

A telomere-to-telomere genome assembly of *Oryza australiensis* reveals transposon driven centromere repositioning and shared EE–DD ancestry

Corresponding Author: Dr Tao Zhang

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

Attachments originally included by the reviewers as part of their assessment can be found at the end of this file.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

This study elucidates the centromere repositioning mechanism of *Oryza australiensis* (Australian wild rice) through telomere-to-telomere (T2T) genome assembly, and clarifies the evolutionary relationship in which the EE genome serves as the direct donor of the DD subgenome using multiple methods. The study features an innovative design and sufficient data volume. Overall, this work is of great importance and deserves publication in Nature Communications. However, several issues need to be clarified prior to its publication:

Major points:

- 1) Confirmation of LAI values: There is a significant discrepancy in the reported LAI (Longest Aligned Index) values. The manuscript reports inconsistent values: 22.67 (Line 32), 20.67 (Line 528), and 22.53 (Table 1). The authors must carefully check their data, identify the correct value, and ensure it is consistently reported throughout the manuscript.
- 2) Sequencing depth consistency: The description of sequencing depth in the main text (Lines 107–109) is inconsistent with the supplementary table. The authors should cross-verify these values and ensure consistency across the manuscript and supplementary materials.
- 3) Comparison with existing genome assemblies: The introduction (Lines 73–87) lists existing *O. australiensis* genomes but does not provide sufficient comparison to highlight the unique value of this new T2T assembly. The authors should perform a comprehensive comparative analysis of assembly quality metrics (e.g., contig N50, BUSCO completeness, LAI, presence of gaps, QV) between the new T2T assembly and previously published versions. Presenting these results in a summary table will help clearly demonstrate the advantages of the new assembly and its contribution to advancing genomic research in *Oryza*.
- 4) Gap closure validation: The closure of gaps is a central achievement of T2T assembly. The manuscript would be significantly strengthened by providing direct visualization evidence. The authors should include, for example, an IGV screenshot showing a representative gap region that was closed, with reads spanning the previously gapped region and illustrating read coverage in the surrounding sequence both before and after gap filling.

Minor points:

- 1) Line 44: “Key word” should be changed to “Keywords”.
- 2) Line 105: Section title “Result” should be changed to “Results”.
- 3) Line 135: “*O australiensis*” should be changed to “*O. australiensis*”.

4) Line 248: "O. australiensi" should be changed to "O. australiensis".

5) Line 350: "three centromere repositioning model" should be changed to "three centromere repositioning models".

6) Line 647: Section title "Method" should be changed to "Methods".

7) In Figure 5, the font size of the legends in panels (c) and (d) is not uniform. Please revise the figures to ensure consistent formatting.

Reviewer #2

(Remarks to the Author)

The authors presented a telomere-to-telomere genome assembly of *Oryza australiensis* (genome type EE) using long and ultra-long ONT reads and Illumina reads. Genome annotation included telomere and centromere regions, transposable elements (TEs) and genes. One main goal was investigating mechanisms of centromere repositioning in *O. australiensis*. The other was providing support to the hypothesis of the EE progenitor being the direct donor of the DD subgenome in the CCDD tetraploids. For this purpose, the authors used two approaches to analyse *O. australiensis* and the DD subgenome of *O. alba* (CCDD) - i.e. they estimated DD-EE divergence time using single-gene phylogenetic analysis and compared LTR-RT sequences and their insertion times in the two species.

Overall, the research questions (the mechanisms of centromere repositioning in *O. australiensis*, and the analysis of LTR-RT elements to investigate the evolutionary history of DD and EE genomes) are of great interest. However, the methods lack description of important details or even main elements necessary to clearly understand the rationale/implementation of the analyses, the chosen methods do not always accurately address the research questions, and the conclusions are not always supported by results.

The conclusion of EE genome being the direct donor of the DD subgenome in the tetraploid CCDD is hard to prove and it also depends on how genome types are defined (i.e. the original definition was based on interspecific hybridization and cytogenetics experiments to assess potential interfertility between species, but the analysis provided here is based on sequence comparison). The methods used by the authors are not described to enough detail to allow full understanding of the analyses. Results (visualized in figure 5 and 6) do not reveal a 'direct' relationship but show (to some extent) a close relationship between EE genome and DD subgenome, providing further analysis in support to previous literature (Ammiraju et al., The Plant Journal. 2010; Bao and Ge, Plant Syst. Evol. 2004).

My main general suggestion to the authors is to: 1) revise the Methods to give a precise, detailed description of all analyses performed and 2) draw conclusions that are clearly and robustly supported by results.

Attached, I provided in-depth comments/suggestions to four sections of the manuscript, and I grouped comments/suggestions across the manuscript into three categories referring to: clarity/rationale, figures and methods.

Reviewer #3

(Remarks to the Author)

The authors present a valuable T2T genome assembly of *O. australiensis*, a significant resource for evolutionary genomics and rice improvement. However, the study's central claims regarding centromere repositioning require substantial strengthening to be convincing.

Major Concerns:

1. Weak Evidence for Centromere Repositioning: The hypothesis of centromere shift/repositioning in *O. australiensis* is not well supported by the current comparisons. Given the significant phylogenetic distance between *O. australiensis* and *O. sativa*, the observed centromeric changes are more likely the result of artificial selection and diversification within the *O. sativa* lineage. To robustly test their hypothesis, the authors must use a closer relative (e.g., *O. rufipogon*) as the primary comparator. A three-way analysis including *O. rufipogon*, *O. sativa*, and *O. australiensis* is essential. As Figures 4 and S6 suggest, the major shifts appear to be a feature of *O. sativa*, not its wild relatives.

2. Unsupported Proposed Mechanism: The model in Figure 7 lacks compelling evidence from the data presented and should be either removed or significantly revised to reflect the more limited and alternative conclusions.

3. The description of the evolution relationship between EE and DD subgenomes in the tetraploid should be simplified. If the assembly quality permits, a more insightful analysis would be to compare centromere positions between the diploid and tetraploid forms.

4. The discussion is overly long and complex. It should be shortened and focused on the core findings. Excessive detail on other retrotransposons should be reduced.

5. The abbreviation "CRM" must be replaced with "CRR" consistently throughout the text and all figures. The CRM means Centromere Retrotransposon of Maize.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The authors have responded to all my comments and the manuscript is now much improved. I have no more questions.

Reviewer #2

(Remarks to the Author)

I appreciated that the authors changed the title to more accurately reflect the content and findings of the present manuscript and their acknowledgement in the rebuttal letter to re-formulate some conclusions in a more cautious way. However, the effort showed in the rebuttal letter not always aligns with the formulation of some statements found in the main manuscript, which remain essentially the same as before, and I am specifically referring to the section about a direct ancestry contribution of EE to the DD subgenome in the O alta species, which in my opinion is not robustly supported.

The conclusive biological insight of a direct genomic contribution of EE as the donor of subgenome DD in the CCDD tetraploids (lines 554-555) cannot be inferred given the proposed hypothesis and the provided evidence. The hypothesis is centred on the CCDD polyploidization time estimate based on radiation time of C genomes (*O. officinalis*, *O. minuta* CC and *O. alta* CC) by Ammiraju et al. (2010), and used here as the reference value to compare TE distributions in *O. australiensis* and *O. alta*. Results presented in Fig. 5c, 5d and Fig. 6c, 6g are overly interpreted and do not provide the conclusive evidence of a direct contribution of EE as the donor of the DD subgenome in CCDD. Detailed argumentation can be found attached.

In conclusion, in my view the present manuscript provides:

- 1) a genomic resource (i.e. the telomere-to-telomere genome reference of *O. australiensis* with relative annotation of genes, TEs and telomeres) that is gap-free and identification of centromeric regions;
- 2) a structural comparison of centromeric regions (and relative satellite repeat and TE content) with other *Oryza* species (where the authors used publicly available genome resources);
- 3) a descriptive comparison of TE distribution and insertion time between *O. australiensis* and *O. alta* supporting their previously described phylogenetic relationship.

Reviewer #3

(Remarks to the Author)

The authors have addressed all my questions and given realistic re-analysis and explanation. My suggestion is accepted.

Version 2:

Reviewer comments:

Reviewer #4

(Remarks to the Author)

I did not review earlier versions of this manuscript. I have reviewed the current revision with a specific focus on whether the concerns raised by the previous Reviewer 2 have been adequately addressed. Many of these concerns related to overstated conclusions about the evolutionary relationships among *Oryza* subgenomes -- specifically, divergence times and patterns of ancestry/descent between the EE lineage and the DD subgenome of the CCDD polyploid lineage.

The revision resolves the previous overstatements to my satisfaction. Other concerns raised by reviewer 2 are also addressed. I appreciate the authors' careful rewording to accurately state the data-supported inferences.

Please note, however, the following recommended rephrasing of the paper's key evolutionary inference to clarify the English:

The following text (present in the abstract and elsewhere) is currently phrased awkwardly as: "...the EE lineage and DD subgenome prior to CCDD formation share common ancestry with an EE-like progenitor."

This idea should be rephrased as follows for greater clarity: "... prior to CCDD formation, the ancestral lineage that contributed the DD subgenome shared common ancestry with an EE-like progenitor."

This inference could also be written as: "...prior to formation of the CCDD polyploid lineage, the ancestral DD subgenome donor shared common ancestry with the ancestor of the EE genome lineage."

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Reviewer1

1. Confirmation of LAI values: There is a significant discrepancy in the reported LAI (Longest Aligned Index) values. The manuscript reports inconsistent values: 22.67 (Line 32), 20.67 (Line 528), and 22.53 (Table 1). The authors must carefully check their data, identify the correct value, and ensure it is consistently reported throughout the manuscript.

Response: We sincerely appreciate your attention to detail and acknowledge that data accuracy is paramount for scientific integrity. The LAI recalculated using our final validated assembly is 23.79. We thoroughly reviewed the manuscript and updated all instances to consistently report this value. Thank you very much!

2. Sequencing depth consistency: The description of sequencing depth in the main text (Lines 107–109) is inconsistent with the supplementary table. The authors should cross-verify these values and ensure consistency across the manuscript and supplementary materials.

Response: Thank you very much for your valuable suggestion. We have carefully verified the sequencing depth values between the main text and the supplementary table and have ensured consistency throughout the manuscript and supplementary materials. Upon reviewing our revision records, we found that updates to the sequencing depth values were not carried through to all sections simultaneously. We have verified these figures against the raw data and have applied the consistent values of 88.71×, 22.65×, and 113.15× throughout both the main text (Lines 115-117, 664-666) and Supplementary Table 2. Thanks a lot!

3. Comparison with existing genome assemblies: The introduction (Lines 73–87) lists existing *O. australiensis* genomes but does not provide sufficient comparison to highlight the unique value of this new T2T assembly. The authors should perform a comprehensive comparative analysis of assembly quality metrics (e.g., contig N50, BUSCO completeness, LAI, presence of gaps, QV) between the new T2T assembly and previously published versions. Presenting these results in a summary table will help clearly demonstrate the advantages of the new assembly and its contribution to advancing genomic research in *Oryza*.

Response: Following your valuable suggestion, we have now compared our T2T assembly against the four previously published *O. australiensis* genomes across the key metrics requested, with the full results presented in Supplementary Table 1. This comparison shows that our assembly is the first for this species to report zero gaps, providing continuous sequence through centromeres and telomeres—regions important for *Oryza* evolutionary studies but difficult to resolve previously. The QV of 41.76 reflects a known trade-off when using ONT ultralong reads to maximize structural

completeness rather than base-level precision; while this is lower than the HiFi-based assembly by Long et al., it remains fully sufficient for phylogenomic and selection analyses, particularly when balanced with the 99.8% BUSCO completeness and LAI of 23.79 that indicate proper assembly across repetitive regions. Our assembly improvements directly enable the core biological findings: the gap-free, phased structure allowed us to identify three distinct centromere repositioning mechanisms, including a burst-like transposon amplification pattern unique to the EE genome, and to provide transposon-based evidence clarifying the DD-EE evolutionary relationship. By framing the comparative results in the Introduction (Lines 99-103), we now more clearly demonstrate how resolving these previously recalcitrant regions yields new mechanistic insights, strengthening both the resource value and its contribution to *Oryza* evolutionary biology. Thank you very much!

Supplementary Table 1 Comparison of Genome Assembly Metrics between this study and previously published

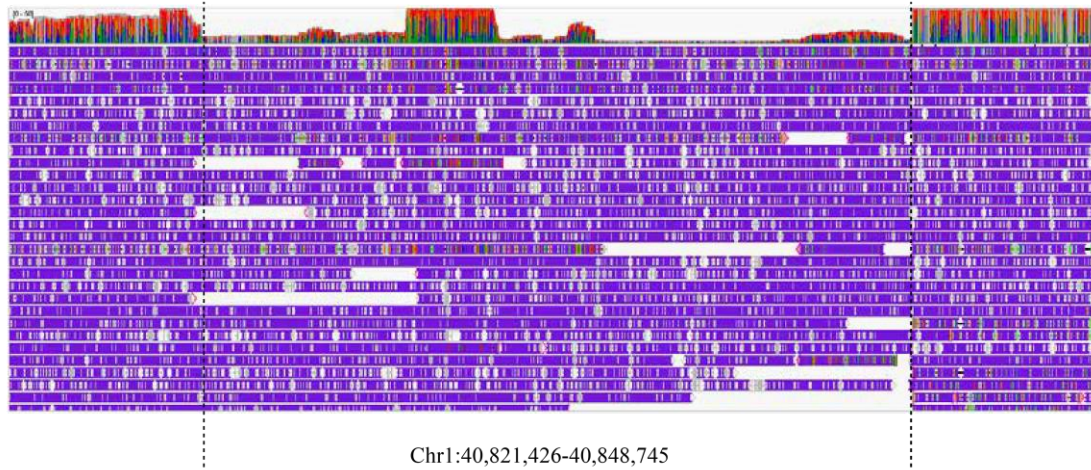
	This study	Long et.al. 2024 <i>Nat. Commun.</i>	Fornasiero, A. et.al. 2025 <i>Nat. Genet.</i>	Phillips, A.L. et.al. 2022 <i>Sci. Rep.</i>	Morrison, S. et.al. 2025 <i>Sci. Rep.</i>
Gaps	0	3	88	0	3
LAI	23.79	22.9	23.77	18.35	21.55
QV	41.76	57.62	36.99	25.18	-
BUSCO	99.8	99.8	99.5	99.3	99.3
N50	76.6 Mb	78.1 Mb	73.8 Mb	2.6 Mb	76.6 Mb
Number of contigs	12	16	60	850	13

Note: In the genome assembly project of *O. australiensis* completed by Morrison, S. et al., the QV is not provided due to the lack of Illumina whole-genome sequencing data.

4. Gap closure validation: The closure of gaps is a central achievement of T2T assembly. The manuscript would be significantly strengthened by providing direct visualization evidence. The authors should include, for example, an IGV screenshot showing a representative gap region that was closed, with reads spanning the previously gapped region and illustrating read coverage in the surrounding sequence both before and after gap filling.

Response: Thank you very much for your valuable suggestions. We have added Supplementary Figure 3 showing an IGV screenshot of a representative previously

gapped region. The figure illustrates a ~27 kb area now spanned by multiple ultra-long ONT reads, with coverage tracks displaying the sequence context before and after closure. The contiguous reads support across several independent reads offers validation for the gap closure. This figure is now referenced in the Supplemental figure 3 (Line 124). Thank you very much!



Supplemental figure 3. Nanopore long reads coverage of closed gap regions in the *O. australiensis* genome.

Minor points:

1) Line 44: "Key word" should be changed to "Keywords".

Response: We have corrected. Thank you very much!

2) Line 105: Section title "Result" should be changed to "Results".

Response: We have corrected. Thank you very much!

3) Line 135: "O australiensis" should be changed to "O. australiensis".

Response: We have corrected. Thank you very much!

4) Line 248: "O. australiensi" should be changed to "O. australiensis".

Response: We have corrected. Thank you very much!

5) Line 350: "three centromere repositioning model" should be changed to "three centromere repositioning models".

Response: We have corrected. Thank you very much!

6) Line 647: Section title "Method" should be changed to "Methods".

Response: We have corrected. Thank you very much!

7) In Figure 5, the font size of the legends in panels (c) and (d) is not uniform. Please revise the figures to ensure consistent formatting.

Response: We have revised Figure 5 to ensure that the font size of the legends in panels (c) and (d) is now uniform and consistent throughout. Thank you very much!

Reviewer2 :

The authors presented a telomere-to-telomere genome assembly of *Oryza australiensis* (genome type EE) using long and ultra-long ONT reads and Illumina reads. Genome annotation included telomere and centromere regions, transposable elements (TEs) and genes. One main goal was investigating mechanisms of centromere repositioning in *O. australiensis*. The other was providing support to the hypothesis of the EE progenitor being the direct donor of the DD subgenome in the CCDD tetraploids. For this purpose, the authors used two approaches to analyse *O. australiensis* and the DD subgenome of *O. alta* (CCDD) - i.e. they estimated DD-EE divergence time using single-gene phylogenetic analysis and compared LTR-RT sequences and their insertion times in the two species.

Overall, the research questions (the mechanisms of centromere repositioning in *O. australiensis*, and the analysis of LTR-RT elements to investigate the evolutionary history of DD and EE genomes) are of great interest. However, the methods lack description of important details or even main elements necessary to clearly understand the rationale/implementation of the analyses, the chosen methods do not always accurately address the research questions, and the conclusions are not always supported by results.

The conclusion of EE genome being the direct donor of the DD subgenome in the tetraploid CCDD is hard to prove and it also depends on how genome types are defined (i.e. the original definition was based on interspecific hybridization and cytogenetics experiments to assess potential interfertility between species, but the analysis provided here is based on sequence comparison). The methods used by the authors are not described to enough detail to allow full understanding of the analyses. Results (visualized in figure 5 and 6) do not reveal a 'direct' relationship but show (to some extent) a close relationship between EE genome and DD subgenome, providing further analysis in support to previous literature (Ammiraju et al., The Plant Journal. 2010; Bao and Ge, Plant Syst. Evol. 2004).

My main general suggestion to the authors is to: 1) revise the Methods to give a precise, detailed description of all analyses performed and 2) draw conclusions that are clearly and robustly supported by results.

Attached, I provided in-depth comments/suggestions to four sections of the manuscript, and I grouped comments/suggestions across the manuscript into three categories referring to: clarity/rationale, figures and methods.

Response: We sincerely appreciate the reviewer's incisive summary and constructive

criticism, which has helped us identify critical areas for improvement. To enhance methodological transparency, we have completely rewritten the Methods section with full technical specifications, parameter details, and rationales. Please see detail in Supplementary Method. Following the reviewer's guidance, we have substantially revised our conclusions regarding the DD-EE relationship throughout the manuscript. The language now accurately reflects that our transposon dynamics and comparative genomic analyses demonstrate a very close evolutionary relationship between the pre-1.7 Mya EE lineage and the DD subgenome, suggesting they share common ancestry with an EE-like progenitor.

We have prepared detailed point-by-point responses to all comments in the attached file, pertaining to clarity/rationale, figures, and methods. These responses detail the specific textual changes, figure modifications, and methodological clarifications made in the revised manuscript. Thank you very much!

Centromere Mapping Reveals Repositioning and Transposon-Driven Neocentromere Formation in *O. australiensis*

1. The satellite repeat distribution analysis in the regions flanking the functional centromeres is lacking a temporal dimension (Fig. 2). How can the authors conclude that satellite repeats 'may represent sequential stages of neocentromere evolution' (line 219) if there is no temporal mean to investigate the sequential occurrence of these events?

Response: We have carefully examined satellite repeat abundance in regions flanking functional centromeres and observed a gradient—from high in Chr2 to intermediate in Chr12 to nearly absent in Chr6—that aligns with theoretical models of centromere turnover. However, this spatial pattern alone cannot establish temporal sequence, as the reviewer correctly notes, and we agree that without direct time-course data, the gradient could alternatively reflect chromosome-specific mutational rates or functional constraints rather than evolutionary stages. To address this limitation, we have revised the text (Lines 221-223) to state that "This gradient in satellite repeat abundance varies among chromosomes and may be consistent with ongoing centromere reorganization, though this pattern cannot demonstrate temporal sequence". Thank you very much!

2. Lines 215-218: "The neocentromere is present in Chr2, Chr6, and Chr12, where Chr2 has a large number of satellite repeats at a distance of 1.5 Mb from the functional centromere, Chr12 contains fewer satellite repeats, while Chr6 contains almost no satellite repeats". How can the authors conclude that a neocentromere is present on Chr 6 if no satellite repeat signal was detected on this chromosome?

Response: Thank you very much for your valuable suggestions. The CENH3 ChIP-seq data on Chr6 shows strong enrichment at a genomic locus that lacks satellite repeats,

which raises a valid question about neocentromere classification. A neocentromere is defined by the establishment of centromere function at a new position, and this function is marked by CENH3 binding regardless of underlying sequence composition. In Chr6, robust CENH3 enrichment demonstrates functional centromere activity, while comparative analysis shows this position differs from the satellite-rich centromeres found in related species (Fig. 2), indicating it has indeed shifted. Plant literature contains several documented cases of functional centromeres specified solely by CENH3 without satellite arrays (Bousios et al., 2025; Yang et al., 2016) , confirming that canonical repeats are not required for centromere identity. To directly incorporate this clarification into our results, we have added the following point to the description in Lines 218-221. Thank you very much!

3. Lines 219-222: "Additionally, functional neocentromere formation also occurs on chromosomes 1, 3, 5, 7, 8, and 10, where CENH3-binding assays revealed that large structural domains containing two distinct satellite repeat clusters serve as novel functional centromeric regions". What are these two satellite repeat clusters? How where they detected?

Response: We are so sorry for the misleading, which may have led to the reviewer's confusion. The phrase "two distinct satellite repeat clusters" refers to two separate genomic regions that are enriched with the 155-bp CentO satellite repeat. These regions were identified computationally: we identified the 155-bp centromeric satellite sequence in the *O. australiensis* genome using RepeatExplorer and then mapped its distribution across the entire genome to identify these enriched regions. To clarify, we have revised the manuscript to explicitly state that these are "two distinct regions enriched for satellite repeats"(Line 226). Thank you very much!

4. Line 227-230: "We further compared the insertion times of LTRs in active and inactive centromeres, and the results showed that active centromeres contain younger LTRs (Fig. 2f), suggesting that TE insertions may contribute to the formation of neocentromere". How were active and inactive centromeres defined? What are their respective genomic coordinates? How were “younger LTRs” defined? Could you show the distribution of these centromeric LTRs along chromosomal coordinates?

Response: We are so sorry for the misleading of those description. In our study, "active centromeres" were defined as the genomic regions exhibiting significant CENH3 binding, as identified by ChIP-seq peaks (detailed in the Methods section "Centromere analysis"). These regions represent the functional centromeres. Conversely, "inactive centromeres" refer to loci that contain abundant centromeric satellite repeats (CentO) but show no significant CENH3 binding signal. These typically represent centromere positions that have been inactivated. The genomic coordinates for both active and

inactive centromeres on all chromosomes have been provided in Supplementary Table 6. The term "younger LTRs" was used in a comparative context. It means that the LTRs within the active centromeres have, on average, a more recent insertion time compared to those within the inactive centromeres. As suggested, we have now provided a detailed view of the LTR distribution within the centromeric regions (Supplementary Figure 8). We have also revised the relevant sentence in the manuscript (Lines 243-246) for greater clarity to read: We further compared the LTR insertion times between active centromeres (defined by CENH3 binding) and inactive centromeres (enriched in satellite repeats but devoid of CENH3). The results showed that active centromeres are enriched with younger LTRs, indicative of more recent insertion events. Thank you very much!

5. Line 214-215: "Unlike *O. sativa*, the EE genome exhibits centromere repositioning and neocentromere formation on multiple chromosomes". Reference(s) supporting the statement about *O. sativa* is/are missing. To improve analysis of centromere repositioning, I suggest reading: Ahmed HI, Heuberger M, Schoen A et al., Nature. 2023. <https://doi.org/10.1038/s41586-023-06389-7>. This work nicely investigated centromere dynamics and centromere-specific retrotransposons insertion time in einkorn.

Response: Thanks for your valuable suggestions. We have incorporated the Ahmed et al. (Ahmed et al., Nature 2023) study, which investigated centromere dynamics and retrotransposon insertion timing in einkorn wheat, into our Discussion section (Lines 594-595) to provide comparative context for our findings on centromere evolution. We also have revised manuscript (Lines 214-215) to remove the unsupported comparison and now state objectively: "The EE genome exhibits centromere repositioning and neocentromere formation on multiple chromosomes". Thank you very much!

Evolutionary trajectory of centromeres

1. To reconstruct centromere evolution in the *Oryza* genus, the authors built a "multi-genome synteny map based on the centromeric core sequences and epigenetic characteristics of the cultivated species (Nipponbare and MH63) and wild species (*O. australiensis* and *O. brachyantha*)" (Lines 298-301). Firstly, the authors should clearly define what are the genomic and epigenetic features used (genes, satellite repeats, CENH3, others?) and how the maps were built (I couldn't find methods for this analysis).

Response: Thank you very much for your valuable suggestion. We have revised the Results and add Supplementary Methods sections to provide a transparent account of our approach. The multi-genome synteny map was primarily constructed based on the collinearity of protein-coding genes across the genomes of *O. sativa* (Nipponbare and MH63), *O. australiensis*, and *O. brachyantha*, using *Leersia perrieri* as the outgroup.

Subsequently, to analyze centromere evolution, we then projected the genomic positions of two key centromeric features--the core centromeric satellite repeats (CentO) and the CENH3-binding domains defined by ChIP-seq--onto this syntenic framework to compare centromere positions and structures across species. In essence, the "map" is the gene synteny, and the "centromeric core sequences and epigenetic characteristics" are the data layers we overlaid onto this map for comparative analysis.

The original description was unclear, so we have thoroughly revised the relevant text in the Results section (Lines 315-320) to eliminate ambiguity and have added a detailed description of this methodology in the "Supplementary Method" section under a new subsection titled "Comparative synteny analysis of centromere evolution". Thank you very much!

2. Secondly, the rationale of the "ancestral centromeres" needs to be clearly explained and justified. The authors defined as ancestral centromeres the syntenic features within centromeres and flanking regions across two *O. sativa* varieties (i.e. Nipponbare and MH6) and *O. brachyantha* (FF). For a start, the authors should describe in detail how synteny was defined for each feature under study, and how sequence homology was measured across species. Then, they should explain that conserved synteny of homologous regions among *O. sativa* varieties and *O. brachyantha* reflects the ancestral centromeric state, and finally they should present the comparative analysis of homologous regions in *O. australiensis*. Given the great evolutionary distance between AA, EE and FF genome types, it is not trivial to analyse homology of non genic regions, and this should be thoroughly described in the methods and results. Furthermore, figure legend for Fig. 4a-c is missing. What are the features connected by the grey lines in these panels?

Response: Thank you very much for your valuable suggestions. We have revised the manuscript to provide a thorough clarification as follows:

Syntenic blocks were identified using the JCVI toolkit. Orthologous gene pairs were first inferred by comparing the coding sequences (CDS) of the two genomes using BLASTN, with the default parameters (e-value < 1e-10). Reciprocal best hits from this search were defined as high-confidence anchors. Subsequently, these anchor pairs were used as input for the MCscan software to delineate syntenic blocks based on the conservation of gene order and density (Supplementary Method: Comparative synteny analysis of centromere evolution). The grey lines in Fig. 4a-c represent these anchor gene pairs connecting homologous loci within syntenic blocks, and we have added this description to the figure legend.

We defined a centromeric region as "ancestral" if it was syntenically conserved in at least two of the three divergent evolutionary lineages analyzed: the AA genome (*O. sativa* varieties), the EE genome (*O. australiensis*), and the FF genome (*O.*

brachyantha). The conservation of a centromeric position (as defined by the synteny of its flanking genic regions) across multiple, independent lineages provides strong evidence that this locus was present in their last common ancestor.

This framework allowed us to directly assess centromere evolution in the EE genome. For each inferred ancestral centromere position, we examined the homologous genic region in the EE genome. We then determined whether the functional CENH3-binding domain and CentO satellites in the EE genome were located within this syntenic region (indicating conservation) or if they had shifted to a new, non-syntenic location (indicating repositioning/neocentromere formation). As the reviewer rightly points out, we did not analyze the homology of the non-genic, highly repetitive centromeric core itself across species, as this is exceptionally challenging. Our analysis focused on the conserved genic flanks to anchor the position. We have now explicitly detailed this rationale and methodology in Supplementary Methods (Comparative synteny analysis of centromere evolution) and Results sections (Lines 320-324). Thank you very much!

Shared Transposon Amplification History Implicates EE as Direct Ancestor of the DD Subgenome

1. Fig. 5c: I struggle seeing the secondary amplification peak at 1.7 Mya in *O. alta* DD. Furthermore, LTR-RT insertion time on the x-axis should be taken as a broad estimate (i.e. confidence intervals were not defined), and not as the precise timing of insertion.

Response: Thanks for your valuable suggestion. We agree that the DD subgenome lacks a clear visual peak at ~1.7 Mya and that LTR insertion times should be treated as estimates. To address this, we have revised our approach in two ways. First, we now use a quantitative enrichment analysis across the 1.7-2.5 Mya window rather than relying on visual peak identification, which shows that both EE and DD genomes contain significantly more insertions in this time frame compared to the CC genome ($p < 0.01$, Fisher's exact test). Second, we have revised the text to use cautious phrasing such as "approximately 1.7-2.5 Mya" and "broad temporal enrichment" rather than implying precise timing (Lines 423-432). Thank you very much!

2. Lines 433-434: "the nearly-parallel trends in the fitted curves of CC genome and CC subgenome, as well as EE genome and DD subgenome, suggest that EE genome may be the direct donor of the DD subgenome". How can the authors support their conclusions given the results showed in Fig. 5d? Why does the "nearly parallel trend of the fitted curves" suggest that EE may be the direct donor of the DD subgenome? To better understand results in Fig. 5d, the authors should include thorough description of the settings, parameters and thresholds used to identify and compare different LTR-RT families(?)/lineages(?) (line 419).

Response: We are sorry for the misleading. We intended to convey that the divergence

trajectories show a highly concordant timing of differentiation, and have revised the Results section accordingly. Specifically, the logic for the DD-EE relationship is based on intersection points of the fitted curves. The CC-DD divergence curve represents independent evolution of these subgenomes after polyploid formation. The CC/CC-subgenome and EE/DD curves both intersect with this reference curve at nearly identical times (~1.73 and ~1.84 Mya), indicating their differentiation was temporally coupled with CCDD polyploid formation. Since CC is the established direct donor of the CC-subgenome, the similar timing of DD-EE divergence suggests an analogous direct donor relationship. Methodological details have been added to the Supplementary Methods (Differential Analysis of LTR Sequences), including full-length LTR identification, orthologous pair alignment parameters, and sequence similarity calculations over time. Thank you very much!

Angela LTR Retrotransposons and a Conserved 275-bp Repeat Validate EE as the Donor of DD

1. Line 455: "This suggests that Angela LTR was gradually lost during the evolution of *Oryza*," I don't think Angela LTR elements were necessarily lost during the evolution of *Oryza*. An alternative simple explanation could be that expansion of Angela LTRs was more successful in the genome of *O. australiensis* than in other genomes. It has indeed been shown that, in *Oryza*, TE evolutionary history is independent 2 from the host genome type (see El Baidoury and Panaud, *Genome Biol. Evol.* 2013, doi: 10.1093/gbe/evt025). Furthermore, Angela LTR RTs are a family of RTs consisting of many different types. What are the sequences considered for this analysis?

Response: Thanks for your valuable suggestions. Regarding the diversity within the *Angela* superfamily, we analyzed elements at superfamily resolution using TEsorter using an E-value cutoff of $<1e-10$ and requiring $\geq 80\%$ coverage of the reference reverse transcriptase domain, and were defined as those classified as *Angela* superfamily containing intact RT domains. This approach captures overall family-level abundance dynamics rather than subfamily-specific variations. We have revised line 483 to read: "This pattern suggests differential evolutionary dynamics of *Angela* LTR-RTs across *Oryza* genomes, characterized by prominent expansion and retention in *O. australiensis* contrasted with comparatively lower abundance in other lineages." The revision aligns with El Baidoury and Panaud (2013), who demonstrated that TE evolutionary histories can be independent of host genome divergence in *Oryza*. We maintain that amplification events prior to ~1.7 Mya in EE and DD contributed to the observed pattern but state this more cautiously.

2. Fig. 6c: Again, I struggle seeing the peak of Angela LTR expansion at 1.7 My in *O. alta* DD that would support the author's claim of EE being the direct donor of the DD

subgenome. The presence of two y-axes with the count of *Angela* LTRs in EE and CCDD, respectively, confuses me. Are these distributions directly comparable (i.e. element count is up to 175 and 25 for *O. australiensis* and *O. alta*, respectively)? Given that *Angela* LTRs were detected in many species with AA, BB, CC (diploid), FF and GG genome types, I suggest adding distributions of *Angela* LTR insertion times for each of these species in the Supplementary figures.

Response: We are sorry for the misleading on our original description of the *Angela* LTR patterns could be more precise. We have corrected this in the revised text (Lines 423-432). Regarding Fig. 6c (*Angela* LTRs specifically), we wish to clarify that it shows a concentration of 12 intact *Angela* LTRs in DD within the 1.7-3 Mya window, not a distinct peak at 1.7 Mya. This pattern indicates temporal overlap in transposon activity between EE and DD within a broad period, rather than synchronized peaks. We have revised the discussion accordingly to reflect this more cautious interpretation. We acknowledge that the dual y-axes in Fig. 6c, while necessary to visualize the distribution shapes given the vast difference in total copy numbers, hinder direct comparability. To address this and the reviewer's excellent suggestion, we have added Supplementary Figure 14, which provides normalized *Angela* LTR insertion time distributions for multiple diploid genomes (BB, CC, FF, GG) to place the EE and DD data in a broader context. Thank you very much!

3. Fig. 6g: I have no idea what this figure refers to. What does "with" and "without" categories in the figure legend mean? How were these two categories defined and measured?

Response: We are so sorry for the misleading. In Fig. 6g, "with" refers to *Angela* LTR-RTs that share significant sequence similarity with the *Angela*-Rep-DDEE consensus element, which we reconstructed from conserved regions across DD and EE subgenomes. "Without" refers to those lacking such similarity. We classified these categories by aligning all intact *Angela* LTR-RTs to the consensus using BLASTN ($E\text{-value} \leq e^{-4}$); sequences meeting these criteria were designated "with," while all others were "without." This analysis was designed to test whether *Angela* elements sharing identity with the DD/EE consensus showed distinct age distributions, potentially reflecting activity during the polyploidization window. We have revised the Fig. 6g legend to state: the labels "with" and "without" have been revised to "similar to *Angela*-Dep-DDEE" and "dissimilar to *Angela*-Dep-DDEE", respectively. We have also added these classification parameters to the Supplementary Methods (Classification of Intact *Angela* LTR Based on *Angela*-Rep-DDEE Similarity). Thank you very much!

4. Fig. 6h: Please justify why the red signal for *Angela*-Rep-DDEE is completely zero

in the CC genome used for FISH (*O. eichingeri*), even if a few Angra repeats were found in the enrichment analysis (Fig. 6f).

Response: Thanks for your good suggestion. The apparent discrepancy is due to the different detection limits of the two methods. The computational analysis in Fig. 6f is highly sensitive and detected trace amounts of the *Angela*-Rep-DDEE repeat in the CC genome. However, its abundance is below the visualization threshold of the FISH experiment in Fig. 6h, where a signal can only form if the tandem repeats are sufficiently concentrated. Therefore, the repeat is present in trace amounts but not visibly detected by FISH. Thank you very much!

5. Lines 438-440: A standardized dataset containing 100,0000 (?) Illumina sequencing reads per genome (AA, BB, CC, EE, FF, GG, CCDD) were used to perform comparative analysis using RepeatExplorer to detect shared/lineage-specific LTR patterns. Why did the author decide to use Illumina reads when they could use ONT ultra-long reads for this analysis? Did the authors verify the results of RepeatExplorer on Illumina reads by checking/searching for LTR patterns in their high-quality ONT reads?

Response: Thank you very much for your valuable suggestion. We used Illumina reads for RepeatExplorer analysis for two main reasons aligned with standard practice in the field. The RepeatExplorer is a specialized, widely-used pipeline designed and optimized for short-read data (like Illumina). It relies on precise, high-accuracy reads to perform comparative clustering and *de novo* assembly of repetitive elements. This approach is established for detecting shared and lineage-specific repeat patterns. While ONT ultra-long reads are superior for genome assembly due to length, their higher per-base error rate makes them less ideal for the precise, comparative sequence clustering and consensus building required for fine-scale, lineage-specific repeat analysis. The repeat landscape patterns identified by RepeatExplorer were consistent with our independent analyses of the complete genome assemblies, providing cross-validation.

Comments on clarity, accuracy, argumentation/rationale:

1. Some wordings are imprecise and bring to wrong conclusions. For example, in line 56: "Due to the extinction of the diploid DD genome, researchers commonly use the DD subgenome of CCDD to study the evolutionary relationship between EE and DD in *Oryza*." Actually, the DD genome species is presumably extinct, but it might not have been discovered yet. To be proven extinct, archaeological remains of a diploid DD genome should be found. It's not DD extinction the connecting factor here, but the lack of a diploid DD genome to be compared with other genome types. On that regard, I would reword the title of the manuscript as well.

Response: Thank you very much for your good suggestion. We have revised the

sentence as "In the absence of an extant diploid DD genome, researchers commonly use the DD subgenome of the CCDD allotetraploid as a proxy to study the evolutionary relationship between the EE and DD lineages in *Oryza*" (Lines 59-62). Thanks a lot. Furthermore, we have also revised the manuscript title as "T2T Assembly of *Oryza australiensis* Reveals Transposon-Driven Centromere Reposition and Shared EE-DD Ancestry".

2. Another example is about divergence time estimation using single-gene phylogenetic analysis (lines 66-69): "The substantial discrepancy in divergence times between studies is perplexing and likely stems from methodological limitations: single-copy genes are prone to biases caused by post polyploidization evolutionary rate heterogeneity, and short sequence fragments (e.g., *Adh1*) inadequately capture deep divergence events". Analysis of multiple single-copy gene phylogeny from Fornasiero et al. (2025) showed estimates of diverge times comparable to those provided by Long et al. (2024). The difference with Ammiraju et al. is due to the methodology: at the time (i.e. 2010), they used a single locus (and a few genes) while today's estimates are based on hundreds/thousands loci. Furthermore, the confidence interval (1.71 – 8.89 My) of the DD-EE divergence time (5.13 My) provided by the authors is very broad and overlaps with estimates from Fornasiero et al. (4.9 My; confidence interval: c.a 4 – 6 My), Long et al. (4.44 My; no confidence interval) and Ammiraju et al. (7.08 My; confidence interval: c.a 2.5 – 8.5 My).

Response: We appreciate your important clarification. Recent phylogenomic studies show remarkably consistent estimates for the DD-EE divergence (Long et al.: 4.44 Mya; Fornasiero et al.: 4.9 Mya) with overlapping confidence intervals. We have revised the paragraph extensively (Lines 70-75) to reflect this consensus. The revised text now explicitly acknowledges that divergence time estimates have largely converged with large-scale phylogenomic data, while appropriately contextualizing the older Ammiraju et al. estimate (7.08 Mya) as reflecting earlier single-locus methodological limitations rather than true biological discordance. We also note that our own estimate of 5.13 Mya (CI: 1.71–8.89 Mya) overlaps substantially with these recent studies, supporting the reviewer's observation. At the same time, we maintain that even robust phylogenomic approaches based on single-copy genes have inherent constraints in polyploid systems, where post-polyploidization rate heterogeneity and selective pressures can introduce complexities. Therefore, while we agree that recent estimates show encouraging convergence, we believe integration with non-coding evidence (e.g., transposon dynamics) remains valuable for independent validation. Thank you very much!

3. Some statements should be rephrased and better contextualized. For example, in lines 73-75: "To date, the chromosome-level reference genome of *O. alta* has been released. However, high-quality *O. australiensis* genome sequence and their annotations are still

lacking.". Actually, high quality *O. australiensis* reference genomes have been released recently by different groups: Fornasiero et al., Nat Gen 2025; Morrison et al., Sci Rep 2025.

Response: We are sorry for those mistakes. We have corrected those parts as suggested. Thank you very much!

4. Not always statements are congruent throughout the manuscript. For example, in line 214 of the Results: "Unlike *O. sativa*, the EE genome exhibits centromere repositioning and neocentromere formation 3 on multiple chromosomes". But then in lines 588-590 in the Discussion: "67 high-quality functional centromeric maps of the *Oryza* AA group revealed reposition and neocentromere formation in multiple genomes". See also comment above on the content of Lines 215-218.

Response: Thank you very much for your careful reviewing. We have revised the relevant statements to ensure clarity and consistency. The comparative claim against *Oryza sativa* has been removed from Line 214, which now objectively states the finding within the EE genome. Furthermore, we have carefully re-evaluated the description of the cited work in the Discussion (original manuscript lines 588-590). That study primarily conducts a genome-wide, population-level analysis of CEN155 array organization patterns and a statistical neocentromere formation tendency (NFT) across AA genomes. While it provides valuable insights into structural patterns and probabilistic trends, its focus and methodology differ from the direct, empirical observation of specific and localized centromere repositioning and neocentromere formation events that we document in the EE genome. To avoid conflation of these distinct analytical perspectives and to maintain precise terminology, we have removed this reference. Correspondingly, we have ensured that the supporting statements in Lines 215-218 are consistent with the revised narrative. Thanks a lot.

5. Another example is in line 359: (Phylogenomic Reconstruction Resolves *Oryza* Divergence Times and EE/DD Clade Affiliation): this title is not consistent with the Author's conclusions. Does phylogenomic reconstruction resolve divergence time in *Oryza* or does it provide ambiguous results (see lines 547-549 in the Discussion) and therefore the authors adopted the LTR-RT insertion time method? Also, the EE/DD clade affiliation is not resolved here, but it is consistent with previous literature.

Response: Thank you very much for the valuable suggestion. We have replaced title with a more accurate one: "Phylogenetic Tree Construction and Divergence Time Estimation in *Oryza*" (Line 380). Thank you very much!

Comments on Figures: - Many figures lack figure legends, and/or proper axis and axis titles/labels, and/or accurate descriptions in the figure captions.

Some examples follow:

Fig. 2a: Genomic coordinates on the x axes are not plotted. It is difficult to tell whether, for each chromosome, the plotted regions have the same length across all panels since they are plotted on the same x-axis length.

Response: We have added the genomic coordinate labels to the x-axes in Fig. 2a. Thank you very much!

Fig. 2b: Sequence comparison of centromeric satellite repeats in the BB, CC, and FF genome types are missing from the figure.

Response: We have added the alignment results of centromeric satellite repeats for AA, BB, CC, CCDD, EE, and FF genomes in Supplementary Figure 7. Thank you very much!

Fig. 2f: How are "active" and "inactive" centromere defined? How were they measured?

Response: We are sorry for the misleading. "active centromeres" were defined as the genomic regions exhibiting significant CENH3 binding, as identified by ChIP-seq peaks. These regions represent the functional centromeres. Conversely, "inactive centromeres" refer to loci that contain abundant centromeric satellite repeats (CentO) but show no significant CENH3 binding signal. We have added the corresponding description in the figure legend. Thank you very much!

Fig. 4: Figure legend for panels a), b), and c) are missing, and the figure caption does not describe the plotted features.

Response: We have added the corresponding legend to Fig. 4. Thank you very much!

Figure 4d) and e) miss a figure legend showing the % of sequence identity and the colour code of the heat map.

Response: We have added a figure legend to both panels (Fig. 4d and 4e) indicating the % of sequence identity and the corresponding colour code for the heat map. Thank you very much!

Fig. 5d: Please add correlation coefficients and confidence levels of the intervals for the three curves.

Response: The confidence intervals are represented by the shaded areas in the figure, and we have described this in the figure legend. The correlation coefficients have also been added to the legend. Thank you very much!

Fig. 6e: Is this the comparison between *O. australiensis* (EE) and the DD subgenome of *O. alta* (see figure 6e) or the comparison "between *O. australiensis* and *O. sativa*

(NIP)" (see figure 6e caption)?

Response: Yes, the figure compares the EE and DD genomes. We have corrected it accordingly. Thank you very much!

Suppl.Fig. 4: Y-axis labels showing values for the coverage are missing. The "gene" track at the bottom of each panel is barely visible and not described in the figure caption.

Response: We have added the Y-axis coordinates. Since the displayed genomic interval is large, individual gene features would appear very small; to avoid potential confusion, we have removed the gene track from the visualization. Thank you very much!

Suppl.Fig. 5: Necat software is mentioned in the figure legend, but it is not described in the methods.

Response: We have described the parameters of the NECAT software in the Supplementary Methods section. Thank you very much!

Some figures were never mentioned in the manuscript (i.e. Suppl. Fig. 8 and 11).

Response: We have cited Suppl. Fig. 8 and Suppl. Fig. 11 in the manuscript. Thank you very much!

Each figure consists of many panels, the size is often too small for details to be clearly seen, and colours are often pale and barely visible (e.g. Fig. 2, Fig. 6).

Response: We have enhanced the color and detail in Figures 2 and 6 to improve visual clarity. Additionally, the panels within these figures have been rearranged to achieve a more logical layout and further enhance visual clarity. Thank you very much!

Specific comments on Suppl. Fig. 5 I struggle understanding Suppl. Fig. 5 and how the results showed support the "consistent assembly results" statement (line 186) in the main text. What is it plotted on the x and y axis (what are the target and query)? The axes titles are too generic/Figure caption is not exhaustive (are these centromere regions or whole chromosomes?) In the plot, there are two alignments that fall in the 65% identity range. What can the decrease of sequence identity be attributable to? To show assembly contiguity (i.e. no sequence gaps) in the centromeric regions, I would plot the ChIP-Seq read depth along centromeres and flanking regions and show the genomic coordinates of the contigs within each assembled pseudomolecule.

Response: Thank you very much for your valuable suggestions. Follow your suggestions, we have revised Supplementary Figure 5 (now Supplementary Figure 6) to clarify its purpose and evidential logic by directly comparing centromeric regions from the independent NECAT and NextDenovo assemblies; the updated axis labels now specify each assembly in megabases and the expanded caption explicitly states this

centromere-focused scope. The continuous diagonal pattern indicates structural concordance in sequence order and contiguity between the two assemblies, which underlies the reliability statement in the main text. The alignments showing approximately 65% identity reflect base-level errors inherent in the unpolished Nanopore reads used for the NECAT assembly at this stage, rather than structural gaps that would disrupt the diagonal alignment. To further support this interpretation, we have added a detailed step-by-step description of the comparative dot plot methodology in the Supplementary Methods. Regarding the ChIP-seq depth suggestion, we agree this provides orthogonal validation and note that Supplementary Figure 5 already displays CENH3 distribution across centromeres and flanking regions; to enhance transparency further, we have now annotated the revised Supplementary Figure 6 with genomic coordinates showing the constituent contig boundaries within each pseudomolecule, making the gap-free structure explicitly visible. Thank you very much!

Comments on Methods: Methods for DNA and RNA isolation and library preparation are not detailed enough (e.g. the voucher used, details on protocols/kits/versions).

Response: Thank you very for the good suggestions. We have added detailed descriptions of DNA/RNA isolation and library preparation in the Supplementally Methods section, including the experimental protocols and kit specifications. Thanks a lot.

Methods for data analysis or settings/parameters used are often missing (see also specific comments above). The sequential/logical order of method description is hard to follow and not always correct, with an impact on clarity and reproducibility. Sometimes incorrect terminology is used. Only one verb tense should be used. - Line 649. Is there any voucher number for the accession of *O. australiensis* used for the sequencing?

Response: Thank you very much for your valuable suggestions. We have systematically revised the Methods section to address these concerns. The descriptions for data analysis, including all software settings and key parameters, have been comprehensively added. The section has been reorganized into a clear, logical sequence to enhance readability and reproducibility. Terminology has been standardized for accuracy, and verb tenses have been unified throughout. Regarding the voucher number for the *O. australiensis* accession used for sequencing, the relevant specimen and source details have been specified in the revised Materials subsection. Thanks a lot.

Line 650. Please add reference to the CTAB protocol used, and details on the ONT and Illumina library preparation protocol used (kit, version, library insert size).

Response: Thank you very much for the good suggestion. We have revised the

"Methods" section to provide a detailed description of the DNA extraction, Oxford Nanopore (ONT), and Illumina library preparation protocols. The protocols are based on the standardized procedures provided by our sequencing service provider, and the description now includes the specification of all commercial kits used (with their versions), an outline of the key steps in library construction. Thanks a lot.

Line 658. Please add a brief description of the sample collection methods. Add protocol, kit and version used for RNA isolation and library preparation. Add amount of data produced per tissue.

Response: We have supplemented the Methods section with a brief description of the sample collection process. Details regarding the protocol, specific kits and their versions used for RNA isolation and library preparation have been added. The amount of sequencing data generated per tissue type is now also specified. Thank you very much!

Line 665. Please specify the company providing the antibody and the peptide sequence of the antigen used to produce the antibody.

Response: We have updated Line 665 (now lines 678-680) to specify the commercial supplier of the CENH3 antibody and have included the peptide sequence of the antigen used to generate this antibody. Thank you very much!

Line 666. Were biological replicates produced?

Response: Regarding biological replicates for the CENH3 ChIP-seq experiment, we clarify that its primary goal was the qualitative identification of CENH3-binding regions, not quantitative comparison. A single, high-quality ChIP-seq dataset is generally considered sufficient for robust locus identification in such studies. To enhance the reliability of our annotations, we conducted an additional, independent CENH3 ChIP-seq experiment. The results of both experiments are fully aligned and mutually corroborative, confirming the same centromeric regions. The new CENH3 ChIP-seq data have been uploaded to the National Genomics Data Center under CRX2406326. Thank you very much!

Gene and repeat annotations: This section is untidy. I suggest reorganizing and reordering the sections to make the reading smoother.

Response: Regarding the "Gene and repeat annotations" section, we have thoroughly reorganized and restructured it to present the annotation pipeline in a logical, step-by-step manner, moving from structural annotation to functional characterization (Lines 695-715). Thank you very much!

Line 690. The Github link doesn't work.

Response: This link works fine now: <https://github.com/meiyang12/Genome-annotation-pipeline>. Thank you very much!

Line 693. I believe there's a typo as a fasta file is not a command.

Response: We are sorry for this mistake. We have corrected this as suggested. Thank you very much!

Line 696. Did the authors mean: Homology based evidence was used (not 'generated') in GenomeThreader to perform gene prediction.

Response: We are sorry for this mistake. We have corrected this as suggested. Thank you very much!

Line 705. Functional annotation should be moved after gene prediction.

Response: We have revised the relevant text in the manuscript as suggested. Thank you very much!

Centromere analysis. - How did authors handle duplicates and secondary alignments?
- How were centromeric region boundaries defined?

Response: To handle duplicates and secondary alignments, we used samtools view -q 30 to retain unique mapped reads, followed by Picard MarkDuplicates to remove PCR duplicate reads. For defining centromeric region boundaries, we first called broad peaks of CENH3 with MACS2 using the parameters: -f BAM -g 9e8 --broad --broad-cutoff 0.1. Peaks with ChIP signals greater than 50 were selected, and Bedtools was used to merge peaks within 150 kb. Finally, the start and end coordinates of each chromosomal centromeric region were determined based on the merged CENH3 ChIP-seq signal (Lines 716-723). Thank you very much!

Identification of centromeric satellite repeat sequences. - The sequence of the identified satellite repeats is missing.

Response: The sequences of the identified satellite repeats have been provided in Supplementary Table 7. Thank you very much!

Lines 723-724. Please specify which species these genome types refer to.

Response: Thanks a lot. We have specified the corresponding *Oryza* species names for all genome types (e.g., AA, BB, CC, EE, FF, GG) mentioned in the manuscript.

Phylogenetic analysis. - Line 735. Protein homology (and not "sequence") was detected by OrthoFinder.

Response: We have been changed "sequence" to "homology" as suggested. Thank you very much.

Differential Analysis of LTR Sequences. - Line 756. There are no methods describing how insertion time was estimated based on differences between flanking LTR sequences in the species.

Response: The insertion times were calculated using LTR_retriever, which estimates the divergence time between the paired 5' and 3' LTR sequences of each intact element based on a molecular clock model. This method and the key parameters used are now clearly stated in the revised manuscript. Thank you very much!

- Line 757. Define fLTR.

Response: To ensure terminological precision and clarity, we have uniformly revised the manuscript by replacing the abbreviation "fLTR" with "5' LTR" throughout the text. This change was made to avoid any potential ambiguity and to explicitly refer to the 5' long terminal repeat sequence extracted from each intact LTR retrotransposon element. The corresponding descriptions have been updated in the Methods section and Supplementary Methods, and consistency has been maintained across the entire manuscript. Thank you very much!

- Lines 762-764. This sentence is hard to follow and the meaning not clear.

Response: We have revised the corresponding sentence in the Methods section to improve its clarity. Thank you very much!

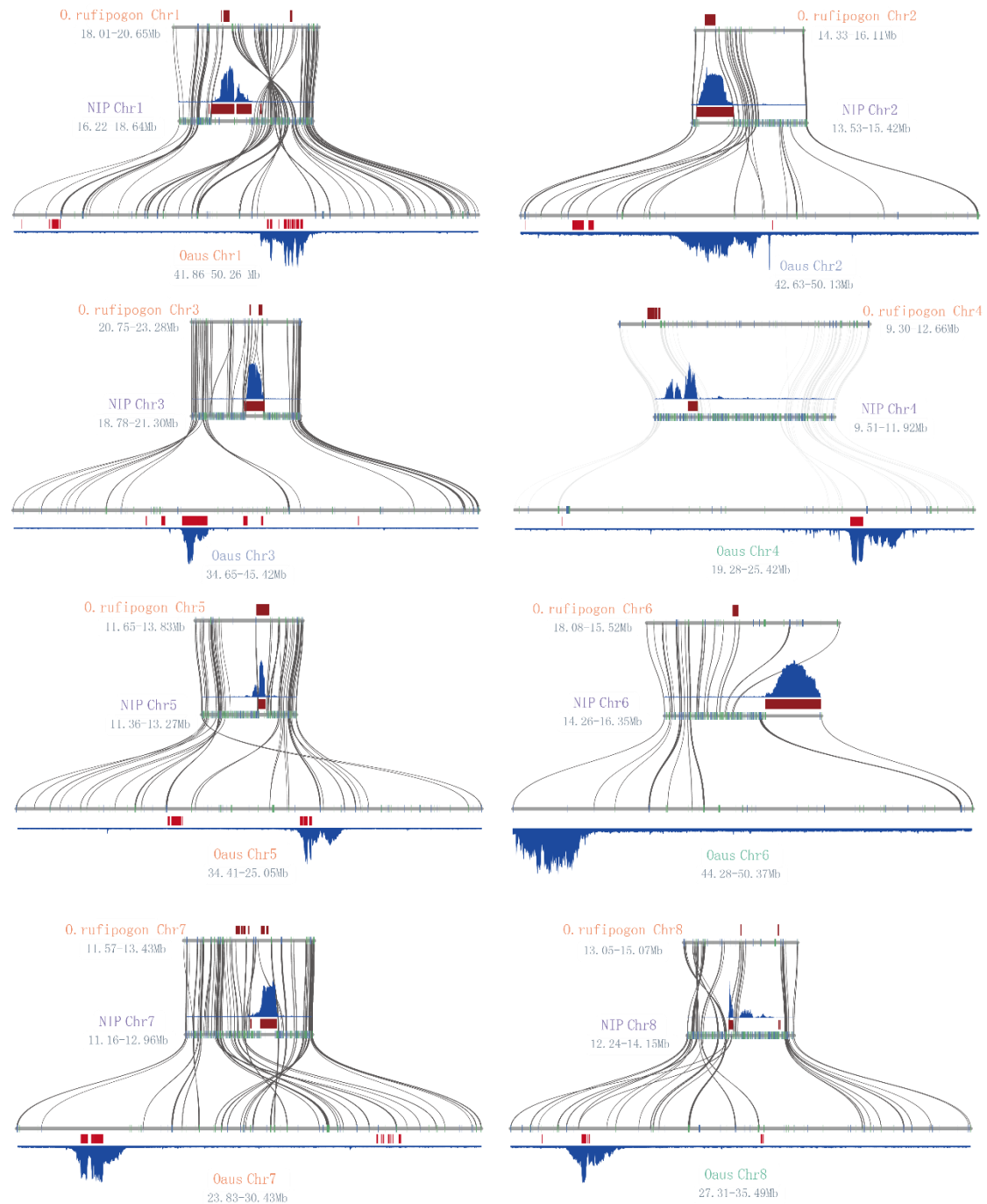
Identification of DD and EE genomic repeat markers - Lines 777-779. Why did the authors use short reads to find the repeat sequences instead of the ONT ultra long reads from their *O. australiensis* assembly and the PacBio reads from the *O. alta* assembly? Did the authors verify whether the sequences identified with RepeatExplorer can be found in the ONT and Pacbio reads?

Response: We selected RepeatExplorer for repeat identification as it is a specialized, widely adopted pipeline optimized for short-read data, which provides high base accuracy essential for precise comparative clustering and de novo assembly of repetitive elements. While ONT and PacBio reads are invaluable for genome assembly due to their length, their higher per-base error rates make them less suitable for the fine-scale, comparative sequence analysis required to identify and differentiate shared and lineage-specific repeats at the whole-genome level. Regarding verification, although we did not explicitly realign the consensus repeats to the raw ONT/PacBio reads, the biological relevance and specificity of the identified EE- and DD-enriched repeats were robustly validated by FISH. Thank you very much!

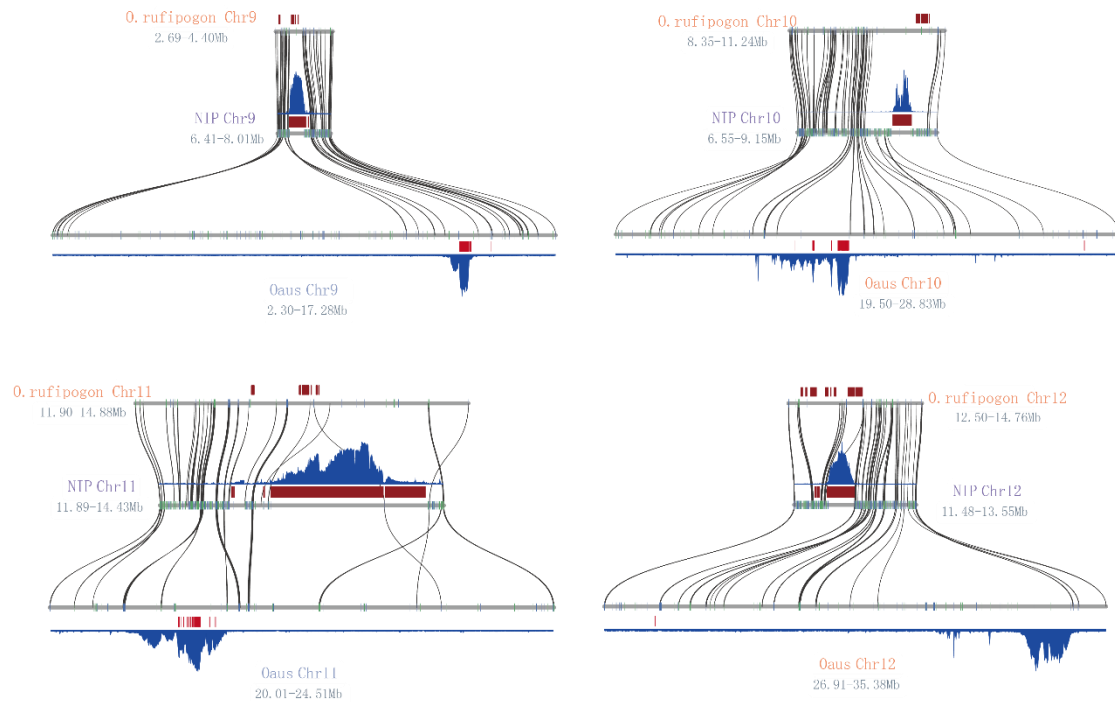
Reviewer3 :

1. Weak Evidence for Centromere Repositioning: The hypothesis of centromere shift/repositioning in *O. australiensis* is not well supported by the current comparisons. Given the significant phylogenetic distance between *O. australiensis* and *O. sativa*, the observed centromeric changes are more likely the result of artificial selection and diversification within the *O. sativa* lineage. To robustly test their hypothesis, the authors must use a closer relative (e.g., *O. rufipogon*) as the primary comparator. A three-way analysis including *O. rufipogon*, *O. sativa*, and *O. australiensis* is essential. As Figures 4 and S6 suggest, the major shifts appear to be a feature of *O. sativa*, not its wild relatives.

Response: Thank you very much for your good comments and valuable suggestions. We have conducted a comparative analysis of centromere positions among *O. rufipogon*, *O. sativa*, and *O. australiensis*. Due to the lack of available CENH3 ChIP-seq data for *O. rufipogon*, we used the CentO enrichment regions as its centromeric loci. The comprehensive analysis across all 12 chromosomes reveals a high degree of conservation in centromere positions between *O. rufipogon* and *O. sativa* (Nipponbare). Specifically, the centromeres are precisely collinear and localized to identical genomic regions in the majority of chromosomes (Chr 1-5, 7-10, and 12). The two exceptions are: (i) on Chr6, where the entire centromeric region is inverted, and the *O. rufipogon* centromere maps to the left side relative to Nipponbare; and (ii) on Chr11, where the Nipponbare centromere shows a slight rightward offset. This indicates that the major shifts we observed are not a specific feature of cultivated *O. sativa* and its domestication history. The significant divergence in centromere positions seen in *O. australiensis*, relative to this conserved wild-cultivated baseline, therefore suggests that centromere repositioning may be a distinctive evolutionary event in the EE genome lineage. These results have been incorporated into the revised manuscript (Lines 614-623) and Supplementary Figure 16-17. Thank you very much!



Supplementary Figure 16. The pairwise synteny visualization of centromere regions (Chr1-Chr8) among *O. australiensis*, *O. sativa* (NIP) and *O. rufipogon*. Red lines represent the distribution of centromeric satellite repeats, while gray lines connect syntenic gene pairs to indicate homologous regions. Histograms (blue peaks) show the distribution of CENH3 ChIP-seq signals, defining functional centromeres in *O. sativa* and *O. australiensis*. Due to the absence of CENH3 data, centromere loci in *O. rufipogon* were inferred based on CentO enrichment. The analysis reveals conserved centromere positioning between *O. rufipogon* and *O. sativa* across Chromosomes 1–8, with notable positional shifts on Chr6.



Supplementary Figure 17. Comparative analysis of synteny and centromere positions across *O. rufipogon*, *O. sativa* (Nipponbare, NIP), and *O. australiensis*. Red lines represent the distribution of centromeric satellite repeats, while gray lines connect syntenic gene pairs to indicate homologous regions. Histograms (blue peaks) show the distribution of CENH3 ChIP-seq signals, defining functional centromeres in *O. sativa* and *O. australiensis*. Due to the absence of CENH3 data, centromere loci in *O. rufipogon* were inferred based on CentO enrichment. The analysis reveals conserved centromere positioning between *O. rufipogon* and *O. sativa* across Chromosomes 9–12, with notable positional shifts on Chr11.

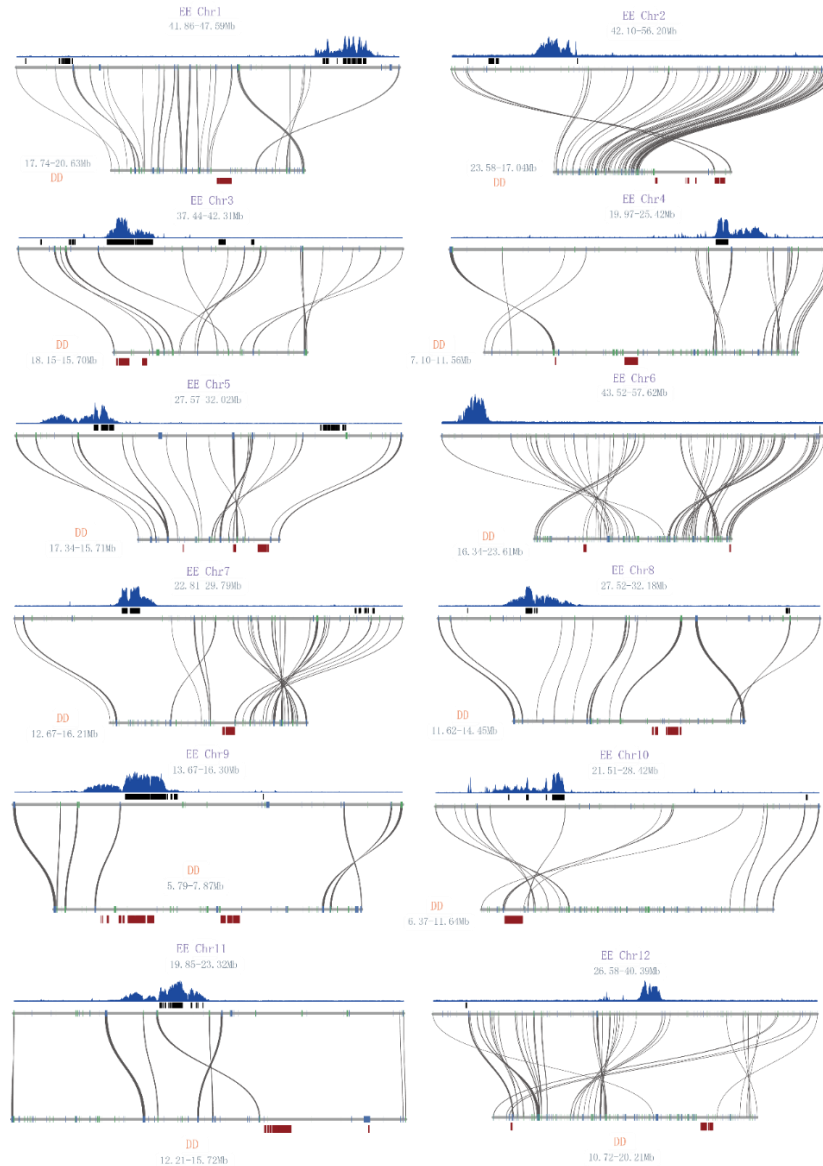
2. Unsupported Proposed Mechanism: The model in Figure 7 lacks compelling evidence from the data presented and should be either removed or significantly revised to reflect the more limited and alternative conclusions.

Response: We agree that the proposed model in the original Figure 7 extended beyond the direct evidence presented in our study. Therefore, in accordance with your suggestion, we have removed Figure 7 from the revised manuscript. This ensures that the conclusions drawn are fully and rigorously supported by the experimental and computational data.

3. The description of the evolution relationship between EE and DD subgenomes in the tetraploid should be simplified. If the assembly quality permits, a more insightful analysis would be to compare centromere positions between the diploid and tetraploid forms.

Response: As advised, we have simplified the description of the EE and DD evolutionary relationship in the revised text. We have also performed the diploid-tetraploid comparison you recommended, analyzing centromere positions between the

diploid EE genome and the DD subgenome within the CCDD tetraploid. Due to CENH3 ChIP-seq data are not available for the tetraploid; thus, we used CentO satellite repeat enrichment as a surrogate indicator for centromere loci, following standard practice in *Oryza* genomics. This comparison across all 12 chromosomes reveals a consistent evolutionary narrative characterized by four distinct patterns: absolute positional conservation (chromosomes 2, 9, 10), lineage-specific shifts (chromosomes 1, 4, 7, 11), and intriguing cases where ancestral CentO loci remain collinear despite neocentromere formation in one or both lineages (chromosomes 3, 5, 6, 8, 12). The prevalence of shared ancestral loci strongly suggests that the EE and DD genomes inherited a congruent centromere architecture prior to divergence, with subsequent modifications driven by lineage-specific transposon activity or genomic reorganization. Given that these results rely on sequence features rather than functional validation, we agree that they are best presented as supporting evidence. These results have been incorporated into the revised manuscript (Lines 624-635) and Supplementary Figure 18.



Supplementary Figure 18. Comparative analysis of centromere positions between the diploid EE genome and the DD subgenome of *O. alta* (CCDD).

For each of the 12 chromosomes, the centromeric regions are depicted. Black and red lines represent the distribution of the centromeric satellite repeat in the EE and DD subgenome, respectively. Grey lines connect syntenic gene pairs, indicating collinear regions. The comparative analysis reveals four distinct patterns: (1) positionally conserved centromeres (Chr 2, 9, 10); (2) positionally divergent centromeres (Chr 1, 4, 7, 11); (3) EE-specific neocentromere formation with a conserved ancestral CentO locus aligned to the DD centromere (Chr 3, 5, 8); and (4) independent neocentromere formation in both lineages with a shared ancestral CentO locus (Chr 6, 12). The prevalence of shared ancestral loci strongly suggests that the EE and DD genomes possessed congruent centromere architecture prior to their divergence, with observed differences likely driven by post-divergence, lineage-specific evolution.

4. The discussion is overly long and complex. It should be shortened and focused on the core findings. Excessive detail on other retrotransposons should be reduced.

Response: We have significantly streamlined the Discussion section (Lines 556-656) as suggested. The revision focused on condensing the core arguments, primarily by: 1) substantially shortening the discussion on the evolutionary relationship between the EE genome and DD subgenome, and 2) tightening the narrative on centromere evolution. Excessive detail on other retrotransposon families has been reduced to maintain a clear focus on the study's key findings.

5. The abbreviation "CRM" must be replaced with "CRR" consistently throughout the text and all figures. The CRM means Centromere Retrotransposon of Maize.

Response: Thanks a lot. We have revised CRM to CRR.

Reference

- Bousios, A., Kakutani, T., & Henderson, I. R. (2025). Centrophilic Retrotransposons of Plant Genomes. *Annu Rev Plant Biol*, 76(1), 579-604. <https://doi.org/10.1146/annurev-arplant-083123-082220>
- Yang, R., Li, Y., Su, Y., Shen, Y., Tang, D., Luo, Q., & Cheng, Z. (2016). A functional centromere lacking CentO sequences in a newly formed ring chromosome in rice. *J Genet Genomics*, 43(12), 694-701. <https://doi.org/10.1016/j.jgg.2016.09.006>

Reviewer #2

Shared Transposon Amplification History Implicates EE as Direct Ancestor of the DD Subgenome

Q1: Lines 506-508 and lines 554-555 strongly state that the EE genome is the direct donor of the DD subgenome in the CCDD species, but the evidence provided do not convincingly support the authors' hypothesis of direct ancestry.

Response: Thank you for this insightful critique. After careful re-evaluation of our evidence, we agree that the statements in lines 506–508 and 554–555 were overly assertive. The data demonstrate significant structural similarity and temporal coupling between the EE genome and the DD subgenome at the transposon level, but these findings support shared ancestry rather than a direct donor relationship.

Accordingly, we have revised the manuscript accordingly, removing all wording that implies direct ancestry, and have consistently revised the conclusion to state: "the EE lineage and DD subgenome prior to CCDD formation share common ancestry with an EE-like progenitor" (Lines 506-507; Lines 552-554). This revised wording accurately reflects the evolutionary relationships revealed by transposon characteristics, including *Angela*-Rep-DDEE enrichment, insertion time distributions, and structural similarities, while acknowledging that we cannot exclude alternative scenarios such as contribution from an extinct EE-like progenitor more closely related to the DD genome than extant *O. australiensis*.

Q2: Line 424. "The critical 1.7 Mya period".

Given the sequencing technology and computational power at the time, Ammiraju et al (2010) used the sequence of a single locus (*Adh*) to estimate divergence times in the *Oryza* genus. They estimated that the diploid and polyploid C genomes (*O. officinalis* (Cd), *O. minuta* CC (CTM) and *O. alta* CC (CTA)) radiated approximately 1.7 million years ago, thus inferring a polyploidization time for *O. alta* (Cd- CTA) of ~1.7 million years ago (with an HPD interval of 0.7636–2.9052, Table 2 in doi: 10.1111/j.1365-313X.2010.04251.x).

The authors of the present manuscript are using this punctual estimate as the critical watershed value to find validation for their transposon-based analysis and support the

hypothesis of EE being the direct ancestor of the DD subgenome in the CCDD species. Given the availability of up-to-date divergence time estimates based on high-quality chromosome-level genomes and complete gene annotations 15 years later, I would not take the 1.7 Mya estimate as the absolute "reference value". I would consider a temporal interval (instead of a timepoint) and provide the transposon-based analysis in *O. australiensis* and *O. alta* DD proposed here as further evidence supporting the known phylogenetic relationship between these genomes.

Summarizing, I agree with the authors' comment in the rebuttal letter which says that "*integration using non-coding evidence (e.g. transposon dynamics) remains valuable*", but this claim is very different from demonstrating a direct ancestry contribution of the EE genome. Furthermore, given the urgency of the authors to reassess the DD-EE relationship using evidence independent of single copy genes (lines 75-77), the choice of referring their transposon-based estimates to a polyploidization time inferred using a single gene sequence is not convincing.

Response: This is an important and constructive critique. Using a point estimate of 1.7 Mya—derived from a single-gene (*Adh*) analysis with a wide HPD interval (0.76–2.90 Mya)—as an absolute reference value is methodologically problematic, especially given the advances in genomic technologies over the past 15 years.

In response, we have fundamentally revised our analytical approach to no longer rely on external time points for calibration. Instead, we use transposon evidence itself to independently estimate the timing of CCDD formation. By calculating LTR sequence similarity across different time intervals (Fig. 5d), we observed that when tracing back to more ancient time points, sequence similarity between the EE genome and the DD subgenome increases, whereas similarity between the CC subgenome and the DD subgenome decreases. Within the time interval of approximately 1.52–2.1 Mya, the fitted curves of EE–DD subgenome and CC–CC subgenome intersect with the CC subgenome–DD subgenome curve at ~1.81 Mya and ~1.73–1.92 Mya, respectively. This high concordance of intersection times indicates that the evolutionary histories of the EE lineage and the DD subgenome were temporally coupled with the formation of the CCDD polyploid.

We have reframed our argument to state that "the EE lineage and DD subgenome prior to CCDD formation share common ancestry with an EE-like progenitor" rather than claiming direct ancestry. This conclusion is supported by multiple independent lines of transposon-based evidence with internal consistency, rather than relying on a single external time point.

We are grateful for your guidance in moving us away from problematic single-gene calibrations toward a more robust, evidence-based framework. Thank you very much!

Q3: Lines 425-427. Given the results of the analysis, the conclusion that can be drawn is only qualitative (as the authors wrote, it shows that EE and DD genomes contain higher proportion of Copia LTR than the CC subgenome), while a quantitative outcome cannot be inferred. Fig 5c shows that LTR distribution over time do not overlap in the EE and DD genomes, suggesting that TE dynamics in these genomes are different. Furthermore, LTR distribution pattern over time is more similar when comparing the two subgenomes of *O. alta* than when comparing *O. australiensis* with either *O. alta* subgenomes. For completeness, I would add *O. officinalis* to the analysis as representative ancestor of the CC subgenome. In addition to the density plot, I would plot the underlying histogram as well.

Response: We really appreciate these valuable and constructive suggestions.

Our original emphasis on the 1.7–2.5 Mya window placed undue weight on a specific time point, potentially giving the impression of quantitative precision that the data do not support. We have removed this section from the revised manuscript to avoid misinterpretation.

Regarding the observation that EE and DD LTR distributions do not overlap: the LTR insertion time distributions of EE and DD show distinct patterns, reflecting their divergent evolutionary trajectories: after the lineages split, EE and DD underwent independent evolution with distinct LTR amplification dynamics. In contrast, as two subgenomes of the CCDD, CC and DD have experienced similar evolutionary processes following polyploidization (including potential transposon reactivation and genome stabilization), resulting in more similar LTR distribution patterns.

Following your recommendations, we have made the following revisions:

1. Added histograms beneath the density plots in Fig. 5c to better visualize the underlying data distribution and provide transparency about sample sizes at different time points.
2. Included LTR insertion time analysis for *O. officinalis* (representative of the CC genome ancestor) in Supplementary Figure 14. This addition provides a more complete comparative framework and confirms that the CC genome shows distinct LTR dynamics from both EE and DD.

These improvements have significantly enhanced the rigor, completeness, and interpretability of our manuscript. Thank you very much!

Q4: Fig 5d: the intersection of the interpolated curves of a few datapoints (calculated using only one individual per species, Fig 5d) at around 1.7 Mya (1.73-1.84 interval) is not a robust and well-supported evidence of EE being the direct donor of the DD subgenome to the CCDD species.

Response: This is a fair concern. Relying solely on single-individual data to infer evolutionary relationships would be insufficiently robust. The purpose of Fig. 5d is not to prove direct ancestry, but to estimate the timing of CCDD formation based on the concordance of intersection points across multiple genome pairs.

To address this, we have made the following revisions:

1. Removed all claims of direct ancestry throughout the manuscript, now consistently stating that "the EE lineage and DD subgenome prior to CCDD formation share common ancestry with an EE-like progenitor".
2. Added multi-individual validation: To address the single-individual sampling concern, we extended our analysis to include two additional CC diploid species (*O. officinalis* and *O. eichingeri*) (Supplementary Figure 15). Results from all three CC species show high consistency: *O. officinalis* ~1.73 Mya, *O. eichingeri* ~1.76 Mya, *O. rhizomatis* ~1.92 Mya. All three fitted curves intersect with the CC–DD subgenome curve within the ~1.73–1.92 Mya interval. While the EE genome is represented by only one extant species, the consistency among multiple CC species substantially strengthens the robustness of the estimated intersection time and supports a CCDD formation event at approximately 1.73–1.92 Mya.

3. Internal consistency check: the high concordance between the EE–DD intersection (~1.81 Mya) and the CC–CC intersections (~1.73–1.92 Mya) provides internally consistent evidence from independent genome comparisons, reducing the impact of single-individual sampling bias.

We thank the reviewer for prompting us to strengthen the robustness of our analysis through multi-species validation.

Q5: Finally, as the authors correctly noted (lines 455-458), the hybridisation time predicted by TE is significantly earlier than their estimate using single-copy genes (5.13 Mya) and the HPD found in the analysis is very broad (1.71 – 8.89, Fig 5a).

Response: We appreciate your noting this important discrepancy. We have given careful consideration to this issue and offer the following interpretation:

1. Potential impact of polyploidization on molecular clock dynamics: Following the formation of allopolyploids such as CCDD, subgenomes often undergo accelerated evolutionary changes, including gene loss, expression divergence and transposon reactivation and bursts. Single-copy genes, as functional elements, are subject to selective constraints that may cause variable evolutionary rates under these conditions, potentially violating molecular clock assumptions. In contrast, LTR sequences are largely neutral after insertion and accumulate mutations at a relatively constant rate, making them suitable as genome-wide molecular clocks.

2. Internal consistency of transposon-based evidence: Our transposon analysis provides multiple internally consistent lines of evidence: the intersection points of the fitted curves for EE–DD subgenome and CC–CC subgenome in Fig. 5d is highly congruent (~1.81 Mya and 1.73–1.92 Mya, respectively), and analyses of multiple CC diploid species further support this temporal framework. This multiple lines of evidence, with internal consistency, provides robust constraints for estimating the timing of CCDD formation.

3. Compatibility rather than contradiction: Regarding our statement that "the TE-predicted time falls within the 95% HPD of the gene-based estimate," we wish to clarify: we are not attempting to validate the TE result using the gene-based estimate. Rather, we aim to demonstrate that—despite being derived from different

molecular markers with distinct evolutionary properties—the TE estimate (~1.7–1.9 Mya) is not incompatible with the gene-based confidence interval (1.71–8.89 Mya). The fact that the TE estimate falls within the HPD range indicates that these two independent lines of evidence are consistent rather than contradictory.

We propose that the discrepancy between TE-based (~1.8 Mya) and gene-based (5.13 Mya) estimates may reflect the differential impact of polyploidization on neutral versus functional genomic elements. The TE-based estimate likely captures the actual hybridization event, while the gene-based estimate may be influenced by post-polyploidization evolutionary rate changes.

Q6: Line 487 and Fig. 6c: How can the authors conclude that amplification curves in EE and DD genomes are similar starting from 1.7 My backwards? The count ratio of *Angela* LTRs for *O. alta* DD : *O. australiensis* EE is approx. 1:2, between 1.2 and 1.9 Mya.

Why choosing exactly 1.7 My as the threshold value? The same 1 : 2 : 0 pattern (*O. alta* DD : EE : *O. alta* CC) begins at approx. 1.3 My in the plot. Besides, for completeness of Fig. 6c, I would add the histogram plot underlying the density plot and the *Angela* LTR insertion time distribution in *O. officinalis*.

Response: We appreciate your careful examination of Fig. 6c and acknowledge that our original phrasing was imprecise.

Regarding the description of amplification patterns: We agree that "similar amplification curves" was not accurately descriptive. The key observation supported by our data is not similarity in amplification magnitude, but temporal association between the amplification histories of EE and DD prior to CCDD formation. We have revised the text accordingly:

"Analysis of *Angela* LTR insertion times reveals that prior to CCDD formation (1.73–1.92 Mya), insertions were abundant in the EE genome and DD subgenome but almost absent in the CC subgenome and *O. officinalis*. This indicates that the amplification histories of *Angela* LTRs in the EE lineage and the DD subgenome were temporally coupled before CCDD formation." (Lines 481-485)

Regarding the 1.7 Mya threshold: The original selection of 1.7 Mya was based on

previously reported CCDD formation time from single-gene estimates. Following your earlier suggestion to avoid reliance on gene-based point estimates, we now refer exclusively to the 1.73–1.92 Mya interval derived from our transposon-based analysis as the timeframe for CCDD formation. All relevant descriptions have been updated accordingly.

Regarding the 1:2:0 pattern beginning at ~1.3 Mya: The reviewer correctly notes that the pattern of *Angela* LTR presence in EE and DD but absence in CC begins at approximately 1.3 Mya. However, we retain 1.73–1.92 Mya as the temporal reference point because this interval corresponds to the CCDD polyploidization event—the core biological process of interest—and is derived from independent transposon-based evidence rather than external calibration. Regardless of when *Angela* LTR activity in the CC subgenome began, the key point is that prior to CCDD formation (i.e., within the >1.73 Mya time window), *Angela* LTR insertions were present in both EE and DD but nearly absent in CC, supporting a shared amplification history between EE and DD before CCDD formation.

Following your recommendations, we have supplemented the density plots in Fig. 6c with underlying histograms to show the data distribution, and have incorporated the *Angela* LTR insertion time distribution for *O. officinalis* to provide a more complete comparative analysis.

Q7: Lines 499 -501 and Fig. 6e. What are the summarized statistics of the whole-genome synteny analysis of *Angela* LTRs? The example provided is ok, but it would be more useful to get a number at the whole genome level since this evidence is used to support the hypothesis of direct ancestry.

Response: Thank you very much for the valuable suggestion. We have added the requested whole-genome statistics. In response to your helpful suggestion to provide additional quantification, we carried out a comprehensive survey of all *Angela* LTR insertions older than 1.7 Mya in the DD subgenome and assessed their syntenic relationships with the EE genome. Among the 12 such LTRs identified in the DD subgenome, 8 (66.7%) were found to reside in syntenic blocks with corresponding LTRs in the EE genome, as detailed in new Supplementary Table 12.

Supplementary Table 12. Angela LTRs inserted prior to 1.7 Mya in the DD subgenome and their syntenic counterparts in the EE genome

DD genomic location	EE genomic location	LTR integrity status
ChrD3:23350000-23357133	Chr3:51492713-51493003	Non-intact
ChrD4:7719158-7727244	Chr4:20304173-20312607	Intact
ChrD4:9966132-9975565	Chr4:17617603-17626575	Intact
ChrD5:14028699-14035959		
ChrD5:16178273-16186361	Chr5:30007020-30007308	Non-intact
ChrD6:12717174-12724985		
ChrD8:9808543-9824011	Chr8:23619624-23619810	Non-intact
ChrD10:7584126-7591865	Chr10:21889623-21889912	Non-intact
ChrD12:7645460-7653337	Chr12:15506723-15515147	Intact
ChrD12:18123460-18125326		
ChrD12:19947852-19955789	Chr12:37559540-37559830	Non-intact
ChrD10:14352701-14359150		

Q8: Lines 521-523 and Fig. 6g. What thresholds/cutoffs make the intact Angela LTRs "similar" or "dissimilar" to the Angela-Rep-DDEE? The definition found in Suppl. Methods is too vague to be reproducible: "An intact Angela LTR was classified as "similar to Angela-Rep-DDEE" if its annotated genomic coordinates overlapped with a significant BLASTN hit."

Response: Thank you very much for pointing out the lack of clarity in our original description. We have revised the classification criteria as follows:

A BLASTN search ($E\text{-value} \leq 1 \times 10^{-4}$) was performed using the *Angela*-Rep-DDEE sequence as query against each target genome. Only hits with >80% sequence identity and >100 bp alignment length were considered significant. An intact *Angela* LTR was classified as "similar to *Angela*-Rep-DDEE" if its annotated genomic coordinates contained at least one significant BLASTN hit that fell entirely within the LTR region (i.e., hit start $\geq$ LTR start AND hit end $\leq$ LTR end). Intact LTRs without any such hit

within their boundaries were classified as "dissimilar to *Angela*-Rep-DDEE".

These detailed parameters have been added to the Supplementary Methods to ensure full reproducibility of our analysis.

Q9: Lines 544-546: This conclusion is not convincing. Why would the analysis in Fig. 6g and described in lines 539-546 represents evidence for direct ancestry of EE in the CCDD genome? This could, at most, show genomic/structural similarity at the TE level between the EE genome and DD subgenome. Only because there is no other extant candidate in the *Oryza* genus to test, it does not necessarily imply that the EE genome is the direct donor of the DD subgenome to the CCDD. There might be an extant DD species not yet discovered, or the direct DD progenitor of CCDD species might be extinct. The logic used to compare insertion time of "matching" Angela LTRs in *O. australiensis* and *O. alta* (lines 551-552: "In general, Angela LTR in EE and DD genomes had similar amplification events and structural characteristics before 1.7 Mya" and Fig. 6g) is not conclusive evidence as one could paradoxically draw the same conclusion by looking at the fraction of Angela LTRs in the EE genome matching Angela-Rep-DDEE in the GG genome between 1.7 and 3 Mya.

Response: Thank you for this important clarification regarding the limits of our evidence. We agree that the current evidence—while demonstrating significant structural similarity and temporal coupling between the EE genome and the DD subgenome at the transposon level—does not conclusively prove that the EE genome is the direct donor of the DD subgenome, nor does it establish a direct ancestral relationship. While our evidence strongly supports common ancestry between EE lineage and DD subgenome, we cannot exclude alternative scenarios, such as the existence of an undiscovered diploid DD species or contribution from an extinct EE-like progenitor that was more closely related to the DD genome than the extant *O. australiensis*. The alternative scenarios suggested by the reviewer (e.g., an undiscovered extant DD species, or an extinct direct progenitor) are entirely plausible scientific possibilities.

In response, we have revised the manuscript to remove any wording that might imply

a direct donor relationship. The conclusion is now consistently phrased as: "the EE lineage and the DD subgenome prior to CCDD formation share common ancestry with an EE-like progenitor" emphasizing common ancestry rather than direct descent.

Regarding Fig. 6g, the key observation is that intact Angela LTRs matching the Angela-Rep-DDEE sequence are predominantly found in insertion time windows older than 1.73 Mya (ranging from 1.73 to 3 Mya) in both the EE genome and the DD subgenome, whereas non-matching LTRs are concentrated in the 0–1.5 Mya window. This pattern is absent in the CC subgenome. While this suggests that the EE and DD lineages shared a distinct class of Angela LTR structures prior to CCDD formation—consistent with the hypothesis of common ancestry—we acknowledge that this alone cannot rule out other evolutionary scenarios.

In the Discussion, we now state: "The evidence supports a close homologous relationship between the EE lineage and the DD subgenome, consistent with their common ancestry from an EE-like progenitor. However, we cannot exclude the possibility that the direct progenitor of the DD subgenome was an extinct lineage more closely related to *O. australiensis* than any extant species, or that an undiscovered diploid DD species exists." (Lines 591-596)

Q10: Additional note. The LTR insertion time analysis (SanMiguel et al. *Nat. Genet.* 20, 43–45. 1998) has been widely used in the literature and it's a nice method to explore insertion time of complete LTR-RTs in genome species based on the nucleotide distance between the two flanking LTR sequences of a complete LTR-RT. The software used here, LTR_Retrieve, uses the above approach to calculate LTR insertion time. A caveat is base call accuracy of the sequencing reads, because sequencing errors are counted as mutations for age calculations and this might affect age distribution.

In the rebuttal letter, the authors argued that analysis of repeat patterns using ONT long reads would not be feasible given "their higher per-base error rates make them less suitable for the fine-scale, comparative sequence analysis required to identify and

differentiate shared and lineage-specific repeats at the whole-genome level". Based on the same logic (and not considering the fact that ONT sequences can be polished), then ONT long reads would not be suitable for calculating LTR insertion time in the present study.

Response: We appreciate this important technical note regarding sequencing accuracy. The LTR insertion time analysis in our study was performed on the assembled reference genomes, not on the raw ONT long reads. Genome assembly underwent rigorous quality assessment:

- Consensus Quality Value (QV): 41.76
- BUSCO completeness: 99.8%
- LTR Assembly Index (LAI): 23.79

These metrics demonstrate that the assembled sequences achieve accuracy levels fully suitable for precise LTR insertion time estimation.

Our previous statement that "ONT long reads are less suitable for repeat analysis" referred specifically to raw ONT reads, whose higher per-base error rates limit fine-scale comparative analysis of repetitive sequences. However, after assembly and polishing, the genome sequences achieve accuracy levels that fully support the analyses conducted in this study.

Centromere analysis.

Q11: Lines 218-221: This sentence is confusing. Are neocentromeres located in "satellite repeat depleted regions" (line 220) or in "regions enriched for satellite repeats" (line 226)?

Response: Thank you very much for pointing out this ambiguous description. To clarify the confusion about whether neocentromeres are located in satellite-depleted or satellite-enriched regions, we turn to the centromere maturation model (Piras et al. 2010; Gong et al. 2012). This model suggests that neocentromeres go through different evolutionary stages—they first appear in satellite-depleted regions (e.g., Chr6) and over time can gradually acquire satellite repeats through amplification, evolving into mature

neocentromeres that are satellite-enriched (e.g., Chr1, 3, 5, 7, 8, 10). Based on this framework, we now explicitly interpret the satellite-enriched functional centromeres on chromosomes 1, 3, 5, 7, 8, and 10 as "mature neocentromeres" in the main text. The revised text reads: "According to the centromere maturation model, neocentromeres can acquire satellite repeats over evolutionary time after their initial formation. Consistent with this model, we interpret the pattern on chromosomes 1, 3, 5, 7, 8, and 10 as representing mature neocentromeres. On these chromosomes, CENH3-binding assays revealed two distinct satellite repeat-enriched regions, with the larger region serving as the functional centromeric region (CENH3-binding site) and the smaller region lacking CENH3 association". In addition, we have further discussed this phenomenon and its evolutionary implications in the Discussion section (Lines 617-627).

Q12: Lines 225-227: Is the presence of satellite repeats outside the functional centromere regions (defined by CENH3 binding) alone sufficient evidence to demonstrate neocentromere formation?

The presence of interspersed satellite repeats outside the functional centromeres might be due to an expansion/novel insertion of TEs in the CENH3 region that pushed away the flanking satellite arrays. This is also stated by the authors in line 245-247 and supplementary Figure 8.

Response: This is an important point regarding evidence interpretation. The presence of satellite repeats outside the functional centromere region alone is not direct evidence for neocentromere formation—as the reviewer noted, this pattern could also result from transposon insertions that "push away" flanking satellite arrays.

However, our subsequent ancestral centromere analysis provides key evidence to distinguish between these two possibilities. If the pattern were due to transposon insertions, the "pushed away" satellite arrays would be located immediately adjacent to the functional centromere region, with continuous syntenic relationship and a small distance (typically <1 Mb). Instead, we observed:

- The satellite repeat-enriched regions lacking CENH3 are highly homologous to ancestral centromere positions identified through multi-species synteny

analysis

- These regions are located 2.3–6 Mb away from the functional centromeres
- The CENH3-associated functional centromeres are separated from the ancestral satellite arrays by multiple genes and extensive non-satellite sequences
- There is no direct syntenic continuity between the functional centromeres and the ancestral satellite arrays

This pattern cannot be explained by transposon insertions pushing satellite arrays. Instead, it fits the neocentromere formation model where new centromeres emerge at distant locations while ancestral centromeres gradually lose CENH3 binding and satellite repeats over evolutionary time.

Therefore, we have revised the main text to present this observation as a scientific interpretation based on established theoretical models, rather than as a direct conclusion from the presence of two satellite repeat-enriched regions alone (Lines 224-227). We appreciate your insight in helping us refine the presentation of this analysis.

Q13: Lines 251 - 253: It is not clear whether this sentence is referring to ref. 31-37-38; other studies (references missing?); or to the previous sentence ("among the 81 genes found by the authors, many (how many?) were found to be conserved and expressed in other studies" - references missing).

Response: We appreciate the reviewer noting this ambiguity. The phrase "Further studies showed that many of these expressed genes were conserved" was indeed unclear and lacked precision.

The term "conserved" in the original text refers to our own analysis based on data generated in this study, not to external studies. We performed syntenic comparison with the Nipponbare (AA) genome and found that 30 of the 34 expressed genes have orthologs in the AA genome, indicating evolutionary conservation across the *Oryza* genus.

We have revised the sentence with precise numbers and clearer terminology:

"Further comparative analysis with the Nipponbare genome revealed that 30 of these 34 expressed genes were syntenic orthologs, and all expressed genes were located at non-CENH3 binding sites." (Lines 254-256)

Q14: Fig 3d: how was TE density calculated in this figure?

Response: Thank you for requesting this clarification. The TE density in Fig. 3d was calculated as follows:

Step 1: Chromosome arm definition

Each chromosome was divided into short arm (p-arm) and long arm (q-arm) based on the centromere position determined by CENH3 ChIP-seq data:

- The midpoint of the centromeric region was used as the boundary
- Region from chromosome start to midpoint = short arm
- Region from midpoint to chromosome end = long arm

Step 2: TE density calculation

For each arm:

- $\text{TE density} = (\text{total length of all TE intervals}) / (\text{total arm length})$

For whole chromosome:

- $\text{TE density} = (\text{total TE coverage length}) / (\text{total chromosome length})$

These calculation details have been added to the Supplementary Methods to ensure full reproducibility.

Q15: Supplementary Figures 9-10: in addition to *O australiensis*, the figures show the CENH3 signal for two *O sativa* varieties and *O brachyantha*, but the methods do not describe ChIP seq experiment for the latter three species. How was the CENH3 signal showed in the plot of Suppl. Fig. 9 and 10 measured/quantified?

Response: We appreciate your pointing out this omission. The CENH3 ChIP-seq signals for *O. sativa* (two varieties: Nipponbare and MH63) and *O. brachyantha* shown in Supplementary Figures 9–10 was obtained from previously published studies.

Data sources:

- *O. sativa* (Nipponbare): SRR25203308
- *O. sativa* (MH63): SRR13285939
- *O. brachyantha*: SRR7351627

These public database accession numbers have been added to the Data Availability section (Lines 855–857) to ensure full traceability and reproducibility.

Q16: Structure and redundancy: methods description needs a clearer and logic narrative.

Methods should be re-organised in a way that the reader is not forced to continuously shift from the main manuscript to the supplementary material to find methods referring to the same experiment/analysis.

The authors should make a choice and decide what methods to put in the supplementary materials and what to keep in the main manuscript.

Response: We appreciate this constructive suggestion regarding manuscript organization. We have reorganized the Methods section to improve clarity and logical flow: core experimental procedures and key parameters are now fully described in the main text, while detailed parameters and secondary details have been consolidated in the Supplementary Materials.

Q17: Methods reproducibility

Although I appreciated the effort of the authors to improve method description, there is still some missing information.

Angela-LTR analysis: What thresholds/cutoffs make the LTRs similar or dissimilar to the Angela-Rep-DDEE? The definition found in Suppl. Methods is too vague to be reproducible: "An intact Angela LTR was classified as "similar to Angela-Rep-DDEE" if its annotated genomic coordinates overlapped with a significant BLASTN hit."

Response: Thank you very much for the good suggestion. Please see Response to Q8 for the detailed BLASTN parameters and classification criteria, which have been added to the Supplementary Methods to ensure full reproducibility.

Q18: "Genome de novo assembly and assessment" in Supplementary methods (Line 60): What was the total contig-level assembly size before the filtering step? How many contigs shorter than 1 Mb were discarded and what was the total size of the discarded portion?

1 Mbp looks like a high value to be used as cutoff. Were the discarded contigs all consisting of organellar DNA?

What were the L50, L90, N50, N90 values at the contig-level after removing organellar DNA? Suppl. Table 1 only shows the N50 value of the scaffold-level assembly.

Response: Thank you for these detailed questions about assembly statistics.

We note that the phrase "1 Mb" in our original description may have been misleading. 1

Mb was not a preset filtering threshold but rather an objective observation of the assembly results: after initial assembly, we obtained 13 contigs longer than 1 Mb, with only one contig falling below this size.

Filtering details:

- Only one contig (148 Kb) was discarded during filtering
- This contig originated from a breakpoint region (a region with ambiguous read coverage suggesting potential misassembly)
- It was subsequently incorporated into the final assembly after gap filling and validation
- It was removed initially as a precautionary measure to ensure assembly quality
- No organellar DNA contigs were discarded (organelle sequences were identified and removed in a separate step prior to nuclear genome assembly)

Contig-level L50, L90, N50, and N90 values have now been added to the Supplementary Table1. We have revised the description in the Supplementary Methods to clarify that 1 Mb was an observation, not a filtering threshold, and to provide complete assembly statistics.

Q19: Lines 63-64 in Supplementary methods: I don't understand how gaps were counted. If you selected 13 contigs for your assembly, you should end up closing only one gap, but the methods here report the presence of multiple gaps filled with LR_Gapcloser (see also comment below).

Supplementary Table 1: The counting of gaps is not correct.

In a linear genome assembly, the relationship between the number of contigs (n) and the number of gaps (g) is defined as: $g=n-1$

Given 12 chromosomes (= 12 scaffolds in a chromosome-level assembly), how can the assembly from Morrison et al have 13 contigs and three gaps? Same for the remaining assemblies?

Response: Thank you very much for raising this point, which has helped us identify an inconsistency in our gap counting methodology.

Upon re-examination, we realize that our gap counting method was unclear because we distinguished between two distinct types of gaps:

Type 1: Structural gaps (between contigs)

- These are the breaks between contigs that need to be joined to form complete chromosomes
- Formula: structural gaps = contigs - chromosomes
- For our assembly: 13 contigs - 12 chromosomes = 1 structural gap

Type 2: Intra-scaffold gaps (within chromosomes)

- These are stretches of consecutive 'N's within a chromosome sequence
- Represent regions of unresolved bases where order and orientation of flanking sequences are known
- These gaps remain after scaffolding and gap-filling efforts

Previously, we only counted intra-scaffold gaps (Type 2) and did not include the structural gap between contigs (Type 1) in our gap count, leading to the discrepancy with the reviewer's expectation.

We have now revised the counting method to include both types of gaps in Supplementary Table 1. Thank you very much!

Minor comments:

- Line 466: Please correct this number.

Response: Thanks. We have corrected the number from "100,0000" to "1,000,000".

- Line 281: Please define "p-arm".

Response: We thank you for this note. "p-arm" refers to the short arm of a chromosome. To improve clarity for a broader readership, we have revised the manuscript to consistently use the more descriptive terms "short-arm" and "long-arm" in place of "p-arm" and "q-arm" throughout the text.

Reviewer #4

Please note, however, the following recommended rephrasing of the paper's key evolutionary inference to clarify the English:

The following text (present in the abstract and elsewhere) is currently phrased awkwardly as: "...the EE lineage and DD subgenome prior to CCDD formation share common ancestry with an EE-like progenitor."

This idea should be rephrased as follows for greater clarity: "... prior to CCDD formation, the ancestral lineage that contributed the DD subgenome shared common ancestry with an EE-like progenitor."

This inference could also be written as: "...prior to formation of the CCDD polyploid lineage, the ancestral DD subgenome donor shared common ancestry with the ancestor of the EE genome lineage."

Response: We thank you for this careful and constructive suggestion. We have adopted the first recommended rephrasing and revised the sentence throughout the manuscript (Abstract, Results, and Discussion) to read: "prior to CCDD formation, the ancestral lineage that contributed the DD subgenome shared common ancestry with an EE-like progenitor."

ROUND 1 REVIEWER 2 ATTACHMENT:

T2T Assembly of *Oryza australiensis* Reveals Centromere Repositioning Mechanisms and Identifies the EE Genome as Direct Donor of the Extinct DD Lineage

The authors presented a telomere-to-telomere genome assembly of *Oryza australiensis* (genome type EE) using long and ultra-long ONT reads and Illumina reads. Genome annotation included telomere and centromere regions, transposable elements (TEs) and genes. One main goal was investigating mechanisms of centromere repositioning in *O. australiensis*. The other was providing support to the hypothesis of the EE progenitor being the direct donor of the DD subgenome in the CCDD tetraploids. For this purpose, the authors used two approaches to analyse *O. australiensis* and the DD subgenome of *O. alta* (CCDD) - i.e. they estimated DD-EE divergence time using single-gene phylogenetic analysis and compared LTR-RT sequences and their insertion times in these two species.

Overall, the research questions (the mechanisms of centromere repositioning in *O. australiensis*, and the analysis of LTR-RT elements to investigate the evolutionary history of DD and EE genomes) are of great interest. However, the methods lack description of important details or even main elements necessary to clearly understand the rationale/implementation of the analyses, the chosen methods do not accurately address the research questions, and the conclusions are not always supported by results.

The conclusion of EE genome being the direct donor of the DD subgenome in the tetraploid CCDD is hard to prove and it also depends on how genome types are defined (i.e. the original definition was based on interspecific hybridization and cytogenetics experiments to assess potential interfertility between species, but the analysis provided here is based on sequence comparison). The methods used by the authors are not described to enough detail to allow full understanding of the analysis. Results (visualized in figure 5 and 6) do not reveal a 'direct' relationship but show (to some extent) a close relationship between EE genome and DD subgenome, providing further analysis that supports previous literature (Ammiraju et al., The Plant Journal. 2010; Bao and Ge, Plant Syst. Evol. 2004).

My main general suggestion to the authors is to: 1) revise the Methods to give a precise, detailed description of all analyses performed and 2) draw conclusions that are clearly and robustly supported by results.

Attached, I provided in-depth comments/suggestions to four sections of the manuscript, and I grouped comments/suggestions across the manuscript into three categories referring to: clarity/rationale, figures and methods.

Centromere Mapping Reveals Repositioning and Transposon-Driven Neocentromere Formation in *O. australiensis*

The satellite repeat distribution analysis in the regions flanking the functional centromeres is lacking a temporal dimension (Fig. 2). How can the authors conclude that satellite repeats 'may represent sequential stages of neocentromere evolution' (line 219) if there is no temporal mean to investigate the sequential occurrence of these events?

Lines 215-218: "The neocentromere is present in Chr2, Chr6, and Chr12, where Chr2 has a large number of satellite repeats at a distance of 1.5 Mb from the functional centromere, Chr12 contains fewer satellite repeats, while Chr6 contains almost no satellite repeats". How can the authors conclude that a neocentromere is present on Chr 6 if no satellite repeat signal was detected on this chromosome?

Lines 219-222: "Additionally, functional neocentromere formation also occurs on chromosomes 1, 3, 5, 7, 8, and 10, where CENH3-binding assays revealed that large structural domains containing two distinct satellite repeat clusters serve as novel functional centromeric regions". What are these two satellite repeat clusters? How where they detected?

Line 227-230: “We further compared the insertion times of LTRs in active and inactive centromeres, and the results showed that active centromeres contain younger LTRs (Fig. 2f), suggesting that TE insertions may contribute to the formation of neocentromere”. How were active and inactive centromeres defined? What are their respective genomic coordinates? How were “younger LTRs” defined? Could you show the distribution of these centromeric LTRs along chromosomal coordinates?

Line 214-215: “Unlike *O. sativa*, the EE genome exhibits centromere repositioning and neocentromere formation on multiple chromosomes”. Reference(s) supporting the statement about *O. sativa* is/are missing.

To improve analysis of centromere repositioning, I suggest reading: Ahmed HI, Heuberger M, Schoen A et al., *Nature*. 2023. <https://doi.org/10.1038/s41586-023-06389-7>. This work nicely investigated centromere dynamics and centromere-specific retrotransposons insertion time in einkorn.

Evolutionary trajectory of centromeres

To reconstruct centromere evolution in the *Oryza* genus, the authors built a “multi-genome synteny map based on the centromeric core sequences and epigenetic characteristics of the cultivated species (Nipponbare and MH63) and wild species (*O. australiensis* and *O. brachyantha*)” (Lines 298-301).

Firstly, the authors should clearly define what are the genomic and epigenetic features used (genes, satellite repeats, CENH3, others?) and how the maps were built (I couldn’t find methods for this analysis). Secondly, the rationale of the “ancestral centromeres” needs to be clearly explained and justified. The authors defined as ancestral centromeres the syntenic features within centromeres and flanking regions across two *O. sativa* varieties (i.e. Nipponbare and MH6) and *O. brachyantha* (FF). For a start, the authors should describe in detail how synteny was defined for each feature under study, and how sequence homology was measured across species. Then, they should explain that conserved synteny of homologous regions among *O. sativa* varieties and *O. brachyantha* reflects the ancestral centromeric state, and finally they should present the comparative analysis of homologous regions in *O. australiensis*. Given the great evolutionary distance between AA, EE and FF genome types, it is not trivial to analyse homology of non-genic regions, and this should be thoroughly described in the methods and results. Furthermore, figure legend for Fig. 4a-c is missing. What are the features connected by the grey lines in these panels?

Shared Transposon Amplification History Implicates EE as Direct Ancestor of the DD Subgenome

Fig. 5c: I struggle seeing the secondary amplification peak at 1.7 Mya in *O. alta* DD. Furthermore, LTR-RT insertion time on the x-axis should be taken as a broad estimate (i.e. confidence intervals were not defined), and not as the precise timing of insertion.

Lines 433-434: “the nearly-parallel trends in the fitted curves of CC genome and CC subgenome, as well as EE genome and DD subgenome, suggest that EE genome may be the direct donor of the DD subgenome”.

How can the authors support their conclusions given the results showed in Fig. 5d? Why does the ‘nearly parallel trend of the fitted curves’ suggest that EE may be the direct donor of the DD subgenome?

To better understand results in Fig. 5d, the authors should include thorough description of the settings, parameters and thresholds used to identify and compare different LTR-RT families(?)/lineages(?) (line 419).

Angela LTR Retrotransposons and a Conserved 275-bp Repeat Validate EE as the Donor of DD

Line 455: “This suggests that Angela LTR was gradually lost during the evolution of *Oryza*,”

I don’t think Angela LTR elements were necessarily lost during the evolution of *Oryza*. An alternative simple explanation could be that expansion of Angela LTRs was more successful in the genome of *O. australiensis* than in other genomes. It has indeed been shown that, in *Oryza*, TE evolutionary history is independent

from the host genome type (see El Baidoury and Panaud, *Genome Biol. Evol.* 2013, [doi: 10.1093/gbe/evt025](https://doi.org/10.1093/gbe/evt025)).

Furthermore, Angela LTR RTs are a family of RTs consisting of many different types. What are the sequences considered for this analysis?

Fig. 6c: Again, I struggle seeing the peak of Angela LTR expansion at 1.7 My in *O. alta* DD that would support the author's claim of EE being the direct donor of the DD subgenome. The presence of two y-axes with the count of Angela LTRs in EE and CCDD, respectively, confuses me. Are these distributions directly comparable (i.e. element count is up to 175 and 25 for *O. australiensis* and *O. alta*, respectively)? Given that Angela LTRs were detected in many species with AA, BB, CC (diploid), FF and GG genome types, I suggest adding distributions of Angela LTR insertion times for each of these species in the Supplementary figures.

Fig. 6g: I have no idea what this figure refers to. What does "with" and "without" categories in the figure legend mean? How were these two categories defined and measured?

Fig. 6h: Please justify why the red signal for Angela-Rep-DDEE is completely zero in the CC genome used for FISH (*O. eichingeri*), even if a few *Angra* repeats were found in the enrichment analysis (Fig. 6f).

Lines 438-440: A standardized dataset containing 100,000 (?) Illumina sequencing reads per genome (AA, BB, CC, EE, FF, GG, CCDD) were used to perform comparative analysis using RepeatExplorer to detect shared/lineage-specific LTR patterns. Why did the author decide to use Illumina reads when they could use ONT ultra-long reads for this analysis? Did the authors verify the results of RepeatExplorer on Illumina reads by checking/searching for LTR patterns in their high-quality ONT reads?

Comments on clarity, accuracy, argumentation/rationale:

Some wordings are imprecise and bring to wrong conclusions. For example, in line 56: "*Due to the extinction of the diploid DD genome, researchers commonly use the DD subgenome of CCDD to study the evolutionary relationship between EE and DD in Oryza.*" Actually, the DD genome species is presumably extinct, but it might not have been discovered yet. To be proven extinct, archaeological remains of a diploid DD genome should be found. It's not DD extinction the connecting factor here, but the lack of a diploid DD genome to be compared with other genome types. On that regard, I would reword the title of the manuscript as well. Another example is about divergence time estimation using single-gene phylogenetic analysis (lines 66-69): "*The substantial discrepancy in divergence times between studies is perplexing and likely stems from methodological limitations: single-copy genes are prone to biases caused by post-polyploidization evolutionary rate heterogeneity, and short sequence fragments (e.g., Adh1) inadequately capture deep divergence events*". Analysis of multiple single-copy gene phylogeny from Fornasiero et al. (2025) showed estimates of diverge times comparable to those provided by Long et al. (2024). The difference with Ammiraju et al. is due to the methodology: at the time (i.e. 2010), they used a single locus (and a few genes) while today's estimates are based on hundreds/thousands loci. Furthermore, the confidence interval (1.71 – 8.89 My) of the DD-EE divergence time (5.13 My) provided by the authors is very broad and overlaps with estimates from Fornasiero et al. (4.9 My; confidence interval: c.a 4 – 6 My), Long et al. (4.44 My; no confidence interval) and Ammiraju et al. (7.08 My; confidence interval: c.a 2.5 – 8.5 My).

Some statements should be rephrased and better contextualized. For example, in lines 73-75: "*To date, the chromosome-level reference genome of O. alta has been released. However, high-quality O. australiensis genome sequence and their annotations are still lacking,*". Actually, high quality *O. australiensis* reference genomes have been released recently by different groups: Fornasiero et al., *Nat Gen* 2025; Morrison et al., *Sci Rep* 2025.

Not always statements are congruent throughout the manuscript. For example, in line 214 of the Results: "*Unlike O. sativa, the EE genome exhibits centromere repositioning and neocentromere formation*

on multiple chromosomes". But then in lines 588-590 in the Discussion: "67 high-quality functional centromeric maps of the *Oryza* AA group revealed reposition and neocentromere formation in multiple genomes". See also comment above on the content of Lines 215-218. Another example is in line 359: (Phylogenomic Reconstruction Resolves *Oryza* Divergence Times and EE/DD Clade Affiliation): this title is not consistent with the Author's conclusions. Does phylogenomic reconstruction resolve divergence time in *Oryza* or does it provide ambiguous results (see lines 547-549 in the Discussion) and therefore the authors adopted the LTR-RT insertion time method? Also, the EE/DD clade affiliation is not resolved here, but it is consistent with previous literature.

Comments on Figures:

- Many figures lack figure legends, and/or proper axis and axis titles/labels, and/or accurate descriptions in the figure captions. Some examples follow:

Fig. 2a: Genomic coordinates on the x axes are not plotted. It is difficult to tell whether, for each chromosome, the plotted regions have the same length across all panels since they are plotted on the same x-axis length.

Fig. 2b: Sequence comparison of centromeric satellite repeats in the BB, CC, and FF genome types are missing from the figure.

Fig. 2f: How are "active" and "inactive" centromere defined? How were they measured?

Fig. 4: Figure legend for panels a), b), and c) are missing, and the figure caption does not describe the plotted features. Figure 4d) and e) miss a figure legend showing the % of sequence identity and the colour code of the heat map.

Fig. 5d: Please add correlation coefficients and confidence levels of the intervals for the three curves.

Fig. 6e: Is this the comparison between *O. australiensis* (EE) and the DD subgenome of *O. alta* (see figure 6e) or the comparison "between *O. australiensis* and *O. sativa* (NIP)" (see figure 6e caption)?

Suppl. Fig. 4: Y-axis labels showing values for the coverage are missing. The "gene" track at the bottom of each panel is barely visible and not described in the figure caption.

Suppl. Fig. 5: Necat software is mentioned in the figure legend, but it is not described in the methods.

- Some figures were never mentioned in the manuscript (i.e. Suppl. Fig. 8 and 11).

- Each figure consists of many panels, the size is often too small for details to be clearly seen, and colours are often pale and barely visible (e.g. Fig. 2, Fig. 6).

Specific comments on Suppl. Fig. 5

I struggle understanding Suppl. Fig. 5 and how the results showed support the "consistent assembly results" statement (line 186) in the main text.

What is it plotted on the x and y axis (what are the target and query)? The axes titles are too generic/Figure caption is not exhaustive (are these centromere regions or whole chromosomes?) In the plot, there are two alignments that fall in the 65% identity range. What can the decrease of sequence identity be attributable to?

To show assembly contiguity (i.e. no sequence gaps) in the centromeric regions, I would plot the ChIP-Seq read depth along centromeres and flanking regions and show the genomic coordinates of the contigs within each assembled pseudomolecule.

Comments on Methods:

Methods for DNA and RNA isolation and library preparation are not detailed enough (e.g. the voucher used, details on protocols/kits/versions). Methods for data analysis or settings/parameters used are often missing (see also specific comments above). The sequential/logical order of method description is hard to follow and not always correct, with an impact on clarity and reproducibility. Sometimes incorrect terminology is used. Only one verb tense should be used.

- Line 649. Is there any voucher number for the accession of *O. australiensis* used for the sequencing?
- Line 650. Please add reference to the CTAB protocol used, and details on the ONT and Illumina library preparation protocol used (kit, version, library insert size).
- Line 658. Please add a brief description of the sample collection methods. Add protocol, kit and version used for RNA isolation and library preparation. Add amount of data produced per tissue.
- Line 665. Please specify the company providing the antibody and the peptide sequence of the antigen used to produce the antibody.
- Line 666. Were biological replicates produced?

Gene and repeat annotations: This section is untidy. I suggest reorganizing and reordering the sections to make the reading smoother.

- Line 690. The Github link doesn't work.
- Line 693. I believe there's a typo as a fasta file is not a command.
- Line 696. Did the authors mean: Homology based evidence was *used* (not 'generated') in GenomeThreader to perform gene prediction.
- Line 705. Functional annotation should be moved after gene prediction.

Centromere analysis.

- How did authors handle duplicates and secondary alignments?
- How were centromeric region boundaries defined?

Identification of centromeric satellite repeat sequences.

- The sequence of the identified satellite repeats is missing.

Lines 723-724. Please specify which species these genome types refer to.

Phylogenetic analysis.

- Line 735. Protein *homology* (and not 'sequence') was detected by OrthoFinder.

Differential Analysis of LTR Sequences.

- Line 756. There are no methods describing how insertion time was estimated based on differences between flanking LTR sequences in the species.
- Line 757. Define fLTR.
- Lines 762-764. This sentence is hard to follow and the meaning not clear.

Identification of DD and EE genomic repeat markers

- Lines 777-779. Why did the authors use short reads to find the repeat sequences instead of the ONT ultra-long reads from their *O. australiensis* assembly and the PacBio reads from the *O. alta* assembly? Did the authors verify whether the sequences identified with RepeatExplorer can be found in the ONT and Pacbio reads?

ROUND 2 REVIEWER 2 ATTACHMENT:

Shared Transposon Amplification History Implicates EE as Direct Ancestor of the DD Subgenome

Lines 506-508 and lines 554-555 strongly state that the EE genome is the direct donor of the DD subgenome in the CCDD species, but the evidence provided do not convincingly support the authors' hypothesis of direct ancestry.

- Line 424. "The critical 1.7 Mya period".

Given the sequencing technology and computational power at the time, Ammiraju et al (2010) used the sequence of a single locus (*Adh*) to estimate divergence times in the *Oryza* genus. They estimated that the diploid and polyploid C genomes (*O. officinalis* (C_d), *O. minuta* CC (C_{TM}) and *O. alta* CC (C_{TA})) radiated approximately 1.7 million years ago, thus inferring a polyploidization time for *O. alta* (C_d - C_{TA}) of ~1.7 million years ago (with an HPD interval of 0.7636–2.9052, Table 2 in doi: 10.1111/j.1365-3113.2010.04251.x).

The authors of the present manuscript are using this punctual estimate as the critical watershed value to find validation for their transposon-based analysis and support the hypothesis of EE being the direct ancestor of the DD subgenome in the CCDD species.

Given the availability of up-to-date divergence time estimates based on high-quality chromosome-level genomes and complete gene annotations 15 years later, I would not take the 1.7 Mya estimate as the absolute "reference value". I would consider a temporal interval (instead of a timepoint) and provide the transposon-based analysis in *O. australiensis* and *O. alta* DD proposed here as further evidence supporting the known phylogenetic relationship between these genomes.

Summarizing, I agree with the authors' comment in the rebuttal letter which says that "*integration using non-coding evidence (e.g. transposon dynamics) remains valuable*", but this claim is very different from demonstrating a direct ancestry contribution of the EE genome. Furthermore, given the urgency of the authors to reassess the DD-EE relationship using evidence independent of single copy genes (lines 75-77), the choice of referring their transposon-based estimates to a polyploidization time inferred using a single gene sequence is not convincing.

Figure 5.

- Lines 425-427. Given the results of the analysis, the conclusion that can be drawn is only qualitative (as the authors wrote, it shows that EE and DD genomes contain higher proportion of Copia LTR than the CC subgenome), while a quantitative outcome cannot be inferred. Fig 5c shows that LTR distribution over time do not overlap in the EE and DD genomes, suggesting that TE dynamics in these genomes are different. Furthermore, LTR distribution pattern over time is more similar when comparing the two subgenomes of *O. alta* than when comparing *O. australiensis* with either *O. alta* subgenomes. For completeness, I would add *O. officinalis* to the analysis as representative ancestor of the CC subgenome. In addition to the density plot, I would plot the underlying histogram as well.

- Fig 5d: the intersection of the interpolated curves of a few datapoints (calculated using only one individual per species, Fig 5d) at around 1.7 Mya (1.73-1.84 interval) is not a robust and well-supported evidence of EE being the direct donor of the DD subgenome to the CCDD species.

- Finally, as the authors correctly noted (lines 455-458), the hybridisation time predicted by TE is significantly earlier than their estimate using single-copy genes (5.13 Mya) and the HPD found in the analysis is very broad (1.71 – 8.89, Fig 5a).

Figure 6.

- Line 487 and Fig. 6c: How can the authors conclude that amplification curves in EE and DD genomes are similar starting from 1.7 My backwards? The count ratio of Angela LTRs for *O. alta* DD : *O. australiensis* EE is approx. 1:2, between 1.2 and 1.9 Mya.

Why choosing exactly 1.7 My as the threshold value? The same 1 : 2 : 0 pattern (*O. alta* DD : EE : *O. alta* CC) begins at approx. 1.3 My in the plot. Besides, for completeness of Fig. 6c, I would add the histogram plot underlying the density plot and the Angela LTR insertion time distribution in *O. officinalis*.

- Lines 499 -501 and Fig. 6e. What are the summarized statistics of the whole-genome synteny analysis of Angela LTRs? The example provided is ok, but it would be more useful to get a number at the whole genome level since this evidence is used to support the hypothesis of direct ancestry.

- Lines 521-523 and Fig. 6g. What thresholds/cutoffs make the intact Angela LTRs “similar” or “dissimilar” to the Angela-Rep-DDEE? The definition found in Suppl. Methods is too vague to be reproducible: “An intact Angela LTR was classified as “similar to Angela-Rep-DDEE” if its annotated genomic coordinates overlapped with a significant BLASTN hit.”

- Lines 544-546: This conclusion is not convincing. Why would the analysis in Fig. 6g and described in lines 539-546 represents evidence for direct ancestry of EE in the CCDD genome? This could, at most, show genomic/structural similarity at the TE level between the EE genome and DD subgenome. Only because there is no other extant candidate in the *Oryza* genus to test, it does not necessarily imply that the EE genome is the direct donor of the DD subgenome to the CCDD. There might be an extant DD species not yet discovered, or the direct DD progenitor of CCDD species might be extinct. The logic used to compare insertion time of “matching” Angela LTRs in *O. australiensis* and *O. alta* (lines 551-552: “In general, Angela LTR in EE and DD genomes had similar amplification events and structural characteristics before 1.7 Mya” and Fig. 6g) is not conclusive evidence as one could paradoxically draw the same conclusion by looking at the fraction of Angela LTRs in the EE genome matching Angela-Rep-DDEE in the GG genome between 1.7 and 3 Mya.

Additional note. The LTR insertion time analysis (SanMiguel et al. *Nat. Genet.* 20, 43–45. 1998) has been widely used in the literature and it’s a nice method to explore insertion time of complete LTR-RTs in genome species based on the nucleotide distance between the two flanking LTR sequences of a complete LTR-RT. The software used here, LTR_Retrieve, uses the above approach to calculate LTR insertion time. A caveat is base call accuracy of the sequencing reads, because sequencing errors are counted as mutations for age calculations and this might affect age distribution.

In the rebuttal letter, the authors argued that analysis of repeat patterns using ONT long reads would not be feasible given “their higher per-base error rates make them less suitable for the fine-scale, comparative sequence analysis required to identify and differentiate shared and lineage-specific repeats at the whole-genome level”. Based on the same logic (and not considering the fact that ONT sequences can be polished), then ONT long reads would not be suitable for calculating LTR insertion time in the present study.

Centromere analysis.

- Lines 218-221: This sentence is confusing. Are neocentromeres located in "satellite repeat depleted regions" (line 220) or in "regions enriched for satellite repeats" (line 226)?

- Lines 225-227: Is the presence of satellite repeats outside the functional centromere regions (defined by CENH3 binding) alone sufficient evidence to demonstrate neocentromere formation?

The presence of interspersed satellite repeats outside the functional centromeres might be due to an expansion/novel insertion of TEs in the CENH3 region that pushed away the flanking satellite arrays. This is also stated by the authors in line 245-247 and supplementary Figure 8.

Lines 251 - 253: It is not clear whether this sentence is referring to ref. 31-37-38; other studies (references missing?); or to the previous sentence ("*among the 81 genes found by the authors, many (how many?) were found to be conserved and expressed in other studies*" - references missing).

- Fig 3d: how was TE density calculated in this figure?

- Supplementary Figures 9-10: in addition to *O. australiensis*, the figures show the CENH3 signal for two *O. sativa* varieties and *O. brachyantha*, but the methods do not describe ChIP seq experiment for the latter three species. How was the CENH3 signal showed in the plot of Suppl. Fig. 9 and 10 measured/quantified?

Structure and redundancy: methods description needs a clearer and logic narrative.

Methods should be re-organised in a way that the reader is not forced to continuously shift from the main manuscript to the supplementary material to find methods referring to the same experiment/analysis.

The authors should make a choice and decide what methods to put in the supplementary materials and what to keep in the main manuscript.

Methods reproducibility

Although I appreciated the effort of the authors to improve method description, there is still some missing information.

- Angela-LTR analysis: What thresholds/cutoffs make the LTRs similar or dissimilar to the Angela-Rep-DDEE? The definition found in Suppl. Methods is too vague to be reproducible: "An intact *Angela* LTR was classified as "similar to *Angela*-Rep-DDEE" if its annotated genomic coordinates overlapped with a significant BLASTN hit."

- "Genome de novo assembly and assessment" in Supplementary methods (Line 60): What was the total contig-level assembly size before the filtering step? How many contigs shorter than 1 Mb were discarded and what was the total size of the discarded portion?

1 Mbp looks like a high value to be used as cutoff. Were the discarded contigs all consisting of organellar DNA?

What were the L50, L90, N50, N90 values at the contig-level after removing organellar DNA? Suppl. Table 1 only shows the N50 value of the scaffold-level assembly.

- Lines 63-64 in Supplementary methods: I don't understand how gaps were counted. If you selected 13 contigs for your assembly, you should end up closing only one gap, but the methods here report the presence of multiple gaps filled with LR_Gapcloser (see also comment below).

- Supplementary Table 1: The counting of gaps is not correct.

In a linear genome assembly, the relationship between the number of contigs (n) and the number of gaps (g) is defined as: $g=n-1$

Given 12 chromosomes (= 12 scaffolds in a chromosome-level assembly), how can the assembly from Morrison et al have 13 contigs and three gaps? Same for the remaining assemblies?

Minor comments:

- Line 466: Please correct this number.
- Line 281: Please define “p-arm”.