

Telomere-to-telomere genome assemblies and population resequencing of diploid and allotetraploid peanut varieties

Corresponding Author: Ms Xiaoqin Liu

Version 0:

Decision Letter:

25th Jun 2024

Dear Professor Liu,

Your Article, "Origin, evolution and domestication of cultivated peanuts" has now been seen by 3 referees. You will see from their comments copied below that while they find your work of considerable potential interest, they have raised quite substantial concerns that must be addressed. In light of these comments, we cannot accept the manuscript for publication, but would be very interested in considering a revised version that addresses these serious concerns.

We hope you will find the referees' comments useful as you decide how to proceed. If you wish to submit a substantially revised manuscript, please bear in mind that we will be reluctant to approach the referees again in the absence of major revisions.

To guide the scope of the revisions, the editors discuss the referee reports in detail within the team, including with the chief editor, with a view to identifying key priorities that should be addressed in revision and sometimes overruling referee requests that are deemed beyond the scope of the current study. In this case, we ask you to address the reviewers' comments in full. We expect you to address all the technical concerns raised by the Reviewers, especially those regarding the analysis of the estimated age of polyploidy (Reviewer #2) and the potential confounding of population structure for the GWAS results (Reviewer #1).

If you choose to revise your manuscript taking into account all reviewer and editor comments, please highlight all changes in the manuscript text file. At this stage we will need you to upload a copy of the manuscript in MS Word .docx or similar editable format.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

If revising your manuscript:

*1) Include a "Response to referees" document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the referees along with the revised manuscript.

*2) If you have not done so already please begin to revise your manuscript so that it conforms to our Article format instructions, available here. Refer also to any guidelines provided in this letter.

*3) Include a revised version of any required Reporting Summary: <https://www.nature.com/documents/nr-reporting-summary.pdf>

It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review.

A revised checklist is essential for re-review of the paper.

Please be aware of our guidelines on digital image standards.

You may use the link below to submit your revised manuscript and related files:

Link Redacted

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

If you wish to submit a suitably revised manuscript we would hope to receive it within 6 months. If you cannot send it within this time, please let us know. We will be happy to consider your revision so long as nothing similar has been accepted for publication at Nature Genetics or published elsewhere. Should your manuscript be substantially delayed without notifying us in advance and your article is eventually published, the received date would be that of the revised, not the original, version.

Please do not hesitate to contact me if you have any questions or would like to discuss the required revisions further.

Nature Genetics is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit <http://www.springernature.com/orcid>.

Thank you for the opportunity to review your work.

Sincerely,
Chiara

Chiara Anania, PhD
Associate Editor
Nature Genetics
<https://orcid.org/0000-0003-1549-4157>

Referee expertise:

Referee #1: plant biology; evolutionary genomics

Referee #2: peanut genomics; evo

Referee #3: signed review

Reviewers' Comments:

Reviewer #1:

Remarks to the Author:

This is an interesting paper, but I find some important attention to detail is missing, and that several findings need better couching as speculative or even whether they're based on a priori evidence. Some points to consider, in no particular order:

- I suppose the major significance of the TE analyses is timing of bloom vs. tetraploidization, one needs to consider mutation rates used for Ks versus TEs. the same or different? One could anticipate that both should be proxies for base neutral mutation rate. For example, do the different species have different generation times?
- Differences in centromeric repeat array lengths. The authors suggest "significant differentiation of centromeric regions after tetraploidization." But they have sequenced modern exemplars of parental sub genomes, so it remains unknown what the arrays were like in the true ancestral genomes. I would rephrase.
- SVs - I strongly suggest that these be binned by molecular mechanism, so that the results make better biological sense (even though much of the literature conflates them). So what about indels, CNVs (tandems), TE insertions/ectopic recombinations, etc. The above can become important for functional annotation of SVs. For example, tandems (and new or sub-functional potential) can have very different functional relevance than indels, which can even include TE or NUMT/NUPT insertions into noncoding regulatory sequence. And of course, TEs hopping into exons and introns can be a different functional class to consider.
- Note that ADMIXTURE will always return K ancestral populations, even when one K is preferred ad hoc or by cross-validation (eg). Thus the results need to be couching in this respect; they are not necessarily "real". It's good that the K separates phylogenetic lineages - highlight this better.
- Selective sweeps. Recent TE or NUMT/NUPT insertions can fool these analyses by presenting highly homozygous

stretches of the genome. Look through your results for these molecular phenomena - at least using a sanity check of about 100 random hits. You may wish to filter out "sweep" hits that better reflect TE or NUPT/NUMT homozygosity before functional enrichment analysis. Also, what happens in terms of enrichment when you consider top 1% instead of 5%? More conservative could be safer

- GWAS: how did you convincingly control for what must be considerable underlying population structure in your sample? For example, among the different phylogenetic lineages.

- Were the candidate nut phenotype genes identified a priori, or solely through anonymous bioinformatics? If the former, and corroborated informatically, admit that please explicitly ;)

- WGD analyses. Please use syntenic analysis, not only Ks, which can be sensitive to underlying demographic differences, eg

- Functional enrichments - what were the exact backgrounds used? They should be included as files for replicability

- Phylogenetic trees should either display support values or parenthetical tree files should be provided in supplemental so investigators can evaluate robustness of results

- Fig. 4B shows a pretty clear horizontal cline. Carefully evaluate how this affected GWAS. And does the PCA suggest admixed populations (intermediate between red and green on X axis) in similar way as your other analyses?

- Where is the TreeMix analysis? Where are the f stats in tabular form? What were the significance values used. Etc.

Reviewer #2:

Remarks to the Author:

The study concerns a large sequencing analysis of peanuts, with several complete genomes and hundreds of peanut accessions characterized by whole genome sequencing. One major problem is that one of the basal analysis on the estimated age of polyploidy is fundamentally flawed, leading to wrong conclusions propagating through the manuscript:

The authors infer the age of divergence of two genomes from peaks in Ks curves constructed using the WGD software. The WGD software is constructed to find evidence for ancient Whole Genome Duplications, by intragenomic (not intergenomic) comparison. When making an intragenomic comparison (a comparison of a genome with itself), it's clear that there will always be 100% self-identity. Since the software is searching for residual, much weaker similarities from a WGD, it filters Ks values at (or close to) zero. However, when making comparisons between two different genomes, there is no self-identity to eliminate. Each gene should be compared to its best match in the other genome (a comparison of orthologs - the authors use the word "ortholog", but in fact, orthologous comparisons are filtered out by the software they use).

The authors kindly added several self-comparisons to their figure. These new self-comparison curves serve as controls for the method. Considering the objectives of the authors, these control curves should give values of zero divergence because the genomes are identical to themselves. In fact, none of the self-comparisons have a peak at zero. The methods used by the author are unable to show that a genome has zero divergence from itself!

In the same way, the method is unable to give a correct divergence of two genomes that are very similar to each other. Reiterating, this is because the method is designed for self-comparisons of a polyploid genome to indicate the age of divergence of the subgenomes. Values at (or close to) zero are filtered out. This is the reason the peak in the S245 (Ahy) vs S245 (Ahy) curve is Ks 0.025 (=1.58 Mya -the age divergence of A and B subgenomes), instead of zero.

For instance, the genome of Y4 (*A. ipaensis*) is, in fact, remarkably similar to the B subgenome of *A. hypogaea*. From the author's Ks data (S83B.cds.fa_Y4.cds.fa.ks.tsv) 27,923 genes are compared between S83 and Y4. Of these 24,541 have a Ks of zero. Indeed 23,584 of the genes have no nucleotide differences at all in their coding regions (Ks and Ka are both zero). The filtering of Ks=0 values combined with some sort of curve regression results in completely wrong Ks=0 values being represented on the graph.

The authors did not rescale the vertical axis as requested (It was relabelled as "Frequency", but the curves remain the same). If a number of gene pairs (=frequency) had been used, it would have been possible to see that data had gone missing from the different areas under the different curves.

I have attached three pdfs to help explain more:

Reviewer #3:

Remarks to the Author:

Overall evaluation

=====

This manuscript presents the sequencing, assembly and analysis of six peanut genomes. It expands the analysis with an

association analysis between two traits, oil content and seed weight and a genotype collection of 521 accessions. The manuscript is clear and easy to follow, although it contains a lot of information of results representing an incredible amount of work. The material and methods and the discussion should be improved in order to have a better support of the results presented. Overall the manuscript is well built, has interest and present relevant results not only for the peanut scientific community but also for the genomic community working in genome evolution and polyploidy.

Point-By-Point Manuscript Evaluation:

A. Summary of the key results

The manuscript titled "Origin, evolution and domestication of cultivated peanuts" describes the evolution of tetraploid peanut (*Arachis hypogaea* L.) genome through the sequencing and analysis of two diploid, and four tetraploid genomes. It also identifies genomic regions associated to domestication and to several agronomical traits through the re-sequencing of more than 500 accessions. Finally, the authors analyze the transcriptomics and metabolomics networks of the different peanut domesticated lineages. The results are summarized in the following points:

- Sequencing, assembly, and annotation of six representative accessions of peanuts, two diploids' progenitors (*A. duranensis*, AA, Y26 and *A. ipaensis*, BB, Y4) and four tetraploids (*A. hypogaea*, var. *fastigiata*, var. *vulgaris*, var. *hypogaea*, var. *hirsute*). All the genomes were sequenced with PacBio HiFi, scaffolded with Hi-C data and the gaps filled with ONT. Genome size ranged from 1.18 Gb to 2.62 Gb. Each of the six assemblies presented excellent assembly metrics in terms of Illumina mapping rate (>99%), BUSCO completeness (>95%), LAI (>20), and Merqury QV (>40). The genome annotation produced 34-35K gene models for the diploid accessions and 71-72K for the tetraploids.
- TE analysis reported a range from 75 to 78% of repetitive elements. B subgenomes have a higher TE content than A being the Gypsy elements the main contributors. Age divergency between both subgenomes was dated 1.58 MYA, meanwhile the divergency between the AA-At and BB-Bt subgenomes was estimated 0.27 and 0.29 MYA in agreement with previously published results. A deeper TE analysis revealed different insertion patterns for the LTR elements after the polyploidization event. Centromeric regions were different between the diploid and the tetraploid accessions. Retard elements expanded after the polyploidization event contributed to the differentiation between AA and At subgenomes.
- The analysis of the SVs between the A, B sub-genomes and their ancestors revealed a higher number of inversions and translocations in the A sub-genome, compared with the B. Genes closed to the SVs were enriched in functions such as flavonoid biosynthesis, brassinosteroid biosynthesis, fatty acid metabolism and linoleic acid metabolism. Comparisons across different *A. hypogaea* subspecies revealed also a different functional enrichment with could contribute to the different phenotypes in some lineages like *hirsuta*.
- Population structure analysis with 521 peanut accessions revealed 5 distinct groups. Gene flow was detected between two different groups: G2-G3 and G4-G5.
- GWAS analysis on peanut oil content reported a significant signal on chromosome 8 with 151 genes. A gene named *AhWRI1* was identified as responsible for the trait. Two haplotypes were identified in the collection with different oil content (~48% hap 1 and ~54% hap 2) associated with DE. Overexpression of this gene on rapeseed and soybean lines produced accumulation of lipid droplets supporting the involvement of this gene in the peanut oil content trait. Y1H assays allowed to associate several genes such as *AhFUS3*, *AhABI3* and *AhLEC1* to the regulation of the expression of this gene. A mutation on the E-Box promoter of this gene was identified as responsible for the different interaction with the *AhFUS3* gene.
- GWAS analysis on seed size and weight revealed a peak on chromosome 16 spanning 130 genes. A gene named *GSA1* was identified as one of the possible candidate genes for this trait based on similar phenotypes in rice. Two different haplotypes were identified for this gene, hap1 with 43 accessions and an average grain size of 650 gr. and hap2 with 294 accessions and an average grain size of 920 gr. The association of this candidate gene with the trait was verified with transcriptomic data as well as some analysis on the promotor of the gene.
- Transcriptomics and metabolomics comparisons between the accessions S83 and S245 identified key genes involved on lipid and anthocyanin content.

B. Originality and significance: if not novel, please include reference

The manuscript is original and novel. It expands our knowledge about the peanut domestication, and it generates important data for the community. It also describes two genes associated to two important agronomical traits.

C. Data & methodology: validity of approach, quality of data, quality of presentation.

The methodology is valid, and the quality of the presentation is good. More details in the material and methods are needed, specially to assess the RNA-Seq and the GWAS analysis that it has been evaluated under the principles the it has the right number of replicates.

D. Appropriate use of statistics and treatment of uncertainties

It is not applicable to this type of research, except for the GWAS and the RNA-Seq which looks fine. Nevertheless more details are needed in the M&M for a full assessment.

E. Conclusions: robustness, validity, reliability

Most of the conclusions are robust and valid although the lack of some information on the M&M (e.g., replicates) makes difficult to have a full assessment.

F. Suggested improvements: experiments, data for possible revision

There are several important improvements that should be addressed if the manuscript is accepted for publication.

- a. Population structure analysis has been slightly treated in the manuscript. There are several publications that describe in detail some of the peanut population structure. This manuscript presents an interesting analysis but it does not compare its results with what it was published before. Although, the manuscript is already too extensive, I found missing a dating of the different splits of the different populations and how this is related with the current status of the species (e.g., domestication, distribution...).
- b. Transcriptomic and metabolomic analysis of the S83 and S245 accessions does not describe the genes involved in the trait. The last result section describes the transcriptomic and metabolomic analysis associated to some secondary compounds, but it does not describe in detail the WGCNA (e.g., number of clusters) or further more the "key genes highly correlated with differentially abundant metabolite content across seed development stages". The transcriptomic analysis is not summarized in the M&M section, so it is difficult to evaluate if it was performed with the right number of replicates and the adequate tools.
- c. Discussion section is more of a conclusion section. Discussion of the manuscript should be improved. Usually discussion section focus is to contextualize the results presented in a manuscript with what it is known. Conclusion is a summary of the results with some highlights. In this manuscript, the discussion section is the second one. The results incorporate some pieces that it could be included in the discussion section, but the manuscript as such lacks on a solid discussion section presenting some of the results, in a wider context. For example, the authors could compare the TE dynamics of peanut with other polyploid species. They could bring to the discussion an interesting topic like the genomic shock and associate it with the different insertion rates in both subgenomes. They could talk about subgenome dominance. Same could be applicable to the population structure section or the selection of the gene candidates. How are those compared with other legumes like soybean?
- d. The manuscript does not have a statement describing that the research presented has followed the Nagoya protocol. There are some samples from different countries which should either meet the Nagoya protocol, or state that they were collected before the implementation of it.
- e. The material and method information could be improved. The M&M information should be enough to reproduce the experiments. In this sense, the M&M lacks from important information. Some examples are:
 - i. The description of the accessions used on the analysis is poor, specially for the 521 accessions. They are summarised in the suppl. table 15, but it has not been specified if they are part of a collection, if the accession ID is the same than the collection. The information about the origin of the samples is poor and it does not include the type of the accession (e.g., cultivated vs wild, subspecies...).
 - ii. The version number should be specified for all the programs used in the manuscript.
 - iii. The gene annotation methods is poorly described. It does not mention how the RNA-Seq data was mapped nor the quality evaluation process. It does not describe the weight to each of the evidences used for EVM. The functional annotation has been poorly described, specially because eggNOG-mapper tool has some limitations when it is used with plant species.
 - iv. There is not description about how the phenotyping of the 521 accessions were performed (oil content, seed size and weight..., number of plants, number of years, replicates in the field...).
- f. The data availability for the project PRJNA1106792 is incomplete. It has only 4 of 6 genome assemblies. No annotation, transcriptomic data or re-sequencing data.

Some discretionary suggestions to improve the data support of the manuscript are:

- a. Add the completeness values of merqury, as well as the K-mer distribution curves to have a better perspective of the resolution of the duplicated regions. Additionally having some description of the possible contaminations with Blobtools could be good, although the comparison between cultivars is a good support for the absence of contaminants.
- b. The evaluation of the gene models could be complicated, but in addition to the BUSCO analysis, a good way to check for the gene models quality is to compare them with the transcriptomic data and report the number of gene models supported by the transcriptomic/protein sequences (from a closed related species) data (e.g., through the distribution of the different codes supplied by Gffcompare, or with the AED from MAKER). Single exon gene rate as well as number of overlapping genes are also good metrics (AGAT toolkit is a good tool to calculate those). It also describe how many gene models have a GO Term annotation or a NCBI NR/TREMBL/SwissProt blast/diamond hit.

G. References: appropriate credit to previous work?

Yes, in the introduction, M&M and results. The discussion lacks references.

H. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions

The manuscript is clear and easy to read, although the discussion should be improved. The abstract summarizes well the results presented, as well as the introduction and the conclusions.

Manuscript reviewed by Aureliano Bombarely on 22/06/2024

Version 1:

Decision Letter:

10th Jan 2025

Dear Professor Liu,

Your Article, "Origin, evolution and domestication of cultivated peanuts" has now been seen by 3 referees. You will see from their comments below that while they find your work of interest, some important points are raised. We are interested in the possibility of publishing your study in Nature Genetics, but would like to consider your response to these concerns in the form of a revised manuscript before we make a final decision on publication.

To guide the scope of the revisions, the editors discuss the referee reports in detail within the team, including with the chief editor, with a view to identifying key priorities that should be addressed in revision and sometimes overruling referee requests that are deemed beyond the scope of the current study. In this case, we ask you to take seriously into account Reviewer #2 concerns and address all the remaining requests. We hope that you will find the prioritized set of referee points to be useful when revising your study. Please do not hesitate to get in touch if you would like to discuss these issues further.

We therefore invite you to revise your manuscript taking into account all reviewer and editor comments. Please highlight all changes in the manuscript text file. At this stage we will need you to upload a copy of the manuscript in MS Word .docx or similar editable format.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

When revising your manuscript:

*1) Include a "Response to referees" document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the referees along with the revised manuscript.

*2) If you have not done so already please begin to revise your manuscript so that it conforms to our Article format instructions, available

http://www.nature.com/ng/authors/article_types/index.html here.

Refer also to any guidelines provided in this letter.

*3) Include a revised version of any required Reporting Summary: <https://www.nature.com/documents/nr-reporting-summary.pdf>

It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review.

A revised checklist is essential for re-review of the paper.

Please be aware of our <https://www.nature.com/nature-research/editorial-policies/image-integrity> guidelines on digital image standards.

EXTENDED DATA FIGURES

When re-submitting your manuscript, please ensure that any supplementary figures and tables that are crucial to the manuscript's conclusions are converted into Extended Data figures and tables to increase visibility of these data. Extended Data figures and tables are online-only (present in the online PDF and full-text HTML versions of the paper), peer-reviewed display items that provide essential background to the article but are not included in the main article due to space constraints. A maximum of ten Extended Data display items (figures and tables) is permitted.

Please use the link below to submit your revised manuscript and related files:

Link Redacted

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We hope to receive your revised manuscript within four to eight weeks. If you cannot send it within this time, please let us know.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Nature Genetics is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please

visit please visit www.springernature.com/orcid.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Sincerely,
Chiara

Chiara Anania, PhD
Associate Editor
Nature Genetics
<https://orcid.org/0000-0003-1549-4157>

Referee expertise:

Referee #1:

Referee #2:

Referee #3:

Reviewers' Comments:

Reviewer #1 (Remarks to the Author):

After enormous revisions, many of which entailed significant new bioinformatic research, this manuscript is now an excellent example of a paper with the interest level and attention to detail requisite for a journal such as Nature Genetics. I am satisfied with how the investigators followed up on my suggestions, and am also particularly impressed at how independent approaches (fastsimcoal being one of them) yielded similar dates. This consistency speaks to both care taken to analyses as well as, of course, the appropriateness of the data to address the questions at hand. My only final suggestion - which I think is very important these days - is to revise all main-text figures such that they use similar (and best, color-blind compatible) color palettes. For example, Figs. 1 and 2 stand out very starkly in "standard" red, blue, and (bright) green colors, while Figs. 3 and 4, e.g., have much more muted colors that are easier to view.

Reviewer #2 (Remarks to the Author):

The following conclusions of the abstract aren't supported:

Lines 42-43

"a 13.23 Mb-long insertion on Chr14 of all the var. hirsute accessions may have contributed to the prostrate growth and narrow leaves of this group"

The data in the paper shows correlation - not causation

Lines 43-44

"Selective sweep and introgression analyses highlighted multiple asymmetric selections between the two subgenomes during breeding improvement."

There is no data to show this. Differences between the subgenomes were almost all accumulated in the diploid state, before the origin of *Arachis hypogaea*

Lines 45-46

"Genome-wide association studies identified 4,638 loci associated with 37 agronomic traits, including candidate genes for oil content, seed size, seed weight, etc."

Almost all of these loci lack any sort of verification.

Lines 50-51

"These findings provide a comprehensive understanding of the origin, evolution and domestication"

It's handwaving - there are few new insights. From this paper, I've learned nothing new of the origin, a bit about the evolution of the tetraploid genome (mainly about some rearrangements), and nothing at all about domestication.

More detailed comments

The authors shouldn't use the abbreviations Y26 and Y4 and instead use the full species names along with their original collection numbers (e.g., *Arachis ipaensis* K30076). This practice is crucial for properly acknowledging prior scientific work and enabling the stepwise accumulation of knowledge within the field.

fastigiata is still incorrectly spelled fastigate in multiple instances in the manuscript

The discovery that *A. ipaensis* is the ancestor of the B subgenome of *A. hypogaea* is wrongly attributed: "Zhuang et al. (2019) compared the sequenced A and B genomes of wild species to those of the cultivated variety Shitouqi and reported high sequence similarity between *A. ipaensis* and the B genome of Shitouqi (> 99.5%), suggesting that it or a close relative is the peanut Bt subgenome progenitor". This conclusion was not a discovery of Zhuang et al. (2019); rather, it is the culmination of decades of careful study that are conspicuously ignored:

Kochert, G. et al. RFLP and cytogenetic evidence on the origin and evolution of allotetraploid domesticated peanut, *Arachis hypogaea* (Leguminosae). *Am. J. Bot.* 83, 1282–1291 (1996).
Moretzsohn, M.C. et al. A study of the relationships of cultivated peanut (*Arachis hypogaea*) and its most closely related wild species using intron sequences and microsatellite markers. *Ann. Bot.* 111, 113–126 (2013).
Seijo, G. et al. Genomic relationships between the cultivated peanut (*Arachis hypogaea*, Leguminosae) and its close relatives revealed by double GISH. *Am. J. Bot.* 94, 1963–1971 (2007).
Robledo, G. & Seijo, G. Species relationships among the wild B genome of *Arachis* species (section *Arachis*) based on FISH mapping of rDNA loci and heterochromatin detection: a new proposal for genome arrangement. *Theor. Appl. Genet.* 121, 1033–1046 (2010).
Grabiele, M., Chalup, L., Robledo, G. & Seijo, G. Genetic and geographic origin of domesticated peanut as evidenced by 5S rDNA and chloroplast DNA sequences. *Plant Syst. Evol.* 298, 1151–1165 (2012).
Bertoli, D.J. et al. The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nat. Genet.* 48, 438–446 (2016).

References 1–5 firmly establish *A. ipaensis* as the B subgenome donor of peanut. Reference 6 further concluded, based on genomic and biogeographic evidence, that the only known accession of *A. ipaensis* K30076 likely was from the same population that gave rise to peanut less than 9,600 years ago. This is easy to spot because the abstract explicitly states: "On the basis of remarkably high DNA identity of the *A. ipaensis* genome and the B subgenome of cultivated peanut and biogeographic evidence, we conclude that *A. ipaensis* may be a direct descendant of the same population that contributed the B subgenome to cultivated peanut."

It is extraordinary - and troubling - that the paper so badly misrepresents this substantial body of research."

The preceding lines of the introduction [Lines 68–80] are also deeply flawed. The authors rely on outdated papers and a poorly executed review to create an artificial sense of uncertainty and confusion. They state, "However, the precise genetic origin of *A. hypogaea* remains undetermined." Really? Can the authors provide even a single example of an ancient polyploid crop whose origin is as thoroughly and conclusively established as that of peanut?

Line 93

What is meant by 'predominantly cultivated ecotypes'? As all *Arachis hypogaea* are domesticated (cultivated), this term seems redundant or misinformed. Could the authors clarify why this phrase was chosen?

Lines 100-102

The authors misattribute fundamental knowledge about the genus *Arachis* and the domestication process to recent genome papers that did not originally discover or describe these insights. For example, they state: "During the process of domestication, selection for larger pod sizes led to a transition from the small pods found in wild accessions to the larger, high-yielding pods characteristic of modern varieties." This observation, while accurate, has been well-documented in much earlier studies on the genus and on domestication and does not originate from genome sequencing efforts. Proper attribution to the foundational studies that established these concepts is essential to accurately represent the development of knowledge in the field. The authors need to cite the proper literature.

Lines 111-113

Some parts of the discussion are not comprehensible "In addition, key candidate QTLs for branching habit, testa color, root nodule development, and NB-LRR genes. have been identified. These QTLs were identified through map-based cloning". QTLs for NB-LRR genes doesn't make sense because QTLs describe quantitative traits rather than specific gene categories. "These QTLs were identified through map-based cloning" appears to be inaccurate in this context. The citations closest to this sentence include three genome papers, one unrelated methodological paper, and a review on NB-LRR genes—none of which seem to describe the use of map-based cloning. Map-based cloning involves isolating genes within a QTL region

through fine mapping and positional cloning, which differs from the methods typically used in genome studies or broader QTL mapping.

The figure following Fig. 1 lacks both a legend and a figure number. (This doesn't really matter because I could use the legend in the previous version). In the text, the authors have adjusted their estimate for the date of peanut's origin in the right direction, but likely for the wrong reasons as the graph is incorrectly plotted and is not a suitable method for estimating such a recent divergence.

To give context to my previous request to add extra self-comparisons. The reason I did this was to demonstrate to the authors how nonsensical this plotting is. The methods used by the authors are unable to show that a sequence is identical to itself, and for the same reasons that the self-comparison fails, a comparison of two closely related genomes fails. I assumed that, seeing this, the authors would totally reform this part of the paper. Unfortunately not.

Now, I will explain why the graph (A) is wrongly plotted.

- 1) The label on the vertical axis "Frequency" isn't defined and covers up the dramatically different numbers of gene comparisons under each curve. To correct the axis, it must have a scale of number of genes in each category.
- 2) The use of a line graph is wrong. The data is a frequency distribution; it isn't continuous data. A scatter-plot or column graph should be used. Although it's common practice to use a line graph for Ks curves, the plotting methods are leftover from when ESTs were used to produce the data. In the days of Ks graphs from ESTs, the data was smoothed and (probably) modeled. In contrast, data from whole genome comparisons needs no smoothing or modeling because it is a complete dataset. There are standards of representation of data required by Nature, and this representation drastically fails those standards. Take, for example, the blue lines; as they move from 0 to higher values, the blue line initially rises. This is a complete misrepresentation because most gene comparisons in the blue curves have zero Ks values (from the authors own data provided in tsv files) - the blue line should move down when initially moving away from zero. Put in other words, the peak values on the blue lines should be zero, not a value higher than zero.

Now, I will explain why the use of Ks values isn't suitable for closely related genomes

Using Ks rather to calculate divergence is ill-suited for closely related genomes with short divergence times. Ks graphs, which measure synonymous substitutions in coding regions, was historically developed to enable comparisons between distantly related sequences where intergenic regions are too divergent to align reliably. This method is highly effective for long-term evolutionary studies, as it focuses on conserved coding regions that are possible to align, and where there is sufficient time for synonymous substitutions to accumulate. However, for recent divergences, the number of synonymous substitutions is too low to generate meaningful averages, resulting in high variability and poor resolution.

For closely related genomes that can be completely aligned (as is the case for these genomes) using SNPs across the whole genome leverages a much larger dataset, capturing variations in both coding and noncoding regions. This provides the granularity necessary for precise divergence estimates over short timescales. Relying on Ks for such analyses is like trying to measure 10 micrometers with a wooden ruler - it's the wrong tool for the scale of the measurement. SNPs, by encompassing genome-wide differences, are far better suited for estimating divergence in recently separated genomes.

To help illustrate, here are calculations to show that two peanut-like subgenomes, each of 1.4Mb, diverged by 9,000, would show most genes as having Ks=0 values

Parameters

- 1) Number of aligned orthologous genes: 28,662 (from the author's data)
- 2) Average alignment length per gene: 1,100 base pairs (bp) (from the author's data)
- 3) Fraction of synonymous sites: 75% of coding sequence sites are synonymous (825 bp per gene).
- 4) Neutral substitution rate: 8.12×10^{-9} Ks/site/year
- 5) Divergence time: 9,000 years (as an illustration)

Calculations

Step 1:

The expected number of synonymous substitutions per gene (λ) is calculated as:

$$\lambda = 2 \times \text{Substitution rate} \times \text{Divergence time} \times \text{Number of synonymous sites per gene}$$

$$\lambda = 2 \times (8.12 \times 10^{-9}) \times 9,000 \times 825 = 0.1201$$

In other words, only about 12% of the genes will have a synonymous substitution. The rest won't!

Step 2

[Strictly speaking, for the next step, we should use the Poisson distribution, but because λ is so low, it really doesn't make much difference]

Therefore, out of 28,622 gene pairs, we would expect $28,622 \times 0.88 = 25,223$ to have Ks=zero. (That number is pretty close to the B vs B comparisons from the paper's data which in one example is ~24,500 gene pairs having Ks=0)

Therefore, the genes with Ks>0 are unusually highly mutated. Filtering Ks=0 numbers is like trying to determine the average speed of cars on a highway by only measuring the fastest ones - you'd ignore the slower majority, leading to a distorted conclusion. The absence of genes with Ks=0 removes the necessary baseline for accurately estimating divergence on a short timescale, as the molecular clock relies on the full distribution of substitutions, including those that have not

accumulated any. Having said this, the best way to date close divergences like this is SNP rates from, whole genome alignments.

Why does Supplementary Figure 14 (B) use two different methods for the divergence dates K_s for the origin and coalescent-based genetic simulation for the others? The latter is giving dates I would expect (although they are stated with absurd precision (eg 9,491 years).

Reviewer #3 (Remarks to the Author):

Overall evaluation on the revised version

=====

The authors have successfully resolved the concerns and questions risen during the first revision. The manuscript, although it contains a huge amount of information, is clear and easy to read. The results are well supported by several data sources (phenome2genome associations, transcriptomics...). As I mentioned in my first revision, the manuscript is well built, has interest and present relevant results not only for the peanut scientific community but also for the genomic community working in genome evolution and polyploidy.

Revised by Aureliano Bombarely on December 23rd, 2024

Version 2:

Decision Letter:

31st Mar 2025

Dear Professor Liu,

Your Article entitled "Telomere-to-telomere genomes of diploids and allotetraploids provide insights into the origin, evolution, and domestication of peanuts" has now been seen by 3 referees, whose comments are attached. In the light of these advice we have decided that we cannot offer to publish your manuscript in Nature Genetics.

We feel that the ongoing technical reservations related to the incomplete assembling of some chromosomes are sufficiently important as to preclude publication of this study in Nature Genetics.

I am sorry that we cannot be more positive on this occasion, but hope that you will find our referees' comments helpful when preparing your paper for submission elsewhere.

Thank you.

Sincerely,
Chiara

Chiara Anania, PhD
Associate Editor
Nature Genetics
<https://orcid.org/0000-0003-1549-4157>

Reviewers' Comments:

Reviewer #1 (Remarks to the Author):

I am entirely satisfied with the authors' acceptance of my suggestion re: colorblind palettes:

"Regarding your final suggestion about figure color palettes, we fully agree with your point about the importance of using color-blind friendly palettes. In the current revised manuscript, we have ensured that all figures adopt consistent and more accessible color schemes, with careful attention to color-blind compatibility. We have updated Fig. 1 and Fig. 2 to replace the bright red, blue, and green with more muted, color-blind friendly alternatives, in line with the colors used in Figs. 3 and 4."

Regarding their responses to the other reviewers, in a short look, they seem well-considered, but I'm afraid I don't have time to carefully evaluate them.

Reviewer #2 (Remarks to the Author):

Please see uploaded file. It contains formatting of text and graphs that can't be entered into this box.
If there are any issues with the uploaded file, please contact me so that we can address the problems.

Reviewer #3 (Remarks to the Author):

The author did resolved most of my concerns and questions in the previous revision. I think that the new version of the manuscript has been improved after a new round of interactions with other reviewers. From my point of view the manuscript it is good to go.

****Although we cannot publish your paper, it may be appropriate for another journal in the Nature Portfolio. If you wish to explore the journals and transfer your manuscript please use our [manuscript transfer portal](#). You will not have to re-supply manuscript metadata and files, unless you wish to make modifications. For more information, please see our [manuscript transfer FAQ](http://www.nature.com/authors/author_resources/transfer_manuscripts.html?WT.mc_id=EMI_NPG_1511_AUTHORTRANSF&WT.ec_id=AUTHOR) page.**

Version 3:

Decision Letter:

IMPORTANT: Please note the reference number: NG-A65542R2-Z Liu. This number must be quoted whenever you communicate with us regarding this paper.

4th Jun 2025

Dear Dr. Liu,

Thank you for your message of 4th Jun 2025, asking us to reconsider our decision on your manuscript "Telomere-to-telomere genomes of diploids and allotetraploids provide insights into the genome organization, evolution, and agronomic traits for peanut improvement". I have now discussed the points of your letter with my colleagues, and we think that you have some valid points. We therefore invite you to revise your manuscript along the lines that you propose.

When preparing a revision, please ensure that it fully complies with our editorial requirements for format and style; details can be found in the Guide to Authors on our website (<http://www.nature.com/ng/>).

Please be sure that your manuscript is accompanied by a separate letter detailing the changes you have made and your response to the points raised. At this stage we will need you to upload:

- 1) a copy of the manuscript in MS Word .docx format.
- 2) The Editorial Policy Checklist:
<https://www.nature.com/documents/nr-editorial-policy-checklist.pdf>
- 3) The Reporting Summary:
<https://www.nature.com/documents/nr-reporting-summary.pdf>

(Here you can read about the role of the Reporting Summary in reproducible science:

<https://www.nature.com/news/announcement-towards-greater-reproducibility-for-life-sciences-research-in-nature-1.22062>)

Please use the link below to be taken directly to the site and view and revise your manuscript:

Link Redacted

With kind wishes,
Chiara

Chiara Anania, PhD
Associate Editor
Nature Genetics
<https://orcid.org/0000-0003-1549-4157>

Version 4:

Decision Letter:

27th Aug 2025

Dear Professor Liu,

I apologize for the delay in sending you the decision, which is due to my recent absence from the office. Thank you for your patience during this process.

Your Article, "Telomere-to-telomere genomes of diploids and allotetraploids provide insights into the genome organization, evolution, and agronomic traits for peanut improvement" has now been seen by 3 referees. You will see from their comments below that Reviewers are persuaded by the quality of the assemblies and support the publication of your study in Nature Genetics. However, some remaining points are raised by the new reviewer that we think should be implemented via textual change before we make our final decision.

To guide the scope of the revisions, the editors discuss the referee reports in detail within the team, including with the chief editor, with a view to identifying key priorities that should be addressed in revision and sometimes overruling referee requests that are deemed beyond the scope of the current study. In this case, please discuss all the points raised by Reviewer #4, including reworking on the writing style. We hope that you will find the prioritized set of referee points to be useful when revising your study. Please do not hesitate to get in touch if you would like to discuss these issues further.

We therefore invite you to revise your manuscript taking into account all reviewer and editor comments. Please highlight all changes in the manuscript text file. At this stage we will need you to upload a copy of the manuscript in MS Word .docx or similar editable format.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

When revising your manuscript:

*1) Include a "Response to referees" document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the referees along with the revised manuscript.

*2) If you have not done so already please begin to revise your manuscript so that it conforms to our Article format instructions, available

[here](http://www.nature.com/ng/authors/article_types/index.html).

Refer also to any guidelines provided in this letter.

*3) Include a revised version of any required Reporting Summary: <https://www.nature.com/documents/nr-reporting-summary.pdf>

It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review.

A revised checklist is essential for re-review of the paper.

Please be aware of our [guidelines](https://www.nature.com/nature-research/editorial-policies/image-integrity) on digital image standards.

EXTENDED DATA FIGURES

When re-submitting your manuscript, please ensure that any supplementary figures and tables that are crucial to the manuscript's conclusions are converted into Extended Data figures and tables to increase visibility of these data. Extended Data figures and tables are online-only (present in the online PDF and full-text HTML versions of the paper), peer-reviewed display items that provide essential background to the article but are not included in the main article due to space constraints. A maximum of ten Extended Data display items (figures and tables) is permitted.

Please use the link below to submit your revised manuscript and related files:

Link Redacted

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We hope to receive your revised manuscript within four to eight weeks. If you cannot send it within this time, please let us know.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Nature Genetics is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit www.springernature.com/orcid.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Sincerely,
Chiara

Chiara Anania, PhD
Associate Editor
Nature Genetics
<https://orcid.org/0000-0003-1549-4157>

Referee expertise:

Referee #4: plant genomics

Reviewers' Comments:

Reviewer #1 (Remarks to the Author):

The authors, in their response, present a convincing case with associated graphics that appears to demonstrate that a reviewer mistakenly analyzed an earlier version of the assembly in question. I have no further comments for improvement.

Reviewer #3 (Remarks to the Author):

As I stated in previous revisions, I think that the authors resolved all the concerns that I rose during my first revision. Furthermore, I think that they did an excellent job answering and resolving not only my questions and concerns but also others risen by the other reviewers. I think that the manuscript has enough merit for its publication.

Reviewer #4 (Remarks to the Author):

Comparative Analysis & Novelty:

Previous studies have already addressed topics such as genome assembly, selective sweeps, evolutionary history, and phenotypic diversification. The authors should represent the unique contributions through the differences between previous studies and this study.

Origin & Evolution Section:

The contents on Origin and Evolution should be toned down, as these aspects have been provided in previous studies. The authors should focus more on their novel findings.

SV-Domestication Relationship:

Authors should strengthen the relationship between structural variations (SVs) and domestication. The current manuscript does not sufficiently explain the functional relationships among genes, SVs, and domestication traits.

Genome Similarities:

I agree with the authors' responses regarding genome similarities.

GWAS Analysis:

In addition to the known loci (FAD and AhRUVBL2), the authors are encouraged to include genetic mechanism analyses for one or two additional phenotypic traits to enhance the GWAS findings.

Title & Writing Style:

The current title is too informal, and the overall writing style lacks academic rigor (e.g., the use of "etc." in the Abstract). A more precise and professional revision should be recommended.

Version 5:

Decision Letter:

Our ref: NG-A65542R4

31st Oct 2025

Dear Dr. Liu,

Thank you for submitting your revised manuscript "Telomere-to-Telomere Genomes of Diploids and Allotetraploids provide Insights into Peanut Genome Evolution and Improvement" (NG-A65542R4). It has now been seen by the original referees and their comments are below. The reviewers find that the paper has improved in revision, and therefore we'll be happy in principle to publish it in Nature Genetics, pending minor revisions to comply with our editorial and formatting guidelines.

If the current version of your manuscript is in a PDF format, please email us a copy of the file in an editable format (Microsoft Word or LaTeX)-- we can not proceed with PDFs at this stage.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements soon. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Thank you again for your interest in Nature Genetics Please do not hesitate to contact me if you have any questions.

Sincerely,

Wei Li, PhD
Senior Editor
Nature Genetics
www.nature.com/ng

Reviewer #4 (Remarks to the Author):

The authors have addressed all comments and suggestion of reviewers, thus, I suggest to accept it for publication..

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/>

To guide the scope of the revisions, the editors discuss the referee reports in detail within the team, including with the chief editor, with a view to identifying key priorities that should be addressed in revision and sometimes overruling referee requests that are deemed beyond the scope of the current study. In this case, we ask you to address the reviewers' comments in full. We expect you to address all the technical concerns raised by the Reviewers, especially those regarding the analysis of the estimated age of polyploidy (Reviewer #2) and the potential confounding of population structure for the GWAS results (Reviewer #1).

If revising your manuscript:

*1) Include a “Response to referees” document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the referees along with the revised manuscript.

*2) If you have not done so already please begin to revise your manuscript so that it conforms to our Article format instructions, available [here](#).

Refer also to any guidelines provided in this letter.

*3) Include a revised version of any required Reporting Summary:
<https://www.nature.com/documents/nr-reporting-summary.pdf>

It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review.

A revised checklist is essential for re-review of the paper.

Reviewers' Comments:

Reviewer #1:

Remarks to the Author:

This is an interesting paper, but I find some important attention to detail is missing, and that several findings need better couching as speculative or even whether they're based on a priori evidence. Some points to consider, in no particular order:

- I suppose the major significance of the TE analyses is timing of blooms vs. tetraploidization, one needs to consider mutation rates used for Ks versus TEs. the same or different? One could anticipate that both should be proxies for base neutral mutation rate. For example, do the different species have different generation times?

Response: We appreciate your insightful comments. As previously noted, *Ks* values and transposable elements (TEs) display distinct mutation rates in *Arachis*¹⁻⁴. Prior studies have shown that the modal *Ks* paralog values since the divergence of the Dalbergioid clade approximately 58 million years ago are around 0.95 for *A. ipaensis* and 0.90 for *A. duranensis*, consistent with findings in *Medicago* (approximately 0.95)¹. Using the method described in "Simple Methods for Estimating the Numbers of Synonymous and Nonsynonymous Nucleotide Substitutions" by Masatoshi Nei and Takashi Gojobori⁵, the average rate of genetic change in *Arachis* genes has been calculated at 8.12×10^{-9} synonymous substitutions per synonymous site per generation¹. This mutation rate has been widely utilized in peanut genome studies, including those on Tifrunner², Fuhuasheng⁶, and Monticolar³.

To estimate the genome-wide mutation rate in *Arachis*, Bertoli et al. (2016) calculated values by mapping *A. ipaensis* Molecule reads to the pseudomolecules of *A. duranensis*. This analysis produced a corrected median DNA identity of 93.11%, consistent with earlier comparisons using BAC sequences⁷. Based on *Ks* values, which suggest that the A and B genomes diverged 2.16 million years ago, the mutation rate was estimated at 1.6×10^{-8} base substitutions per site per generation¹. This rate aligns with the range of $1-2 \times 10^{-8}$ reported for other plants⁸ and has been extensively used in *Arachis* genomic studies.

Considering that TEs constitute more than 75% of the peanut genome, their analysis offers greater resolution and less granularity than *Ks*-based analyses, involving substantially more data relative to the total length of genes. Consequently, the mutation rate used in TE analysis corresponds to the overall genome mutation rate.

In our study, different mutation rates of *Ks* and TE analyses in peanuts were employed. And the mutation rates are consistent with those in previous studies.

- Differences in centromeric repeat array lengths. The authors suggest "significant differentiation of centromeric regions after tetraploidization." But they have sequenced modern exemplars of parental sub genomes, so it remains unknown what the arrays were like in the true ancestral genomes. I would rephrase.

Response: Thank you for your valuable comments. In our centromere analysis, we evaluated centromere length, repeat unit length, and the number of repeat units across the six assembled

genomes. The results indicate that, compared to diploids, the total length and number of repetitive elements in tetraploid At are generally consistent with those in diploid Ad (**Supplemental Table 12**). However, Bt total length is only 59% of Bd total length. When the lengths of the repetitive elements are largely similar, the reduced number of repetitive elements in tetraploid Bt explains this difference. As you noted, the difficulty in identifying the direct ancestral species of the cultivated tetraploids means that the statement in the manuscript regarding the "significant differentiation of centromeric regions after tetraploidization" may not be entirely accurate. We appreciate this valuable feedback and have revised this section accordingly, as shown in **lines 259-263**.

- SVs - I strongly suggest that these be binned by molecular mechanism, so that the results make better biological sense (even though much of the literature conflates them). So, what about indels, CNVs (tandems), TE insertions/ectopic recombinations, etc. The above can become important for functional annotation of SVs. For example, tandems (and new or sub-functional po/ential) can have very different functional relevance than indels, which can even include TE or NUMT NUPT insertions into noncoding regulatory sequence. And of course, TEs hopping into exons and introns can be a different functional class to consider.

Response: Thank you for your insightful feedback. In the revised analysis, we categorized and statistically analyzed various types of structural variations (SVs) using the SYRI (v.1.5.3)⁹ and SVMU (v.4.0.0beta2)¹⁰ software, including INDELs, INVs, CNVs, and TRANSs (**Figure 1 (Supplementary Figure 7)**). Our analysis of TE-derived SVs, conducted with a comprehensive whole-genome TE library, showed that approximately 60% of SVs originated from TEs, and ~55% of SV-associated genes were classified as TE-SV genes. These findings indicate significant TE mobilization (**Figure 2 (Supplementary Figure 8), Supplementary Tables 13 & 14**). Further analysis demonstrated that TEs are distributed within introns, exons, and regions 2 kb upstream and downstream of genes (**Figure 2 (Supplementary Figure 8)**), suggesting key roles in conditional SVs and SV-gene functions. We also performed statistical analyses along with GO and KEGG enrichment for genes located within 2 kb of SV breakpoints (**Figure 2 (Supplementary Figure 8), Supplementary Tables 17 & 18**). As you suggested, separate analyses and enrichment of different SV types is important for efforts to understand their evolutionary roles in peanut varieties. These revisions are reflected in the main text (**lines 282-301 & 311-327**) and supplementary materials (**Supplementary Figure 8), Supplementary Tables 13, 14, 17 & 18**). Additionally, we have expanded the Methods section to include the standards and approaches used for SV identification.

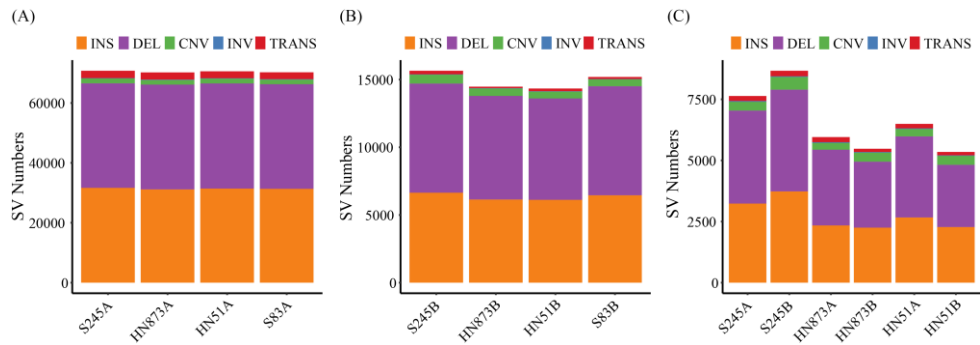
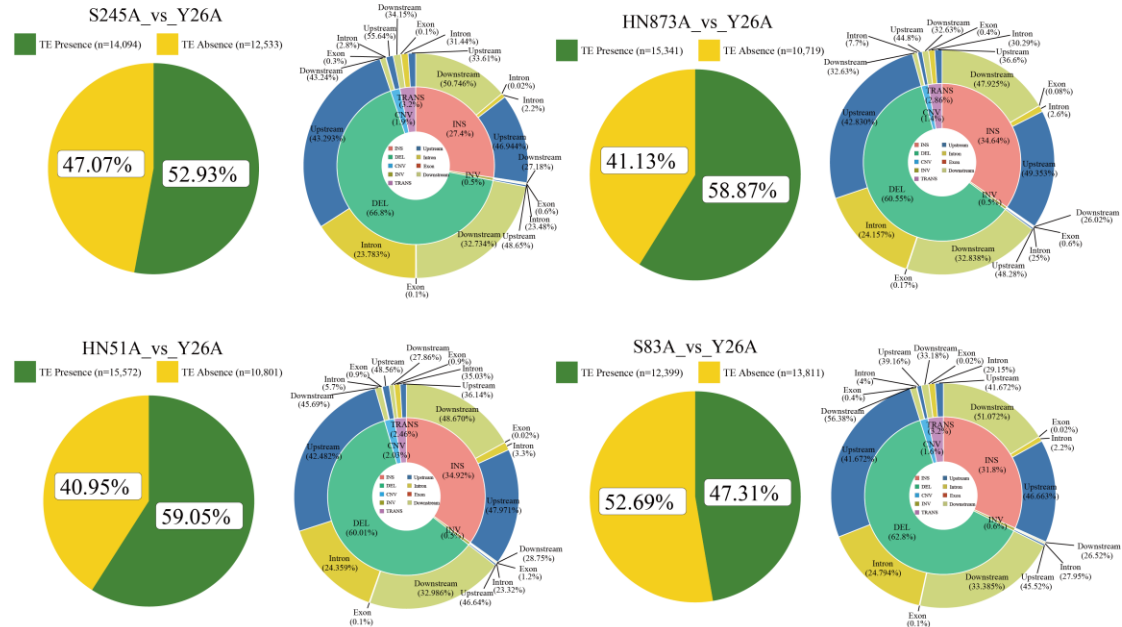


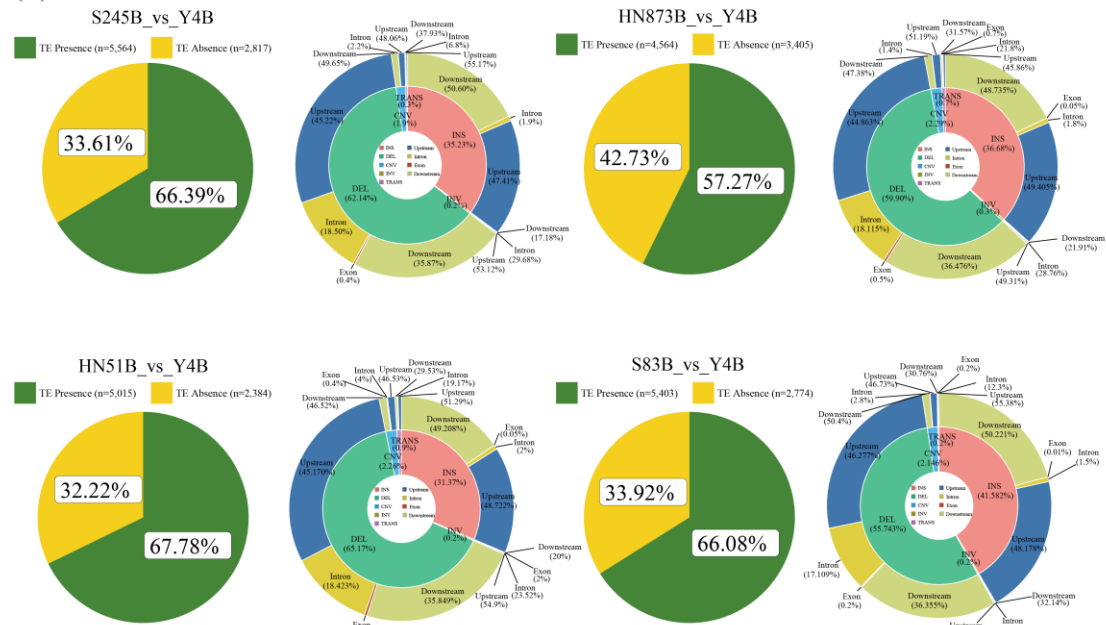
Figure 1 (Supplementary Figure 7). Structural Variations Among Peanut Genomes. The number of structural variant types identified in each genome compared to (A) Y4, (B) Y26, and

(C) S83. The structural variant types include INS (insertion), DEL (deletion), CNV (copy number variation), INV (inversion), and TRANS (translocation).

(A)



(B)



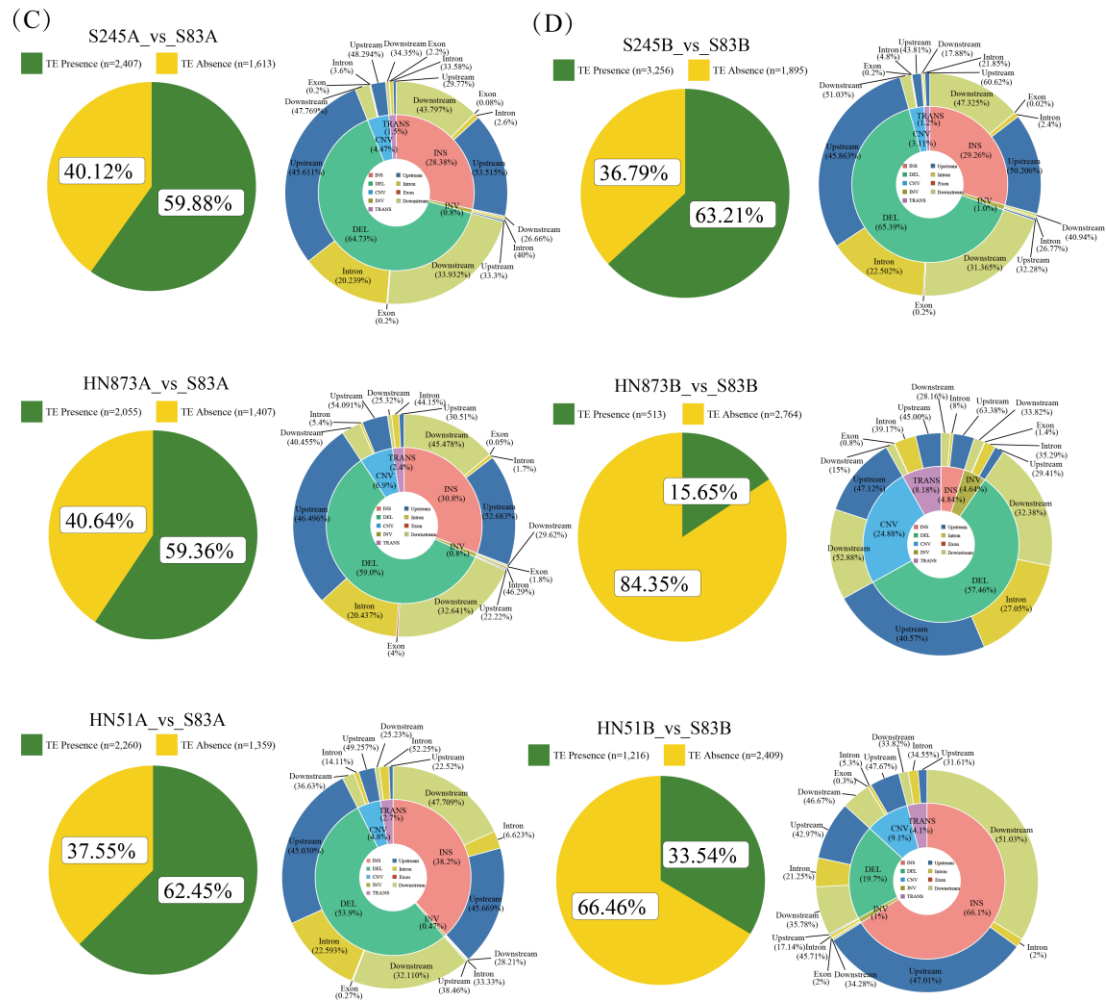


Figure 2 (Supplementary Figure 8). Transposable elements (TEs) and their associations with structural variations (SVs) in the peanut genome. The proportions of SVs associated with TEs (left) and the percentage of genomic features with high TE-derived SVs (right) are shown for (A) At vs. Ad (Y26), (B) Bt vs. Bd (Y4), (C) At vs. At (S83), and (D) Bt vs. Bt (S83). The left pie chart displays TE presence (green) and absence (yellow) within SVs. The right pie chart illustrates the distribution of TE-SV types in relation to genomic features, including upstream regions (2 kb upstream of the gene), downstream regions (2 kb downstream of the gene), introns, and exons.

Note that ADMIXTURE will always return K ancestral populations, even when one K is preferred ad hoc or by cross-validation (eg). Thus, the results need to be couched in this respect; they are not necessarily "real". It's good that the K separates phylogenetic lineages - highlight this better.

Response: Thank you for your insightful observation regarding ADMIXTURE. We used the

ADMIXTURE software to calculate and plot results for K values ranging from 2 to 20. Cross-validation (CV) error analysis showed that the rate of decrease in CV error was most pronounced at K=5 (Figure 3 (Supplementary Figure 14A)). This result is consistent with clustering observed in the phylogenetic tree and aligns with classifications based on peanut phenotypic information. Therefore, we selected K=5 as the optimal number of clusters; this value effectively separates phylogenetic lineages and reflects underlying biological differences. In the revised manuscript, we have updated this section (lines 358-363, (Supplementary Figure 14A)).

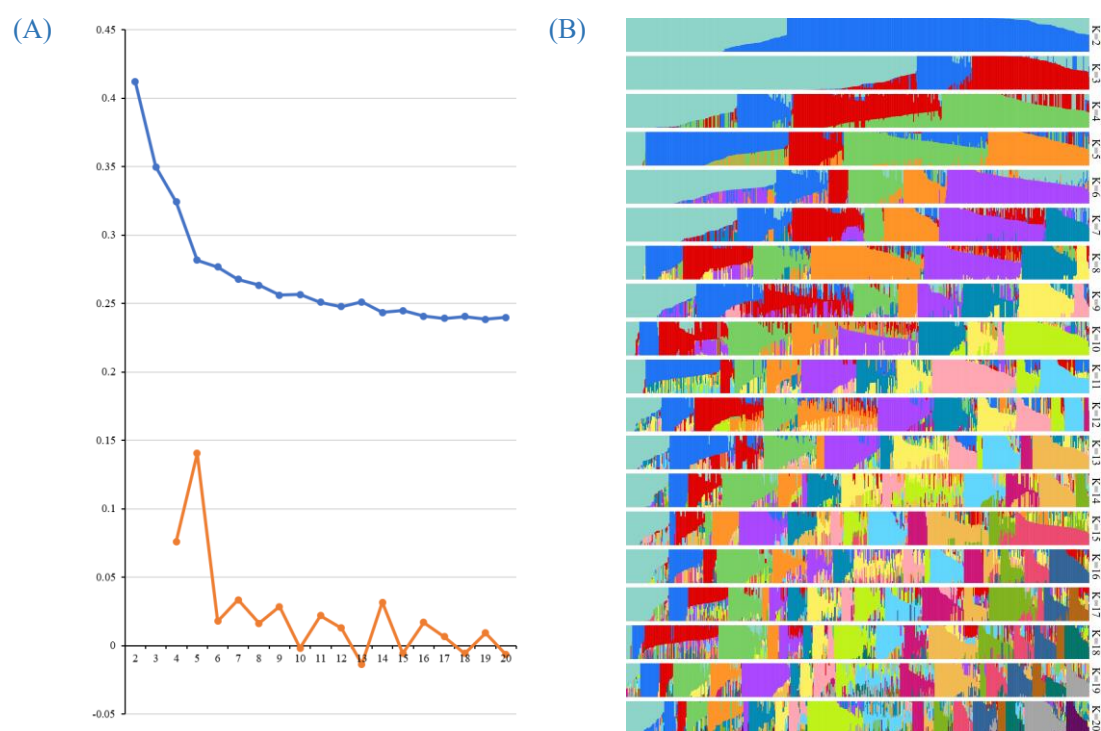


Figure 3 (Supplementary Figure 14A). Population structure of peanut populations. (A) Cross-validation (CV) error (top) and the rate of decrease in CV error (bottom). (B) Population structure plot.

- **Selective sweeps.** Recent TE or NUMT/NUPT insertions can fool these analyses by presenting highly homozygous stretches of the genome. Look through your results for these molecular phenomena – at least using a sanity check of about 100 random hits. You may wish to filter out “sweep” hits that better reflect TE or NUPT/NUMT homozygosity before functional enrichment analysis. Also, what happens in terms of enrichment when you consider top 1% instead of 5%? More conservative could be safer.

Response: Thank you for your insightful feedback. To evaluate the potential impact of TE insertions on the selected sweep regions, we categorized TE insertions into five time periods based on insertion times calculated using LTR_retriever software: < 10,000 years ago, 10,001-50,000 years ago, 50,001-150,000 years ago, 150,001-300,000 years ago, and > 300,000 years ago (**Supplementary Tables 21 & 22**). Combined with selected sweep intervals, revealed that TE insertions are associated with selected regions and genes. The results also indicate that recent TE insertions have a much smaller impact on selected regions and genes relative to earlier insertions (**Figure 4** (**Supplementary Fig. 15**)). Additionally, we adjusted the threshold for enrichment analysis to the top 1% and excluded regions affected by TE insertions (**Supplementary Fig. 16 & 17**). We have updated the relevant results in both the main text (**lines 380-388**) and supplementary materials (**Supplementary Tables 21 & 22; Supplementary Fig. 15-17**).

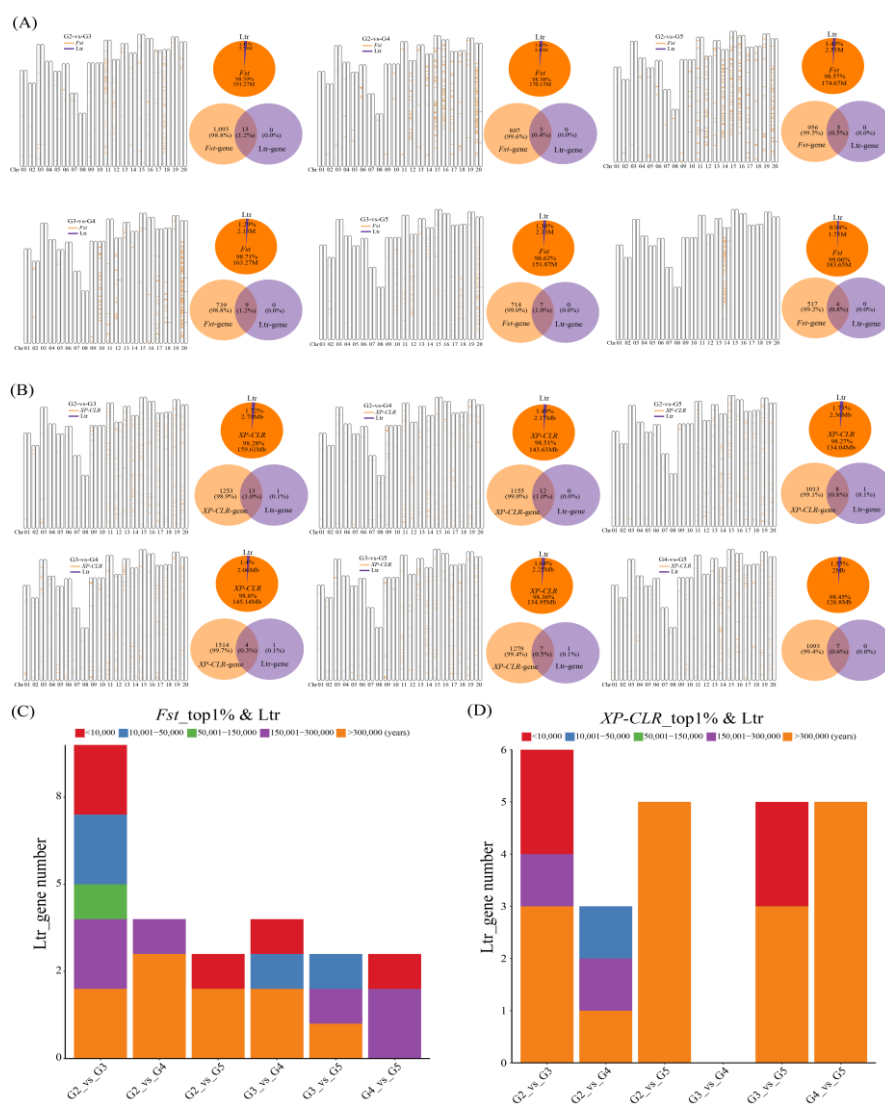


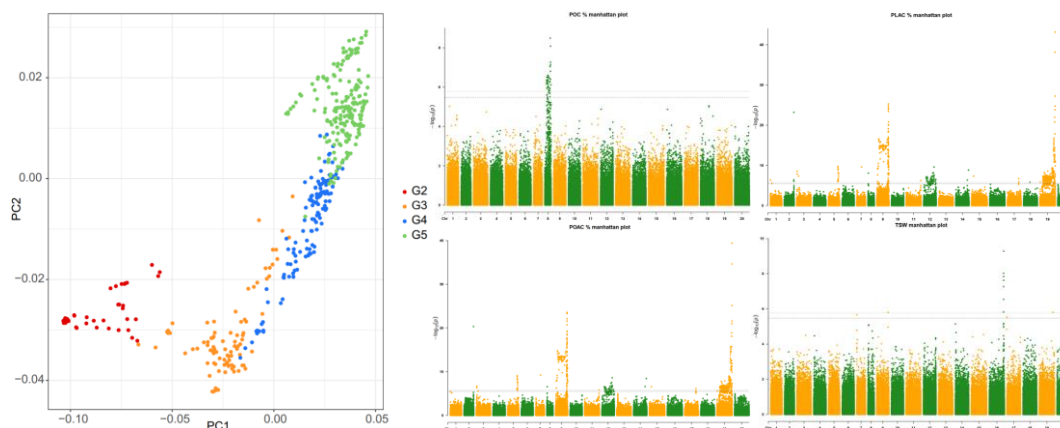
Figure 4 (Supplementary Figure 15). LTR insertions in selected sweep regions. Selective sweep analysis across six populations was performed using *Fst* and XP-CLR, including comparisons of G2 vs. G3, G2 vs. G4, G2 vs. G5, G3 vs. G4, G3 vs. G5, and G4 vs. G5. The chromosome schematic in panels A and B illustrates selective sweep regions (left) and selective sweep regions containing LTR insertion sites (right). Pie charts in panels A and B display the proportions of selective sweep intervals and LTRs, along with the corresponding number of genes. LTR insertions were categorized into five time periods based on insertion times calculated using LTR_retriever software: < 10,000 years ago, 10,001-50,000 years ago, 50,001-150,000 years ago, 150,001-300,000 years ago, and > 300,000 years ago. *Fst*_top1% (C) and XP-CLR_top1% (D) represent the top 1% intervals identified as candidate regions based on *Fst* and XP-CLR values, respectively. LTR-genes are defined as genes that overlap with LTR elements.

- GWAS: how did you convincingly control for what must be considerable underlying population structure in your sample? For example, among the different phylogenetic lineages.

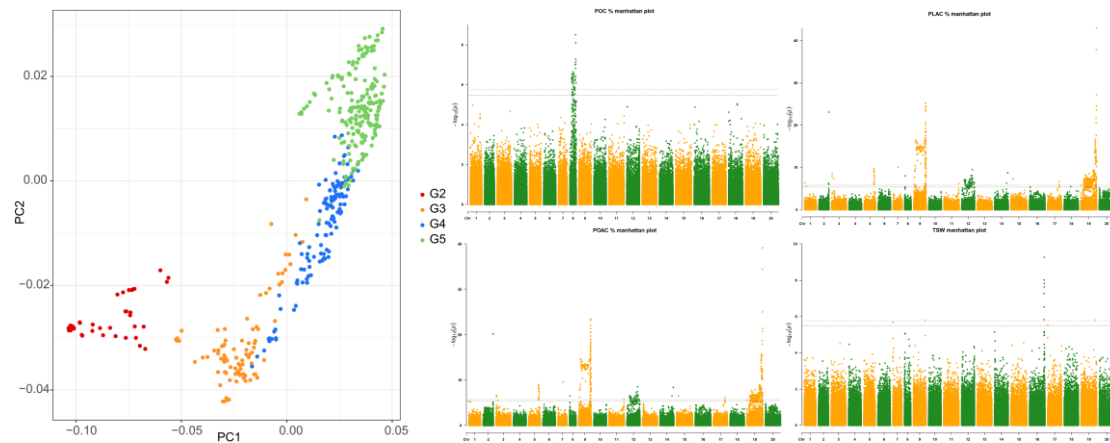
Response: Thank you for your insightful comments. To assess and mitigate the impact of population structure on GWAS results, we used PCA results as covariates to reduce the influence of population structure. Furthermore, to validate the effect of mixed samples on GWAS outcomes, we incrementally removed confounding samples based on PCA results. The GWAS analysis illustrated in **Figures 5** included 476, 474, 456, and 432 accessions, respectively.

From left to right, the analyses include PCA, oil content, oleic acid content, linoleic acid content, and thousand grain weight. The results show that after the removal of confounding samples, peak points remained consistent, confirming that PCA effectively controlled the influence of population structure on the outcomes. Accordingly, we conducted GWAS analysis using the entire population without excluding any samples. The results are shown as Manhattan and QQ plots in **Figure 5 (Supplementary Figure 18)**.

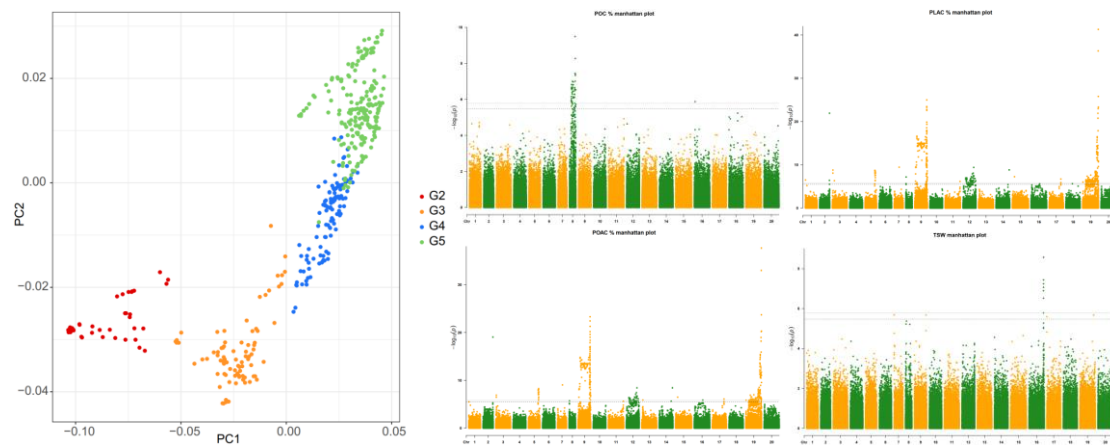
(A)



(B)



(C)



(D)

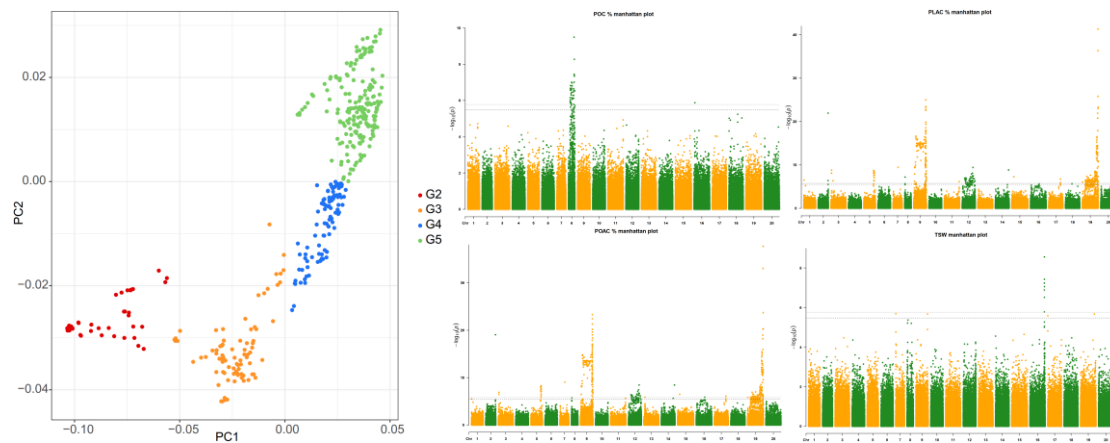


Figure 5. PCA and GWAS analysis. (A) PCA and GWAS Manhattan plot after removing G1 and admixed samples. (B) GWAS results after removing G2 and G3 admixed samples based on (A). (C) GWAS results after removing G3 and G4 admixed samples based on (A) and (B). (D) GWAS results after removing G4 and G5 admixed samples based on (A), (B), and (C).

- Were the candidate nut phenotype genes identified a priori, or solely through anonymous bioinformatics? If the former, and corroborated informatically, admit that please explicitly;

Response: Thank you for your valuable comment. To identify candidate genes, we applied a stepwise bioinformatics approach. First, we filtered variants from the population dataset to retain high-quality genotypes. Using GAPIT software, we identified signal intervals by integrating these genotypes with phenotypic data. Homologous comparisons were conducted with known functional genes in model crops such as rice and *Arabidopsis*; candidate genes were confirmed using phenotype data. To refine the selection, we focused on candidate nut phenotype genes within 200 kb of the peak regions and prioritized those with annotated functions related to the observed phenotypes. Gene haplotype analysis was performed using population data, and the relationship between haplotypes and phenotypes was used to finalize candidate genes. Functional validation was carried out in *Brassica napus* and *Arabidopsis*.

- WGD analyses. Please use syntenic analysis, not only Ks, which can be sensitive to underlying demographic differences, eg

Response: Thank you for the suggestion. We have incorporated syntenic analysis alongside *Ks* to address potential demographic sensitivities and provide a more comprehensive assessment. These revisions are detailed in **Figure 6 (Supplementary Figure 4)** and **lines 192-196**.

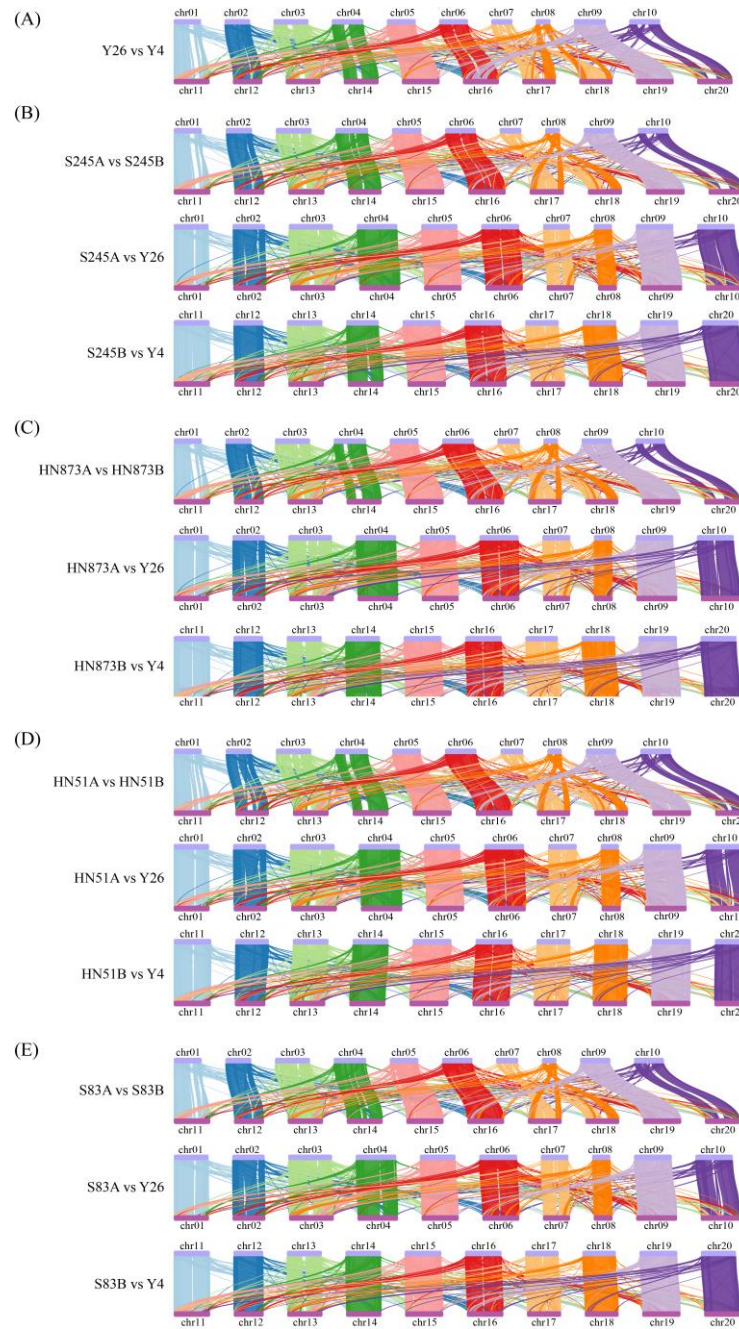


Figure 6 (Supplementary Figure 4). Syntenic analyses between Ad vs. Bd, At vs. Ad, and Bt vs. Bd. (A) Comparison between the diploid genomes Ad and Bd. (B) to (E) Syntenic analysis of the A and B subgenomes of *A. hypogaea* and their respective ancestral diploids. Homoeologous blocks comprising ≥ 5 gene pairs between Chr01–Chr10 and Chr11–Chr20 are represented by connecting lines.

- Functional enrichments - what were the exact backgrounds used? They should be included as files for replicability

Response: Thank you for pointing this out. We have added the exact background files used for functional enrichment analyses to **Supplemental Table 16** to ensure replicability. to ensure replicability. The background file includes all genes in each sample, along with their corresponding

KO and GO numbers, or any missing results for each gene (**Supplemental Table 16**). Significantly enriched pathways or terms were identified by comparing the proportion of candidate genes in each pathway or term to their proportion among all genes, using a threshold of $p < 0.05$.

- Phylogenetic trees should either display support values or parenthetical tree files should be provided in supplemental so investigators can evaluate robustness of results

Response: Thank you for highlighting the issue with displaying support values in the phylogenetic trees. Due to the large number of samples, the rectangular phylogenetic tree format could not clearly display support values. Accordingly, we created a circular plot that includes both support values and groupings. The revised figure has been added as **Supplementary Fig. 13**.

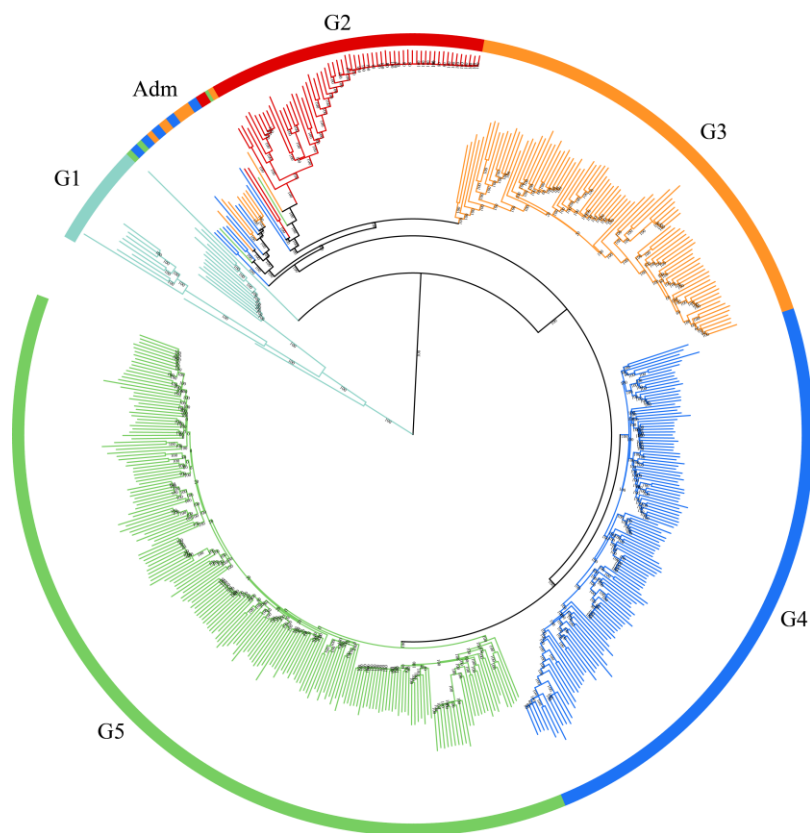


Figure 7. (Supplementary Fig. 13). Phylogenetic tree of 521 peanut accessions. The neighbor-joining tree, including support values, divides the 521 peanut accessions into five main groups: wild diploid (G1), *var. fastigiata* (G2), *var. vulgaris* (G3), *var. hypogaea* (G4), and *var. hypogaea* + *var. hirsuta* (G5).

- Fig. 4B shows a pretty clear horizontal cline. Carefully evaluate how this affected GWAS. And does the PCA suggest admixed populations (intermediate between red and green on X axis) in similar way as your other analyses?

Response: Thank you for your valuable comments. To assess the impact of the horizontal cline on

GWAS outcomes (as shown in **Figure 5**), we evaluated PCA results and progressively excluded samples with confounding factors. We then extracted genotypes from the remaining samples and used peak signal intensity to evaluate the accuracy of the results, which allowed us to identify optimal signal loci.

Analysis of PCA-confounding samples showed that they belong to an admixed population, consistent with findings from the phylogenetic tree and population structure analysis, which indicate a high degree of admixture. The materials include cultivated varieties that have undergone improvement through the introduction of wild germplasm (e.g., yuania9102). Due to their significant proportion of cultivated ancestry, these samples continue to cluster with cultivated types.

- Where is the TreeMix analysis? Where are the f stats in tabular form? What were the significance values used. Etc.

Response: Thank you for your comments. We apologize for the oversight in failing to update the method description after TreeMix-related analyses had been removed from the main text and supplementary materials. In the revised text, we have removed all TreeMix-related content. Gene flow analysis between groups was conducted by calculating the fd statistic in sliding windows of 100 kb with a 10 kb step. Windows with fewer than three informative SNPs were excluded, as were windows with negative Patterson's (D) statistic values (related to fd) and those with $fd > 1$. The top 1% of introgressed regions were retained for subsequent gene selection and enrichment analyses. The fd statistic results have been added to **Supplementary Table 24**.

Reviewer #2:

Remarks to the Author:

The study concerns a large sequencing analysis of peanuts, with several complete genomes and hundreds of peanut accessions characterized by whole genome sequencing. One major problem is that one of the basal analysis on the estimated age of polyploidy is fundamentally flawed, leading to wrong conclusions propagating through the manuscript:

- The authors infer the age of divergence of two genomes from peaks in Ks curves constructed using the WGD software. The WGD software is constructed to find evidence for ancient Whole Genome Duplications, by intragenomic (not intergenomic) comparison. When making an intragenomic comparison (a comparison of a genome with itself), it's clear that there will always be 100% self-identity. Since the software is searching for residual, much weaker similarities from a WGD, it filters Ks values at (or close to) zero. However, when making comparisons between two different genomes, there is no self-identity to eliminate. Each gene should be compared to its best match in the other genome (a comparison of orthologs - the authors use the word "ortholog", but in fact, orthologous comparisons are filtered out by the software they use).

- The authors kindly added several self-comparisons to their figure. These new self-comparison curves serve as controls for the method. Considering the objectives of the authors, these control curves should give values of zero divergence because the genomes are identical to themselves. In fact, none of the self-comparisons have a peak at zero. The methods used by the author are unable to show that a genome has zero divergence from itself!

- In the same way, the method is unable to give a correct divergence of two genomes that are very similar to each other. Reiterating, this is because the method is designed for self-comparisons of a polyploid genome to indicate the age of divergence of the subgenomes. Values at (or close to) zero are filtered out. This is the reason the peak in the S245 (Ahy) vs S245 (Ahy) curve is Ks 0.025 (=1.58 Mya -the age divergence of A and B subgenomes), instead of zero.

- For instance, the genome of Y4 (*A. ipaensis*) is, in fact, remarkably similar to the B subgenome of *A. hypogaea*. From the author's Ks data (S83B.cds.fa_Y4.cds.fa.ks.tsv) 27,923 genes are compared between S83 and Y4. Of these 24,541 have a Ks of zero. Indeed 23,584 of the genes have no nucleotide differences at all in their coding regions (Ks and Ka are both zero). The filtering of Ks=0 values combined with some sort of curve regression results in completely wrong Ks=0 values being represented on the graph.

- The authors did not rescale the vertical axis as requested (It was relabelled as "Frequency", but the curves remain the same). If a number of gene pairs (=frequency) had been used, it would have been possible to see that data had gone missing from the different areas under the different curves.

Response: Thank you for your valuable comments, that helped us to improve our manuscript significantly on the divergence estimation. We agree that WGD software is more appropriate for

identifying ancient whole genome duplications between species (e.g., At vs. Bt & Ad vs. Bd). We appreciate your suggestion to retain $Ks = 0$ when calculating interspecies differentiation. In the updated results, we have preserved the $Ks = 0$ values as suggested.

To accurately determine the timing of tetraploid formation, we used the alignment software JCVI (v.1.3.4)¹¹ and the KaKs calculation software KaKs_Calculator (v.3.0)¹² to analyze genetic differentiation between diploid and tetraploid subgenomes (Materials and Methods section (**lines 869-877**)). We conducted biased cross-validation to identify the optimal bandwidth for the results, followed by curve fitting. First, we used JCVI (v.1.3.4) to identify orthologous gene pairs. We then utilized KaKs_Calculator (v.3.0) to calculate the Ka and Ks values for these orthologous gene pairs in the A and B genomes/subgenomes (Ad vs. Bd & At vs. Bt) across the six assembled genomes, as well as in the published cultivated variety Tifrunner². Additionally, we calculated the Ka and Ks values between the tetraploid subgenomes and diploid genomes (At vs. Bd & Bt vs. Bd). Using the biased cross-validation method, we determined the optimal bandwidth for different comparisons and fitted the Ks results to estimate the timing of genetic differentiation. The following results were obtained:

- (1) Between the AB genomes/subgenomes, we identified an average of 22,419 gene pairs, of which 20,948 exhibited differentiation ($Ks \neq NA(0)$) (**Table 1 (Supplemental Table 10)**). Notably, the proportion of gene pairs with $Ks = NA(0)$ was significantly higher in the tetraploid subgenomes (average of 7.5%) than in the wild diploids (1.61%). This finding suggests gene exchange or chromosomal rearrangements among the tetraploid subgenomes, consistent with previous studies².
- (2) By performing biased cross-validation to identify optimal peaks (Bcv-peaks) and incorporating the peanut mutation rate of 8.12×10^{-9} , we calculated the average divergence time ($T = Ks/2r$) between the A and B genomes/subgenomes to be approximately 1,721,000 years, consistent with previous research (**Table 1 (Supplemental Table 10, Figure 2A)**).
- (3) Further analysis revealed significant differences between At vs. Ad and Bt vs. Bd in terms of the number of gene pairs and the proportion of $Ks = NA(0)$ (**Table 1 (Supplemental Table 10, Figure 8 (Figure 2A))**). For At vs. Ad, an average of 25,050 gene pairs were identified, with $Ks = NA(0)$ representing 36.95% on average. In contrast, for Bt vs. Bd, an average of 28,111 gene pairs were identified, with $Ks = NA(0)$ representing 79.55% on average. These results suggest greater similarity between Y4 and the tetraploid B subgenome, indicating that Y4 may be a source of the tetraploid B subgenome. Based on the optimal Bcv-peak results, the average divergence time between At and Ad is approximately 155,000 years, whereas the divergence time between Bt and Bd is approximately 11,000 years.
- (4) Regarding the Y-axis label, we have verified this section and confirmed that the vertical axis represents the frequency of gene pairs. This correction has been included in the revised text.

Overall, the Ks analysis results suggest that the formation of the cultivated tetraploid occurred within the past around 11,000 years, consistent with previous findings^{2,3}.

Thank you again for your valuable suggestions, which have been instrumental in advancing our understanding of tetraploid peanut formation.

Table 1 (Supplemental Table 10) *Ks* values and divergence times for At vs. Bt, Ad vs. Bd, At vs. Ad, and Bt vs. Bd.

	Total number of KS	Number of KS≠NA(0)	Number of NA(0)	Percentage of NA(0)	Bcv.peak	Years
S245A_S245B	20994	19732	1262	6.01%	0.0286405	1763578
HN873A_HN873B	23488	21802	1686	7.18%	0.02756844	1697564
HN51A_HN51B	23365	21389	1976	8.46%	0.02771358	1706501
S83A_S83B	23547	21714	1833	7.78%	0.02781789	1712924
TifA_TifB	21242	19526	1716	8.08%	0.02801186	1724868
Y26_Y4	21875	21522	353	1.61%	0.02800128	1724217
S245A_Y26	24921	15931	8990	36.07%	0.002608347	160613
S245B_Y4	28247	6236	22011	77.92%	0.00022127	13625
HN873A_Y26	26101	16104	9997	38.30%	0.002346743	144504
HN873B_Y4	28568	5617	22951	80.34%	0.000175027	10778
HN51A_Y26	25477	15802	9675	37.98%	0.002257026	138979
HN51B_Y4	28860	5460	23400	81.08%	0.000173256	10668
S83A_Y26	24911	15851	9060	36.37%	0.002509033	154497
S83B_Y4	28213	6048	22165	78.56%	0.00017723	10913
TifA_Y26	23839	15253	8586	36.02%	0.002884051	177589
TifB_Y4	26669	5372	21297	79.86%	0.000187491	11545

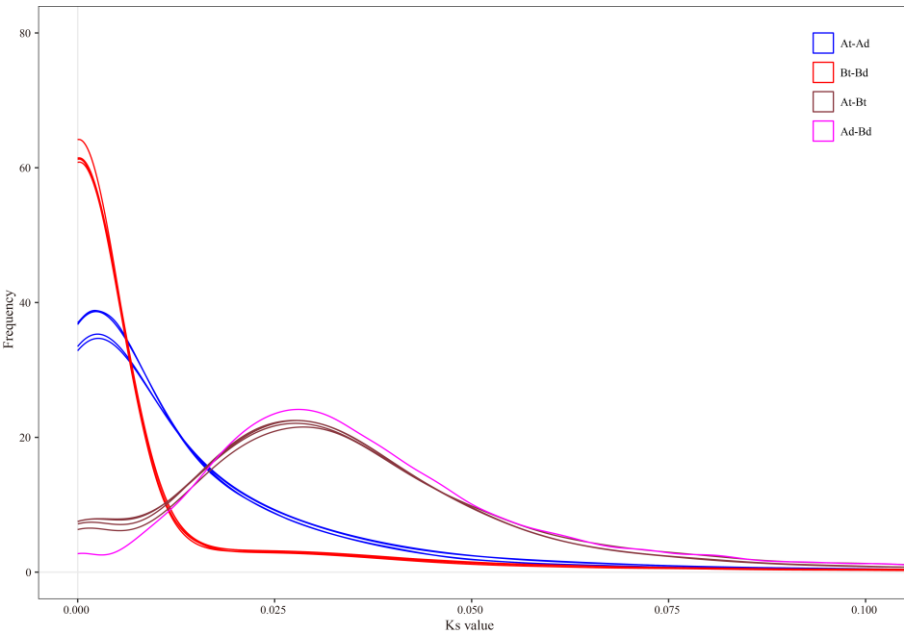


Figure 8 (Figure 2A). Distribution of *Ks* values for orthologous genes. Distribution of *Ks* values for orthologous genes among the subgenomes of *A. hypogaea*, including the A subgenomes (S245A, HN873A, HN51A, and S83A) and the B subgenomes (S245B, HN873B, HN51B, and S83B), as well as their suspected progenitor genomes (Y26 and Y4).

Reviewer #3:

Remarks to the Author:

Overall evaluation

This manuscript presents the sequencing, assembly and analysis of six peanut genomes. It expands the analysis with an association analysis between two traits, oil content and seed weight and a genotype collection of 521 accessions. The manuscript is clear and easy to follow, although it contains a lot of information of results representing an incredible amount of work. The material and methods and the discussion should be improved in order to have a better support of the results presented. Overall the manuscript is well built, has interest and present relevant results not only for the peanut scientific community but also for the genomic community working in genome evolution and polyploidy.

Point-By-Point Manuscript Evaluation:

A. Summary of the key results

The manuscript titled “Origin, evolution and domestication of cultivated peanuts” describes the evolution of tetraploid peanut (*Arachis hypogaea* L.) genome through the sequencing and analysis of two diploid, and four tetraploid genomes. It also identifies genomic regions associated to domestication and to several agronomical traits through the re-sequencing of more than 500 accessions. Finally, the authors analyze the transcriptomics and metabolomics networks of the different peanut domesticated lineages. The results are summarized in the following points:

- Sequencing, assembly, and annotation of six representative accessions of peanuts, two diploids’ progenitors (*A. duranensis*, AA, Y26 and, *A. ipaensis*, BB, Y4) and four tetraploids (*A. hypogaea*, var. *fastigiata*, var. *vulgaris*, var. *hypogaea*, var. *hirsute*). All the genomes were sequenced with PacBio HiFi, scaffolded with Hi-C data and the gaps filled with ONT. Genome size ranged from 1.18 Gb to 2.62 Gb. Each of the six assemblies presented excellent assembly metrics in terms of Illumina mapping rate (>99%), BUSCO completeness (>95%), LAI (>20), and Merqury QV (>40). The genome annotation produced 34-35K gene models for the diploid accessions and 71-72K for the tetraploids.
- TE analysis reported a range from 75 to 78% of repetitive elements. B subgenomes have a higher TE content than A being the Gypsy elements the main contributors. Age divergency between both subgenomes was dated 1.58 MYA, meanwhile the divergency between the AA-At and BB-Bt subgenomes was estimated 0.27 and 0.29 MYA in agreement with previously published results. A deeper TE analysis revealed different insertion patterns for the LTR elements after the polyploidization event. Centromeric regions were different between the diploid and the tetraploid accessions. Retand elements expanded after the polyploidization event contributed to the differentiation between AA and At subgenomes.
- The analysis of the SVs between the A, B sub-genomes and their ancestors revealed a higher number of inversions and translocations in the A sub-genome, compared with the B. Genes closed to the SVs were enriched in functions such as flavonoid biosynthesis, brassinosteroid biosynthesis,

fatty acid metabolism and linoleic acid metabolism. Comparisons across different *A. hypogaea* subspecies revealed also a different functional enrichment which could contribute to the different phenotypes in some lineages like *hirsuta*.

- Population structure analysis with 521 peanut accessions revealed 5 distinct groups. Gene flow was detected between two different groups: G2-G3 and G4-G5.

- GWAS analysis on peanut oil content reported a significant signal on chromosome 8 with 151 genes. A gene named *AhWRI1* was identified as responsible for the trait. Two haplotypes were identified in the collection with different oil content (~48% hap 1 and ~54% hap 2) associated with DE. Overexpression of this gene on rapeseed and soybean lines produced accumulation of lipid droplets supporting the involvement of this gene in the peanut oil content trait. Y1H assays allowed to associate several genes such as *AhFUS3*, *AhABI3* and *AhLEC1* to the regulation of the expression of this gene. A mutation on the E-Box promoter of this gene was identified as responsible for the different interaction with the *AhFUS3* gene.

- GWAS analysis on seed size and weight revealed a peak on chromosome 16 spanning 130 genes. A gene named *GSA1* was identified as one of the possible candidate genes for this trait based on similar phenotypes in rice. Two different haplotypes were identified for this gene, hap1 with 43 accessions and an average grain size of 650 gr. and hap2 with 294 accessions and an average grain size of 920 gr. The association of this candidate gene with the trait was verified with transcriptomic data as well as some analysis on the promoter of the gene.

- Transcriptomics and metabolomics comparisons between the accessions S83 and S245 identified key genes involved on lipid and anthocyanin content.

B. Originality and significance: if not novel, please include reference

The manuscript is original and novel. It expands our knowledge about the peanut domestication, and it generates important data for the community. It also describes two genes associated to two important agronomical traits.

C. Data & methodology: validity of approach, quality of data, quality of presentation.

The methodology is valid, and the quality of the presentation is good. More details in the material and methods are needed, specially to assess the RNA-Seq and the GWAS analysis that it has been evaluated under the principles that it has the right number of replicates.

D. Appropriate use of statistics and treatment of uncertainties

It is not applicable to this type of research, except for the GWAS and the RNA-Seq which looks fine. Nevertheless more details are needed in the M&M for a full assessment.

E. Conclusions: robustness, validity, reliability

Most of the conclusions are robust and valid although the lack of some information on the M&M (e.g., replicates) makes difficult to have a full assessment.

F. Suggested improvements: experiments, data for possible revision

There are several important improvements that should be addressed if the manuscript is accepted for publication.

a. Population structure analysis has been slightly treated in the manuscript. There are several publications that describe in detail some of the peanut population structure. This manuscript presents an interesting analysis but it does not compare its results with what it was published before. Although, the manuscript is already too extense, I found missing a dating of the different splits of the different populations and how this is related with the current status of the species (e.g., domestication, distribution...).

Response: Thank you for your detailed review and valuable feedback regarding our manuscript. We appreciate your observations and have made substantial revisions to address your suggestions.

We have provided a detailed summary of recent studies that utilized whole-genome resequencing data to analyze peanut population structure. Liu et al. (2022) categorized 203 resequenced accessions into four genetic groups. Zheng et al. (2024) performed a population structure analysis using chloroplast genomes from 113 accessions (36 wild *Arachis* and 77 cultivated accessions) and whole-genome resequencing of 77 cultivated accessions. They identified two main varieties: Ahh (*var. hypogaea*) and Ahf (*var. fastigiata*), with further subdivisions within these varieties.

Our population structure analysis demonstrates that the resequencing results are consistent with those reported by Liu et al. (2022) and Zheng et al. (2024). The resequenced peanuts were classified into two subspecies (*var. hypogaea* and *var. fastigiata*) and six botanical varieties (*var. hypogaea*, *var. hirsuta*, *var. fastigiata*, *var. vulgaris*, *var. aequatoriana*, and *var. peruviana*). These classifications align with widely accepted categorizations based on morphological traits and geographical origin. This consistency with botanical classifications enhances the accuracy of our classification and provides valuable insights concerning peanut origin and domestication.

Additionally, we used fastsimcoal2 (v.2.8)¹³ to estimate divergence times (Materials and Methods section (**lines 907-917**)), which revealed the following: a divergence between *var. fastigiata* and *var. hypogaea* approximately 9,500 years ago; a subsequent divergence between *var. fastigiata* (G3) and *var. vulgaris* (G2) around 7,200 years ago; and a divergence of *var. hypogaea* into *var. hypogaea* (G4) and *var. hirsuta* (G5) approximately 8,900 years ago (**Figure 9 (Supplementary Figure 14B)**).

The differentiation among populations has significantly influenced phenotypic traits and their functional roles. In the revised Discussion, we made the following speculations (**lines 593-606**). “Based on the results of the phylogenetic tree, population structure, and population dynamics, we inferred that the differentiation of peanuts into two subspecies and four major cultivated varieties was primarily driven by yield and usage requirements. The differentiation between *var. fastigiata* (G2+G3) and *var. hypogaea* (G4+G5) is largely attributed to significant differences in seed size,

which directly impact yield. *Var. fastigiata* (G2+G3) types are primarily cultivated for consumption; breeding efforts are focused on the development of varieties with a higher seed count (e.g., G2 multi-seed types) to increase yield. In contrast, *var. hypogaea* (G4+G5) types are predominantly grown for oil extraction. The differentiation between G4 and G5 mainly arises from differences in planting methods. G4, with its upright growth habit, supports higher planting densities, thereby increasing yield. Due to its creeping or semi-creeping growth habit, G5 enhances yield by producing more effective fruiting nodes. The population dynamics results provide a clear representation of these breeding processes.”

Thank you again for your thoughtful suggestions. We believe these revisions enhance the depth and clarity of our paper. We look forward to further refining our research based on your valuable feedback.

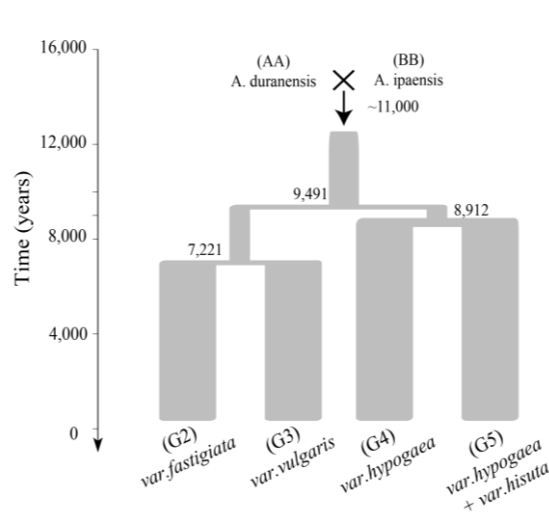


Figure 9 (Supplementary Figure 14B). Demographic history of peanut populations. Schematic representation of divergence times estimated using fastsimcoal2. Vertical axis represents the estimated timing of population divergence.

b. Transcriptomic and metabolomic analysis of the S83 and S245 accessions does not describe the genes involved in the trait. The last result section describes the transcriptomic and metabolomic analysis associated to some secondary compounds, but it does not describe in detail the WGCNA (eg., number of clusters) or furthermore the “key genes highly correlated with differentially abundant metabolite content across seed development stages”. The transcriptomic analysis is not summarized in the M&M section, so it is difficult to evaluate if it was performed with the right number of replicates and the adequate tools.

Response: We appreciate your thorough review and the valuable comments regarding our transcriptomic and metabolomic analysis. We recognize the need for greater clarity and detail in presenting this data. Please find our point-by-point response below:

(1) Description of key genes correlated with metabolite content:

We apologize for the lack of clarity in this regard. We have expanded the Results section to include

a detailed description of the genes associated with the metabolite of interest. Specifically, we have added new **Supplementary Tables 33 & 35**, which provide a comprehensive list of these key genes, their correlation coefficients, and their known or predicted functions.

For oil content, we identified several fatty acid desaturases (FAD) and ketoacyl-ACP synthases (KAS) through correlation analysis of metabolite abundance and gene expression levels. We suspect that these genes play crucial roles in determining metabolite content. Additionally, we analyzed the phylogenetic relationships of FAD and KAS enzymes in peanut (**Supplementary Fig. 24**), including those reported in previous studies^{14–16}.

For anthocyanin content, we focused on MYB and bHLH transcription factors. Previous studies have shown that these transcription factors influence peanut seed coat color^{17–19}. We conducted phylogenetic analyses of these transcription factors, similar to those conducted for other MYB and bHLH genes (**Supplementary Fig. 24**).

Based on these findings, we have provided a more detailed description of the key genes in the Results section (**lines 513-524 and 536-552** in the revised manuscript).

(2) WGCNA details:

Thank you for pointing out this oversight. We have now included a more detailed description of WGCNA in the Results section. This includes the number of clusters identified (11 clusters & 20 clusters), the soft-thresholding power used (power = 16 & 6), and the minimum module size (50 genes). Additionally, we have added a supplementary figure displaying the cluster dendrogram and module colors (**Supplementary Fig. 23**).

(3) Transcriptomic analysis in Materials and Methods:

We sincerely apologize for this omission. We have now added a comprehensive description of the transcriptomic analysis to the Materials and Methods section (**lines 985-1018**). Due to the journal's word count limitations, we have included these results and data analyses in **Supplementary Table 31**. We believe these additions allow a more thorough evaluation of our methodologies and results.

We have made these changes in the revised manuscript and hope they adequately address your concerns. We greatly appreciate your insightful comments, which have undoubtedly improved the clarity and quality of our paper. If you have any further questions or need additional information, please do not hesitate to ask.

Thank you for your time and expertise in reviewing our manuscript.

c. Discussion section is more of a conclusion section. Discussion of the manuscript should be improved. Usually discussion section focus is to context ulize the results presented in a manuscript with what it is known. Conclusion is a summary of the results with some highlights. In this manuscript, the discussion section is the second one. The results incorporate some pieces that it could be included in the discussion section, but the manuscript as such lacks on a solid discussion section presenting some of the results, in a wider context. For example, the authors could compare the TE dynamics of peanut with other polyploid species. They could bring to the discussion an interesting topic like the genomic shock an associate it with the different insertion rates in both subgenomes. They could talk about subgenome

dominance. Same could be applicable to the population structure section or the selection of the gene candidates. How are those compared with other legumes like soybean?

Response: Thank you for your feedback and suggestions. In the revised text, we have streamlined the content to integrate the results more effectively into the Discussion section. We have added a segment concerning the impact of TE insertions on the genomes and key trait genes in peanut, cotton, and soybean. Additionally, in response to your suggestions, we have addressed the relationship between population differentiation and phenotypic traits. Thank you again for your valuable input; these revisions have substantially enhanced the overall quality of our article.

d. The manuscript does not have a statement ad scribing that the research presented has followed the Nagoya protocol. There are some samples from different countries which should either meet the Nagoya protocol, or state that they were collected before the implementation of it.

Response: Thank you for your suggestion. We have included the statement, "The research presented has followed the Nagoya Protocol," in the Ethics Declarations section (line 1140).

e. The material and method information could be improved. The M&M information should be enough to reproduce the experiments. In this sense, the M&M lacks from important information. Some examples are:

i. The description of the accessions used on the analysis is poor, especially for the 521 accessions. They are summarised in the suppl. table 15, but it has not been specified if they are part of a collection, if the accession ID is the same than the collection. The information about the origin of the samples is poor and it does not include the type of the accession (eg., cultivated vs wild, subspecies...).

Response: Thank you for your feedback regarding the description of the accessions used in the analysis. We acknowledge that the initial information, particularly for the 521 accessions, was insufficient. We have updated **Supplementary Table 19** and the Materials and Methods section (lines 674-686) to include detailed information about accession sources, such as collection details, type (wild, landrace, and cultivated), ploidy (diploid and tetraploid), subspecies (*var. hypogaea* and *var. fastigiata*), and variety (*var. fastigiata*, *var. vulgaris*, *var. hypogaea*, *var. hirsuta*, *var. peruviana*, and *var. aequatoriana*). These updates improve the clarity and completeness of the accession descriptions in the manuscript.

ii. The version number should be specified for all the programs used in the manuscript.

Response: Thank you for highlighting this issue. We have updated the manuscript to specify the version numbers for all programs used in the analysis, thereby ensuring clarity and reproducibility.

iii. The gene annotation methods is poorly described. It does not mention how the RNA-Seq data was mapped nor the quality evaluation process. It does not describe the weight to each

of the evidences used for EVM. The functional annotation has been poorly described, especially because eggNOG-mapper tool has some limitations when it is used with plant species.

Response: Thank you for your feedback. In the revised text, we have updated the descriptions of RNA-seq data mapping and quality evaluation steps (lines 835-865). Specifically, for RNA-seq annotation, we sequenced various developmental stages (embryo, cotyledon, hypocotyl, axillary bud, lateral branch, root nodule, root, stem, leaf, flower, aerial peg, subpeg, peanut shell, pod, and seed) using Illumina technology, with three biological replicates for each tissue, along with mixed-sample third-generation transcriptome sequencing. For second-generation sequencing, reads were aligned using HISAT2, BAM files were merged with SAMtools, and transcripts were assembled with StringTie to produce GTF files. Longest transcripts were extracted using TransDecoder-v5.5.0 and compared against the UniProt_SwissProt database for optimal alignment, generating high-confidence GFF3 files. Third-generation full-length transcripts were assembled using the Iso-Seq3 pipeline. The Evidence Modeler (EVM) weighting scheme was configured as follows: Liftoff homology-set (weight :10); ISO-set (weight :10); TransDecoder (weight :8); PASA (weight :5); Ab-set (weight :2) Homology-set (weight :1).

We have also improved gene function annotation by integrating data from the NR (non-redundant, NCBI), Pfam, InterPro, and Swiss-Prot databases (Table 2 (Supplemental Table 7)), enhancing the accuracy and comprehensiveness of the annotations. Gene Ontology term annotations and protein domains were identified using InterProScan. KEGG annotations were performed using kofam_scan-1.3.0 with the parameter -e 1e-5.

Table 2 (Supplemental Table 7) Functional annotations of predicted genes across the six peanut genomes.

Database	Y26		Y4		S245		HN873		HN51		S83	
	Gene number	Percentage (%)	Gene number	Percentage (%)	Gene number	Percentage (%)	Gene number	Percentage (%)	Gene number	Percentage (%)	Gene number	Percentage (%)
Annotation	33970	98.69%	35408	98.88%	71368	98.56%	74248	98.81%	73222	98.80%	71334	98.74%
GO	25098	72.92%	26629	74.36%	52331	72.27%	54708	72.81%	54271	73.23%	53252	73.71%
KEGG	17470	50.76%	18467	51.57%	36473	50.37%	38012	50.59%	37734	50.92%	36847	51.00%
Nr	33749	98.05%	35201	98.30%	70854	97.85%	73709	98.09%	72666	98.05%	70880	98.11%
Pfam	24540	71.30%	25791	72.02%	50701	70.02%	52432	69.78%	52161	70.38%	51849	71.77%
Interpro	28050	81.49%	29349	81.96%	57618	79.57%	59997	79.84%	59439	80.20%	58751	81.32%
PANTHER	29272	85.04%	30456	85.05%	59811	82.60%	62364	82.99%	61584	83.10%	60978	84.40%
Swiss-Prot	33382	96.98%	34809	97.20%	69792	96.38%	72833	96.93%	71845	96.94%	70040	96.95%

iv. There is not description about how the the phenotyping of the 521 accessions were performed (oil content, seed size and weight..., number of plants, number of years, replicates in the field...).

Response: We appreciate your feedback and suggestions. In the revised text, we have included a detailed description of the material phenotyping process (lines 955-966). This description outlines the instruments used for phenotypic measurements, the setup of planting plots, the number of sample replicates, and the timing and location of the surveys. Additionally, we utilized the BLUP package to fit multi-year and multi-location phenotypic traits, adjusting for year and location effects. This approach allowed us to generate robust phenotypic data suitable for GWAS analysis.

f. The data availability for the project PRJNA1106792 is incomplete. It has only 4 of 6 genome assemblies. No annotation, transcriptomic data or re-sequencing data.

Response: Thank you for pointing out the issue with data availability for project PRJNA1106792. We have addressed this concern, and the genomic data for all six materials is now accessible on NCBI under accession number PRJNA1106792. The raw transcriptomic and resequencing data have also been uploaded and will be released after publication of the article. Updated genomic data can be accessed here: <https://www.ncbi.nlm.nih.gov/datasets/genome/?bioproject=PRJNA1106792>. The annotation files are still under review. We are actively working to expedite this process and will provide updates as soon as they become available. Thank you for your patience.

Some discretionary suggestions to improve the data support of the manuscript are:

- a. **Add the completeness values of merquary, as well as the K-mer distribution curves to have a better perspective of the result of the duplicated regions. Additionally, having some description of the possible contaminations with Blobtools could be good, although the comparison between cultivars is a good support for the absence of contaminants.**

Response: We appreciate your feedback and suggestions. Based on your suggestion, we analyzed and plotted the K-mer distribution curves for the six assembled genomes (**Figure 1 (Supplementary Figure 2)**). Additionally, the potential contamination was assessed. Due to the large genome size, we divided the genome into 500kb segments and used DIAMOND to align the genome sequences to the Reference_Proteomes of UniProt database, with the parameters ‘max-target-seqs 1’ and ‘evalue 1e-25’. The results showed that over 98% of the sequences were aligned to species in the legume family (**Table 3**).

Thank you again for your valuable suggestions, which have been crucial in ensuring successful construction of the T2T genome.

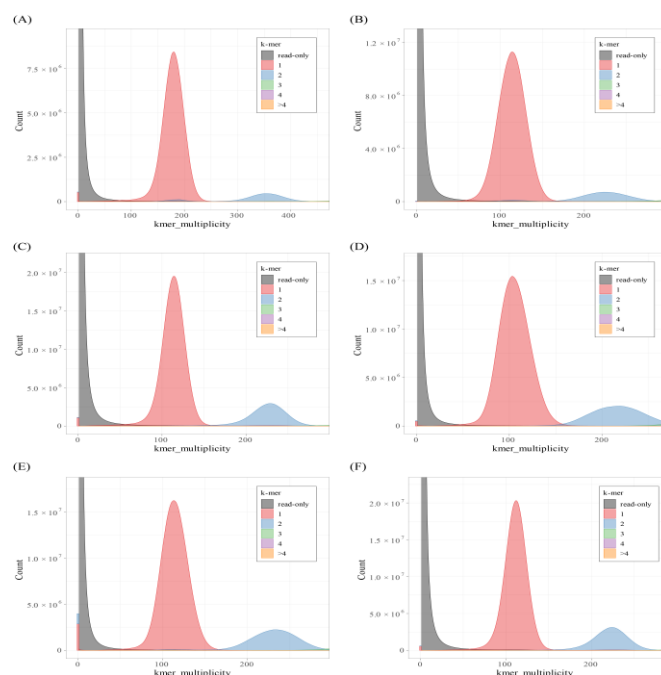


Figure 10 (Supplementary Figure 3). Merquary copy number spectrum plots for the six assembled peanut genomes. The copy number spectrum (spectra-cn) of identical K-mers is

displayed as stacked histograms, color-coded by copy numbers found in the combined assembly. The samples are: (A) Y26; (B) Y4; (C) S245; (D) HN873; (E) HN51; and (F) S83.

Table 3 The assessment results of the potential contamination among the six assembled genomes.

Organism	Y26	Y4	S245	HN873	HN51	S83
<i>Acacia crassiparva</i>	778	769	1489	1502	1510	1499
<i>Acer saccharum</i> (Sugar maple)	0	0	0	0	1	0
<i>Actinidia chinensis</i> var. <i>chinensis</i> (Chinese soft-hair kiwi)	0	5	6	7	6	6
<i>Arachis duranensis</i> (Wild peanut)	760	686	1321	1297	1321	1277
<i>Arachis hypogaea</i> (Peanut)	637	1345	2055	2120	2120	2058
<i>Brassica oleracea</i> var. <i>oleracea</i>	1	0	0	1	0	0
<i>Cajanus cajan</i> (Pigeon pea) (<i>Cajanus indicus</i>)	38	28	65	60	61	51
<i>Centaurea solstitialis</i>	54	52	122	126	130	117
<i>Cinnamomum micranthum</i> f. <i>kanehirae</i>	0	1	0	0	1	0
<i>Coffea arabica</i> (Arabian coffee)	0	0	1	0	0	0
<i>Cucumis melo</i> var. <i>makuwa</i>	0	0	1	1	1	0
<i>Cuscuta europaea</i> (European dodder)	5	5	13	7	7	10
<i>Dendrobium nobile</i> (Orchid)	0	0	0	1	1	1
<i>Escallonia rubra</i>	0	0	0	1	1	1
<i>Glycine max</i> (Soybean) (<i>Glycine hispida</i>)	6	5	7	3	4	5
<i>Glycine soja</i> (Wild soybean)	6	3	6	7	2	10
<i>Gossypium anomalum</i>	0	3	4	4	4	4
<i>Handroanthus impetiginosus</i>	0	0	0	1	0	1
<i>Helianthus annuus</i> (Common sunflower)	4	8	22	20	16	18
<i>Hibiscus trionum</i> (Flower of an hour)	1	1	2	2	0	1
<i>Lupinus albus</i> (White lupine) (<i>Lupinus termis</i>)	0	0	0	1	1	1
<i>Lupinus angustifolius</i> (Narrow-leaved blue lupine)	0	1	0	1	1	1
<i>Medicago truncatula</i> (Barrel medic) (<i>Medicago tribuloides</i>)	8	2	3	9	12	3
<i>Microthlaspi erraticum</i>	1	0	0	0	1	2
<i>Olea europaea</i> subsp. <i>europaea</i>	1	0	6	5	4	3
<i>Phoenix dactylifera</i> (Date palm)	2	0	3	0	1	2
<i>Populus alba</i> x <i>Populus x berolinensis</i>	2	0	3	2	3	5
<i>Quillaja saponaria</i> (Soap bark tree)	1	0	0	0	0	1
<i>Rhamnella rubrinervis</i>	0	1	1	1	2	1
<i>Rhododendron simsii</i> (Sims's rhododendron)	9	9	19	18	23	18
<i>Rhododendron williamsianum</i>	0	0	1	1	0	0
<i>Rosa chinensis</i> (China rose)	0	0	0	0	0	1
<i>Salix dunnii</i>	0	0	0	0	0	1
<i>Salvia splendens</i> (Scarlet sage)	0	0	0	0	0	1
<i>Sesamum alatum</i>	0	0	0	1	0	0
<i>Solanum commersonii</i> (Commerson's wild potato) (<i>Commerson's nightshade</i>)	0	2	0	0	0	2
<i>Sphenostylis stenocarpa</i>	0	0	1	0	1	1
<i>Tagetes erecta</i> (African marigold)	1	0	0	1	0	0
<i>Tetracentron sinense</i> (Spur-leaf)	0	0	0	1	0	1
<i>Trifolium medium</i>	1	0	1	0	2	0
<i>Trifolium pratense</i> (Red clover)	1	6	8	4	5	4
<i>Trifolium subterraneum</i> (Subterranean clover)	1	1	2	0	1	1
<i>Vitis rotundifolia</i> (Muscadine grape)	0	1	0	1	0	0
Total	2318	2934	5162	5206	5243	5108

b. The evaluation of the gene models could be complicated, but in addition to the BUSCO analysis, a good way to check for the gene models quality is to compare them with the transcriptomic data and report the number of gene models supported by the transcriptomic/protein sequences (from a closed related species) data (e.g., through the distribution of the different codes supplied by Gffcompare, or with the AED from MAKER). Single exon gene rate as well as number of overlapping genes are also good metrics (AGAT toolkit is a good tool to calculate those). It also describe how many gene models have a GO Term annotation or a NCBI NR/TREMBL/SwissProt blast/diamond hit.

Response: Thank you for your feedback and suggestions. Based on your recommendations, we used Gffcompare (v0.12.6) to compare the annotation results of full-length transcripts obtained from ONT sequencing with those from EVM. The results evaluated sensitivity and precision across six feature levels, as summarized in **Table 4**. The analysis revealed an average sensitivity of 94.65% and precision of 65.02% at the base level, 69.12% sensitivity and 63.73% precision at the exon level,

and 92.40% sensitivity and 85.38% precision at the intron level. At the same time, we used the MAKER software to perform statistical analysis of the AED values for the annotated genes, and found that, on average, 96.58% of the genes had an AED < 0.5 (Tbale 5).

Table 4 Comparison of gene models between Iso-Seq transcripts and EVM annotation.

	Y26.iso_vs_Y26.evm (ref)		Y4.iso_vs_Y4.evm (ref)		S245.iso_vs_S245.evm (ref)		HN873.iso_vs_HN873.evm (ref)		HN51.iso_vs_HN51.evm (ref)		S83.iso_vs_S83.evm (ref)	
	Sensitivity	Precision	Sensitivity	Precision	Sensitivity	Precision	Sensitivity	Precision	Sensitivity	Precision	Sensitivity	Precision
Base level (%):	91.5	68	95.5	63.7	96.1	64.8	94.3	64.6	94.5	64.2	96	64.8
Exon level (%):	67.6	65	70	62.9	70.2	64.8	68.4	62.6	68.7	62.6	70.1	64.5
Intron level (%):	89.7	87.4	93.2	84	93.7	86.3	91.9	84.3	92.1	84.1	93.8	86.2
Intron chain level (%):	58.6	42	57.1	40.7	60.4	47.2	59.5	45.5	59.4	44.4	61.4	48.1
Transcript level (%):	55.1	42	52.1	40.6	55.5	46.9	55.5	45.2	55.2	44.2	56.9	47.9
Locus level (%):	55.1	54.7	52.1	51.6	55.5	55.3	55.5	53.1	55.2	53	56.9	56.1

Table5 Accuracy assessment of genome annotation by AED values.

	Y26	Y4	S245	HN873	HN51	S83
AED<=0.5	33,009 (95.90%)	34718 (96.95%)	70002 (96.67%)	72302 (96.22%)	71545 (96.54%)	70209 (97.18%)
0.5<AED<1	697	473	1146	1,352	1,197	1,149
AED=1	714	619	1,262	1,489	1,369	889
Total gene numbers	34,420	35,810	72,410	75,143	74,111	72,247

In accordance with your suggestions, we utilized AGAT (v1.4.0) to perform a statistical analysis of the single exon gene rate and the number of overlapping genes in several published species (Table 6). The findings indicated an average single exon gene rate of 24.01% without overlapping genes, consistent with results reported for other species (Table 6).

Table 6 Single exon gene rate and number of overlapping genes in representative plant species.

Species	Number of gene	Number of mrna	Number gene overlapping	Number of single exon gene	Number of single exon mrna	% of single exon gene	Reference
<i>Arachis duranensis</i> (Y26)	34420	34420	0	8439	8439	24.52%	This Study
<i>Arachis ipaensis</i> (Y4)	35810	35810	0	8878	8878	24.79%	This Study
<i>Arachis hypogaea</i> (S245)	72410	72410	0	18281	18281	25.25%	This Study
<i>Arachis hypogaea</i> (HN873)	75143	75143	0	17100	17100	22.76%	This Study
<i>Arachis hypogaea</i> (HN51)	74111	74111	0	16697	16697	22.53%	This Study
<i>Arachis hypogaea</i> (S83)	72247	72247	0	17476	17476	24.19%	This Study
<i>Actinidia chinensis</i> (Donghong)	40977	40977	0	7622	7622	18.60%	Han et al. 2023; <i>Molecular Plant</i>
<i>Actinidia latifolia</i> (Kuoye)	41317	41317	0	8375	8375	20.27%	
<i>Aegilops tauschii</i> (DD)	44513	44513	193	11878	11878	26.68%	Zhou et al. 2021; <i>Nature Plants</i>
<i>Annona cherimola</i> (Fino de Jete)	41413	44597	84	8390	8625	20.26%	Talavera et al. 2023; <i>Plants People Planet</i>
<i>Arabidopsis thaliana</i> (TAIR11)	27655	48456	3559	5382	6336	19.46%	Cheng et al. 2017; <i>The Plant Journal</i>
<i>Capsicum annuum</i> (G1-36576)	34428	34428	475	6801	6801	19.75%	Chen et al.2024; <i>Nature Communication</i>
<i>Capsicum rhomboideum</i> (PI645680)	33512	33512	1079	5233	5233	15.62%	
<i>Carya illinoensis</i> (Lakota)	31911	68161	827	5820	7257	18.24%	Lovell et al. 2021; <i>Nature Communications</i>
<i>Citrullus lanatus</i> (PI 673135)	26952	26952	0	7897	7897	29.30%	
<i>Citrullus lanatus</i> (PI 670011)	24205	24205	0	6078	6078	25.11%	Zhang et al.2024; <i>Nature Genetics</i>
<i>Citrullus lanatus</i> (PI 673135)	23964	23964	0	5490	5490	22.91%	
<i>Glycine max</i> (Wm82-NJAU)	55497	88847	1545	11100	12282	20.00%	Wang et al. 2023; <i>Molecular Plants</i>
<i>Gossypium raimondii</i> (D5-3)	53167	53167	208	14221	14221	26.75%	Huang et al.2024; <i>Nature Genetics</i>
<i>Lilium davidii</i> (Unicolor)	87501	87501	0	16329	16329	18.66%	Xu et al. 2024; <i>The Innovation</i>
<i>Lupinus albus</i> (AMIGA)	38258	38258	0	8457	8457	22.11%	Hufnagel et al. 2020; <i>Nature Communications</i>
<i>Oryza sativa</i> (ZS97)	60935	88091	3195	14020	14346	23.01%	
<i>Oryza sativa</i> (MH63)	59903	89040	3771	13814	14071	23.06%	Song et al. 2021; <i>Molecular Plant</i>
<i>Oryza sativa</i> (Nipponbare)	57359	67790	876	13570	14087	23.66%	
<i>Setaria italica</i> (C1)	47421	59916	3663	10700	16644	22.56%	Shang et al.2023; <i>Molecular Plant</i>
<i>Setaria italica</i> (C16)	45623	56305	3449	10530	15922	23.08%	
<i>Setaria italica</i> (C22)	45137	55481	3460	10357	15685	22.95%	He et al.2023; <i>Nature Genetics</i>
<i>Zea mays</i> (Mo17)	42580	45358	199	9011	9019	21.16%	

On average, 98.75% of the identified genes were functionally annotated as orthologs using data from the NR (non-redundant, NCBI), Pfam, InterPro, and Swiss-Prot databases. Additionally, 73.22%

of the genes have GO annotations, whereas 50.87% have KEGG annotations (**Table 2 (Supplemental Table 7)**).

Thank you again for your valuable suggestions. Your insights have been instrumental in assuring the quality of the genome annotation and ensuring the reliability of subsequent analysis results.

G. References: appropriate credit to previous work?

Yes, in the introduction, M&M and results. The discussion lacks references.

Response: Thank you very much for your suggestions. In the revised text, we have updated the discussion section based on your recommendations and incorporated additional references.

H. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions

The manuscript is clear and easy to read, although the discussion should be improved. The abstract summarizes well the results presented, as well as the introduction and the conclusions.

References

1. Bertoli, D. J. *et al.* The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nat Genet* **48**, 438–446 (2016).
2. Bertoli, D. J. *et al.* The genome sequence of segmental allotetraploid peanut *Arachis hypogaea*. *Nat Genet* **51**, 877–884 (2019).
3. Yin, D. *et al.* Comparison of *Arachis monticola* with Diploid and Cultivated Tetraploid Genomes Reveals Asymmetric Subgenome Evolution and Improvement of Peanut. *Advanced Science* **7**, (2020).
4. Zhuang, W. *et al.* The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. *Nat Genet* **51**, 865–876 (2019).
5. Nei, M. & Gojoborit, T. *Simple Methods for Estimating the Numbers of Synonymous and Nonsynonymous Nucleotide Substitutions*. <https://academic.oup.com/mbe/article/3/5/418/988012>.
6. Chen, X. *et al.* Sequencing of Cultivated Peanut, *Arachis hypogaea*, Yields Insights into Genome Evolution and Oil Improvement. *Mol Plant* **12**, 920–934 (2019).
7. Bertoli, D. J. *et al.* The repetitive component of the A genome of peanut (*Arachis hypogaea*) and its role in remodelling intergenic sequence space since its evolutionary divergence from the B genome. *Ann Bot* **112**, 545–559 (2013).
8. Jiao, Y. *et al.* Genome-wide genetic changes during modern breeding of maize. *Nat Genet* **44**, 812–815 (2012).
9. Goel, M., Sun, H., Jiao, W. B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol* **20**, (2019).
10. Chakraborty, M., Emerson, J. J., Macdonald, S. J. & Long, A. D. Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nat Commun* **10**, (2019).
11. Tang, H. *et al.* JCVI: A versatile toolkit for comparative genomics analysis. *iMeta* **3**, (2024).
12. Zhang, Z. KaKs_Calculator 3.0: Calculating Selective Pressure on Coding and Non-coding Sequences. *Genomics Proteomics Bioinformatics* **20**, 536–540 (2022).
13. Excoffier, L. *et al.* Fastsimcoal2: Demographic inference under complex evolutionary scenarios. *Bioinformatics* **37**, 4882–4885 (2021).
14. Hudson, K. A. & Hudson, M. E. Genetic Variation for Seed Oil Biosynthesis in Soybean. *Plant Molecular Biology Reporter* vol. 39 700–709 Preprint at <https://doi.org/10.1007/s11105-020-01276-1> (2021).
15. Wang, J. *et al.* Global Transcriptome Analyses Provide Into Several Fatty Acid Biosynthesis-related Genes in Peanut (*Arachis hypogaea* L.). *Trop Plant Biol* **14**, 267–282 (2021).
16. Lu, Q. *et al.* A genomic variation map provides insights into peanut diversity in China and associations with 28 agronomic traits. *Nat Genet* **56**, 530–540 (2024).
17. Xu, W., Dubos, C. & Lepiniec, L. Transcriptional control of flavonoid biosynthesis by MYB-bHLH-WDR complexes. *Trends in Plant Science* vol. 20 176–185 Preprint at <https://doi.org/10.1016/j.tplants.2014.12.001> (2015).
18. Zhao, Y. *et al.* Whole-genome resequencing-based QTL-seq identified AhTc1 gene encoding a R2R3-MYB transcription factor controlling peanut purple testa colour. *Plant Biotechnol J* **18**, 96–105 (2020).
19. Chen, H. *et al.* Fine-mapping and gene candidate analysis for AhRt1, a major dominant locus responsible for testa color in cultivated peanut. *Theoretical and Applied Genetics* **134**, 3721–3730 (2021).

Reviewers' Comments:

Reviewer #1 (Remarks to the Author):

After enormous revisions, many of which entailed significant new bioinformatic research, this manuscript is now an excellent example of a paper with the interest level and attention to detail requisite for a journal such as Nature Genetics. I am satisfied with how the investigators followed up on my suggestions, and am also particularly impressed at how independent approaches (fastsimcoal being one of them) yielded similar dates. This consistency speaks to both cares taken to analyses as well as, of course, the appropriateness of the data to address the questions at hand. My only final suggestion - which I think is very important these days - is to revise all main-text figures such that they use similar (and best, color-blind compatible) color palettes. For example, Figs. 1 and 2 stand out very starkly in "standard" red, blue, and (bright) green colors, while Figs. 3 and 4, e.g., have much more muted colors that are easier to view.

Response:

We appreciate your summary of manuscript and encouraging comment.

Regarding your final suggestion about figure color palettes, we fully agree with your point about the importance of using color-blind friendly palettes. In the current revised manuscript, we have ensured that all figures adopt consistent and more accessible color schemes, with careful attention to color-blind compatibility. We have updated Fig. 1 and Fig. 2 to replace the bright red, blue, and green with more muted, color-blind friendly alternatives, in line with the colors used in Figs. 3 and 4.

We sincerely thanks for your thoughtful feedback, which has significantly improved the presentation of our work.

Reviewer #2 (Remarks to the Author):

The following conclusions of the abstract aren't supported:

Lines 42-43

"a 13.23 Mb-long insertion on Chr14 of all the var. hirsute accessions may have contributed to the prostrate growth and narrow leaves of this group"

The data in the paper shows correlation - not causation

Response:

We greatly appreciate your thoughtful comments.

In the synteny analysis of the B genome, we identified a 13.23 Mb-long insertion on Chr14, and visualized its coverage using both HiFi and Illumina reads. The accuracy of this structural variation (SV) was further confirmed through Hi-C heatmap analysis. Statistical analysis revealed that all samples in the G5 population contained this SV, while 23% and 15% of samples in the G4 and G3 populations contained it, respectively. No samples in the G2 population contained this SV (**Supplementary Figures 10A-10E**). These results suggested that this SV could be associated with the differentiation between the G5 and G4 populations, which both belong to the *var. hypogaea* subspecies.

In this SV region, we identified 53 homologous genes with over 70% sequence identity and an e-value of less than $1e-10$, including several rice functional genes related to plant architecture and organ development, such as *OsA7*, *OsWOX3A*, *OsWOX3B*, *OsIAA17*, *OsLAC*, and *OsRLCK107*. Among these, two genes, chr14.1804 and chr14.1807, were found to be tissue-specific, particularly expressed in roots, stems, and leaves but not in cotyledons, shells, seeds, or pegs (**Supplementary Fig. 12**). These two genes are homologous to *OsLAC*, which is a target gene of OsMiR397. Studies suggest that *OsLAC* regulates various developmental processes in rice by modulating the brassinosteroid (BR) signaling pathway. Overexpression of *OsLAC* leads to shortened stems, upright leaves, and reduced grain size, phenotypes resembling those seen in BR-deficient plants. Moreover, OsMiR397, which negatively regulates *OsLAC* expression, enhances BR signaling and alters plant architecture by increasing leaf angles. These findings indicate that *OsLAC*, through its interaction with the BR pathway, plays a vital role in regulating plant growth and development, offering potential strategies for improving crop architecture and yield through the manipulation of hormone signaling pathways¹.

Another gene, chr14.1829, which is homologous to the *WOX3A/B* gene. Studies indicate that both *OsWOX3A* and *OsWOX3B* genes play crucial roles in the organ development of rice, especially in the development of leaves, tillers, lateral roots, and lateral organs. In the case of the *OsWOX3A* mutant, the plant exhibits narrow and curled leaves, an increase in tiller number, a decrease in lateral roots, and narrower grains. This phenotype is associated with restricted lateral growth of the leaves, fewer longitudinal veins, and the enlargement of bubble-like cells. Moreover, *OsWOX3A* also plays a key role in the development of the outer glume (lemma), with mutations leading to abnormal glume morphology, affecting the seed shape². At the molecular level, *OsWOX3A* is primarily

expressed in the vascular tissues of rice, and mutations lead to changes in the expression of genes related to leaf development, hormone transport, and other pathways, resembling the phenotypes seen in PIN gene mutants.

On the other hand, *OsWOX3B* plays an essential role in the symmetry of leaf development. The *OsWOX3B* mutant displays asymmetrical leaves and sometimes forked leaf shapes, which leads to abnormal development of lateral organs, affecting leaf width and morphology. Mutants also show an increase in tillers, longer panicles, and a greater number of grains per panicle. Furthermore, the vascular bundles in the flag leaf (the leaf above the panicle) increase in number, and the spacing between small vascular bundles widens, further affecting the growth and development of rice³.

Otyama et al. (2020) classified 787 accessions from the U.S. Peanut core collection and 12 commercial varieties into Clades 1 to 4 using the *Arachis*_Axiom2 SNP array data⁴. Phenotypic observations showed that in Clade 1, "most accessions are classified as the *hypogaea* botanical variety and exhibit a 'mixed' or *hypogaea* pod shape (60.0%, 44.4%, 40.0%). The growth type varies widely, with an even distribution between erect, bunch, spreading-bunch, mixed, and prostrate types."

Further analysis reveals that Clade 1 in our study corresponds to the G4+G5 populations. Based on phylogenetic relationships and material information, we found that both G4 and G5 belong to *var. hypogaea*. Field surveys indicate that the G5 population has a larger tillering angle (with a higher proportion of prostrate types), whereas the G4 population has a larger proportion of erect types. This is consistent with the observed variation in growth types within Clade 1.

Combining the functional annotation and expression results of the 109 SV-genes, we hypothesize that the growth type difference between the G4 and G5 populations may be linked to the presence of this structural variant (SV).

Thank you for your valuable suggestions. We totally agree with your comment that the presence of this SV just showed a correlation with architectural differences. We have revised our manuscript accordingly to reflect this clarification in **lines 41-43**.

Lines 43-44

"Selective sweep and introgression analyses highlighted multiple asymmetric selections between the two subgenomes during breeding improvement."

There is no data to show this. Differences between the subgenomes were almost all accumulated in the diploid state, before the origin of *Arachis hypogaea*.

Response:

We sincerely appreciate your valuable suggestion.

In our selective sweep and introgression analyses, we observed notable differences within and between the subspecies of *Arachis* in terms of selection and introgression across the A and B subgenomes (**Table 1; Supplementary Tables 21 & 23**). Specifically, our statistical analysis

revealed that the B subgenome contained 379.37 Mb of selective sweep regions, significantly larger than the 66.09 Mb in the A subgenome. Further comparison of variant groups showed substantial differences: the G2 and G3 groups identified 522 intervals (19.91 Mb) in the A subgenome, while the B subgenome had 2,019 intervals (64.66 Mb). In contrast, the G4 and G5 groups detected only 20 intervals (560 kb) in the A subgenome, compared to 2,368 intervals (46.96 Mb) in the B subgenome.

Additional analysis revealed that the G2 and G3 groups exhibited selective regions across all 20 chromosomes, while the G4 and G5 groups identified selective regions only on chr04 in the A subgenome and on chr11, chr12, chr14, chr15, chr17, and chr18 in the B subgenome (**Table 1; Supplementary Tables 21 & 23**). Similar patterns were found in other groups, further supporting the significant selection differences between the two subgenomes.

We also observed notable variations in the counts, sizes, and chromosomal distribution of introgression intervals between the two subgenomes, further highlighting the complexity of genomic interactions during domestication and breeding (**Table 1; Supplementary Tables 21 & 23**).

Table 1. Total Length of Selective Sweep and Introgression Regions in the A and B Subgenomes Across Different Groups.

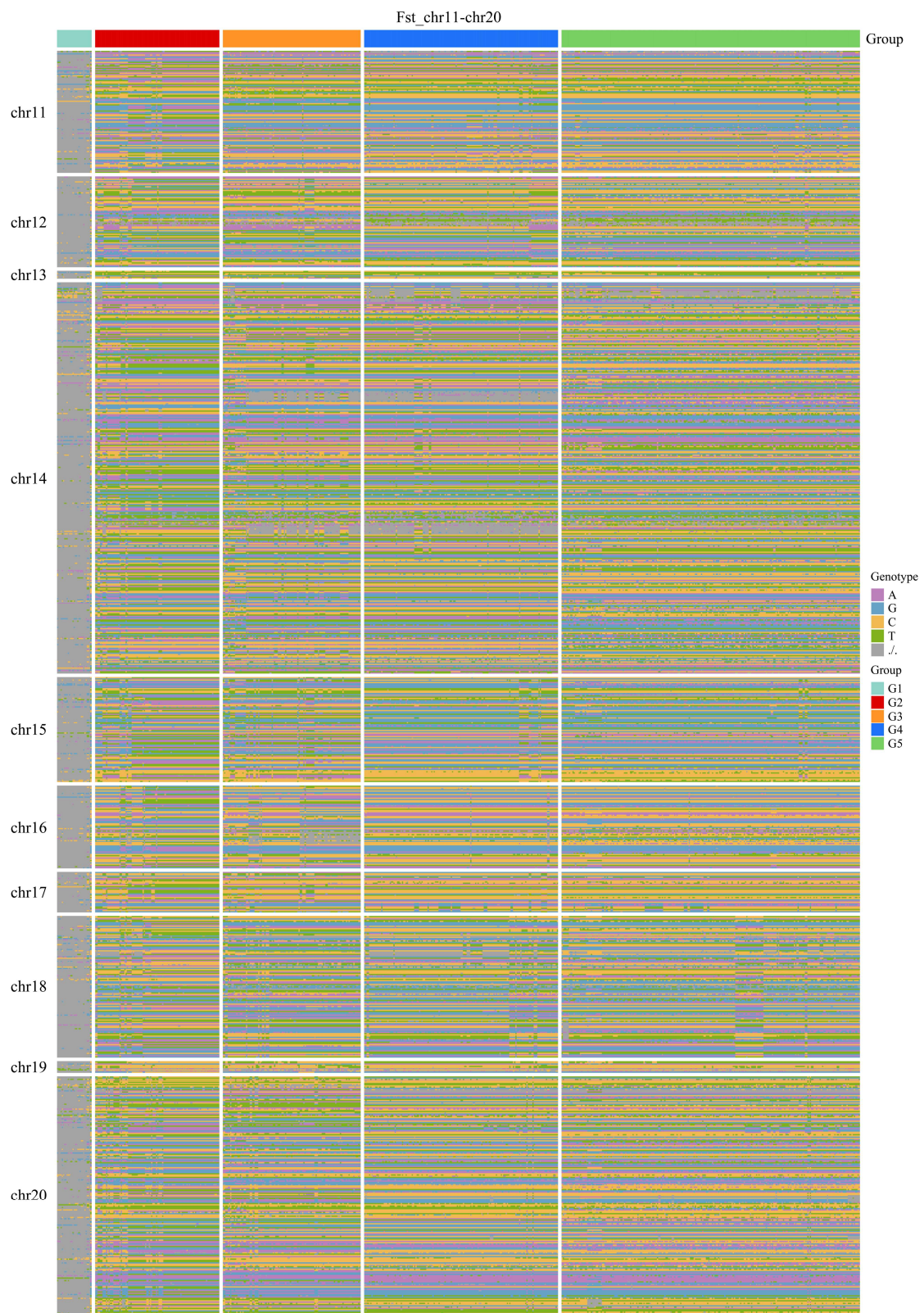
	G4	G3	G2	G5	G4	G2	G4	G5	G2	G5	G4	G3	G4	G5	G3	G3	G5	G4
subA_ <i>fst</i> (bp)	19,910,000			11,140,000			6,960,000			15,570,000			11,950,000			560,000		
subB_ <i>fst</i> (bp)	64,660,000			70,940,000			79,000,000			53,780,000			64,030,000			46,960,000		
subA_ introgression (bp)	41,200,000			65,560,000			6,610,000			64,680,000			1,990,000			6,890,000		
subB_ introgression (bp)	43,060,000			10,170,000			790,000			20,160,000			340,000			35,640,000		

Further, we visually represent the variation types of SNPs in the top1% of selected and introgressed regions across different populations (**Figure 1 (Supplementary Figure 17)**). And we performed a statistical analysis of the sources of variation in the selected and introgressed regions, using the 23 wild accessions as references. Any genotypes present in these 23 wild species that also appeared in the tetraploid accessions were defined as shared variation between the diploid and tetraploid genomes, while those unique to the tetraploid accessions were classified as tetraploid-specific variation. The results are shown in **Figure 2 (Supplementary Figure 18)**.

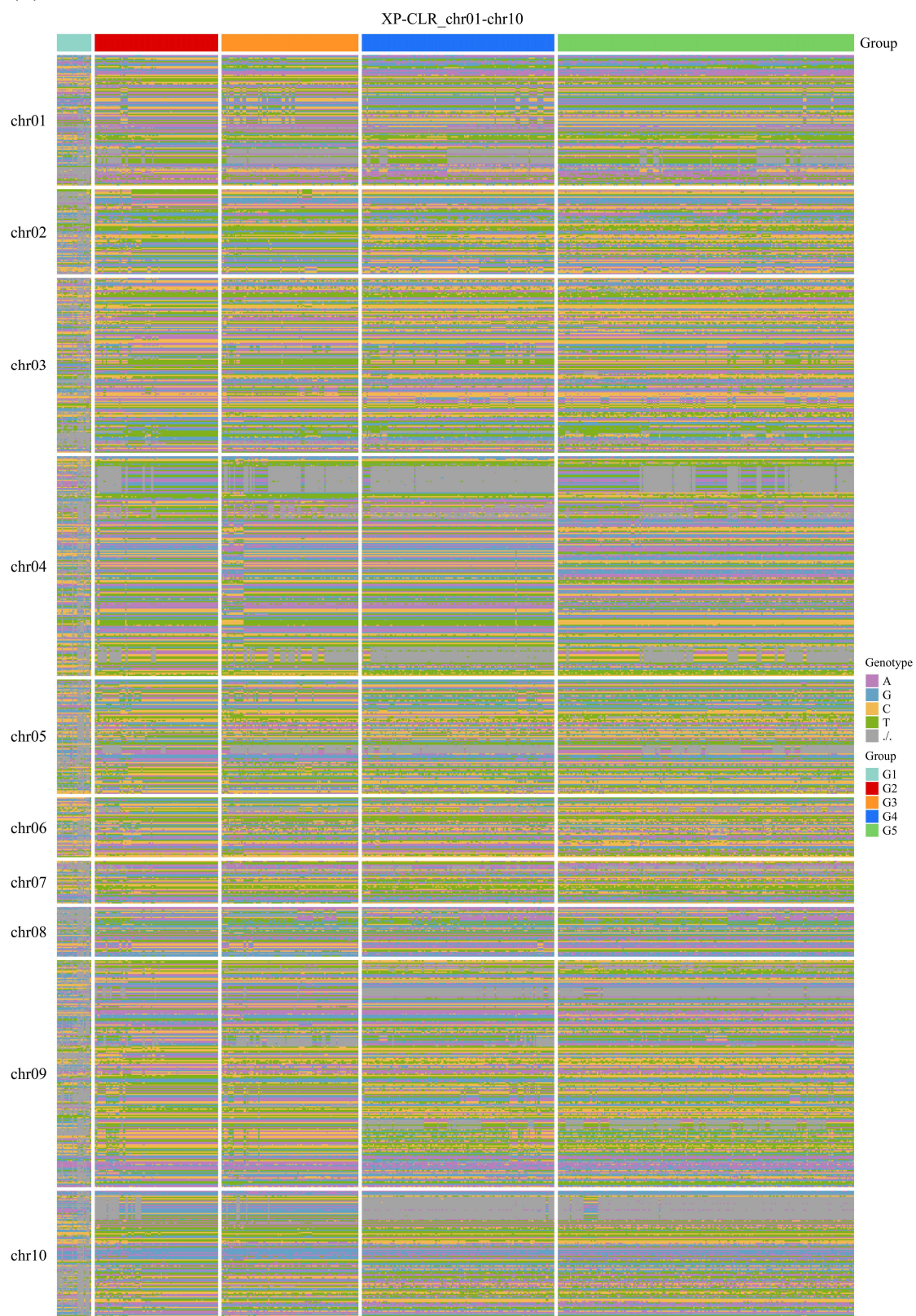
The results show that the variation in tetraploid populations not only originates from the diploid ancestors but also exhibits population-specific variation during the domestication and improvement of tetraploid crops. We greatly appreciate yours’s suggestions, and in the revised version, we have expanded and clarified this part of the content in **lines 372-374**. This provides direct data support for understanding the imprint of tetraploid selection and introgression in peanut population improvement.

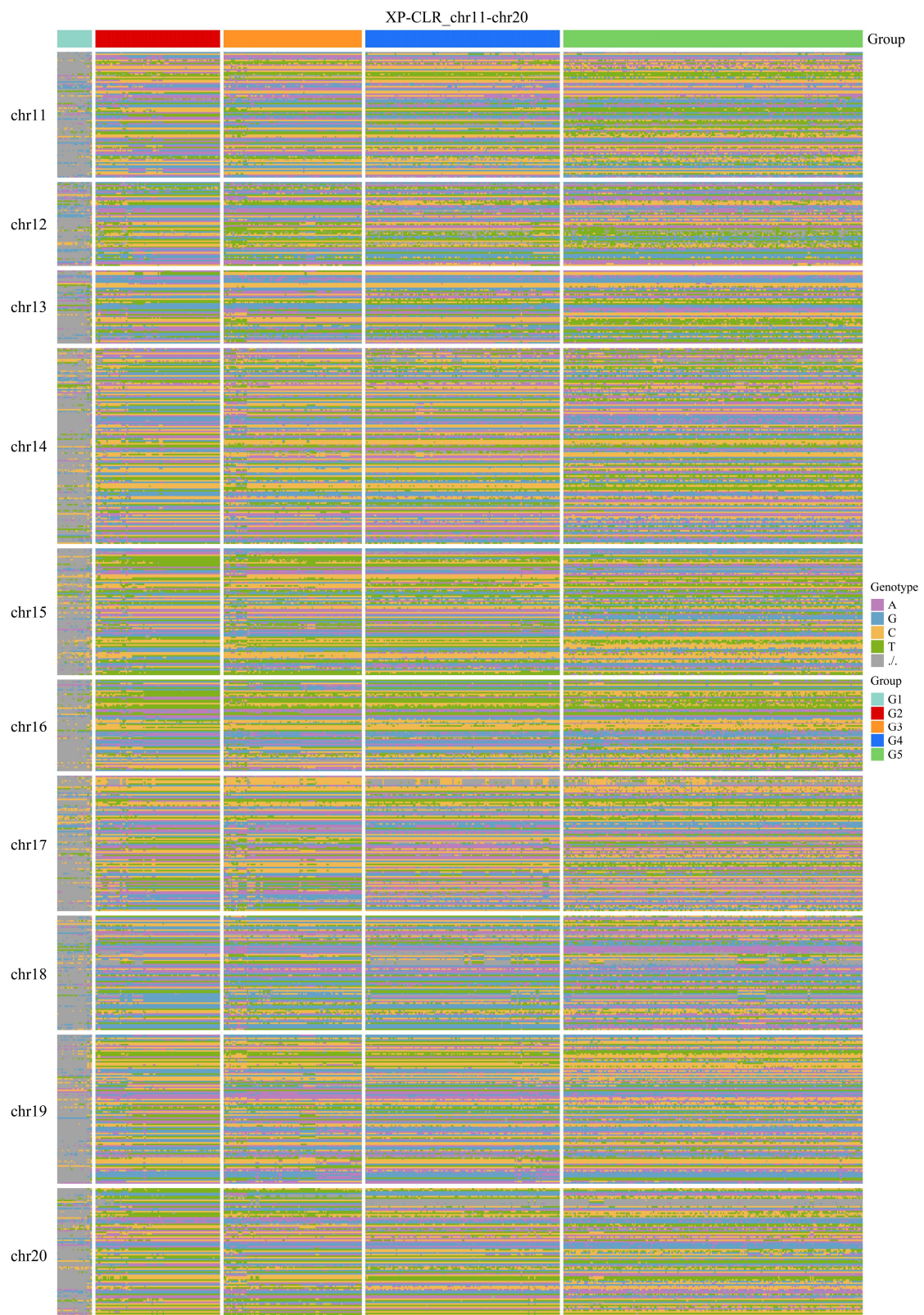
(A)



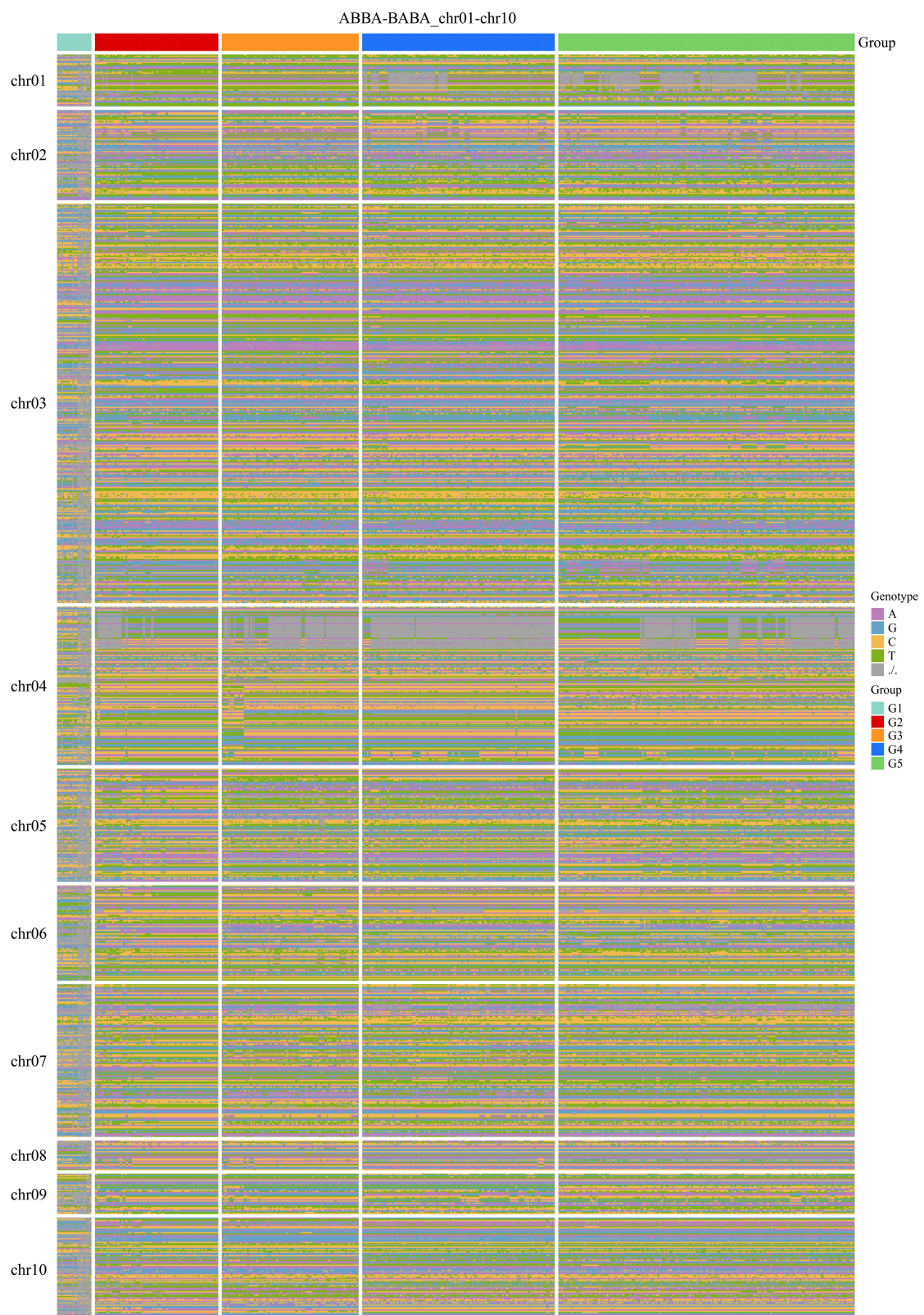


(B)





(C)



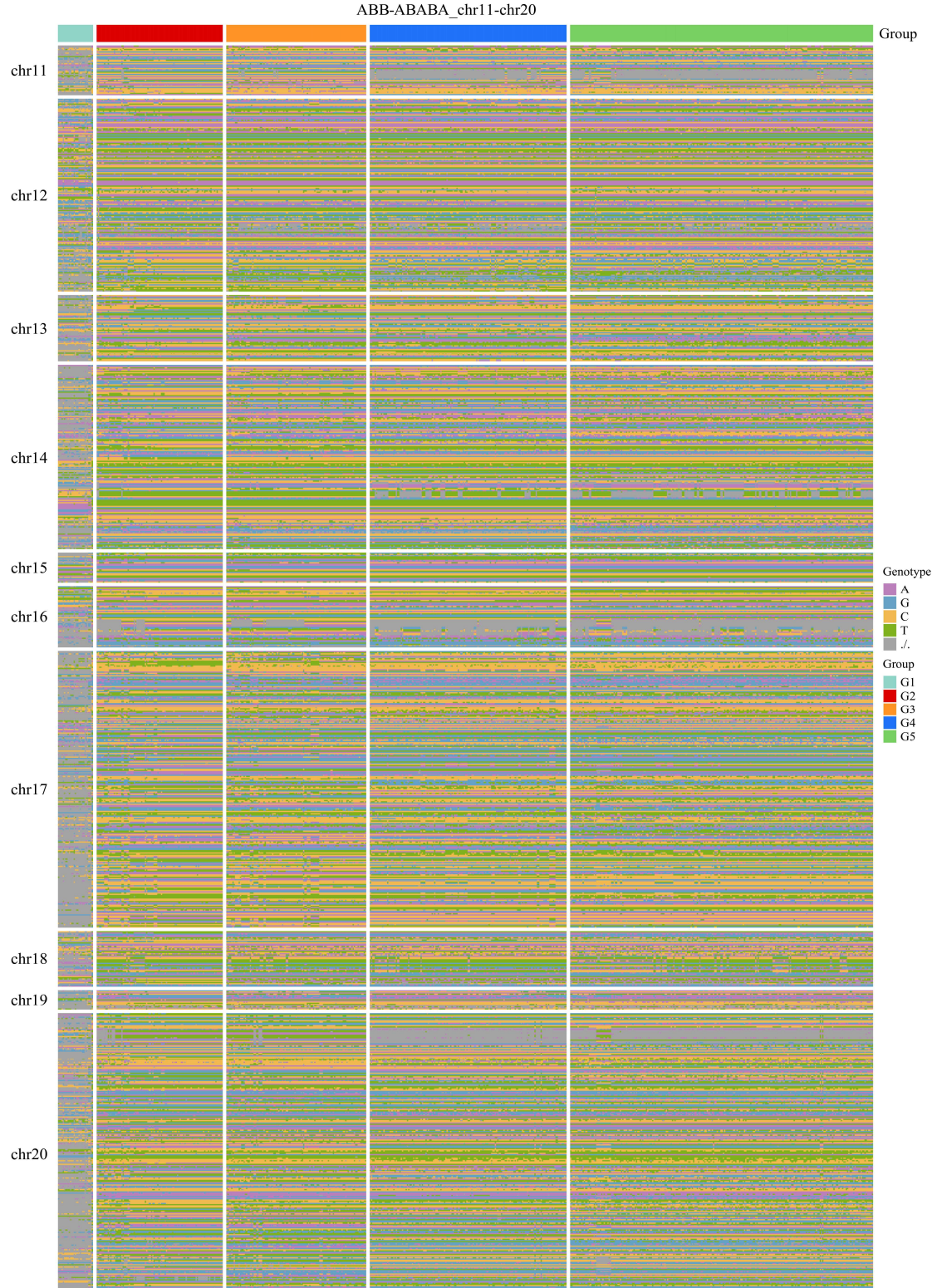


Figure 1. SNP Variation Types in Selected and Introgressed Regions. The distribution of SNP variants (A, G, C, T, and .) across Group1 to Group5 is visualized with different colors. The S245 reference genome is used for comparison. (A) Top 1% Fst regions. (B) Top 1% XP-CLR regions. (C) Top 1% Introgressed regions.

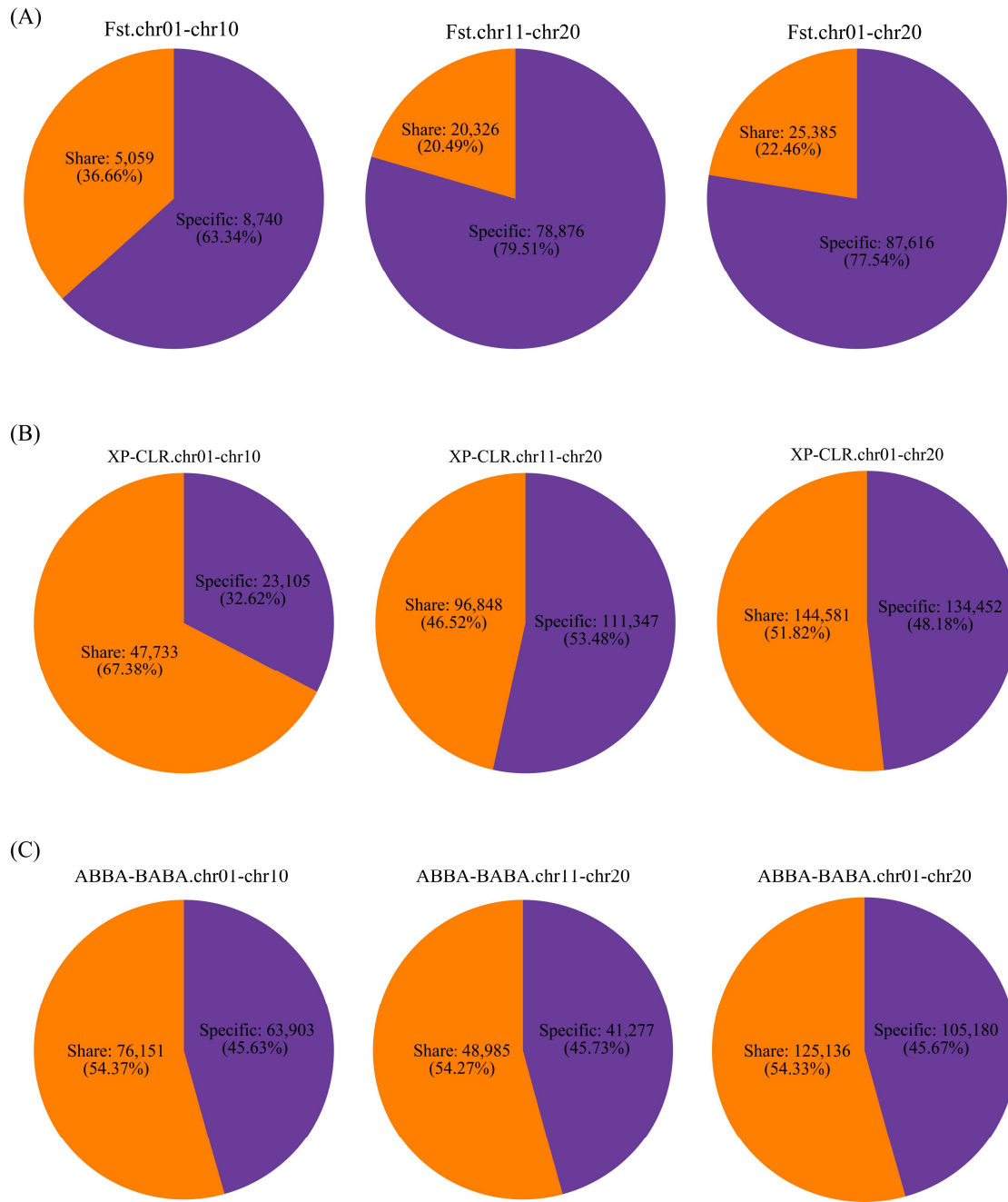


Figure 2. Percentage of share and specific variation between diploids and tetraploids. The variation type in tetraploids is defined as shared variation if it exists in the wild accessions; otherwise, it is considered tetraploid-specific variation. (A) Top 1% Fst regions. (B) Top 1% XP-CLR regions. (C) Top 1% introgressed regions. From left to right, represent the At (chr01-chr10), Bt (chr11-chr20), and AtBt (chr01-chr20) genomes.

Therefore, we infer that, following polyploidization, these distinct variants continued to experience differential selection pressures in the A and B subgenomes. These selective forces, along with introgressive exchanges, have played a significant role in breeding improvements.

Lines 45-46

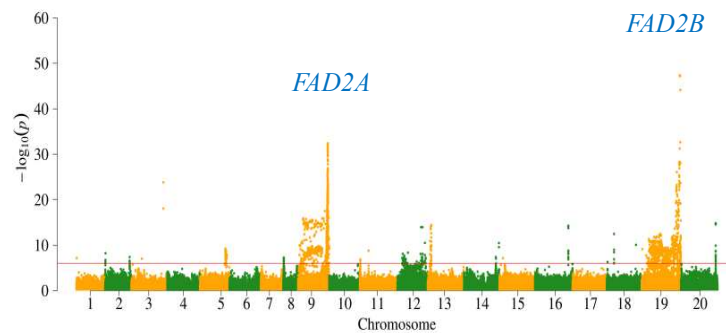
"Genome-wide association studies identified 4,638 loci associated with 37 agronomic traits, including candidate genes for oil content, seed size, seed weight, etc."

Almost all of these loci lack any sort of verification.

Response:

Thank you very much for your valuable suggestions. In our genome-wide association study (GWAS) on 37 agronomic traits, we also included traits related to oleic and linoleic acid content as controls. Significant associations were identified on chromosomes 9 and 19 (**Figure 3**). Further gene analysis revealed candidate genes that align with the previously reported *FAD2A* and *FAD2B* genes^{5,6}, and our haplotype analysis supports these findings (the details of the correlation analysis are not included in the article).

(A)



(B)

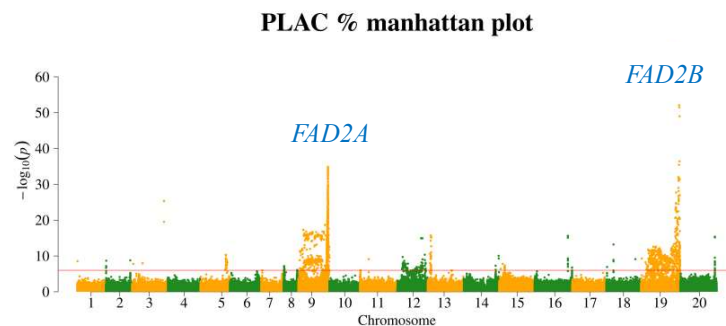


Figure 3. Manhattan plot of GWAS results for oleic acid and linoleic acid. (A) Oleic acid, (B) Linoleic acid. The causation gene *FAD2A* is located at the chr09, while *FAD2B* is located at the chr19.

Additionally, during the GWAS analysis of thousand-grain weight, we identified a strong signal on chromosome 16, and significant peak signals were also observed on chromosome 7. Further analysis revealed that the major-effect QTL, *qAHPS07*, which controls seed size, and the key gene *AhRUVBL2* are located within this region⁷ (**Figure 4**). This result also provides support for the accuracy of our GWAS findings.

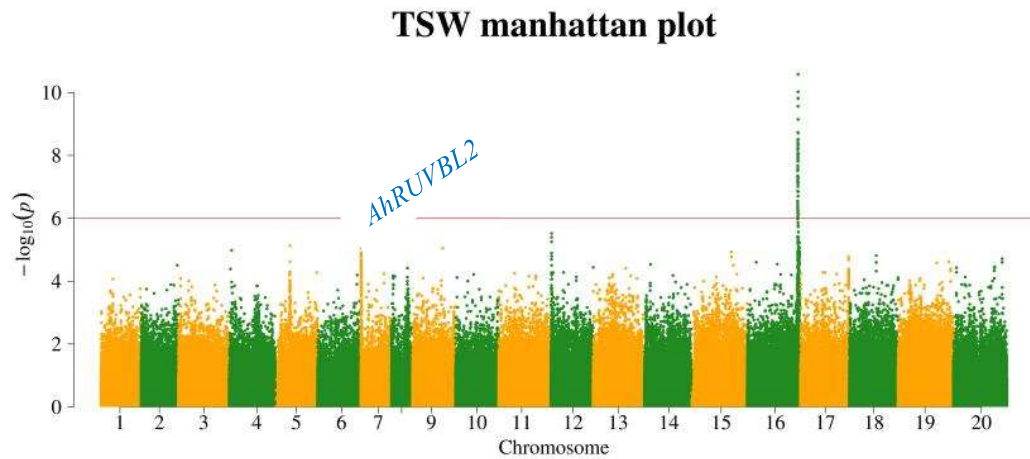


Figure 4. Manhattan plot of GWAS results for Thousand Seed Weight (g). The signal of the Pod size-controlling gene *AhRUVBL2* on chr12.

Although some of the identified loci have not yet been experimentally verified, the GWAS approach is supported by the identification of well-established candidate genes with known roles. Given the importance of oil content, seed size, and seed weight in peanut breeding, we have prioritized these representative traits for candidate gene mining, haplotype analysis, and experimental validation.

Once again, thank you for your constructive feedback.

Lines 50-51

"These findings provide a comprehensive understanding of the origin, evolution and domestication"

It's handwaving - there are few new insights. From this paper, I've learned nothing new of the origin, a bit about the evolution of the tetraploid genome (mainly about some rearrangements), and nothing at all about domestication.

Response:

Thank you for the comments.

#####Origin

While the origin of peanuts has been the subject of extensive study, our research aims to provide a more nuanced, comparative analysis using T2T genome sequencing on key representative materials. This allows us to contribute additional data, offering a deeper and more comprehensive understanding of peanut origins, building on previous findings.

By leveraging SNP variations within the T2T genome, we refine the timeline of tetraploid formation to approximately 10,000 years ago, consistent with earlier studies⁸.

Furthermore, our centromere length analysis offers new insights into structural differences between subgenomes. Specifically, we find that the centromere length of the At subgenome in the tetraploid (46.11 Mb) is comparable to that of the diploid (44.70 Mb), whereas the Bt subgenome (25.70 Mb)

is significantly smaller than the BB subgenome (43.30 Mb). This difference primarily results from a reduction in the number of repetitive units in the Bt subgenome. Given that the diploid BB genome is larger than the diploid AA genome, the substantial change in centromere length may be a key factor in ensuring proper homologous pairing during tetraploid formation.

Additionally, our investigation of the centromeric regions has provided new insights into their phylogenetic relationships. The analysis suggests that a diploid B species may have diverged from a diploid A species before hybridizing with an A genome to form the tetraploid. This refined understanding of the centromeric regions enhances our knowledge of the origin of peanuts

#####Evolution

In our study of genome evolution, we focused not only on genome rearrangements but also on a detailed investigation of LTR (Long Terminal Repeat) retrotransposons. This included examining the various types of LTR insertion events, their timing, and the differences across subgenomes.

We found that *Gypsy*-type LTR retrotransposons showed distinct insertion times between the At and Bt subgenomes. Specifically, the *Althila*, *Retand*, *CRM*, and *Ogre* types underwent two expansions in the BB genome and Bt subgenome, while the *Reina* type expanded twice in the AA genome and At subgenome. These active transposon insertions likely contributed to the increased *Gypsy* content, which in turn played a significant role in the larger size of the B subgenome.

Furthermore, we categorized *Gypsy* and *Copia* transposable elements (TEs) based on their insertion times into four periods: S1 (1-10,000 years ago), S2 (10,001-100,000 years ago), S3 (100,001-300,000 years ago), and S4 (>300,000 years ago). The TE counts for the AA genome were 286, 1417, 5322, and 9582, while the BB genome showed counts of 460, 1426, 3434, and 15892 for the same periods. The major TE insertion events in the A genome occurred during the S3 period, with *Gypsy* TEs being the primary contributors. In both tetraploid subgenomes, *Gypsy* TEs were more abundant in the At subgenome than the Bt subgenome, especially during the S2 and S3 phases. These findings highlight the significant role of *Gypsy* transposons in the genomic evolution of peanut.

Additionally, we performed a comparative analysis of LTR insertions between diploid and polyploid species. This analysis revealed that the insertion times of the two major *Gypsy* transposable element types, *Althila* and *Retand*, were significantly shorter in the AA and At genomes compared to the BB and Bt genomes. This difference may have played a crucial role in the formation of the tetraploid genomes. Moreover, some *Althila* and *Retand* TEs present in the diploid genomes were absent from the tetraploid subgenomes, suggesting complex evolutionary dynamics and potential differential retention or loss of TEs during polyploidization.

#####Domestication

In this study, we explored structural variations (SVs) between diploid and tetraploid T2T peanut genomes, as well as variations within tetraploid variants, to better understand their role in domestication and breeding divergence. Our analysis yielded several important insights (**Table 2; Supplemental Table 13-18**):

The total number of SVs was significantly higher in the comparison of At_vs_V14167 (average: 70,439) than in Bt_vs_K30076 (average: 14,917). However, within the tetraploids, the number of SVs was similar, with averages of 6,694 for At_vs_S83A and 6,496 for Bt_vs_S83B.

Specific SV types revealed notable differences. In At_vs_V14167, the number of SVs was greater across all five types compared to At_vs_S83A. In Bt_vs_K30076 and Bt_vs_S83B, SVs related to transposons (TRANS), copy number variations (CNV), and inversions (INV) were similar, with significant differences seen in deletions (DEL) and insertions (INS).

Comparing At to the diploid V14167 genome, the SV count remained stable (70,188 to 70,754). For Bt_vs_K30076, the S245 variant had the highest SV count (15,653), followed by S83 (15,195), HN873 (14,477), and HN51 (14,342).

In the tetraploid comparison, S245A had the highest number of SVs (7,635), surpassing HN51A (6,494) and HN873A (5,954). In the Bt subgenome, S245B also showed more SVs (8,667) than HN873B (5,476) and HN51B (5,346).

Notably, the S245 botanical variety, classified as *var. fastigiata*, exhibited considerable divergence in both direction and extent compared to other varieties. It is characterized by smaller seeds and a higher seed count (≥ 3), likely a result of artificial domestication.

Table 2. The Number of SVs Between Diploid and Tetraploid Genome Comparisons.

Q VS Ref	TRANS	DEL	CNV_Q	INS	INV	Total
S245A_vs_V1416	2441	34893	1600	31662	158	70754
HN873A_vs_V14167	2381	34980	1521	31158	148	70188
HN51A_vs_V14167	2309	35108	1583	31423	151	70574
S83A_vs_V14167	2288	34944	1519	31340	147	70238
Average	2355	34981	1556	31396	151	70439
S245A_vs_S83A	202	3803	360	3236	34	7635
HN873A_vs_S83A	206	3097	280	2341	30	5954
HN51A_vs_S83A	181	3315	303	2668	27	6494
Average	196	3405	314	2748	30	6694
S245B_vs_K30076	258	8064	659	6645	27	15653
HN873B_vs_K30076	104	7656	546	6142	29	14477
HN51B_vs_K30076	197	7489	506	6121	29	14342
S83B_vs_K30076	150	8056	505	6458	26	15195
Average	177	7816	554	6342	28	14917
S245B_vs_S83B	225	4161	520	3734	27	8667
HN873B_vs_S83B	130	2697	371	2252	26	5476
HN51B_vs_S83B	137	2543	367	2277	22	5346
Average	164	3134	419	2754	25	6496

We also analyzed population structure, selection, and gene flow regions and conducted a Genome-Wide Association Study (GWAS) to explore the roles of selected candidate genes.

The divergence driven by domestication is also reflected in significant haplotype changes in key traits, such as oil content and seed size/weight. For instance, haplotype analysis of the *AhWRII* gene,

which is associated with oil content, revealed that haplotypes linked to low oil content were mainly found in the G2 group, with some occurrence in the G3 group. In contrast, the high oil content haplotype, Hap2, was primarily present in the G3, G4, and G5 groups, suggesting significant genotypic variation in oil content between the G2 group and the others. Similarly, for seed size/weight, haplotypes of the *AhGSA1* gene, associated with smaller 1,000-seed weight, were mostly found in G2 and G3 groups. On the other hand, the haplotype linked to larger seed weight, Hap2, was predominantly observed in the G4 and G5 groups. This indicates that domestication has led to distinct haplotype selection influencing these key agronomic traits

Additionally, by integrating transcriptomic and metabolomic data, we gained deeper insights into the molecular processes influencing peanut seed development and pigmentation.

We believe that this integrated genomic analysis offers a comprehensive understanding of peanut origin, evolution, and domestication, adding valuable insights to prior research.

Thank you once again for your helpful feedback.

More detailed comments

The authors shouldn't use the abbreviations Y26 and Y4 and instead use the full species names along with their original collection numbers (e.g., *Arachis ipaensis* K30076). This practice is crucial for properly acknowledging prior scientific work and enabling the stepwise accumulation of knowledge within the field.

Response:

Thank you for your valuable suggestion. We fully recognize the importance of properly acknowledging prior scientific work and facilitating the stepwise accumulation of knowledge within the field. In the revised manuscript, we have replaced the abbreviations Y26 and Y4 with the full species names V14167 (*A.duranensis*) and K30076 (*A.ipaensis*) throughout the manuscript, including the supplementary tables and figures, in accordance with the materials cited in the 2016 publication⁸. This change ensures clearer references and accurately attributes the materials used in our study.

fastigiata is still incorrectly spelled fastigate in multiple instances in the manuscript

Response:

Thank you for pointing this out. We have noticed the incorrect spelling of *fastigiata* and appreciate your careful review. We have thoroughly checked the manuscript and ensured that all instances of the term are corrected to the proper spelling, *fastigiata*, as you suggested. This correction will help maintain the accuracy and consistency of the manuscript.

Thank you again for bringing this to our attention.

The discovery that *A. ipaensis* is the ancestor of the B subgenome of *A. hypogaea* is wrongly attributed: "Zhuang et al. (2019) compared the sequenced A and B genomes of wild species to

those of the cultivated variety Shitouqi and reported high sequence similarity between *A. ipaensis* and the B genome of Shitouqi (> 99.5%), suggesting that it or a close relative is the peanut Bt subgenome progenitor". This conclusion was not a discovery of Zhuang et al. (2019); rather, it is the culmination of decades of careful study that are conspicuously ignored:

Kochert, G. et al. RFLP and cytogenetic evidence on the origin and evolution of allotetraploid domesticated peanut, *Arachis hypogaea* (Leguminosae). *Am. J. Bot.* 83, 1282–1291 (1996).

Moretzsohn, M.C. et al. A study of the relationships of cultivated peanut (*Arachis hypogaea*) and its most closely related wild species using intron sequences and microsatellite markers. *Ann. Bot.* 111, 113–126 (2013).

Seijo, G. et al. Genomic relationships between the cultivated peanut (*Arachis hypogaea*, Leguminosae) and its close relatives revealed by double GISH. *Am. J. Bot.* 94, 1963–1971 (2007).

Robledo, G. & Seijo, G. Species relationships among the wild B genome of *Arachis* species (section *Arachis*) based on FISH mapping of rDNA loci and heterochromatin detection: a new proposal for genome arrangement. *Theor. Appl. Genet.* 121, 1033–1046 (2010).

Grabiele, M., Chalup, L., Robledo, G. & Seijo, G. Genetic and geographic origin of domesticated peanut as evidenced by 5S rDNA and chloroplast DNA sequences. *Plant Syst. Evol.* 298, 1151–1165 (2012).

Bertioli, D.J. et al. The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nat. Genet.* 48, 438–446 (2016).

References 1–5 firmly establish *A. ipaensis* as the B subgenome donor of peanut. Reference 6 further concluded, based on genomic and biogeographic evidence, that the only known accession of *A. ipaensis* K30076 likely was from the same population that gave rise to peanut less than 9,600 years ago. This is easy to spot because the abstract explicitly states: "On the basis of remarkably high DNA identity of the *A. ipaensis* genome and the B subgenome of cultivated peanut and biogeographic evidence, we conclude that *A. ipaensis* may be a direct descendant of the same population that contributed the B subgenome to cultivated peanut."

It is extraordinary - and troubling - that the paper so badly misrepresents this substantial body of research."

Response:

Thank you for pointing out the misattribution regarding the discovery of *A. ipaensis* as the ancestor of the B subgenome of *A. hypogaea*. We appreciate the opportunity to clarify this matter.

As you correctly noted, the conclusion that *A. ipaensis* is the progenitor of the B subgenome of cultivated peanut is not a new discovery made by **Zhuang et al. (2019)**, but rather the result of decades of comprehensive research. Key studies by **Kochert et al. (1996)**, **Moretzsohn et al. (2013)**, **Seijo et al. (2007)**, **Robledo and Seijo (2010)**, **Grabiele et al. (2012)**, and **Bertioli et al. (2016)** have laid the foundation for this conclusion, providing genomic and cytogenetic evidence that establishes *A. ipaensis* as the B subgenome donor^{9–13}.

In particular, the study by Bertioli et al. (2016) provides genomic and biogeographic evidence supporting the relationship between the *A. ipaensis* accession K30076 and the population from

which the B subgenome of *A. hypogaea* originated⁸. This is explicitly stated in the abstract, where the remarkable DNA identity between the *A. ipaensis* genome and the B subgenome of cultivated peanut is highlighted.

Thank you once again for your valuable suggestion. In the revised version of the manuscript, we have re-described the origin of peanuts and ensured accurate citations of the relevant literature to better reflect prior research in **lines 81-88**.

The preceding lines of the introduction [Lines 68–80] are also deeply flawed. The authors rely on outdated papers and a poorly executed review to create an artificial sense of uncertainty and confusion. They state, "However, the precise genetic origin of *A. hypogaea* remains undetermined." Really? Can the authors provide even a single example of an ancient polyploid crop whose origin is as thoroughly and conclusively established as that of peanut?

Response:

We appreciate your concerns regarding the accuracy and clarity of the origin of *A. hypogaea*. We understand that the statement about the "precise genetic origin of *Arachis hypogaea* (peanut)" may have caused some confusion.

In response to your feedback, we have revised the relevant portion of the introduction and have removed the phrase "precise genetic origin remains undetermined." Additionally, we have provided a more accurate and detailed description of the main origins of peanut in **lines 81-88**.

Thank you once again for your valuable suggestion. It is crucial for us to accurately present and understand the origin of peanuts.

Line 93

What is meant by 'predominantly cultivated ecotypes'? As all *Arachis hypogaea* are domesticated (cultivated), this term seems redundant or misinformed. Could the authors clarify why this phrase was chosen?

Response:

Thank you for your valuable comment regarding the term "predominantly cultivated ecotypes." You are absolutely correct that all *A. hypogaea* (peanut) are domesticated, and the use of this term may indeed cause confusion.

The phrase "predominantly cultivated ecotypes" was intended to refer to the primary cultivars, distinguishing them from other varieties. However, we recognize that this phrasing might not be the most accurate or clear. To avoid any potential misunderstanding, we have removed this term in the revised version of the manuscript in **line 96**.

Lines 100-102

The authors misattribute fundamental knowledge about the genus *Arachis* and the

domestication process to recent genome papers that did not originally discover or describe these insights. For example, they state: "During the process of domestication, selection for larger pod sizes led to a transition from the small pods found in wild accessions to the larger, high-yielding pods characteristic of modern varieties." This observation, while accurate, has been well-documented in much earlier studies on the genus and on domestication and does not originate from genome sequencing efforts. Proper attribution to the foundational studies that established these concepts is essential to accurately represent the development of knowledge in the field. The authors need to cite the proper literature.

Response:

Thank you for your detailed and thoughtful comment. You are absolutely correct that the observation regarding the selection for larger pod sizes during the domestication of *A. hypogaea* has been well-documented in earlier studies. It is essential to properly attribute this knowledge to the foundational works, rather than to more recent genome sequencing papers.

In response, we have revised the manuscript to ensure that the transition from small pods in wild accessions to the larger, high-yielding pods in modern varieties is accurately credited to the earlier studies that originally described this phenomenon. We have carefully updated the references^{14–18} to include the key literature that established these concepts and have removed any misleading attribution to genome sequencing efforts for this well-established insight.

Thank you for highlighting this issue. The citations have been corrected accordingly in the revised version of the manuscript **in line 103**.

Lines 111-113

Some parts of the discussion are not comprehensible "In addition, key candidate QTLs for branching habit, testa color, root nodule development, and NB-LRR genes. have been identified. These QTLs were identified through map-based cloning". QTLs for NB-LRR genes doesn't make sense because QTLs describe quantitative traits rather than specific gene categories. "These QTLs were identified through map-based cloning" appears to be inaccurate in this context. The citations closest to this sentence include three genome papers, one unrelated methodological paper, and a review on NB-LRR genes—none of which seem to describe the use of map-based cloning. Map-based cloning involves isolating genes within a QTL region through fine mapping and positional cloning, which differs from the methods typically used in genome studies or broader QTL mapping.

Response:

Thank you for your insightful feedback. We acknowledge that some parts of the discussion are unclear, particularly regarding the mention of QTLs for NB-LRR genes and the reference to map-based cloning.

As you correctly pointed out, QTLs refer to quantitative traits, not specific gene categories. We have revised the sentence to more accurately describe the identification of QTLs related to traits such as branching habit, testa color, and root nodule development, without incorrectly associating them with specific gene categories like NB-LRR.

Additionally, we understand your concern regarding the reference to map-based cloning. We appreciate that this method, which involves isolating genes within QTL regions through fine mapping and positional cloning, is not typically associated with genome-wide studies or broad QTL mapping. We have updated the manuscript to reflect the correct methodology used in the cited genome studies and removed any inaccurate references to map-based cloning in **lines 113-115**.

Thank you again for your valuable suggestions, and we have ensured these changes are made to enhance the clarity and accuracy of the manuscript.

The figure following Fig. 1 lacks both a legend and a figure number. (This doesn't really matter because I could use the legend in the previous version).

Response:

Thank you for pointing out that the figure following Fig. 1 lacks both a legend and a figure number, and we apologize for any inconvenience this may have caused. In the revised version of the manuscript, we have reviewed the figure and correctly labeled it with both a figure number and a legend.

Thank you again for your valuable feedback.

In the text, the authors have adjusted their estimate for the date of peanut's origin in the right direction, but likely for the wrong reasons as the graph is incorrectly plotted and is not a suitable method for estimating such a recent divergence.

To give context to my previous request to add extra self-comparisons. The reason I did this was to demonstrate to the authors how nonsensical this plotting is. The methods used by the authors are unable to show that a sequence is identical to itself, and for the same reasons that the self-comparison fails, a comparison of two closely related genomes fails. I assumed that, seeing this, the authors would totally reform this part of the paper. Unfortunately, not.

Now, I will explain why the graph (A) is wrongly plotted.

- 1) The label on the vertical axis "Frequency" isn't defined and covers up the dramatically different numbers of gene comparisons under each curve. To correct the axis, it must have a scale of number of genes in each category.
- 2) The use of a line graph is wrong. The data is a frequency distribution; it isn't continuous data. A scatter-plot or column graph should be used. Although it's common practice to use a line graph for Ks curves, the plotting methods are left over from when ESTs were used to produce the the data. In the days of Ks graphs from ESTs, the data was smoothed and (probably) modeled. In contrast, data from whole genome comparisons needs no smoothing or modeling because it is a complete dataset. There are standards of representation of data required by Nature, and this representation drastically fails those standards. Take, for example, the blue lines; as they move from 0 to higher values, the blue

line initially rises. This is a complete misrepresentation because most gene comparisons in the blue curves have zero Ks values (from the authors own data provided in tsv files) - the blue line should move down when initially moving away from zero. Put in other words, the peak values on the blue lines should be zero, not a value higher than zero.

Now, I will explain why the use of Ks values isn't suitable for closely related genomes

Using Ks rather to calculate divergence is ill-suited for closely related genomes with short divergence times. Ks graphs, which measure synonymous substitutions in coding regions, was historically developed to enable comparisons between distantly related sequences where intergenic regions are too divergent to align reliably. This method is highly effective for long-term evolutionary studies, as it focuses on conserved coding regions that are possible to align, and where there is sufficient time for synonymous substitutions to accumulate. However, for recent divergences, the number of synonymous substitutions is too low to generate meaningful averages, resulting in high variability and poor resolution.

For closely related genomes that can be completely aligned (as is the case for these genomes) using SNPs across the whole genome leverages a much larger dataset, capturing variations in both coding and noncoding regions. This provides the granularity necessary for precise divergence estimates over short timescales. Relying on Ks for such analyses is like trying to measure 10 micrometers with a wooden ruler - it's the wrong tool for the scale of the measurement. SNPs, by encompassing genome-wide differences, are far better suited for estimating divergence in recently separated genomes.

To help illustrate, here are calculations to show that two peanut-like subgenomes, each of 1.4Mb, diverged by 9,000, would show most genes as having Ks=0 values

Parameters

- 1) Number of aligned orthologous genes: 28,662 (from the author's data)
- 2) Average alignment length per gene: 1,100 base pairs (bp) (from the author's data)
- 3) Fraction of synonymous sites: 75% of coding sequence sites are synonymous (825 bp per gene).
- 4) Neutral substitution rate: 8.12×10^{-9} Ks/site/year
- 5) Divergence time: 9,000 years (as an illustration)

Calculations

Step 1:

The expected number of synonymous substitutions per gene (λ) is calculated as:

$$\lambda = 2 \times \text{Substitution rate} \times \text{Divergence time} \times \text{Number of synonymous sites per gene}$$

$$\lambda = 2 \times (8.12 \times 10^{-9}) \times 9,000 \times 825 = 0.1201$$

In other words, only about 12% of the genes will have a synonymous substitution. The rest won't!

Step 2

[Strictly speaking, for the next step, we should use the Poisson distribution, but because λ is so low, it really doesn't make much difference]

Therefore, out of 28,622 gene pairs, we would expect $28,622 \times 0.88 = 25,223$ to have $K_s=0$. (That number is pretty close to the B vs B comparisons from the paper's data which in one example is ~24,500 gene pairs having $K_s=0$)

Therefore, the genes with $K_s>0$ are unusually highly mutated. Filtering $K_s=0$ numbers is like trying to determine the average speed of cars on a highway by only measuring the fastest ones - you'd ignore the slower majority, leading to a distorted conclusion. The absence of genes with $K_s=0$ removes the necessary baseline for accurately estimating divergence on a short timescale, as the molecular clock relies on the full distribution of substitutions, including those that have not accumulated any. Having said this, the best way to date close divergences like this is SNP rates from, whole genome alignments.

Response:

Thank you for your detailed and insightful feedback on our work. I greatly appreciate your careful review of the figures and your explanations regarding the flaws in our plotting methods. The issues you raised, such as the incorrect labeling of the vertical axis, the inappropriate use of line charts to display frequency distributions, and the misrepresentation of the blue line, have been duly noted, and we agree that these elements need to be corrected to align with standard scientific data presentation. I also recognize that the K_s value is not suitable for closely related genomes, and I appreciate your suggestion to use SNP data from whole-genome alignments to better estimate the level of genomic divergence. We have removed the K_s -based results in the revised version in **lines 191-199&791-803**.

I apologize for the errors with the line chart color labeling and the misunderstanding between the last response version and the original publication. We regret any confusion this may have caused.

Additionally, as per your suggestion, we have calculated the time of polyploidization in the peanut genome using SNP data. By analyzing the SNP divergence between the A subgenome (aligned with V14167) and the B subgenome (aligned with K30076), we have estimated the polyploidization time.

SNP identification was performed using MUMmer. The modal divergence was statistically analyzed with a 10 kb window and a 1 kb overlap. The results indicated that the A subgenome had an average divergence of 21.5 (ranging from 19 to 23), while the B subgenome exhibited a significantly lower average divergence of only 1.25 (ranging from 1 to 2). Furthermore, the modal SNP frequency was also calculated, with a modal SNP frequency of 60.02 per 10,000 bases (ranging from 59.57 to 60.54) in the At subgenome, and a modal SNP frequency of 4.01 per 10,000 bases (ranging from 3.48 to 4.68) in the Bt subgenome (**Table 3 (Supplemental Table 10)**).

Further analysis suggests that the AA-At divergence occurred around 190,000 years ago, while the BB-Bt divergence took place approximately 10,000 years ago (**Table 3 (Supplemental Table 10)**). These results indicate that the whole-genome duplication event occurred approximately 10,000 years ago, which is consistent with previous estimates⁸.

These updated results have been incorporated into the revised version, specifically in **lines 200-209&787-790**. We believe this provides a more accurate and meaningful analysis of the divergence

time scale.

Once again, thank you for your valuable suggestions and thorough guidance, which have greatly improved the quality and accuracy of our research.

Table 3. The SNP rates in whole genome alignments.

Ref_query	All snps	The modal divergence (10KB)	Genome size bp	SNPs/10KB	years
V14167_S245A	7148685	22	1200000000	59.57	187466
V14167_HN873A	7265256	22	1200000000	60.54	190522
V14167_HN51A	7220666	19	1200000000	60.17	189353
V14167_S83A	7177163	23	1200000000	59.81	188212
K30076_S245B	701635	2	1500000000	4.68	14720
K30076_HN873B	607475	1	1500000000	4.05	12744
K30076_HN51B	522102	1	1500000000	3.48	10953
K30076_S83B	575081	1	1500000000	3.83	12065

Why does Supplementary Figure 14 (B) use two different methods for the divergence dates Ks for the origin and coalescent-based genetic simulation for the others? The latter is giving dates I would expect (although they are stated with absurd precision (eg 9,491 years)).

Response:

Thank you very much for your insightful feedback. We initially intended to integrate the origin and divergence times of peanuts, which led us to combine both aspects. However, thanks to your guidance, we now understand that the KS results are not suitable for estimating divergence times in this context. As a result, in the new version, we have made corrections and retained only the results related to population divergence (**Figure 5(Supplementary Figure 14B)**). We sincerely appreciate your valuable suggestions.

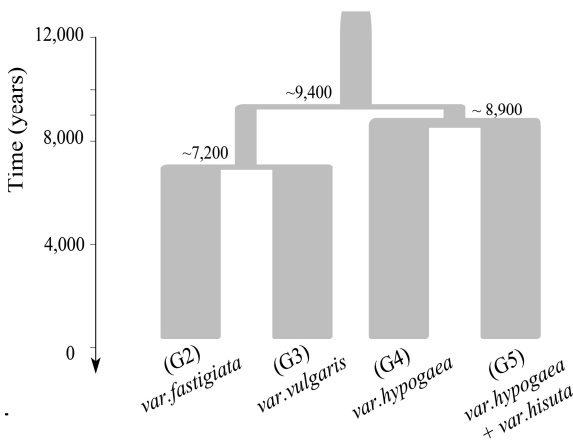


Figure 5. Demographic history of peanut populations.

Reviewer #3 (Remarks to the Author):

Overall evaluation on the revised version

The authors have successfully resolved the concerns and questions risen during the first revision. The manuscript, although it contains a huge amount of information, is clear and easy to read. The results are well supported by several data sources (phenome2genome associations, transcriptomics...). As I mentioned in my first revision, the manuscript is well built, has interest and present relevant results not only for the peanut scientific community but also for the genomic community working in genome evolution and polyploidy.

Revised by Aureliano Bombarely on December 23rd, 2024

Response:

Thank you for your positive feedback and for recognizing the revisions made in response to the first review. We are pleased that the manuscript is clear and well-supported by various data sources. As you mentioned, we are glad that the manuscript's structure and findings are relevant not only to the peanut scientific community but also to the broader genomic community focused on genome evolution and polyploidy. Your feedback has been invaluable in improving the manuscript.

Once again, thank you for your thoughtful comments and support.

References

1. Zhang, Y. C. *et al.* Overexpression of microRNA OsmiR397 improves rice yield by increasing grain size and promoting panicle branching. *Nat Biotechnol* 31, 848–852 (2013).
2. Cho, S. H. *et al.* The rice narrow leaf2 and narrow leaf3 loci encode WUSCHEL-related homeobox 3A (OsWOX3A) and function in leaf, spikelet, tiller and lateral root development. *New Phytologist* 198, 1071–1084 (2013).
3. Dai, M., Hu, Y., Zhao, Y., Liu, H. & Zhou, D. X. A WUSCHEL-LIKE HOMEBOX gene represses a YABBY gene expression required for rice leaf development. *Plant Physiol* 144, 380–390 (2007).
4. Otyama, P. I. *et al.* Genotypic characterization of the U.S. Peanut core collection. *G3: Genes, Genomes, Genetics* 10, 4013–4026 (2020).
5. Jung, S. *et al.* The High Oleate Trait in the Cultivated Peanut [*Arachis Hypogaea* L.]. I. Isolation and Characterization of Two Genes Encoding Microsomal Oleoyl-PC Desaturases.
6. Chu, Y., Holbrook, C. C. & Ozias-Akins, P. Two alleles of ahFAD2B control the high oleic acid trait in cultivated peanut. *Crop Sci* 49, 2029–2036 (2009).
7. Yang, H. *et al.* Fine mapping of qAHPS07 and functional studies of AhRUVBL2 controlling pod size in peanut (*Arachis hypogaea* L.). *Plant Biotechnol J* 21, 1785–1798 (2023).
8. Bertoli, D. J. *et al.* The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nat Genet* 48, 438–446 (2016).
9. Kochert, G. *et al.* RFLP AND CYTOGENETIC EVIDENCE ON THE ORIGIN AND EVOLUTION OF ALLOTETRAPLOID DOMESTICATED peanut, *Arachis hypogaea* (Leguminosae). *Am J Bot* 83, 1282–1291 (1996).
10. Moretzsohn, M. C. *et al.* A study of the relationships of cultivated peanut (*Arachis hypogaea*) and its most closely related wild species using intron sequences and microsatellite markers. *Ann Bot* 111, 113–126 (2013).
11. Seijo, G. *et al.* Genomic relationships between the cultivated peanut (*Arachis hypogaea*, Leguminosae) and its close relatives revealed by double GISH. *Am J Bot* 94, 1963–1971 (2007).
12. Robledo, G. & Seijo, G. Species relationships among the wild B genome of *Arachis* species (section *Arachis*) based on FISH mapping of rDNA loci and heterochromatin detection: A new proposal for genome arrangement. *Theoretical and Applied Genetics* 121, 1033–1046 (2010).
13. Grabile, M., Chalup, L., Robledo, G. & Seijo, G. Genetic and geographic origin of domesticated peanut as evidenced by 5S rDNA and chloroplast DNA sequences. *Plant Systematics and Evolution* 298, 1151–1165 (2012).
14. Leal-Bertioli, S. C. M. *et al.* Phenotypic effects of allotetraploidization of wild *arachis* and their implications for peanut domestication. *Am J Bot* 104, 379–388 (2017).
15. Krapovickas, A. & Gregory, W. C. Taxonomy of the Genus *Arachis* (Leguminosae). *Bonplandia* 16 (SUPL.), 1–205 (2007).
16. Freitas, F., Francisco, J. & Valls, M. *Genetic Variability of Brazilian Indian Landraces of Arachis Hypogaea* L. <https://www.researchgate.net/publication/5798151>.
17. Dillehay, T. D., Rossen, J., Andres, T. C. & Williams, D. E. Preceramic adoption of peanut, squash, and cotton in Northern Peru. *Science* (1979) 316, 1890–1893 (2007).
18. Simpson, E. *et al.* History Of *Arachis* Including Evidence Of *A. Hypogaea* L. Progenitors History and Origin of the Genus.

Authors' responses to Reviewers Comments

We would like to express our sincere thanks for the insightful suggestions provided by the editor and the Reviewers, which have significantly contributed to the improvement of our manuscript. The following are our responses to each of the Reviewers' comments (in blue colour).

Reviewer #1 (Remarks to the Author):

I am entirely satisfied with the authors' acceptance of my suggestion re: colorblind palettes: "Regarding your final suggestion about figure color palettes, we fully agree with your point about the importance of using color-blind friendly palettes. In the current revised manuscript, we have ensured that all figures adopt consistent and more accessible color schemes, with careful attention to color-blind compatibility. We have updated Fig. 1 and Fig. 2 to replace the bright red, blue, and green with more muted, color-blind friendly alternatives, in line with the colors used in Figs. 3 and 4."

Regarding their responses to the other reviewers, in a short look, they seem well-considered, but I'm afraid I don't have time to carefully evaluate them.

Authors' response: We are thankful to the Reviewer # 1 for appreciating the revised version of the MS.

Reviewer #2 (Remarks to the Author):

Please see uploaded file. It contains formatting of text and graphs that can't be entered into this box. If there are any issues with the uploaded file, please contact me so that we can address the problems.

Authors' response: We are thankful to the Reviewer # 2. Detailed response to other points is provided in the following sections.

Reviewer #3 (Remarks to the Author):

The author did resolve most of my concerns and questions in the previous revision. I think that the new version of the manuscript has been improved after a new round of interactions with other reviewers. From my point of view the manuscript it is good to

go.

Authors' response: Sincere thanks to the Reviewer # 3 for appreciating the revision.

Reviewer2' Comments:

Review of: “Telomere-to-telomere genomes of diploids and allotetraploids provide insights into the origin, evolution, and domestication of peanuts”

Despite its volume, the paper offers no new insights

The multitude of tables, graphs, and verbose text only serves to obscure the absence of substantive new conclusions. With the advent of long-read sequencing, assembling large and complex genomes have become relatively routine. The paper is poorly referenced, represents previously known findings as new, misrepresents prior research, and contains numerous errors, unsupported statements, and wild speculations. It has undergone very substantial changes to its fundamental conclusions during review—this removes some of the more obvious mistakes—but how can the credibility of the work remain intact? Indeed, after looking further, I highlight below more major problems below:

The paper's current title doesn't reflect its content. A suitable title could be: “**Extreme similarity of peanut genomes confirms a recent single origin**”

Authors' response: Thank you very much for your suggestion. We have addressed suggestions of the Reviewer 2 while revising the MS. We have also updated the title to: 'Telomere-to-telomere genomes of diploids and allotetraploids provide insights into genome organization, evolution, and agronomic traits for peanut improvement.'

Origin

There are no new insights into the origin of *Arachis hypogaea* in this paper. In reply to my question what are the insights, the authors replied in three points:

- 1) “we refine the timeline of tetraploid formation to approximately 10,000 years ago, consistent with earlier studies” – this was previously reported by Bertoli et al 2016.
- 2) “our centromere length analysis offers new insights into structural differences between subgenomes.” – this isn’t an insight about the origin, it’s about the structural differences of the component subgenomes.
- 3) “The analysis suggests that a diploid B species may have diverged from a diploid A species before hybridizing with an A genome to form the tetraploid.” - The divergence of the A and B subgenomes is well known, as is their union into the allotetraploid. The principle of divergence is obvious, it’s implied by the concept of evolution. The allotetraploid nature of peanut via the union of A and B subgenomes was first reported by Husted in 1936. The species origin of the A and B subgenomes has been the subject of extensive research that is now summarized in the new version of the paper in lines 62-88. The dates of divergence of the A and B subgenomes have been previously estimated as 3.5 Mya (Nielsen et al 2011), then refined to 2.3–2.9 Mya (Moretzsohn et al 2012), then further refined to ~2.16 Mya by Bertoli et al 2016).

There are no new insights about the origin of *Arachis hypogaea* in this paper.

Authors’ response: Thank you very much for your suggestion.

1) We agree with Reviewer 2’s observation that the time of the first tetraploid formation, approximately 10,000 years ago, was reported in the article by Bertoli et al. (2016) ¹. However, subsequent studies have shown discrepancies regarding the formation time of the tetraploid. For instance, Chen et al. (2019) assembled the Fuhuasheng genome and compared it with previously published wild genomes, suggesting that the tetraploid formation time was less than 0.18 million years ago². Yin et al. (2020) analyzed the wild tetraploid Monticolar and wild materials published in 2019, estimating the tetraploid formation time of Monticolar to be around 11,690 years ago³. Additionally, in 2019 and 2020, Zhuang et al. and Bertoli et al. conducted assemblies of Shitouqi and Tifrunner, respectively. Based on the divergence relationship between tetraploid subgenomes and diploid genomes, they estimated significantly different times for the tetraploid divergence⁴⁻⁶. Zhuang et al. (2019 & 2020) estimated the tetraploid formation time to be around 45 million years ago, whereas Bertoli et al. (2016 & 2020) suggested

a time of approximately 10,000 years ago. Meanwhile, Zheng et al. (2024) re-sequenced the chloroplast and whole genome of peanuts and inferred the tetraploid divergence time. They proposed that the differences in tetraploid formation time between Ahh (<10,000 years ago) and Ahf (0.42-0.47 million years ago) might be attributed to the fact that the reference genomes come from different subspecies⁷. Thus, the time of tetraploid formation remains a subject of ongoing debate.

Additionally, due to the limitations of sequencing and assembly technologies, many published diploid and tetraploid genomes contain numerous contigs and gaps, which significantly affect the accurate identification of SNP variations. Moreover, since peanut has two subspecies and six varieties, predicting the tetraploid formation time based on a single subspecies or variety is inaccurate. We should mention that we should not just remain focused only studies from one group (Moretzsohn et al 2012; Bertoli et al 2016). We should consider all other studies that have also been published in top class journals such as Nature Genetics, Molecular Plant, etc.

In the present study, we constructed six T2T genomes, including two wild diploids and four varieties. This approach not only allows for accurate identification of variations between the tetraploid subgenomes and their diploid progenitors, but also enables the analysis of whether differences among subspecies/varieties affect the prediction of tetraploid formation time. Therefore, we believe that our analysis is crucial and necessary for accurately determining the tetraploid formation time. This result represents the application of complete genomes for variant identification and divergence time analysis in peanuts.

2) Compared to the centromere length in the diploid genome, there are significant differences in the centromere length and structure of the tetraploid B subgenome. As no complete centromeres of diploids and allotetraploids were reported and analyzed, our study provides insights of centromere evolution, enabling the formation and stable inheritance of the tetraploid.

3) As the Reviewer mentioned, there is general agreement in different studies regarding the divergence time of the wild diploid AB genome, which is approximately around 2 million years ago. However, the **origin relationship** between the A and B genomes (A to B or B to A) has been rarely reported. Using centromere repeat units for phylogenetic tree analysis, we have preliminarily identified that the diploid B species **may have originated from** a diploid A species. This result represents the first attempt to explore the divergence relationship of the diploid subgenomes (A and B) using centromere repeat units.

Building upon previous research, we have conducted a systematic analysis of the origin of multiple peanut varieties using multiple T2T genomes. In particular, we have further clarified the tetraploid formation time and analyzed the divergence relationship of the diploid genomes through centromere structural variation. This study holds significant importance for understanding the origin of peanuts.

Therefore, we don't think that it is a fair assessment of the MS and mentioning that "There are no new insights into the origin of *Arachis hypogaea* in this paper".

Evolution

There are no new insights into the evolution of *Arachis hypogaea* in this paper. In reply to the question as to what are the insights, the authors replied:

"Gypsy-type LTR retrotransposons showed distinct insertion times between the At and Bt subgenomes..." - Distinct insertion times of component genomes are to be expected with a divergence of more than 2 Mya, and have been reported before. Nielen et al (2010) reported at least threefold higher copy number of the most abundant Ty3-gypsy retrotransposon in the A-genome ancestor compared to the B-genome ancestor. Bertoli et al (2013) reported of the ten most abundant LTR retrotransposons that "Almost all dateable transposition events occurred <3.5 million years ago, the estimated date of the divergence of A and B genomes." and "The activity of these retrotransposons has been a very significant driver of genome evolution since the evolutionary divergence of the A and B genomes."

How insightful can the paper's evolutionary analysis be, if it failed to alert the authors that the original manuscript version had overestimated the origin of peanut by more than a quarter of a million years?

Authors' response: Thank you very much for your suggestion.

We carefully reviewed the articles mentioned by the Reviewer and summarized them as follows: Nielen et al. (2010) focused on the FIDEL-aretrovirus-like retrotransposon in the AB subgenomes and emphasized the role of FIDEL in the evolution and divergence of different *Arachis* genomes⁸. Bertoli et al. (2013) utilized 27 *A. duranensis* BAC clones as probes and employed the BAC-FISH method. They found that the sequences of 14 BAC clones revealed complete and truncated copies of ten abundant long terminal repeat (LTR) retrotransposons. These ten LTR transposition events occurred less than 3.5 million years ago and have been a significant driver of genome evolution since the evolutionary divergence of the A and B genomes⁹.

Due to limitations in sequencing and assembly technologies, the studies by Nielen et al. (2010) and Bertoli et al. (2013) could not provide a comprehensive genome-wide analysis of various transposon types and insertion times, nor could they explore the roles of different transposons in genome evolution. Additionally, due to material constraints, these studies were unable to perform identification across multiple varieties or subspecies.

In contrast, our study, based on the polyploid formation timeline, subdivides different types of transposable elements (TEs) based on their insertion times. We identified insertion events that may have occurred before polyploid formation and compared the changes in TEs before and after tetraploid formation. Our findings showed that multiple TEs present in the diploid genomes were absent from the tetraploid subgenomes, suggesting complex evolutionary dynamics and potential differential retention or loss of TEs during polyploidization. These results, which have not been previously reported, highlight the importance of the T2T genome assembly for further research on TE dynamics in peanut evolution.

We believe that our study, based on T2T genomes, provides a better understanding of the expansion of LTR retrotransposons in the AB subgenomes and will significantly contribute to advancing research on peanuts. Moreover, our findings offer valuable insights that could be applied to the study of TEs in other species as well. As mentioned earlier, if our studies provide different/ better results than reported by Bertoli et al. doesn't mean that our study is useless and doesn't provide new insights.

Domestication

There are no new insights into the domestication of *Arachis hypogaea* in this paper. In reply to the question as to what are the insights, the authors replied:

“The total number of SVs was significantly higher in the comparison of At_vs_V14167 (average: 70,439) than in Bt_vs_K30076 (average: 14,917)...”, their analysis then builds off this. - A greater number of SVs is completely expected between At_vs_V14167 than Bt_vs_K30076, because *A. duranensis* V14167 is more distantly related to the A subgenome donor than *A. ipaënsis* is to the B subgenome ancestor.

Closer representatives of the A subgenome donor, *A. duranensis* from Rio Seco, have been identified. Why didn't the authors sequence one of these?

How insightful can the paper's analysis of domestication be, given that it failed to alert the authors that their inference of “multiple origins of the B subgenome” in the original version of their manuscript was incorrect?

Authors' response: Thank you very much for your suggestion.

In terms of domestication, our analysis focused on comparing structural variations (SVs) between subgenomes, which included comparisons not only with wild diploids but also between different tetraploid varieties. In addition to using the wild diploid as the reference, we also used the cultivated tetraploid S83 as the reference to identify variations between tetraploid varieties. These varieties, during the domestication process, exhibit diverse phenotypes, such as increased pod number, seed size, oil content, and plant architecture. These phenotypic differences are influenced by various selective pressures. The primary cause of these phenotypic variations can be traced back

to genotype differences. The construction of the T2T genome not only enables the accurate identification of large structural variation (SV) differences but also allows for the discovery of genes related to these variations, followed by functional validation.

In this study, as per Reviewer 1's suggestion, we categorized the different SVs and conducted separate enrichment analyses, which will be crucial for the subsequent identification of key genes in peanut domestication. This will provide valuable insights into the genetic underpinnings of important traits that have been selected for during the domestication of peanuts.

Based on the SV analysis, we identified a 13.23 Mb-long insertion on Chr14, and visualized its coverage using both HiFi and Illumina reads. The accuracy of this structural variation (SV) was further confirmed through Hi-C heatmap analysis. Statistical analysis revealed that all samples in the G5 population contained this SV, while 23% and 15% of samples in the G4 and G3 populations contained it, respectively. No samples in the G2 population contained this SV (Supplementary Figures 10A-10E). These results suggested that this SV could be associated with the differentiation between the G5 and G4 populations, which both belong to the var. *hypogaea* subspecies. In this SV region, we identified 53 homologous genes with over 70% sequence identity and an e-value of less than $1e-10$, including several rice functional genes related to plant architecture and organ development, such as OsA7, OsWOX3A, OsWOX3B, OsIAA17, OsLAC, and OsRLCK107. Among these, two genes, chr14.1804 and chr14.1807, were found to be tissue-specific, particularly expressed in roots, stems, and leaves but not in cotyledons, shells, seeds, or pegs (Supplementary Fig. 12). These two genes are homologous to OsLAC, which is a target gene of OsmiR397. Studies suggest that OsLAC regulates various developmental processes in rice by modulating the brassinosteroid (BR) signaling pathway. Overexpression of OsLAC leads to shortened stems, upright leaves, and reduced grain size, phenotypes resembling those seen in BR-deficient plants. Moreover, OsmiR397, which negatively regulates OsLAC expression, enhances BR signaling and alters plant architecture by increasing leaf angles. These

findings indicate that OsLAC, through its interaction with the BR pathway, plays a vital role in regulating plant growth and development, offering potential strategies for improving crop architecture and yield through the manipulation of hormone signaling pathways¹⁰.

Another gene, chr14.1829, which is homologous to the WOX3A/B gene. Studies indicate that both OsWOX3A and OsWOX3B genes play crucial roles in the organ development of rice, especially in the development of leaves, tillers, lateral roots, and lateral organs. In the case of the OsWOX3A mutant, the plant exhibits narrow and curled leaves, an increase in tiller number, a decrease in lateral roots, and narrower grains. This phenotype is associated with restricted lateral growth of the leaves, fewer longitudinal veins, and the enlargement of bubble-like cells. Moreover, OsWOX3A also plays a key role in the development of the outer glume (lemma), with mutations leading to abnormal glume morphology, affecting the seed shape¹⁰. At the molecular level, OsWOX3A is primarily expressed in the vascular tissues of rice, and mutations lead to changes in the expression of genes related to leaf development, hormone transport, and other pathways, resembling the phenotypes seen in PIN gene mutants.

On the other hand, OsWOX3B plays an essential role in the symmetry of leaf development. The OsWOX3B mutant displays asymmetrical leaves and sometimes forked leaf shapes, which leads to abnormal development of lateral organs, affecting leaf width and morphology. Mutants also show an increase in tillers, longer panicles, and a greater number of grains per panicle. Furthermore, the vascular bundles in the flag leaf (the leaf above the panicle) increase in number, and the spacing between small vascular bundles widens, further affecting the growth and development of rice¹¹.

Otyama et al. (2020) classified 787 accessions from the U.S. Peanut core collection and 12 commercial varieties into Clades 1 to 4 using the *Arachis*_Axiom2 SNP array data⁴¹². Phenotypic observations showed that in Clade 1, "most accessions are classified as the hypogaea botanical variety and exhibit a 'mixed' or hypogaea pod shape (60.0%, 44.4%, 40.0%). The growth type varies widely, with an even distribution between erect, bunch,

spreading-bunch, mixed, and prostrate types."

Further analysis reveals that Clade 1 in our study corresponds to the G4+G5 populations. Based on phylogenetic relationships and material information, we found that both G4 and G5 belong to var. hypogaea. Field surveys indicate that the G5 population has a larger tillering angle (with a higher proportion of prostrate types), whereas the G4 population has a larger proportion of erect types. This is consistent with the observed variation in growth types within Clade 1.

Combining the functional annotation and expression results of the 109 SV-genes, we hypothesize that the growth type difference between the G4 and G5 populations may be linked to the presence of this structural variant (SV).

Therefore, we believe that comparing the structural variations (SVs) between the diploid genome and different tetraploid subgenomes is of great significance for exploring genomic changes during the tetraploid domestication process. At the same time, SV analysis between different tetraploid varieties plays a crucial role in identifying key domestication candidate genes that have been subject to selection. This comparative approach not only helps to elucidate the genomic alterations that have occurred during the domestication of peanuts but also aids in pinpointing specific genetic changes that have contributed to the selection of important traits. Identifying these SVs and their associated genes will provide valuable insights into the underlying mechanisms of domestication and could inform future breeding strategies for enhancing peanut crops.

Above explanations indicate and prove our point about new insights on domestication.

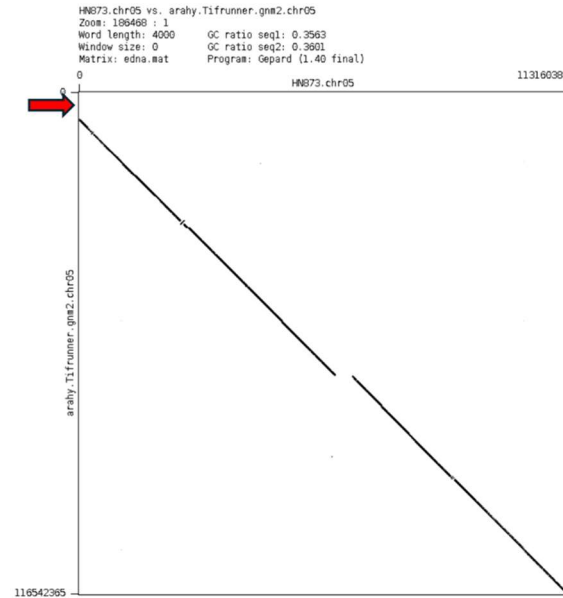
The assemblies are incomplete

A key feature of the paper is that it reports "telomere to telomere assemblies".

In fact, four half-chromosome arms are missing:

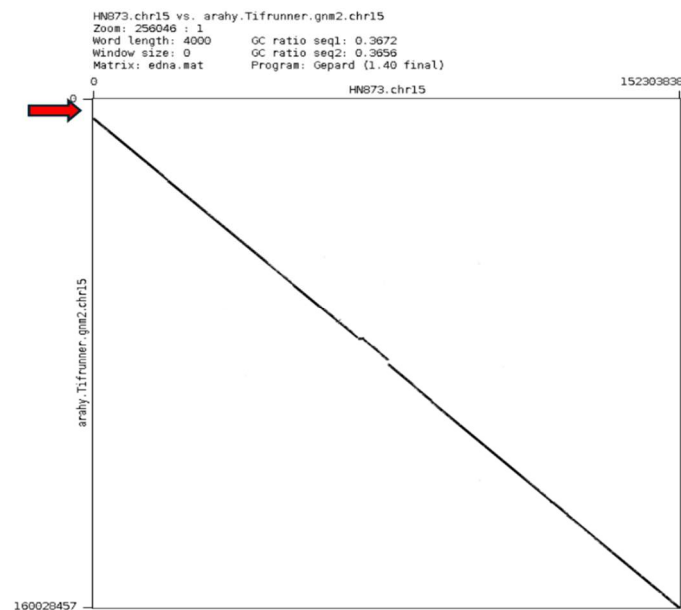
Missing region 1)

More than 6 Mb (about half the euchromatic chromosome arm, where most of the chromosome's genes are) is missing from chr05 of HN873. The sequence is indicated by the red arrow in the below plot of HN873 vs the Tifrunner reference genome.



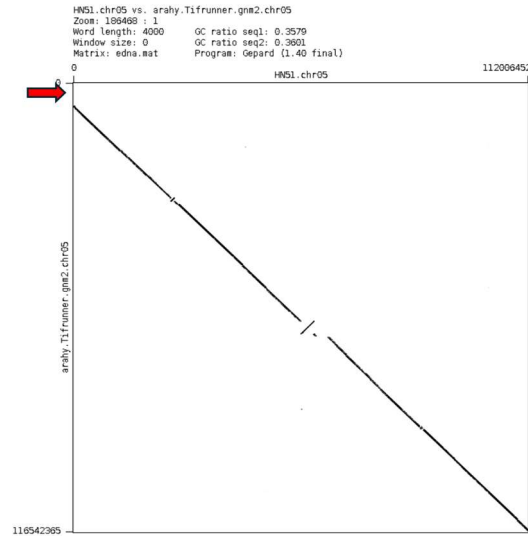
Missing region 2)

About half the euchromatic chromosome arm is missing from chr15 of HN873. The sequence is indicated by the red arrow in the below plot.



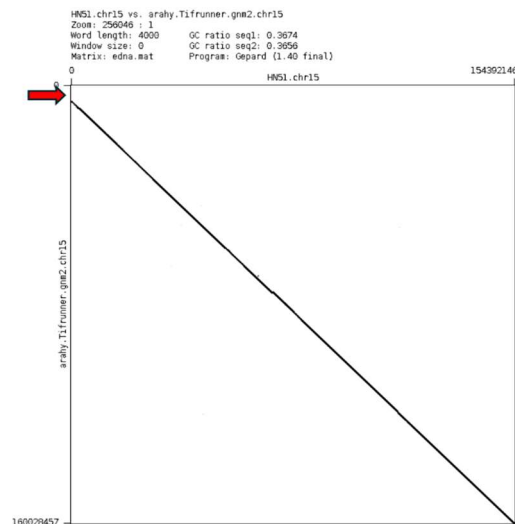
Missing region 3)

About half the euchromatic chromosome arm is missing from chr05 of HN51. This is indicated by the red arrow in the below plot.



Missing region 4)

About half the euchromatic chromosome arm is missing from chr15 of HN51. This is indicated by the red arrow in the below plot of HN51.

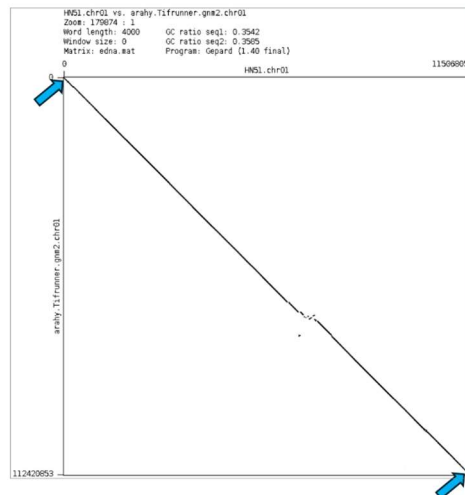


The terminal regions of all these four chromosomes are also completely lacking characteristic telomeric sequences. The explicit claim made in the paper that there are 40 telomere sequences in the four sequenced tetraploid genomes is simply incorrect. As is their gross misrepresentation that the previously sequenced genomes: Tifrunner, Shitouqui, Fuhuasheng, *A. duranensis* and *A. ipaënsis* have 0 telomeric sequences! (see

eg. Table 1) – Many of the chromosome ends of previous assemblies do, in fact, have the telomeric sequences.

For comparison here is a plot of two complete chromosomes.

Note how the line in the plot runs from corner to corner, indicated by blue arrows:



A major selling point of the paper is its analysis of genomic differences—but how reliable and informative could that analysis have been since the absence of four half-chromosome arms went unnoticed?

The paper misrepresents genome similarities

These plots reveal another major shortcoming of the paper: it gives a misleading impression of extensive, complex differences between the genomes. In reality, the genomes are strikingly similar to current reference genomes. The lines in the plots above are composed of dots indicating 4,000 bp windows of identical sequence. Aside from the centromeric regions, which are highly repetitive and gene-poor, and a few deletions and reciprocal exchanges, the genome sequences are almost identical to the existing diploid and tetraploid reference sequences. The authors have presumably “talked-up” the differences to make their hypothesis-free research seem more significant.

Authors’ response: We greatly appreciate your thorough analysis.

Upon receiving this feedback, we immediately verified the data and discovered that the

issue arose because the Reviewer had used the **V1-gapfree version** of the genome uploaded on May 30, 2024. However, prior to our first revision (2024.12.08, NG-A65542R), we had updated our assemblies and the **latest T2T genomes** were provided in the Data Availability section (lines 1103-1108). It seems that the Reviewer used the older version of the assembly and made his/her remarks.

Please note that all related data analyses, present in this MS (version) is based on the update version of genome assemblies. However, we wonder, that if the Reviewer has not used the correct version of genome assemblies (probably he/she just wants to kill the paper, and without even going through the update genome assemblies, he/she just made negative comments on the basis of previous version.

When we compared our T2T genomes with the Tifrunner.gnm2 genome used by the Reviewer, we observed a high degree of collinearity between these chromosomes. Below are the details of the genome update process and the associated synteny analysis results, including data processing and alignment information:

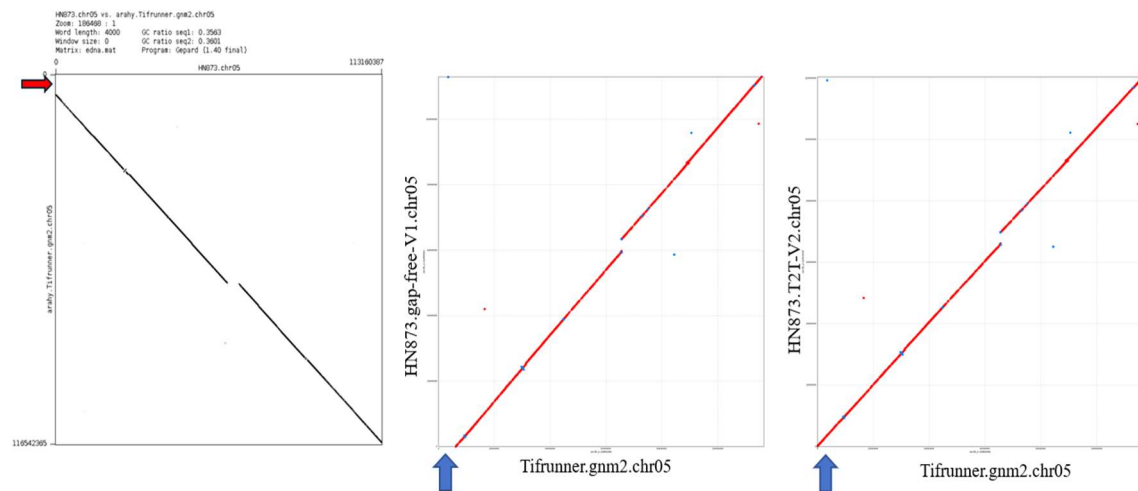
Telomere Deletion Validation:

Based on the Reviewer's alignment results, we verified the genomic collinearity and discovered that the Reviewer used the **gapfree-V1** version of the genome from May 30, 2024. This version has incomplete telomere sequences, which resulted in the telomere deletions mentioned by the reviewer (**shown in the left image in the Figure below**). The middle image shows our verification results using the gapfree-V1 version. And the right image shows the alignment results from the **T2T-V2** version with the same genome used by the Reviewer (arahy.Tifrunner.gnm2.J5K5.genome_main.fna), and it showed higher consistency, with no fragment deletions present.

Collinearity Results of Different Genome Versions with the Tifrunner.gnm2 Genome. gapfree-V1: Left and Middle Figures, T2T-V2: Right Figure, Chromosome start: Red and Blue arrows indicate.

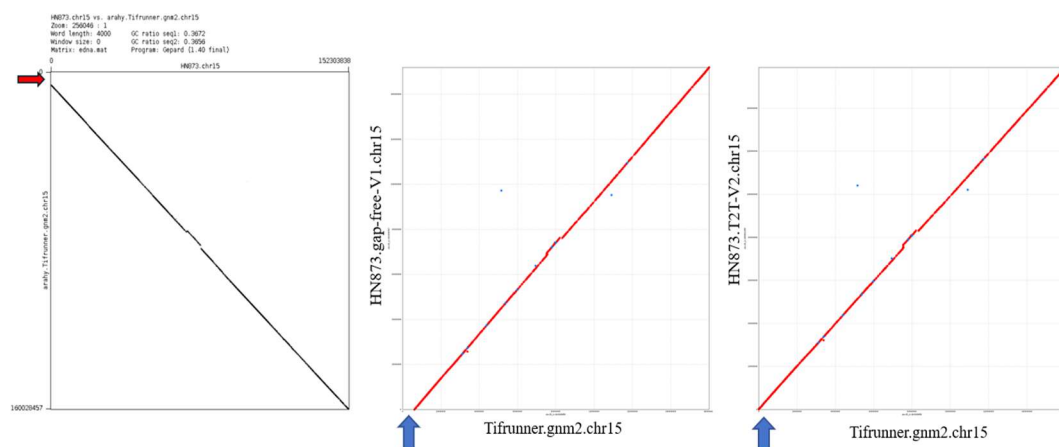
1) HN873.chr05-VS-Tifrunner.gnm2.chr05

The image on the right shows the latest version of the genome, with no deletions observed.



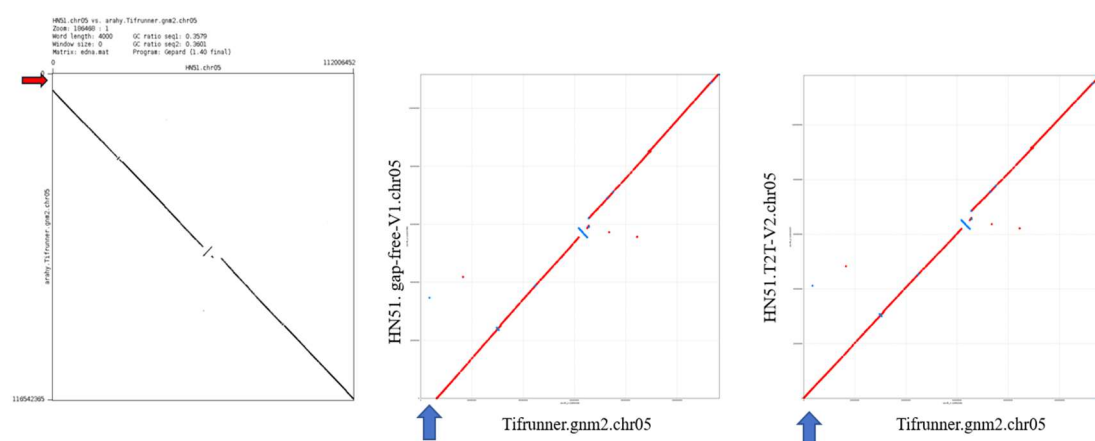
2) HN873.chr15-VS-Tifrunner.gnm2.chr15

The image on the right shows the latest version of the genome, with no deletions observed.



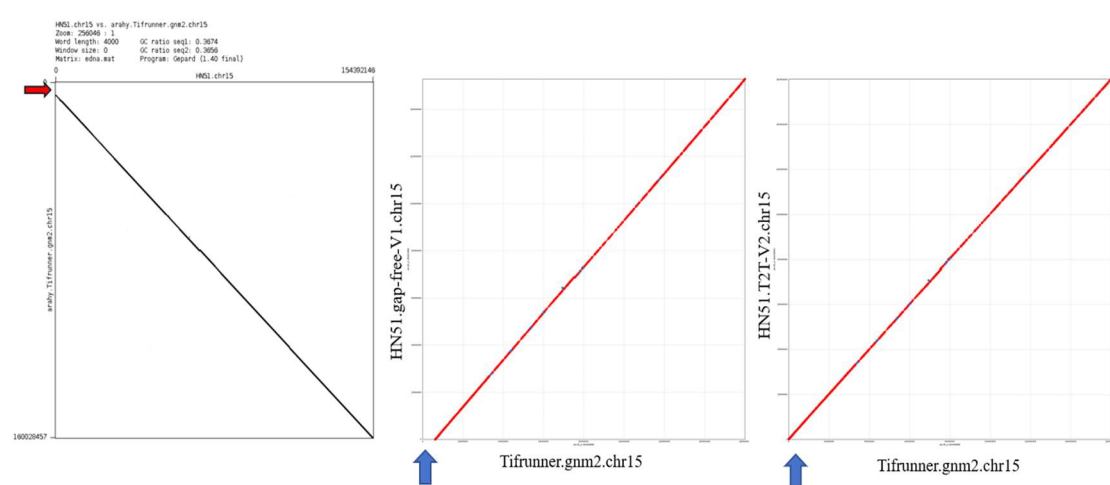
3) HN51.chr05-VS-Tifrunner.gnm2.chr05

The image on the right shows the latest version of the genome, with no deletions observed.

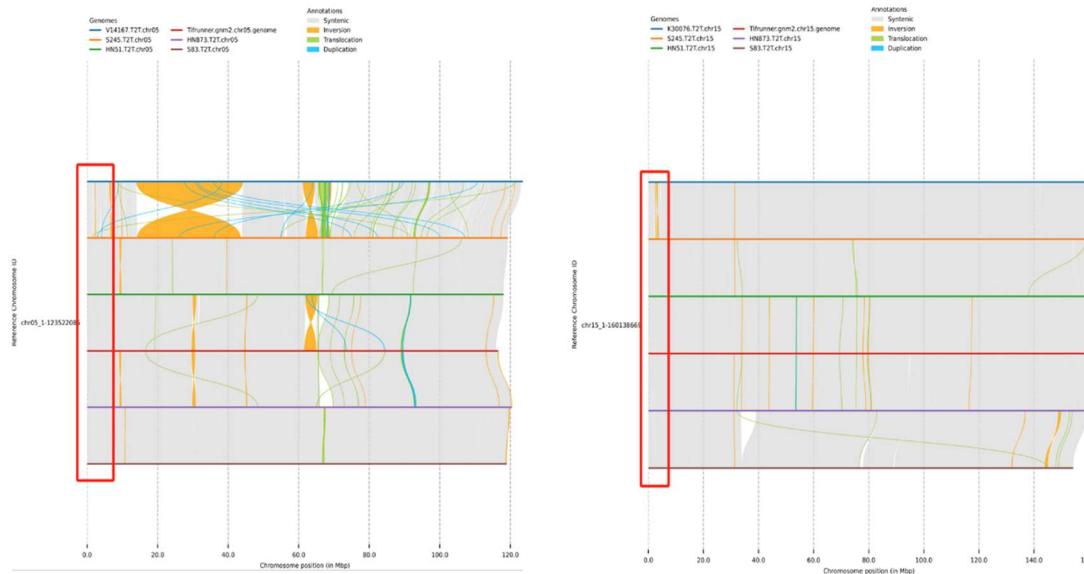


4) HN51.chr15-VS-Tifrunner.gnm2.chr15

The image on the right shows the latest version of the genome, with no deletions observed.



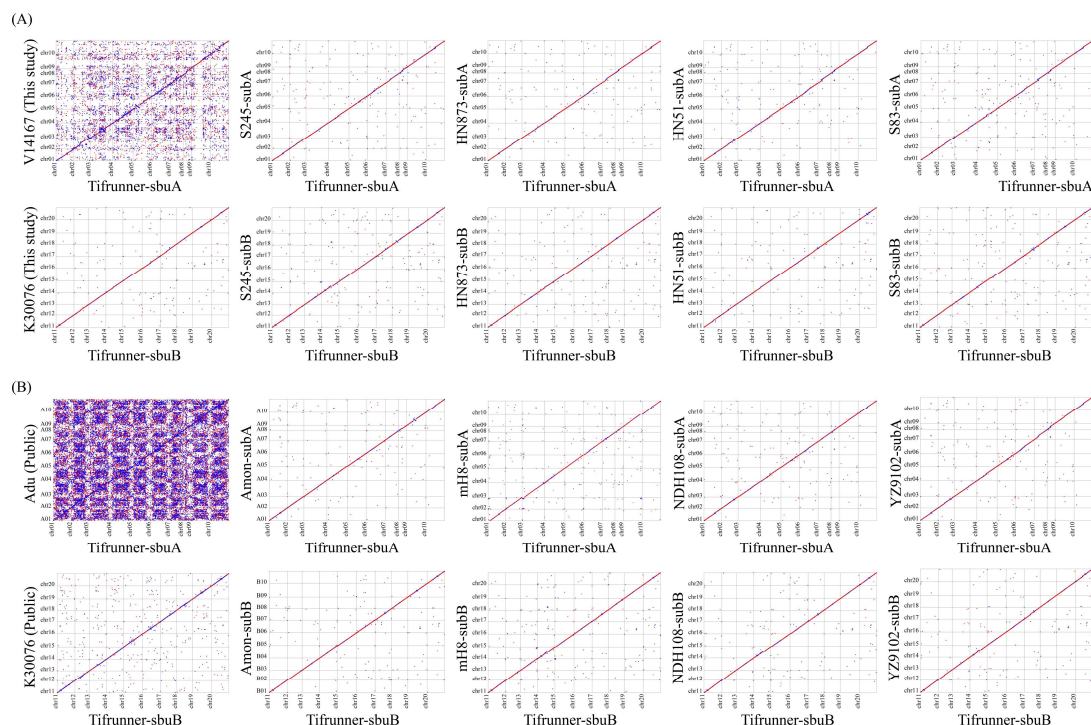
5) The collinearity of different accessions on chr05 and chr15 shows that the V2 version of the genome exhibits a high level of consistency in chromosomal collinearity, with no telomere deletions (regions in red boxes).



Additionally, we performed a syntenic analysis using the published Tifrunner.gnm2 genome as the reference, comparing it with our assembled 6 T2T genomes and 6 published genomes (the K30076 (public) genome from Bertoli et al. (2016)¹; Adu (public), Amon, mH8, and NHD108 genomes from Zhao et al. (2025)¹³; YZ9102 from Wang et al. (2025))¹⁴. Except for YZ9102, which is a T2T genome, the contig

sequences in the other five genomes were deleted prior to alignment, leaving only chromosome sequences.

The results showed a high degree of synteny with Tifrunner, with no evidence of telomere deletions. This result is also added in **Supplementary Figure 4**. The results are as follows:



Based on the above results, we believe that the issue of telomere deletions mentioned was caused by the incorrect use of the gapfree-V1 version (2024.05.30) of the genome, and there is no issue with assembly completeness in the updated T2T-V2 genomes. The genomes used in the above analysis can also be downloaded from the following databases and links:

1. **Figshare:**
https://figshare.com/articles/dataset/T2T_genome_assembly_in_peanut/28883636
2. **NGDC:**
https://ngdc.cncb.ac.cn/gwh/search/advanced/result?search_category=&search_term=&source=0&query_box=PRJCA026588

3. Google Drive:

https://drive.google.com/drive/folders/1Y_HoDMkFaPXgOMXDKHs3SyWvUlhFAfZe

Once again, we appreciate the Reviewer's feedback, and we believe that viewing and analysis of the updated T2T genome can address the Reviewer's concerns.

Genome-wide association studies yielded no new verified associations

In response to my criticism that nearly all of the 4,638 loci reported as associated with agronomic traits lack verification, the authors cite just two FAD gene marker-trait associations-both of which are already well known and have been published in multiple prior studies.

The FAD gene sequences and their role in the high oleic trait were reported in the early 2000s by Jung et al (2000), López et al 2000 and López et al 2002 (see also Refs 35-37 in the paper).

The authors also give the case of *AhRUVBL2*, but without any details of verification, only saying “we have prioritized these representative traits for candidate gene mining, haplotype analysis, and experimental validation.”

Authors' response: Thank you very much for your suggestion.

Regarding the results of the GWAS analysis, we considered the points exceeding the threshold as candidate loci and performed statistical analysis, a method commonly used in several studies. By combining the 200kb linkage interval around the GWAS Manhattan peak points, we selected previously functionally validated genes, *FAD*¹⁵⁻¹⁷ and *AhRUVBL2*¹⁸, as positive controls to confirm the accuracy of our GWAS results. Furthermore, based on the population haplotype analysis, we identified a significant correlation between the variation in the *FAD* gene and the high-oil and low-oil populations, which is consistent with previous studies.

For the *AhRUVBL2* gene, we observed a clear signal in the GWAS Manhattan results, and it lies within the 200kb linkage interval around the peak points. Previous studies have shown that this gene is a key regulator of seed size. We used these two genes as

positive controls to demonstrate the accuracy of our GWAS results.

Regarding the Reviewer 2's comment about "prioritizing these representative traits for candidate gene mining, haplotype analysis, and experimental validation," we selected major effect genes that influence oil content, a key trait for peanut quality, and genes that regulate seed size, which is critical for yield and development, for a series of validation work. The results of these validations are presented in the main text.

It is also important to mention that GWAS analysis provides large number of genetic loci/genes associated with various agronomic traits, and verification/ validation can be done only for few genes as examples. Therefore, such a criticism that validation has been done only for two genes and that too for well-established genes is not justified.

The comments above highlight some of the most important problems with the paper. Below is a selection of comments on the narrative from the Discussion

Representing previously known findings as new

Bian et al, the first results discussed in Discussion section lines 530-532 "Based on the analysis of TE insertion timing, the insertion of Gypsy elements was identified as the primary factor contributing to the larger size of the B genome/subgenome relative to the A genome/subgenome."

Compare to results from Bertoli et al 2016 "All *A. ipaensis* pseudomolecules were larger than their *A. duranensis* counterparts... This is partly because of a greater frequency of local duplications and higher transposon content in *A. ipaensis*." Furthermore, this is because of deletions after inversions (not just insertions) "Size differences between homeologous chromosomes that were differentiated by large rearrangements tended to be greater than those between collinear ones ($r(6) = 0.65$, $P < 0.05$; Supplementary Table 14). Because the *A. duranensis* chromosomes that have undergone inversions are smaller than expected, it is evident that, in this dynamic, on balance, the elimination of DNA has predominated over its accumulation."

This is gone over in more detail by Ren et al 2018.

Bian et al, Lines 554-555 “These findings suggest that a diploid B species diverged from a diploid A species and subsequently hybridized with an A genome to form a tetraploid.”

– this is, of course, completely established

Bian et al, Lines 562-563 “Based on the results of the phylogenetic tree and population structure, we inferred that the differentiation of cultivated peanuts into four major groups.” – this is, of course, completely established

Authors’ response: Thank you very much for your valuable suggestions. We have revised discussion sections based on your feedback.

Wild Speculation

Lines 535-537 “Significant TE insertion events were also observed in the A genome/subgenome during the S3 period (100,001–300,000 years ago), suggesting that these events created favorable genomic conditions for tetraploid formation.”

Lines 539-540 “This discrepancy may largely be attributed to differences in genome stability and the functional redundancy of TEs.”

Lines 548-549 “The difference in centromere length between cultivated peanuts and wild diploids related to chromosome pairing during meiosis.”

Authors’ response: Thank you very much for your valuable suggestions. Although we don’t agree with comments of the Reviewer, we have considered some suggestions and revised the MS accordingly.

Errors

Lines 568-570 “var. *fastigiata* types are primarily cultivated for consumption ... In contrast, var. *hypogaea* types are predominantly grown for oil extraction.” In fact, virtually the entire production of the USA, Argentina and Brazil is var. *hypogaea* types for direct consumption.

Lines 571-573 “Furthermore, the differentiation between G4 and G5 primarily arises from differences in planting methods. G4, with its upright growth habit, supports higher

planting densities, thereby increasing yield.” This is simply untrue; prostrate growth habitpeanuts have a longer growing season and, consequently, generally a much higher yield.

Their cultivation is generally preferred where the rainy season is long enough, especially where harvest is mechanized.

Authors’ response: Thank you very much for your valuable suggestions. We have revised and corrected this section based on your feedback.

References

1. Bertoli, D. J. *et al.* The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nat Genet* 48, 438–446 (2016).
2. Chen, X. *et al.* Sequencing of Cultivated Peanut, *Arachis hypogaea*, Yields Insights into Genome Evolution and Oil Improvement. *Mol Plant* 12, 920–934 (2019).
3. Yin, D. *et al.* Comparison of *Arachis monticola* with Diploid and Cultivated Tetraploid Genomes Reveals Asymmetric Subgenome Evolution and Improvement of Peanut. *Advanced Science* 7, (2020).
4. Zhuang, W. *et al.* The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. *Nat Genet* 51, 865–876 (2019).
5. Zhuang, W. *et al.* Reply to: Evaluating two different models of peanut’s origin. *Nature Genetics* vol. 52 560–563 Preprint at <https://doi.org/10.1038/s41588-020-0627-0> (2020).
6. Bertoli, D. J., Abernathy, B., Seijo, G., Clevenger, J. & Cannon, S. B. Evaluating two different models of peanut’s origin. *Nature Genetics* vol. 52 557–559 Preprint at <https://doi.org/10.1038/s41588-020-0626-1> (2020).
7. Zheng, Z. *et al.* Chloroplast and whole-genome sequencing shed light on the evolutionary history and phenotypic diversification of peanuts. *Nat Genet* 56, 1975–1984 (2024).
8. Nielen, S. *et al.* FIDEL-a retrovirus-like retrotransposon and its distinct evolutionary histories in the A- and B-genome components of cultivated peanut. *Chromosome Research* 18, 227–246 (2010).
9. Bertoli, D. J. *et al.* The repetitive component of the A genome of peanut (*Arachis hypogaea*) and its role in remodelling intergenic sequence space since its evolutionary divergence from the B

- genome. *Ann Bot* 112, 545–559 (2013).
10. Zhang, Y. C. *et al.* Overexpression of microRNA OsmiR397 improves rice yield by increasing grain size and promoting panicle branching. *Nat Biotechnol* 31, 848–852 (2013).
 11. Dai, M., Hu, Y., Zhao, Y., Liu, H. & Zhou, D. X. A WUSCHEL-LIKE HOMEBOX gene represses a YABBY gene expression required for rice leaf development. *Plant Physiol* 144, 380–390 (2007).
 12. Otyama, P. I. *et al.* Genotypic characterization of the U.S. Peanut core collection. *G3: Genes, Genomes, Genetics* 10, 4013–4026 (2020).
 13. Zhao, K. *et al.* Pangenome analysis reveals structural variation associated with seed size and weight traits in peanut. *Nat Genet* (2025) doi:10.1038/s41588-025-02170-w.
 14. A telomere-to-telomere genome assembly of the cultivated peanut. doi:10.6084/m9.
 15. Jung, S. *et al.* *The High Oleate Trait in the Cultivated Peanut [Arachis Hypogaea L.]. I. Isolation and Characterization of Two Genes Encoding Microsomal Oleoyl-PC Desaturases.*
 16. Chu, Y., Holbrook, C. C. & Ozias-Akins, P. Two alleles of ahFAD2B control the high oleic acid trait in cultivated peanut. *Crop Sci* 49, 2029–2036 (2009).
 17. Nawade, B. *et al.* Insights into the Indian peanut genotypes for ahFAD2 gene polymorphism regulating its oleic and linoleic acid fluxes. *Front Plant Sci* 7, (2016).
 18. Yang, H. *et al.* Fine mapping of qAHPS07 and functional studies of AhRUVBL2 controlling pod size in peanut (*Arachis hypogaea* L.). *Plant Biotechnol J* 21, 1785–1798 (2023).

Authors' responses to Reviewers Comments

We sincerely thank the Editor and Reviewers for their insightful feedback, which has greatly contributed to the improvement of our manuscript. Below are our responses to each reviewer's comments (highlighted in blue).

Reviewer #1 (Remarks to the Author):

The authors, in their response, present a convincing case with associated graphics that appears to demonstrate that a reviewer mistakenly analyzed an earlier version of the assembly in question. I have no further comments for improvement.

Authors' response: We sincerely thank Reviewer #1 for your thoughtful feedback and for recognizing the revisions we made in the manuscript. We greatly appreciate the time and effort invested in reviewing our work.

Reviewer #3 (Remarks to the Author):

As I stated in previous revisions, I think that the authors resolved all the concerns that I rose during my first revision. Furthermore, I think that they did an excellent job answering and resolving not only my questions and concerns but also others risen by the other reviewers. I think that the manuscript has enough merit for its publication.

Authors' response: Thank you very much for your kind words and for recognizing the efforts we have made in revising the manuscript. We truly appreciate your constructive feedback, which has been invaluable in improving the quality of our work. We are grateful for your support and look forward to the manuscript's potential publication.

Reviewer #4 (Remarks to the Author):

Comparative Analysis & Novelty:

Previous studies have already addressed topics such as genome assembly, selective

sweeps, evolutionary history, and phenotypic diversification. The authors should represent the unique contributions through the differences between previous studies and this study.

Authors' response: We appreciate your insightful suggestions. In the revised version, we have added a comparison with previous studies on genome assembly, particularly emphasizing our novel contribution of assembling the T2T genomes of two wild diploid species and four representative cultivated varieties. A comparison with published genomes revealed that, aside from the previously published YZ9102 genome, which has already achieved a T2T level, the other genomes still contain numerous contig sequences. Importantly, no T2T-level genome for wild diploid species has been published to date, marking a significant advancement in accurately identifying the timing of tetraploid formation and tetraploid variations. For instance, we found chromosomal exchanges on chr07, chr08, chr17, and chr18 in the tetraploid, which were already present in the ancestral diploid, not occurring after the formation of the tetraploid.

Regarding the population analysis section, we have summarized key previous studies on peanut populations. Previous research primarily focused on the following:

1. **Improved Analysis from Landrace to Cultivated Varieties:** For example, Liu et al. (2022) performed a population analysis of 203 cultivated varieties. In their selective sweeps analysis, they used *Fst* and *XP-CLR* to examine landrace and cultivated varieties, identifying signals of peanut improvement.
2. **Comparison of Peanuts in Northern and Southern China:** Lu et al. (2024) analyzed a population of 390 peanut materials, focusing on selective elimination between northern and southern China, as well as non-Chinese materials, identifying relevant regions and genetic diversity.
3. **Phenotypic Differences between Two Subspecies:** Zheng et al. (2025) divided 353 tetraploid cultivated peanut materials into two subspecies, *Ahh* and *Ahf*, and analyzed candidate genes related to traits such as flowering patterns and seed inner tegument color.

In our study, based on the T2T genome, we have expanded upon previous work by identifying variations within a population of 521 representative accessions. For the first time, we compared genetic differentiation and introgression (gene flow) regions among the four distinct cultivated varieties, identifying significant asymmetric differentiation and introgression patterns. Additionally, in response to Reviewer 1's suggestion, we evaluated the impact of transposable elements (TEs) on the top 1% of these selection and introgression regions.

Once again, thank you for your valuable suggestions. These revisions have greatly contributed to the improvement of our manuscript.

Origin & Evolution Section:

The contents on Origin and Evolution should be toned down, as these aspects have been provided in previous studies. The authors should focus more on their novel findings.

Authors' response: We thank the Reviewer for this suggestion. In the revised version, we have streamlined and revised the sections on peanut origin and evolution to minimize overlap with previous studies, thereby placing greater emphasis on our novel findings.

SV-Domestication Relationship:

Authors should strengthen the relationship between structural variations (SVs) and domestication. The current manuscript does not sufficiently explain the functional relationships among genes, SVs, and domestication traits.

Authors' response: We appreciate your valuable suggestion. In addition to identifying genes related to plant architecture in the 13.2 Mb variation on chromosome 14, we have further identified and enriched four representative peanut-specific SV variations in the revised version, using the wild diploid as the reference (**lines 281-301**). We have also performed enrichment analyses of the specific genes from different variations, which allowed us to further elucidate the pivotal role of SVs in the domestication process.

Thank you once again for your input. These additions significantly enhance our understanding of peanut domestication.

Genome Similarities:

I agree with the authors' responses regarding genome similarities.

Authors' response: Thank you very much for your recognition. We appreciate your acknowledgment of our responses regarding genome similarities.

GWAS Analysis:

In addition to the known loci (FAD and AhRUVBL2), the authors are encouraged to include genetic mechanism analyses for one or two additional phenotypic traits to enhance the GWAS findings.

Authors' response: Thank you very much for your valuable suggestion. In the revised version, we have expanded our analysis to include two additional phenotypic traits: Kernel Dehydration Rate (KDR) and Arachidic Acid content in peanuts. KDR is crucial for preventing premature germination and the effects of aflatoxin contamination after maturation, ensuring peanuts reach a safe dehydration level. Arachidic acid, a key component of saturated fatty acids, has been linked to increased risk of atherosclerosis and cardiovascular disease when consumed in excess. Therefore, reducing its content is vital for human health.

In response to your suggestion, we have conducted a thorough GWAS analysis for these two phenotypes, identifying candidate genes (**lines 496-544; Supplementary Fig. 25&26**). Further, we have incorporated gene annotation, structural analysis, haplotype analysis, population-based haplotype variation correlations with phenotypes, which confirm the essential role of these candidate genes in regulating KDR and Arachidic Acid content.

Once again, thank you for your constructive feedback, which significantly enriched our manuscript.

Title & Writing Style:

The current title is too informal, and the overall writing style lacks academic rigor (e.g., the use of "etc." in the Abstract). A more precise and professional revision should be recommended.

Authors' response: Thank you for your valuable feedback. We agree that the current title and writing style could benefit from a more formal and precise approach. In the revised version, we have refined the title to reflect the academic rigor required, and we have also revised the writing style to avoid informal language such as 'etc.' in the Abstract. We hope these changes align better with the expectations for academic writing.

In view of above, we believe that Editor and Reviewers are happy with our responses and revision of the MS.