

Exploring the genomic population structure and history of Austroasiatic speakers in Mainland Southeast Asia

Corresponding Author: Dr Ke Wang

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

Version 0:

Decision Letter:

Dear Dr. Wang,

Thank you for your patience during the peer review process. Your manuscript entitled "New insights into the genomic population structure and history of Austroasiatic speakers in Mainland Southeast Asia" has now been seen by four external referees, whose comments are appended below. You will see from their comments below that while they find your work of considerable interest, some important points are raised. We are interested in the possibility of publishing your study in Communications Biology, 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.

The four reviewers, with diverse expertise and knowledge of populations from MSEA, highly valued the data presented in the study, as well as the complexity and large number of methods employed. However, all of them agreed that the length of the text is notably substantial, and the main text should be reduced or moved to supplementary information. Additionally, reviewers highlighted the need for modifications and corrections to the presentation of the results and methodological approaches employed, as well as the inclusion of additional explanations for non-specialist readers. Also, some key points in the interpretation of DyStruct vs. ADMIXTURE and the use of qpGraph could be further contextualised. We therefore invite you to revise and resubmit your manuscript, taking into account the points raised.

Please highlight all changes in the manuscript text file.

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

At the same time, we ask that you ensure your manuscript complies with our editorial policies. Specifically:

For all graphs depicting a single point value (e.g., mean) with error bars, you must add individual data points or convert the graph to a boxplot or dot-plot to show data distribution.

It's mandatory to provide access to the numerical source data for graphs and charts either through a repository or by providing the data in a Supplementary Data file (in Excel format).

All blots/gels must be accompanied by size markers in every figure panel. Uncropped and unedited blot/gel images must be included as Supplementary Figure(s) in the Supplementary Information pdf.

Please ensure that you have complied with the data deposition policies at the Nature Portfolio, please see here.

Please ensure that you have complied with our policies on research involving animals and humans, see here

Please follow the ARRIVE guidelines for reporting animal experiments. Please fully complete an <a

[https://arriveguidelines.org/sites/arrive/files/documents/Author Checklist - Full.pdf](https://arriveguidelines.org/sites/arrive/files/documents/Author%20Checklist%20-%20Full.pdf)>ARRIVE checklist including both the essential and recommended set of items (adding information to the manuscript where needed) and upload this with your revised manuscript.

Please also see <https://www.nature.com/documents/CommsBio-file-checklist-revision.pdf>>our revision checklist for guidance on formatting the manuscript and complying with our policies. A comprehensive guide to our formatting requirements for final submissions is also available for your reference <https://www.nature.com/documents/commsj-life-style-formatting-guide-accept.pdf>>here.

Please use the following link to submit your revised manuscript, point-by-point response to the referees' comments (which should be in a separate document to the cover letter) and any additional files:

Link Redacted

**** This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first ****

When submitting the revised version of your manuscript, please pay close attention to our <https://www.nature.com/commsbio/editorial-policies/image-integrity>>Digital Image Integrity Guidelines.

We would like to receive your revision within 4 weeks, but appreciate that every situation is unique. We look forward to receiving your revised manuscript when it is ready, and will not enforce a hard deadline on this revision.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further. We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Best regards,

Cesar Fortes-Lima, PhD
External Editor
Communications Biology

Reviewers' comments:

Reviewer #1 (Remarks to the Author):

It is encouraging to see another manuscript -- "New insights into the genomic population structure and history of Austroasiatic speakers in Mainland Southeast Asia" by Yin et al; which dissects the population history, demography and evolution in MSEA. It is all the more exciting as the central theme revolves around delineating the intricacies of the under-represented AA speakers.

MAJOR COMMENTS:

The manuscript is very dense and the researchers have performed a lot of analysis. Some of these can be moved to the supplementary section to ease reading. I am however more concerned about the design and I think a major problem has not been appropriately addressed.

The problem arises because the authors considered only one Indian AA population for the analyses but tries to infer a lot about the dispersal of AA in South Asia. In the process the authors also are sold to one particular hypothesis, which relates AA origin to Southern regions of Yangtze River and their dispersal and spread as Expansion related to rice cultivation.

There is indeed a lot of literature that challenges this view and multiple models have been proposed. Indeed genetic analysis of S Asian AA as well as SE Asian AA have strongly proposed them as the original inhabitants of their respective regions. The presence of the indigenous Hoabinhian ancestry component among South Asian AA populations is not completely conforming with a neolithic expansion model around rice cultivation. The presence of the indigenous Hoabinhian ancestry in other present-day AA speakers such as Mon and Kharia may also suggest a widespread distribution of this ancestry in South Asia and MSEA, and not necessarily dispersal from MESA.

I would suggest that the authors drop the South Asian component of the study and concentrate on delineating the details of AA populations in MESA. Else they may consider including more Indian AA populations may increase the robustness and generate more confidence on the connection between South and Southeast Asian populations. In that case they have to appropriately test the other alternative hypotheses of the AA population history and demography.

The fine analysis of the AA within MESA is well done. I would think a little bit more detail of dispersal can be worked out

from the IBD sharing data.

A comparison with the findings of the following paper (Nature. 2025 Jul;643(8071):417-426. doi: 10.1038/s41586-025-08998-w. Epub 2025 May 14.) is important.

MINOR POINTS:

Strong IBD sharing between Mlabri and other Northern Mon-Khmer groups have been observed, despite substantial drift in Mlabri population in TREEMIX analysis along with distinct genetic ancestry in ADMIXTURE analysis. What might be the reasons for that?

Most Khmuic populations (except Khmu) do not exhibit increase in effective population sizes around 125-150 generations ago, that has been connected to the hypothesized rice agricultural dispersal around 4 kya, and rather show a declining trend. The pattern of increasing effective population size in Khmuic populations is mostly driven by Khmu, which has substantially higher sample size than all other Khmuic populations. Can you explain why did the effective population sizes of these AA populations did not increase even with the dispersal that likely introduced AA languages in Southeast Asia?

The quality of Fig1 is really bad. However, in PCA Fig 1c, Onge (Andamanese speakers) individuals are positioned between the individuals from Kharia (Indian Austroasiatic) and Maniq (Southeast Asian Austroasiatic) populations. Did you check why Onge, despite belonging to a different linguistic family, occupies this position in PCA?

Reviewer #2 (Remarks to the Author):

Review of "New insights into the genomic population structure and history of Austroasiatic speakers in Mainland Southeast Asia" by Yin et al.

Summary

This study investigates the origins, genetic structure, and historical interactions of Austroasiatic (AA) language speakers in Mainland Southeast Asia (MSEA), using the largest AA dataset to date that includes newly generated genome-wide data from seven AA and two Sino-Tibetan populations in Thailand, Laos, and Myanmar, alongside ancient DNA. The key questions addressed include when and where AA groups formed, how their genetic patterns align with their linguistic affiliations, and how historical contacts with East Asian and South Asian populations shaped their genomes. The findings reveal that AA groups exhibit significant genetic heterogeneity that broadly aligns with linguistic branches and geography, with northern highland groups showing greater East Asian ancestry and links to Iron Age populations, while southern lowland groups display additional South Asian ancestry and affinities with Neolithic populations. Surprisingly, South Asian-related ancestry was also detected in Neolithic MSEA samples, suggesting the presence of an unsampled, deeply diverged lineage shared across ancient South, East, and Southeast Asia. The study confirms that AA is likely the oldest linguistic and genetic layer in the region, predating other language families such as Sino-Tibetan and Tai-Kadai as AA speakers show genetic cladality with Neolithic samples while other language families show cladality only with later samples from Iron age and Historical era.

Overall impression of the work

The overall impression of this paper is that it is a thorough, well-written and well-executed study. It stands out for its impressively wide range of methods and brings valuable clarity to a complex topic. We liked that the introduction was very clear and to the point.

1. The manuscript is a substantial and impressive body of work, but in its current form it can feel overwhelming due to the number and density of analyses. As a result, it becomes difficult to follow the main narrative and identify the key findings. To help the central results stand out more clearly, we suggest streamlining the main text - possibly by moving some of the more technical analyses to the supplementary materials.

In particular, the Refined IBD section ("Intra-population and Inter-regional structure among AA speakers and MSEA populations") could benefit from a more concise presentation. Currently, the section includes an extensive description of the total amount and number of IBD blocks shared between populations. While informative, this level of detail may not be necessary for supporting the main conclusions. Especially when considering the section about overall length and count of shared IBD segments between population pairs (that can be influenced by several different population historical events), not the average segment lengths, which are often more informative for interpreting population contact time depth.

Similarly, the 5th chapter of results ("The genomic formation of ancient MSEA and legacy in present-day AA speakers") is quite long. For the sake of clarity, we suggest dissecting this chapter into two chapters, e.g. "The genomic formation of ancient MSEA" and "Legacy in present-day AA speakers".

2. We suggest more care in wordings related to interpretations of language group movements based on genetic data. See for example specific comment nr 1.

3. One of the key topics in the paper is the (early) presence of SA admixture in MSEA. In this regard it would seem valuable to have more than one source from SA in analyses like SOURCEFIND. This would enable to disentangle genetic

contributions from SA to MSEA before and after the Steppe migration. Yes, all SA groups today are mixtures of ASI and ANI but different proportions of those in different sources might help shed more light on the SA admixture into MSEA. Also including Onge as one source would make sense as they don't have ANI. Furthermore, it may be a good idea to include a population with Steppe ancestry but no ANI or ASI ancestry. Basically, some European population as a population from Central Asia would have East Asian ancestry, which would also confuse the picture. Currently the fact that Kharia draws all ancestry from a Brahmin group seems strange when AA sources as alternatives are also available.

Specific comments

Abstract:

1) 50 "The Austroasiatic (AA) language family represents the earliest known linguistic presence in Mainland Southeast Asia (MSEA) dating back to the Neolithic period,". – What is the direct evidence of AA language family in the Neolithic period? What is probably meant here is that people speaking AA languages display genetic component(s) that are detectable in the region already in the Neolithic period.

Introduction:

2) 77 "AA is thought to be the most ancient language family in MSEA" – add reference here

3) 78 "AA speakers are primarily distributed across mainland MSEA" - mainland 2x in a row in this sentence

4) "The AA language family comprises two primary branches - Mon-Khmer and Munda – and Southern Monic, Viet-Muong, Monic and Aslian branches." This sentence under Sup. Figure 1 is a bit confusing as this division of languages doesn't stand out from the figure depicting AA language phylogeny. Therefore, it seems like the figure has certain branches and the caption underneath describes different branches.

5 chapters of results:

I. Genetic heterogeneity within modern Austroasiatic speakers in MSEA

II. Relationships between modern AA groups and ancient samples

5) 251 "It is likely that this reflects a mismodeled case in which an ancestral population, which has affinities with several descendant populations that are highly drifted and hence highly differentiated, is inferred to be a mixture of the descendant ancestries, rather than as directly ancestral to the descendant ancestries." -Cool that this textbook example happened here and you were able to interpret it.

III. Intra-population and Inter-regional structure among AA speakers and MSEA populations

6) Supplementary Figure 5 has lots of little plots. Unfortunately, the numbers on their scales are not easily readable. They can be guessed, but it doesn't help that the scales differ between figures, which makes it especially hard to compare the already blurry scales with each other.

IV. South Asian and East Asian admixture in modern AA-speaking groups

7) 452 "the Munda Kharia was completely surrogated by SA ancestry." – I see a tiny bit of Kuay ancestry in Kharia on Fig. 4a and Sup. Fig. 11. but yes, it is very suspicious that Kharia is almost completely Brahmin_Tiwari.

8) 470 "Focusing on geographical locations (Fig. 4a), we found that populations shifted towards EA populations are all from northern MSEA, while those exhibiting shared genetic drift with SA populations are situated in southern MSEA." – this is interesting

9) 515 "when using Karen as the proximal surrogate for EA ancestry, the Karen admixture date in both Lawa groups is rather recent," –Karen is actually AA with language shift to ST and is therefore not a suitable proxy for EA ancestry (Mishra, R. K., Singh, P. P., Rai, N., Desai, S., Pandey, P., Tiwary, S. K., Tamang, R., Suravajhala, P., Shrivastava, P., Thangaraj, K., van Driem, G., & Chaubey, G. (2024). Reconstructing the population history of the Nicobarese. *European Journal of Human Genetics*, 1–7. <https://doi.org/10.1038/s41431-024-01720-w>)

V. The genomic formation of ancient MSEA and legacy in present-day AA speakers

10) 541 "exact" -should probably be "extract"

11) 598 "Overall, we observed a declining trend of indigenous MSEA ancestry from the Neolithic to the Iron Age and Historical era in MSEA (Fig. 5), likely reflecting later population admixture with people from East Asia and South Asia as observed in modern-day MSEA populations." – cool that the trend is visible with the amount of ancient samples

12) 636 "The estimated presence of this deeply diverged ancestry in Neolithic individuals (ca. 4000 years ago) and Bronze-Iron Age individuals (ca. 1800-2200 years ago) suggests either a persistent genetic link between the South Asia lineage and ancient MSEA, multiple prehistorical mixing events, or both." – Intriguing. Then the SA-AA admixture event(s) didn't necessarily happen in SA. For example, in the Indian AA speakers, some of the SA ancestry could be from earlier admixture that happened in MSEA and that could change the implication of admixture time = time of migration to SA.

Discussion

13) 861 "We found that only modern MSEA-AA groups show cladality with ancient samples from the Neolithic; specifically, the Eastern Mon-Khmer and Southern Monic lowland groups. In contrast, all other language groups show cladality only with samples from the Iron Age or Historical era periods. This supports the general consensus that AA is the most ancient language family in MSEA. Intriguingly, we identified a signal of ~20% ancient SA ancestry (using Indus_Periphery as a proxy, Methods) in the late Neolithic samples MaiDaDieu, ManBac, and Oakaie (Fig. 5), all dating to ~3-4 kya, which has not been reported before." - Interesting!

Reviewer #3 (Remarks to the Author):

This manuscript "New insights into the genomic population structure and history of Austroasiatic speakers in Mainland Southeast Asia" by Zi Yin et al. presents an important and well-executed genomic analysis of Austroasiatic-speaking populations in Mainland Southeast Asia. The authors combine dense genome-wide SNP data with a range of analytical approaches (PCA, ADMIXTURE, DyStruct, f-statistics, IBD, and qpGraph) to investigate both deep population structure and recent demographic histories.

The results contribute meaningfully to our understanding of linguistic, cultural, and genetic diversity in the region. In particular, the fine-scale resolution of population relationships among Austroasiatic speakers is novel and highly valuable. The manuscript is data-rich, methodologically up-to-date, and mostly clear in its exposition.

That said, there are several areas where improvement would enhance the clarity and accessibility of the paper, both scientifically and linguistically. I list below the different points that should be improved:

On the main text

1 Clarity of Language and Style

The manuscript occasionally uses passive voice and awkward constructions that obscure the meaning or weaken the narrative. Examples include:

- "Is thought to be among the oldest extant language families" (page 1, line 7) → could be simplified to "likely one of the oldest surviving language families."
- "IBD segments >3cM were used..." (page 4, line 9) → rewrite as "We used IBD segments longer than 3cM..."
- "Admixture graph topologies were modeled..." (page 8, line 17) → rewrite as "We modeled admixture graph topologies using..."

2 A full proofreading pass for stylistic consistency would strengthen the manuscript.

3 Interpretation of DyStruct vs. ADMIXTURE

While both methods are presented, the paper would benefit from a more explicit comparison between the two, including how different values of K were selected and how the outputs agree or diverge.

4 f4-Statistics and qpGraph Explanation

The rationale behind specific f4 tests and graph models is clear to specialists, but a brief paragraph explaining the biological implications of these statistics for a broader audience would improve accessibility. Also, residuals and Z-scores from alternative graph models should be included, even in supplementary files.

Minor Points and Edits

- Define acronyms such as IBD, PCA, and SNP at first use.
- The phrase "in agreement with historical records" is repeated several times and could be rephrased for variation (e.g., "consistent with historical data").
- Clarify whether ancient DNA (aDNA) was considered or excluded, and if so, why.
- Add a brief statement on ethical approvals for sampling and data use.

On the Supplementary Materials

The supplementary materials are rich and well-structured. A few improvements are suggested:

- 1 Methodological Descriptions: The IBD-based demographic inference would benefit from clearer mention of generation time assumptions and phasing quality.
- 2 ADMIXTURE and DyStruct: Include a comparative summary of K selection and interpretation.
- 3 qpGraph Topologies: Provide Z-scores and residuals for alternative models.
- 4 Metadata: Clarify dialectal or subpopulation distinctions in Table S1.
- 5 Figures: Improve legibility of S6–S10 (axis clarity and contrast).
- 6 Ethics: Add fieldwork permit and ethical clearance info, even in supplement. Or say where they are deposited.

In conclusion, this is a strong and significant contribution to the field of population genetics and Southeast Asian prehistory. With modest revisions to improve clarity, stylistic consistency, and methodological transparency, the manuscript will be ready for publication.

Reviewer #4 (Remarks to the Author):

I co-reviewed this manuscript with Reviewer #2, who provided the listed reports. This is part of the Communications Biology initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

** See the Nature Portfolio author and referees' website at www.nature.com/authors for information about policies, services and author benefits

Communications Biology 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' create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System 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 Manuscript Tracking System by clicking on 'Modify my Springer Nature account' and following the instructions in the link below. Please also inform all co-authors that they can add their ORCIDs to their accounts and that they must do so prior to acceptance.

<https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research>

For more information, please visit <http://www.springernature.com/orcid>

If you experience problems in linking your ORCID, please contact the [Platform Support Helpdesk](http://platformsupport.nature.com/).

Version 1:

Decision Letter:

** Please ensure you delete the link to your author homepage in this email if you wish to forward it to your coauthors **

Dear Dr. Wang,

Thank you for your patience during the peer review process. Your manuscript entitled "New insights into the genomic population structure and history of Austroasiatic speakers in Mainland Southeast Asia" has now been seen again by our referees, whose comments appear below. In light of their advice I am delighted to say that we are happy, in principle, to publish a suitably revised version in Communications Biology.

We therefore invite you to revise your paper one last time to comply with our format requirements and to maximise the accessibility and therefore the impact of your work.

* Please see the attached document for editorial requests for the final version (.docx file). Please ensure a completed version of this file is uploaded as a Related Manuscript with your final submission.

* Please review our [final submission file checklist](https://www.nature.com/documents/commsj-file-checklist.pdf) to ensure all necessary files are present with your final submission and to avoid delays in accepting your manuscript. For your reference, a style and formatting guide is available [here](https://www.nature.com/documents/commsj-life-style-formatting-guide-accept.pdf) and includes all of our style requirements.

It is important that you pay careful attention to the requests in these documents to avoid a delay in formal acceptance of the article.

If you have any questions or concerns about any of our requests, please do not hesitate to contact me.

Open access

Communications Biology is a fully open access journal. Articles are made freely accessible on publication. For further information about article processing charges, open access funding, and advice and support from Nature Research, please visit <https://www.nature.com/commsbio/open-access>

At acceptance, you will be provided with instructions for completing the open access licence agreement on behalf of all authors. This grants us the necessary permissions to publish your paper. Additionally, you will be asked to declare that all required third party permissions have been obtained, and to provide billing information in order to pay the article-processing charge (APC).

Please use the following link to upload your revised files:

Link Redacted

** This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first **

We hope to hear from you within two weeks. If you expect the process to take longer than one month, please let us know.

Congratulations on an excellent paper!

Best regards,

George Inglis

George Inglis, PhD
Deputy Editor
Communications Biology

on behalf of
Michele Repetto, PhD
Associate Editor
Communications Biology
orcid.org/0000-0001-5583-7287

and
Cesar Fortes-Lima, PhD
Editorial Board Member
Communications Biology
orcid.org/0000-0002-9310-5009

PS: At acceptance, the corresponding author will be provided with instructions for completing the license on behalf of all authors. This grants us the necessary permissions to publish your paper. Additionally, you will be asked to declare that all required third party permissions have been obtained, and to provide billing information in order to pay the article-processing charge (APC).

REVIEWERS' COMMENTS:

Reviewer #1 (Remarks to the Author):

The revised manuscript looks satisfactory to me.

Reviewer #2 (Remarks to the Author):

The authors have addressed all our comments. We are happy to see this manuscript accepted and printed.

Reviewer #3 (Remarks to the Author):

This manuscript "New insights into the genomic population structure and history of Austroasiatic speakers in Mainland Southeast Asia" by Zi Yin et al. presents an important and well-executed genomic analysis of Austroasiatic-speaking populations in Mainland Southeast Asia. The authors combine dense genome-wide SNP data with a range of analytical approaches (PCA, ADMIXTURE, DyStruct, f-statistics, IBD, and qpGraph) to investigate both deep population structure and recent demographic histories.

The results contribute meaningfully to our understanding of linguistic, cultural, and genetic diversity in the region. In particular, the fine-scale resolution of population relationships among Austroasiatic speakers is novel and highly valuable. The manuscript is data-rich, methodologically up-to-date, and clear in its exposition.

I understand that parts of my first-round review caused confusion. In several places, I placed paraphrased methodological or interpretive points in quotation marks, which may have given the impression that I was citing your text verbatim. This was not my intention — my aim was only to formulate suggestions and highlight areas where additional clarification might have been useful. I regret the lack of precision in how I presented it.

Having carefully read your thorough responses to my questions, I am satisfied that you have successfully addressed the concerns raised. Your revisions have strengthened the manuscript, and I now fully support its publication.

Thank you again for your careful work and for your patience with this clarification.

Reviewer #4 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Communications Biology initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

** See the Nature Portfolio author and referees' website at www.nature.com/authors for information about policies, services and author benefits

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/>

Point-by-point Responses

We thank the editors and reviewers for the thoughtful and constructive feedback on our manuscript. We have carefully reviewed all the comments and substantially revised the manuscript in response.

In particular, we have adjusted the main text and moved some detailed content to the Supplementary Information to improve clarity and readability. We have also revised the presentation of the results and associated methodologies to make the main text more accessible to a broader readership.

Furthermore, we have expanded our explanation of the methodologies used and elaborated on our novel findings compared to previously published research in this field. Here follows the original comments from reviewers in black, and our responses in blue text. The corresponding changes in the main text are also highlighted in blue text.

Reviewer #1 (Remarks to the Author):

It is encouraging to see another manuscript -- "New insights into the genomic population structure and history of Austroasiatic speakers in Mainland Southeast Asia" by Yin et al; which dissects the population history, demography and evolution in MSEA. It is all the more exciting as the central theme revolves around delineating the intricacies of the under-represented AA speakers.

We appreciate the positive feedback from the reviewer 1 and happy to see that the reviewer 1 recognizes the significance of our study on the under-represented AA speakers in mainland southeast Asia.

MAJOR COMMENTS:

The manuscript is very dense and the researchers have performed a lot of analysis. Some of these can be moved to the supplementary section to ease reading. I am however more concerned about the design and I think a major problem has not been appropriately addressed.

Thanks for pointing out that the manuscript might be too dense for readers. We have now made the following adjustment in the text:

Lines 274–314:

We have shortened the IBD analysis section by moving some information to the Supplementary Information, retaining only the key conclusions in the main text.

Lines 452–454:

We have condensed the ancestry modelling strategy for modeling ancient MSEA populations, and moved some of the details to the Methods and Supplementary Information section.

The problem arises because the authors considered only one Indian AA population for the analyses but tries to infer a lot about the dispersal of AA in South Asia. In the process the authors also are sold to one particular hypothesis, which relates AA origin to Southern regions of Yangtze River and their dispersal and spread as Expansion related to rice cultivation.

Thank you for raising this important point. We would like to respectfully clarify that our intention was not to infer the dispersal or origin of AA populations in South Asia, which indeed would require broader representation of South Asian AA groups.

Instead, our emphasis has consistently been on the potential gene flow between South Asia and Neolithic and later MSEA, rather than the reverse.

To clarify that, we have explicitly added the following sentence at lines 367 to 369:

“We note that we are not concerned here with the origin/dispersal of AA-speaking populations in South Asia, but rather we are interested in the impact of South Asian-related ancestry in MSEA.”

There is indeed a lot of literature that challenges this view and multiple models have been proposed. Indeed genetic analysis of S Asian AA as well as SE Asian AA have strongly proposed them as the original inhabitants of their respective regions. The presence of the indigenous Hoabinhian ancestry component among South Asian AA populations is not completely conforming with a neolithic expansion model around rice cultivation. The presence of the indigenous Hoabinhian ancestry in other present-day AA speakers such as Mon and Kharia may also suggest a widespread distribution of this ancestry in South Asia and MSEA, and not necessarily dispersal from MSEA.

I would suggest that the authors drop the South Asian component of the study and concentrate on delineating the details of AA populations in MESA. Else they may consider including more Indian AA populations may increase the robustness and generate more confidence on the connection between South and Southeast Asian populations. In that case they have to appropriately test the other alternative hypotheses of the AA population history and demography.

To further clarify on the main questions in our manuscript, at the beginning of the Discussion section we added at lines 586 to 589:

“However, as our dataset included only one SA-AA group, we do not consider the origin/dispersal of SA-AA groups, but rather we focused our analyses on the genetic structure and history of MSEA-AA groups.”

Considering the documented historical interactions between South Asia and MSEA, along with the detectable South Asian-related genetic ancestry in present-day MSEA AA populations, we investigated the potential genetic connections between prehistorical SA and MSEA using published ancient data, *instead of SA AA group.*

We have added at lines 509 to 512:

“The estimated presence of this deeply diverged ancestry in Neolithic individuals (ca. 4000 years ago) and Bronze-Iron Age individuals (ca. 1800-2200 years ago) suggests either a persistent genetic link between the South Asia lineage and ancient MSEA, multiple events of gene flow from SA to MSEA, or both.”

Since our analyses do not attempt to reconstruct the origins of South Asian AA groups, we would prefer to retain the newly reported genetic evidence for the connection between ancient South Asian and Southeast Asians.

The fine analysis of the AA within MESA is well done. I would think a little bit more detail of dispersal can be worked out from the IBD sharing data.

We have expanded the interpretation at lines 298–308 to better highlight patterns of interaction and potential dispersal among AA groups, as well as between AA groups and other populations.

“we found that AA speakers share more IBD segments with AN/TK/ST speakers than with HM speakers (Fig. 3).”

115
116 *“Similarly, we found that the differential sharing with neighboring AA branches*
117 *emerged in the length range of 3 to 5 cM”*

118
119 *“AN speakers (Malay, Ede, and Giarai), ST Karen speakers, and TK speakers*
120 *(Kalueang, Laolsan, Nyaw, and Phutai), showed more IBD sharing with AA speakers*
121 *in the range of 5 to 10 cM (Fig. 3c, 3e), indicating their genetic interactions with AA*
122 *groups remained active until recently.”*

123
124 A comparison with the findings of the following paper (Nature. 2025
125 Jul;643(8071):417-426. doi: 10.1038/s41586-025-08998-w. Epub 2025 May 14.) is
126 important.

127
128 We appreciate the reviewer’s suggestion and have incorporated a detailed
129 comparison with the findings of He et al. as follows:

130
131 1. Genetic heterogeneity and geographic structure: Both studies observed high
132 genetic heterogeneity across MSEA populations and a genetic structure aligned
133 with geography. While He et al. highlighted the clustering of genetic patterns
134 according to geographic regions, our study further expanded on this by exploring
135 the underlying driving forces. These include genetic drift due to small population
136 isolation and interactions among groups from different language families shaped by
137 geographical factors, and different gene flow from South and East Asia.

138
139 At line 589 to 598, we have now incorporated such comparisons:

140 *“Based on genetic clustering, previous MSEA genomic studies of populations from*
141 *Vietnam and Thailand suggested a good correlation between genetic structure and*
142 *language families^{14,20}, while a recent study with a broader sampling across MSEA*
143 *showed that geography played a more important role¹⁹. Here, with a denser*
144 *sampling of AA groups, we integrated modern and ancient genomes, as well as*
145 *haplotype-based approaches and allele-sharing statistics, to gain new insights from*
146 *a finer-scale and temporally resolved characterization of the genetic structure and*
147 *population history of MSEA AA groups and its implication to the overall genetic*
148 *structure and history of MSEA.”*

149
150 We also compared the methodology and evidence for the geographic-genetic
151 structure from different analytical approaches at lines 684–692:

152 *“To revisit the important role of geography in shaping the observed structure as*

suggested by the recent MSEA study using Procrustes statistics to compare the geographic and PCA locations¹⁹, we employed a more sophisticated method – MAPS (Methods) – which estimates the potential migration pattern quantitatively across space for the correlation between genetic structure and geographic distribution in MSEA. As shown, the patterns of population isolation and contact described above are further evident in the geographic pattern of effective population migration rates (Supplementary Fig. 10b).”

2. Language-based genetic structure: He et al. reported that most MSEA populations do not genetically cluster according to language families. In contrast, our results reveal that within AA populations, the genetic structure does correlate with their labeled linguistic subbranches, particularly among people associated to Eastern Mon-Khmer and Southern Monic language groups, while notable genetic heterogeneity exists among Northern Mon-Khmer groups.

At lines 600 to 603, we have added a more detailed explanation to clarify: “Our PCA analysis (Fig. 1d) revealed genetic heterogeneity among AA groups from MSEA that partially corresponds to their linguistic relationships and geographic locations. This suggests that the genetic makeup, language, and geography are interconnected in these populations to some extent.”

3. East Asian ancestry and ancient admixture models: we have incorporated ancient DNA data from MSEA, EA and SA, together with the newly generated modern data, which identified the temporal pattern of genomic formation on ancient MSEA. We wrote at line 468-472:

“Overall, we observed a declining trend of indigenous MSEA ancestry from the Neolithic to the Iron Age and Historical era in MSEA (Fig. 5), likely reflecting later population admixture with people from East Asia and South Asia⁴ as observed in modern-day MSEA populations.”

4. South Asian ancestry and deep ancestral links: He et al. identified shared ancestry between MSEA populations and Andamanese/Bengali groups based in the clustering pattern, implying some genetic connections.

While, we directly detected genetic contact between South Asia and modern MSEA AA populations, and further dated the admixture events to around 1,000 years ago.

Moreover, for the first time, in our study we identified South Asian ancestry in

Neolithic MSEA individuals, providing new evidence for deep genetic links among ancient Southeast Asians, South Asians, and East Asians.

In summary, despite some expected overlap between our findings and those of He et al., our study zoomed into the population history of MSEA AA speakers in particular, meanwhile revealed compelling evidence including the temporal and geographic pattern for the EA and SA population admixture and effective migration rates across ancient and modern MSEA populations.

In particular, we highlight the following key findings in our study:

- 1) fine-scale genetic structure decomposition within MSEA AA groups;
- 2) multiple admixture events and active intergroup interactions across AA speakers and other MSEA, SA, EA populations;
- 3) genetic continuity from Neolithic MSEA to present-day AA populations;
- 4) early South and East Asian influence on MSEA populations prior to the Neolithic.

Together, our findings contribute to a nuanced and comprehensive understanding of the consecutive population dynamics in MSEA from Neolithic to the present.

MINOR POINTS:

Strong IBD sharing between Mlabri and other Northern Mon-Khmer groups have been observed, despite substantial drift in Mlabri population in TREEMIX analysis along with distinct genetic ancestry in ADMIXTURE analysis. What might be the reasons for that?

Thank you for this insightful question. Despite the Mlabri have a small population size and an isolated demographic history (as reflected by the substantial genetic drift observed in both the ADMIXTURE and TREEMIX), the Mlabri is classified within the Northern Mon-Khmer linguistic branch. The strong IBD sharing observed between the Mlabri and other Northern Mon-Khmer groups likely reflects the *rather recent genetic contact* within the population associated with the Northern Mon-Khmer linguistic branch (Oota, H. et al. 2005; Rischel, J. 1995).

Oota, H. et al. (2005). *Recent Origin and Cultural Reversion of a Hunter–Gatherer Group*. *PLOS Biol.* 3, e71.

Rischel, J. (1995). *Minor Mlabri: A Hunter-Gatherer Language of Northern Indochina*.
(Museum Tusculanum, Copenhagen).

We have reworded the related sentence on this phenomenon at lines 666 to 670:

(previous phrasing)

*“Despite the extreme bottleneck, intriguingly, the genetic relationship between the
Mlabri and the HtinMal and HtinPray is still evident in the IBD sharing analysis
(Fig.3), supporting linguistic and historical sources that state that the Mlabri are
related to the Htin⁵⁰”*

(current phrasing)

*“Despite the extreme bottleneck, intriguingly, the genetic relationship between the
Mlabri and the HtinMal and HtinPray is still evident in the IBD sharing analysis (Fig.3,
Supplementary Fig. 8), supporting a recent origin of the Mlabri from these Northern
Mon-Khmer speakers (or a related group) as indicated by genetic, linguistic and
historical sources^{46,47}”*

Most Khmuic populations (except Khmu) do not exhibit increase in effective
population sizes around 125-150 generations ago, that has been connected to the
hypothesized rice agricultural dispersal around 4 kya, and rather show a declining
trend. The pattern of increasing effective population size in Khmuic populations is
mostly driven by Khmu, which has substantially higher sample size than all other
Khmuic populations. Can you explain why did the effective population sizes of these
AA populations did not increase even with the dispersal that likely introduced AA
languages in Southeast Asia?

Thank you for noting this point. Indeed, the pattern of increasing effective population
size among Khmuic populations is primarily driven by the Khmu group, which has a
substantially larger sample size. The other Khmuic-speaking groups, i.e. the HtinMal,
HtinPray, Lua, and Mlabri, did not present a similar expansion around 125–150
generations ago but exhibited the highest levels of within-group IBD sharing, implying
recent genetic bottlenecks that could obscure the genetic evidence of previous
population expansions.

We think other Khmuic population did not experience population expansion likely due
to geographic isolation as most of them are hill tribes, as evidenced by MAPS (used
for estimating effective migration rates across space). Line 330-339, we wrote:

"It appears that the mountainous region of northwestern and eastern coastal MSEA has consistently exhibited lower dispersal rates from at least around 2600 years ago (Supplementary Fig. 10). These dispersal rates seem to be aligned with differentiated genetic structure between highland (i.e. those located in the mountainous regions of the northwest MSEA, at altitudes exceeding 500 meters above sea level, such as Northern Mon-Khmer groups) and lowland populations (i.e. those located in the southern plain regions, at altitudes less than 500 meters above sea level, such as Monic, Southern Monic, and Eastern Mon-Khmer groups) (Supplementary Fig. 10a, Fig. 4)."

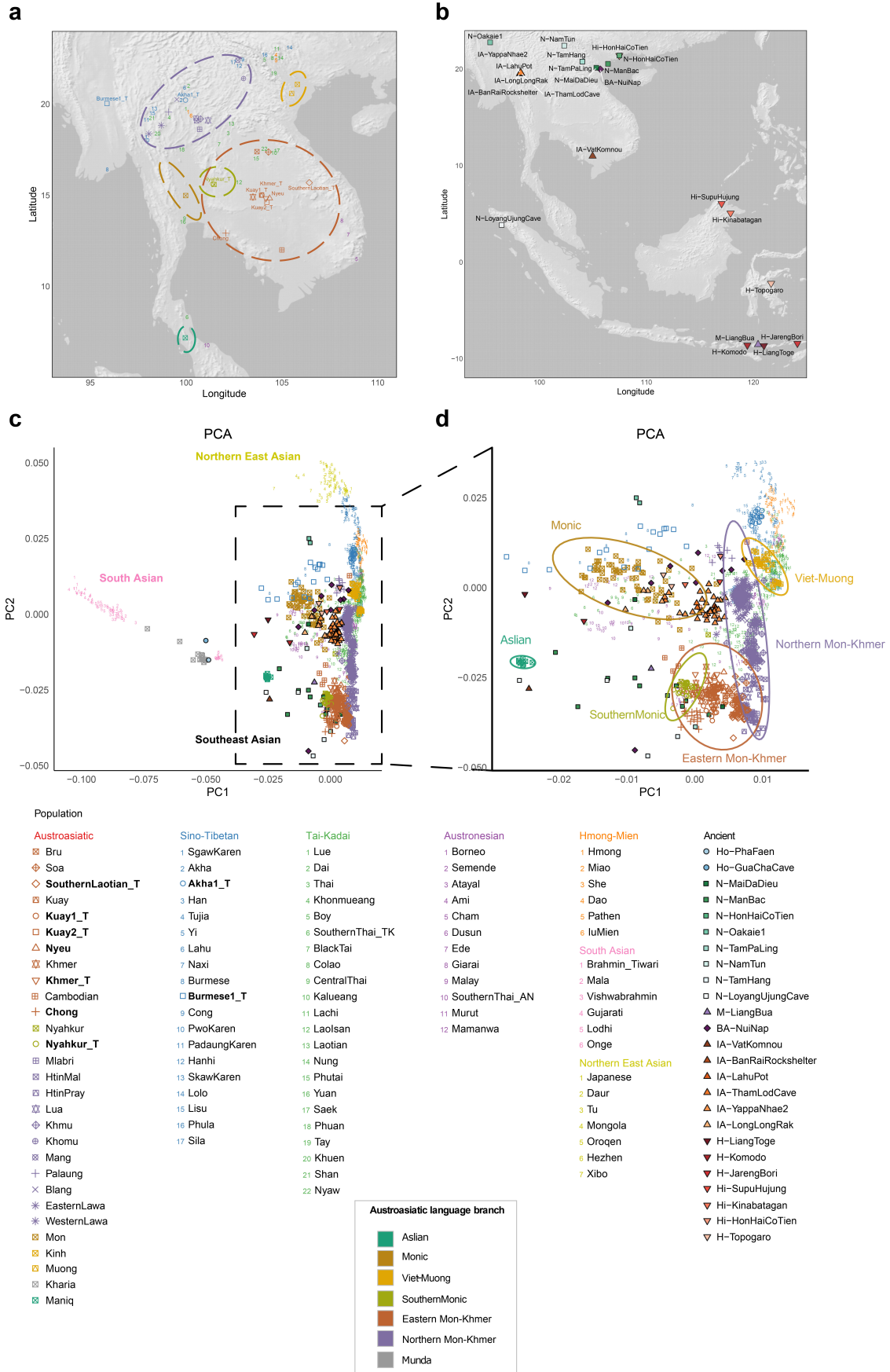
Meanwhile, we have openly discussed about the potential caveat of population size estimation from software HapNe-IBD in our main text at lines 1026 to 1029:

"However, the absolute estimates of intensity and timing from hapNe may be biased by population demographic history, and the exact timing of expansions or bottlenecks should be interpreted with caution."

The quality of Fig1 is really bad. However, in PCA Fig 1c, Onge (Andamanese speakers) individuals are positioned in between the individuals from Kharia (Indian Austroasiatic) and Maniq (Southeast Asian Austroasiatic) populations. Did you check why Onge, despite belonging to a different linguistic family, occupies this position in PCA?

Thank you for your careful review. The Onge individuals were projected to avoid strong drift distortion of plotting. Alternatively, their position may also reflect the deep genetic connection we observed between South Asians and the Hoabinhian individual, who was projected to a similar position as the Onge.

We have improved the quality of figure from 300ppi to 900 ppi in the revised manuscript. Here we attach the high-resolution figure as follows:



298

299

Reviewer #2 (Remarks to the Author):

Review of “New insights into the genomic population structure and history of Austroasiatic speakers in Mainland Southeast Asia” by Yin et al.

Summary

This study investigates the origins, genetic structure, and historical interactions of Austroasiatic (AA) language speakers in Mainland Southeast Asia (MSEA), using the largest AA dataset to date that includes newly generated genome-wide data from seven AA and two Sino-Tibetan populations in Thailand, Laos, and Myanmar, alongside ancient DNA. The key questions addressed include when and where AA groups formed, how their genetic patterns align with their linguistic affiliations, and how historical contacts with East Asian and South Asian populations shaped their genomes. The findings reveal that AA groups exhibit significant genetic heterogeneity that broadly aligns with linguistic branches and geography, with northern highland groups showing greater East Asian ancestry and links to Iron Age populations, while southern lowland groups display additional South Asian ancestry and affinities with Neolithic populations. Surprisingly, South Asian-related ancestry was also detected in Neolithic MSEA samples, suggesting the presence of an unsampled, deeply diverged lineage shared across ancient South, East, and Southeast Asia. The study confirms that AA is likely the oldest linguistic and genetic layer in the region, predating other language families such as Sino-Tibetan and Tai-Kadai as AA speakers show genetic cladality with Neolithic samples while other language families show cladality only with later samples from Iron age and Historical era.

Overall impression of the work

The overall impression of this paper is that it is a thorough, well-written and well-executed study. It stands out for its impressively wide range of methods and brings valuable clarity to a complex topic. We liked that the introduction was very clear and to the point.

We appreciate such positive feedback from reviewer! We aim to present a compelling and well-structured scientific study on this complex topic with a clear introduction and a wide range of cutting-edge methods, making it accessible to both the general public and experts in the field.

1. The manuscript is a substantial and impressive body of work, but in its current form it can feel overwhelming due to the number and density of analyses. As a result, it becomes difficult to follow the main narrative and identify the key findings. To help the central results stand out more clearly, we suggest streamlining the main text -

possibly by moving some of the more technical analyses to the supplementary materials.

We are aware that the presented results in the main text might be too extensive. Hence following your suggestion, we have now shortened the following sections:

Lines 274–314:

We have shortened the IBD analysis section by moving some information to the Supplementary Information, retaining only the key conclusions in the main text.

Lines 452–454:

We have condensed the ancestry modelling strategy for modeling ancient MSEA populations, and moved some of the details to the Methods and Supplementary Information section.

In particular, the Refined IBD section (“Intra-population and Inter-regional structure among AA speakers and MSEA populations”) could benefit from a more concise presentation. Currently, the section includes an extensive description of the total amount and number of IBD blocks shared between populations. While informative, this level of detail may not be necessary for supporting the main conclusions. Especially when considering the section about overall length and count of shared IBD segments between population pairs (that can be influenced by several different population historical events), not the average segment lengths, which are often more informative for interpreting population contact time depth.

We have previously presented the average IBD segment lengths (normalized fraction) in Supplementary Fig. 8. As suggested by the reviewer, we have replaced Fig. 3 (the total amount and number of IBD blocks shared between populations) in the main text with the previous Supplementary Fig. 8, which presents the normalized fraction of IBD sharing across different segment length ranges, as this measure can also reflect the time depth of population contact.

Below is our updated Fig. 3:

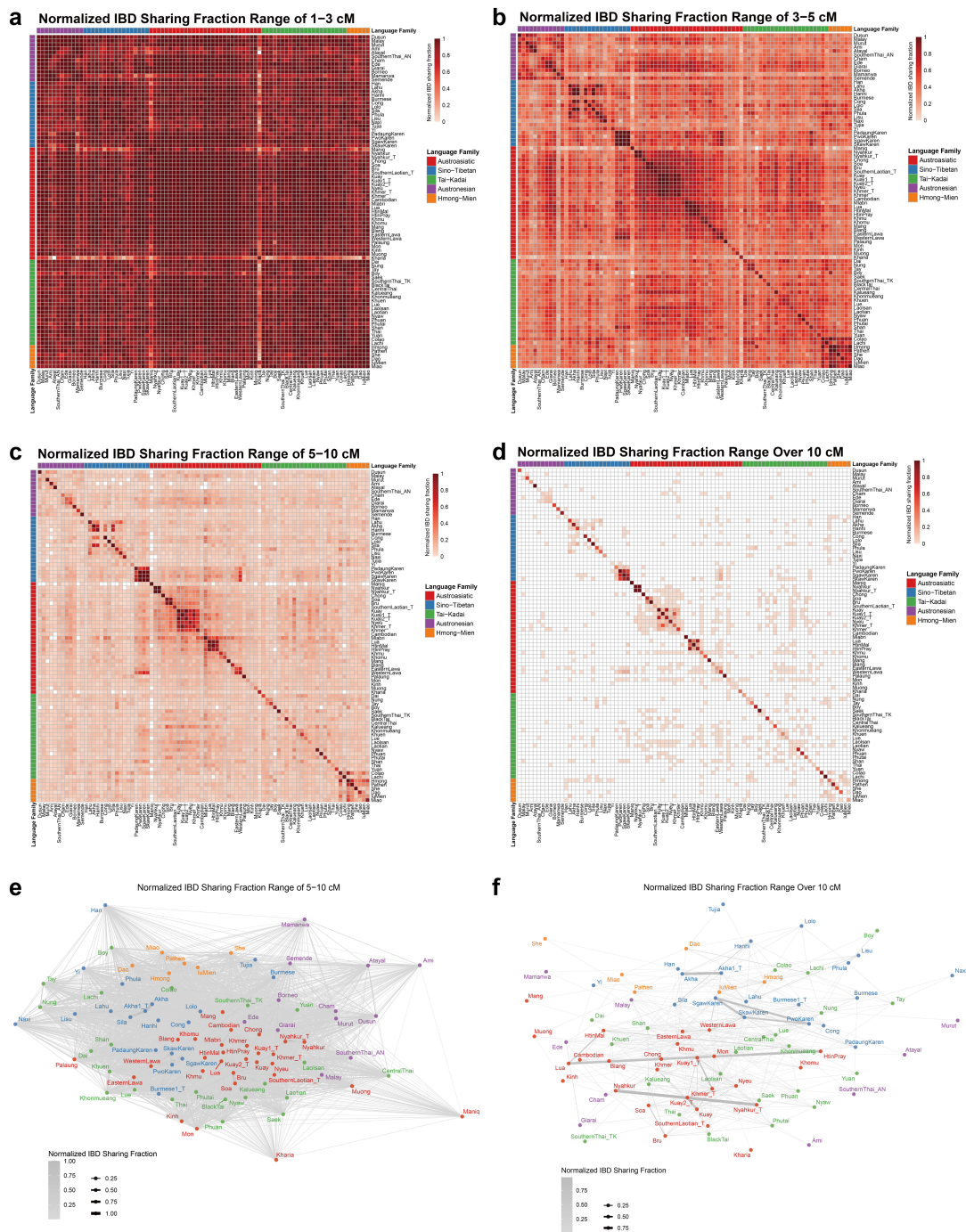


Fig. 3 | The heatmap and network visualization of IBD sharing across SEA, EA

and SA populations. The identified IBD blocks were categorized into the range of 1

to 3 cM, 3 to 5 cM, 5 to 10 cM and over 10 cM, in SEA, SA and EA populations. **a-d.**

Corresponding heatmaps for each segment length category. For each pair of

populations, we plot the fraction of all pairs of individuals that share at least one IBD

block in the size range, which we obtained by dividing the total number of pairs that

share such IBD segments by the total number of all possible pairs. **e-f.** Network

visualizations for the 5 to 10 cM and over 10 cM categories. In these networks,

populations are clustered based on their normalized IBD sharing fraction; populations

with higher IBD sharing are located closer to each other.

Accordingly, the text at lines 299–308 has also been revised to the following to more informatively interpret population contact time depth:

"We found that AA speakers share more IBD segments with AN/TK/ST speakers than with HM speakers. Notably, the differential sharing between AA speakers and neighboring AN/TK/ST speakers appeared to be prominent in the length range of 3 to 5 cM (Fig. 3). Similarly, we found that the differential sharing with neighboring AA branches emerged in the length range of 3 to 5 cM (Supplementary Fig. 9). Taken together, the interaction between neighboring AA branches or other language family groups that has played a key role in driving the structure of AA groups dates back to at least 1160 years ago. Furthermore, AN groups Malay, Ede, and Giarai, ST Karen groups, and TK groups Kalueang, Laolsan, Nyaw, and Phutai, showed more IBD sharing with AA speakers in the 5 to 10 cM range (Fig. 3g), indicating that their genetic interactions with AA groups remained active until recently."

The figure and interpretation on the total amount and number of IBD blocks shared between populations have been moved to the Supplementary Information.

Similarly, the 5th chapter of results ("The genomic formation of ancient MSEA and legacy in present-day AA speakers") is quite long. For the sake of clarity, we suggest dissecting this chapter into two chapters, e.g. "The genomic formation of ancient MSEA" and "Legacy in present-day AA speakers".

We agreed, and we have now rephased the 5th chapter into the 5th chapter "The genomic formation of ancient MSEA" and the 6th chapter "The genetic legacy in present-day AA speakers" chapters in the main text as suggested.

2. We suggest more care in wordings related to interpretations of language group movements based on genetic data. See for example specific comment nr 1.

We have revised the descriptions accordingly. For example, we use "AA speakers", "AA-speaking populations" instead of "AA populations", and we use "people associated to the ST language family" "the Katuic-speaking Kuay and Nyeu" "the Khmeric-speaking Khmer" "Katuic speakers" instead of "populations"

Here are some more of the changes we have made in the text:

Lines 51 to 54: *"The Austroasiatic (AA) language family is supported by multidisciplinary evidence as the earliest known linguistic presence in Mainland Southeast Asia (MSEA) dating back to the Neolithic period, likely preceding the presence of other language families like Sino-Tibetan (ST)."*

Lines 77 to 78: *"Among these, AA is likely one of the oldest surviving language families¹"*

Lines 107 to 109: *"These dynamic population movements associated with multiple language families into MSEA suggest a complicated picture of AA-speaking population diversification since the Neolithic."*

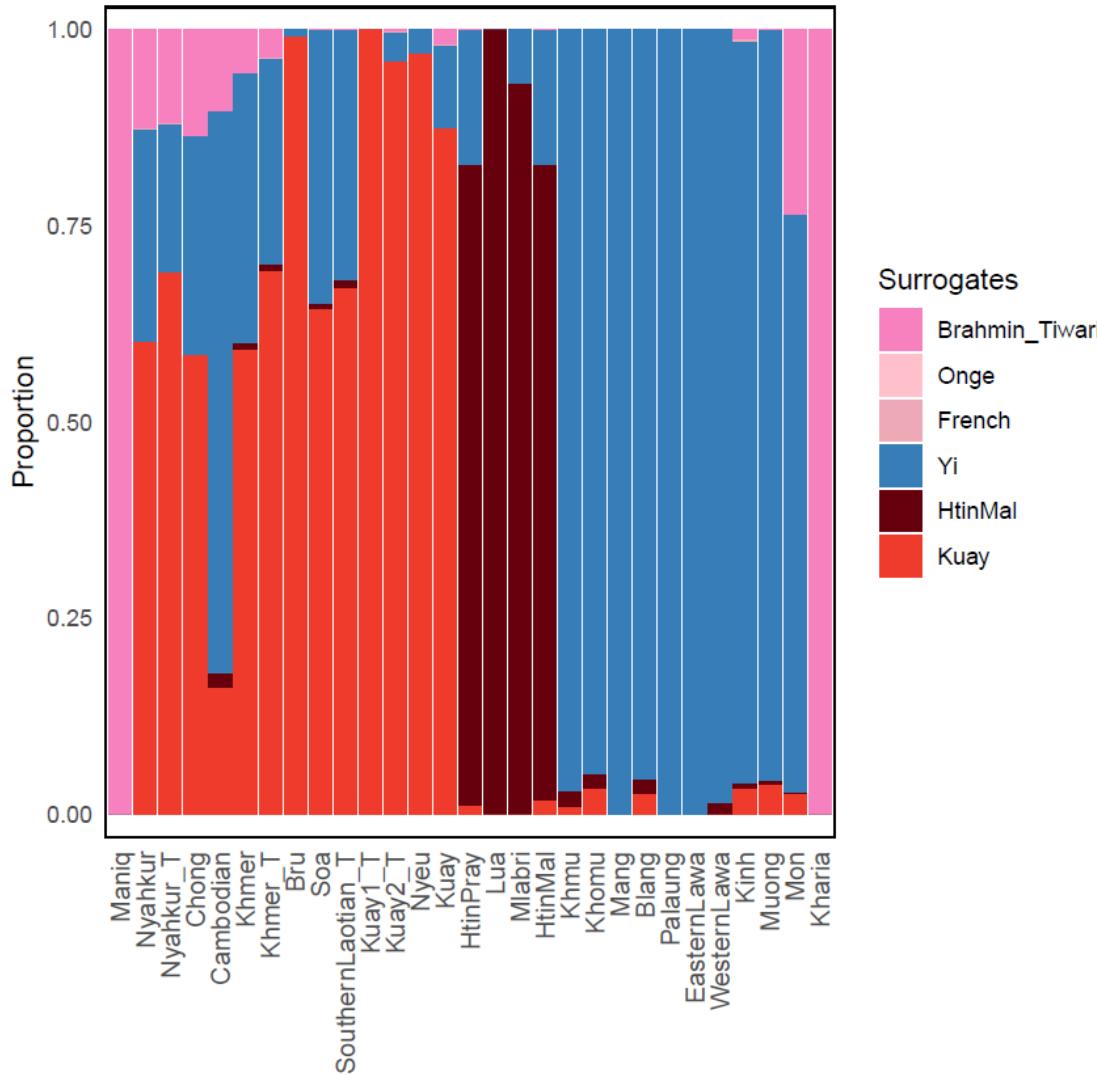
Lines 119 to 121: *"To what extent does the genetic heterogeneity of MSEA AA groups reflect the historical diversification of AA groups versus differential genetic contact with other MSEA language groups?"*

Line 442: *"As AA languages are hypothesized to be the oldest language family in MSEA"*

3. One of the key topics in the paper is the (early) presence of SA admixture in MSEA. In this regard it would seem valuable to have more than one source from SA in analyses like SOURCEFIND. This would enable to disentangle genetic contributions from SA to MSEA before and after the Steppe migration. Yes, all SA groups today are mixtures of ASI and ANI but different proportions of those in different sources might help shed more light on the SA admixture into MSEA. Also including Onge as one source would make sense as they don't have ANI. Furthermore, it may be a good idea to include a population with Steppe ancestry but no ANI or ASI ancestry. Basically, some European population as a population from Central Asia would have East Asian ancestry, which would also confuse the picture. Currently the fact that Kharia draws all ancestry from a Brahmin group seems strange when AA sources as alternatives are also available.

Following the suggestion, we reran the SOURCEFIND analysis with Onge and French as additional surrogates for ASI- and Steppe-related ancestry, respectively. As shown below, neither contributed significantly to the ancestry of AA groups. Kharia remains represented completely by Brahmin_Tiwari, and Maniq is also represented as 100% Brahmin_Tiwari. Compared to the previous analysis, some southern AA

groups show increased South and East Asian ancestry. Several factors may explain the full Brahmin_Tiwari assignment in Kharia and Maniq: 1). ASI- or Steppe-related ancestry may be too subtle to detect and treated as background noise by SOURCEFIND. 2). Both Onge and Maniq are highly drifted populations, which may prevent Onge from being assigned as a source and lead to misassignment of Maniq. These results are likely constrained by limited surrogate availability, the low proportions of detectable ancestry, and the denoising strategy of SOURCEFIND, which may reduce resolution in distinguishing ANI and ASI ancestry in modern AA groups. Therefore, we propose to retain the original SOURCEFIND results.



Specific comments

Abstract:

1) 50 “The Austroasiatic (AA) language family represents the earliest known linguistic presence in Mainland Southeast Asia (MSEA) dating back to the Neolithic

period.”. – What is the direct evidence of AA language family in the Neolithic period?
What is probably meant here is that people speaking AA languages display genetic
component(s) that are detectable in the region already in the Neolithic period.

For the direct evidence of AA language family in the Neolithic period, linguistically, we
meant that previous linguistic studies based on Bayesian phylogenetic analysis has
suggested that the emergence and initial diversification of the AA language family
occurred approximately 7 kya and 4 kya respectively (Van Driem 2007)

*Van Driem, G. (2007). Austroasiatic phylogeny and the Austroasiatic homeland in
light of recent population genetic studies. Mon-Khmer Studies, 37, 1-14.*

The hypothesized linguistic expansion time period aligns well with the Neolithic
period of MSEA according to archaeological evidence (Higham et al 1998).

*Higham, C. F. W. Archaeology, Linguistics and the Expansion of the East and
Southeast Asian Neolithic. in Archaeology and Language II (Routledge, 1998).*

Second, genetically, we newly discovered the genetic similarities between Neolithic
MSEA individuals and present-day MSEA AA speaking populations. Such finding
adds direct genetic evidence for the hypothesis that people speaking AA languages
display genetic component(s) that are detectable in the region already in the
Neolithic period.

To be more rigorous, we have reworded the related sentence to the following:
“The Austroasiatic (AA) language family is supported by multidisciplinary evidence as
the earliest known linguistic presence in Mainland Southeast Asia (MSEA) dating
back to the Neolithic period,”

Introduction:

2) 77 “AA is thought to be the most ancient language family in MSEA” – add
reference here

We have added a reference to support this statement.

Reference:

*Sidwell, P., & Blench, R. (2011). The Austroasiatic Urheimat: the Southeastern
Riverine Hypothesis. In E. N. J. (Ed.), Dynamics of human diversity : the case of*

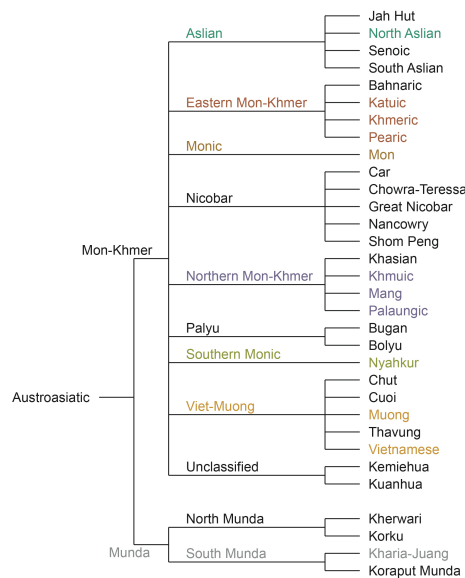
mainland Southeast Asia (1st ed., Vol. 1, pp. 315-343). *Pacific Linguistics*.

3) 78 “AA speakers are primarily distributed across mainland MSEA” - mainland 2x in a row in this sentence

Thank you for the reminder. We have removed the redundant "mainland".

4) “The AA language family comprises two primary branches - Mon-Khmer and Munda – and Southern Monic, Viet-Muong, Monic and Aslian branches.” This sentence under Sup. Figure 1 is a bit confusing as this division of languages doesn’t stand out from the figure depicting AA language phylogeny. Therefore, it seems like the figure has certain branches and the caption underneath describes different branches.

Thank you for the suggestion. We have revised the Supplementary Figure 1 caption to the following to align with the language phylogeny structure in the figure.



“Supplementary Figure 1. Linguistic classification of the Austroasiatic language family in *Ethnologue*¹. The AA language family comprises two primary branches - Mon-Khmer and Munda. We analysed speakers from the Aslian, Eastern Mon-Khmer, Northern Mon-Khmer, Monic, Southern Monic, Viet-Muong subbranches of the Mon-Khmer branch, as well as the South Munda subbranch of the Munda branch, which are highlighted in colored text in the phylogeny tree.”

5 chapters of results:

I. Genetic heterogeneity within modern Austroasiatic speakers in MSEA

II. Relationships between modern AA groups and ancient samples

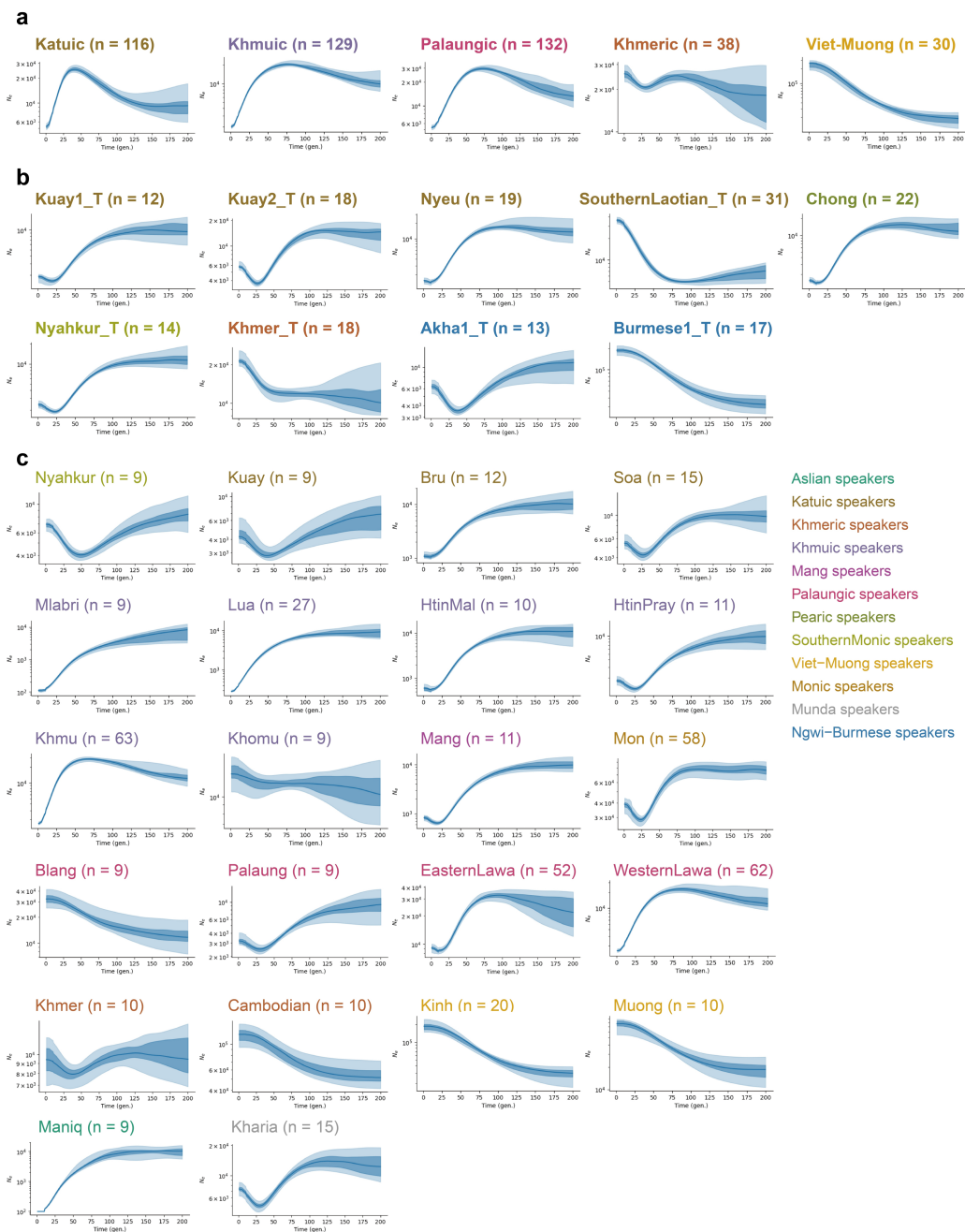
5) 251 “It is likely that this reflects a mismodeled case in which an ancestral population, which has affinities with several descendant populations that are highly drifted and hence highly differentiated, is inferred to be a mixture of the descendant ancestries, rather than as directly ancestral to the descendant ancestries.” -Cool that this textbook example happened here and you were able to interpret it.

Thank you! We appreciate your positive evaluation.

III. Intra-population and Inter-regional structure among AA speakers and MSEA populations

6) Supplementary Figure 5 has lots of little plots. Unfortunately, the numbers on their scales are not easily readable. They can be guessed, but it doesn't help that the scales differ between figures, which makes it especially hard to compare the already blurry scales with each other.

We have rearranged the figure layout and increased the resolution to 900 ppi to ensure that the scale numbers are clearer and easier to read.



IV. South Asian and East Asian admixture in modern AA-speaking groups

7) 452 “the Munda Kharia was completely surrogated by SA ancestry.” – I see a tiny bit of Kuay ancestry in Kharia on Fig. 4a and Sup. Fig. 11. but yes, it is very suspicious that Kharia is almost completely Brahmin_Tiwari.

As noted in our previous response regarding the SOURCEFIND analysis with additional surrogates representing ANI and ASI ancestries, Kharia still appears to be entirely represented by Brahmin_Tiwari. This may be attributed to SOURCEFIND’s denoising algorithm, which tends to remove subtle signals and treat them as background noise, or possibly because Kharia share more detectable ancestry with

Brahmin_Tiwari than with the available AA surrogates—an observation also supported by the ChromoPainter results (Supplementary Fig. 7).

8) 470 “Focusing on geographical locations (Fig. 4a), we found that populations shifted towards EA populations are all from northern MSEA, while those exhibiting shared genetic drift with SA populations are situated in southern MSEA.” – this is interesting

Thank you! We are glad you find it intriguing.

9) 515 “when using Karen as the proximal surrogate for EA ancestry, the Karen admixture date in both Lawa groups is rather recent,” –Karen is actually AA with language shift to ST and is therefore not a suitable proxy for EA ancestry (Mishra, R. K., Singh, P. P., Rai, N., Desai, S., Pandey, P., Tiwary, S. K., Tamang, R., Suravajhala, P., Shrivastava, P., Thangaraj, K., van Driem, G., & Chaubey, G. (2024). Reconstructing the population history of the Nicobarese. *European Journal of Human Genetics*, 1–7. <https://doi.org/10.1038/s41431-024-01720-w>)

Thank you for your suggestion. Accordingly, we have revised our interpretation: instead of interpret the signals as the Lawa groups received their EA ancestry directly from the Karen, we now interpret the signals as reflecting recent interactions between the two groups.

To investigate the exact source of EA ancestry in the Lawa, we actually used ancient EA population (Nepal_Chokhopani_2800BP.SG and China_YR_MN) as potential sources in the DATES analysis, and the estimated admixture dates fall around 1,000 years ago for Eastern Lawa and 1,200 to 1,400 years ago for Western Lawa.

We have revised the related sentences to the following:

Line 429 to 439:

“To account for the potential proximal sources transmitting the EA ancestry to the two Lawa groups, we tested alternative proxies based on demographic history³¹ by employing Karen as the proximal surrogate for the retained EA ancestry (Mlabri as the other surrogate). In such test, the estimation from fastGLOBETROTTER suggested rather recent contact at around 300 years ago (Supplementary Table 2). Alternatively, we employed DATES using ancient EA sources for admixture time estimates, which for the two Lawa groups fell between 1000 and 1400 years ago,

aligning with the estimated admixture time for SgawKaren (~1288 years ago), indicating that the EA ancestry in Lawa groups may reflect contact with SgawKaren (Supplementary Table 2)."

V. The genomic formation of ancient MSEA and legacy in present-day AA speakers
10) 541 "exact" -should probably be "extract"

Thanks. We have now corrected the spelling.

11) 598 "Overall, we observed a declining trend of indigenous MSEA ancestry from the Neolithic to the Iron Age and Historical era in MSEA (Fig. 5), likely reflecting later population admixture with people from East Asia and South Asia as observed in modern-day MSEA populations." – cool that the trend is visible with the amount of ancient samples

Thank you. It is also encouraging for us that the trend is clear even with the currently available ancient samples. We hope future thorough sampling would further refine this pattern.

12) 636 "The estimated presence of this deeply diverged ancestry in Neolithic individuals (ca. 4000 years ago) and Bronze-Iron Age individuals (ca. 1800-2200 years ago) suggests either a persistent genetic link between the South Asia lineage and ancient MSEA, multiple prehistorical mixing events, or both." – Intriguing. Then the SA-AA admixture event(s) didn't necessarily happen in SA. For example, in the Indian AA speakers, some of the SA ancestry could be from earlier admixture that happened in MSEA and that could change the implication of admixture time = time of migration to SA.

*Thank you for raising this point. It is indeed possible that some admixture between ancient SA and MSEA populations occurred prior to the migration of AA groups into SA. However, our intention was *not* to suggest gene flow from ancient MSEA into ancient SA populations, but rather the reverse—gene flow from SA into MSEA. We have revised the text to clarify at lines 508 to 512:*

"The estimated presence of this deeply diverged ancestry in Neolithic individuals (ca. 4000 years ago) and Bronze-Iron Age individuals (ca. 1800-2200 years ago) suggests either a persistent genetic link between the South Asia lineage and ancient MSEA, multiple events of gene flow from SA to MSEA, or both."

Discussion

13) 861 “We found that only modern MSEA-AA groups show cladality with ancient samples from the Neolithic; specifically, the Eastern Mon-Khmer and Southern Monic lowland groups. In contrast, all other language groups show cladality only with samples from the Iron Age or Historical era periods. This supports the general consensus that AA is the most ancient language family in MSEA. Intriguingly, we identified a signal of ~20% ancient SA ancestry (using Indus_Periphery as a proxy, Methods) in the late Neolithic samples MaiDaDieu, ManBac, and Oakaie (Fig. 5), all dating to ~3-4 kya, which has not been reported before.” - Interesting!

[Thank you! We appreciate your interests in this result.](#)

Reviewer #3 (Remarks to the Author):

This manuscript "New insights into the genomic population structure and history of Austroasiatic speakers in Mainland Southeast Asia" by Zi Yin et al. presents an important and well-executed genomic analysis of Austroasiatic-speaking populations in Mainland Southeast Asia. The authors combine dense genome-wide SNP data with a range of analytical approaches (PCA, ADMIXTURE, DyStruct, f-statistics, IBD, and qpGraph) to investigate both deep population structure and recent demographic histories.

The results contribute meaningfully to our understanding of linguistic, cultural, and genetic diversity in the region. In particular, the fine-scale resolution of population relationships among Austroasiatic speakers is novel and highly valuable. The manuscript is data-rich, methodologically up-to-date, and mostly clear in its exposition.

That said, there are several areas where improvement would enhance the clarity and accessibility of the paper, both scientifically and linguistically. I list below the different points that should be improved:

[We sincerely appreciate your thoughtful and constructive comments, and have revised the text accordingly. See examples of revised text listed below.](#)

On the main text

1 Clarity of Language and Style

The manuscript occasionally uses passive voice and awkward constructions that

obscure the meaning or weaken the narrative. Examples include:

- “Is thought to be among the oldest extant language families” (page 1, line 7) → could be simplified to “likely one of the oldest surviving language families.”

We failed to find this sentence unfortunately.

- “IBD segments >3cM were used...” (page 4, line 9) → rewrite as “We used IBD segments longer than 3cM...”
- “Admixture graph topologies were modeled...” (page 8, line 17) → rewrite as “We modeled admixture graph topologies using...”

We failed to find these sentences unfortunately.

For the mentioned “IBD segments >3cM”, we have now carefully revised throughout the text for the usage of “>”, and revised the related words to “We used IBD segments longer than”

2 A full proofreading pass for stylistic consistency would strengthen the manuscript.

Thank you for the suggestion. We have carefully proofread the manuscript and revised several expressions to improve stylistic consistency. For example:

Line 141: “*Austroasiatic-speaking groups*” to “*AA-speaking groups*”.

Line 143: “*Non-Austroasiatic*” to “*Non-AA*”.

Line 164: “*Austroasiatic-speaking groups*” to “*AA speakers*”

3 Interpretation of DyStruct vs. ADMIXTURE

While both methods are presented, the paper would benefit from a more explicit comparison between the two, including how different values of K were selected and how the outputs agree or diverge.

Thanks for your advice. We have revised the relevant sections about the interpretations of the two analyses at lines 233 to 235:

“We further investigated this clustering pattern by DyStruct, which is an ADMIXTURE-like method utilizing the temporal information of ancient populations”

In the previous round of submission, we have included our rationale for K selection in

the supplementary figure legend: “*DyStruct analysis was conducted using 142,139 SNPs. a DYSTRUCT analysis conducted across modern populations from Austroasiatic (AA) and ancient individuals, ranging from $K = 2$ to $K = 15$; b The plot demonstrates the Log likelihood, revealing three prominent peaks in Log likelihood at $K = 5$, $K = 9$, and $K = 12$.*”

4 f₄-Statistics and qpGraph Explanation

The rationale behind specific f₄ tests and graph models is clear to specialists, but a brief paragraph explaining the biological implications of these statistics for a broader audience would improve accessibility. Also, residuals and Z-scores from alternative graph models should be included, even in supplementary files.

Thank you for the suggestion. We have now included a brief explanation of the biological implications of the f₄ statistics to aid accessibility for a broader audience.

At lines 385 to 389:

“We also carried out allele sharing f₄(W, Maniq; SA/EA, Mbuti) tests for comparing the genetic affinity of AA groups to SA and EA populations (Fig. 4b). Assuming the African Mbuti is an outgroup, a significantly positive f₄ value in this context indicates that, relative to Maniq, population W is genetically closer to the respective SA or EA group.”

In the caption of Supplementary Fig. 15-17, we have provided the residuals and Z-scores for all alternative graph models

Minor Points and Edits

- Define acronyms such as IBD, PCA, and SNP at first use.

In the *previous round* of submission, 1) we defined IBD in the chapter of “Intra-population and Inter-regional structure among AA speakers and MSEA populations”: “we analyzed relationships based on shared Identical by-Descent (IBD) segments.” 2) and we defined PCA in Figure 1 caption: “Geographic distribution of the newly studied groups in MSEA and Principal Component Analysis (PCA)”.

After examining acronyms throughout the text, we revised a sentence in Figure 1 caption to spell out acronyms of LD and SNP: “PCA was performed using linkage disequilibrium (LD) pruned data, comprising 195,752 unlinked single nucleotide polymorphisms (SNPs) across the whole genome (Methods).”

756

757 • The phrase “in agreement with historical records” is repeated several times and
 758 could be rephrased for variation (e.g., “consistent with historical data”).

759

760 We failed to find such phrase in our submitted documents to *Communications*
 761 *Biology*.

762

763 • Clarify whether ancient DNA (aDNA) was considered or excluded, and if so, why.

764

765 In the IBD analyses, we did not consider ancient DNA for the reason that its quality
 766 was insufficient for reliable haplotype-based analysis.

767

768 We have added clarification regarding the exclusion of ancient populations from the
 769 IBD analysis, which was due to the limited data quality, at lines 275 to 278:

770

771 “we analyzed relationships based on the shared Identical-by-Descent (IBD)
 772 segments in modern populations (as the data quality of available ancient MSEA
 773 samples was insufficient for reliable IBD detection) (Supplementary Information)”

774

775 In other methodologies used in the manuscript, we have provided related explanation
 776 regarding the consideration of ancient data. For example, at line 327-330:

777

778 “We estimated the population dispersal rate surface across the landscape by the
 779 software MAPS, using genome-wide data from all modern-day MSEA populations”

780

781 • Add a brief statement on ethical approvals for sampling and data use.

782

783 In the previous round of submission, the statement of ethical approvals for sampling
 784 and data use was included in the Method section (subsection “Samples and genomic
 785 data”).

786

787 We have now separated the ethical statement into a separate subsection to make it
 788 more visible.

789

790 **“Ethic Statement**

791 *Ethical approval for this study was granted by the Khon Kaen University Ethics*
 792 *Committee (HE622027) and the Institutional Review Board of Naresuan University*
 793 *(P1-0156/2566).”*

794

795 On the Supplementary Materials

796 The supplementary materials are rich and well-structured. A few improvements are

797 suggested:

798 1 Methodological Descriptions: The IBD-based demographic inference would benefit

799 from clearer mention of generation time assumptions and phasing quality.

800

801 In the previous round of information, the generation time has been clearly defined as

802 29 years per generation with reference to *Fenner et al 2005* (at line 296 in the

803 Results section, and at line 887 in the Method section). Regarding phasing quality,

804 we have clearly written in the Method section that

805 “we used Hap-IBD v.1.076 following the recommended data phasing pipeline using

806 Beagle1067 v.5.4 with GrCh38 genetic map” with reference to the HapIBD

807 publication where the phasing strategy and quality were originally described.

808

809 In response to reviewer’s suggestion, we have now added the generation information

810 to the caption of Supplementary Fig. 10 as well.

811

812 “The indicated time was estimated based on the length of IBD segments, assuming a

813 generation time of 29 years.”

814

815 2 ADMIXTURE and DyStruct: Include a comparative summary of K selection and

816 interpretation.

817

818 In the previous round of submission, such information was reported in

819 Supplementary Fig. 3 and Fig 4.

820

821 3 qpGraph Topologies: Provide Z-scores and residuals for alternative models.

822

823 In the previous round of submission, we have provided the residuals and Z-scores for

824 all reported graph models in the caption of Supplementary Fig. 15-17.

825

826 4 Metadata: Clarify dialectal or subpopulation distinctions in Table S1.

827

828 In the previous round of submission, we have included the language family and

829 branch information (suppose that is the dialectal information the reviewer meant).

830

831 In response to reviewer’s suggestion, we have newly added information on the

language subgroups in Table S1.

5 Figures: Improve legibility of S6–S10 (axis clarity and contrast).

As requested, we have updated the figure quality for improved clarity.

6 Ethics: Add fieldwork permit and ethical clearance info, even in supplement. Or say where they are deposited.

In the previous round of submission, we have made an ethical statement in the Method section (currently at line 826-829).

In conclusion, this is a strong and significant contribution to the field of population genetics and Southeast Asian prehistory. With modest revisions to improve clarity, stylistic consistency, and methodological transparency, the manuscript will be ready for publication.

Thanks again for highlighting the significance of our study and all constructive suggestions!

Reviewer #4 (Remarks to the Author):

I co-reviewed this manuscript with Reviewer #2, who provided the listed reports. This is part of the Communications Biology initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Thanks.

Rebuttal Letter

Regarding the suggestions from the reviewers, find our point-by-point responses below.

Reviewer #1 (Remarks to the Author):

The revised manuscript looks satisfactory to me.

We are thankful for the positive assessment and are pleased that the revised manuscript is satisfactory.

Reviewer #2 (Remarks to the Author):

The authors have addressed all our comments. We are happy to see this manuscript accepted and printed.

We are grateful for the supportive remarks and are delighted that the revisions have addressed all previous concerns.

Reviewer #3 (Remarks to the Author):

This manuscript "New insights into the genomic population structure and history of Austroasiatic speakers in Mainland Southeast Asia" by Zi Yin et al. presents an important and well-executed genomic analysis of Austroasiatic-speaking populations in Mainland Southeast Asia. The authors combine dense genome-wide SNP data with a range of analytical approaches (PCA, ADMIXTURE, DyStruct, f-statistics, IBD, and qpGraph) to investigate both deep population structure and recent demographic histories.

The results contribute meaningfully to our understanding of linguistic, cultural, and genetic diversity in the region. In particular, the fine-scale resolution of population relationships among Austroasiatic speakers is novel and highly valuable. The manuscript is data-rich, methodologically up-to-date, and clear in its exposition.

I understand that parts of my first-round review caused confusion. In several places, I placed paraphrased methodological or interpretive points in quotation marks, which may have given the impression that I was citing your text verbatim. This was not my intention — my aim was only to formulate suggestions and highlight areas where additional clarification might have been useful. I regret the lack of precision in how I presented it.

Having carefully read your thorough responses to my questions, I am satisfied that you have successfully addressed the concerns raised. Your revisions have strengthened the manuscript, and I now fully support its publication.

Thank you again for your careful work and for your patience with this clarification.

We sincerely thank Reviewer #3 for the thorough evaluation and the supportive comments on our manuscript. We are grateful for the reviewer's thoughtful reconsideration and are pleased that our revisions have successfully addressed the earlier concerns.

Reviewer #4 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Communications Biology initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

We appreciate the note about the co-review process. We thank both the reviewer and the co-reviewer for their time and contribution.