Thank you for carefully review. We had corrected it to "The genome size of G genome type (*O. meyeriana)* was predominantly influenced by Retand LTR amplification, which counted as 8.66% of the genome size (Fig 3c), and the 10.17% abundance of Tekay in *O. glumaepatula* surpassed that 2.78% in the *O. sativa* resulted in the increase of genome size." in lines 232-236 in this revision.

224 – “it has risen to expansion of geographical range” is pure speculation and not backed up by any data, only a correlation. Genetic drift is equally plausible.

Response: Thank you very much for bringing this up and help us to improvement our manuscript. We acknowledge that the conclusion was not sufficiently rigorous, and we have removed it in this revision.

276 – how does “substantial variation in SV sizes” indicate “an enrichment of SVs in repetitive DNA regions”?

Response: Thank you for your comments. We have rewritten the sentence in lines 331-335as follows:

However, the size of structural variation in wild rice, particularly in *O. australiensis* and *O. meyeriana*, was found to be 614.70 Mb and 711.90 Mb, respectively, was overlapped with repeat sequences. This is significantly higher than the 274.43 Mb and 258.09 Mb associated with non-repeat sequences (Supplementary Fig. 17).

Line 288 – “This gene might be de novo birth to contribute to the ability of wild rice to adapt to poor and problem soil in Sri Lanka.” – this statement is confusing but also completely speculative. What does the analysis indicate the function of this gene is? Is it supported by expression data?

Response: We agree with your suggestion. The evolution of PSTOL1 was speculative, the more direct evidence was lacking. After our comprehensive consideration, we have removed this statement in this revised manuscript.

Line 290 – “The phylogenetic results revealed that the gene originated from O.

eichingeri and then transferred to O. punctata, providing evidence that O. eichingeri was the progenitor of tetraploid O. punctata, consistent with our chloroplast evolution results." > where is the evidence it was transferred to punctata? Isn't this simply because punc couldn't have evolved without eich? How does it back up the cpDNA results, these are completely different things. The parents are shown by the nuclear genome and the female parent by the cpDNA. The cpDNA has nothing to do with determining the parents. Surely the parents were already known too prior to this analysis. This could just be removed, it doesn't mean anything.

Response: We agree with your suggestion. We have removed these sentences.

L296 – 301 – picking one gene out of a QTL region is again entirely speculation that it is involved in resistance. What is the evidence that this gene is the likely causative gene? How many genes were in this region? The gene structure looks very different between sat and off, is this really an orthologue?

Response: We agree with your suggestion. we have removed the *Bph3* as an example in this revision. And subsequent complementary verification will be demonstrated in our following studies.

L302-6 – I have no idea what is the relevance of this.

Response: We have removed it.

L311 – what are "whole-genome alleles"?

Response: The term "whole-genome alleles" refers to genome-wide allelic in all *Oryza* species in comparison to the NIP reference genome. Our objective is to investigate the

genome-wide allelic variation in wild rice, as the functions of most genes are well-documented in the reference genome. We have corrected it in this revision.

L315 – "decreased, ranging from 18,463 to 23,812," – this sounds like an increase. Please reword.

Response: Corrected.

L333-7 – I don't understand this. It sounds like the genes in collinear regions were translated and their domains compared. But then 7 clusters were found. Surely there are more than 7 clusters of proteins. Do you mean 7 members per cluster? Ext Fig 11d does nothing to help me understand.

Response: We revised these sentence as "A genome-wide protein cluster was created based on their domain similarity, revealing that the protein in wild rice had 7.71 clusters in average, whereas only grouped to 1.78 in cultivated rice." in lines 367-369.

Line 342-7 – this seems like any genome region could have the same properties – this statement simply says that the different species have different SVs. But nothing about selection during domestication. In addition, it is already known that Rc controls pericarp colour, so what is the need to focus on it here?

Response: Thank you for your suggestions. We just selected one example to illustrated that wild rice had more variations than cultivar in protein sequence, and it is nothing about selection during domestication. We have revised it in lines 374-376.

L382 – "This suggests that the immune system in wild rice has a more diverse evolutionary history compared to cultivated rice." – the comparison is two cv and many wilds, it is therefore not surprising there is this difference. Also the cv rice is

diploid and the wild contains polyploid therefore at least double the number of genes is expected in punctata.

Nip has been sequenced before. How does your Nip genome compare to the previously sequenced Nip?

Response: We are sorry for the confusion. That is a mistake of writing from our side. In our opinon, *Oryza* genus can divided into 25 species, which including two cultivated species, *O. glaberrima*, and *O. satiava*. Wherea *O. sativa* contains two subspecies: *O. sativa.* ssp. *indica* and *O. sativa.* ssp. *japonica*. and the remain species all belong to wild rice. We have rewritten the sentence "This suggests that the immune system in wild rice has a more diverse evolutionary history compared to cultivated rice." into "which revealed a significant expansion of NLRs in the cultivar rice".

We identified the R gene number in two subgenomes (diploid genomes) instead of the allotertraploid wild rice (polyploid genomes). Thus, all analysis was conducted in diploid genome level.

We did not sequence the Nipponbare genome; instead, we used NIP (Qin et al., 2021, Cell) as the reference genome.

Qin, P. et al. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* **184**, 3542–3558 e3516 (2021).

Reviewer #3 (Remarks to the Author):

The paper describes the generation of reference-level genomes of 13 wild rice accessions (including BB, BBCC, CC, EE, FF, and GG) using HiFi technology. The genome assemblies are generally of high quality and will be an important resource for

utilizing useful alleles in wild rice gene pools. However, there are several major concerns about the work, which are listed below:

Thank you for your valuable comments and suggestions helping us to improve the manuscript. We have carefully revised and improved the manuscript.

1. RNA sampling of diverse tissues is critical for genome annotation. Currently, the data seem to not have been comprehensively collected and are not well utilized.

Response: We appreciate your comments and suggestions. All tissues, including leaf, stem, root, and spikelet, have been collected for the 13 wild rice species in this study, except for the species *O. meyeriana* (Supplementary Data 3). We collected only one plant of *O. meyeriana*, which has low biomass; therefore, the RNA data for this species has not been comprehensively collected (only the root is included). However, we have obtained public RNA data for *O. meyeriana* (SRR16192096 and SRR1685731) and the closest relatives *O. granulata* (PRJNA756291) for genome annotation. Furthermore, using the transcriptome data, we investigated gene expression in wild rice to enhance our understanding of gene expression supporting the core, dispensable, and private genome genes. In our copy number variation (CNV) analysis, we assessed whether gene expression is influenced by copy number variations. We identified 207 genes related to important agricultural traits (Supplementary Data 12). Additionally, we calculated the expression of NLR genes in both wild and cultivated species. Taking all of the above into consideration, we believe we have effectively utilized the RNA data, including both our manuscript data and publicly available data.

2. Most of the findings, including allelic and regulatory element variations, are based

purely on sequences, and there is a lack of genetic evidence using intercross populations and functional validation through transgenic experiments. Without substantial results, the conclusions should be provided more carefully.

Response: Thank you for point out these questions. Our manuscript presents the generation of reference-level genomes of 13 wild rice accessions (including B, BC, C, D, E, F, and G) using HiFi and Hic sequencing. The rice genus super pangenome is an important resource for utilizing useful alleles in wild rice gene pools.

However, we indeed lack of the genetic evidence for the gene functions. Hence, After Comprehensive consideration, the conclusions for the function gene had been removed in this revision.

3. The investigation of genomic heterozygosity, phasing, and intraspecific variation is completely lacking.

Response: Thank you for your suggestion and help us improving our manuscript. We have investigated the genomic heterozygosity in supplementary Table 2 by using GCE software (Table R4).

Additionally, we performed haplotype assembly using Hifiasm with parameters -h1 -h2 -ul, producing one primary genome and two haplotype draft contig genomes, the valid interaction pairs of HI-C data were used for further assembled to primary, Hap1, and Hap2 contigs genomes. The primary assembly chromosome ID and orientation were adjusted in alignment with the R498 rice genome by RagTAG. 3D-DNA was used to connect and order contigs forming pseudomolecules for the Hap1 and Hap2 genomes. The information regarding intraspecific variation information was listed in Table R5.

Table R4: Genome heterozygosity of the 13 wild rice.

Number	Rice species	Genome heterozygosity(%)
W-18	O. alta	1.0396
W-26	O. australiensis	0.31
W-59	O. brachyantha	0.073
W-71	O. eichingeri	0.1116
W-97	O. glumaepatula	0.1148
W-100	O. grandiglumis	0.147
W-171	O. malampuzhaensis	0.2033
W-190	O. minuta	0.1553
W-323	O. officinalis	0.5927
W-363	O. punctata	0.133
W-375	O. rhizomatis	0.5141
M-6	O. meyeriana	0.2993
W-128	O. latifolia	0.1905

Table R5. The intraspecific variation between the 13 wild rice assemblies (19 diploid genomes).

Species	Syntenic regions(Number)	Syntenic regions(Hap1/Hap2,Mb)	Inversion(Number)	Inversion(Hap1/Hap2,Mb)	Translocations (Number)	Translocations (Length,Mb)	SNP	Insertions(number/Length,Mb)	Deletion(number/Length,Mb)
O. glumaepatula	120	342.35/337.69	0	/	18	0.82/0/82	31756	4480/0.48	2477/0.50
O. punctata|B_t	36	430.81/428.71	0	/	10	0.77/0.77	18870	1169/0.06	1059/0.003
O. punctata|C_t	45	485.37/481.96	1	0.10/0.10	17	1.49/1.31	53315	2224/0.005	2149/0.005
O. malampuzhaensis|B_t	59	460.03/455.91	5	0.45/0.45	23	1.30/1.30	282151	13304/0.05	16706/0.05
O. malampuzhaensis|C_t	65	501.81/498.59	3	0.08/0.08	18	0.89/0.89	153079	5992/0.02	6214/0.02
O. minuta|B_t	50	433.21/432.03	1	0.008/0.008	18	0.48/0.48	191158	46162/0.06	18281/0.06
O. minuta|C_t	93	528.61/524.67	1	0.11/0.09	167	35.27/35.12	507967	107120/0.14	41279/0.14
O. rhizomatis	1446	568.03/563.25	55	7.90/8.13	425	3.50/3.47	2257727	138972/8.27	125527/8.62
O. officinalis	2372	526.63/527.29	68	17.64/17.91	811	7.27/7.28	2399716	159186/14.09	135082/14.16
O. eichingeri	2086	339.87/336.68	112	31.18/19.43	1140	18.11/17.03	4064028	358214/6.93	289957/7.46
O. alta|C_t	1150	370.39/376.19	88	51.28/43.00	394	2.83/2.77	2993400	303809/15.30	229606/14.59
O. alta|D_t	1657	375.11/366.26	86	37.38/36.93	654	4.50/4.49	3504563	307778/19.75	253034/19.20
O. grandigumis|C_t	81	424.63/423.63	0	/	28	2.02/2.03	46815	3298/0.02	2651/0.01
O. grandigumis|D_t	84	422.28/419.39	2	1.38/1.36	24	1.18/1.20	53563	3398/0.11	2907/0.44
O.latifolia|C_t	80	545.16/539.79	0	/	18	1.39/1.39	20050	1471/0.003	904/0.06
O.latifolia|D_t	89	504.49/500.25	1	0.02/0.02	51	43.25/42.14	14451	875/0.002	674/0.002
O. australiensis	948	827.93/841.86	37	14.97/19.13	209	22.01/21.91	1100303	103936/8.26	90954/ 3.28
O. branchyantha	17	150.29/158.92	3	8.57/8.57	2	0.003/0.003	9424	638/0.0001	792/0.85
O. meyeriana	2116	501.91/493.93	167	145.58/123.67	878	29.27	10524288	548069/29.05	413047/31.67

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

I am very glad to see that the current manuscript has been largely improved. Most of my comments on the last version of the manuscript have been well addressed. However, there are still some small issues should be fixed.

1. The authors claimed that “The NLR identification in this study was used the combination of denovo annotation of the genomes and RGAugry software” as revealed in their response to Reviewer #1. But I don’t find such description under the subtitle of “The analysis of pan-NLRome” in the M&M section. The de novo annotation of NLR genes should be updated with details in the M&M section, accordingly.

2. The authors claimed that they used the Nipponbare genome (Qin et al., 2021) as the reference. However, Qin et al. 2021 did not assemble the genome for Nipponbare, instead, they just used the IRGSP-1.0 (Kawahara, Y. et al., 2013), which is a high-quality assembly of Nippponbare genome, for their analysis. Thus, it is not suitable to cite Qin et al., 2021 here. This also suggests that the authors may lack sufficient investigation on the rice genomics. Please check throughout the manuscript to avoid such common-sense mistakes.

3. Line 61, “A total of 20045 gene families”. The thousandth separator is missing here.

4. The Latin names of some species are inconsistently written, e.g., the abbreviation of Oryza, O., was used in Lines 678-680 before the use of its full name in Lines 683-685 (the caption of Fig. 4). Please check throughout the manuscript to revise such typesetting errors.

Reviewer #2 (Remarks to the Author):

Based on the response to reviewers I can see the authors have made the revisions I requested in the first version. A cursory look over some of the other edits however finds a few errors and as such I think a small revision is still required:

1. Oryza sataiva spelled incorrectly in Supp Figure 4. Why is diploid punctata not included in Supp Fig 4 and only in the response to reviewers? Also needs bootstrap values adding.
2. A small point but for accuracy use the same number of decimal places for a set of data, for example in table R2 (Supp Table 1) all values in a column should have the same number of dp (e.g. 93.0 not 93).
3. Is table R5 in the main ms or only in the response. It seems important to include, but it is currently very confusing. For example “syntenic regions” and “inversions” – compared to what? Similarly “SNP” – what about them? Again if its compared to NIP how useful is this,

especially for the very distantly related species? See also point above about being consistent in the number of dp among individuals in a column.

Reviewer #3 (Remarks to the Author):

The manuscript has been enhanced based on the reviewer's feedback. One additional recommendation: include the accession number for each RNA-seq dataset in Supplementary Data 3, and ensure that all genomic (raw reads and genome assemblies) and RNA-seq data are publicly accessible and thoroughly described.

Point-by-point response

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

I am very glad to see that the current manuscript has been largely improved. Most of my comments on the last version of the manuscript have been well addressed. However, there are still some small issues should be fixed.

1. The authors claimed that "The NLR identification in this study was used the combination of denovo annotation of the genomes and RGAugry software" as revealed in their response to Reviewer #1. But I don't find such description under the subtitle of "The analysis of pan-NLRome" in the M&M section. The de novo annotation of NLR genes should be updated with details in the M&M section, accordingly.

Response: We thank you for pointing out this issue. We have added the descriptions of the de novo annotation of NLR genes in the M&M sections in line 275 to 278 in this revision.

2. The authors claimed that they used the Nipponbare genome (Qin et al., 2021) as the reference. However, Qin et al. 2021 did not assemble the genome for Nipponbare, instead, they just used the IRGSP-1.0 (Kawahara, Y. et al., 2013), which is a high-quality assembly of Nippponbare genome, for their analysis. Thus, it is not suitable to cite Qin et al., 2021 here. This also suggests that the authors may lack sufficient

investigation on the rice genomics. Please check throughout the manuscript to avoid such common-sense mistakes.

Response: We apologize for the mistake. We agree with the reviewer that we used the IRGSP-1.0 (Kawahara, Y. et al., 2013) as the reference. However, we used Qin's re-annotated genome file to construct a gene-based super pangenome and we described it in the "Analyses of the protein-coding-gene based *Oryza* super pangenome" section of M&M. We had missed the cite of IRGSP-1.0 (Kawahara, Y. et al., 2013) and have corrected it in line 78 and 89 in the revised manuscript and in line 202, 219, and 238 in the revised M&M. We have also revised the cite in the manuscript throughout in this revision. Finally, we sincerely appreciate your rigorous attitude towards scientific research, which helped us to further improve the manuscript.

3. Line 61, "A total of 20045 gene families". The thousandth separator is missing here.

Response: We have corrected it and carefully checked the whole manuscript to revised such mistakes in the revised manuscript.

4. The Latin names of some species are inconsistently written, e.g., the abbreviation of Oryza, O., was used in Lines 678-680 before the use of its full name in Lines 683-685 (the caption of Fig. 4). Please check throughout the manuscript to revise such typesetting errors.

Response: Thank you for your detailed and careful reviews. We have corrected it in line 743-745 and checked throughout the manuscript and supplementary files. And we thank you again for improving our manuscript.

Reviewer #2 (Remarks to the Author):

Based on the response to reviewers I can see the authors have made the revisions I requested in the first version. A cursory look over some of the other edits however finds a few errors and as such I think a small revision is still required:

1. Oryza sataiva spelled incorrectly in Supp Figure 4. Why is diploid punctata not included in Supp Fig 4 and only in the response to reviewers? Also needs bootstrap values adding.

Response: Thank you for your constructive suggestion. In our opinion, it is appropriate to correspond to the sample which used in the nuclear genome phylogenetic tree construction, so we did not add the bb genome. We have added the bootstrap values in Supplementary Fig 4 in this revision.

2. A small point but for accuracy use the same number of decimal places for a set of data, for example in table R2 (Supp Table 1) all values in a column should have the same number of dp (e.g. 93.0 not 93).

Response: Thank you for your great suggestion, very appreciated. We have corrected it in Supplementary Table 1 and checked throughout the manuscript.

3. Is table R5 in the main ms or only in the response. It seems important to include, but it is currently very confusing. For example “syntenic regions” and “inversions” – compared to what? Similarly “SNP” – what about them? Again if its compared to NIP how useful is this, especially for the very distantly related species? See also point above

about being consistent in the number of dp among individuals in a column.

Response: We apologize for the confusion. The R5 data were exclusively used to respond to Reviewer 3. The information presented in Table 5 illustrates the variation between the two haplotype assemblies of the 13 wild rice species. We have thoroughly checked and revised the number of data points among individuals in each column throughout the manuscript.

Reviewer #3 (Remarks to the Author):

The manuscript has been enhanced based on the reviewer's feedback. One additional recommendation: include the accession number for each RNA-seq dataset in Supplementary Data 3, and ensure that all genomic (raw reads and genome assemblies) and RNA-seq data are publicly accessible and thoroughly described.

Response: Thank you for your recommendation. We had updated the accession number for each RNA-seq dataset in Supplementary Data 3. And the genomic and RNA-seq data were share to public according to your suggestion.

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

The current manuscript has been further improved and all my comments on the last version of the manuscript have been well addressed.

Reviewer #2 (Remarks to the Author):

The authors have made the small edits I suggested.

Reviewer #1 (Remarks to the Author):

The current manuscript has been further improved and all my comments on the last version of the manuscript have been well addressed.

Response: Thank you! Thank the reviewer for the review, comments and recognition of our work.

Reviewer #2 (Remarks to the Author):

The authors have made the small edits I suggested.

Response: Thanks!