

Genome diversity and tropical adaptation of Southeast Asians

Corresponding Author: Professor Bing Su

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version. Additionally, Reviewer#3 Version 1 comments are attached at the end of the file.

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)

In this manuscript, the authors from the Consortium of Anthropological Research in Southeast Asia and Southwest China (CASEAC) generated whole-genome sequencing (WGS) data from 3,023 individuals from 30 Mainland Southeast Asia (MSEA) and long-read WGS data from 37 representative MSEA individuals. They analyzed this "SEA3K genome dataset (Phase-I)" and identified 71.91 million short variants and 43,462 structural variants, of which many have not been reported previously. They also generated genome assemblies of the MSEA populations, adding 339.36 Mb novel sequences to the current human reference genomes. They claim to have discovered strong signatures of Darwinian positive selection (based on the tropical environment of the MSEA populations) that led to their distinctive physical characteristics, and evidence of at least two distinct instances of Denisovan admixture into modern humans in Asia.

As expected, the basic analyses of the SEA3K dataset yielded many novel SNVs, INDELs, and SVs, adding to the global reference sequence data and pointed to the value of sampling more populations in genome research. Principal component analysis established the genetic relatedness of the 30 MSEA populations and their relationship to other populations. There are major issues the authors should address and clarify.

1. While the 37 genome assemblies are sorely needed to expand the number of global reference genomes, these assemblies are billed as "haplotypic assemblies" but it is unclear what they are. Because each sample is diploid, there should be 74 haploid assemblies from the 37 individuals. If only one assembly is constructed from each individual, the assemblies are average/composite assemblies and cannot be labeled as of a quality "comparable with the two versions of the reference genome (GRCh38 and T2T-CHM13)". The authors should adopt the approach of the Human Pangenome Project to construct 2 haploid assemblies for each individual. Until then, the current assemblies are not contiguous assemblies but large stretches of phased contigs.

2. The CMS (composite of multiple signals) approach has been used successfully to identify regions (or even nucleotides) under positive selection, but CMS is a "screening tool" that cannot be used to draw definitive conclusions that the genes/variants under selection are responsible for specific traits even when the gene functions are known. Functional studies and much larger population studies must be conducted to make that connection. For example, some physical features of these populations are so distinctive and universal (such as hair shape) that they cannot be explained by the ~20% allele frequencies of the positively selected variants. The authors should tabulate the positively selected regions and the genes/variants identified in these regions without drawing the conclusions that they are the cause of population traits without further and more in depth studies. The authors are encouraged to conduct a formal study of "tropical adaptation" but the current dataset and its analyses do not provide sufficient evidence for the conclusions the authors are drawing.

3. Similarly, the discovery of "pathogenic variants" in these populations should not be used to draw conclusions about the health of the populations, as most sample donors are likely to be without serious rare genetic diseases. As it is commonly known that all humans harbor multiple "pathogenic variants" without the associated diseases, there is no basis for drawing broad conclusions about the MSEA populations based on incidental findings of "pathogenic variants".

4. With 3,023 individuals, this is the largest study of the MSEA populations but, once divided into 30 populations, the average cohort size for each population is only ~100, rendering the studies of individual populations ones of low statistical power. With the highly diverse genetic profiles across the populations, the authors need to justify why the analytical methods chosen are appropriate for the composite cohort of such a diverse group of populations.

Referee #2

(Remarks to the Author)

Yaoxi He and colleagues present a large-scale sequencing study of 3023 human genomes from mainland Southeast Asia, collectively referred to as the SEAK3K Phase-I dataset. The study presents a fairly large variety of analyses that range from descriptive, to exploratory and up to population genetics and the potential medical relevance of the dataset at hand. Unfortunately, many of the analyses remain somewhat superficial and the conclusions drawn seem too assertive or at least lack a proper validation with orthogonal data (in particular regarding potential functional implications). It is certainly the case that this dataset closes several (regional) blindspots in the current maps of global genetic variation, putatively adding thousands or millions [depending on variant type] of new variants. The value of this resource could be increased by making data and products (e.g. callsets) openly available to the scientific community. Apart from that, I have several concerns about the reporting and evaluation of the results relative to current standard catalogs of human variation that I listed below:

1. Your claim that SEA populations are currently underrepresented in the various large-scale efforts to characterize human genetic diversity you are listing in the Introduction is certainly correct, yet it would help motivate your study if you could make this more precise in one or two sentences. Can you make a statement how much (how many [local] populations) you are adding that are entirely missing so far?
2. Your section about reference datasets seems a bit off: the problem of a single reference genome is the focal point of initiatives such as the HPRC, and their pangenomes / pangenome graphs are meant to represent human genetic diversity in a more comprehensive and accurate way. If I assume that by "reference genome datasets are needed ..." you meant population-specific reference genome datasets - did you? - then this is exactly what a pangenome reference is aiming at. However, of course, efforts such as your study are then required to make the pangenome more complete. I suggest you rephrase that section to better emphasize the relevance of your contribution.
3. "The research themes of CASEAC cover ..." - I fail to see the relevance of that statement for the study at hand.
4. I do not understand what "rigorous" read mapping is supposed to be? In the Methods, I can see that you used bwa-mem, FASTQC but no further details are provided. Did you follow an established (best-practice) protocol, or just used the defaults ...? The "rigorous" here is meaningless and would better be replaced with an explicit reference to established procedures (I see that GATK Best Practices are mentioned in the Methods / Variant calling).
5. Please also evaluate the novelty of your SNV/INDEL calls relative to other population-scale references such as the HGSC and the HPRC callsets; or have these been incorporated into the dbSNP release you are comparing to?
6. I cannot match the numbers in the main (62.38M SNV/9.53M INDEL) to the post-QC numbers in the Methods: the "Variant QC" section states 60.5M variants (only SNVs? Or SNV and INDEL?) were left after QC. Please correct whatever went wrong here.
7. "the SEAK3K panel significantly improved ..." - I don't see any statement/calculation of the statistical significance here.
8. I do not fully support your conclusion that the varying - and not surprising - imputation performance highlights the necessity for population-specific reference panels. Why would more diverse reference panels not achieve a similar performance? A fundamental point, e.g., about the 1KGP panel is its public and free availability.
9. Based on the reference you are citing, the HGSC callset you are comparing to is the HGSC phase 2 callset (not 3). The reference is missing when you first mention the callset.
10. I have trouble following the way you constructed the SV callset. When looking at the Methods, it seems like you restricted the SV space essentially to breakpoints common to all samples, and then force-called this set in all samples again? However, the merge step with SURVIVOR seems very lenient by allowing a 1 kbp distance between breakpoints, which means a larger (kb-scale) SV could be merged with a much smaller one, effectively making the smaller one then disappear from your callset? This would explain the comparatively low number of SVs you are detecting per sample (~15k). In the same context, I am really surprised that you used such an outdated version of Sniffles (1.0.12) given that the major version upgrade to v2+ happened in early 2022 (and the last bugfix update to that in July 2023). Since I know how much work is required to update a callset that is the beginning for so many analyses downstream, please do not read that point as "rerun everything with Sniffles v2", but maybe you can justify your decision?
11. "falling in the annotated regulatory regions" - no reference given; what annotation was used here?
12. Same as above: please also include the HPRC set in your evaluation of novelty for your SV callset.
13. "they are comparable with the two versions of the reference genome with an average quality value of 50.31". The T2T genome is virtually error-free and has an estimated error rate of 1 per 10 Mbp (with local uncertainties), which is far beyond a QV value of ~50. A "comparison" would simply indicate that the assemblies are worse, but this is obviously as expected, so the relevance of your statement is unclear to me. Also, the GRC reports the hg38 N50 online as ~67.8 Mbp - how did you get to your number of ~58 Mbp?
14. Your conclusion that the unmappable sequences are likely (!) novel seems far-fetched. There are inherently difficult regions to align to (e.g., on chr16, the acrocentrics, centromeres etc.), hence not everything that does not align can be assumed to be novel in its entirety. Including the incomplete hg38 reference in this analysis, which is full of unresolved

sequence gaps, does not support your statement at all, at least not at the level of depth you are presenting here. To make this more plausible, you would have to check the unaligned sequences in more detail and confirm a correct assembly. Next, you would need to find the regions where these sequences putatively should align to (assuming that the unaligned sequences are large enough to reason about their relative placement in the assembled chromosomes) and show that there is indeed substantial sequence divergence relative to the reference. Given that consortia such as the HGSVC and the HPRC have released numerous assemblies, a claim of “novelty” is still prohibitive unless you also check against those datasets; a relevant point if you put your work in perspective relative to “the current pangenome of human populations”.

15. I do not understand what insight the short paragraph on the minigraph/pangenome construction is supposed to deliver.

16. “12 non-coding PS-SNVs falling in the known” - your conclusion about the concerted adaptation is very strong w/o additional evidence. How do you know that these regulatory regions are actually relevant in this context?

17. I do not find a proper description of how you obtained the functional label in Table S10 that seems the basis for your functional interpretation in that section (how you went from the GO enrichment to the one label shown). In particular, I am also unclear about the background distribution of GO terms in the 44 non-coding PS-SNV regions you are investigating. How many other genes are in those regions? What is their function?

18. “suggesting the potential involvement of ... no significant GO enrichment was detected”. Since you did not provide any additional evidence, I think “suggesting” is too strong here. It is certainly a viable hypothesis, though.

19. “One notable example is a 7439 bp deletion ...” - a notable example for what? For tropical adaptation? What is the evidence to make this a “notable” example?

20. In general, the results you are presenting that aim at emphasizing the importance of the identified variants for adaptation or medical conditions lack evidence of actually having a functional effect (i.e., are there detectable changes in gene expression, be it because of a direct or an indirect effect via differential regulation?).

Minor:

Some of your sentences could be restructured and shortened to ease readability. I am assuming you are going to revise the text anyway, so I am not highlighting all of them; just to give you two examples from the Introduction: “Mainland Southeast Asia (MSEA, also known as Indochina Peninsula), as the ...” and “Utilizing the SEA3K dataset, we characterized ...”. When you first introduce a new data type in the text, please be more explicit about the technology: “...high-coverage (average 43.5×) WGS of 3,023...”. I can conclude from the number of individuals that this is a short-read dataset, but please make that explicit.

Referee #3

(Remarks to the Author)

Please find my comments in the attached file.

Referee #4

(Remarks to the Author)

Nature MS 471894

April 2024

The authors present results from a variety of analyses on genomic samples collected from 30 pops in six southeast Asian countries including persons from minority Chinese communities. Their goal was to build a reference panel based on individuals underrepresented in previously established panels. The authors find genetic diversity not represented in previous panels.

If we assume that the data was collected ethically and within boundaries of modern genomic methodology, the analyses presented here may be informative. However, the interpretation of results and subsequent inference made are unfit for publication

Data & Methodology

The analyses generally appear to be robust and without obvious flaw. The authors are applauded in their gaining informed consent, and the inclusion of authors from sampled countries. However, problematic elements of this manuscript may have been avoided if members of subject minority MSEA communities were included as collaborators and coauthors.

Results and Appropriate Statistics

Most figures are well designed and informative. The usefulness of Figure 3E is not apparent. Despite the figure caption stating that it shows “historical changes of effective population sizes (N_e) of MSEA populations over time,’ in reality, the figure does not reflect that; instead it compares this measure between Tribal and non-Tribal populations. For what scientific

purpose is the comparison of IBD between Tribal and non-Tribal IBD? On page 16 the authors state that “some tribal populations... kept ancient genomic features inherited by their hunter and gatherer ancestors.” This statement is not supported by a citation, and it is irrelevant to note that a human population retains DNA from their ancestors. The authors desired goal to produce a panel for MSEA populations is not consistent with comparisons of measures between Tribal and non-Tribal peoples. While measures of IBD can be informative, when applied to comparisons of Tribal and non-Tribal populations, it generates the need to examine why the authors found that to be relevant or interesting. Historical abuses and contemporary social power differentials between Tribal and non-tribal peoples around the globe create a reality in which researchers must be or become aware of the stigma and damages of these kinds of comparisons.

The portion of the manuscript referring to “tropical adaptation” is flawed. In the most sophomorically erroneous portion, the authors make several troubling claims. They suggest that curly hair “is probably a local adaptation to tropical heat since curly hair may provide better ventilation.” They go on to support this statement by citing a paper by L’Oreal researchers that outlines geometric measures to describe human hair strand morphology. The De la Mettrie et al. paper does not in any way support or claim to support a causal relationship between curly hair and an evolutionary adaptation to tropical heat.

Similarly, their claims on short stature are not accurate. Human height is very highly polygenic with the majority of sites being individually of very low effect. Height is highly heritable, but genomewide studies have failed to identify most loci with significant contributions (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3232052/>). But, it is known that many loci have very small effects; it is unclear that the authors’ assertion that shorter height is adaptive based on their results. They assert that their candidate regions are enriched for loci contribution to height variation, but this is untested: if they sample 44 regions randomly, how many loci that effect height would overlap? The inference made by the authors on their untested claims on stature are inappropriate.

Conclusions

Hair, skin color, and height are known to be very highly polygenic, the few loci reported here would only be a fraction of the genotype contributing to phenotype. The nature of the genes underlying these classically quantitative traits is such that these conclusions are unsupported under the presented experimental design. Data presented in this manuscript have value and should be in the literature. The interpretation of results and subsequent inference made are unfit for publication. In particular, the claims of genotype to phenotype linkage and adaptation to various biotic and abiotic features of the environments experienced by MSEA peoples are untestable and unsupported by any data presented. Instead, they are based on stereotypes and dated assumptions of physiological benefits of specific traits. It is recommended that the authors of this manuscript revisit their findings to eliminate bias reflecting anachronistic approaches to genetic studies in Tribal populations. See approaches described by Fox et al. 2020 (<https://doi.org/10.3390/genes11010088>) and Claw et al. 2018 (<https://doi.org/10.1038/s41467-018-05188-3>). The addition of MSEA representation in authorship would be a positive step. The addition of MSEA representation in authorship would be a positive step.

Version 2:

Reviewer comments:

Referee #1

(Remarks to the Author)

The authors have made an effort to address the issues I raised and partially resolved them. Of note, they have now withdrawn the unsupported conclusions such as medical implications of the rare variants identified, high quality reference genomes, variants that contribute to “adaptations” such as short stature and other physical features. They have also produced 74 haplotype assemblies from the 37 long-read sequencing datasets.

However, the authors still make broad conclusions/speculations based on a very small dataset (short-read sequencing data from ~100 individuals from each subpopulation and genome assemblies from 1 individual from 37 subpopulations) without further studies. While I understand the desire to make the dataset seem more important than it currently stands, unsubstantiated claims will likely lead to confusion and diminish the credibility of the project. It is best to remove the speculative part of the discussion unless additional studies are done to validate them.

Referee #2

(Remarks to the Author)

He et al. present a substantially revised manuscript of their comprehensive study of 3,023 samples from mainland Southeast Asia. Overall, I find that the manuscript has much improved in focus, in detailing the analysis steps and a more careful interpretation of the results. The newly added text and figures add a lot of clarity but not in all remaining aspects. I have a few more questions and remarks stated below and most of which, I believe, can be answered with little effort.

Major:

P7: the added detail about your discovered short variants helps a lot to put the numbers into perspective. However, I think I didn’t see a section in the Methods that describes how you assess novelty; for SNVs, there is no room for interpretation (it’s just an exact match or it isn’t), but for the short indels, did you just overlap and require identity to decide on known vs novel? This is relevant information because your number of putatively novel (short) indels seems somewhat high when compared to, e.g., the 1000 Genomes cohort.

P10: I cannot entirely follow the section of the 37 assembled genomes. You state in the text that you produced phased assemblies, but the hifiasm call in the Methods lacks the appropriate input to produce a (truly) phased assembly. Given that you have family trios in your dataset, I would have expected to see the respective parental k-mer databases as additional input (i.e., using the so-called trio-binning mode). According to the hifiasm readme, your assemblies are only partially phased? Please clarify. And if I am not mistaken, you make the more precise statement about the assembly status in your answer to my previous question (13).

P10: "complex SVs" - what is the (working) definition of a complex SV in your analysis?

P46: in your explanation about how you derived the set of known SVs using HGSVC and HPRC references, what exactly is the operation behind the word 'and' in this statement: "... the combined SVs from HGSVC3 and HPRC was regard as the known SV set." For me, "and" implies an intersection, which I think is consistent with your use of bedtools stated at the beginning of the sentence. If you used the intersection instead of the union between HPRC and HGSVC to define a known set, you have created a very conservative set of known variants because both the HPRC and the HGSVC certainly performed internal QC before releasing their callsets. How many novel MSEA SVs remain if you consider the union of the two callsets?

Minor:

p5/6 and supp. Fig. 1: I think, overall, the numbers seem plausible and provide a great summary of the current status quo emphasizing the importance of your contribution. However, could be that readers may be confused due to the view limited on WGS only - I think, though, that this is absolutely the right perspective given the focus of your study - which then excludes samples that were only characterized with, e.g., WES or proteomics data; and many of the popular projects from your list (HPRC, SGDP, 1000 Genomes etc.) only have a small sample size (relative to the total ~1.4M that you counted) with quite a focus on high-diversity samples (essentially, the AFR continental group). I also randomly checked some of the references 1-41 (as per supp. Fig. 1), and I found some examples that do not make sense to me, e.g., the number 32 (Tan/Timpson, Annu Rev Genomics Hum Genet 2022) is apparently a review paper about the UK Biobank and I don't see how this adds to your sample count; number 3 (Brucato, iScience 2022) lists data sources in their Table S1, all of which seem to refer to other studies, which thus means this paper also does not add anything (or that samples were accidentally counted twice). I leave it up to you how to deal with this, but you may want to make your counting more transparent by adding a simple supplementary table stating how many samples characterized with WGS you added to the total from each publication stratified by continental group (what you show in the left pie chart in suppl. Fig. 1).

P7: You are referencing items 43 (Liao et al., Nature; the HPRC flagship paper) and 44 (an associated perspectives article that does not represent any source) for the HGSVC phase 3 (= identical citations for HPRC). Obviously, that is the wrong attribution and it seems the HGSVC phase 3 has at least a preprint available that you should cite here:

<https://biorxiv.org/cgi/content/short/2024.09.24.614721v1>

This is related to your answer to my previous question (9).

P11: given that reference-style SV callsets produced, e.g., by the HPRC and the HGSVC, are still work in progress, it would add perspective to also report the inverse statistic, i.e. how many SVs in those catalogs have no equivalent in your MSEA callset?

P28 / your answer to my previous question (8): I find your response somewhat beyond the point because global reference panels that are being expanded over time by adding more and more diverse populations will simply subsume population-specific reference panels. That does not deny that population-specific panels outperform global panels before that point is reached, but my point is that the collective effort of the scientific community should be targeted at creating globally diverse panels. However, since you are now offering the imputation server, your panel is at least accessible in that way, and I am fine with that.

P44 Methods: the yak call to build the k-mer databases lists the same input twice ("sample.clean.fq.gz"). I assume that is a typo, but you may want to double-check that the correct input was used.

P45 Methods: please add the version information for Jasmine

Typos:

P40: typo - "BLUEPRI>N<T"

Suppl. P25 / panel c: arch>ai<c

Use of Kb / "k": I assume the journal has guidelines for this, but, generally, the SI prefix for "kilo" is a lowercase "k" and I would hence suggest to use "kbp" as the unit for 1,000 bp throughout the manuscript; same for "Mbp" and "Gbp" where applicable

Referee #3

(Remarks to the Author)

In this revision, the authors have removed many overstatements in the previous version and focused on characterization of genetic variation in the MESA populations. These changes make the paper more suitable for publication, but the story also became less attractive because many suggestions from the reviewers were put as future work (and we know that "later" often means "never"). Anyway, most of my comments have been properly addressed by the authors. I have a few more suggestions for the authors to consider.

1) While I still feel the long-read sequencing part is a bit unconnected to most of the paper based on the short-read sequencing data, the authors have made a good effort to strengthen the connection by performing SV calling using the short-read sequencing data for SVs identified by the long-read data.

With the SV data from all individuals, can we combine SVs and small variants into an imputation reference panel? That is, can we use this resource to better impute SVs when we only have SNP array data? In addition, because SVs are likely to

have greater functional impacts than small variants, they should be included in the analysis of medical implications.

2) Regarding the imputation, results show that the SEA3K panel is only useful for improving imputation of MSEA populations, most of which are indigenous. This result is not surprising, but it makes the MSEA panel unlikely to be widely used given the difficulty to perform large-scale GWAS in these populations. A better way is to combine SEA3K with other datasets to create a MEGA panel.

3) In Supplementary Figure 3b and c, does "global populations" mean the combined data of all 56 HGDP populations? These populations are highly diverse and should not be pooled together for the calculation of MAF and R², which might make some results difficult to interpret. For example, SEA3K seems to perform worse than the SG10K panel when imputing variants with MAF < 0.2 in Supplementary Figure 3b, but the R² statistic shown in Supplementary Figure 3c is much higher than the SG10K panel. Why? Please add some discussion on the imputation results. For example, why GAsP can impute global populations almost as good as TOPMed and better than 1KGP, given that GAsP has much smaller sample size and only focuses on Asian populations. These results are difficult to explain.

4) Regarding results in Figure 1e and 1f, please make clear that the target population is MSEA populations and the better performance of SEA3K and GAsP is likely because these two panels include more indigenous SEA populations than the other panels.

5) For the Procrustes analysis, the authors reported the M2 statistic, which was defined as the sum of squared differences between the two coordinates (page 15). Do you mean the Procrustes t statistic instead? The t statistic is a similarity score ranged between 0 and 1, which the M2 statistic is a distance score. Please check and make sure. For more about the definition of these statistics, please refer to Noah Rosenberg's paper in 2010 (PMID: 20196748).

6) In Supplementary Figure 15, for better comparison, please use the same color code and y-axis scale for results based on SMC++ and MSMC2. When two methods were applied on the same dataset, are the results consistent? Please mention in the main text and be clear that which method is chosen as the main method and why.

Version 3:

Reviewer comments:

Referee #2

(Remarks to the Author)

He et al. have fully addressed my previous remarks and I am glad to see that more errors and ambiguities could be corrected by the authors.

I have thus no further comments.

Referee #3

(Remarks to the Author)

The authors have addressed all my comments. I only have 1 more minor question:

In Figure R6, when imputing EAS populations, SG10K seems to outperform 1KGP in terms of R² across all frequency bins (panel b), but 1KGP can impute many more variants with R² > 0.8 (panel a). How to explain these results, which seem inconsistent? Better to check if any calculation went wrong.

Referee #5

(Remarks to the Author)

I appreciate the effort the authors have put into compiling and analyzing this extensive dataset. While I am not a specialist in DNA or genetics, my review focuses on the manuscript's treatment of Indigenous communities and potential language that might be deemed pejorative or problematic in reference to the peoples who are the subject of this study.

Comments:

Recent Studies on Island Southeast Asia (ISEA)

The manuscript would benefit from incorporating recent genetic studies that parallel its findings, particularly:

Larena, Maximilian, Federico Sanchez-Quinto, Per Sjödin, James McKenna, Carlo Ebeo, Rebecca Reyes, Ophelia Casel et al. "Multiple migrations to the Philippines during the last 50,000 years." *Proceedings of the National Academy of Sciences* 118, no. 13 (2021): e2026132118.

Karmin, Monika, Rodrigo Flores, Lauri Saag, Georgi Hudjashov, Nicolas Brucato, Chelzie Crenna-Darusallam, Maximilian Larena et al. "Episodes of diversification and isolation in Island Southeast Asian and Near Oceanian male lineages." *Molecular Biology and Evolution* 39, no. 3 (2022): msac045.

These studies offer relevant comparative studies (within the region) into genetic diversification and migrations, complementing the findings of the manuscript.

Terminology: "Tribal"

The term "tribal" carries connotations linked to outdated evolutionary hierarchies and is often perceived as implying a lower stage of social complexity. A more neutral alternative, such as "ethnolinguistic groups" or simply "communities," would be preferable.

Defining "Indigenous"

The manuscript uses the term "Indigenous" but does not clarify its specific application within the context of Southeast Asia. Given the political weight of this term in the region, the authors should clearly define whether they mean local populations in general or groups that are historically marginalized and legally recognized as Indigenous. Currently, the manuscript presents these populations as static and unchanging, which does not reflect the dynamic and adaptive nature of these societies.

Framing of "Hunting-Gathering Lifestyle"

The use of the term "hunting-gathering lifestyle" reinforces outdated assumptions about certain communities being "primitive" or "backward." The manuscript would be more precise and less value-laden if it referred to "traditional subsistence strategies" instead. There are no pure hunter-gatherers today; most communities practice mixed economies, even if they engage in seasonal foraging.

Scope of Mainland Southeast Asia (MSEA)

The manuscript defines MSEA in a way that excludes significant parts of the region. MSEA is not limited to the Indochina Peninsula—it also includes Myanmar, Thailand, and Peninsular Malaysia. The authors should ensure that their geographic framing aligns with established regional definitions.

Clarity of Research Questions and Hypotheses

The manuscript presents findings in a largely descriptive manner without clearly articulating the central research questions or hypotheses. Are the authors attempting to identify common patterns across MSEA populations, or are they simply reporting genomic results? A stronger discussion on the "so what" aspect of the findings is needed. What are the broader implications of their conclusions? What are the consequences of their findings, and how do they contribute to addressing contemporary challenges faced by local communities in MSEA? A clearer articulation of these points would significantly improve the manuscript.

The study makes a valuable contribution to our understanding of the genetic diversity of Southeast Asia. However, a more careful approach to language and framing will enhance its impact and avoid potential misinterpretations. Thank you for the opportunity to review this work and hope my comments prove useful to the authors.

Version 4:

Reviewer comments:

Referee #5

(Remarks to the Author)

Thank you for addressing concerns from the last round of review.

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/>

Editor's comment

You will see that the reviewers agree that this is an important dataset, however, they feel that you have not showcased the full potential of it. We therefore ask that a revision will include an improved analysis of the data, that goes deeper for less superficial insights. Importantly, please pay very close attention to reviewer #4's concerns about the interpretation of data. Further, while we understand that human genetic data should be under controlled access to protect the participants, we ask that you make as much of the data available as possible (can you make callsets and summary data available in addition to the raw sequencing data, with less restrictive access conditions?) to maximize the usefulness to the scientific community. Access restrictions must be stated clearly and transparently. It will also be crucial to improve the methods description, so that the reviewers can fully assess their validity.

Should further experiments and/or analyses allow you to address these and all other criticisms, we would be happy to consider a revised manuscript (unless something similar has been accepted at Nature or appeared elsewhere in the meantime). We hope to receive your revised paper within four to six months. If you cannot complete the required revisions within this time frame, please let us know when you would anticipate being able to submit a revised manuscript. We also strongly suggest that your revised manuscript has tracked changes, which is increasingly requested by referees to aid in their re-review.

We are committed to providing a fair and constructive peer-review process. Please 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.

Authors response: We have carefully considered the questions and issues raised by the four referees, which are indeed helpful for improving our manuscript. In the revision, we have conducted the following analyses to showcase the full potential of the SEA3K dataset:

1) Using the long-read data, we re-assembled the 37 MSEA genomes at phased and diploid level. For structural variations (SVs), following the reviewer's suggestion, we updated the callset by utilizing the up-to-date analytical tools, including Sniffles2, PAV and SVision-pro. Finally, we obtained a full-scale SV spectrum of MSEA populations using the long-read data and the phased assemblies, and we also obtained the SV frequency spectrum at population level by genotyping these SVs in the 2,183 unrelated MSEA individuals using the deep-depth short-read WGS data.

2) We re-assessed the novelty and significance of our SEA3K dataset in multiple aspects: including re-evaluation of the novel variants (SNV, INDEL and SV) by covering more public datasets (e.g. datasets from HPRC and HGSC), re-evaluation of the imputation power by comparing with more Asian reference panels (e.g. the SG10K panel), re-analyses of archaic introgression in the MSEA genomes using high-coverage WGS data of 1KGP with statistical supports, and re-evaluation of the medical implications of the SEA3K dataset by assessing the prevalence of the reported disease

risk variants in MSEA populations.

3) We re-analyzed the results of natural selection in MSEA populations by incorporating a comprehensive functional annotation using public databases, including GWAS Catalog, GTEx and ENCODE, and we provided the statistical significance for the enriched functional terms. To be objective in interpreting the selective signals, we tuned down the previous claim of tropical adaptation since we did not perform functional experiments to validate the speculated functional/phenotypical relevance, which will be a separate project in the future. Following the reviewer's suggestion, we objectively listed the variants/genes/regions with putative signals of positive selection in MSEA populations, together with the annotations of functions and traits/disease of these selected genes based on previous studies and public databases. Importantly, we tested polygenic selection on multiple traits using a series of statistics, and proved that short stature has been subject to polygenic selection in MSEA populations.

4) For data sharing, according to the Editor's suggestion, all summary data including variant callsets, variant frequency and all statistics of natural selection analyses will be open access. All original data including individual sequencing data and genomes of the SEA3K participants will also be available upon request for research purposes. For raw data (WGS data of short reads and long reads) and genome assemblies, according to the regulations of the People's Republic of China on the administration of human genetic resources, and to protect participant confidentiality, we deposited the raw data at Genome Sequence Archive for Human (GSA-Human: <https://ngdc.cncb.ac.cn/gsa-human/>) as controlled-access. In brief, investigators who are interested in obtaining the SEA3K data would need to submit an application that includes a detailed research proposal and an IRB approval from the applicant's home institute to the Data Access Committee. Once approved, researchers will be able to obtain the raw data for the approved research projects during the approved time period. In addition, to accelerate evaluation and broad utility of our SEA3K data, we generated the SEA3K Imputation Server (<https://ngdc.cncb.ac.cn/SEA3K/>) by taking MSEA populations as the reference panel. Users can have free access to this server and conduct imputation analyses. We hope this server will help the scientific community with a better use of the SEA3K dataset and promote genetic and medical studies of MSEA populations. We provided a detailed description of data access in the Data Accessibility section of the revised manuscript.

5) In particular, to address the concerns of data interpretation by Reviewer #4, we adjusted our presentation to avoid overstatement. To avoid potential ethical issue, we removed the comparisons between tribal and non-tribal groups. We downplayed the conclusion and interpretation of "tropical adaptation" derived from the analyses of natural selection and objectively listing the variants/genes/regions with putative signals of positive selection in MSEA populations. At the same time, we conducted comprehensive analyses of polygenic adaptation and we proved that short stature was subjected to polygenic selection in MSEA populations.

Reviewer-1

In this manuscript, the authors from the Consortium of Anthropological Research in Southeast Asia and Southwest China (CASEAC) generated whole-genome sequencing (WGS) data from 3,023 individuals from 30 Mainland Southeast Asia (MSEA) and long-read WGS data from 37 representative MSEA individuals. They analyzed this "SEA3K genome dataset (Phase-I)" and identified 71.91 million short variants and 43,462 structural variants, of which many have not been reported previously. They also generated genome assemblies of the MSEA populations, adding 339.36 Mb novel sequences to the current human reference genomes. They claim to have discovered strong signatures of Darwinian positive selection (based on the tropical environment of the MSEA populations) that led to their distinctive physical characteristics, and evidence of at least two distinct instances of Denisovan admixture into modern humans in Asia.

As expected, the basic analyses of the SEA3K dataset yielded many novel SNVs, INDELs, and SVs, adding to the global reference sequence data and pointed to the value of sampling more populations in genome research. Principal component analysis established the genetic relatedness of the 30 MSEA populations and their relationship to other populations.

There are major issues the authors should address and clarify.

1. While the 37 genome assemblies are sorely needed to expand the number of global reference genomes, these assemblies are billed as "haplotypic assemblies" but it is unclear what they are. Because each sample is diploid, there should be 74 haploid assemblies from the 37 individuals. If only one assembly is constructed from each individual, the assemblies are average/composite assemblies and cannot be labeled as of a quality "comparable with the two versions of the reference genome (GRCh38 and T2T-CHM13)". The authors should adopt the approach of the Human Pangenome Project to construct 2 haploid assemblies for each individual. Until then, the current assemblies are not contiguous assemblies but large stretches of phased contigs.

Authors response: We understand the reviewer's point, and we agree that constructing two haploid assemblies per individual is the ultimate goal of dissecting sequence variations in a genome, just like Human Pangenome Project is working on. However, according to the current sequencing technologies and algorithm tools, besides of a high-coverage long-read sequencing data, *i.e.* the PacBio HiFi reads, obtaining fully phased assemblies requires Hi-C data or trio-binning data. The Human Pangenome Reference Consortium (HPRC) produced fully phased diploid assemblies using three data types, including long-read HiFi data, trio-binning data and Hi-C data (Liao et al., 2023). Unfortunately, for our 37 MSEA individuals, the available blood samples only allowed us to generate high-coverage long-read HiFi data.

Therefore, following the reviewer's suggestion, we utilized the Hifiasm tool (Cheng et al., 2021), in the mode of "HiFi-only Assembly" to produce two partially phased assemblies for each sample, *i.e.* each individual having a haplotype-1 and a

haplotype-2 assemblies. According to the description of Hifiasm (<https://hifiasm.readthedocs.io/en/latest/pa-assembly.html#hifi-only-assembly>) and the published assemblies from Chinese Pangenome Consortium (Gao et al., 2023), this pair of files can be considered representing the two haplotypes in a diploid genome, though with occasional switch errors. Finally, after polishing using both long-read and short-read data, a total of 74 high-quality haploid assemblies were eventually produced for the 37 individuals. We evaluated these phased assemblies and they showed high qualities for completeness, contiguity and accuracy, with an average total genome length of 3.00 Gb (range, 2.91-3.04 Gb), the contig N50 of 49.98 Mb (range, 29.06 - 79.16 Mb), and the quality value of 49.83 (range, 46.47-52.75) (**Figure R1** and **Supplementary Table 5**). In the future, we will make effort to collect trio samples and eventually produce fully phased assemblies.

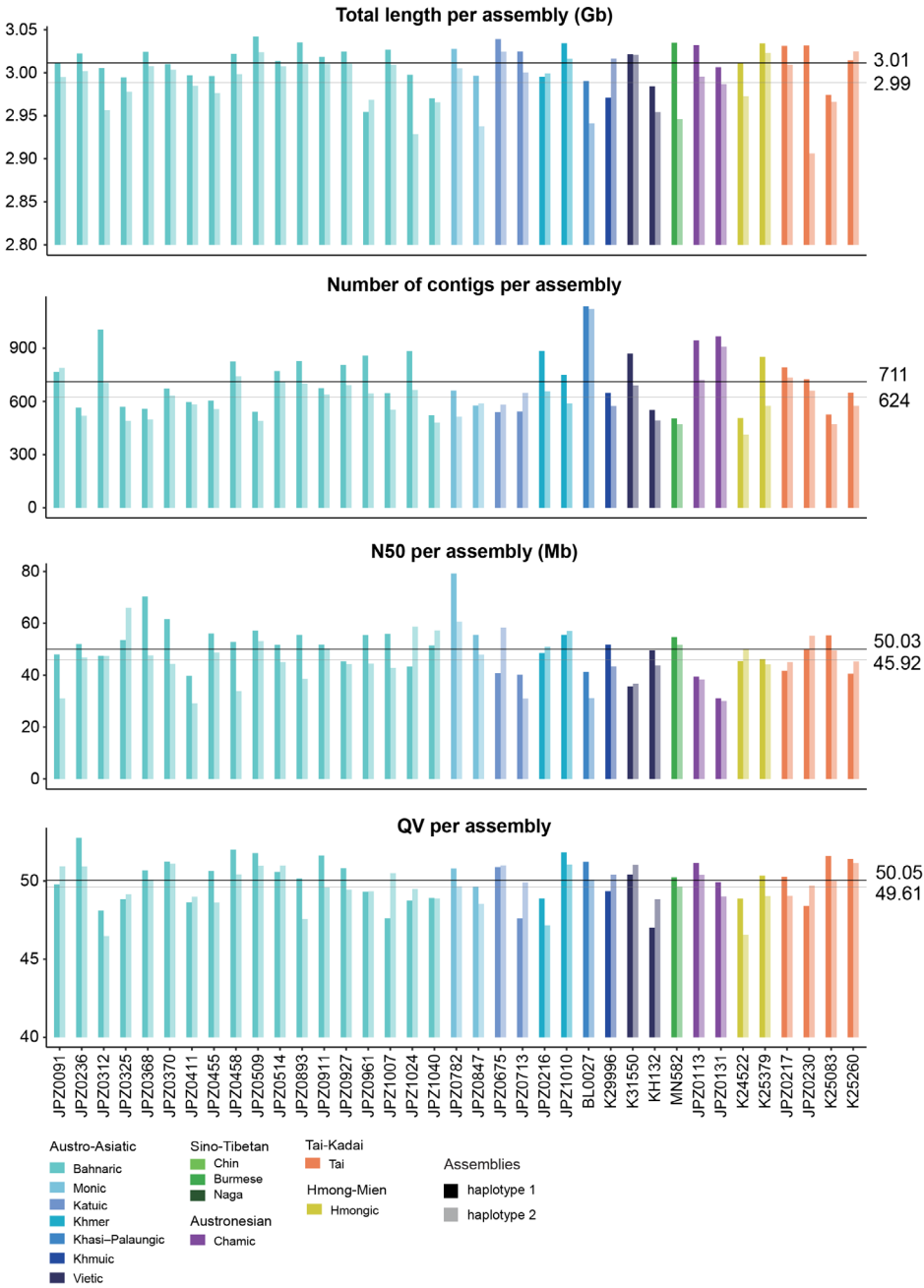


Figure R1. Assembly statistics of the 37 MSEA samples with long-read HiFi data. Each individual has two haploid assemblies (haplotype-1 and haplotype-2), plotted as two bars. A total of 74 haploid assemblies were produced for the 37 individuals. Total length, number of contigs, contig N50, and quality value (QV) per assembly is shown in the plots, respectively. The detailed statistics of the 74 assemblies are provided in **Supplementary Table 5**.

2. The CMS (composite of multiple signals) approach has been used successfully to identify regions (or even nucleotides) under positive selection, but CMS is a "screening tool" that cannot be used to draw definitive conclusions that the genes/variants under selection are responsible for specific traits even when the gene functions are known. Functional studies and much larger population studies must be conducted to make that connection. For example, some physical features of these populations are so distinctive and universal (such as hair shape) that they cannot be explained by the ~20% allele frequencies of the positively selected variants. The authors should tabulate the positively selected regions and the genes/variants identified in these regions without drawing the conclusions that they are the cause of population traits without further and more in-depth studies. The authors are encouraged to conduct a formal study of "tropical adaptation" but the current dataset and its analyses do not provide sufficient evidence for the conclusions the authors are drawing.

Authors response: We agree with the reviewer that CMS analysis is just a screening tool, and more functional and large-scale population studies need to be conducted to establish robust connections between the detected selective signals and the adaptive traits. In the revised manuscript, following the reviewer's suggestion, we tuned down the previous claims of tropical adaptation. For the article title, the word of "tropical adaptation" is now changed to "natural selection". We updated the list variants/genes under positive selection and their annotated functions without drawing conclusions on their contribution to adaptation (**Supplementary Table 9**). We also annotated all genes located in the 44 positive selection regions by GWAS Catalog, HPO and MGI mutant phenotypes using GeneCards database. This information has been updated in the **Supplementary Table 10**. Besides, in the Manhattan plot of the CMS scores (**Figure 4a**), we removed the color-coded panels indicating the connections between the genes under selection and the adaptive traits.

For the reviewer's point on polygenic effect of traits, we employed multiple strategies to test the speculated polygenic adaptation on these traits in MSEA populations, including body height, skin pigmentation, hair shape and nasal shape. We detected a significant signature of polygenic selection on height decreasing in MSEA populations (P -value = 0.004; **Supplementary Figure 18, Methods**). In other words, natural selection favors a lower body height in MSEA populations. For the other traits, no significant signature was detected (**Supplementary Table 11**).

3. Similarly, the discovery of "pathogenic variants" in these populations should not be used to draw conclusions about the health of the populations, as most sample donors are likely to be without serious rare genetic diseases. As it is commonly known that all

humans harbor multiple "pathogenic variants" without the associated diseases, there is no basis for drawing broad conclusions about the MSEA populations based on incidental findings of "pathogenic variants".

Authors response: We agree the reviewer’s comment that, at individual level, a person usually harbors multiple “pathogenic” variants without serious diseases. For example, we found almost everyone in our SEA3K cohort carried at least one variant designated as “pathogenic” (level 5) based on the ClinVar Database (<http://www.ncbi.nlm.nih.gov/clinvar/>), with each person carrying a median of 5 (ranging from 2 to 11) risk alleles (**Figure R2a**, added as **Figure 2B** in the revised manuscript). This number is similar with that reported in African individuals from the H3Africa Project, a median of 7 (ranging from 2 to 14) risk alleles (Choudhury et al., 2020).

However, in rare disease genetics, the population-based datasets can be used to filter out those reported “risk” variants for rare diseases (usually by studying patients), since common variants in the populations are unlikely to be responsible for highly penetrant rare disorders. Hence, in the absence of appropriate population reference datasets, some of the reported “risk” rare alleles inferred from a specific population can be misclassified and may lead to false disease associations or toward overestimating pathogenicity (Shah et al., 2018). At the same time, the classification of variants as either benign or pathogenic is challengeable without diverse population sets, given that the effect of disease-causal variants can vary across populations (All of Us Research Program Genomics, 2024). Therefore, for the 10 annotated “pathogenic” variants that are common in our SEA3K dataset but rare in other populations of the published datasets (e.g. 1KGP, gnomAD and TOPMed) (**Figure R2b**), a more plausible explanation is that many of these variants were likely misclassified as “pathogenic” by the ClinVar database. Such observation and interpretation have been reported in other non-European populations, including H3Africa Project (Choudhury et al., 2020) and GenomeAsia Pilot project (GenomeAsia, 2019).

To clearly state the usefulness of the SEA3K dataset for clinical genetics, in the revised manuscript, we rephrased this part and pointed out the medical implications of our SEA3K dataset and the reason for a valuable resource for clinical genetics, which is also suggested in comment#8 of Reviewer #3.

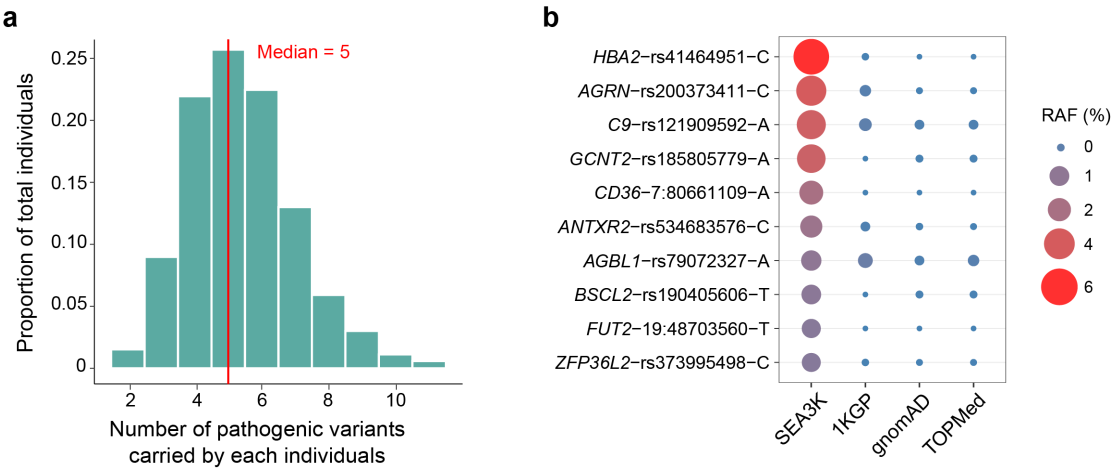


Figure R2. Assessment of the 604 pathogenic ClinVar variants in the SEA3K dataset at individual level and population level. a) Counts of the annotated pathogenic variants per individual in the SEA3K dataset. The median value is five indicated by the vertical line in red. **b)** The top ten annotated disease risk alleles with relatively high frequencies in MSEA populations (risk allele frequency, RAF > 1%), but rare in other world populations (RAF < 0.5%).

4. With 3,023 individuals, this is the largest study of the MSEA populations but, once divided into 30 populations, the average cohort size for each population is only ~100, rendering the studies of individual populations ones of low statistical power. With the highly diverse genetic profiles across the populations, the authors need to justify why the analytical methods chosen are appropriate for the composite cohort of such a diverse group of populations.

Authors response: We agree with the reviewer that sample size might influence the statistic power of population analysis. Indeed, in some analyses, a diverse composite of populations might affect the results when combining all samples into the analysis. To avoid the influence of population substructure, in our population genetic analyses (i.e. allele frequency or LD information statistics), we chose to use a more homogenous genetic population as the unit to make inference, and thus dividing the 3,023 individuals into 30 populations.

In our study, two analyses involved calculation of allele frequency, including the CMS calculation for natural selection detection and the archaic introgression assessment, which need to consider the influence of small sample size.

For the introgression analysis, the two analytical methods, SPrime and IBDmix, in the original publications (Browning et al., 2018; Chen et al., 2020), have both used simulation data to prove that panels with 15 and 10 individuals respectively show good robustness for statistical power (**Figure R3a**). Therefore, the sample size of dozens of people is enough to meet the analytical needs of these two methods. To minimize the influence of sample size on the results, we only selected populations with sample size >20 (23 MSEA populations in total) for introgression analyses.

For the CMS calculation, only the two statistics: differentiation of derived allele frequency (ΔDAF) and population branch statistic (PBS), require a large sample size, because they are both based on calculation of allele frequency to estimate genetic differentiation (i.e. fixation index: F_{ST}). There have been a large number of previous studies exploring the statistical power of minimum sample size for accurate estimation of genetic differentiation, and it is commonly believed that a sample size of $n > 20$ are sufficient to yield reliable estimates of genetic differentiation (Holsinger and Weir, 2009; Kalinowski, 2005; Morin et al., 2009; Willing et al., 2012) (**Figure R3b**). In our CMS analyses, we only included populations with a sample size of $n > 20$ (23 MSEA populations in total).

The CMS score of each SNV is calculated on the basis of all unrelated individuals separately on a population basis, and then are averaged among all the populations, and there were 44 identified positive-selection regions. When we recalculated the CMS

score using only the 23 large-sample-size populations, we identified 55 positive-selection regions, and found that all 44 previous positive-selection regions were included. The more regions inferred from the large-sample-size populations is due to the reduced number of the denominator compared to the use of all populations (23 vs. 30) when calculating the average CMS scores. To be conservative, in the revised manuscript, we kept including the 44 regions for downstream analysis.

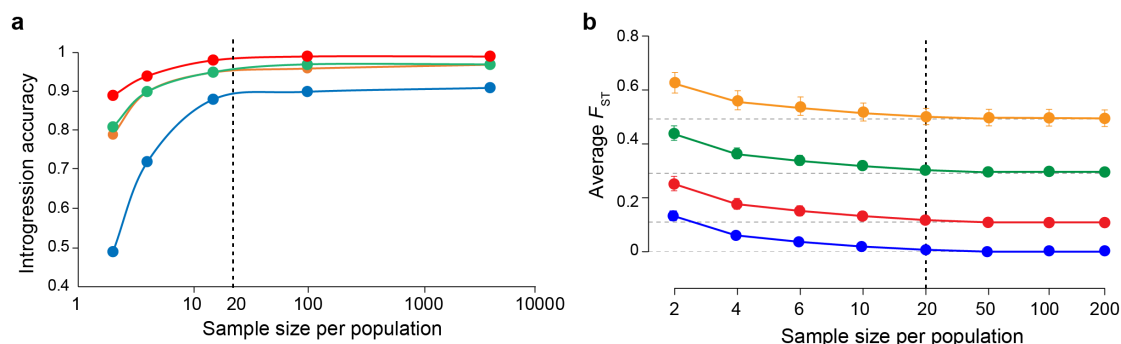


Figure R3. Assessment of statistic power of introgression and natural selection detection based on different population size. a) Power and accuracy of introgression detection (by Sprime) in the simulated data as a function of sample size. The colors refer to the four different detection modes, and the figure was adapted from the published study (Browning et al., 2018). **b)** Influence of accuracy of natural selection detection (by F_{ST}) with increase of sample size from 2 to 200 individuals from each population. The colors refer to the different levels of genetic differentiation (F_{ST} =0, 0.1, 0.2, 0.4), and the figure was adapted from the published study (Kalinowski, 2005; Willing et al., 2012).

Reviewer-2

Yaoxi He and colleagues present a large-scale sequencing study of 3023 human genomes from mainland Southeast Asia, collectively referred to as the SEAK3K Phase-I dataset. The study presents a fairly large variety of analyses that range from descriptive, to exploratory and up to population genetics and the potential medical relevance of the dataset at hand. Unfortunately, many of the analyses remain somewhat superficial and the conclusions drawn seem too assertive or at least lack a proper validation with orthogonal data (in particular regarding potential functional implications). It is certainly the case that this dataset closes several (regional) blindspots in the current maps of global genetic variation, putatively adding thousands or millions [depending on variant type] of new variants. The value of this resource could be increased by making data and products (e.g. callsets) openly available to the scientific community. Apart from that, I have several concerns about the reporting and evaluation of the results relative to current standard catalogs of human variation that I listed below.

Authors response: We appreciate the reviewer's approval of our study and the constructive comments. In the revised manuscript, we have addressed these points and the detailed explanations are provided below.

1. Your claim that SEA populations are currently underrepresented in the various large-scale efforts to characterize human genetic diversity you are listing in the Introduction is certainly correct, yet it would help motivate your study if you could make this more precise in one or two sentences. Can you make a statement how much (how many [local] populations) you are adding that are entirely missing so far?

Authors response: We appreciate the reviewer's suggestion. We summarized the whole-genome sequencing (WGS) data of global human populations from the published studies (**Figure R4**). Among the 1.4 million global samples for WGS, 65.21% are Europeans, and 13.65% are Asians, and among Asians, the great majority are East Asians and South Asians, and only 5.21% of the Asian samples are Southeast Asians, involving 9,795 samples. However, 9,051 out of the 9,795 SEA samples are Singaporean from the SG10K project (Chan et al., 2022; Wu et al., 2019). In particular, only 3% (233 samples) of the SEA samples are indigenous people, given the fact that SEA is home to the largest number of indigenous peoples with an estimated population size of 150 million out of the 470 million original inhabitants worldwide (Dhir et al., 2020). In the revised manuscript, we added two sentences in the Introduction section to highlight the point of underrepresentation of MSEA populations (**Page 6, first paragraph**). Also, we added Supplementary **Figure 1 (Figure R4)** to demonstrate the stated underrepresentation of MSEA populations in current WGS datasets worldwide.

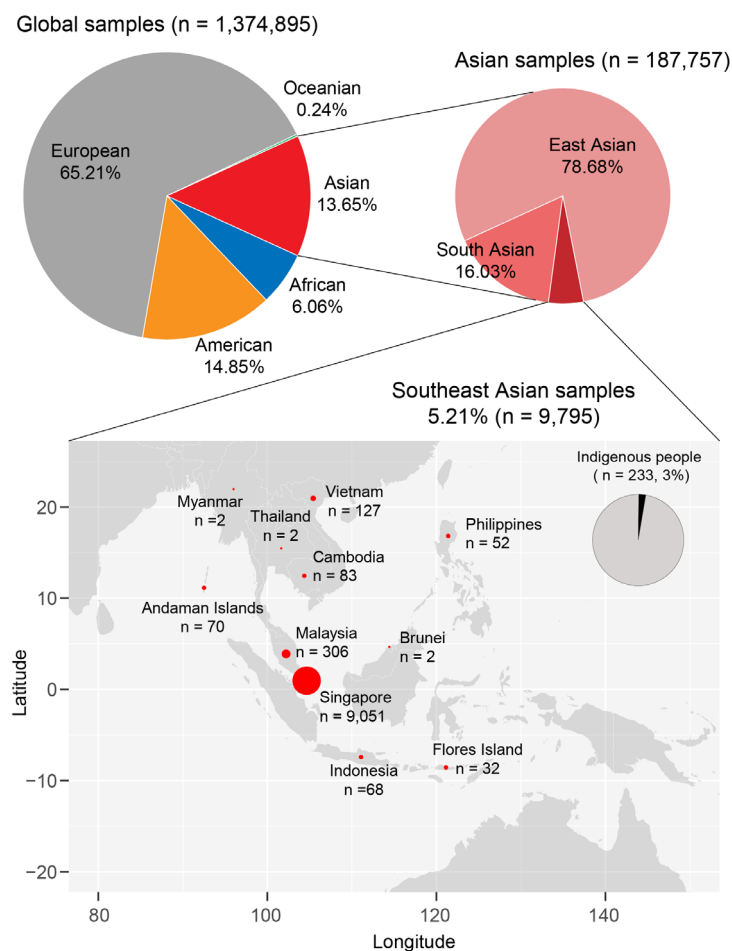


Figure R4. Summary of human individuals with whole-genome sequencing data in global populations. The data were retrieved from the published literatures (Bergström et al., 2020; Beyter et al., 2021; Browning et al., 2018; Brucato et al., 2022; Brucato et al., 2021; Byrska-Bishop et al., 2022; Chan et al., 2022; Chen et al., 2023; Choin et al., 2021; Cong et al., 2022; Deng et al., 2022; Du et al., 2019; Fatumo et al., 2022; Halldorsson et al., 2022; Hao et al., 2021; Ioannidis et al., 2020; Jacobs et al., 2019; Jeon et al., 2020; Karczewski et al., 2020; Kawai et al., 2023; Liu et al., 2022; Malaspinas et al., 2016; Mallick et al., 2016; Mondal et al., 2016; Naslavsky et al., 2022; Razali et al., 2021; Reis et al., 2023; Scheinfeldt et al., 2019; Silcocks et al., 2023; Taliun et al., 2021; Tan and Timpson, 2022; Walter et al., 2015; Wang et al., 2022; Wong et al., 2023; Wong et al., 2013; Zhang et al., 2021; Zhang et al., 2022; Ziyatdinov et al., 2023).

2. Your section about reference datasets seems a bit off: the problem of a single reference genome is the focal point of initiatives such as the HPRC, and their pangenomes / pangenome graphs are meant to represent human genetic diversity in a more comprehensive and accurate way. If I assume that by “reference genome datasets are needed ...” you meant population-specific reference genome datasets - did you? - then this is exactly what a pangenome reference is aiming at. However, of course, efforts such as your study are then required to make the pangenome more complete. I suggest you rephrase that section to better emphasize the relevance of your contribution.

Authors response: We agree with reviewer that the limitation of a single reference genome for human genomic research should be improved by constructing the pangenome with diverse populations, as HPCR is doing. However, construction of pangenome reference for MSEA populations was not the goal of our Phase I genome project because due to sample limitation, we only generated long-read HiFi data, which was not sufficient for constructing a pangenome reference. Our current study aimed to provide a genome-wide landscape of genetic variations, by leveraging the 3,023 deep-coverage short-read WGS data for SNV/INDEL discovery, together with the 37 deep-coverage long-read WGS data for SV discovery. Therefore, following the reviewer’s suggestion, in the revised manuscript, to better highlight the contribution of our SEA3K (Phase I) dataset, we rephrased the section of the long-read dataset by emphasizing SV detection instead of pangenome reference construction (**Page 6, Line 138-141**).

3. “The research themes of CASEAC cover ...” - I fail to see the relevance of that statement for the study at hand.

Authors response: Considering the reviewer’s comment, we moved this part to the Methods section, where we described ethical approval and the contribution of CASEAC.

4. I do not understand what “rigorous” read mapping is supposed to be? In the Methods, I can see that you used bwa-mem, FASTQC but no further details are provided. Did you follow an established (best-practice) protocol, or just used the defaults ...? The “rigorous” here is meaningless and would better be replaced with an explicit reference to established procedures (I see that GATK Best Practices are mentioned in the Methods

/ Variant calling).

Authors response: Indeed, as we described in Methods, we identified SNVs and INDELs based on the GATK Best Practices of germline short variant discovery, and we employed the standard pipeline without extra changes. Therefore, we removed the word “rigorous” in the revised manuscript.

5. Please also evaluate the novelty of your SNV/INDEL calls relative to other population-scale references such as the HGSVC and the HPRC callsets; or have these been incorporated into the dbSNP release you are comparing to?

Authors response: According to the reviewer’s suggestion, we added more public datasets to evaluate the novelty of SNV/INDEL calls of SEA3K, including HGSVC3, HPRC, high-coverage 1KGP Phase 3, HGDP, gnomAD, as well as two callsets with Southeast Asian samples: GAsP and SG10K. As expected, compared to our previous assessment using only the dbSNP callset, the novel variants of SEA3K decrease from 32.97 million to 24.33 million, especially for common variants (minor allele frequency > 1%), decreasing from 2.10 million to 1.14 million. Nevertheless, the SEA3K callset still shows over one million novel common variants, which have not been detected in the public callsets. We updated this part in the revised manuscripts:

(Page 7, Line 172-177):

“Compared with current public variant repositories, including dbSNP (v156), Human Genome Structural Variation Consortium (HGSVC) Phase 3, Human Pangenome Reference Consortium (HPRC) (v1.1), high-coverage 1KGP, HGDP, gnomAD (v3.1.2), GAsP, and SG10K (Pilot version), the SEA3K dataset contains 24.33 million novel variants, including 19.61 million SNVs and 4.72 million INDELs ...”

6. I cannot match the numbers in the main (62.38M SNV/9.53M INDEL) to the post-QC numbers in the Methods: the “Variant QC” section states 60.5M variants (only SNVs? Or SNV and INDEL?) were left after QC. Please correct whatever went wrong here.

Authors response: We appreciate the reviewer’s reminding. It was our unclear description about variant counts. In the revised, we checked the number of the variants of each step and presented them in a clear way.

(Page 7, Line 169-171):

“After read mapping, variant calling, and quality control, we identified 79.63 million small variants among the 3,023 MSEA samples, including 65.02 million SNVs and 14.61 million short insertions and deletions (INDELs, 1bp to 49 bp) (Methods).”

Variant filtering in Methods (Page 33, Line 833-837):

“Finally, we identified 89,226,548 variant sites for the 3,023 individuals, and 71,908,240 variant sites were flagged as “PASS” by VQSR and retained for

downstream analyses. After splitting of multi-allelic sites (4,838,041), the final dataset contained 79,634,419 million small variants, including 65,021,530 SNVs and 14,612,889 INDELs.”

7. “the SEAK3K panel significantly improved ...” - I don’t see any statement/calculation of the statistical significance here.

Authors response: According to the reviewer’s comment, to avoid overstatement, we removed the word “significantly” in the revised manuscript.

8. I do not fully support your conclusion that the varying - and not surprising - imputation performance highlights the necessity for population-specific reference panels. Why would more diverse reference panels not achieve a similar performance? A fundamental point, e.g., about the 1KGP panel is its public and free availability.

Authors response: The basic principle of imputation is that, by leveraging the haplotype structure determined by the genotyped variants on the array, we can infer the un-genotyped variations through multi-marker imputation into the sequenced reference panels of whole genomes (Li et al., 2009; Marchini and Howie, 2010). Indeed, a more diverse reference panel could enhance the imputation power for global populations, but the premise is that the target population, which you want to impute should be included in the global reference panel. The imputed genotypes are prone to errors in the absence of haplotype structure of the target population in the global reference panel (Choi et al., 2023). Therefore, increasing diversity and specificity of the reference panel will help improve imputation accuracy of genotypes, especially for the underrepresented populations which are not included in the current reference panel.

The 1KGP3 is one of the gold standard panels for a variety of genetic studies and is known to evenly cover global populations including 1,186 Asians. Besides, efforts have been devoted to construct reference panels for Asians in recent years, for example, the GenomeAsia 100K panel (GenomeAsia, 2019) and Singapore 10K (SG10K) (Chan et al., 2022; Wu et al., 2019), which contains more diverse Asian populations. However, Southeast Asians are still underrepresented in these reference panels.

Another reason for the necessity of population-specific imputation reference panel is that, the low-frequency variants tend to be population specific because they are mostly evolutionarily recent (Taliun et al., 2021). Population-specific reference panels created from WGS data are essential for accurately genotyping these low-frequency variants.

As expected, our SEA3K dataset outperforms the current reference panel in imputation accuracy of Southeast Asian populations, especially for those rare variants (**Figure 1e, 1f** and **Supplementary Figure 3**). These results highlight the usefulness of a population-specific imputation reference panel for accurately imputing genotypes.

9. Based on the reference you are citing, the HGSVC callset you are comparing to is the HGSVC phase 2 callset (not 3). The reference is missing when you first mention

the callset.

Authors response: We thank the reviewer for raising this point. Indeed, the callset we used to compare with our SV callset was HGSVC phase 2 callset, which was described in the publication by (Ebert et al., 2021). We noticed that the HGSVC phase 3 callset have been released recently, which expanded to additional human genomes using more complete phased assemblies including longer and more accurate long-read sequence data. Therefore, we updated our comparison analysis using the latest callset of HGSVC Phase 3 to assess the novelty of our SV callset.

(http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC3/working/20240415_Freeze4/)

Considering that the HGSVC3 dataset has not been formally published, in the revised manuscript, we cited the two main papers of HGSVC (Chaisson et al., 2019; Ebert et al., 2021), and added the corresponding ftp address of the latest HGSVC callset (**Page 46, Line 1179-1193**).

10. I have trouble following the way you constructed the SV callset. When looking at the Methods, it seems like you restricted the SV space essentially to breakpoints common to all samples, and then force-called this set in all samples again? However, the merge step with SURVIVOR seems very lenient by allowing a 1 kbp distance between breakpoints, which means a larger (kb-scale) SV could be merged with a much smaller one, effectively making the smaller one then disappear from your callset? This would explain the comparatively low number of SVs you are detecting per sample (~15k). In the same context, I am really surprised that you used such an outdated version of Sniffles (1.0.12) given that the major version upgrade to v2+ happened in early 2022 (and the last bugfix update to that in July 2023). Since I know how much work is required to update a callset that is the beginning for so many analyses downstream, please do not read that point as “rerun everything with Sniffles v2”, but maybe you can justify your decision?

Authors response: As the reviewer pointed out, considering Sniffles2 outperforms the old version, we updated our SV callset using Sniffles2 (v2.2) (Smolka et al., 2024). We also employed an additional SV caller (PAV, <https://github.com/EichlerLab/pav>) (v2.3.4) (Ebert et al., 2021) based on assembly comparison to detect the SVs in the 37 MSEA samples (74 phased assemblies). The details of SV calling of Sniffles2 and PAV are provided in the Methods section of the revised manuscript.

For SV merging, we agree with reviewer that using 1000-bp distance between breakpoints of two SVs in SURVIVOR is too lenient. Therefore, we combined our SV calls with the use of Jasmine (<https://github.com/mkirsche/Jasmine/>) (Kirsche et al., 2023), a commonly used tool to merge SV across samples based on forest algorithm, which was developed by the research group of Michael Schatz and Sergey Aganezov. The merging strategy of our call sets was shown in the schematic diagram below (Figure R5). We merged the Sniffles2 calls and the PAV calls of each individual one by one firstly using Jasmine, then merged the 37 individual calls into a population SV callset by Jasmine again.

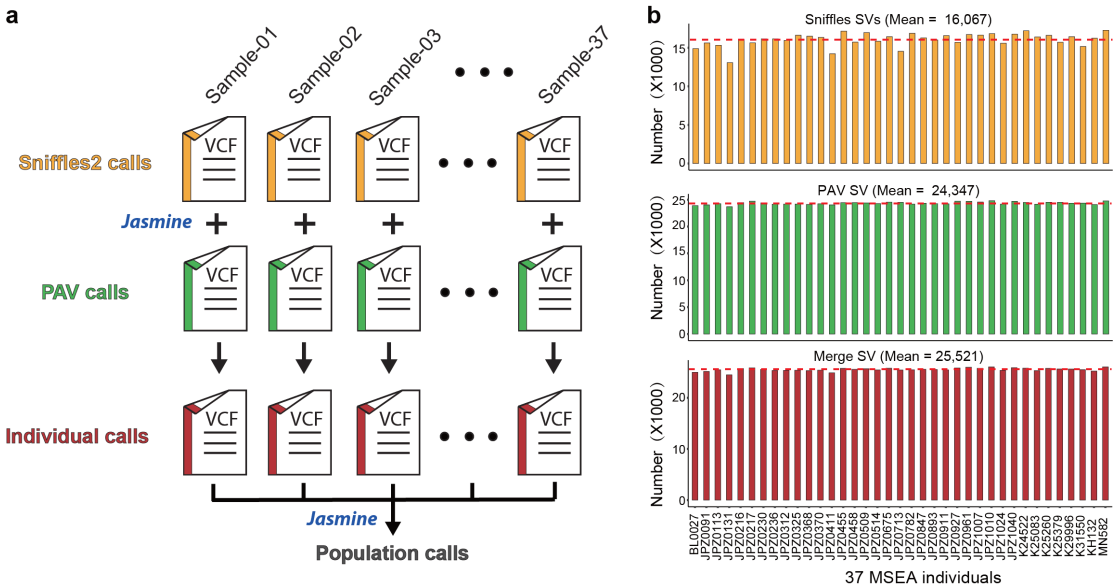


Figure R5. Schematic diagram of SV merging. **a)** The SV calls from Sniffles2 and PAV of each individual were merged using Jasmine, and then we merged the 37 individual calls to obtain the population calls (here refers to simple SVs). **b)** The counts of individual SVs in the Sniffles2 calls, the PAV calls and the merged calls. The mean SV counts per individual at each step are labeled and denoted using dashed red lines.

We employed Jasmine under the default parameters to merge the calls by two different tools, as well as the individual calls, except that to be conservative, we changed the ‘*min_overlap*’ (the reciprocal overlap for two SVs) from the default value of 0 to 0.5, and the used parameters are listed below:

Parameters	Value	Description
max_dist_linear	[0.5]	- the proportion of the length of each variant to set distance threshold to
max_dist	[inf]	- the maximum distance variants can be apart when being merged
min_dist	[100]	- the minimum distance threshold a variant can have when using max_dist_linear
min_overlap	[0.5]	- the minimum reciprocal overlap for DEL/INV/DUP SVs
kd_tree_norm	[2]	- the power to use in kd-tree distances (1 is Manhattan, 2 is Euclidean, etc.)
min_seq_id	[0]	- the minimum sequence identity for two insertions to be merged
k_jaccard	[9]	- the kmer size to use when computing Jaccard similarity of insertions
max_dup_length	[10k]	- the maximum length of duplication that can be converted to an insertion

min_support	[1]	- the minimum number of callsets a variant must be in to be output
spec_reads	[10]	- the minimum number of reads a variant needs to be in the specific callset
spec_len	[30]	- the minimum length a variant needs to be in the specific callset

When merging SVs, it is important to determine for a given SV pair whether or not the two SVs are sufficiently close together in terms of their breakpoints to be considered the same SV. The Jasmine tool sets a default parameter for distance threshold of 100 bp (*min_dist*), which is highlighted in the snapshot above.

As expected by the reviewer, our previous callsets had relatively low numbers of SVs per sample (15,139 SVs), and the updated callsets generated higher number of SVs per sample (25,522 SVs), which is comparable with the previous results from HGSVC or HPRC. For example, HGSVC1 identified 22,755 per sample using SMRT-SV caller (Audano et al., 2019). HGSVC2 identified 24,653 SVs per sample using PAV caller (Ebert et al., 2021), and HPRC identified 16,800 SVs and 17,400 SVs per haplotype on average in the MC graph and PGGB graph, respectively (Liao et al., 2023).

Using these new SV sets of MSEA populations, we updated all downstream SV analyses, including SV novelty evaluation, SV annotation, SV genotyping in the population-scale short-read samples and the MSEA-enriched SVs. All these updates have been described in the revised manuscripts.

11. “falling in the annotated regulatory regions” - no reference given; what annotation was used here?

Authors response: We annotated functional consequence of the identified PS-SNVs with Variant Effect Predictor (VEP) (McLaren et al., 2016) in the Ensembl database (<https://www.ensembl.org/Tools/VEP>). The VEP tool annotates regulatory regions based on the Ensembl Regulatory Build (Zerbino et al., 2015), which uses data from large-scale projects like ENCODE (EncodeProjectConsortium, 2012), NIH Epigenomics Roadmap (Romanoski et al., 2015), and BLUEPRIT (Adams et al., 2012). It mainly includes a series of epigenetic markers observed in various cell types, including H3KMe1, H3KMe3, H3K27Ac, open chromatin by DNase hypersensitivity, ATAC-seq, and CTCF binding sites by Hi-C data.

We have rephrased the descriptions and references about annotations in the revised manuscript (**Page 34, Line 839-848**).

12. Same as above: please also include the HPRC set in your evaluation of novelty for your SV callset.

Authors response: As we explained above, in the revised version, we updated our SV results using Sniffles2 and PAV caller by the mapped HiFi reads or the assembled contigs of 37 MSEA samples to the human reference genome GRCh38 (see **Methods**). In the end, we identified 94,658 simple SVs, including 38,531 deletions, 55,073 insertions, 134 duplications, 920 inversions. Then, following the reviewer’s suggestion, we evaluated the SV novelty using public SV callsets from HGSVC3 (freeze4 version, including 176,231 SVs):

http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC3/working/20240415_Freeze4/, and HPRC (version 1.1, 79,781 SVs after collapsed by Truvari for decomposed VCF file from Minigraph-Cactus graph based on GRCh38): <https://s3-us-west-2.amazonaws.com/human-pangenomics/pangenomes/freeze/freeze1/minigraph-cactus/hprc-v1.1-mc-grch38/hprc-v1.1-mc-grch38.vcfbub.a100k.wave.vcf.gz> by adopting a criterion based on reciprocal overlap of at least 50% (see details in **Methods**). Only deletions and insertions were included in SV novelty evaluation. Consequently, we found 68,982 SVs (73.7%) have been reported in HGSVC3 and HPRC (**Figure R6**), a high concordance with the previous SV calls, implying that our SV calls are reliable. Of note, although HGSVC3 and HPRC cover 65 and 47 worldwide individuals (including 18 and 6 Asian individuals, respectively), we still identified 24,622 (26.3%) novel SVs in our data that have not been reported in the public datasets, suggesting MSEA populations are underrepresented in the current SV call sets.

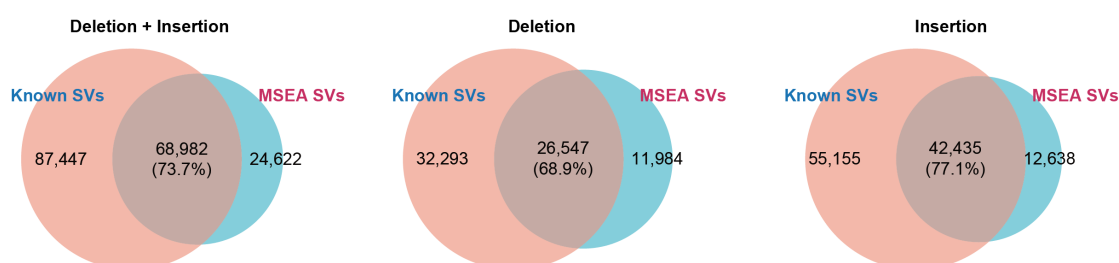


Figure R6. The novelty evaluation of the updated SV callset in MSEA populations. The HGSVC3 and HPRC SV callsets are considered as the reference panel. The detailed information of evaluation procedure is provided in the Methods section of the revised manuscript. The proportion of overlapped SV in MSEA SV set are labeled in the intersected regions.

Besides of simple SVs (only one type in a sequence region, *i.e.* deletion, insertion, duplication and inversion), we also called the complex SVs (typically comprise several internal simple SV subcomponents) using SVision-pro caller (Wang et al., 2024). It turns out that each person carries 158 complex SVs on average, consistent with the previous reports (Lin et al., 2022; Wang et al., 2024). However, considering there is no reported set of complex SVs for global populations, we did not include them in the SV novelty evaluation.

In the revised manuscript, we provided detailed information of SV calling, merging, filtering and assessing in the Methods section.

13. “they are comparable with the two versions of the reference genome with an average quality value of 50.31”. The T2T genome is virtually error-free and has an estimated error rate of 1 per 10 Mbp (with local uncertainties), which is far beyond a QV value of ~50. A “comparison” would simply indicate that the assemblies are worse, but this is obviously as expected, so the relevance of your statement is unclear to me. Also, the GRC reports the hg38 N50 online as ~67.8 Mbp - how did you get to your number of ~58 Mbp?

Authors response: We reassembled our 37 assemblies into 74 phased haploid genomes as suggested by Reviewer #1 (Comment-1). Even so, considering these assemblies are only partially phased assemblies solely relying on the HiFi-read data, their qualities are incomparable with GRCh38 or T2T-CHM13. Therefore, we rephrased the sentence by removing the word “comparable” in the revised manuscript.

For the inconsistent problem of N50 for the hg38 assembly, the contig N50 of hg38 assembly is 57.9 Mb, reported from NCBI: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.40/ (**Figure R7a**). This N50 value is similar to the statistics (GRCh38 contig N50 = 56.41 Mb) from Telomere-to-Telomere (T2T) Consortium (Nurk et al., 2022) (**Figure R7b**). We speculate that the 67.8 Mb N50 the reviewer mentioned is scaffold N50 of hg38, rather than contig N50. Considering only contig N50 was produced for our HiFi-read only assemblies, for a fair comparison, we used contig N50 to make comparison between our assemblies and GRCh38.

a

GRCh38 assembly statistics (from NCBI)

	RefSeq	GenBank
Genome size	3.1 Gb	3.1 Gb
Total ungapped length	2.9 Gb	2.9 Gb
Gaps between scaffolds	349	349
Number of chromosomes	24	24
Number of organelles	1	1
Number of scaffolds	470	473
Scaffold N50	67.8 Mb	67.8 Mb
Scaffold L50	16	16
Number of contigs	996	999
Contig N50	57.9 Mb	57.9 Mb
Contig L50	18	18
GC percent	41	41
Assembly level	Chromosome	Chromosome

b

GRCh38 assembly statistics (from Nurk et al. 2022)

STATISTICS	GRCH38	T2T-CHM13
Summary		
Assembled bases (Gbp)	2.92	3.05
Unplaced bases (Mbp)	11.42	0
Gap bases (Mbp)	120.31	0
Number of contigs	949	24
Contig NG50 (Mbp)	56.41	154.26
Number of issues	230	46
Issues (Mbp)	230.43	8.18

Figure R7. Assembly statistics of GRCh38 from NCBI and T2T consortium. The two tables are snapshots of the assembly statistics of the GRCh38 genome from NCBI (https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.40/) (Panel-a), and the published paper of T2T-CHM13 (Nurk et al., 2022) (Panel-b). The value of contig N50 of the GRCh38 assembly is highlighted.

14. Your conclusion that the unmappable sequences are likely (!) novel seems far-fetched. There are inherently difficult regions to align to (e.g., on chr16, the acrocentrics, centromeres etc.), hence not everything that does not align can be assumed to be novel in its entirety. Including the incomplete hg38 reference in this analysis, which is full of unresolved sequence gaps, does not support your statement at all, at least not at the level of depth you are presenting here. To make this more plausible, you would have to check the unaligned sequences in more detail and confirm a correct assembly. Next, you would need to find the regions where these sequences putatively should align to (assuming that the unaligned sequences are large enough to reason about their relative placement in the assembled chromosomes) and show that there is indeed substantial sequence divergence relative to the reference. Given that consortia such as the HGSC and the HPRC have released numerous assemblies, a claim of “novelty” is still prohibitive unless you also check against those datasets; a relevant point if you put your work in perspective relative to “the current pangenome of human populations”.

Author response: We agree with the reviewer that the depth of HiFi-read data we presented in this study is insufficient to accurately characterize the novel sequences of MSEA samples relative to the current reference genomes (e.g. GRCh38 or T2T-CHM13), or to the reference pangenomes (e.g. assemblies from HGSC or HPRC). Besides, given the main task of this study is to describe the variation spectrum of MSEA populations using deep short-read and long-read data, rather than to assemble the reference genomes, or to construct the pangenome, or to evaluate the novel sequence of them. Therefore, we removed the statement of “novel sequence” in the revised manuscript.

15. I do not understand what insight the short paragraph on the minigraph/pangenome construction is supposed to deliver.

Author response: Given the limitations of the HiFi read only data, it is difficult to build a high-quality pangenome using our current data, and at the same time, pangenome reference construction was not the goal of our Phase I genome project. Therefore, we removed this paragraph in the revised manuscript. In the future, we plan to collect trio samples from Southeast Asian populations, which can help to build high-quality pangenomes, the goal of our Phase II project.

16. “12 non-coding PS-SNVs falling in the known” - your conclusion about the concerted adaptation is very strong w/o additional evidence. How do you know that these regulatory regions are actually relevant in this context?

Author response: We understand the reviewer’s concern that the regulatory function

of non-coding PS-SNVs need be tested if we propose that concerted adaptation plays a dominant role in this region. Therefore, we re-analyzed this region (Chr1:152033580-152476512, in total 443 Kb). We found this region contains 262 PS-SNVs with MSEA-enriched alleles, including 11 coding PS-SNVs (missense), and 12 non-coding PS-SNVs located in enhancers or promoters based on ENCODE database, which are also annotated as regulatory variants in the Ensembl database (**Figure R8a**, added as **Figure 4b** in the revised manuscript). The LD pattern among the PS-SNVs fits the expectation of an extended MSEA-specific haplotype resulting from selective sweep (**Figure R8b**, added as **Supplementary Figure 20a** in the revised manuscript), suggesting there might be multiple functional variants and they form a MSEA-specific haplotype (**Figure R8c**, added as **Supplementary Figure 20b** in the revised manuscript) to affect protein function and/or gene expression regulation. However, future functional experiments need be conducted to confirm the causality.

Combined with the comments of Reviewer-3 that the conception of concerted adaptation is not a new and main contribution of this study, we removed the previous statement about concerted adaptation in the revised manuscript, and tuned down the interpretation of putative function of the identified PS-SNVs.

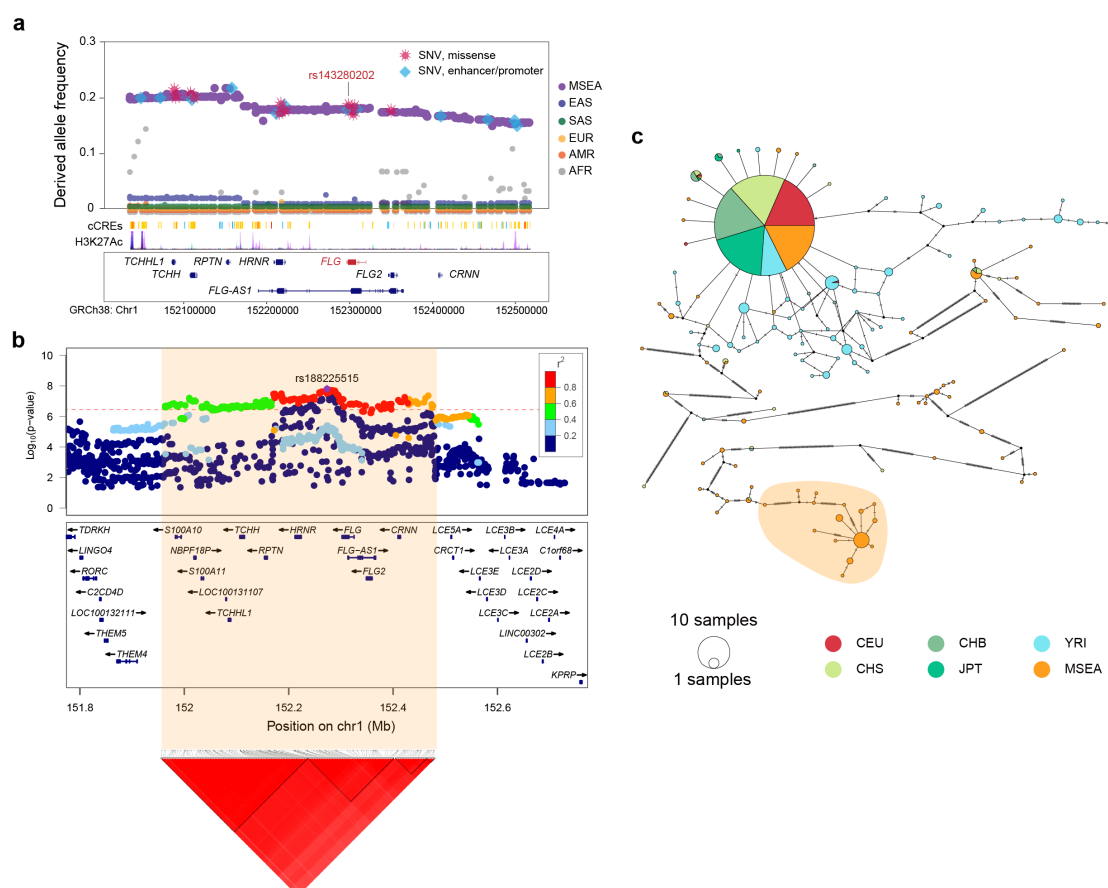


Figure R8. Signature of natural selection in the *FLG* gene region. a) The Manhattan plots showing the differences of the derived allele frequencies (DAF) of the PS-SNVs in the *FLG* region among MSEA and other populations from 1KGP. **b)** The regional plot of CMS scores and recombination rates in the *FLG* region, in which the peaks indicate the selective signals. The peak SNVs are marked with colors. The bottom panel

shows the LD blocks of the 262 PS-SNVs with MSEA-enriched alleles. c) Median-joining network of the *FLG* gene region showing a MSEA-specific haplogroup. Each node represents a haplotype, and the size is proportional to its frequency. The MSEA-specific haplogroup is highlighted.

17. I do not find a proper description of how you obtained the functional label in Table S10 that seems the basis for your functional interpretation in that section (how you went from the GO enrichment to the one label shown). In particular, I am also unclear about the background distribution of GO terms in the 44 non-coding PS-SNV regions you are investigating. How many other genes are in those regions? What is their function?

Authors response: Actually, we did not perform GO enrichment analysis for each positive selection region (PS-region). Previously, we presented the functional label of each candidate PS-region by choosing the functional term of the lead gene (based on GeneCard database) in each PS-region (the gene mapped or nearest to the peak PS-SNV), and the chosen term was relevant to the trait putatively undergone tropical adaptation.

Given this strategy might bring artificial bias with missing information of the other genes located in each PS-region, we re-conducted the functional annotations of the 44 PS-regions. Firstly, we tabulated the PS-regions and the genes identified in these regions. Those genes flanking the upstream and downstream 500Kb regions of the peak PS-SNV, and only the genes covering at least one PS-SNV in a PS-region were considered as the mapped genes in a PS-region (see **Methods**). To avoid overstatement, in the revised manuscript, we did not draw conclusions on the phenotypic outcomes of these genes. The functional annotations of the lead gene were obtained from the GeneCards database, the GWAS Catalog (<https://www.ebi.ac.uk/gwas/>) and Human Phenotype Ontology (HPO, <https://hpo.jax.org/>). Only the top five traits/diseases with 'Best Score' were presented, and all functional terms of these lead genes can be found in GeneCards database (<https://www.genecards.org/>). This information has been updated in **Supplementary Table 9**.

In addition to the lead gene in each PS-region, we also annotated the function of all genes located in PS-regions, including 89 genes mapped to the 44 independent candidate PS-regions. Gene functions are annotated by phenotypes from GWAS Catalog, HPO and MGI mutant phenotypes using GeneCards database (<https://www.genecards.org/>). This information has been updated in the **Supplementary Table 10**. We presented the top20 of gene functions based on GWAS Catalog, HPO and MGI, respectively, and ranked them based on counts of each functional term. All involved phenotypes or diseases were classified into six categories, including physical/physiological measures (20 terms), immunity/inflammation response (14 terms), metabolism measures (9 terms), nervous system (8 terms), cardiovascular system (3 terms) and other traits (4 terms) (**Supplementary Table 10**).

Considering we found a significant excess of genes related to body height, with 21 height-associated SNVs in the PS-SNV set, and 59 genes out of 89 genes in the 44 PS-regions, we performed a permutation test to evaluate the statistical significance

compared to the 44 randomly generated independent regions in the genome. The results showed a statistically significant enrichment of body-height-related genes under selection (P -value = 0.005 based on 1000 permutation tests, **Supplementary Figure 18b**).

18. “suggesting the potential involvement of ... no significant GO enrichment was detected”. Since you did not provide any additional evidence, I think “suggesting” is too strong here. It is certainly a viable hypothesis, though.

Authors response: Considering that we updated our SV set using more comprehensive approaches, we re-analyzed the GO enrichment analysis using the new SV set (involving 529 associated genes). We found one functional term (GTPase activator activity) was significantly enriched (FDR = 0.006). We updated the information in the revised manuscript and added as **Supplementary Table 13 (Page 21, Line 500-502)**.

Table R1. GO enrichment results of genes associated with the MSEA-enriched SVs. The table only shows the top 10 terms, and the complete information of the GO results can be found in **Supplementary Table 13**.

Category	Term	Gene counts	P Value	Bonferroni P value	Benjamini P value	FDR
Molecular function	GO:0005096~ GTPase activator activity	19	1.09E-05	0.006	0.006	0.006
Biological process	GO:0090630~ activation of GTPase activity	11	1.83E-05	0.032	0.033	0.033
Cellular component	GO:0031012~ extracellular matrix	15	1.38E-04	0.059	0.060	0.060
Molecular function	GO:0005085~ guanyl-nucleotide exchange factor activity	14	6.52E-04	0.309	0.184	0.184
Cellular component	GO:0045202~ synapse	21	1.02E-03	0.360	0.223	0.222
Biological process	GO:0016236~ macroautophagy	8	3.73E-04	0.487	0.333	0.333
Cellular component	GO:0005886~ plasma membrane	127	3.19E-03	0.753	0.436	0.434
Cellular component	GO:0005829~ cytosol	127	4.62E-03	0.868	0.436	0.434
Cellular component	GO:0005813~ centrosome	23	5.96E-03	0.927	0.436	0.434
Cellular component	GO:0046581~ intercellular canaliculus	3	6.53E-03	0.943	0.436	0.434

19. “One notable example is a 7439 bp deletion ...” - a notable example for what? For tropical adaptation? What is the evidence to make this a “notable” example?

Authors response: We chose this 7439-bp deletion as the representative case because it is the largest and the top SV showing allelic divergence between MSEA and other

populations. It is relatively common in MSEA (14%), but rare in East Asians (3.5%) and Central Asians (0.8%), and absent in other world populations (**Figure R9**, as **Figure 4d** in the revised manuscript). To be clear, we rephrased the sentence in the revised manuscript.

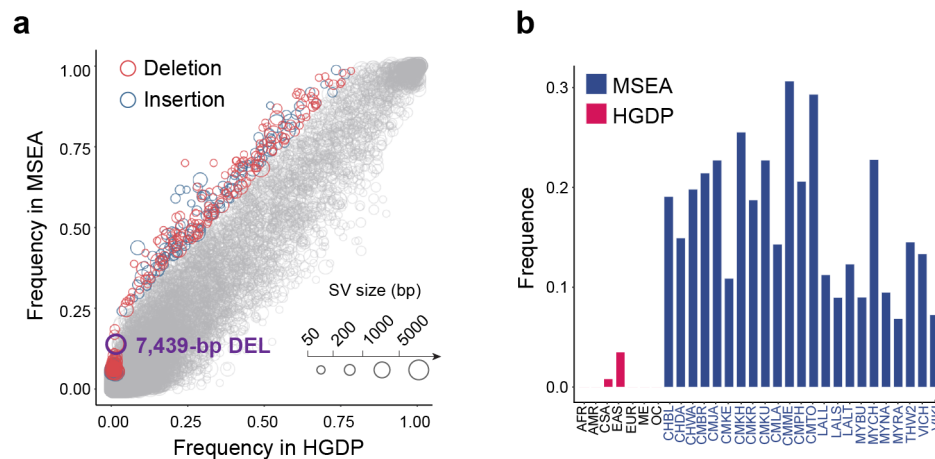


Figure R9. A 7,439-bp MSEA-enriched deletion. a, The dot plot showing the MSEA-enriched SVs. The 785 candidate SVs are highlighted since they are enriched in MSEA populations compared to other global populations. The deletions and insertions are denoted in red and blue, respectively. The MSEA-specifically enriched SV, a 7,439-bp deletion in the *PEXI4* gene is labeled by purple circle. **b,** Comparison of allele frequencies of the 7,439-bp deletion shows regional enrichment in MESA compared to the HGDP populations. EAS-East Asians, AMR-Americans, CAS-Central South Asians, EUR-Europeans, AFR-Africans, ME-Middle East, OC-Oceania.

20. In general, the results you are presenting that aim at emphasizing the importance of the identified variants for adaptation or medical conditions lack evidence of actually having a functional effect (i.e., are there detectable changes in gene expression, be it because of a direct or an indirect effect via differential regulation?).

Authors response: We agree with reviewer that our data is insufficient to establish assertive connections between the detected selective signals and the adaptive traits, and more functional and large-scale population studies need to be conducted in the future. In the revised manuscript, we conducted more functional annotations using the public databases, including GWAS Catalog for genes/variants associated to human traits/diseases, GTEx for variants effect on gene expression and ENCODE for variants involving regulation function. Also, to be objective in interpreting the selective signals, we tuned down the previous claim of tropical adaptation since we did not perform functional experiments to validate the speculated functional/phenotypical relevance, which will be a separate project in the future. In the revised manuscript, we objectively listed the variants/genes/regions with putative signals of positive selection in MSEA populations, together with the annotated functions and traits/disease of these selected genes based on previous studies and public databases. We also tested polygenic selection on multiple traits using a series of statistics, and found that short stature has been subject to polygenic selection in MSEA populations.

Minor:

Some of your sentences could be restructured and shortened to ease readability. I am assuming you are going to revise the text anyway, so I am not highlighting all of them; just to give you two examples from the Introduction: “Mainland Southeast Asia (MSEA, also known as Indochina Peninsula), as the ...” and “Utilizing the SEA3K dataset, we characterized ...”.

Author response: We appreciate the reviewer’s advice for restructuring the sentences. We have checked and polished the text in the revised manuscript.

When you first introduce a new data type in the text, please be more explicit about the technology: “...high-coverage (average 43.5×) WGS of 3,023...”. I can conclude from the number of individuals that this is a short-read dataset, but please make that explicit.

Author response: According to the reviewer’s suggestion, we checked and presented these descriptions in an explicit way in the revised manuscript.

Reviewer-3

This paper by He et al. presents the SEA3K genome dataset (Phase I), consisting of 3023 short-read WGS samples and 40 long -read WGS samples from 30 mainland Southeast Asian (MSEA) populations.

The authors performed a series of standard population genetic analyses based on the short-read WGS data, including characterization of variants, imputation accuracy, population structure, demographic history, and selection. In addition, based on the long-read WGS data, the authors characterized structural variants (SVs) and constructed the first pangenome of MSEA populations. This SEA3K genome dataset is a valuable resource because MSEA consists of rich information of human genetic diversity and MSEA populations are currently underrepresented in human genetic studies.

Overall, it seems to me that the two components of this dataset, the short-read and long-read sequencing data, had little connection in this paper. When analyzing the short-read sequencing data, the authors didn’t make full use of their advantage of having a MSEA pangenome from long-read data. And the paper was not organized in a logical way – the long-read WGS data were inserted in the middle of short-read WGS results without clear connection. Ideally, the paper should start with a pangenome constructed from long-read sequencing data, including SVs. Then, short-read sequencing data should be mapped to the pangenome, followed by variant calling and all the downstream analyses. Thus, short-read WGS data analysis can take full advantage of the long-read data. It would be very interesting to find out how much the long-read data can facilitate short-read data analysis. For example, can we perform better in calling SVs when the short-read WGS data were mapped to the pangenome, and subsequently imputing SVs when

using the short-read WGS data as the imputation reference panel? Otherwise, why do we need a pangenome?

Authors response: We understand the reviewer's point that constructing a pangenome reference of MSEA populations will bring many advantages for exploration of genomic variants, including improvement of the read mapping, variant calling and discovery of population-specific sequences without reference biases (Gao et al., 2023; Liao et al., 2023; Wang et al., 2022). However, achieving this goal would require multi-platform data and complex manual curations to make a high-quality pangenome meeting the criteria of a reference genome (Liao et al., 2023). Although the Human Pangenome Reference Consortium (HPRC) has assembled a high-quality pangenome based on 47 fully phased, diploid assemblies based on abundant data and comprehensive analyses, this pangenome still contains assembly errors and unfixed gaps, and HPRC only considered it as a draft human pangenome reference (Liao et al., 2023). At the same time, although a human pangenome reference can be created with many efforts, the current algorithms and tools for efficient analyses of graph-based genome are still under development. The application of graph-based pangenome is still challenging, for example, variant discovery from pangenome-based mapping and calling, especially the normalization of "fuzzy" breakpoint of structural variants (Liao et al., 2023; Wang et al., 2022).

In this study, due to the limited volume of blood samples, we only generated long-read HiFi data for the 37 MSEA individuals. Currently, we are incapable of generating fully-phased assemblies to construct a high-quality pangenome reference for MSEA populations. In the absence of a high-quality pangenome reference, any downstream analyses would be unreliable. In addition, the construction of a pangenome reference for MSEA populations was not the goal of our Phase-I genome project. In this study, we aimed to provide a genome-wide landscape of genetic variations, leveraging the 3,023 deep-coverage short-read WGS data for SNV/INDEL discovery, together with the 37 deep-coverage long-read WGS data for SV discovery. Therefore, to keep our results robust and better highlight our major contributions, in the revised manuscript, we removed the section of pangenome, and instead we focused on SV discovery using the long-read sequencing data. In the future, we plan to collect trio samples of MSEA populations so that a high-quality pangenome can be constructed, which is the goal of our Phase II genome project.

Other specific comments:

1. Imputation. Figure 1E and Figure S2 are highly similar to Figure 5 of Wu et al. 2019 (Cell, PMID: 31626772). Clearly, the authors had read the paper and tried to follow. But why didn't they make a direct comparison with Wu et al. (2019), which presented the Phase I dataset of the SG10K Project, consisting of WGS data of 4,810 Singaporeans, including 903 Malays? The paper Wu et al. was only cited improperly as "small-scale genome data in Malays" in the Introduction (line 2 on page 5). The SG10K Project has been expanded to 9,051 genomes (Nat Commun in 2022, PMID: 36335097) but was not even mentioned. Did the SEA3K dataset performed better than

the SG10K (or the Phase I dataset by Wu et al.) on the imputation in Asian populations? In addition, it is recommended to provide key information of the reference panels, especially the sample size. For example, what is the sample size of the GAsP panel? There are two versions of GAsP on the UM imputation server, consisting of 1,654 and 6,461 individuals, respectively. Which one did the authors used?

Authors response: Following the reviewer's suggestion, we obtained the SG10K Phased I dataset (4,810 samples) (Wu et al., 2019), and we compared its imputation power with our SEA3K dataset. We show that the SEA3K panel outperforms the SG10K panel in Southeast Asians for all evaluation methods (Log2FC of error rate, R^2_{Pec} and R^2_{Est}) (**Figure R10**), so does it for some of the East Asian populations (e.g. Dai, Lahu, Yi, Naxi and Miao). In contrast, the SG10K panel outperforms the SEA3K panel for South Asian, European and some East Asian populations (e.g. Han, Japanese, Mongolian and She) (**Figure R10a**). This is expected, because the two main populations of the SG10K Phase I dataset are Indians and Chinese (Wu et al., 2019). Both SEA3K and SG10K (Phase I) showed poor performance for Africans (**Figure R10a** and **Supplementary Figure 3**), highlighting the necessity of population-specific reference panel for accurate genotype imputation. These results have been added in the revised manuscript.

Unfortunately, the 9,051 genomes of SG10K (Chan et al., 2022) is currently unavailable, though the SG10K consortium provided a Imputation Server based on these 9,051 samples (<https://npm.a-star.edu.sg/>), but the service was currently only available for Singaporean researchers. Therefore, we could not compare it with our SEA3K dataset. However, given its larger sample size and more diverse samples, we would expect a better imputation power than the SG10K (Phase I) panel for Asian populations. Considering the SG10K (Phase I) and SG10K_Health Project are two important datasets for genotype imputation of Southeast Asian population, we cited these two papers in the revised manuscript.

For the problem of lacking key information of imputation panel, following the reviewer's recommendation, we added detailed information of the reference panels (e.g. sample size, version and website of imputation server, and the corresponding references) in the revised manuscript.

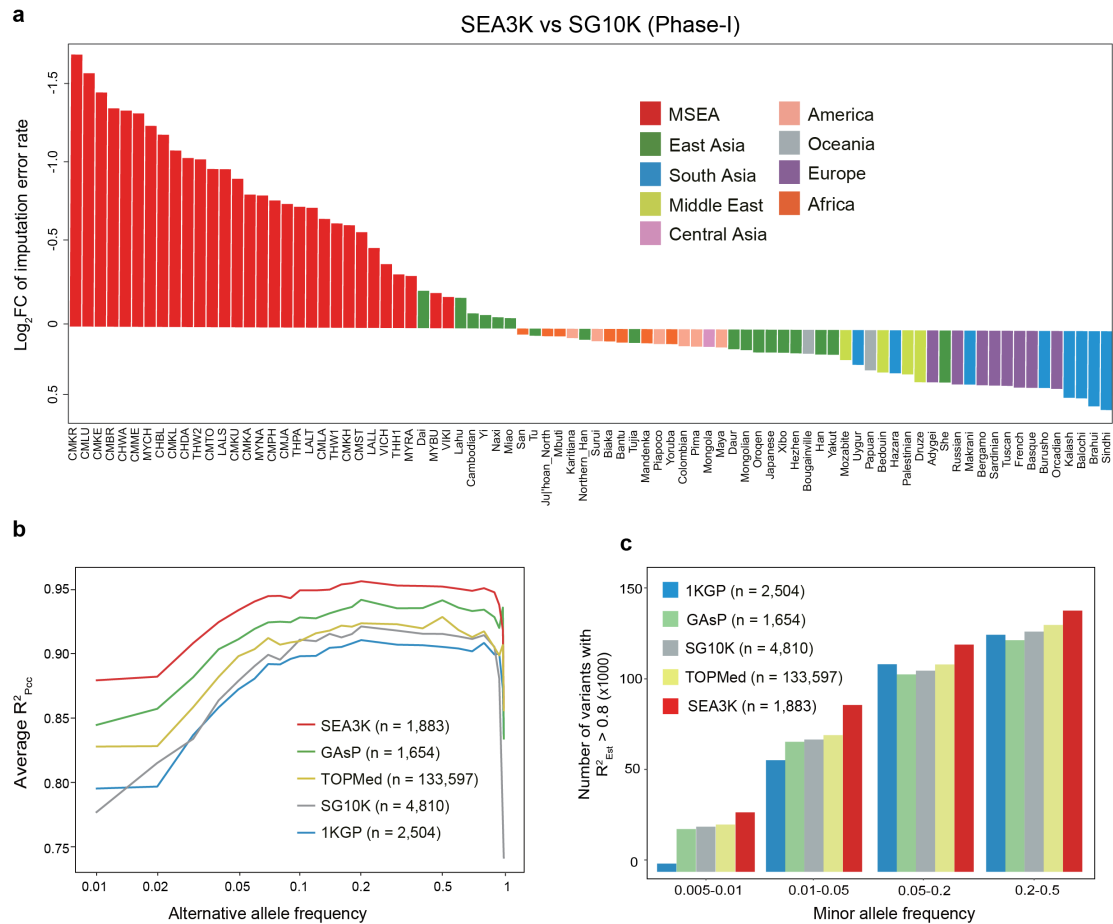


Figure R10. Comparisons of imputation performance between the SEA3K panel and the SG10K (Phased I) reference panel by multiple statistics. a) Comparison of the fold changes ($\log_2FC$) of the imputation error rates for the worldwide populations using the SEA3K panel and the SG10K Phased I panel. **b)** Comparisons of imputation accuracy of Southeast Asian populations using squared Pearson correlation coefficient (R^2_{Pcc}) between the SEA3K panel and the other panels (1KGP, TOPMed, GAsP and SG10K). **c)** Comparisons of high-quality imputed variants (using estimated imputation accuracy from Minimac4: $R^2_{Est} > 0.8$) in different frequency ranges for Southeast Asian populations between the SEA3K panel and the other panels (1KGP, TOPMed, GAsP and SG10K). The sample size of each panel is labeled in the plots.

2. Again, on imputation. What is the target population of Figure 1F? There is no clear description in the figure caption, main text and even in the method section. Based on the method section, it seems the target population is 300 individuals randomly selected from the SEA3K dataset? But this information was only mentioned in a different paragraph. Let's assume this is correct. Were these 300 individuals phased together with the other SEA3K samples? If so, they were not independent and might cause overestimation of the imputation accuracy of the SEA3K reference panel.

Authors response: We understand the reviewer's concern. In this analysis, we randomly subsampled 300 individuals from the 2,183 unrelated individuals of the SEA3K dataset to create a target panel, and then we used the remaining 1,883 samples

to conduct haplotype phasing independently, and utilized it as a proxy for the new Southeast Asian reference panel. To avoid confusion, in the revised manuscript, we provided clear descriptions in **Methods** and **Results**, as well as in the caption of **Figure 1f**.

3. Population structure. In Figure 3A, the authors showed that the PCA plot mimics the geographic map followed by some rotation transformation. This can be done in a formal and quantitative way using the Procrustes analysis (please see Wang et al. 2012 PLoS Genet, PMID: 22927824).

Authors response: According to the suggestion of the reviewer, we applied the Procrustes analysis by the R package *vegan* (<https://github.com/vegandevs/vegan>), to quantitatively evaluate the similarity between the PCA map and the geographic map.

Figure R11 shows the optimal transformation that maximizes the similarity between the PCA map and the geographic map according to the *vegan* analysis. Procrustes M^2 , the sum of squared differences between the two ordinations, is 0.732, and the significance of similarity is $P\text{-value} < 1 \times 10^{-5}$ based on 100,000 permutation tests. Consistent with the previous finding by studying populations from different geographic regions (Wang et al., 2022), the MSEA populations show significant similarity between genes and geography. The PC1 and PC2 from the rotated PCA map mirror the west-east cline and the south-north cline, respectively.

In the revised manuscript, we updated the PCA map based on the Procrustes analysis as **Figure 3a**, and cited the paper of Wang et al. (2012) in the reference.

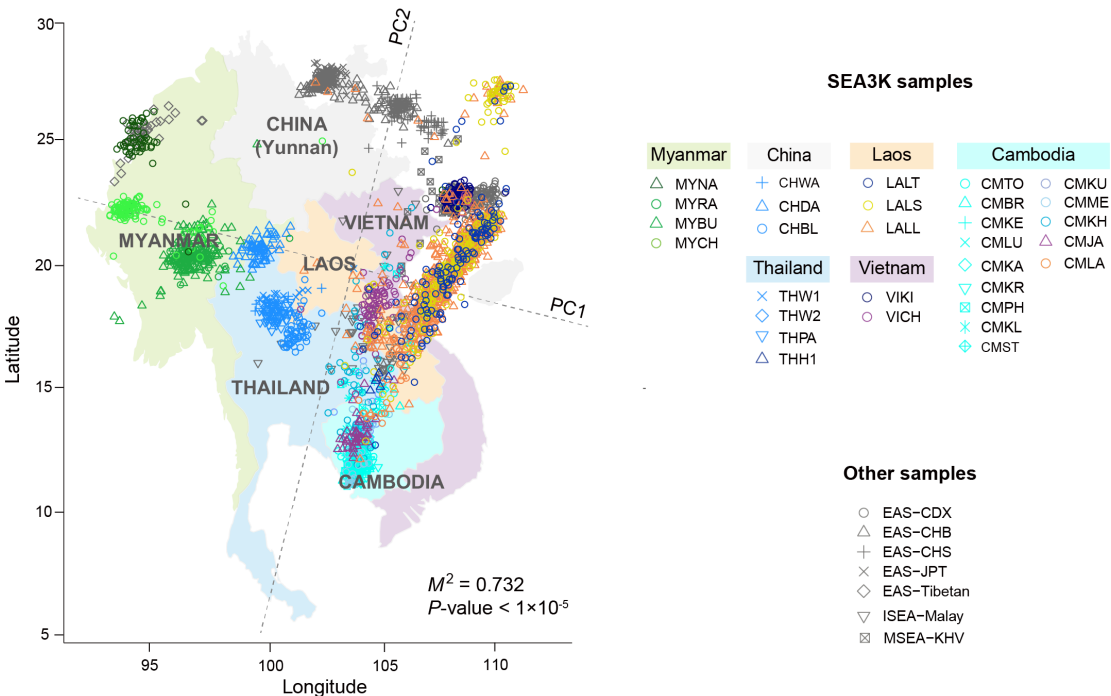


Figure R11. Procrustes analysis of genetic and geographic coordinates of MSEA populations. The Procrustes-transformed PCA plot of genetic variants is presented here. The Procrustes analysis was conducted by the R package *vegan* using unrelated 3,678 individuals of 45 populations from MSEA and East Asia. PC1 and PC2 are

indicated by the dotted lines, which account for 1.02% and 0.53% of the total variance, respectively. The Procrustes similarity $M^2 = 0.732$ (P -value $< 1 \times 10^{-5}$ and the rotation angle of the PCA map is 14 degrees clockwise.

4. Population structure. The authors over-interpret results from ADMIXTURE analysis with demographic history, such as the origin of different ancestry components and the evolutionary timing (second paragraph on page 13). ADMIXTURE is just a non-supervised clustering algorithm. It is well-known that many different demographic models can yield the same clustering patterns in the ADMIXTURE results (please see Lawson et al. 2018 Nat Commun, PMID: 30108219).

Authors response: We agree with reviewer that interpretation of ADMIXTURE should be cautious, because the same admixture history plot inferred by ADMIXTURE might result from different demographic events, such as recent admixture, bottlenecks or unobserved ancient admixture. Therefore, following the reviewer's suggestion, we employed badMIXTURE (Lawson et al., 2018) to assess the goodness of fit of the model in order to avoid misleading interpretations of the ADMIXTURE results.

In brief, to exam the four major components observed in MSEA populations, we set up the $K = 4$ for ADMIXTURE, and ran badMIXTURE to visualize the model fit according to the recommended analysis procedure (Lawson et al., 2018). The results showed that there is no systematic pattern to the residuals (**Figure R12**, also added as **Supplementary Figure 12** in the revised manuscript). Consequently, the results by ADMIXTURE constitute robust inferences of the ancestries.

Besides, to avoid over-interpreting the results of ADMIXTURE mentioned by the reviewer, in the revised manuscript, we removed the assertive interpretations for each component in MSEA populations, including origin of different ancestry components and the evolutionary timing, and we only described the ancestry composition objectively based on the ADMIXTURE plot. The reconstruction of population demographic history in MSEA is another story and will be explored in the future.

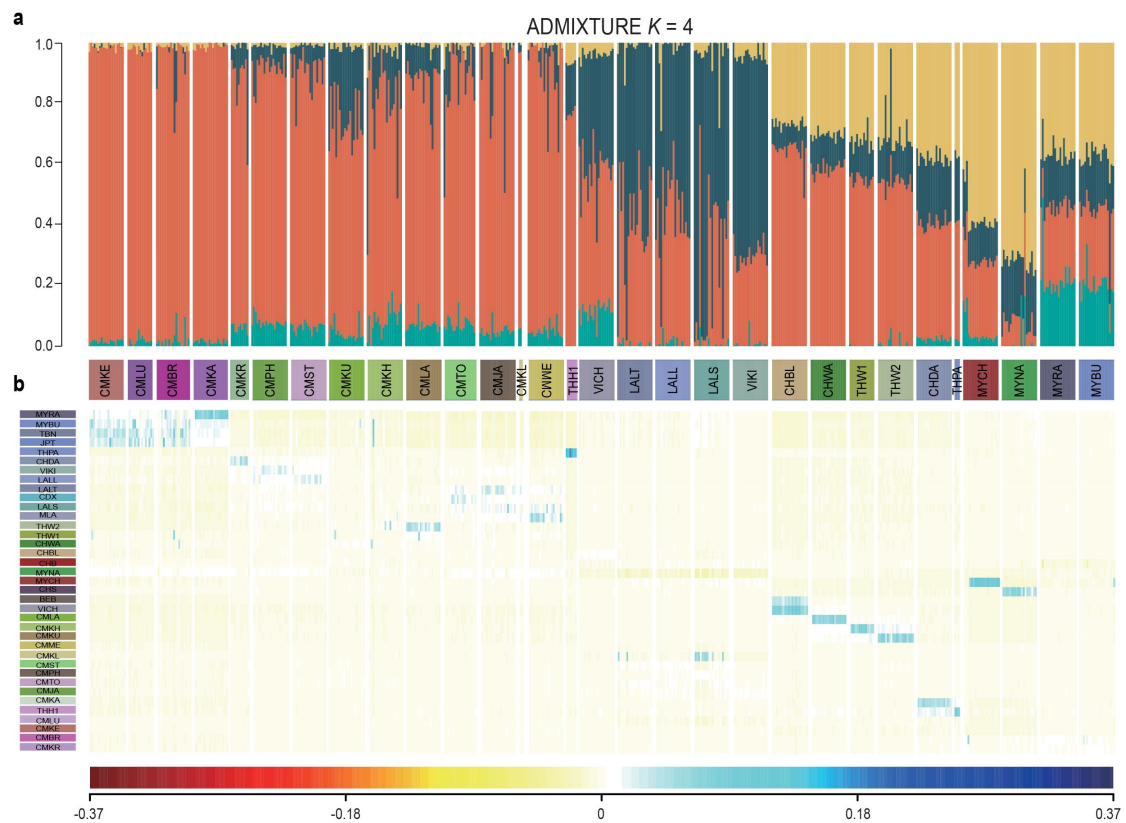


Figure R12. Inspection of model fitting by badMIXTURE at $K = 4$. a) Ancestry proportions estimated through ADMIXTURE. b) The residual palette under the best-fit ancestral population admixture model. No systematic pattern of residuals was observed.

5. Positive selection. The authors tried to claimed novel selection signals at the SNV level (for example, they claimed 7 out of 13 SNVs were novel selection signals). Traditionally, positive selection signals were reported at the gene level because genes are the functional unit on which selection is acted on. Thus, claiming novelty at the SNV level might be inappropriate and overstated. Better to report how many genes or loci harbor novel selection signals.

Authors response: According to the suggestion of the reviewer, we updated the presentation, and we assessed the novel selection signals based on genes rather than variants, and we rewrote the part of selection signals in the revised manuscript (**Page 17, the section of natural selection**).

6. Medical implications. The authors provide a list of “pathogenic” variants that are common in the SEA3K dataset but rare in other populations, and a list of homozygous PTVs found in the SEA3K populations. But they didn’t explain clearly the medical implications and why these are “valuable resource”. For the “pathogenic” variants common in the SEA3K population, the author only provided an example of balancing selection at the HBA2 gene. Are they suggesting that all the others are likely to be under balancing selection as well? It seems a more plausible explanation is that many of variants were misclassified as “pathogenic” by the ClinVar database because they can

reach a high frequency in MSEA populations. Regarding the homozygous PTVs, they should examine related phenotypes in the carriers in order to understand the functional consequences of these “natural knockouts” in humans.

Authors response: We appreciate the reviewer’s constructive comments. For “pathogenic variants”, we agree with the reviewer’s comments that these high-frequency variants might be misclassified as “pathogenic” by the ClinVar database. Classification of variants as either benign or pathogenic is challenging without diverse population sets, given that the effect of disease-causal variants can vary across populations. This point has been proven in a recent study of All of Us Research Program with diverse group of 245,388 individuals across USA (All of Us Research Program Genomics, 2024). Therefore, in the absence of appropriate population reference datasets, allele frequencies can be misclassified and may lead to false disease associations or toward overestimating pathogenicity (Shah et al., 2018). This observation highlights the significant medical implications and valuable resource for the SEA3K dataset, which have also been reported in the genomic data of non-European populations, including H3Africa Project (Choudhury et al., 2020) and GenomeAsia Pilot project (GenomeAsia, 2019).

Among the 10 identified “pathogenic” variants specifically enriched in MSEA populations, most of them were likely misclassified as “pathogenic” in the ClinVar database. However, there might be exceptions because pathogenic variants under balancing selection can also show high frequency in regional populations. In the manuscript, we provided a well-described balancing selection case, i.e. the Thalassemia risk variants that are common in African populations because they provide resistance to malaria (Verrelli et al., 2002). For MSEA populations, among the 10 enriched “pathogenic” variants, the *HBA2* variant (rs41464951) might be true pathogenic since there have been reported data showing its association with alpha Thalassemia (Efremov et al., 1990; Kanavakis et al., 1996; Kimura et al., 2009). Consequently, its high frequency in MSEA populations is better explained by balancing selection and it may contribute to malaria resistance in MSEA populations.

For homozygous PTVs, it is not plausible to examine the related phenotypes in human PTV carriers, because we did not collect the phenotypes of the sampled subjects. However, we did know that no serious diseases were observed in our sampled subjects, suggesting that these homozygous PTVs should not be lethal. Nevertheless, the phenotype outcomes of these homozygous PTVs are worth to be studied in the future to dissect their phenotypic and/or pathogenic consequences.

As the reviewer suggested, we rephrased this part in the revised manuscript (**Page 26, Line 623-627**).

7. Discussion, page 28, “In MSEA populations, we showed that standing variations play a dominant role”. I am not sure which result support this claim?

Authors response: In fact, we distinguished the mode of selection for the 5,505 PS-SNVs based on their allele frequency spectrum. In other words, if the derived allele frequency of a selected variant is common (>5%) in MSEA but rare (<1%) in other

global populations, we classified them as the mode of selection due to *de novo* mutations (SDN). If the derived allele frequency >1% in global populations, we classified them as the mode of selection due to standing variant (SSV). As expected, we observed that >99% (5,459/5,505) mutations under selection are in the SSV mode, and only 46 shows the mode of SDN (**Figure R13**). This is consistent with previous findings that compared to SDN, SSV is much more prevalent in human local adaptations (Fan et al., 2016; Rees et al., 2020). A reasonable explanation is that the time duration of local selection on non-Africans is relatively short (<60,000 years), and there has been not enough time for the emergence of many *de novo* mutations, and most of the variants were inherited from our African ancestors.

Considering that we just made a simple classification of selection mode by frequency difference, without further statistical test (i.e. the ages of the selected alleles or the selection coefficients acting on them). Therefore, we tuned down the description of mode of selection in the Discussion section of the revised manuscript (**Page 29, Line 696-711**).

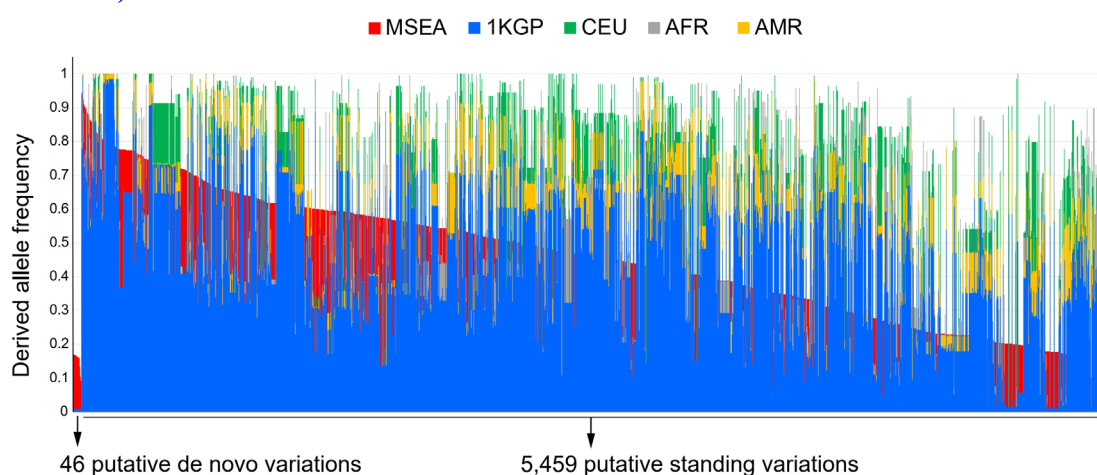


Figure R13. Classification of the 5,505 PS-SNVs in MSEA populations based on frequency pattern. The x-axis refers to the 5,505 PS-SNVs, and the 46 variants under SDN and the 5,459 variants under SSV are shown, respectively. The y-axis refers to their derived allele frequencies in MSEA populations (in red) and European populations (in green), African populations (in gray), American populations (in orange), and the average frequencies of 1KGP dataset (in blue).

8. Discussion, page 28, regarding the contribution of regulatory variants and protein-coding variants to adaptive evolution. Can the authors provide a recent reference on the debate, in addition to reference 81, which was published in 1975? My impression is that the research community has already reached an agreement that both types variants are important in the evolution since over 10 years ago. Furthermore, it seems an overclaim of “a concerted adaptation attributed to both protein functional alterations and changes of gene expression regulation” because it is not surprising to see a locus with both types of variants, especially when strong positive selection can lead to extended haplotypes.

Authors response: We agree with the reviewer that the conception of concerted adaptation in protein functional alterations and changes of gene expression regulation

is not new, and certainly not the main contribution of this study. Therefore, we removed this part in the revised manuscript.

9. Quality controls. PCA and ADMIXTURE should be performed on unrelated samples. But it seems that the authors only excluded related samples for the selection analysis. Closely related samples are likely to show as outliers in the PCA, which might explain some of the PCA outliers identified in the QC. If the authors excluded one sample from a related pair first, the other one would not show as an outlier.

Authors response: We agree with the reviewer's comments. As the reviewer suggested, we updated our PCA and ADMIXTURE analyses only using the unrelated samples (n = 2,183) in the revised manuscript. We observed that the results of PCA and ADMIXTURE based on all samples (n = 3,023) and on unrelated samples only (n = 2,183) are similar.

10. QC for long-read sequencing. Three samples (MN269, MN422, and MN582) were removed in the QC because of low sequencing quality and abnormal GC contents (page 39). But why were these low-quality samples included in the analysis of genome assembly? In the genome assembly analysis, MN024, MN269 and MN422 were removed based on assembly metrics, but MN582 was retained. Despite the low sequencing quality of MN582, the authors included it as one of the 37 "high-quality" assemblies.

Authors response: It was a typo in the main text. As we presented in **Supplementary Figure 5**, MN269, MN422 and MN024 were excluded because of poor quality. The reason we included these three samples in assembly was to check the effect of genome assembly from poor long-read sequencing data. In the revised manuscript, we rephrased in the Methods that we only assembled 37 genomes out of 40 samples and the three samples with poor quality were excluded.

11. Introgression analysis. This study used the low-coverage WGS data of the 1KGP (phase 3 version 5a). Why didn't they use the high-coverage WGS data of 1KGP (Cell 2022, PMID: 36055201)? In particular, the authors "assumed that YRI is monomorphic for the SEA3K major allele for variants called in the SEA3K but not called in the 1KGP" (page 41, second last paragraph), and this assumption might not hold because low-coverage sequencing data have lower sensitivity to detect rare variants. Using the high-coverage 1KGP data will be more suitable, especially when the SEA3K WGS data are high-coverage.

Authors response: We understand the reviewer's concern on low-coverage data for introgression detection. The reason why we adopted the low-coverage WGS data of 1KGP is that the version is based on hg19 coordinates, and it is the mostly used version in the published studies of archaic introgression studies, considering all currently released genotypes of archaic humans were based on the hg19 coordinates. Therefore, to keep the coordinates consistent with previous studies and to make the results comparable, we utilized the low-coverage WGS data of 1KGP to detect the

introgression signals.

We agree that low-coverage sequencing data have lower sensitivity to detect rare variants, thus the low-confident genotype of YRI-specific variants will affect the evaluation of archaic introgression.

Following the reviewer's suggestion, we re-analyzed the archaic introgression of MSEA populations using high-coverage short-read WGS data of 1KGP (Byrsk-Bishop et al., 2022), and we found the results are similar between these two versions, including: 1) the identified introgressed sequences (a high consistency between the two versions, Pearson correlation $R = 0.956$, $p = 7.1\text{e-}23$ for Neanderthal introgression, $R = 0.964$, $p = 1.3\text{e-}24$ for Denisovan introgression) (**Figure R14**); 2) the inferred introgression waves (both low-coverage and high-coverage detect the same five MSEA populations showing significant two-pulse Denisovan introgressions) (**Table R2**); 3) the adaptive introgression detection (both low-coverage and high-coverage detect five Neanderthal introgressed signals and two Denisovan Introgressed signals, respectively).

In the revised manuscript, we updated all results of archaic introgression based on the results using the high-coverage short-read WGS data of 1KGP.

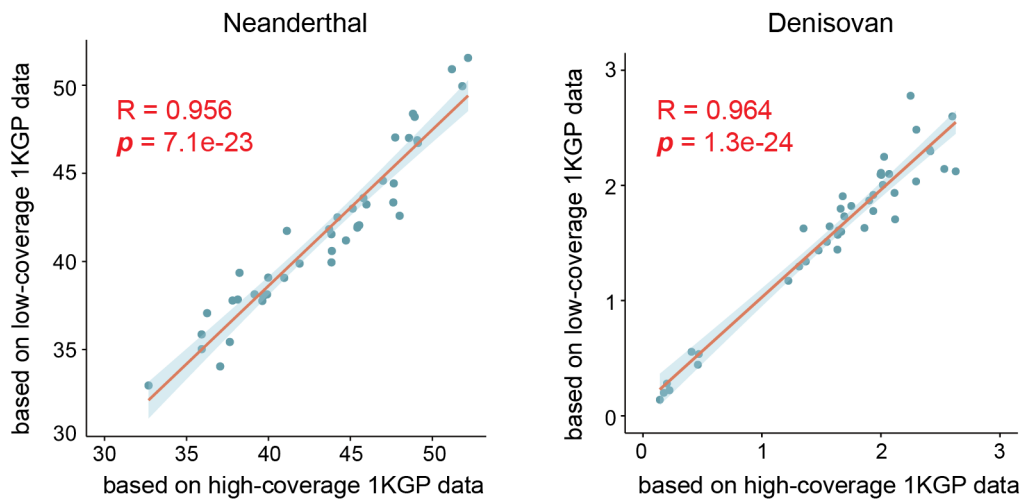


Figure R14. Comparison of the detected introgressed sequences per individual in a population based on the high-coverage and the low-coverage 1KGP data. The dots represent the 23 investigated MSEA populations. The results of the Altai Neanderthal (left panel) and the Altai Denisovan (right panel) are shown separately.

Table R2. Comparison of the two distinct Denisovan components in the 23 MSEA populations based on the high-coverage and the low-coverage 1KGP data.

Population	Based on high-coverage 1KGP data						Based on low-coverage 1KGP data					
	$\mu1$	$\mu2$	$\sigma1$	$\sigma2$	p	P-value	$\mu1$	$\mu2$	$\sigma1$	$\sigma2$	p	P-value
LALS	0.48	0.86	0.18	0.06	0.76	1.9E-05	0.46	0.86	0.18	0.06	0.76	1.6E-06
LALL	0.52	0.86	0.18	0.06	0.82	1.1E-04	0.52	0.86	0.18	0.06	0.8	4.2E-05
LALT	0.34	0.78	0.2	0.1	0.6	4.4E-04	0.36	0.8	0.22	0.1	0.66	1.6E-03
VIKI	0.46	0.8	0.18	0.06	0.72	5.1E-04	0.44	0.82	0.2	0.06	0.72	8.5E-05
MYCH	0.52	0.88	0.16	0.04	0.82	8.7E-04	0.5	0.88	0.18	0.06	0.78	1.1E-03
THW2	0.5	0.86	0.14	0.06	0.74	2.9E-03	0.48	0.88	0.18	0.06	0.78	5.3E-03
CHWA	0.52	0.9	0.18	0.04	0.86	0.013	0.5	0.88	0.16	0.06	0.8	7.3E-03

CHBL	0.48	0.86	0.18	0.06	0.78	0.025	0.5	0.86	0.16	0.06	0.76	0.020
CMTO	0.48	0.86	0.18	0.04	0.88	0.035	0.46	0.86	0.18	0.04	0.86	0.006
CMKH	0.5	0.84	0.14	0.06	0.84	0.053	0.5	0.86	0.16	0.06	0.88	0.060
MYRA	0.52	0.84	0.16	0.08	0.78	0.054	0.54	0.9	0.18	0.04	0.9	0.074
MYBU	0.54	0.9	0.16	0.06	0.88	0.057	0.54	0.88	0.16	0.06	0.88	0.183
CMKE	0.44	0.82	0.2	0.06	0.82	0.061	0.48	0.84	0.18	0.06	0.82	0.054
CHDA	0.54	0.9	0.18	0.04	0.9	0.063	0.52	0.9	0.18	0.06	0.86	0.067
CMBR	0.38	0.86	0.22	0.06	0.86	0.068	0.4	0.86	0.2	0.06	0.84	0.025
VICH	0.48	0.86	0.2	0.04	0.92	0.112	0.5	0.88	0.2	0.04	0.92	0.105
CMKU	0.46	0.9	0.2	0.04	0.92	0.135	0.46	0.88	0.18	0.06	0.92	0.424
CMLA	0.42	0.8	0.2	0.08	0.82	0.148	0.44	0.84	0.2	0.06	0.86	0.073
CMJA	0.52	0.86	0.18	0.04	0.92	0.189	0.52	0.86	0.16	0.04	0.86	0.023
CMPH	0.5	0.84	0.18	0.08	0.84	0.214	0.52	0.88	0.18	0.06	0.9	0.358
MYNA	0.5	0.82	0.14	0.1	0.7	0.248	0.46	0.74	0.1	0.12	0.52	0.338
CMME	0.48	0.82	0.14	0.1	0.84	0.382	0.5	0.84	0.12	0.08	0.82	0.184
CMKH	0.5	0.82	0.14	0.08	0.84	0.400	0.5	0.82	0.14	0.08	0.84	0.516

*The five populations with significant *P*-values after Bonferroni correction ($0.05/23=0.0022$) are highlighted.

Minor points:

1. Page 15. “Linkage disequilibrium (LD) decay indicates the genetic structure and associates with genetic diversity of a population.” --- the meaning of this sentence is unclear. In general, LD decay pattern is a consequence of demographic history, including admixture, effective population size, and many others.

Authors response: We agree with reviewer that decay of LD is affected by many demographic events, except for the indication of genetic diversity, and we rephrased the interpretation of LD-decay pattern in the revised manuscript (**Page 16, Line 374-375**).

2. Figure 5B, it might be better for visualization to plot the contour plots with color indicating the density heights.

Authors response: According to the reviewer’s suggestion, we updated the contour plots with colors to indicate the density heights (see **Figure 5b** and **Supplementary Figure 22** in the revised manuscript).

3. Figure 5C, it is very hard for me to see the green lines (introgression segments). What is the difference between dark green and light green in the plot? I just don’t understand the meaning of this plot.

Authors response: Actually, there is no dark green and light green in the plot, they were vision disparity because of density of lines (i.e. the introgressed segments). To present more explicitly, we updated the plot with a MSEA-combined version, rather than a population-separated version. We utilized a barplot to show the ratio of the introgressed segment in each MSEA population (see **Figure 5c** in the revised

manuscript).

4. Figure 5D, it is almost impossible to visualize the tiny blue bar indicating the Neanderthal sequence, especially when printed on an A4 paper.

Authors response: We redid the plot and highlighted the Neanderthal haplotypes in the revised manuscript (see **Figure 5d** in the revised manuscript).

5. PCA, page 35, line 3. Please define “outliers” (i.e., how many SDs from the mean?).

Authors response: The original presentation is “we performed a series of PCAs by gradually removing ‘outliers’ on the basis of a plot of the PC1 and PC2 and reanalyzing the remaining samples on the basis of the same set of SNVs”. Actually, it is a way to describe that we gradually zoom in the global PCA map into local region involving MSEA populations, which helps to explore the substructure and subtle relationship among populations we focused. Therefore, there is actually no outliers. To be more explicit, we rephrased the sentence in the revised **Methods**.

6. Selection, pages 37-38, please stated clearly how the P values were defined in the DAF, PBS, XP-EHH and iHS analysis. Current method section only describe how the statistics were calculated but the CMS was calculated based on P values.

Authors response: We calculated *P*-values of each statistic (DAF, PBS, XP-EHH and iHS) by Chi-squared distribution. Each statistic is represented by its $-\log_{10}(P\text{-value})$. For composite multiple signals (CMS), we used Fisher’s method (the formula is shown in **Methods**) to combine all the four statistics together and calculated an overall *P*-value by Chi-squared distribution.

To be more explicit, we added the details in the revised **Methods (Page 38, Line 967-969)**.

7. Duplicated references of 32 and 47. And all the consortium names were not displayed properly in the reference list. Please check your Endnote format and the references carefully.

Authors response: We have corrected the reference duplication in the revised manuscript.

8. There are many typos, just to list a few:

(1) Page 13, line 5, “from Sino-Tibetans from the north” → “to Sino-Tibetans from the north”

(2) In Figure 3D, the y-axis ticks, “ESA” → “EAS”

(3) Page 15, first line of 2nd paragraph, “examed” → “examined”

(4) Page 18, Figure 4G, the y-axis label, “Frequence” → “Frequency”

(5) Page 37 line 11, “previous study” → “a previous study”

Authors response: In the revised manuscript, we have carefully checked the text and have corrected these typos.

Reviewer-4

The authors present results from a variety of analyses on genomic samples collected from 30 pops in six southeast Asian countries including persons from minority Chinese communities. Their goal was to build a reference panel based on individuals underrepresented in previously established panels. The authors find genetic diversity not represented in previous panels.

If we assume that the data was collected ethically and within boundaries of modern genomic methodology, the analyses presented here may be informative. However, the interpretation of results and subsequent inference made are unfit for publication.

Data & Methodology

The analyses generally appear to be robust and without obvious flaw. The authors are applauded in their gaining informed consent, and the inclusion of authors from sampled countries. However, problematic elements of this manuscript may have been avoided if members of subject minority MSEA communities were included as collaborators and coauthors.

Authors response: We totally agree with the reviewer's opinion about ethical framework, and the suggestion for inclusion of representatives from the minority MSEA communities as collaborators/coauthors in the manuscript.

In establishing the Consortium of Anthropological Research in Southeast Asia and Southwest China (CASEAC), one core objective of generating the SEA3K dataset was to promote international collaborations, instead of "helicopter research". When we started the project, we contacted and collaborated with members of indigenous groups and local scientists frequently, and we exchanged the information and discussed the sampling design together. We achieved an efficient collaboration with local scientists and members of the minority MSEA communities. We had a face-to-face discussion of genetic results in March, 2023 at Chiang Mai of Thailand, and in December, 2023 at Kunming, China, respectively. At the same time, we started a Joint Education Program to train the young local scientists for bioinformatics skill, providing the opportunity for them to lead local research projects in the future, and Dr. Somphad Sochampa in our co-author list is an example for the training program.

According to the reviewer's recommendation, besides of the inclusion of local scientists as co-authors, we added 7 members from the minority MSEA communities as collaborators and coauthors. All added authors have been informed and we obtained their consent. In the revised manuscript, we stated their contributions on ethnic approval work, personnel coordination and volunteer organization in the collection work.

Mr. Kittisak Rattanakrajangsri is the president of Asia Indigenous Peoples Pact

(AIPP), a regional organization founded in 1992 by Asian indigenous peoples' movements (<https://aippnet.org/>).

Ms. Chutima Morlaeku is the president of Inter Mountain Peoples Education and Culture in Thailand Association (IMPECT) (<https://www.iwgia.org/en/iwgia-partners/64-inter-mountain-peoples-education-and-culture-in-thailand-association.html>), an indigenous and tribal non-governmental development organization founded and staffed entirely by representatives of indigenous communities involved in the Association.

Both of them provided help for sample collection and other local affairs.

Dr. Angkhana Inta is of Northern Thai descent, she has been working closely with Austroasiatic and other ethnic groups in Thailand.

Dr. Wittaya Pongamornkul belongs to the Lawa ethnic group and has spent his life in the Lawa community.

Both have been crucial in providing information for selecting our Austroasiatic samples.

Mr. Xai Xiong, Mr. Tany Sengvilai, and Mr. Somchin Sitthideth are three members of the minority community in Laos, who belong to the three ethnic groups of Laos in this study: Lao Song, Lao Loum and Lao Theung, respectively. In the revised manuscript, we added them into the CASEAC given their contribution in this study and the upcoming collaboration for indigenous groups in Laos.

Results and Appropriate Statistics

Most figures are well designed and informative. The usefulness of Figure 3E is not apparent. Despite the figure caption stating that it shows “historical changes of effective population sizes (N_e) of MSEA populations over time,’ in reality, the figure does not reflect that; instead it compares this measure between Tribal and non-Tribal populations. For what scientific purpose is the comparison of IBD between Tribal and non-Tribal IBD? On page 16 the authors state that “some tribal populations... kept ancient genomic features inherited by their hunter and gatherer ancestors.” This statement is not supported by a citation, and it is irrelevant to note that a human population retains DNA from their ancestors. The authors desired goal to produce a panel for MSEA populations is not consistent with comparisons of measures between Tribal and non-Tribal peoples. While measures of IBD can be informative, when applied to comparisons of Tribal and non-Tribal populations, it generates the need to examine why the authors found that to be relevant or interesting. Historical abuses and contemporary social power differentials between Tribal and non-tribal peoples around the globe create a reality in which researchers must be or become aware of the stigma and damages of these kinds of comparisons.

Authors response: Firstly, we agree with reviewer’s comments that Figure 3E was not sufficiently informative, and we removed this panel in the revised manuscript.

For IBD analysis, the initial motivation was to see the genetic architectures among populations with different characters. For example, since we included many indigenous populations mostly with small population sizes and high levels of isolation, we would expect that endogamy paired with these populations will greatly increase IBD scores, as reported in the published studies of Southeast Asian and South Asian populations (Kutanan et al., 2021; Nakatsuka et al., 2017). Therefore, we simply compared Tribal and non-Tribal groups to test the influence of small population size to IBD estimate. It turned out that tribal groups had IBD scores that were skewed upwards from nontribal groups, consistent with the result reported by GenomeAsia 100K Project, in which they compared the IBD scores between Tribal and non-Tribal groups of Asians (GenomeAsia, 2019). On the other hand, for comparison of Tribal and non-Tribal groups in effective population size (N_e), we wanted to compare hunter-gatherers (mostly tribal populations) and agriculturalists (mostly non-tribal populations) in MSEA, which can help evaluating the scale of the demographic expansion from East Asia into Southeast Asia during the Neolithic transition to farming (Lipson et al., 2018; McColl et al., 2018).

To avoid potential ethical concerns, we removed the comparison results between Tribal and non-Tribal groups in the revised manuscript, including Figure 3E and the IBD comparison. Considering the result of the IBD measurement itself is highly informative in understanding population genetic structure, we kept the population-level comparison without grouping them by tribal or non-tribal.

The portion of the manuscript referring to “tropical adaptation” is flawed. In the most sophomorically erroneous portion, the authors make several troubling claims. They suggest that curly hair “is probably a local adaptation to tropical heat since curly hair may provide better ventilation.” They go on to support this statement by citing a paper by L’Oreal researchers that outlines geometric measures to describe human hair strand morphology. The De la Mettrie et al. paper does not in any way support or claim to support a causal relationship between curly hair and an evolutionary adaptation to tropical heat.

Authors response: We agree with the reviewer’s comments on overstating “tropical adaptation” based on our natural selection results.

Southeast Asians have been living in tropical rainforest, and adapted to diverse selection pressures such as the environment with high temperature, high humidity, strong ultraviolet radiation and pathogen exposure, which drives their local adaptation and shapes a series of adaptive traits among them (Fan et al., 2016). However, currently almost all reported studies of “tropical adaptation” have been focused on African populations, and “tropical adaptation” of SEA are understudied. The biggest obstacle is lack of genomic data of SEA populations, which serves as the foundation for solving the problem of genetic mechanism of adaptive traits. Consequently, we tried to explore the “tropical adaptation” using our MSEA dataset. Of course, we agree with the reviewer that the current dataset and our analyses do not provide sufficient evidence for the conclusion of “tropical adaptation”.

For the trait of curly hair, we agree that the cited study by De la Mettrie et al. only presented the classification hair type with curl index (hair morphology are classified into eight types with regard to curliness, with straight hair as type-I and the curliest hair as type-VIII), rather than explored the causal relationship between curly hair and evolutionary adaptation. However, the De la Mettrie et al. paper measured 1,442 subjects from 18 different countries, and showed that Southeast Asians mainly have mildly curly hair (78.00%, type-II and III hair), more prevalent than that in East Asians (55.00%). Additionally, during our field work, we observed a large number of our MSEA participants with curly hair while we were sampling, although we did not make a statistical assessment. These two points set our motivation to discuss the curly hair and evolutionary adaptation of MSEA populations.

As for the relationship between curly hair and tropical adaptation, a thermoregulatory hypothesis has been previously proposed. Jablonski and Chaplin suggested that the tightly curled hair—which is common in many African populations—may have an advantage in reducing heat gain from exposure to sunlight (Jablonski and Chaplin, 2014). A recent study presented experimental evidence on the potential evolutionary function of human scalp hair and variation in its morphology, and concluded that the tightly curled hair provides the most effective protection for the scalp against solar radiation, while minimizing the need for sweat to offset heat gain (Lasisi et al., 2023). However, there is no experimental data to bridge a causal relationship between curly hair and evolutionary adaptation to tropical heat.

In conclusion, as the reviewer mentioned, we realized it is improper to draw assertive conclusions that the genes we identified under selection are responsible for specific traits of tropical adaptation (such as curly hair). Functional studies and much larger population studies must be conducted to make that connection. In the future, we plan to conduct an independent study on the genetic mechanism of “tropical adaptation”.

In the revised manuscript, we rewrote the section of selective signals by objectively describing the known functions of the identified variants/genes under positive selection, without drawing assertive conclusions on their contribution to tropical adaptation. In the revised discussion, we emphasized the importance of future exploration in evolutionary explanations of adaptive traits and positive selection in MSEA populations. At the same time, we emphasized the important role of our dataset as the valuable resource and key foundation for these future studies. In the future, we plan to study the functional effects of the identified positive selection variants using both the reverse-genetics approach (base editing in cell lines and organoids), and the forward genetics approach (genome-wide association analysis of MSEA populations by collecting large-scale phenotype data).

Similarly, their claims on short stature are not accurate. Human height is very highly polygenic with the majority of sites being individually of very low effect. Height is highly heritable, but genome-wide studies have failed to identify most loci with significant contributions (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3232052/>). But, it is known that many loci have very small effects; it is unclear that the authors' assertion that shorter height is adaptive based on their results. They assert that their

candidate regions are enriched for loci contribution to height variation, but this is untested: if they sample 44 regions randomly, how many loci that effect height would overlap? The inference made by the authors on their untested claims on stature are inappropriate.

Authors response: Firstly, short stature is another distinctive phenotype thought to be adaptive to a tropical rainforest environment (commonly referred to as a “pygmy” phenotype) with more genetic evidence. It is reported that the short stature trait is an example of convergent evolution in rainforest hunter-gatherer populations across Africa, Asia, and South America (Perry and Dominy, 2009). Selection for small body size may be due to limited food resources, resistance to heat stress, immune response, and/or a trade-off between early onset of reproduction and cessation of growth (Jarvis et al., 2012; Perry et al., 2014; Tucci et al., 2018). Several genes showing natural selection signatures in African Pygmies have been suggested associated with short stature phenotype of Pygmies, including *CISH*, *DOCK3*, *MAPKAPK3*, *GHR*, *IGF1*, *BMP4*, *BMP6*, *ANKRD11*, *TRPS1*, *HESX1* and *ACAN* (Fan et al., 2023; Jarvis et al., 2012; Lachance et al., 2012). These studies on rainforest hunter-gatherer populations in central Africa suggest that their short stature is mostly driven by strong selection affecting a relatively small number of loci with moderate to strong effects, in contrast to the thousands of loci associated with European height variation with each of them only having a small effect (Lango Allen et al., 2010; Yengo et al., 2022).

For short-stature population in South America, Asgari et al. reported a gene with extremely high effect size in Peruvian individuals. They identified a population-specific, missense variant in the *FBNI* gene that is significantly associated with lower height and under positive selection, and each copy of the minor allele reduces height by 2.2 cm (4.4 cm in homozygous individuals). This is the largest effect size known for a common height-associated variant in human populations.

For human pygmy populations in SEA, Tucci et al. tested signatures of recent positive selection on Flores pygmies in Indonesia, but no gene associated to height was detected with significant signature of positive selection. Therefore, they proposed that polygenic selection acting on standing variations might explain their short-stature phenotype (Tucci et al., 2018).

In our study, we agree with the reviewer that very likely the selection of stature in MSEA populations is polygenic and most of the involved variants have small effects. Therefore, we employed systematical analyses to detect polygenic adaptation of multiple traits (body height, skin pigmentation, hair shape and nasal shape) in MSEA populations (see **Methods**). Consequently, we detected a significant signature of polygenic selection on height decreasing in MSEA populations (P -value = 0.004), but not on the other three traits (**Figure R15**, added as **Supplementary Figure 19** and **Table R3**, added as **Supplementary Table 11** in the revised manuscript).

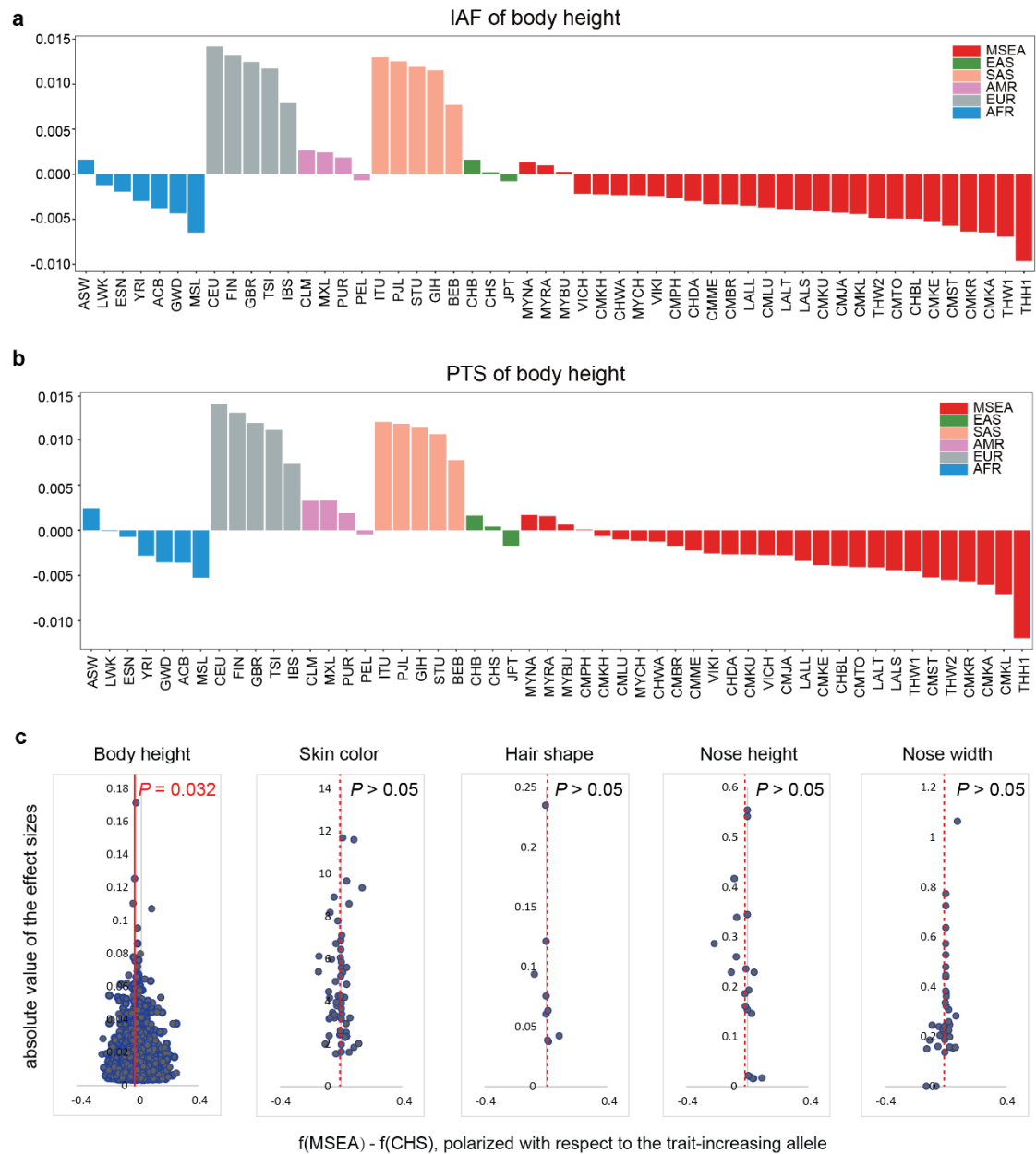


Figure R15. Detection of polygenic effect on multiple traits in MSEA populations. **a)** Comparison of height-increasing allele frequency (IAF) among world populations. The great majority of MSEA populations show negative values. **b)** Polygenic trait scores (PTS) across populations show a similar pattern with IAF. The plots show the mean IAF and PTS of 9,972 independent height-associated SNVs among MSEA and other world populations in 1KGP, and the height-increasing allele was defined as the allele significantly associated with increased height in the GWAS data set. **c)** The absolute values of the effect sizes of the trait-associated SNVs for body height, skin color, hair shape, nose height and nose width, as a function of the difference in frequency between MSEA and CHS (southern Han Chinese): $f(\text{MSEA}) - f(\text{CHS})$. The panel frequencies are polarized with respect to the trait-increasing allele. The vertical lines refer to the mean value, which solid line and dashed lines denote to significant and non-significant P -value under 1000 permutation tests (**Methods**). For body height, the mean value of $f(\text{MSEA}) - f(\text{CHS})$ is -0.00163, which reaches statistical significance

under 1000 permutation tests (P -value = 0.032). This result suggests that the sum of height-increasing allele frequency is significantly depleted in MSEA populations.

Table R3. The results of polygenic adaptation detection of body height in MSEA populations. The asterisk (*) refers to the results when MSEA all populations were combined as one in the analysis. The results of the other traits are provided in **Supplementary Table 11**.

Population	dIAF statistics		beta	Q _B statistics		Q _x statistics	
	f(MSEA)-f(CHS)	P-value		Pval	RandPval	Pval	RandPval
MSEA*	-0.00163	0.032	-2.68	0.007	0.004	0.092	0.044
VIKI	-0.00130	0.055	-3.29	0.001	0.000	0.017	0.014
VICH	-0.00113	0.071	-2.44	0.015	0.011	0.147	0.082
THW2	-0.00249	0.003	-2.78	0.005	0.006	0.063	0.054
CHBL	-0.00256	0.003	-2.73	0.006	0.006	0.072	0.054
LALS	-0.00215	0.012	-3.56	0.000	0.000	0.008	0.006
LALL	-0.00185	0.024	-2.77	0.006	0.004	0.080	0.042
CMTO	-0.00260	0.002	-2.61	0.009	0.010	0.097	0.071
CMPH	-0.00149	0.040	-1.47	0.141	0.115	0.380	0.310
CMME	-0.00184	0.024	-1.57	0.117	0.118	0.355	0.279
CMLA	-0.00270	0.001	-2.48	0.013	0.008	0.132	0.083
CMKU	-0.00213	0.013	-1.40	0.160	0.146	0.396	0.311
CMKR	-0.00325	0.000	-2.25	0.024	0.028	0.168	0.122
CMKH	-0.00130	0.055	-1.58	0.115	0.114	0.376	0.269
CMKE	-0.00253	0.003	-2.36	0.018	0.020	0.144	0.113
CMJA	-0.00227	0.006	-1.50	0.133	0.128	0.377	0.287
CMBR	-0.00180	0.028	-2.02	0.044	0.056	0.238	0.176
CHWA	-0.00115	0.071	-1.23	0.218	0.228	0.446	0.322
CHDA	-0.00166	0.031	-2.15	0.031	0.030	0.208	0.125
MYRA	0.00033	0.340	-1.11	0.269	0.236	0.462	0.373
MYNA	0.00037	0.630	-0.84	0.399	0.419	0.496	0.365
MYCH	-0.00126	0.059	-1.61	0.108	0.080	0.344	0.244
MYBU	0.00005	0.520	-1.10	0.271	0.255	0.476	0.332
LALT	-0.00204	0.017	-3.06	0.002	0.000	0.040	0.017

Additionally, we performed a series of simulation analyses to evaluate the statistical significance of the overlap between the 44 positive selection regions and the height-associated loci from GWAS databases. Among the 5,505 PS-SNVs in the 44 regions, we found a significant excess of variants related to body height, with 21 height-associated SNVs falling in the PS-SNV set (P -value = 0.005 based on 1000 permutation tests, **Figure R16**, added as **Supplementary Figure 18** in the revised manuscript).

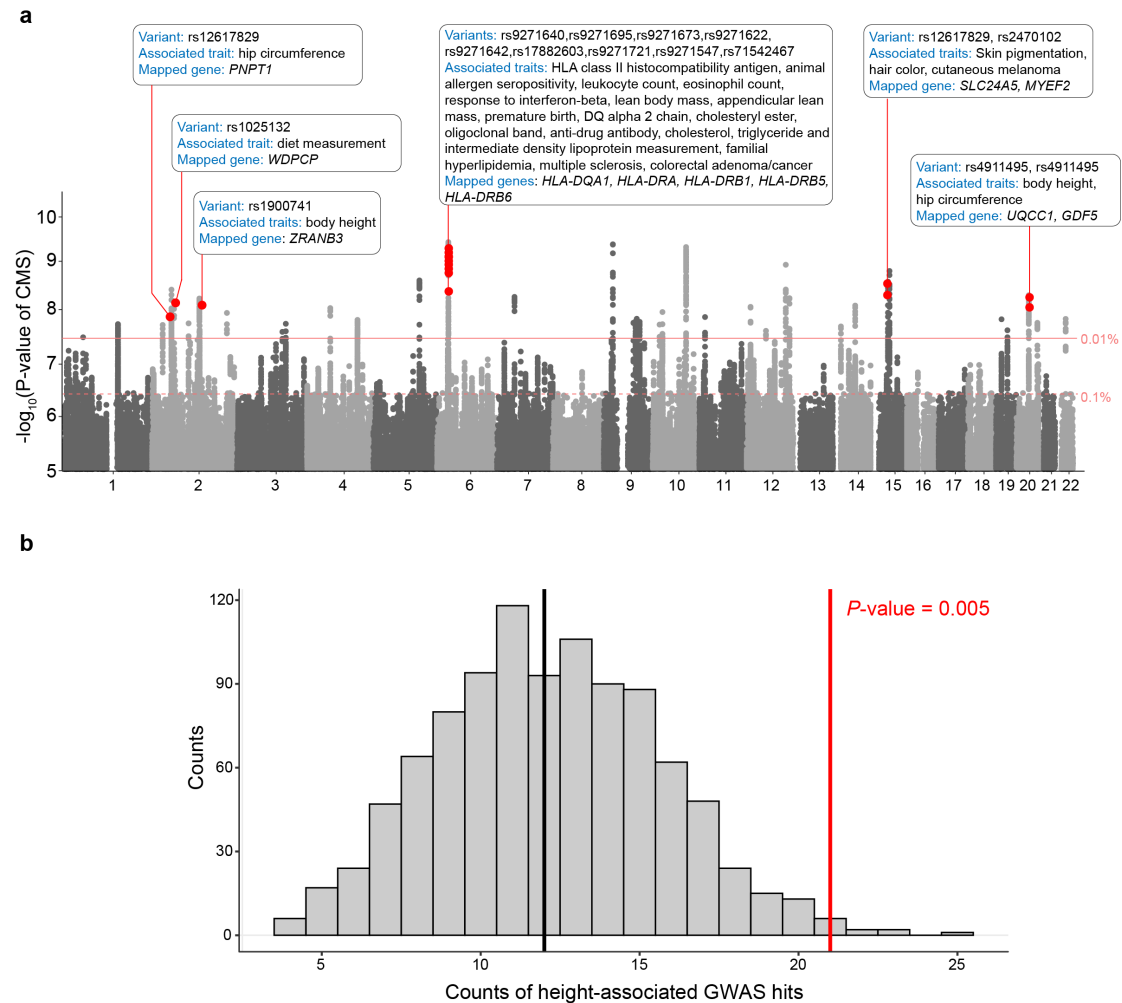


Figure R16. GWAS annotation and simulation of PS-SNVs in MSEA populations.
a) Manhattan plot of the CMS scores of the genome-wide SNVs in MSEA populations. The reported GWAS hits are highlighted by red dots and labeled with GWAS information (variants, associated traits and mapped genes). There are six positive-selection regions harboring the GWAS hits of the top PS-SNVs (with the top 0.01% CMS scores), including two regions covering genes (*GDF5*, *UQC1* and *ZRANB3*) related to body height, one region covering genes (*SLC24A5* and *MYEF2*) related to skin pigmentation, one region covering *PNPT1* related to hip circumference, one region covering *WDPCP* related to diet measurement, and the MHC region related to diverse phenotypes. **b)** Permutation test of height-associated SNVs. Among the 5,505 PS-SNVs in the 44 regions, we found a significant excess of variants related to body height, with 21 height-associated SNVs falling in the PS-SNV set ($P\text{-value} = 0.005$ based on 1000 permutation tests) (see details in **Methods**).

Conclusions

Hair, skin color, and height are known to be very highly polygenic, the few loci reported here would only be a fraction of the genotype contributing to phenotype. The nature of the genes underlying these classically quantitative traits is such that these conclusions

are unsupported under the presented experimental design. Data presented in this manuscript have value and should be in the literature. The interpretation of results and subsequent inference made are unfit for publication. In particular, the claims of genotype to phenotype linkage and adaptation to various biotic and abiotic features of the environments experienced by MSEA peoples are untestable and unsupported by any data presented. Instead, they are based on stereotypes and dated assumptions of physiological benefits of specific traits. It is recommended that the authors of this manuscript revisit their findings to eliminate bias reflecting anachronistic approaches to genetic studies in Tribal populations. See approaches described by Fox et al. 2020 (<https://doi.org/10.3390/genes11010088>) and Claw et al. 2018 (<https://doi.org/10.1038/s41467-018-05188-3>). The addition of MSEA representation in authorship would be a positive step.

Authors response: We agree with reviewer that the current signatures of positive selection is certainly insufficient to explain the underlying adaptive phenotypes in MSEA populations, considering such population-enriched phenotypes (e.g. the curly hair, short stature and dark skin color etc.) might be highly polygenic. It is necessary to conduct large-scale phenotype survey to determine which phenotypes are related to tropical adaptation, and deep functional investigations to uncover the molecular mechanism of evolutionary adaptation. However, bridging the genotype to phenotype linkage using GWAS analysis, or dissecting the functional mechanism of positive selected genes using experimental approach are not the main task of this study. Our study aims to provide a comprehensive variant spectrum of MSEA populations, covering the full-scale variations, including SNV, INDELs, SVs, as well as complex variants using long-read data and high-quality assemblies. At the same time, based on such a novel and unprecedented data resource of MSEA populations, we assessed genomic diversity, genetic structure, population history, medical implications, archaic introgression and footprint of natural selection, adding a great value to the global scientific community and local indigenous community.

In the revised manuscript, we rewrote the part of natural selection to avoid assertive conclusions and statements for the untested genotype to phenotype linkage.

1) We tuned down the inference of tropical adaptation based on natural selection detection. The title of the manuscript is now changed to “3,023 human genomes from mainland Southeast Asia disclose hidden genetic diversity and signatures of **natural selection**”, instead of the previous title of “3,023 human genomes from mainland Southeast Asia disclose hidden genetic diversity and signatures of **tropical adaptation**”.

2) We objectively presented the regions/genes/variants showing significant signatures of positive selection in our dataset, and we made comprehensive functional annotations of these genes according to the available databases (such as ENCODE) and published papers without claiming untested genotype to phenotype linkage.

3) For body height, we performed comprehensive statistical tests with the use of

the published GWAS height-associated variants, we proved a significant excess of variants related to body height among PS-SNVs in MSEA population (P -value = 0.005), and we confirmed the signal of polygenic adaptation for body height (P -value = 0.004), leading to a decreased height in MSEA populations. We also tested three other traits (short stature, darker skin color and curly hair), and we did not see signals of polygenic adaptation.

In addition, as recommended by the reviewer, we revisited our other findings to eliminate bias for genetic explanation of tribal populations. We removed all the comparisons between Tribal and Non-Tribal populations, including IBD measurements, efficient population size evaluation and interpretation of genetic structure.

Finally, we appreciate the reviewer's recommendation of two papers about constructing the framework for functional investigations of the genetic variants identified in indigenous populations. As mentioned in Claw et al. (2018), we agree that indigenous populations remain underrepresented and understudied in genomic, genetic and clinical health research. Hopefully, our SEA3K data closes several (regional) blindspots in the current maps of global genetic variation.

In the future, we plan to collect more populations samples from Southeast Asia, and at the same time, to collect phenotypic data so that genetic association analyses could be applied although we do realize that it is often difficult to conduct a GWAS study in indigenous populations due to the usually small population size. Given this GWAS limitation, we will also apply the approach of reverse genetics to establish the link between genotype and phenotype.

In the revised manuscript, we added a brief discussion on future research directions by referring to the approaches suggested by the two studies (Claw et al. 2018; Fox et al. 2020).

References

- Adams, D., Altucci, L., Antonarakis, S.E., Ballesteros, J., Beck, S., Bird, A., Bock, C., Boehm, B., Campo, E., Caricasole, A., *et al.* (2012). BLUEPRINT to decode the epigenetic signature written in blood. *Nature biotechnology* 30, 224-226.
- All of Us Research Program Genomics, I. (2024). Genomic data in the All of Us Research Program. *Nature* 627, 340-346.
- Audano, P.A., Sulovari, A., Graves-Lindsay, T.A., Cantsilieris, S., Sorensen, M., Welch, A.E., Dougherty, M.L., Nelson, B.J., Shah, A., Dutcher, S.K., *et al.* (2019). Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* 176, 663-675 e619.
- Bergström, A., McCarthy, S.A., Hui, R., Almarri, M.A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., *et al.* (2020). Insights into human genetic variation and population history from 929 diverse genomes. *Science* 367.
- Beyter, D., Ingimundardottir, H., Oddsson, A., Eggertsson, H.P., Bjornsson, E., Jonsson, H., Atlason, B.A., Kristmundsdottir, S., Mehringer, S., Hardarson, M.T., *et al.* (2021). Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nature genetics* 53, 779-786.
- Browning, S.R., Browning, B.L., Zhou, Y., Tucci, S., and Akey, J.M. (2018). Analysis of

- Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. *Cell* 173, 53-61.e59.
- Brucato, N., Andre, M., Hudjashov, G., Mondal, M., Cox, M.P., Leavesley, M., and Ricaut, F.X. (2022). Chronology of natural selection in Oceanian genomes. *iScience* 25, 104583.
- Brucato, N., Andre, M., Tsang, R., Saag, L., Kariwiga, J., Sesuki, K., Beni, T., Pomat, W., Muke, J., Meyer, V., *et al.* (2021). Papua New Guinean Genomes Reveal the Complex Settlement of North Sahul. *Mol Biol Evol* 38, 5107-5121.
- Byrska-Bishop, M., Evani, U.S., Zhao, X., Basile, A.O., Abel, H.J., Regier, A.A., Corvelo, A., Clarke, W.E., Musunuri, R., Nagulapalli, K., *et al.* (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* 185, 3426-3440.e3419.
- Chaisson, M.J.P., Sanders, A.D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E.J., Rodriguez, O.L., Guo, L., Collins, R.L., *et al.* (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature communications* 10, 1784.
- Chan, S.H., Bylstra, Y., Teo, J.X., Kuan, J.L., Bertin, N., Gonzalez-Porta, M., Hebrard, M., Tirado-Magallanes, R., Tan, J.H.J., Jeyakani, J., *et al.* (2022). Analysis of clinically relevant variants from ancestrally diverse Asian genomes. *Nature communications* 13, 6694.
- Chen, C.Y., Chen, T.T., Feng, Y.A., Yu, M., Lin, S.C., Longchamps, R.J., Wang, S.H., Hsu, Y.H., Yang, H.I., Kuo, P.H., *et al.* (2023). Analysis across Taiwan Biobank, Biobank Japan, and UK Biobank identifies hundreds of novel loci for 36 quantitative traits. *Cell Genom* 3, 100436.
- Chen, L., Wolf, A.B., Fu, W., Li, L., and Akey, J.M. (2020). Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals. *Cell* 180, 677-687.e616.
- Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* 18, 170-175.
- Choi, J., Kim, S., Kim, J., Son, H.Y., Yoo, S.K., Kim, C.U., Park, Y.J., Moon, S., Cha, B., Jeon, M.C., *et al.* (2023). A whole-genome reference panel of 14,393 individuals for East Asian populations accelerates discovery of rare functional variants. *Science advances* 9, eadg6319.
- Choin, J., Mendoza-Revilla, J., Arauna, L.R., Cuadros-Espinoza, S., Cassar, O., Larena, M., Ko, A.M., Harmant, C., Laurent, R., Verdu, P., *et al.* (2021). Genomic insights into population history and biological adaptation in Oceania. *Nature* 592, 583-589.
- Choudhury, A., Aron, S., Botigue, L.R., Sengupta, D., Botha, G., Bensellak, T., Wells, G., Kumuthini, J., Shriner, D., Fakim, Y.J., *et al.* (2020). High-depth African genomes inform human migration and health. *Nature* 586, 741-748.
- Cong, P.K., Bai, W.Y., Li, J.C., Yang, M.Y., Khederzadeh, S., Gai, S.R., Li, N., Liu, Y.H., Yu, S.H., Zhao, W.W., *et al.* (2022). Genomic analyses of 10,376 individuals in the Westlake BioBank for Chinese (WBBC) pilot project. *Nat Commun* 13, 2939.
- Deng, L., Pan, Y., Wang, Y., Chen, H., Yuan, K., Chen, S., Lu, D., Lu, Y., Mokhtar, S.S., Rahman, T.A., *et al.* (2022). Genetic Connections and Convergent Evolution of Tropical Indigenous Peoples in Asia. *Molecular biology and evolution* 39.
- Dhir, R.K., Cattaneo, U., Ormaza, M.V.C., Coronado, H., and Oelz, M. (2020). Implementing the ILO Indigenous and Tribal Peoples Convention No. 169: Towards an inclusive, sustainable and just future (International Labour Organization).
- Du, Z., Ma, L., Qu, H., Chen, W., Zhang, B., Lu, X., Zhai, W., Sheng, X., Sun, Y., Li, W., *et al.* (2019). Whole Genome Analyses of Chinese Population and De Novo Assembly of A Northern Han Genome. *Genomics Proteomics Bioinformatics* 17, 229-247.
- Ebert, P., Audano, P.A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M.J., Sulovari, A., Ebler, J., Zhou, W., Serra Mari, R., *et al.* (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* 372.
- Efremov, G.D., Josifovska, O., Nikolov, N., Codrington, J.F., Oner, C., Gonzalez-Redondo, J.M., and Huisman, T.H. (1990). Hb Icaria-Hb H disease: identification of the Hb Icaria mutation through analysis of amplified DNA. *British journal of haematology* 75, 250-253.

- EncodeProjectConsortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57-74.
- Fan, S., Hansen, M.E., Lo, Y., and Tishkoff, S.A. (2016). Going global by adapting local: A review of recent human adaptation. *Science* 354, 54-59.
- Fan, S., Spence, J.P., Feng, Y., Hansen, M.E.B., Terhorst, J., Beltrame, M.H., Ranciaro, A., Hirbo, J., Beggs, W., Thomas, N., *et al.* (2023). Whole-genome sequencing reveals a complex African population demographic history and signatures of local adaptation. *Cell* 186, 923-939 e914.
- Fatumo, S., Yakubu, A., Oyedele, O., Popoola, J., Attipoe, D.A., Eze-Echesi, G., Modibbo, F.Z., Ado-Wanka, N., Salako, O., Nashiru, O., *et al.* (2022). Promoting the genomic revolution in Africa through the Nigerian 100K Genome Project. *Nat Genet* 54, 531-536.
- Gao, Y., Yang, X., Chen, H., Tan, X., Yang, Z., Deng, L., Wang, B., Kong, S., Li, S., Cui, Y., *et al.* (2023). A pangenome reference of 36 Chinese populations. *Nature* 619, 112-121.
- GenomeAsia, K.C. (2019). The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature* 576, 106-111.
- Halldorsson, B.V., Eggertsson, H.P., Moore, K.H.S., Hauswedell, H., Eiriksson, O., Ulfarsson, M.O., Palsson, G., Hardarson, M.T., Oddsson, A., Jensson, B.O., *et al.* (2022). The sequences of 150,119 genomes in the UK Biobank. *Nature* 607, 732-740.
- Hao, M., Pu, W., Li, Y., Wen, S., Sun, C., Ma, Y., Zheng, H., Chen, X., Tan, J., Zhang, G., *et al.* (2021). The HuaBiao project: whole-exome sequencing of 5000 Han Chinese individuals. *J Genet Genomics* 48, 1032-1035.
- Holsinger, K.E., and Weir, B.S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nat Rev Genet* 10, 639-650.
- Ioannidis, A.G., Blanco-Portillo, J., Sandoval, K., Hagelberg, E., Miquel-Poblete, J.F., Moreno-Mayar, J.V., Rodriguez-Rodriguez, J.E., Quinto-Cortes, C.D., Auckland, K., Parks, T., *et al.* (2020). Native American gene flow into Polynesia predating Easter Island settlement. *Nature* 583, 572-577.
- Jablonski, N.G., and Chaplin, G. (2014). The evolution of skin pigmentation and hair texture in people of African ancestry. *Dermatologic clinics* 32, 113-121.
- Jacobs, G.S., Hudjashov, G., Saag, L., Kusuma, P., Darusallam, C.C., Lawson, D.J., Mondal, M., Pagani, L., Ricaut, F.X., Stoneking, M., *et al.* (2019). Multiple Deeply Divergent Denisovan Ancestries in Papuans. *Cell* 177, 1010-1021.e1032.
- Jarvis, J.P., Scheinfeldt, L.B., Soi, S., Lambert, C., Omberg, L., Ferwerda, B., Froment, A., Bodo, J.M., Beggs, W., Hoffman, G., *et al.* (2012). Patterns of ancestry, signatures of natural selection, and genetic association with stature in Western African pygmies. *PLoS Genet* 8, e1002641.
- Jeon, S., Bhak, Y., Choi, Y., Jeon, Y., Kim, S., Jang, J., Jang, J., Blazyte, A., Kim, C., Kim, Y., *et al.* (2020). Korean Genome Project: 1094 Korean personal genomes with clinical information. *Sci Adv* 6, eaaz7835.
- Kalinowski, S.T. (2005). Do polymorphic loci require large sample sizes to estimate genetic distances? *Heredity* 94, 33-36.
- Kanavakis, E., Traeger-Synodinos, J., Papasotiriou, I., Vrettou, C., Metaxotou-Mavromati, A., Stamoulakatou, A., Lagona, E., and Kattamis, C. (1996). The interaction of alpha zero thalassaemia with Hb Icaria: three unusual cases of haemoglobinopathy H. *British journal of haematology* 92, 332-335.
- Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., *et al.* (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434-443.
- Kawai, Y., Watanabe, Y., Omae, Y., Miyahara, R., Khor, S.S., Noiri, E., Kitajima, K., Shimanuki, H., Gatanaga, H., Hata, K., *et al.* (2023). Exploring the genetic diversity of the Japanese population: Insights from a large-scale whole genome sequencing analysis. *PLoS Genet* 19, e1010625.
- Kimura, E.M., Oliveira, D.M., Fertrin, K., Pinheiro, V.R., Jorge, S.E., Costa, F.F., and de Fatima Sonati, M. (2009). Hb H disease resulting from the association of an alpha-thalassaemia allele [-(alpha)] with an unstable alpha-globin variant [Hb Icaria]: First report

- on the occurrence in Brazil. *Genetics and molecular biology* 32, 712-715.
- Kirsche, M., Prabhu, G., Sherman, R., Ni, B., Battle, A., Aganezov, S., and Schatz, M.C. (2023). Jasmine and Iris: population-scale structural variant comparison and analysis. *Nat Methods* 20, 408-417.
- Kutanan, W., Liu, D., Kampuansai, J., Srikummool, M., Srithawong, S., Shoocongdej, R., Sangkhano, S., Ruangchai, S., Pittayaporn, P., Arias, L., *et al.* (2021). Reconstructing the Human Genetic History of Mainland Southeast Asia: Insights from Genome-Wide Data from Thailand and Laos. *Molecular biology and evolution* 38, 3459-3477.
- Lachance, J., Vernot, B., Elbers, C.C., Ferwerda, B., Froment, A., Bodo, J.M., Lema, G., Fu, W., Nyambo, T.B., Rebbeck, T.R., *et al.* (2012). Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell* 150, 457-469.
- Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., *et al.* (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467, 832-838.
- Lasisi, T., Smallcombe, J.W., Kenney, W.L., Shriver, M.D., Zydney, B., Jablonski, N.G., and Havenith, G. (2023). Human scalp hair as a thermoregulatory adaptation. *Proc Natl Acad Sci U S A* 120, e2301760120.
- Lawson, D.J., van Dorp, L., and Falush, D. (2018). A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature communications* 9, 3258.
- Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009). Genotype imputation. *Annual review of genomics and human genetics* 10, 387-406.
- Liao, W.W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J.K., Monlong, J., Abel, H.J., *et al.* (2023). A draft human pangenome reference. *Nature* 617, 312-324.
- Lin, J., Wang, S., Audano, P.A., Meng, D., Flores, J.I., Kusters, W., Yang, X., Jia, P., Marschall, T., Beck, C.R., *et al.* (2022). SVision: a deep learning approach to resolve complex structural variants. *Nat Methods* 19, 1230-1233.
- Lipson, M., Cheronet, O., Mallick, S., Rohland, N., Oxenham, M., Pietrusewsky, M., Pryce, T.O., Willis, A., Matsumura, H., Buckley, H., *et al.* (2018). Ancient genomes document multiple waves of migration in Southeast Asian prehistory. *Science* 361, 92-95.
- Liu, D., Peter, B.M., Schiefenhovel, W., Kayser, M., and Stoneking, M. (2022). Assessing Human Genome-wide Variation in the Massim Region of Papua New Guinea and Implications for the Kula Trading Tradition. *Mol Biol Evol* 39.
- Liu, X., Koyama, S., Tomizuka, K., Takata, S., Ishikawa, Y., Ito, S., Kosugi, S., Suzuki, K., Hikino, K., Kido, M., *et al.* (2024). Decoding triancestral origins, archaic introgression, and natural selection in the Japanese population by whole-genome sequencing. *Science advances* 10, eadi8419.
- Lo, Y.H., Cheng, H.C., Hsiung, C.N., Yang, S.L., Wang, H.Y., Peng, C.W., Chen, C.Y., Lin, K.P., Kang, M.L., Chen, C.H., *et al.* (2021). Detecting Genetic Ancestry and Adaptation in the Taiwanese Han People. *Molecular biology and evolution* 38, 4149-4165.
- Luo, H., Zhang, P., Zhang, W., Zheng, Y., Hao, D., Shi, Y., Niu, Y., Song, T., Li, Y., Zhao, S., *et al.* (2023). Recent positive selection signatures reveal phenotypic evolution in the Han Chinese population. *Science bulletin* 68, 2391-2404.
- Malaspinas, A.S., Westaway, M.C., Muller, C., Sousa, V.C., Lao, O., Alves, I., Bergstrom, A., Athanasiadis, G., Cheng, J.Y., Crawford, J.E., *et al.* (2016). A genomic history of Aboriginal Australia. *Nature* 538, 207-214.
- Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., *et al.* (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* 538, 201-206.
- Marchini, J., and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature reviews Genetics* 11, 499-511.
- McColl, H., Racimo, F., Vinner, L., Demeter, F., Gakuhari, T., Moreno-Mayar, J.V., van Driem, G., Gram Wilken, U., Seguin-Orlando, A., de la Fuente Castro, C., *et al.* (2018). The

- prehistoric peopling of Southeast Asia. *Science* 361, 88-92.
- McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome biology* 17, 122.
- Mondal, M., Casals, F., Xu, T., Dall'Olio, G.M., Pybus, M., Netea, M.G., Comas, D., Laayouni, H., Li, Q., Majumder, P.P., *et al.* (2016). Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation. *Nature genetics* 48, 1066-1070.
- Morin, P.A., Martien, K.K., and Taylor, B.L. (2009). Assessing statistical power of SNPs for population structure and conservation studies. *Molecular ecology resources* 9, 66-73.
- Nakatsuka, N., Moorjani, P., Rai, N., Sarkar, B., Tandon, A., Patterson, N., Bhavani, G.S., Girisha, K.M., Mustak, M.S., Srinivasan, S., *et al.* (2017). The promise of discovering population-specific disease-associated genes in South Asia. *Nat Genet* 49, 1403-1407.
- Naslavsky, M.S., Scliar, M.O., Yamamoto, G.L., Wang, J.Y.T., Zverinova, S., Karp, T., Nunes, K., Ceroni, J.R.M., de Carvalho, D.L., Da, S.S.C.E., *et al.* (2022). Whole-genome sequencing of 1,171 elderly admixed individuals from Sao Paulo, Brazil. *Nat Commun* 13, 1004.
- Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., *et al.* (2022). The complete sequence of a human genome. *Science* 376, 44-53.
- Okada, Y., Momozawa, Y., Sakaue, S., Kanai, M., Ishigaki, K., Akiyama, M., Kishikawa, T., Arai, Y., Sasaki, T., Kosaki, K., *et al.* (2018). Deep whole-genome sequencing reveals recent selection signatures linked to evolution and disease risk of Japanese. *Nat Commun* 9, 1631.
- Perry, G.H., and Dominy, N.J. (2009). Evolution of the human pygmy phenotype. *Trends in ecology & evolution* 24, 218-225.
- Perry, G.H., Foll, M., Grenier, J.C., Patin, E., Nedelec, Y., Pacis, A., Barakatt, M., Gravel, S., Zhou, X., Nsoby, S.L., *et al.* (2014). Adaptive, convergent origins of the pygmy phenotype in African rainforest hunter-gatherers. *Proc Natl Acad Sci U S A* 111, E3596-3603.
- Razali, R.M., Rodriguez-Flores, J., Ghorbani, M., Naeem, H., Aamer, W., Aliyev, E., Jubran, A., Clark, A.G., Fakhro, K.A., and Mokrab, Y. (2021). Thousands of Qatari genomes inform human migration history and improve imputation of Arab haplotypes. *Nat Commun* 12, 5929.
- Rees, J.S., Castellano, S., and Andres, A.M. (2020). The Genomics of Human Local Adaptation. *Trends in genetics : TIG* 36, 415-428.
- Reis, A.L.M., Rapadas, M., Hammond, J.M., Gamaarachchi, H., Stevanovski, I., Ayuputeri, K.M., Chintalaphani, S.R., Dissanayake, D.S.B., Siggs, O.M., Hewitt, A.W., *et al.* (2023). The landscape of genomic structural variation in Indigenous Australians. *Nature* 624, 602-610.
- Romanoski, C.E., Glass, C.K., Stunnenberg, H.G., Wilson, L., and Almouzni, G. (2015). Epigenomics: Roadmap for regulation. *Nature* 518, 314-316.
- Scheinfeldt, L.B., Soi, S., Lambert, C., Ko, W.Y., Coulibaly, A., Ranciaro, A., Thompson, S., Hirbo, J., Beggs, W., Ibrahim, M., *et al.* (2019). Genomic evidence for shared common ancestry of East African hunting-gathering populations and insights into local adaptation. *Proc Natl Acad Sci U S A* 116, 4166-4175.
- Shah, N., Hou, Y.C., Yu, H.C., Sainger, R., Caskey, C.T., Venter, J.C., and Telenti, A. (2018). Identification of Misclassified ClinVar Variants via Disease Population Prevalence. *American journal of human genetics* 102, 609-619.
- Silcocks, M., Farlow, A., Hermes, A., Tsambos, G., Patel, H.R., Huebner, S., Baynam, G., Jenkins, M.R., Vukcevic, D., Easta, S., *et al.* (2023). Indigenous Australian genomes show deep structure and rich novel variation. *Nature* 624, 593-601.
- Smolka, M., Paulin, L.F., Grochowski, C.M., Horner, D.W., Mahmoud, M., Behera, S., Kalef-Ezra, E., Gandhi, M., Hong, K., Pehlivan, D., *et al.* (2024). Detection of mosaic and population-level structural variants with Sniffles2. *Nature biotechnology*.
- Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G.,

- Corvelo, A., Gogarten, S.M., Kang, H.M., *et al.* (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* 590, 290-299.
- Tan, V.Y., and Timpson, N.J. (2022). The UK Biobank: A Shining Example of Genome-Wide Association Study Science with the Power to Detect the Murky Complications of Real-World Epidemiology. *Annu Rev Genomics Hum Genet* 23, 569-589.
- Tucci, S., Vohr, S.H., McCoy, R.C., Vernot, B., Robinson, M.R., Barbieri, C., Nelson, B.J., Fu, W., Purnomo, G.A., Sudoyo, H., *et al.* (2018). Evolutionary history and adaptation of a human pygmy population of Flores Island, Indonesia. *Science* 361, 511-516.
- Verrelli, B.C., McDonald, J.H., Argyropoulos, G., Destro-Bisol, G., Froment, A., Drousiotou, A., Lefranc, G., Helal, A.N., Loiselet, J., and Tishkoff, S.A. (2002). Evidence for balancing selection from nucleotide sequence analyses of human G6PD. *Am J Hum Genet* 71, 1112-1128.
- Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R., Xu, C., Futema, M., Lawson, D., *et al.* (2015). The UK10K project identifies rare variants in health and disease. *Nature* 526, 82-90.
- Wang, S., Lin, J., Jia, P., Xu, T., Li, X., Liu, Y., Xu, D., Bush, S.J., Meng, D., and Ye, K. (2024). De novo and somatic structural variant discovery with SVision-pro. *Nature biotechnology*.
- Wang, T., Antonacci-Fulton, L., Howe, K., Lawson, H.A., Lucas, J.K., Phillippy, A.M., Popejoy, A.B., Asri, M., Carson, C., Chaisson, M.J.P., *et al.* (2022). The Human Pangenome Project: a global resource to map genomic diversity. *Nature* 604, 437-446.
- Willing, E.M., Dreyer, C., and van Oosterhout, C. (2012). Estimates of genetic differentiation measured by F(ST) do not necessarily require large sample sizes when using many SNP markers. *PloS one* 7, e42649.
- Wong, E., Bertin, N., Hebrard, M., Tirado-Magallanes, R., Bellis, C., Lim, W.K., Chua, C.Y., Tong, P.M.L., Chua, R., Mak, K., *et al.* (2023). The Singapore National Precision Medicine Strategy. *Nat Genet* 55, 178-186.
- Wong, L.P., Ong, R.T., Poh, W.T., Liu, X., Chen, P., Li, R., Lam, K.K., Pillai, N.E., Sim, K.S., Xu, H., *et al.* (2013). Deep whole-genome sequencing of 100 southeast Asian Malays. *Am J Hum Genet* 92, 52-66.
- Wu, D., Dou, J., Chai, X., Bellis, C., Wilm, A., Shih, C.C., Soon, W.W.J., Bertin, N., Lin, C.B., Khor, C.C., *et al.* (2019). Large-Scale Whole-Genome Sequencing of Three Diverse Asian Populations in Singapore. *Cell* 179, 736-749 e715.
- Yengo, L., Vedantam, S., Marouli, E., Sidorenko, J., Bartell, E., Sakaue, S., Graff, M., Eliassen, A.U., Jiang, Y., Raghavan, S., *et al.* (2022). A saturated map of common genetic variants associated with human height. *Nature* 610, 704-712.
- Zerbino, D.R., Wilder, S.P., Johnson, N., Juettemann, T., and Flicek, P.R. (2015). The ensembl regulatory build. *Genome Biol* 16, 56.
- Zhang, P., Luo, H., Li, Y., Wang, Y., Wang, J., Zheng, Y., Niu, Y., Shi, Y., Zhou, H., Song, T., *et al.* (2021). NyuWa Genome resource: A deep whole-genome sequencing-based variation profile and reference panel for the Chinese population. *Cell Rep* 37, 110017.
- Zhang, X., Liu, Q., Zhang, H., Zhao, S., Huang, J., Sovannary, T., Bunnath, L., Aun, H.S., Samnom, H., Su, B., *et al.* (2022). The distinct morphological phenotypes of Southeast Asian aborigines are shaped by novel mechanisms for adaptation to tropical rainforests. *Natl Sci Rev* 9, nwab072.
- Zheng, W., He, Y., Guo, Y., Yue, T., Zhang, H., Li, J., Zhou, B., Zeng, X., Li, L., Wang, B., *et al.* (2023). Large-scale genome sequencing redefines the genetic footprints of high-altitude adaptation in Tibetans. *Genome Biol* 24, 73.
- Ziyatdinov, A., Torres, J., Alegre-Diaz, J., Backman, J., Mbatchou, J., Turner, M., Gaynor, S.M., Joseph, T., Zou, Y., Liu, D., *et al.* (2023). Genotyping, sequencing and analysis of 140,000 adults from Mexico City. *Nature* 622, 784-793.

Referee #1

The authors have made an effort to address the issues I raised and partially resolved them. Of note, they have now withdrawn the unsupported conclusions such as medical implications of the rare variants identified, high quality reference genomes, variants that contribute to "adaptations" such as short stature and other physical features. They have also produced 74 haplotype assemblies from the 37 long-read sequencing datasets. However, the authors still make broad conclusions/speculations based on a very small dataset (short-read sequencing data from ~100 individuals from each subpopulation and genome assemblies from 1 individual from 37 subpopulations) without further studies. While I understand the desire to make the dataset seem more important than it currently stands, unsubstantiated claims will likely lead to confusion and diminish the credibility of the project. It is best to remove the speculative part of the discussion unless additional studies are done to validate them.

Authors response: We understand the reviewer's concerns that some conclusions or speculations in our study might be supported due to the limited sample size of each population or lack of functional validations.

For the sample size issue, we have responded in the last version of "response to reviewer's comments". We evaluated the potential influence of sample size on two analyses (natural selection test and introgression detection) that involved calculation of allele frequency (sample size sensitive). As proposed in previous studies (Browning, et al. 2018; Chen, et al. 2020), a sample size of $n > 20$ is sufficient to yield reliable estimates. In our study, to minimize the influence of sample size, we only included those MSEA populations with a sample size of $n > 20$ in the analyses of selection and introgression.

For the functional validation issue, according to the reviewer's suggestion, we have removed the speculative parts in the discussion section, and we also shortened the discussion on adaptation and medical implications (**Page 29, Line 730-731; Line 739-744**).

We do realize that with 3,023 individuals, we report the largest study of the MSEA populations, but the representation of MSEA population remains unideal considering the rich diversity of ethnic populations in this region. In the discussion, we added two sentences to state the limitation of this study, and call for further functional studies and much larger population studies, as well as genetic associations studies on phenotype and disease, to uncover the mechanism of tropical adaptation and disease prevalence in MSEA populations (**Page 29, Line 730-731**).

Referee #2

He et al. present a substantially revised manuscript of their comprehensive study of 3,023 samples from mainland Southeast Asia. Overall, I find that the manuscript has much improved in focus, in detailing the analysis steps and a more careful interpretation of the results. The newly added text and figures add a lot of clarity but not in all remaining aspects. I have a few more questions and remarks stated below and most of which, I believe, can be answered with little effort.

Authors response: We appreciate the reviewer's approval of our study and the constructive comments, we have addressed these points below and updated the related parts in the revised manuscripts accordingly.

Major:

P7: the added detail about your discovered short variants helps a lot to put the numbers into perspective. However, I think I didn't see a section in the Methods that describes how you assess novelty; for SNVs, there is no room for interpretation (it's just an exact match or it isn't), but for the short indels, did you just overlap and require identity to decide on known vs novel? This is relevant information because your number of putatively novel (short) indels seems somewhat high when compared to, e.g., the 1000 Genomes cohort.

Authors response: For novelty assessment of short variants (SNVs + short indels), we adopted a commonly used tool SnpSift (<https://github.com/pcingola/SnpSift>) (Cingolani, et al. 2012) to annotate the SEA3K variants based on a known short variant set, which combined HGSVC3, HPRC, high-coverage 1KGP Phase 3, HGDP, gnomAD, GAsP and SG10K. The command was:

"SnpSift.jar annotate known.vcf variants.vcf > variants_annotated.vcf"

We rechecked the indels annotated by SnpSift, and found that only exact match (100% reciprocal overlap: same length and same sequence) indels were annotated as known indels. We agree with reviewer that it is too conservative which may lead to an overestimation of novel indels. To resolve this issue, we took a loose threshold (50% reciprocal overlap) to annotate the short indels. Under this criterion, the number of novel indels in the SEA3K dataset dropped to 3.20 million, accounting for 22.4% of all identified indels (**Figure R1**). We also tried a more relaxed threshold (overlap of 1 bp), and the novel indels dropped to 2.25 million, accounting for 15.8% of all identified indels.

In the revised manuscript, to be objective, we chose the 50% reciprocal overlap as the threshold in novelty assessment of indels, which is also the threshold commonly

adopted in SV analyses. We updated the results accordingly and added the description of novelty assessments for short variants (SNVs and indels) in the method section (**Page 33, Line 850-857**).

“Novelty assessment: we used SnpSift to annotate SNVs in SEA3K by comparing with the union of the public variant repositories, including dbSNP (v156), HGSC (Phase 3), HPRC (v1.1), high-coverage 1KGP, HGDP, gnomAD (v3.1.2), GAsP, and SG10K (Pilot version). For the indels, we used the BEDtools (v2.31.0) with command bedtools intersect -f 0.5 -r to determine the novel indels by comparing the SEA3K indels with the public variant repositories. An effective overlap was defined with an overlapping length of indels reaching at least 50% of the reciprocal similarities.”

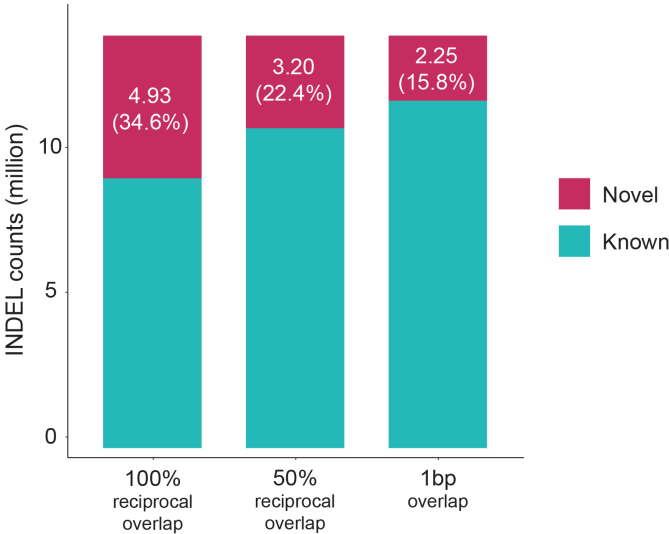


Figure R1. Comparison of the annotation results of short indels under different criteria. The number and proportion of novel indels in each criterion are indicated.

P10: I cannot entirely follow the section of the 37 assembled genomes. You state in the text that you produced phased assemblies, but the hifiasm call in the Methods lacks the appropriate input to produce a (truly) phased assembly. Given that you have family trios in your dataset, I would have expected to see the respective parental k-mer databases as additional input (i.e., using the so-called trio-binning mode). According to the hifiasm readme, your assemblies are only partially phased? Please clarify. And if I am not mistaken, you make the more precise statement about the assembly status in your answer to my previous question (13).

Authors response: Considering that we only generated the long-read HiFi data (without trio-binning data or Hi-C data), we only produced partially phased assemblies for the 37 assembled genomes.

To be clear, according to the reviewer’s suggestion, we rephrased the statement that we generated 74 partially phased haploid assemblies for the 37 individuals in the revised

manuscript (Page 10, Line 259; Page 13, Line 306).

P10: “complex SVs” - what is the (working) definition of a complex SV in your analysis?

Authors response: We adopted SVision-pro (Wang, et al. 2024) to detect complex SVs using our long-read data of 37 MSEA individuals. The complex SVs in this study are defined by the two papers of Prof. Ye’s group, who developed SVision (Lin, et al. 2022) and SVision-pro (Wang, et al. 2024) for complex SV detection based on long-read data or assemblies. Complex SVs refer to the SVs containing multiple breakpoints and often missed or misinterpreted by the detection algorithms of simple SVs (Lin, et al. 2022). Typically, complex SVs comprise several internal simple SV subcomponents (Wang, et al. 2024). We added a sentence to explicitly define the complex SVs in the revised manuscript (Page 11, Line 266-270).

“In this study, simple SVs refer to the SVs involve two breakpoints and one junction, including deletions, insertion, duplications and inversions. Complex SVs refer to the SVs containing multiple breakpoints that cluster together, and often missed or misinterpreted by the detection algorithms of simple SVs.”

We do realize that there have been other forms of complex SVs defined and reported in previous studies. For example, the structurally divergent regions (SDRs) (Mao, et al. 2024) and DNA rearrangement involving multiple breakpoints (Zhou, et al. 2024), which had been poorly characterized or missed by previous genome analyses due to detection challenges. Therefore, in the revised manuscript, we added a sentence in the discussion section to call for further study to explore other forms of complex SVs (Page 30, Line 771-773).

“... and other forms of complex SVs (Mao, et al. 2024; Zhou, et al. 2021) should be studied by conducting telomere-to-telomere (T2T) genome sequencing and establishing a high-quality pangenome reference of MSEA populations.”

P46: in your explanation about how you derived the set of known SVs using HGSVC and HPRC references, what exactly is the operation behind the word ‘and’ in this statement: “... the combined SVs from HGSVC3 and HPRC was regard as the known SV set.” For me, “and” implies an intersection, which I think is consistent with your use of bedtools stated at the beginning of the sentence. If you used the intersection instead of the union between HPRC and HGSVC to define a known set, you have created a very conservative set of known variants because both the HPRC and the HGSVC certainly performed internal QC before releasing their callsets. How many novel MSEA SVs remain if you consider the union of the two callsets?

Authors response: We agree with the reviewer's point. In fact, we used union of the two callsets (HPRC + HGSVC3) as the known callset to assess the novelty of MSEA SVs as the reviewer suggested. To be clear, we rephased the sentence in the revised manuscript (**Page 11, Line 287**).

"...We overlapped our SV set with the union of the reported SV sets from HGSVC3 and HPRC".

Minor:

p5/6 and supp. Fig. 1: I think, overall, the numbers seem plausible and provide a great summary of the current status quo emphasizing the importance of your contribution. However, could be that readers may be confused due to the view limited on WGS only - I think, though, that this is absolutely the right perspective given the focus of your study - which then excludes samples that were only characterized with, e.g., WES or proteomics data; and many of the popular projects from your list (HPRC, SGDP, 1000 Genomes etc.) only have a small sample size (relative to the total ~1.4M that you counted) with quite a focus on high-diversity samples (essentially, the AFR continental group). I also randomly checked some of the references 1-41 (as per supp. Fig. 1), and I found some examples that do not make sense to me, e.g., the number 32 (Tan/Timpson, Annu Rev Genomics Hum Genet 2022) is apparently a review paper about the UK Biobank and I don't see how this adds to your sample count; number 3 (Brucato, iScience 2022) lists data sources in their Table S1, all of which seem to refer to other studies, which thus means this paper also does not add anything (or that samples were accidentally counted twice). I leave it up to you how to deal with this, but you may want to make your counting more transparent by adding a simple supplementary table stating how many samples characterized with WGS you added to the total from each publication stratified by continental group (what you show in the left pie chart in suppl. Fig. 1).

Authors response: According to the reviewer's suggestion, we rechecked the counts and references of our previous survey about whole-genome sequencing (WGS) data of global populations from the published studies. We found some improper datasets or reference papers included in our previous list by mistake. Therefore, in the revised manuscript, we updated the list and the plot (**Figure R2**, updated as **Supplementary Figure 1**), and following the reviewer's suggestion, we added a supplementary table (**Table R1** and **Table R2**, added as **Supplementary Table 1**) to show the sample size of different super-population or populations from each reference.

Table R1. Summary list of global populations with published WGS data. The sample counts stratified by super-population or population from each publication are presented. (The full information of this table can be seen in **Supplementary Table 1** of the revised manuscript). **Figure R2 (Supplementary Figure 1)** is plotted based on this table.

ID	Total	AFR (African)	AMR (American)	EUR (European)	OCEA (Oceanian)	ASIA (Asian)	Asian groups				Projects	Publications
							SAS (South Asian)	EAS (East Asian)	SEA (Southeast Asian)	WAS/CAS (West/Central Asian)		
	674,876	99,959	105,753	388,762	2,507	77,095	5,984	54,487	10,588	6,836	-	-
[1]	-	893	490	633	28	1,186	601	370	215	0	1KGP	2022 (Cell) High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios
[2]	-	104	61	155	0	581	0	205	18	358	HGDP	2019 (Science)-Insights into human genetic variation and population history from 929 diverse genomes
[3]	-	49	28	0	25	198	49	35	12	102	SGDP	2016 (Nature)-The Simons Genome Diversity Project: 300 genomes from 142 diverse populations
[4]	-	0	28,391	41,081	0	5,181	2,419	2,604	0	158	gnomAD	2024 (Nature)-A genomic mutational constraint map using variation in 76,156 human genomes
[5]	-	47,020	23,640	63,420	962	13,860	0	13,860	0	0	TOPMed	2021 (Nature)-Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program
[6]	-	0	0	150,119	0	0	0	0	0	0	UK Biobank	2022 (Nature)-The sequences of 150,119 genomes in the UK biobank
[7]	-	51,365	41,938	125,843	237	7,422	2,121	5,301	0	0	AFRiUs	2024 (Nature)-Genomic data in the All of Us Research Program
[8]	-	102	26	114	74	1,385	724	351	310	0	GA100K	2019 (Nature)-The GenomeAsia 100K Project enables genetic discoveries across Asia
[9]	-	0	0	0	0	9,770	0	0	9,770	0	SG10K	2019 (Cell)-Large Scale Whole-Genome Sequencing of Three Diverse Asian Populations in Singapore
[10]	-	0	0	0	0	9,770	0	0	9,770	0	SG10K	2022 (Nature Communications)-Analysis of clinically relevant variants from ancestrally diverse Asian genomes
[11]	-	0	0	0	0	9,770	0	0	9,770	0	SG10K	2023 (Nature Genetics)-The Singapore National Precision Medicine Strategy
[12]	-	0	0	0	0	2,999	0	2,999	0	0	NyuWa	2021 (Cell Report)-NyuWa Genome resource: A deep whole-genome-sequencing-based variation profile and referencepan
[13]	-	0	0	0	0	4,535	0	4,535	0	0	WBBC	2021 (Nature Communications)-Genomic analyses of 10,376 individuals in theWestlake BioBank for Chinese (WBBC) pilot
[14]	-	0	0	0	0	2,000	0	2,000	0	0	TWBC	2022 (Cell Genomics)-Taiwan Biobank: A rich biomedical research database of the Taiwanese population
[15]	-	0	0	0	0	10,000	0	10,000	0	0	CKB	2022 (NAR)-A high-resolution haplotype-resolved reference panel constructed from the China Kadoorie Biobank Study
[16]	-	0	0	3,775	0	0	0	0	0	0	FinnGen	2023 (Nature)-FinnGen provides genetic insights from a well-phenotyped isolated population
[17]	-	426	0	0	0	0	0	0	0	0	H3Africa	2020 (Nature)-High-depth African genomes inform human migration and health
[18]	-	0	0	3,622	0	0	0	0	0	0	-	2021 (Nature Genetics)-Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in hu
[19]	-	0	0	0	108	0	0	0	0	0	-	2016 (Nature)-A genomic history of Aboriginal Australia
[20]	-	0	0	0	317	0	0	0	0	0	-	2021 (Nature)-Genomic insights into population history and biological adaptation in Oceania
[21]	-	0	0	0	159	0	0	0	0	0	-	2023 (Nature)-Indigenous Australian genomes show deep structure and rich novel variation
[22]	-	0	0	0	139	0	0	0	0	0	-	2023 (Nature)-The landscape of genomic structural variation in Indigenous Australians
[23]	-	0	0	0	161	0	0	0	0	0	-	2019 (Cell)-Multiple Deeply Divergent Denisovan Ancestries in Papuans
[24]	-	0	0	0	58	0	0	0	0	0	-	2021 (MBE)-Papua New Guinean Genomes Reveal the Complex Settlement
[25]	-	0	0	0	239	0	0	0	0	0	-	2022 (iScience)-Chronology of natural selection in Oceanian genomes
[26]	-	0	0	0	0	70	70	0	0	0	-	2016 (Nature Genetics)-Genomic analysis of Andamanese provides insights into ancient human migration into Asia and ac
[27]	-	0	0	0	0	32	0	0	32	0	-	2018 (Science)-Evolutionary history and adaptation of a human pygmy population of Flores Island
[28]	-	0	0	0	0	100	0	0	100	0	-	2013 (AJHG)-Deep whole-genome sequencing of 100 southeast Asian Malays
[29]	-	0	0	0	0	50	0	0	50	0	-	2022 (MBE)-Genetic Connections and Convergent Evolution of Tropical Indigenous Peoples in Asia
[30]	-	0	0	0	0	1,094	0	1,094	0	0	-	2020 (Science Advances)-Korean Genome Project: 1094 Korean personal genomes with clinical information
[31]	-	0	0	0	0	6,218	0	0	0	6,218	-	2021 (Nature Communications)-Thousands of Qatari genomes inform human migration history and improve imputation of /
[32]	-	0	10,008	0	0	0	0	0	0	0	-	2023 (Nature)-Genotyping, sequencing and analysis of 140,000 adults from Mexico City
[33]	-	0	1,171	0	0	0	0	0	0	0	-	2022 (Nature Communications)-Whole-genome sequencing of 1,171 elderly admixed individuals from Brazil
[34]	-	0	0	0	0	597	0	597	0	0	-	2019 (Genomics Proteomics Bioinformatics)-Whole Genome Analyses of Chinese Population and De Novo Assembly of A
[35]	-	0	0	0	0	248	0	248	0	0	-	2022 (Cell Report)-The Tibetan-Yi region is both a corridor and a barrier for human gene flow
[36]	-	0	0	0	0	9,287	0	9,287	0	0	-	2023 (Plos Genetics)-Exploring the genetic diversity of the Japanese Population: Insights from a Large Scale Whole Geno
[37]	-	0	0	0	0	81	0	0	81	0	-	2022 (National Science Review)-The distinct morphological phenotypes of Southeast Asian aborigines are shaped by nov
[38]	-	0	0	0	0	1,001	0	1,001	0	0	-	2023 (Genome Biology)-Large-scale genome sequencing redefines the genetic footprints of high-altitude adaptation in Tib

Table R2. Summary list of the published WGS data of Southeast Asian populations
(The full information of this table can be seen in **Supplementary Table 1** of the revised manuscript).

Projects	Publish year	References	Populations	Counts	Groups	Counts
1KGP	2022-(Cell)-[1]	[1]	KHV/CDX	215	Non-tribal	10,425
HGDP	2019-(Science)-[2]	[2]	Cambodian/Dai	18	Non-tribal	
SGDP	2016-(Nature)-[3]	[3]	Cambodian/Dai/Burmese/Kinh/Thai	12	Non-tribal	
GA100K	2019-(Nature)-[8]	[8]	Myanmar/Cambodian/Vietnam/Philippines/Malays/Brunei	310	Non-tribal	
SG10K	2019-(Cell)-L[9]	[9]	Singaporean	9,770	Non-tribal	
	2022-(Nature)-[10]	[10]				
	2023-(Nature)-[11]	[11]				
-	2013-(AJHG)-[27]	[27]	Indigenous Malays	100	Non-tribal	163
-	2018-(Science)-[28]	[28]	Flores pygmy	32	Tribal	
-	2022-(MBE)-[29]	[29]	Orang Asli	50	Tribal	
-	2022-(Nation)-[37]	[37]	Indigenous Cambodians	81	Tribal	
-	-	-	Total	10,588	-	10,588

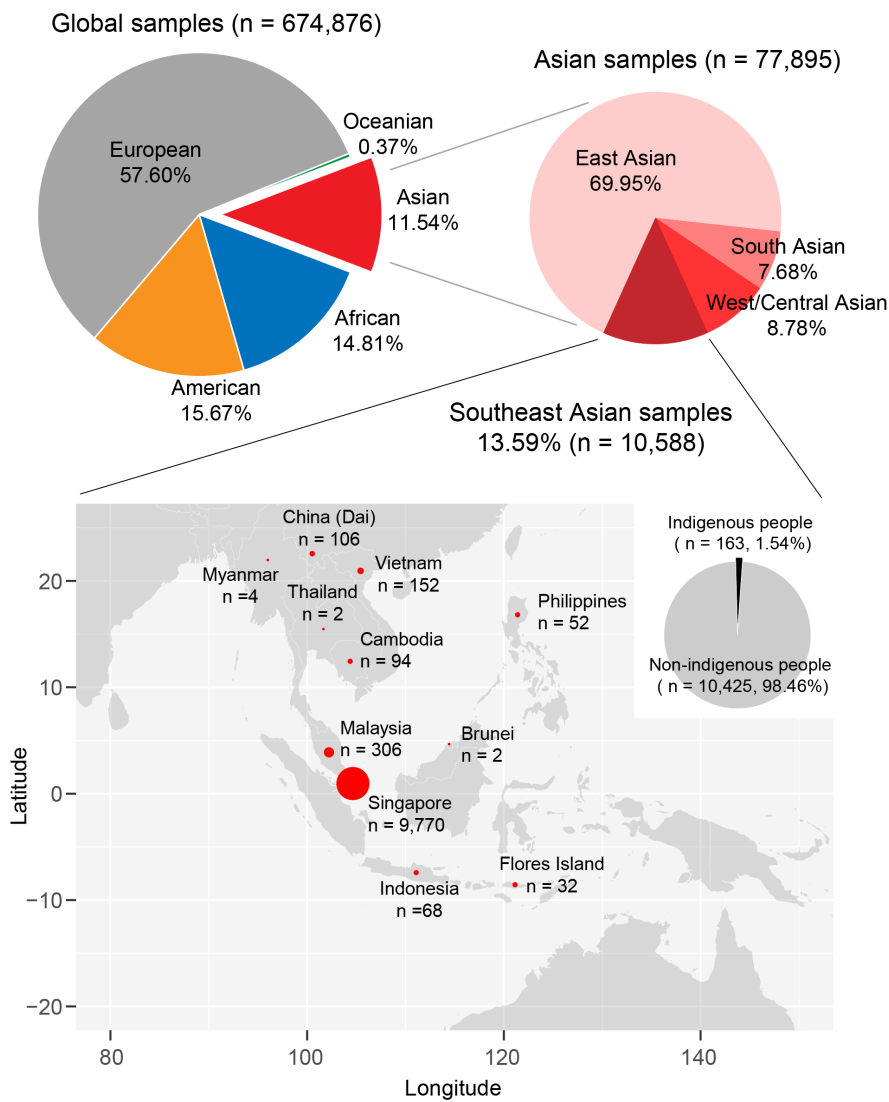


Figure R2. Summary of global population with published WGS data. The data were retrieved from the published literatures listed in **Supplementary Table 1**.

P7: You are referencing items 43 (Liao et al., Nature; the HPRC flagship paper) and 44 (an associated perspectives article that does not represent any source) for the HGSVC phase 3 (= identical citations for HPRC). Obviously, that is the wrong attribution and it seems the HGSVC phase 3 has at least a preprint available that you should cite here: <https://biorxiv.org/cgi/content/short/2024.09.24.614721v1>
This is related to your answer to my previous question (9).

Authors response: We thank the reviewer for pointing out the citation error, and we corrected it accordingly in the revised manuscript (**Page 9, Line 176**).

P11: given that reference-style SV callsets produced, e.g., by the HPRC and the HGSVC, are still work in progress, it would add perspective to also report the inverse statistic, i.e. how many SVs in those catalogs have no equivalent in your MSEA callset?

Authors response: There are 87,477 SVs in the known set (union of HPRC and HGSVC) not detected in our MSEA SV set, which can be seen in **Figure 2i**. To be clear, in the revised manuscript, we added a sentence stating this inverse statistic in the main text (**Page 12, Line 292-293**).

“Conversely, 87,477 SVs in HPRC and HGSVC3 were not detected in our MSEA SV set (Fig. 2i)”

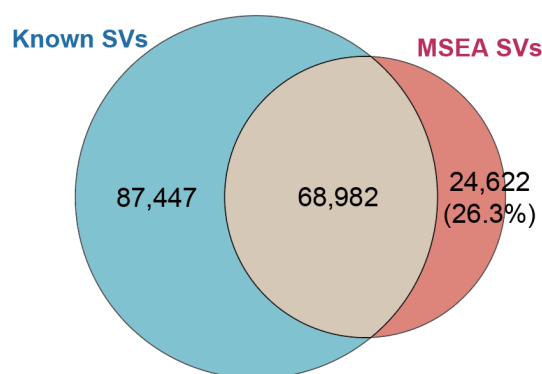


Figure R3. Intersection of the SV sets (insertions and deletions) between the known SVs (union of HGSVC3 and HPRC) and the MSEA SV set. Among the known SVs, 87,447 were not detected in the MSEA set. Among the MSEA set, 24,622 SVs (accounting for 26.3%) were not reported in the known set, namely the newly identified SVs in this study.

P28: your answer to my previous question (8): I find your response somewhat beyond the point because global reference panels that are being expanded over time by adding

more and more diverse populations will simply subsume population-specific reference panels. That does not deny that population-specific panels outperform global panels before that point is reached, but my point is that the collective effort of the scientific community should be targeted at creating globally diverse panels. However, since you are now offering the imputation server, your panel is at least accessible in that way, and I am fine with that.

Authors response: We agree with the reviewer that the goal for genotype imputation is creating a globally diverse panel that can cover all regional populations. As also suggested by Reviewer 3, we tried to generate a MEGA reference panel by combining multiple available panels, including our SEA3K panel, the SG10K panel and the 1KGP panel, and we observed a higher imputation power than using a single panel (**Figure R5**). We have added this polit MEGA panel in the SEA3K imputation server.

Considering that building a MEGA imputation panel of global populations is not the mission of this study, and at the same time, and it takes tremendous efforts to collect all public WGS data. We can share our SEA3K panel with the scientific groups who want to achieve a global MEGA panel in the future.

P44 Methods: the yak call to build the k-mer databases lists the same input twice (“\$sample.clean.fq.gz”). I assume that is a typo, but you may want to double-check that the correct input was used.

Authors response: We rechecked the code of yak in github, and confirmed its accuracy. We noticed that this issue has also been discussed in github, and the twice input was confirmed by the author (Professor Heng Li at Harvard University) (<https://github.com/lh3/yak/issues/10>).

Here is README of yak in github (<https://github.com/lh3/yak>):

#for paired end: to provide two identical streams

yak count -b37 -t32 -o sr.yak <(zcat sr.fq.gz) <(zcat sr*.fq.gz)*

P45 Methods: please add the version information for Jasmine

Authors response: The version is Jasmine (v1.1.5), and we added the version information in the revised manuscript (**Page 46, Line 1209**).

Typos:

P40: typo - “BLUEPRI>N<T”

Suppl. P25 / panel c: arch>ai<c

Use of Kb / “k”: I assume the journal has guidelines for this, but, generally, the SI prefix for “kilo” is a lowercase “k” and I would hence suggest to use “kbp” as the unit for 1,000 bp throughout the manuscript; same for “Mbp” and “Gbp” where applicable

Authors response: In the revised manuscript, we have carefully checked the text, and corrected these typos and unified the abbreviation usage (e.g. Kbp => kbp; Mb/Gb => Mbp/Gbp; INDEL => indel).

Referee #3

In this revision, the authors have removed many overstatements in the previous version and focused on characterization of genetic variation in the MESA populations. These changes make the paper more suitable for publication, but the story also became less attractive because many suggestions from the reviewers were put as future work (and we know that “later” often means “never”). Anyway, most of my comments have been properly addressed by the authors. I have a few more suggestions for the authors to consider.

1) While I still feel the long-read sequencing part is a bit unconnected to most of the paper based on the short-read sequencing data, the authors have made a good effort to strengthen the connection by performing SV calling using the short-read sequencing data for SVs identified by the long-read data.

With the SV data from all individuals, can we combine SVs and small variants into an imputation reference panel? That is, can we use this resource to better impute SVs when we only have SNP array data? In addition, because SVs are likely to have greater functional impacts than small variants, they should be included in the analysis of medical implications.

Authors response: We appreciate the reviewer’s suggestion on combining SVs and small variants into the imputation reference panel. Following reviewer’s suggestion, we updated the SEA3K imputation reference panel combining the SVs and small variants. This new panel can impute up to 31.91 million SNVs/indels and 42,286 SVs. For MSEA populations, this panel shows a 0.6-2.2 fold increase of imputation accuracy for all variants, and a 0.3-1.7 fold increase for SVs compared with the current available panels (e.g. 1KGP) (**Figure R4**).

We added this results in the revised manuscript (**Figure R4**, added as **Supplementary Figure 9** in the revised manuscript), and at the same time, we added this SV-combined panel into the SEA3K Imputation Server (<https://ngdc.cncb.ac.cn/SEA3K/>) (**Page 10, Line 243-246; Page 12, Line 298-300**).

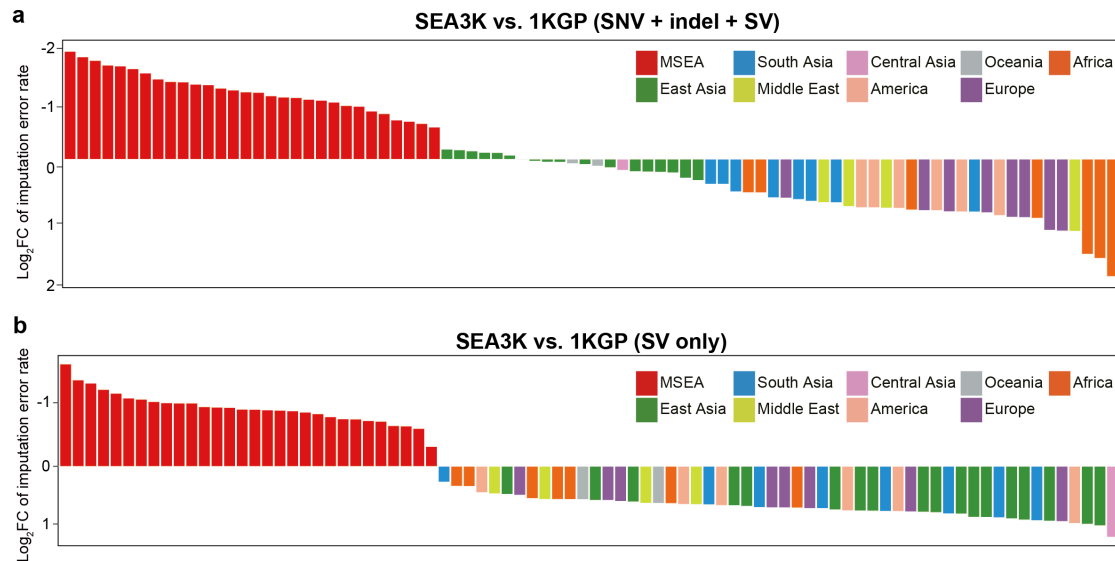


Figure R4. Evaluation of imputation performance using the SEA3K reference panel. Comparison of the fold changes ($\log_2FC$) of the imputation error rates for the worldwide populations using either the SEA3K panel or the 1KGP panel. **(a)** Comparison of imputation performance for all variant combined (SNV + indel + SV) (updated as **Figure 1e** in the revised manuscript). **(b)** Comparison of imputation performance for SV only.

For the medical implications of SVs, following the reviewer's suggestion, we annotated the SVs in the SEA3K dataset with ClinVar Database (Methods). However, we did not find any SV falling in the list of the ClinVar reported disease-related SVs.

We speculate that there are two possible reasons for this observation:

1) For characterization of SVs in the SEA3K samples, our strategy is that we first resolved the precise breakpoints of SVs using the 37 long-read WGS data and phased assemblies, then we genotyped these SVs in all SEA3K samples with deep short-read WGS. This strategy only allows us to discover relatively common SVs in MSEA populations ($>2\%$, 1/74 haplotypes). Considering that almost all SVs reported in the ClinVar database are very rare (often by case reported with a few samples), therefore, our SV sets is hard to capture these rare mutations.

2) Another reason is that currently there is too little data for disease-related SVs in the ClinVar database compared to the disease-related short variants (4,054 SVs vs. 2,231,347 short variants; data resource (v20231015): https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/archive_2.0/2023/clinvar_2023_1015.vcf.gz).

When we relaxed our cutoff to intersect the SEA3K SVs and the ClinVar SVs

(reciprocal 50% overlap SV region without exact match SV breakpoints), 8 SVs in MSEA populations were annotated as “pathogenic” (level 5 in clinical significance) (**Table R3**, added as Supplementary Table 19 in the revised manuscript), involving 11 genes (*HBA1*, *HBA2*, *HBQ1*, *LOC106804612*, *LOC106804613*, *LOC130058090*, *LOC130058091*, *FGF14*, *MARCHF6*, *RSPH1* and *USH1C*). The related diseases include alpha thalassemia, cerebellar ataxia, epilepsy, familial adult myoclonic, primary ciliary dyskinesia, spinocerebellar ataxia and Usher syndrome type 1. We added this information in the revised manuscript (**Page 26, Line 650-654**):

“Additionally, we annotated the SEA3K SVs using the ClinVar database (Methods), and we did not find any exactly matched SVs in the ClinVar reported disease-related SVs, though there are eight SEA3K SVs showing partial reciprocal overlap (>50%) with the ClinVar SV regions (Supplementary Table 19).”

Table R3. Eight SVs overlapped with the ClinVar disease-related SV regions. The detailed information can be found in **Supplementary Table 19** of the revised manuscript.

8 SVs in SEA3K					Overlapped SVs in ClinVar									
Chr	Start	End	SV type	SV length (bp)	Chr	Start	End	SV type	SV length (bp)	Sequence	CLNHGVS	Classification	ClinVar_Genes	ClinVar Disease
16	165396	184700	deletion	19,304	16	171844	181554	deletion	9,710	AGGGATGCA	NC_000016.10:g.171844-181554	Pathogenic	HBA1 HBA2 HBQ1 LOC1068	Alpha Thalassemia
16	173679	177491	deletion	3,812	16	172003	177199	deletion	5,196	TACAACCAAG	NC_000016.10:g.172003-177199	Pathogenic	HBA1 HBA2 LOC1068	Alpha Thalassemia
					16	172870	176674	deletion	3,804	GCCGGCACT	NC_000016.10:g.172870-176674	Pathogenic	HBA1 HBA2 LOC1068	Alpha-thalassemia and related disorders
21	42477110	42477224	deletion	114	21	42477152	42477291	deletion	139	CCCACAGCC	NC_000021.9:g.42477152-42477291	Pathogenic	RSPH1	Primary ciliary dyskinesia
21	42477182	42477255	deletion	73	21	42477152	42477291	deletion	139	CCCACAGCC	NC_000021.9:g.42477152-42477291	Pathogenic	RSPH1	Primary ciliary dyskinesia
11	17527112	17527382	insertion	270	11	17527112	17527427	insertion	315	ACCTGCTCC	NC_000011.10:g.17527112-17527427	Pathogenic	USH1C	Usher syndrome type 1
13	102161675	102162052	insertion	377	13	102161574	102162174	insertion	600	AAAGAAGAA	NC_000013.11:g.102161574-102162174	Pathogenic	FGF14	Spinocerebellar ataxia
13	102161699	102162504	insertion	805	13	102161574	102162174	insertion	600	AAAGAAGAA	NC_000013.11:g.102161574-102162174	Pathogenic	FGF14	Spinocerebellar ataxia
					13	102161574	102162471	insertion	897	AAAGAAGAA	NC_000013.11:g.102161574-102162471	Pathogenic	FGF14	Cerebellar ataxia
5	10356402	10360919	insertion	4,517	5	10356343	10359483	insertion	3,140	ATTTATTTTA	NC_000005.10:g.10356343-10359483	Pathogenic	MARCHF6	Epilepsy/Familial adult myoclonic epilepsy

2) Regarding the imputation, results show that the SEA3K panel is only useful for improving imputation of MSEA populations, most of which are indigenous. This result is not surprising, but it makes the MSEA panel unlikely to be widely used given the difficulty to perform large-scale GWAS in these populations. A better way is to combine SEA3K with other datasets to create a MEGA panel.

Authors response: We agree with the reviewer that a MEGA panel would be much more powerful for imputing diverse populations worldwide. To make the MSEA panel more useful, we tried to generate a MEGA reference panel by combining SEA3K with the other reference panels.

We combined SEA3K, 1KGP and SG10K panel as a MEGA panel (the other panels such as TOPMed and GAsP were not included due to the controlled data access). Because over half of variants are unique to either SEA3K, SG10K or 1KGP, a simple intersection would lead to a loss of most variants. Therefore, we employed the reciprocal imputation strategy to merge two datasets to their union set of variants

according to the previous studies (Huang, et al. 2015; Wu, et al. 2019). We combined the SEA3K and 1KGP first as “SEA3K+1KGP panel”, and then combined “SEA3K+1KGP panel” with SG10K as “SEA3K+1KGP+SG10K panel”, namely the MEGA panel.

The “SEA3K+1KGP panel” improved imputation compared with the 1KGP-only panel in all Southeast Asian, East Asian, some Oceanian and South Asian populations, as well as in the other global populations (**Figure R5**). When SG10K was added, the MEGA panel further improved the imputation for Asian and Oceanian populations, especially for South Asian populations. For the 84 tested populations, 60 populations showed better imputation when using the MEGA panel compared to the 1KGP-only panel, and the other global populations showed slightly worse imputation than the 1KGP-only panel (**Figure R5**).

Although our MEGA panel shows better imputation for Asian and Oceanian populations, this panel was not ideal for imputing samples from Africa, the Middle East, Europe, and America (**Figure R5**). This observation is consistent with previous studies, likely due to the errors introduced by the reciprocal imputation approach (Huang, et al. 2015; Wu, et al. 2019). Therefore, future efforts to combine multiple large WGS datasets into a substantially larger haplotype resource should need be made and will likely increase imputation performance. We understand the final goal of imputation study is to generate a MEGA panel that can be applied universally to impute worldwide populations. This needs to combine all WGS data sets with a large sample size, population diversity and variant coverage, and it takes tremendous efforts to collect all these public WGS data. Therefore, in this study we tried to generate a MEGA reference panel by combining three data set we can obtain. A better way is that we share our MSEA panel with the scientific groups who want to achieve a global MEGA panel in the future.

We added this result in the revised manuscript and added the MEGA panel into the SEA3K imputation server.

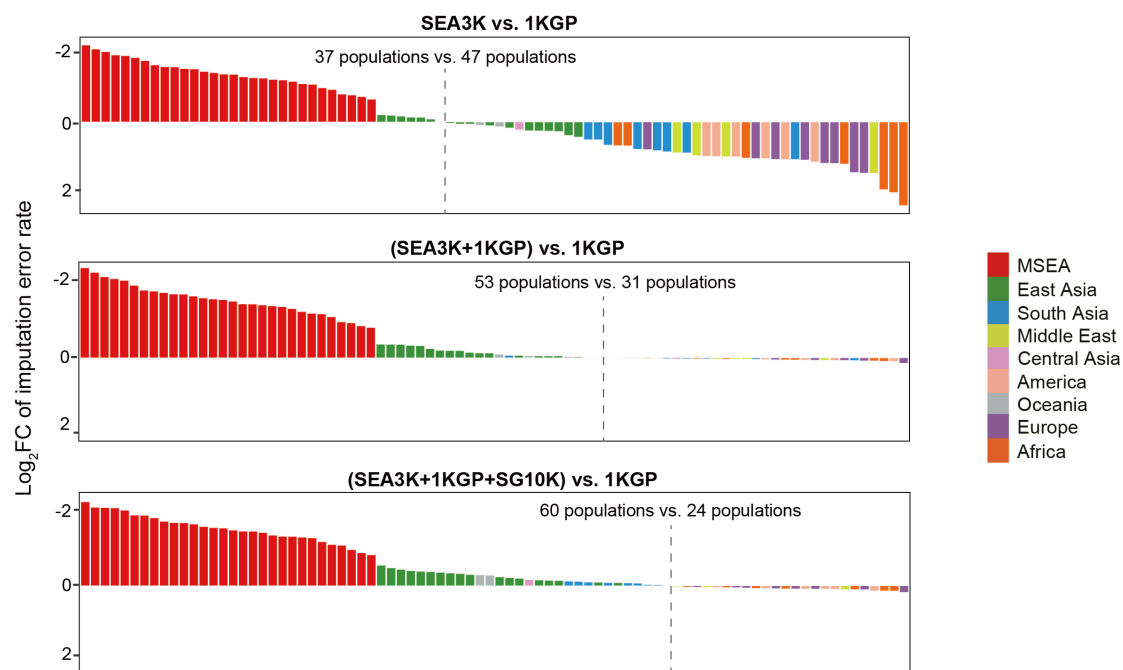


Figure R5. Evaluation of imputation performance using different combinations of reference panels. Comparison of the fold changes ($\log_2FC$) of the imputation error rates for the worldwide populations using either the SEA3K, SEA3K+1KGP, and SEA3K+1KGP+SG10K panel (MEGA panel) or the 1KGP panel. The population number with better imputation performance by different panels (SEA3K, SEA3K+1KGP, and SEA3K+1KGP+SG10K) or 1KGP panel are indicated, respectively.

3) In Supplementary Figure 3b and c, does “global populations” mean the combined data of all 56 HGDP populations? These populations are highly diverse and should not be pooled together for the calculation of MAF and R², which might make some results difficult to interpret. For example, SEA3K seems to perform worse than the SG10K panel when imputing variants with MAF<0.2 in Supplementary Figure 3b, but the R² statistic shown in Supplementary Figure 3c is much higher than the SG10K panel. Why? Please add some discussion on the imputation results. For example, why GAsP can impute global populations almost as good as TOPMed and better than 1KGP, given that GAsP has much smaller sample size and only focuses on Asian populations. These results are difficult to explain.

Authors response: We agree with reviewer that global populations should not be pooled together considering their high genetic diversity, which is not an equal comparison to MSEA population and might brought the results difficult to interpret. Therefore, we picked three representative population groups from global populations of HGDP, including East Asian (EAS, n = 205), European (EUR, n = 155) and African

(AFR, $n = 104$), and we took these three groups as the target populations and compared with MSEA populations (MSEA, $n = 300$ from SEA3K), respectively.

We observed a consistent pattern between the imputation accuracy evaluations using the two methods (namely, average Pearson correlation coefficient (R^2_{PCC}) and counts of high-quality imputed variants ($R^2_{\text{Est}} > 0.8$)). For example, when the target population is EAS population, SG10K panel outperforms the SEA3K panel when imputing variants with $\text{MAF} < 0.2$, which is consistent with the R^2 statistic of **Figure R6a** and **Figure R6b** (updated as **Supplementary Figure 3b and 3c** in the revised manuscript). When the target population is EUR or AFR, the SEA3K and SG10K panel also exhibit consistent patterns in different frequency ranges. We reasoned that the inconsistent pattern in the previous version was brought by our inappropriate strategy of pooling all populations together. In the previous version, we randomly extracted 300 individuals from HGDP dataset as “global population”, and we tested them in different imputation panels, which might lead to biased results if these 300 randomly picked samples contain more samples of one ancestry (e.g. East Asian).

The issue about the GAsP and TOPMed can be explained according to our new results. TOPMed outperforms other panels when the target population is EUR and AFR, and the GAsP and MEGA panel show better performance than the other panels when the target population is EAS.

We updated **Supplementary Figure 3b and 3c** in the revised manuscript accordingly.

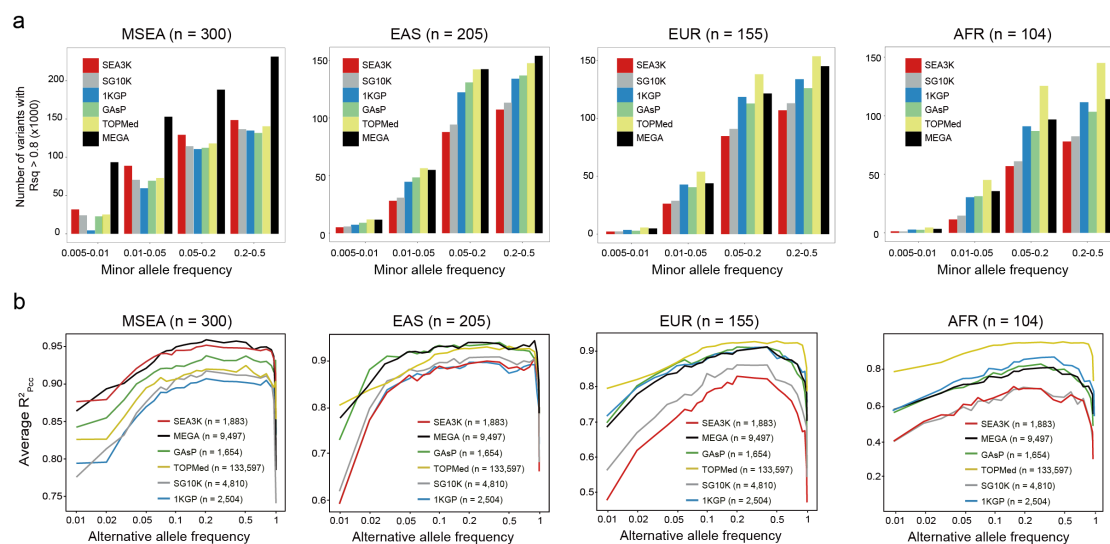


Figure R6. Evaluation of imputation performance using the split populations as target populations. (a) Number of high-quality imputed variants (using estimated imputation accuracy from Minimac4: $R^2_{\text{Est}} > 0.8$) in different minor allele frequency (MAF) bins for MSEA and global populations. The sample size of the target population is indicated, including 300 MSEA samples from SEA3K, 205 East Asian (EAS)

individuals, 155 European (EUR) individuals and 104 African (AFR) individuals from HGD dataset. The MEGA panel is the combination of the SEA3K, 1KGP and SG10K datasets (Methods). **(b)** The imputation accuracy evaluation using squared Pearson correlation coefficient (R^2_{Pcc}) between genotypes called from the sequence reads and the imputed genotypes as a function of alternative allele frequency for four reference panels.

4) Regarding results in Figure 1e and 1f, please make clear that the target population is MSEA populations and the better performance of SEA3K and GAsP is likely because these two panels include more indigenous SEA populations than the other panels.

Authors response: According to the reviewer's suggestion, in the revised manuscript, we added a sentence in the main text and legend of **Figure 1f**, and we also added a label on **Figure 1f** to clearly indicate MSEA populations being the target population in the analyses. We also added a sentence to describe that a better performance of SEA3K and GAsP is observed because of more indigenous MSEA populations in these two panels were included than the other panels.

Main text (**Page 10, Line 232-235**):

“As expected, a better performance of SEA3K and GAsP was observed when the target population is MSEA population, because these two panels include more indigenous SEA samples than the other panels (Fig. 1f)

5) For the Procrustes analysis, the authors reported the M2 statistic, which was defined as the sum of squared differences between the two coordinates (page 15). Do you mean the Procrustes t statistic instead? The t statistic is a similarity score ranged between 0 and 1, which the M2 statistic is a distance score. Please check and make sure. For more about the definition of these statistics, please refer to Noah Rosenberg's paper in 2010 (PMID: 20196748).

Authors response: According to the suggestion of the reviewer, we also calculated the Procrustes similarity statistic t according to the equation in (Wang, et al. 2010). The Procrustes similarity statistic $t_0 = 0.548$ (P -value $< 1 \times 10^{-5}$ based on 100,000 permutation tests), and the rotation angle of the PCA map $\theta = 16.11^\circ$ (**Figure R7**). This result is similar with the result of M^2 statistics ($M^2 = 0.732$, P -value $< 1 \times 10^{-5}$, $\theta = 14.0^\circ$).

For the definition of M^2 statistics, according to the original paper (Gower 1975), the M^2 refers to the sum of the squared deviations between landmarks when the plots are optimally aligned. The M^2 statistics ranges from 0 to 1, and a smaller M^2 means a higher similarity between two plots.

For the definition of t statistics, according to (Wang, et al. 2010; Wang, et al. 2012), the Procrustes similarity statistic $t_0 = \sqrt{1 - D}$, in which D is the minimum sum of squared Euclidean distances between corresponding points in the two plots when the plots are optimally aligned. The similarity statistic t_0 ranges from 0 to 1, and a bigger t_0 means a higher similarity between two plots.

Considering t statistic is more intuitive in the Procrustes context, we updated the results of Procrustes analysis using t_0 instead of M^2 , and cited the methodology papers (Wang, et al. 2010; Wang, et al. 2012) in the revised manuscript. The code for calculating the t statistics were uploaded onto github: <https://github.com/YurunLu/ProcrustesAnalysisT/>.

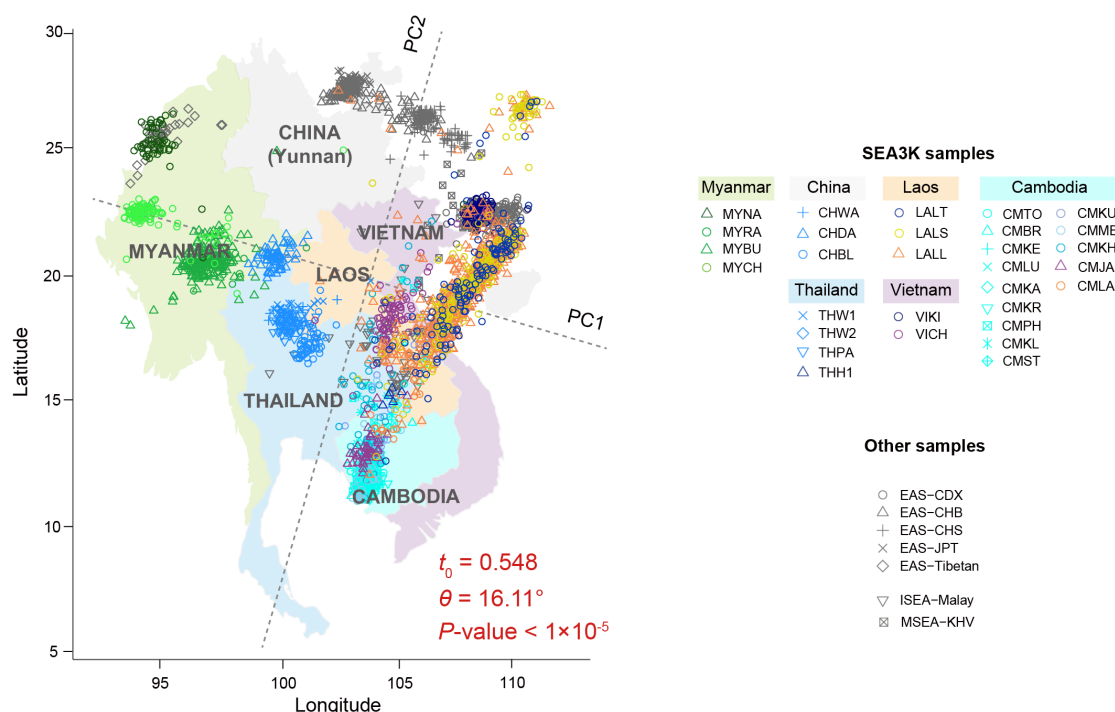


Figure R7. Procrustes analysis of genetic and geographic coordinates of MSEA populations using t statistics. The Procrustes-transformed PCA plot of genetic variants is presented. The Procrustes analysis was conducted based on Procrustes similarity statistic t described in (Wang, et al. 2010). The Procrustes similarity statistic $t_0 = 0.548$, $P\text{-value} < 1 \times 10^{-5}$ based on 100,000 permutation tests, and the rotation angle of the PCA map is 16.11 degrees clockwise.

6) In Supplementary Figure 15, for better comparison, please use the same color code and y-axis scale for results based on SMC++ and MSMC2. When two methods were applied on the same dataset, are the results consistent? Please mention in the main text and be clear that which method is chosen as the main method and why.

Authors response: According to the reviewer's suggestion, for better comparison, we

kept the same color code for populations and y-axis (effective population size, N_e) in the plots based on SMC++ and MSMC2 (**Figure R8**, updated as Supplementary Figure 18 in the revised manuscript).

In general, we observed consistent patterns of inferred population size history based on SMC++ and MSMC2 (**Figure R8**, updated as Supplementary Figure 18 in the revised manuscript). However, we should emphasize that the input datasets are not the same between SMC++ and MSMC2 when inferring population size history for a population, due to the different algorithms of these two methods.

In brief, MSMC2 uses structural/linkage-based information in its estimation of historical demography (Malaspina, et al. 2016). Although MSMC2 allows for demographic inference using multiple individuals, it is computationally limited in the number of haplotypes it can take as input (maximum of eight haplotypes, or four diploid individuals) (Malaspina, et al. 2016; Patton, et al. 2019). Therefore, for a population analysis, we randomly chose four individuals from the analyzed population with phased genotypes to run MSMC2.

By contrast, SMC++ uses structural/linkage-based information, and at the same time incorporates the site-frequency spectrum (SFS) in its estimation, and it requires unphased genotypes and can analyze over thousands of individuals (Terhorst, et al. 2017). Therefore, we input all individuals with unphased genotypes of the analyzed population to run SMC++ (Methods in the revised manuscript).

For a certain population, the input data for SMC++ and MSMC2 was not exactly same considering different sample size and phased/unphased genotype. Despite this, we still observed generally consistent patterns between SMC++ and MSMC2. The inconsistency for several populations during recent time scale might reflect the varied performance using different algorithms.

We chose to present the MSMC2 result due to two reasons:

- 1) SMC++ is an SFS-based method, which exclusively relies on reliability of the constructed frequency spectrum of the investigated population (Terhorst, et al. 2017). Hence, SMC++ is highly sensitivity to sample size. It was reported that SMC++ showed high accuracy in big datasets (>20 haploid genomes), but poorly performed when using smaller datasets (Patton, et al. 2019; Sellinger, et al. 2021). Considering the sample sizes of MSEA populations vary greatly (from 2 to 636 individuals), MSMC2 is more appropriate.

- 2) Another reason for choosing MSMC2 is the robustness of methods. Although all methods for inferring population size history have limited resolution for inferring the

recent past, multiple independent studies have reported that SMC++ displayed higher variance when evaluating N_e of recent time, reflecting a poor robustness of SMC++ for diverse datasets (Hu, et al. 2023; Sellinger, et al. 2021).

According to the reviewer's suggestion, in the revised manuscript, we added a few sentences to describe the comparison of SMC++ and MSMC2, and explained why we chose MSMC2 (**Page 38, Line 988-998**):

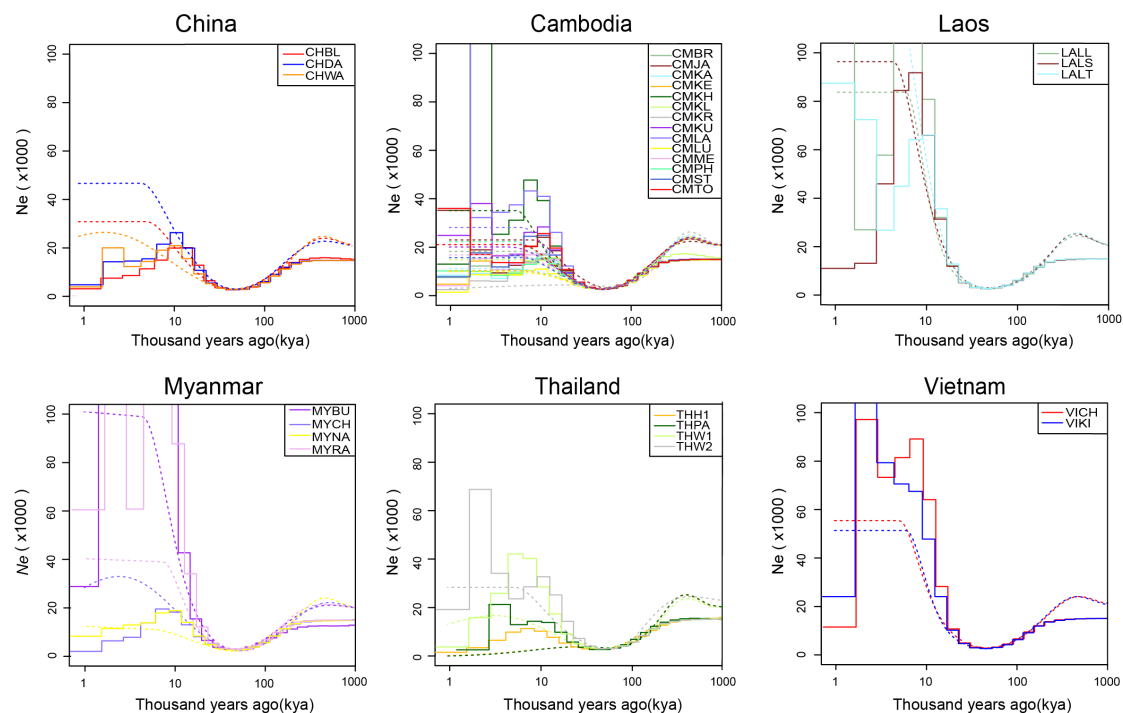


Figure R8. Historical population dynamics of MSEA populations. The plots show the results of the estimated effective population size (N_e) of the ancestral populations in MSEA populations (stratified by countries) using SMC++ (dotted line) and MSMC2 (solid line) methods. The generation time and the mutation rate per generation per site (μ) in the analyses were set as 29 years and 1.25×10^{-8} in both MSMC2 and SMC++.

References

- Browning, S. R., et al. 2018 Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. *Cell* 173(1):53-61.e9.
- Chen, L., et al. 2020 Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals. *Cell* 180(4):677-687.e16.
- Cingolani, P., et al. 2012 Using *Drosophila melanogaster* as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. *Front Genet* 3:35.
- Gower, J. C. 1975 Generalized procrustes analysis. *Psychometrika* 40(1):33-51.
- Hu, W., et al. 2023 Genomic inference of a severe human bottleneck during the Early to Middle Pleistocene transition. *Science* 381(6661):979-984.

- Huang, J., et al. 2015 Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat Commun* 6:8111.
- Lin, J., et al. 2022 SVision: a deep learning approach to resolve complex structural variants. *Nat Methods* 19(10):1230-1233.
- Malaspinas, A. S., et al. 2016 A genomic history of Aboriginal Australia. *Nature* 538(7624):207-214.
- Mao, Y., et al. 2024 Structurally divergent and recurrently mutated regions of primate genomes. *Cell* 187(6):1547-1562 e13.
- Patton, A. H., et al. 2019 Contemporary Demographic Reconstruction Methods Are Robust to Genome Assembly Quality: A Case Study in Tasmanian Devils. *Mol Biol Evol* 36(12):2906-2921.
- Sellinger, T. P. P., D. Abu-Awad, and A. Tellier 2021 Limits and convergence properties of the sequentially Markovian coalescent. *Mol Ecol Resour* 21(7):2231-2248.
- Terhorst, J., J. A. Kamm, and Y. S. Song 2017 Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat Genet* 49(2):303-309.
- Wang, C., et al. 2010 Comparing spatial maps of human population-genetic variation using Procrustes analysis. *Stat Appl Genet Mol Biol* 9(1):Article 13.
- Wang, C., S. Zollner, and N. A. Rosenberg 2012 A quantitative comparison of the similarity between genes and geography in worldwide human populations. *PLoS Genet* 8(8):e1002886.
- Wang, S., et al. 2024 De novo and somatic structural variant discovery with SVision-pro. *Nat Biotechnol*.
- Wu, D., et al. 2019 Large-Scale Whole-Genome Sequencing of Three Diverse Asian Populations in Singapore. *Cell* 179(3):736-749 e15.
- Zhou, B., et al. 2024 Detection and analysis of complex structural variation in human genomes across populations and in brains of donors with psychiatric disorders. *Cell* 187(23):6687-6706 e25.
- Zhou, D., Q. Wang, and H. Liu 2021 A novel mutation in RAB3GAP1 gene in Chinese patient causing the Warburg micro syndrome: A case report. *Medicine (Baltimore)* 100(2):e22902.

Referees' comments:

Referee #3 (Remarks to the Author):

The authors have addressed all my comments. I only have 1 more minor question:

In Figure R6, when imputing EAS populations, SG10K seems to outperform 1KGP in terms of R^2 across all frequency bins (panel b), but 1KGP can impute many more variants with $R^2 > 0.8$ (panel a). How to explain these results, which seem inconsistent? Better to check if any calculation went wrong.

Authors response: We rechecked the imputation results of EAS populations. We found that the bar plot about statistics of high-quality imputed variants ($R^2 > 0.8$ by Minimac4) was an old version by mistake. In the revised manuscript, we replaced the bar plot of EAS with the updated version in **Supplementary Figure 2 (Figure R1)**, and the pattern of imputation accuracy is consistent between the two R^2 statistics.

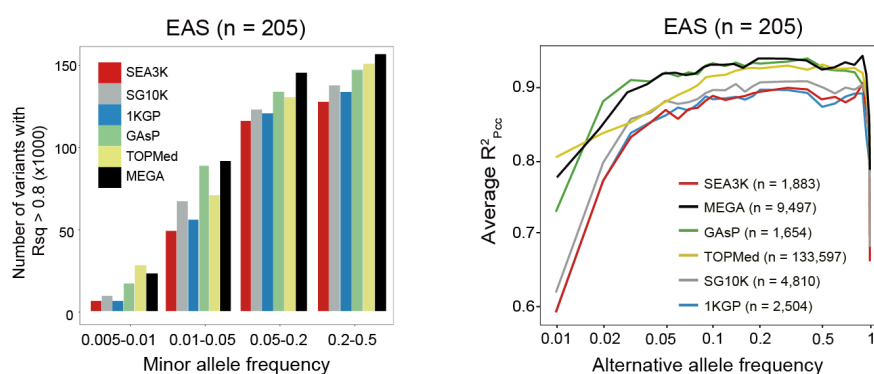


Figure R1. Evaluation of imputation performance in EAS populations. (a) Updated bar plot shows the number of high-quality imputed variants (using estimated imputation accuracy from Minimac4: $R^2_{Est} > 0.8$) in different minor allele frequency bins for East Asian (EAS) populations ($n = 205$). **(b)** The imputation accuracy evaluation using squared Pearson correlation coefficient (R^2_{Pcc}) between genotypes called from the sequence reads and the imputed genotypes as a function of alternative allele frequency for four reference panels.

Referee #5 (Remarks to the Author):

I appreciate the effort the authors have put into compiling and analyzing this extensive dataset. While I am not a specialist in DNA or genetics, my review focuses on the manuscript's treatment of Indigenous communities and potential language that might be deemed pejorative or problematic in reference to the peoples who are the subject of this study.

Authors response: We appreciate the reviewer's approval of our study and the constructive comments, we have addressed these points below and updated the related parts in the revised manuscripts accordingly.

Comments:

Recent Studies on Island Southeast Asia (ISEA)

The manuscript would benefit from incorporating recent genetic studies that parallel its findings, particularly:

Larena, Maximilian, Federico Sanchez-Quinto, Per Sjödin, James McKenna, Carlo Ebeo, Rebecca Reyes, Ophelia Casel et al. "Multiple migrations to the Philippines during the last 50,000 years." Proceedings of the National Academy of Sciences 118, no. 13 (2021): e2026132118.

Karmin, Monika, Rodrigo Flores, Lauri Saag, Georgi Hudjashov, Nicolas Brucato, Chelzie Crenna-Darusallam, Maximilian Larena et al. "Episodes of diversification and isolation in Island Southeast Asian and Near Oceanian male lineages." Molecular Biology and Evolution 39, no. 3 (2022): msac045.

These studies offer relevant comparative studies (within the region) into genetic diversification and migrations, complementing the findings of the manuscript.

Authors response: We agree with the reviewer that the findings of genetic studies on island Southeast Asia (ISEA) will help to complement the discoveries of our study, especially the two recent studies mentioned by the reviewer. By undertaking extensive new sampling of genomic variants, ancient DNA and Y-chromosome in ISEA-Oceania region, these studies elucidated the complex demography history of ISEA and Oceania population, including multiple migrations, rapid diversifications and extensive expansions. These studies provide genetic evidence of early movements between MSEA and ISEA well predating the Neolithic period. At the same time, these studies highlight the importance of sample diversity and high-density variants for conducting population genetics research, consistent with our research on the MSEA population.

According to the reviewer's suggestion, in the Discussion section of our revised manuscript, we added sentences and references to echo the published studies of ISEA.

(Line 385-388)

"We showed that MSEA populations hold rich genomic diversity, largely due to their shared deep history and the Neolithic admixture with East Asian farmers, which was also seen in the genomic features of island Southeast Asians (Larena et al. 2021; Karmin et al. 2022)"

Terminology: "Tribal"

The term “tribal” carries connotations linked to outdated evolutionary hierarchies and is often perceived as implying a lower stage of social complexity. A more neutral alternative, such as “ethnolinguistic groups” or simply “communities,” would be preferable.

Authors response: According to the reviewer’s suggestion, we removed the term “tribal” from the manuscript, and used the term “indigenous”. At the same time, considering the term “indigenous” has various definitions in different contexts, we defined the “indigenous” in our manuscript as the reviewer suggested in the comment below.

Defining "Indigenous"

The manuscript uses the term “Indigenous” but does not clarify its specific application within the context of Southeast Asia. Given the political weight of this term in the region, the authors should clearly define whether they mean local populations in general or groups that are historically marginalized and legally recognized as Indigenous. Currently, the manuscript presents these populations as static and unchanging, which does not reflect the dynamic and adaptive nature of these societies.

Authors response: We totally agree with the reviewer that there is no generally accepted definition of indigenous peoples, and the term “indigenous” has various definitions in different contexts. Therefore, there is a need of defining “indigenous” within the context of SEA in our manuscript. In fact, the indigenous people in our study are more consist with the definition mentioned by the reviewer that those groups who are historically marginalized and legally recognized as indigenous.

In the revised manuscript, we defined the “indigenous people” according to the United Nations (https://www.un.org/esa/socdev/unpfii/documents/5session_factsheet1.pdf).

(Line 105-106)

“the indigenous populations, who are distinct from those of the dominant societies in which they live”.

Framing of "Hunting-Gathering Lifestyle"

The use of the term "hunting-gathering lifestyle" reinforces outdated assumptions about certain communities being "primitive" or "backward." The manuscript would be more precise and less value-laden if it referred to “traditional subsistence strategies” instead. There are no pure hunter-gatherers today; most communities practice mixed economies,

even if they engage in seasonal foraging.

Authors response: According to the reviewer's suggestion, we rephased the "hunting-gathering lifestyle" as "traditional subsistence strategies" in the revised manuscript.

Scope of Mainland Southeast Asia (MSEA)

The manuscript defines MSEA in a way that excludes significant parts of the region. MSEA is not limited to the Indochina Peninsula—it also includes Myanmar, Thailand, and Peninsular Malaysia. The authors should ensure that their geographic framing aligns with established regional definitions.

Authors response: We thank the reviewer for pointing out the inaccurate description on scope of MSEA, and we corrected it accordingly in the revised manuscript (**Line 100**).

Clarity of Research Questions and Hypotheses

The manuscript presents findings in a largely descriptive manner without clearly articulating the central research questions or hypotheses. Are the authors attempting to identify common patterns across MSEA populations, or are they simply reporting genomic results? A stronger discussion on the "so what" aspect of the findings is needed. What are the broader implications of their conclusions? What are the consequences of their findings, and how do they contribute to addressing contemporary challenges faced by local communities in MSEA? A clearer articulation of these points would significantly improve the manuscript.

The study makes a valuable contribution to our understanding of the genetic diversity of Southeast Asia. However, a more careful approach to language and framing will enhance its impact and avoid potential misinterpretations. Thank you for the opportunity to review this work and hope my comments prove useful to the authors.

Authors response: We agree with the reviewer that the current manuscript has a lot of descriptive results. However, despite this is common presentation of data-driven research, our study is by no means just to simply describe the common genomic pattern across MSEA populations. Instead, our study has important research hypotheses and scientific questions based on the background of previous studies. It can be summarized into three major points:

1) Genomic pattern: What is the basic genomic pattern of MSEA populations? Although there have been some previous studies based on array genotypes, chromosome Y, and mitochondrial DNA, MSEA populations are still understudied in the global genome research. Therefore, our study aimed to provide the first genome-

wide landscape of genetic variations of MSEA populations, which sets the root of exploring other scientific questions, and it is the most important contribution of this work.

2) Population structure and history: According to previous genetic and anthropological evidence, MSEA has a deep and complex history of modern human inhabitation in this region, and can be traced back as early as 65 thousand years ago. However, this proposal remains a question mark due to lack of genome-wide data. Using the SEA3K genome data, for the first time, we delineated the genomic structure of MSEA populations with quantitative evaluation of the varied genetic components due to shared ancient ancestry, Neolithic population expansion from mainland East Asia and recent local population admixture.

3) Environmental adaptation and disease: Southeast Asians have been mostly living in tropical rainforest, e.g. high temperature, high humidity, strong ultraviolet radiation and prevalence of parasites and other pathogens. Do they have adaptive variations for adapting to these environments in their genomes? How have the genomic variations shaped the phenotypes and disease characteristics of current MSEA populations? We identified 89 candidate genes under positive selection in MSEA populations. The known functions of these genes are related with multiple phenotypes, including facial morphology, body height, skin pigmentation, and immune system.

To clarify the implications of this genomic study and address the “so what” comment of the reviewer, in the revised manuscript, we restructured the “Discussion” section by emphasizing the major findings and the broad impact of this study.

(Line 388-395)

“ The rich genomic diversity in MSEA populations have been shaped by different evolutionary forces, including geographic isolation, regional admixture, archaic introgression and natural selection, and it holds the genetic basis of the characteristic phenotypes (presumably adaptive) and disease prevalence of Southeast Asian populations that call for future in-depth studies. Collectively, the SEA3K genome data sets up a basis for new discoveries and application of precision medicine in Southeast Asia, and it also brings genomic insights that eventually will benefit the health and wellbeing of Asian and global populations.”

This paper by He et al. presents the SEA3K genome dataset (Phase I), consisting of 3023 short-read WGS samples and 40 long-read WGS samples from 30 mainland Southeast Asian (MSEA) populations. The authors performed a series of standard population genetic analyses based on the short-read WGS data, including characterization of variants, imputation accuracy, population structure, demographic history, and selection. In addition, based on the long-read WGS data, the authors characterized structural variants (SVs) and constructed the first pangenome of MSEA populations. This SEA3K genome dataset is a valuable resource because MSEA consists of rich information of human genetic diversity and MSEA populations are currently underrepresented in human genetic studies.

Overall, it seems to me that the two components of this dataset, the short-read and long-read sequencing data, had little connection in this paper. When analyzing the short-read sequencing data, the authors didn't make full use of their advantage of having a MSEA pangenome from long-read data. And the paper was not organized in a logical way – the long-read WGS data were inserted in the middle of short-read WGS results without clear connection. Ideally, the paper should start with a pangenome constructed from long-read sequencing data, including SVs. Then, short-read sequencing data should be mapped to the pangenome, followed by variant calling and all the downstream analyses. Thus, short-read WGS data analysis can take full advantage of the long-read data. It would be very interesting to find out how much the long-read data can facilitate short-read data analysis. For example, can we perform better in calling SVs when the short-read WGS data were mapped to the pangenome, and subsequently imputing SVs when using the short-read WGS data as the imputation reference panel? Otherwise, why do we need a pangenome?

Other specific comments:

1. Imputation. Figure 1E and Figure S2 are highly similar to Figure 5 of Wu et al. 2019 (Cell, PMID: 31626772). Clearly, the authors had read the paper and tried to follow. But why didn't they make a direct comparison with Wu et al. (2019), which presented the Phase I dataset of the SG10K Project, consisting of WGS data of 4,810 Singaporeans, including 903 Malays? The paper Wu et al. was only cited improperly as “small-scale genome data in Malays” in the Introduction (line 2 on page 5). The SG10K Project has been expanded to 9,051 genomes (Nat Commun in 2022, PMID: 36335097) but was not even mentioned. Did the SEA3K dataset performed better than the SG10K (or the Phase I dataset by Wu et al.) on the imputation in Asian populations? In addition, it is recommended to provide key information of the reference panels, especially the sample size. For example, what is the sample size of the GAsP panel? There are two versions of GAsP on the UM imputation server, consisting of 1,654 and 6,461 individuals, respectively. Which one did the authors used?
2. Again, on imputation. What is the target population of Figure 1F? There is no clear description in the figure caption, main text and even in the method section. Based on the method section, it seems the target population is 300 individuals randomly selected from the SEA3K dataset? But this information was only mentioned in a different paragraph. Let's assume this is correct. Were these 300 individuals phased together with the other SEA3K samples? If so, they were not independent and might cause overestimation of the imputation accuracy of the SEA3K reference panel.
3. Population structure. In Figure 3A, the authors showed that the PCA plot mimics the geographic

map followed by some rotation transformation. This can be done in a formal and quantitative way using the Procrustes analysis (please see Wang et al. 2012 PLoS Genet, PMID: 22927824).

4. Population structure. The authors over-interpret results from ADMIXTURE analysis with demographic history, such as the origin of different ancestry components and the evolutionary timing (second paragraph on page 13). ADMIXTURE is just a non-supervised clustering algorithm. It is well-known that many different demographic models can yield the same clustering patterns in the ADMIXTURE results (please see Lawson et al. 2018 Nat Commun, PMID: 30108219).
5. Positive selection. The authors tried to claimed novel selection signals at the SNV level (for example, they claimed 7 out of 13 SNVs were novel selection signals). Traditionally, positive selection signals were reported at the gene level because genes are the functional unit on which selection is acted on. Thus, claiming novelty at the SNV level might be inappropriate and overstated. Better to report how many genes or loci harbor novel selection signals.
6. Medical implications. The authors provide a list of “pathogenic” variants that are common in the SEA3K dataset but rare in other populations, and a list of homozygous PTVs found in the SEA3K populations. But they didn’t explain clearly the medical implications and why these are “valuable resource”. For the “pathogenic” variants common in the SEA3K population, the author only provided an example of balancing selection at the *HBA2* gene. Are they suggesting that all the others are likely to be under balancing selection as well? It seems a more plausible explanation is that many of variants were misclassified as “pathogenic” by the ClinVar database because they can reach a high frequency in MSEA populations. Regarding the homozygous PTVs, they should examine related phenotypes in the carriers in order to understand the functional consequences of these “natural knockouts” in humans.
7. Discussion, page 28, “In MSEA populations, we showed that standing variations play a dominant role”. I am not sure which result support this claim?
8. Discussion, page 28, regarding the contribution of regulatory variants and protein-coding variants to adaptive evolution. Can the authors provide a recent reference on the debate, in addition to reference 81, which was published in 1975? My impression is that the research community has already reached an agreement that both types variants are important in the evolution since over 10 years ago. Furthermore, it seems an overclaim of “a concerted adaptation attributed to both protein functional alterations and changes of gene expression regulation” because it is not surprising to see a locus with both types of variants, especially when strong positive selection can lead to extended haplotypes.
9. Quality controls. PCA and ADMIXTURE should be performed on unrelated samples. But it seems that the authors only excluded related samples for the selection analysis. Closely related samples are likely to show as outliers in the PCA, which might explain some of the PCA outliers identified in the QC. If the authors excluded one sample from a related pair first, the other one would not show as an outlier.

10. QC for long-read sequencing. Three samples (MN269, MN422, and MN582) were removed in the QC because of low sequencing quality and abnormal GC contents (page 39). But why were these low-quality samples included in the analysis of genome assembly? In the genome assembly analysis, MN025, MN269 and MN422 were removed based on assembly metrics, but MN 582 was retained. Despite the low sequencing quality of MN582, the authors included it as one of the 37 “high-quality” assemblies.
11. Introgression analysis. This study used the low-coverage WGS data of the 1KGP (phase 3 version 5a). Why didn’t they use the high-coverage WGS data of 1KGP (Cell 2022, PMID: 36055201)? In particular, the authors “assumed that YRI is monomorphic for the SEA3K major allele for variants called in the SEA3K but not called in the 1KGP” (page 41, second last paragraph), and this assumption might not hold because low-coverage sequencing data have lower sensitivity to detect rare variants. Using the high-coverage 1KGP data will be more suitable, especially when the SEA3K WGS data are high-coverage.

Minor points:

1. Page 15. “Linkage disequilibrium (LD) decay indicates the genetic structure and associates with genetic diversity of a population.” --- the meaning of this sentence is unclear. In general, LD decay pattern is a consequence of demographic history, including admixture, effective population size, and many others.
2. Figure 5B, it might be better for visualization to plot the contour plots with color indicating the density heights.
3. Figure 5C, it is very hard for me to see the green lines (introgression segments). What is the difference between dark green and light green in the plot? I just don’t understand the meaning of this plot.
4. Figure 5D, it is almost impossible to visualize the tiny blue bar indicating the Neanderthal sequence, especially when printed on an A4 paper.
5. PCA, page 35, line 3. Please define “outliers” (i.e., how many SDs from the mean?).
6. Selection, pages 37-38, please stated clearly how the P values were defined in the DAF, PBS, XP-EHH and iHS analysis. Current method section only describe how the statistics were calculated but the CMS was calculated based on P values.
7. Duplicated references of 32 and 47. And all the consortium names were not displayed properly in the reference list. Please check your Endnote format and the references carefully.
8. There are many typos, just to list a few:
 - (1) Page 13, line 5, “from Sino=Tibetans from the north” → “to Sino=Tibetans from the north”
 - (2) In Figure 3D, the y-axis ticks, “ESA” → “EAS”
 - (3) Page 15, first line of 2nd paragraph, “examed” → “examined”
 - (4) Page 18, Figure 4G, the y-axis label, “Frequence” → “Frequency”
 - (5) Page 37 line 11, “previous study” → “a previous study”