natureresearch

Peer Review File

Manuscript Title: Genomic Insights into the Formation of Human Populations in East Asia

Editorial Notes:

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referee #1 (Remarks to the Author):

In Wang et al., the authors generated and analyzed an exciting aDNA dataset from samples distributed throughout Eastern Eurasia. The sample size of the aDNA dataset is decent, sampled from regions of the world that are currently still understudied. Consistent with a long-line of pioneering work from these teams of senior investigators, the methods applied in this manuscript are established and have been shown to provide insights to historical migration events in the past.

A lot of analyses and results were presented, and the authors adequately guided the readers through the narrative to arrive at the conclusions. In general I have no strong concerns of the methodology and interpretation. My main comments are generally related to presentation and clarifications.

1. In general, the narrative is structured as a disconnected list of analyses from different parts of Eurasia. First starting with Eastern Eurasia/Mongolia, the narrative then pivot to the Amur Basin, then Tibet, Taiwan, China, and then Japan. The bulk of the analysis and insights seem to be from Eastern Eurasia/Mongolia and in China, with the remaining regions loosely connected. I see value in generating and making available aDNA data from all these different parts of Eastern Eurasia. But if there is no major insight from some of these regions (e.g. line 381-399, analysis of the Taiwan individual is really a single analysis of systematic f4 statistics; line 443-453 is a single analysis of admixture f3 using Jomon), perhaps the narrative should set up the expectation early on that this is a general survey of the new data generated, with emphasis on a couple of regions where new insights are produced.

2. To me, the most important figure that summarized the main conclusions of this manuscript is Figure 3 – I kept referring between the text and Figure 3. If this is a survey of the aDNA from multiple regions in Eastern Eurasia, the author should consider putting together an overall or multipanel figure for each of the regions they investigated (where the current Figure 3 would be a subpanel). Such a figure, though undoubtedly will be updated as new data come out in the future, will be referred to repeatedly.

3. From the abstract, the authors suggested that a first wave of Yamnaya ancestry (associated with Afanasievo culture) was nearly completely displaced by a second wave (represented by the Sintashta ancestry), with the exception of samples in Western China (represented by Xinjiang Shirenzigou data). This set up an expectation that present day individuals in this general region carries largely Sintashta ancestry (again, with the exception of Xinjiang). However, interpretation from Figure 3 (and the text starting from line 228 or so) is that most individuals sampled in this region during the early bronze age and late bronze age were actually two-way admixed. The Yamnaya ancestries, whether through Afanasievo or Sintashta, are not as prevalent as they are in modern Europeans, and arguably are more exceptions than the rules. Perhaps in Figure 3 the authors can estimate ancestral proportions of modern populations in these regions, where data is more abundant and geographically diverse, to show whether Yamnaya-related ancestry is actually prevalent today. In fact, it would be nice for Figure 3 to also include Han Chinese ancestry (a fifth

component) in the time transect, which was only briefly discussed in the text (line 278-285).

4. Line 238-253: the male Chalcolithic Afanasievo child who had no Yamnaya ancestry is touted as a striking finding. Although I find it interesting, and understand that it is the first of its case, I failed to appreciate the archaeological significance of this finding. I assume cultural diffusion is not so exceedingly rare. The authors should discuss further why this is surprising (and thus significant), or perhaps this finding would be better located in the Supplement.

5. Line 247: the Chalcolithic Afanasievo child showed genetic continuity with Ulgii individuals. How exactly is genetic continuity demonstrated?

6. Beginning at line 401, the narrative turned towards the second major insight from this manuscript. I would like the authors to clarify what is the model of the origin of Han Chinese in their mind. From the abstract it seems to imply the source population originated from upper and middle Yellow River Basin and then moved West/South into the Tibetan plateau, as well as more East to the Central Plain. Can the authors explained the logic behind the inferred outward direction of expansion from the upper and middle Yellow River Basin, rather than a uni-direction expansion from East coast towards the West and South? Also, are these models consistent with that suggested in Li et al. MBE 2019 (10.1093/molbev/msz072) where river valleys shaped the genetic landscape?

7. Line 426-432: The authors implied that varying proportions of Neolithic Wuzhuangguoliang ancestry can explain the North-South differentiation observed in modern Han Chinese. However, this is not apparent from Figure 4. Can the authors demonstrate the degree to which Wuzhuangguoliang explains the N-S differentiation?

8. I am interested in the finding of West Eurasian ancestry in Han_N_China sample (Line 423-425). We and others have reported the potential admixture signals in modern Han Chinese (Liu et al. Cell 2018 [10.1016/j.cell.2018.08.016]; Chiang et al. MBE 2018 [10.1093/molbev/msy170]). Chiang et al. estimated similar admixture proportions and dates. Is there more information on the geographical origin of the Han_N_China samples? Both Liu et al. and Chiang et al. are much more recent and updated investigations of population structure and admixture history of Han Chinese that should be cited along with the two current citations from over a decade ago (Chen et al. 2009 and Xu et al. 2009).

9. Line 443-453: In the analysis using the Jomon sample, why not use Wuzhuangguoliang sample? Particularly since the narrative left an open question whether a mainland ancestry in modern Japanese is due to an ancient ancestry that contributed significantly to both Han Chinese and Koreans (Line 451-453).

Specific presentation comments:
1. Figure 1: It would be helpful to include the sample size in the legend. Also, to this colorblind reviewer I cannot distinguish the two different colors for the Mongolian aDNA samples, please use two colors with more contrast than the current “orange”.

2. Throughout the paper, multiple geographical locations are referred to without any explicit indication where they are situated. If locations such as the Tibetan plateau, Amur River Basin, Yellow River Basin, Tarim Basin, etc. could be indicated on Figure 1 will help orient the readers tremendously.

3. I will respect the editorial / journal policy on this, but I always felt that it would be a nice gesture to include the archaeological site name in the native language, at least in parenthetical statements at first usage in the supplement, in deference to the local culture.

Additional minor comments:

1. Line 166-167: how exactly were individuals “clustered”?

2. Line 354-356: similarly, how was this clustering conducted? By hierarchical clustering? Or by visual inspection?

3. Line 178-180: I do not think ALDER estimates multiple episodes of admixture; it assumes a single pulse event. I think authors meant multiple populations each having a signature consistent with past admixture. Should clarify the writing to avoid confusion.

4. Line 425: I cannot find results that estimate 2-4% admixture in Northern Han Chinese from Tables S14 or S15.

5. Can the MSMC results be shown in the supplement?

6. In general, the authors can refer more specifically to supplemental or online tables for the reader. Some of these tables are extensive and a generic reference is not helpful. For example: at Line 207, I believe the relevant reference is Online Table 5B

7. Details of the statistical analyses are generally short in the Methods. In contrast, the authors seem to pack quite a bit of the descriptions in the supplemental Table titles and legends (e.g. Table S14 and S15). I think it makes more sense to centrally describe the statistical analyses in the Methods, although perhaps the authors are limited by length constraints. Regardless, I do think the authors can do a bit more explaining of the interpretation of various statistics. For example, in qpWave, what does rank 1 imply, what does rank 2 imply? In a particular formulation of f4, what does positive values imply, what does negative values imply?

8. Line 204: Mongolia_East_N is labeled Monglia_N_East in Figure 2.

9. Line 257: “showing”

10. Line 393: I believe it is Table S5, not Table S8.

Referee #2 (Remarks to the Author):

This is one of the most important papers on the human population history of Asia to appear in recent years. 191 ancient and 343 modern genomes are analysed, with the oldest going back as much as 8000 years. The results are overwhelming in the detail, although many of the conclusions that are drawn equate closely with current understanding of population history in the general region. The paper does not produce any new controversies, but it does reinforce a number of observations made previously, for instance about the importance of the Yangtze and Yellow river valleys in China, links between China and Southeast Asia, and relationships of the Jomon of Japan.

Since I am not a geneticist I cannot comment on the accuracy of the genomic calculations and comparisons presented. The text is very dense, and it is unlikely that anyone unfamiliar with the genomics of East Asian populations and their history and archaeology would be able to understand it easily. Paragraphs are very long, and embedded with many site and people names, statistical terms, and genetic categories. As I read it, I found myself wishing for more aids to understanding.

One would be a list of all the abbreviations used in the text. For instance, I could discover that WSHG means West Siberian hunter-gatherer, but I still do not know what SG means.

Another would be to reduce the wonderful and highly informative figure S2 to something that can be more easily read. I needed a magnifying glass, since if this figure is magnified sufficiently to read, it becomes too large on a screen to navigate easily. Would it be possible to produce a simplified second version in which related populations are grouped in some way into many fewer columns, for instance, living Tibetans, living Europeans, Bronze Age Central Asians, prehistoric Southeast Asians, and so forth? Do we need to have every single population labelled and put in a separate column? Yes, maybe we do in the main chart, but information of this density could hopefully be summarised for presentation purposes.

Another very useful addition for this table would be some idea of what the different ancestry component colours actually mean in terms of real populations. I can see that blue is West Eurasian, as in Anatolian Neolithic and derived populations, whereas orange presumably has something to do with the Neolithic in East Asia. Red is obviously Neolithic and Bronze Age steppes, green is Iran Neolithic, yellow is sub-Saharan African, and so forth. There is a discussion about each K number in the supplementary information, but it is hard to relate to the figure when the information is presented on this huge scale.

I noticed a few typos and I expect these will be picked up later in proof reading. However, line 128 appears to misspell Yukaghiric and Kamchatkan, and the so-called ‘Altaic’ listed here, together with Koreanic and Japonic, is referred to as ‘Transeurasian’ by the only author on the list that I perceive to be a linguist, namely Martine Robbeets. In the supplementary section around line 183 there are many (wrongly) lower case genus names.

I think my overall reaction to this paper is that it is a statement of the enormous power of modern genomic research into both ancient and modern populations. But presenting the information to people who are not geneticists or statisticians is a daunting task. Some of the paragraphs in this paper are almost 2 pages long, and the sheer density of information is very difficult to absorb. The paper has no discussion or conclusion, and indeed no subheadings at all. The abstract is a clear statement, however.

In terms of the population sampling for ancient DNA, I did notice the complete absence of samples from Southern China (excluding Taiwan). The potential importance of the Yangzi River population is mentioned in the article from time to time, yet there is no ancient DNA from there. This is mentioned in the very last sentence, which also contains the strange statement ‘… in particular to understand if dispersals of people in Southeast Asia do or do not correlate to ancient movements of people.’ But what is it is a dispersal, if it does not involve a movement?

In this regard, I notice that one of the authors on this paper is Hirofumi Matsumura, who in my view has presented clear craniometric information to suggest that Neolithic populations in SE Asia were part of a major movement, just as they were in Europe. I know that a paper such as this cannot easily discuss findings made by other disciplines, but it would be good to have some author opinions apart from those of the geneticists alone.

I strongly support publication, but it would be good if the article could be more user-friendly for non-geneticåist readerships.

Referee #3 (Remarks to the Author):

The manuscript of Wang, Yeh, Popov, Zhang et al. brings to the published record ancient data from East Asia, a region where the amount of available ancient DNA data is very limited. Yet such data are crucial for our understanding of population demography of this and adjacent areas. Several sub-regions were analyzed: 52 samples come from Taiwan, 89 from Mongolia, 20 from a Neolithic site in China (Yellow River region) and 30 from other regions throughout East Asia. The authors also present genotype data from 383 present day individuals for comparison purposes. This data will be a major addition to the data already produced from populations in East Asia and this dataset should be published.

The analyses (laboratory and computational) are performed very well and the manuscript was easy to read and carefully written. The data will be likely used in the future by a number of researchers and the conclusions of this manuscript will be of interest to a large number of archaeologists and anthropologists working in different parts of East Asia.

While the data from East Asia are limited, it should be noted that some of the populations, regions and even sites have been analyzed for ancient DNA before (e.g. Jomon individuals, individuals from Amur River Basin, Neolithic individuals from Devil's Gate, individuals from Mongolia). This does not invalidate the importance of the data as such a vast, anthropologically complex region requires very dense sampling. The scarcity of the data in the region makes many conclusions speculative, especially when based on qualitative descriptions and linked to linguistic theories (see below) but I particularly enjoyed the contextualized results (especially the case of the Afanasievo boy).

The diverse nature of the samples analyzed in this study is however one of the major weaknesses of the manuscript. The connection between different regions and periods that their dataset covers is very loose. No hypothesis or results seem to justify analyzing these particular samples together, except that they all are located in some region of East Asia. This negatively influences many aspects of the manuscript and it is difficult to easily fix. This also heavily undervalues the actual results of the study and their understanding for the readers. My suggestion would be to either rewrite the manuscript considerably (addressing the issues described in detail below) or to split this manuscript to two or more self-contained studies. For the benefit of the readers, I would prefer the latter solution but I understand that the authors may have their reasons to describe the data together in this manner. I have also some additional concerns regarding clustering of the individuals to groups, contamination and presentation of the results.

Major points

Temporal and geographical diversity of the samples negatively affects most parts of the manuscript:
Title: Given the ambiguity of the some results and unequal sampling over vast regions and chronological periods, it is not appropriate to title the manuscript as being able to comment on the formation of the populations in the whole region. Yet I see that the lack of a result connecting the samples makes finding a better title difficult.
Abstract: The abstract, as the manuscript itself, contains rather disconnected statements. Some of these statements are very vague and therefore hard to grasp or disapprove. Specifically connections to the spread and branching of language groups are quite weak and are not sufficiently discussed in the manuscript itself: there is no description of different theories regarding the linguistic variability, no hypotheses are formed and there is no discussion how different aspects of the analysis could influence them. The authors might be right that some qualitative aspects of their analysis are not inconsistent with some of the linguistic hypotheses mentioned. However without providing details of the linguistic discussion on the topic, alternative scenarios or discussion on the role of their results on the ability to reject these scenarios, it is not possible to make these claims, however carefully they are written. Especially not in the abstract where there is a little space for nuance. This is crucial as such statements might be used to support claims in other fields and scholars not familiar with genetics might interpret these narrative connections as scientific proofs (since they are so prominent in the abstract). This could be quite likely due to a lack of space in the manuscript because of the large number of different topics that need to be covered (as the context variability is so high).
If the authors want to make these claims, they should provide at least hypothesis descriptions and discussion on alternative scenarios in the manuscript and add a section regarding linguistics and its connection to their results to the supplementary material.
Introduction: I find the beginning of the manuscript difficult for the reader to follow: the first paragraph gives some generic information about East Asia and then it continues with description of the data acquired with the emphasis on their counts. Then there are some methodological details

that should be (and actually sufficiently are, no need to reiterate them in this section) placed in the methods, especially since no new methods have been introduced in this study. Then there are results justifying the grouping of individuals and the manuscript further continues with more results with embedded interpretations that are rather divided by regions, each starting with a mini-introduction of the sub-region of their own (see lines 348, 381, 401 especially). What is seriously lacking is the comprehensive (even if short) description of the state of the art, the results of previous studies of ancient and modern genetic variability that comment on the forming of populations in these regions. Therefore the manuscript does not provide any hypotheses in the beginning or identify gaps in the knowledge that need to be filled. It is then hard for the reader to follow the flow of the results and the authors' reasoning.
Disjoint parts: There are parts in the manuscript that are completely disconnected to others (see paragraphs starting with lines 348, 381, 401, 443). Those could be deleted or placed elsewhere without any damage to the article. Again, as the readers go through the article, this easily creates confusion. The authors should describe in the beginning what they intend to tell the readers and use subheadings.

Contamination issues:
I do not understand the use of "questionable" samples. The authors made a very detailed and clear Online Table 1 with information about the samples and they mark samples with problematic date, contamination estimate or other issue as "questionable". For contamination filtering, they used the upper limit of contamMix estimation (of the amount of endogenous sequences) and they should definitely switch to median or to be conservative (as many other researchers) to the lower bound. What I find even more surprising is that some samples that were marked as questionable continue to be used in various further analysis. For example in Figure 3, several of the samples that were marked as questionable are displayed (among others, sample I13958 with contamination estimated to 16.9-40.8% as can be seen in Online Table 1). This should be at least mentioned in the caption of the figure and the contamination estimate (lower bound or the median) or the reason for doubting this sample should be specified. In general, the questionable samples should be excluded from all the analysis and if an exception is necessary, the reader should be made aware. If problematic samples are further used alongside samples without any issues, I do not see a point of any contamination analysis or dating evaluation. The readers expect that after the contamination analysis was performed, problematic samples are excluded and the data are hence believable. If I understand correctly, the problematic samples were also included in the counts of samples throughout the manuscript and abstract and in the analysis where groups were required (e.g. qpAdm analysis), potentially putting all those in question. Out of 191 ancient samples analyzed, 52 are marked as "questionable" and 9 even as "questinable_critical": this puts 31.9% of their samples into serious doubt. Similarly, the samples with high contamination estimates were used to support genetic contextualization of other samples (that is especially problematic for groups with low number of samples: e.g. Mongolia_N_East sample I7031 with a contamination estimated to 10.9-21% as can be seen in Online Table 1). This is circular and problematic because there could have been a cross-contamination in the lab, the site or at the storage facility (not unlikely and understandable given the number of samples, as evidenced for example by misdating of some samples due to storage issues that is very well described and discussed in Online Table 1). Thus the contaminated samples have to be excluded from this analysis as well. If the authors want to use these samples anywhere e.g. because there are no other samples available from this particular site, the readers need to be warned of the contamination issues repeatedly in the main text and the implications for the results need to be thoroughly discussed. Contamination estimation for Wuzhuangguoliang female samples has not been performed at all even though there are methods to do so (e.g. based on LD or PMD patterns).

Cluster assignment and number of samples per group
The number of samples in the study is large, however given the size of the region and the chronological variability, this means that at times, the authors used only a few samples to represent whole populations at various time points. That, by itself, is standard when there is no

other data available. However, this might result in an increased number of alternative interpretations of their analysis and the authors do not discuss this in sufficient detail. This is even exaggerated by the use of low coverage hence pseudohaploid data that further limit the amount of the data available for the analysis. While they mention the scarcity of their data regarding the interpretation of qpGraph results, it is not mentioned elsewhere. For example, the Mongolian clusters and the interpretation of the different waves of the Yamnaya-related admixture is dependent on the grouping of populations composed of a few individuals. While I do trust their qpAdm analysis (the Afanasievo being in the outgroups when Sintashta is the source indeed points to some later gene flow), I am hesitant to agree that composing the groups based on genetic similarity and then treating them as populations even though they rather overlap chronologically is a good course of action. This analysis (Figure 3) should instead be performed on an individual basis. A similar point is actually argued by the authors themselves in another part of the manuscript: in the case of Heishui_Mohe samples (see Supplementary Section 2), they refuse to make claims regarding the West Euroasian admixture in this population because the two samples they have from this period fall at different locations in PCA plot. Yet, if they would sample only one of these individuals, they would be confident (as the Mongolian example above suggests) to make completely opposite claims depending on which individual was sequenced. That demonstrates the dangers of overinterpreting results from a very low number of individuals and this needs to be addressed in the manuscript whenever this is the case. It should be also noted that the authors generally disregard a hypothesis that individuals that do show different genetic patterns (on PCA for example) still could be from the same group but that such a group is in a process of admixture (they form groups based on genetic similarity). Additionally, they repeatedly favour hypotheses of demographic events as migrations (“waves”) over continuous gene flow that are actually quite likely.

Presentation of the results
The graphical material added to the manuscript has some issues. That is especially difficult for evaluation of the clusters the authors defined. This is a major point because many other analysis and conclusions are based on this grouping.
Figure 1: the colours used in this map should be of more variety than red, orange and brown, it would then be easier to distinguish the points, also the authors could add some additional graphical elements to this map to better illustrate what is known and what they try to uncover (e.g. arrows).
Figure 2: this is a very crucial figure for the text: among others, it should clearly demonstrate the grouping of the individuals. This is however very difficult to see at the moment. It would help if the authors added another figure with simplified legend and visually marked the individuals belonging to their clusters (or make other adjustments to the figure). I also suggest that more PCAs containing only samples from similar periods would be very insightful (instead of projecting all ancient samples to so many modern samples).
Figure 3: This figure has some issues I mentioned elsewhere (contaminated samples and it should be done on the individual basis), I also think the ancestry proportions along the time axis are a bit difficult to compare (their size depends on the date) and it would be better to use for example the other axis to this purpose. Graphically, it seems to me of low quality (the figure resolution).
Figure 4: I fail to see an apparent cline on this figure as stated in the text (line 432). For example, the most upper point is actually with more orange than those below; also the differences observable on the pie charts seem to be small. But otherwise it is an informative figure and the authors should consider if they do not want to add more figures of this type to illustrate other results of qpAdm that are otherwise not so easy to see.
Figure 5: I must say I quite agree with the authors that “this admixture graph is an oversimplification” (line 489) and while I think it is a good addition to the paper, the authors might think about if it needs to discuss it in such detail in the text.
Figure 6: This figure lacks any labels on the x axis and I have some issues with the median of the first group, it rather looks like that the Fst is only a bit higher than in the Neolithic but one population is quite different. The caption should also mention in which table are the values underlying this figure placed.

Figure S1: This figure of Fst results is also crucial for the clustering and the clusters should be marked very clearly. I must say I am a bit worried that the authors use this figure for interpretation of some results while they themselves notice discrepancies for other populations (Papuans, as mentioned in the SI).
Figure S2: This admixture result is very hard to examine. Since this is again crucial for the group assignments, important findings should be presented in subfigures made from this main figure. However, I very much like the textual description of differences between analysis under different K in the SI.

Other points

Ethical concerns
I have concerns regarding ethics and sampling of the ancient individuals. While there is no formal requirement for the authors to contact local communities and get their approval for the sampling of the individuals discovered at archaeological excavations, it is certainly highly desirable and especially so if the ancient DNA studies (e.g. from adjacent regions and from some of the same authors) can be considered problematic in this regard. In my opinion, a publication aspiring to be at a high-level journal has to address this in the supplementary material and in the ethical statement (required by Nature) to make sure there are no ethical issues with the material (both from the research community and from the public). It should be noted however that the authors got proper ethical approvals for contemporary samples and with the ancient samples they did not breach any standard and that authors of many other publications do not go through the process of getting approval from the communities etiher. But it would result in avoiding any dangers of bringing further distrust to the field of human population genomics.

Archaeological information
Given the high number of regions and periods covered, the archaeological supplementary material is rather long, containing information essential to the manuscript. I especially like the care taken to prepare very informative Online Tables (I applaud the Online Tables 1 and 2 especially). However, the quality level of SI is inconsistent: some archaeological sites are described with a lot of attention to detail, some are barely mentioned (e.g. compare the information about 18 individuals from Boisman-2 site and one individual from Nevelsk 2 site). The information about the archaeological context of the ancient individuals should be added to all sites (any graphical material would also be welcome). Also there is quite some lack of references (e.g. no reference at all for the information provided for Slab Grave culture graves, Mongol graves or one reference for Xiongnu burials). I understand that it is difficult to provide information for all the sites and individuals because of the number of samples and the diverse nature of the region and periods. But since this is essential for the grouping of the individuals (a basis of most of their analysis), this cannot be taken lightly. The authors should provide clear references for all the sites, in case some sites are not published, then the authors should provide enough information for other scholars to evaluate them (archaeological documentation of the finds and the site itself). Otherwise, the samples can be used incorrectly in future analysis of researchers using this dataset and it might be also hard to evaluate the current analysis.
I also have some reservations about the anthropometric descriptions used for interpretations of similarities of populations in this part of SI, they are rather dated (e.g. Slab Grave culture: "anthropological typing of this culture suggests they are 'Mongoloid' or similarly in the description Boisman site).

Different pipelines
Samples have been analyzed with 3 different pipelines. Already a part of the data has been processed differently in the lab and I wonder why the authors additionally bias the analysis by using different pipelines. It has been shown previously how even a small amount of bias can severely impact genetic inferences (Günther & Nettelblad 2019 PLOS Genetics) and here they treat the data with different versions of the same tool or even completely different tools. While the

authors claim that the difference in results between the pipelines is not large and cite the analysis of Fernandez et al. that showed that some of the conclusions in that paper are not biased while using different pipelines, I strongly disagree. The f4 statistics analysis of Fernandez et al. between the pipelines can hardly be generalized to other studies and samples and even in that study, it is hard to generalize for all the results. Furthermore, the Wuzhuangguoliang samples were treated with a third, completely different pipeline that has not been compared to the other pipelines at all.. Reanalysis of the data with the same pipeline should not be a major issue for the authors: the pipeline is scripted and the scripts are available and the data are quite low coverage and hence rather small. Therefore, all samples have to be analyzed as similarly as possible.

Genetic results correlate with linguistic and geographic patterns
There are mentions of correlation between genetic, linguistic and geographic patterns in the region in the abstract, the manuscript and in the SI. The authors use only “qualitative” (as they state) assessments to make this case (mainly a PCA plot and Fst distance tree). But correlation cannot be mentioned when no statistical test was performed. While the authors might be right, it is very difficult to assess, given the figures provided: the plots contain too many groups and are not informative about linguistic or geographical assignments. The authors have to support these claims by one of many tests that would allow them to compare geographical, linguistic and genetic diversity. I would suggest using a simple Mantel test on the respective distances.

Incorporation of previously published results
Previous studies on the topic are mostly mentioned only when their data were analyzed together with the presented dataset. Or together with the results of the authors when they reach similar or the same conclusion without explicitly describing what new insight is brought even when there is one (e.g. line 432). The authors should be more precise in distinguishing their (indisputably important) findings from previous work. Additionally, some sections are missing mentions of previously published work. For example in lines 443-453 the authors discuss the formation of the modern Japanese population without mentioning any previous study that exists on the topic (there are many), not even the studies from which they use the data from (and are cited in the manuscript elsewhere).

Minor points and typos

- Supplementary section 3, 2nd and 3rd paragraphs: several typos (“implanted” instead of “implemented”; “to further differentiated” instead of “to further differentiate”; “constrained the model used MCMC”)
- Line 263: derived
- Data merging - it might be renamed to something like “reference panel preparation” (merging usually suggests merging of reads or data per sample).
- The authors sometimes use their codes instead of the labels they decided to use, e.g. in lines 239 and 248 where the reader needs to consult supplementary tables to identify the individuals at Figure 3 that are being referenced.
- Table S1: the caption should mention these are the modern individuals
- Line 227: qpAdm mentioned for the first time, even though stated that it was mentioned “again”
- Line 249: individual instead of individuals
- Line 257: showing instead of showning
- The authors treat all Han as one group while there is very well described variability between Southern and Northern Han (among many others, see Liu et al. 2018 Cell)
- Lines 256-260: all in one sentence, it is a bit hard to read
- Line 321: the authors use term genetic continuity very loosely: if they want to use it they should statistically test for continuity. Similar Y and mtDNA haplogroups and approximately the same spot on the PCA is highly insufficient to claim this.
- Bwa versions used are not stated
- There is a mistake in the Online Table 1 or 2 for the individual I6365. In one of them, the contamination on X chromosome is estimated as 20% and this is not mentioned in the other table.

Probably the contamination method was used for a female while this works only on males and the result was not deleted after sexing.

Referee #4 (Remarks to the Author):

1. Title: “The Genomic Formation of Human Populations in East Asia”. This title does not accurately reflect the content of the manuscript, which aims to reconstruct the demographic history of East Asian populations through the lens of the genome, rather than to trace the formation of the genomes themselves.
The title also does not accurately reflect the scope of the paper, which addresses particular aspects of the genetic formation of particular Asian populations; much of the existing genetic variation in East Asia is not treated at all. A large part of the paper (lines 194-315) deals with the eastward expansion of people associated with steppe cultures such as the Yamanya/Afanasievo and Sintastha/Andronovo. This aspect of the manuscript, although very interesting, is not central to the genetic history of most East Asians.
2. Concerning the abstract in general: it is difficult to discern a unifying thread connecting the many disparate pieces of evidence presented throughout the text. It would be helpful if the hypotheses motivating the study were clearly stated at the outset.
3. Line 94: “We document how 6000-3600 BCE people of Mongolia and the Amur River Basin were from populations that expanded over Northeast Asia, likely dispersing the ancestors of Mongolic and Tungusic languages“.
The significance of this statement is not clear: does the observation that ancient people from Mongolia and North China were spreading over Northeast Asia violate a commonly-held expectation? Second, it would be appropriate to state why the authors believe that these historical genomes represented the carriers of the two language groups. Perhaps a short statement motivating this inference would be helpful, along the lines of "...as inferred from the chronological/geographical appearance/distribution of..."? The same for the appearance of genetic signals associated with the Afanasievo culture in Mongolia “... plausibly acting as the source of the early- splitting Tocharian branch of Indo-European languages”. While the authors have clearly established that genetic signals from further west appear as far east as Mongolia, there are several other conceivable explanations for the spread of Trocharian into this region, and the origin of the language in that region could be considerably older—or younger. The claim is based on the assumptions that people associated with the Yamnaya and Afanasievov cultures spoke Indo-European languages, language transfer is primarily associated with the spread of groups of people, and an unbroken line of gene and language transfer has taken place from the European steppes to the eastern parts of Eurasia. While such tantalizingly simple models are not excluded by the present evidence, neither are they the sole obvious inferences in light of the current analyses. These claims are more appropriately understood as hypotheses requiring further testing.
The same is true for: “Analyzing 20 Yellow River Basin farmers dating to ~3000 BCE, we document a population that was a plausible vector for the
spread of Sino-Tibetan languages both to the Tibetan Plateau and to the central plain ...” I would suggest reformulating this as one possible hypothesis while also addressing other models that could explain the data. A brief mention of the limitations of this analysis would be appropriate here.
4. I am particularly surprised by the statement: “Yangtze Valley first farmers who likely spread Austronesian, Tai-Kadai and Austroasiatic languages across Southeast and South Asia. The basis for this claim (genomic or other evidence? Previously published work?) is not clear, and it is not an obvious inference from the data presented in the manuscript.
5. I don´t think that the following sentence will be clear to all readers because it uses nomenclature particular to archaeological specialists: “In a time transect of 89 Mongolians, we reveal how Yamnaya steppe pastoralist spread from the west by 3300-2900 BCE in association with the Afanasievo culture”. The key findings seem to be that the graves from “Shatar Chuluu kurgan” in Mongolia belong culturally to Afanasievo, and the genomes found there are identical to those in the core Afanasievo region further west. This is exciting, but I think that a broader

readership will not be familiar with these particular archaeological terms. I suggest describing this finding in more accessible terms that better convey its significance, at least in the abstract.
6. Given the geographic and chronological gap, is it appropriate to speak of the appearance of the Yamnaya Steppe pastoralists in Mongolia? How is the phylogenetic “Yamnaya” signal related to the actual processes under study?
I have a similar question concerning the following statement: “The second spread of Yamnaya-derived ancestry came via groups that harbored about a third of their ancestry from European farmers, which nearly completely displaced unmixed Yamnaya-related lineages in Mongolia in the second millennium BCE, but did not replace Afanasievo lineages”. It is interesting that the authors can detect several waves of gene flow from populations originating in Western Eurasia: the fact that these newcomers bear signs of deeper ancestry from southern Russia— or even from prehistoric populations in Europe— can help to better understand later population developments . But they are not the message itself. What could it mean that signals from European farmers and the western steppe are appearing in Mongolia? The fact that human populations are mobile and connected through a band of shared ancestry is interesting, but unenlightening without an understanding of the historical process that generated it.
7. I find the following statement too anecdotal to appear in the abstract: “we also document a boy buried in an Afanasievo barrow with ancestry entirely from local Mongolian hunter-gatherers, representing a unique case of someone of entirely non-Yamnaya ancestry interred in this way“ The fact that an individual from a local population was buried in a kurgan from groups of immigrants is not entirely surprising.
8. “Groups in Northwest China, Nepal, and Siberia deviate towards West Eurasians in the PCA ..., reflecting multiple episodes of West Eurasian- related admixture “ The fact that populations further west show more west Eurasian signals may simply reflect the natural spatial variation: how do the authors distinguish between an “admixture” event and other explanations for spatial variation? Are these "western signals" absent in earlier populations of the same area?
The same is true for “The other seven Neolithic hunter-gatherers from northern Mongolia ... can be modeled as having 5.4% ± 1.1% ancestry from a source related to previously reported West Siberian Hunter-gatherers” The authors write that they form “part of an east-west Neolithic admixture cline in Eurasia with increasing proximity to West Eurasians in groups further west”. It would be helpful if the authors provided their criteria for distinguishing “ancestry complexity” from clinal variation under a null model of genetic variation.
9. Lines 182 to 189, “The “Amur Basin Cluster correlates geographically with .. .. southern parts of China speaking Austroasiatic, Tai-Kadai and Austronesian languages”: The authors use terms like “correlates”, “is most strongly represented”, and “is maximized”. Would it be possible to attach some numbers to these terms so that the reader can see how significant these associations are? The same comment applies to the following: “one falls closer to ancient individuals from the Amur Basin Cluster (‘East’ based on their geography), and the second clusters toward ancient individuals of the Afanasievo culture (‘West’), while a few individuals take intermediate positions between the two”: please clarify how many genomes and individuals are involved in each of these clustering analyses.
10. The paragraph starting in line 317 contains a lot of fascinating results. I think it would be helpful if the essential questions and hypotheses discussed here were presented right at the beginning. As it stands, it seems a bit disconnected from the section of the paper immediately preceding it.
11. Regarding the sentences that start with line 340: "Some present-day populations...": the referent of these statements is not clear; perhaps the authors can clarify and re-phrase slightly?
12. Regarding line 368: “We estimate that the mixture occurred 60-80 generations ago ...” In my opinion, the single admixture event model on which this estimate is based is unrealistic, which has important consequences for the putative admixture event and further arguments following from this estimate. Can the authors test other models and compare them to their “model of a single pulse of admixture“?
13. Regarding the “formation” of the Japanese (line 443): Can the authors please explain the actual significance of the admixture proportions estimated herein for Japanese demographic history? As the authors state, it is a signal from the deeper past, "that it is from an ancestral

population related to those that contributed in large proportion to Han Chinese as well as to Korea". What can be learned from this, except that the Japanese had other East Asian ancestors? What was the initial hypothesis? Is it somehow confirmed or disproved by the admixture estimates?
14. Line 455: Maybe the authors can explain right at the outset of this paragraph what they used qpGraph for? (I infer that it is to “identify a parsimonious working model for the deep history of key lineages“). I assume the authors use “lineage history” to infer population history, but it would be helpful if the authors stated what features of human population history they expect to be illuminated by reconstructing lineage histories. The genetic composition of human populations is uninteresting to most readers unless it can be used to elucidate real historical processes.

Author Rebuttals to Initial Comments:

We thank the four anonymous referees for the constructive feedback and suggestions. We have taken all comments and suggestions on board. Please find our responses below.

Referee #1 (Remarks to the Author):

In Wang et al., the authors generated and analyzed an exciting aDNA dataset from samples distributed throughout Eastern Eurasia. The sample size of the aDNA dataset is decent, sampled from regions of the world that are currently still understudied. Consistent with a long-line of pioneering work from these teams of senior investigators, the methods applied in this manuscript are established and have been shown to provide insights to historical migration events in the past.

A lot of analyses and results were presented, and the authors adequately guided the readers through the narrative to arrive at the conclusions. In general I have no strong concerns of the methodology and interpretation. My main comments are generally related to presentation and clarifications.

1. In general, the narrative is structured as a disconnected list of analyses from different parts of Eurasia. First starting with Eastern Eurasia/Mongolia, the narrative then pivot to the Amur Basin, then Tibet, Taiwan, China, and then Japan. The bulk of the analysis and insights seem to be from Eastern Eurasia/Mongolia and in China, with the remaining regions loosely connected. I see value in generating and making available aDNA data from all these different parts of Eastern Eurasia. But if there is no major insight from some of these regions (e.g. line 381-399, analysis of the Taiwan individual is really a single analysis of systematic f4 statistics; line 443-453 is a single analysis of admixture f3 using Jomon), perhaps the narrative should set up the expectation early on that this is a general survey of the new data generated, with emphasis on a couple of regions where new insights are produced.

We agree that this study has important insights about Mongolian and Chinese population history, but in fact, the newly reported data from additional regions and notably Taiwan and Japan and the Amur River Basin also contributes critically to the findings. Arguably the most important part of our manuscript is the global picture we present of the structure of human variation in East Asia, for which the Mongolian and Chinese data are a part, but which relies even more centrally on the data from Japan, Taiwan, and the Amur River Basin, which are included in our global analyses of broad population structure, and are included in the admixture graph model that now begins the manuscript. We also include focused paragraphs specifically on these regions.

In response to the referee’s comment about the disconnecting ordering of the originally submitted manuscript—a critique with which we agree—we have completely revised our manuscript as follows.

(A) Introduction: We start the manuscript by reviewing the results of previous studies and highlighting open debates related to population history in East Asia in each region in turn. We then outline our paper’s structure and a list of the questions we address. Throughout the paper, we have added section headers to help the reader navigate through the manuscript and understand how they are related.

(B) Data set: We describe the data and this section concludes with an overview of population structure (PCA and ADMIXTURE).

(C)Results Section 1 – Deep time: We re-ordered the manuscript to start with a global admixture graph model.

(D) Results Section 2 – Three Holocene Population Expansions in East Asia

- *The Amur River Basin Expansion*
- *The Upper and Middle Yellow River Farming Expansion (Han and Tibetan population structure arise from this and are discussed within this context)*
- *The Southern Ancestry Expansion (we document that this must have occurred, and hypothesize that it may be associated to Yangtze River farmers while highlighting that a test of this hypothesis needs to await data from early Yangtze River farmers)*

(E) Results Section 3 – Interactions Between West and East Eurasians. This section describes mixtures between West and East Eurasians at the fringe of East Asia, with a focus on Mongolia and Xinjiang

(F) Future Prospects (short conclusion)

2. To me, the most important figure that summarized the main conclusions of this manuscript is Figure 3 – I kept referring between the text and Figure 3. If this is a survey of the aDNA from multiple regions in Eastern Eurasia, the author should consider putting together an overall or multipanel figure for each of the regions they investigated (where the current Figure 3 would be a subpanel). Such a figure, though undoubtedly will be updated as new data come out in the future, will be referred to repeatedly.

This is done in our revision (Figure 5) and we agree it is an improvement.

3. From the abstract, the authors suggested that a first wave of Yamnaya ancestry (associated with Afanasievo culture) was nearly completely displaced by a second wave (represented by the Sintashta ancestry), with the exception of samples in Western China (represented by Xinjiang Shirenzigou data). This set up an expectation that present day individuals in this general region carries largely Sintashta ancestry (again, with the exception of Xinjiang). However, interpretation from Figure 3 (and the text starting from line 228 or so) is that most individuals sampled in this region during the early bronze age and late bronze age were actually two-way admixed. The Yamnaya ancestries, whether through Afanasievo or Sintashta, are not as prevalent as they are in modern Europeans, and arguably are more exceptions than the rules. Perhaps in Figure 3 the authors can estimate ancestral proportions of modern populations in these regions, where data is more abundant and geographically diverse, to show whether Yamnaya-related ancestry is actually prevalent today. In fact, it would be nice for Figure 3 to also include Han Chinese ancestry (a fifth component) in the time transect, which was only briefly discussed in the text (line 278-285).

We certainly did not mean to imply that Mongolians today harbor largely western Steppe pastoralist ancestry. We have revised the relevant sentence in the abstract to highlight the predominant impact of eastern populations on Mongolian group: "Yamnaya Steppe pastoralist ancestry which was likely a vector for spreading late-proto-Indo-European languages arrived after ~3000 BCE in western Mongolia but in contrast to Europe where it persisted, it was displaced by previously established lineages"

Following the referee's suggestion and the suggestions of the other referees and other manuscripts that have become available since the original submission of our paper, we have expanded our admixture proportion analysis in three ways: (1) by extending the Mongolian time transect to more recent times by allowing Han Chinese related admixture as a fifth ancestry source, (2) by adding additional relevant reference populations to the qpAdm analysis allowing more refined inferences about ancestry sources, and (3) by showing maps of ancestry proportions in different time slices including the present (Figure 3).

4. Line 238-253: the male Chalcolithic Afanasievo child who had no Yamnaya ancestry is touted as a striking finding. Although I find it interesting, and understand that it is the first of its case, I failed to appreciate the archaeological significance of this finding. I assume cultural diffusion is not so exceedingly rare. The authors should discuss further why this is surprising (and thus significant), or perhaps this finding would be better located in the Supplement.

This is an important observation in light of the previous literature—and our highlighting of this observation was explicitly identified by Referee #3 as a particularly strong feature of our study—but we had not explained the significance adequately in our original submission. The significance of this observation is that it challenges previous more simplistic narratives about the nature of the Yamnaya expansion. We clarify this in the revised manuscript, but have removed discussion of this finding from the abstract where it was previously featured.

5. Line 247: the Chalcolithic Afanasievo child showed genetic continuity with Ulgii individuals. How exactly is genetic continuity demonstrated?

The evidence for continuity is our demonstration that both the Ulgii individual and the Chalcolithic Afanasievo child were mixtures of two sources that were already established in Mongolia (what we call Mongolia_N_East and WSHG) prior to the arrival of Yamnaya-associated ancestry. The admixture proportions from these two sources are within 6% of each other highlighting the genetic similarity and we now mention this in the revised text.

6. Beginning at line 401, the narrative turned towards the second major insight from this manuscript. I would like the authors to clarify what is the model of the origin of Han Chinese in their mind. From the abstract it seems to imply the source population originated from upper and middle Yellow River Basin and then moved West/South into the Tibetan plateau, as well as more East to the Central Plain. Can the authors explained the logic behind the inferred outward direction of expansion from the upper and middle Yellow River Basin, rather than a uni-direction expansion from East coast towards the West and South? Also, are these models consistent with that suggested in Li et al. MBE 2019 (10.1093/molbev/msz072) where river valleys shaped the genetic landscape?

Our north-to-south finding is really about Sino-Tibetan languages in the Central Plain and not a specific inference about the expansion of the Han ethnicity itself, but at the same time there are circumstantial hints of a north-to-south Han expansion. We clarify this in the following revised paragraph:

"Our results support the 'northern origins hypothesis' for both Sino-Tibetan languages and for the primary ancestry of Han Chinese, and provide evidence against the 'southern origins' theory of an early Holocene pre-farming spread from the Tibetan-Yi corridor. In the northern origins hypothesis, expansions of farmers from the Upper and Middle Yellow River Basin spread both languages and ancestry to the Tibetan Plateau and the China Central plain[15,53], a model that is supported by the specific link we detect between Sino-Tibetan speakers today and Upper and Middle Yellow River farmers, and is further supported by the Y chromosome evidence of a shared common haplogroup Oα-F5 between Han and Tibetans that derives from a single male ancestor who lived ~5,800 years ago[54]. The simplest explanation for the increasing proportion of Han Chinese ancestry related to ancient Liangdao among southern Han groups is thus that early Han came into contact with southern groups and mixed with them as they spread to southern China as recorded in historical literature[55]. However, not all of the southern-related ancestry in the Han can plausibly derive from events in the Han period: populations from southern China are genetically closer to Late Neolithic Yellow River farmers than to earlier Middle Neolithic ones[26] and there was a significant increase of rice farming in the middle and lower Yellow River regions between the Middle and Late Neolithic periods, documenting how this process of north-south genetic and cultural exchange had earlier antecedents."

Our model is consistent with the results from mitochondrial DNA in Li et al (Yu-Chun Li, et al., MBE 2019), and we cite this paper our revision. First, Li et al inferred the fastest population growth at sites associated with agriculture along river valleys including the Upper and Middle Yellow River Basin. Second, Li et al observed close genetic affinities between populations from Yellow River valley and other rivers (e.g., Haihe and Songliao rivers), consistent with expansion of human populations and millet agriculture from the Yellow River to surrounding regions. Third, Li et al detected a genetic difference between southern and northern Han populations, which is also evident in our genome-wide data from modern groups.

7. Line 426-432: The authors implied that varying proportions of Neolithic Wuzhuangguoliang ancestry can explain the North-South differentiation observed in modern Han Chinese. However, this is not apparent from Figure 4. Can the authors demonstrate the degree to which Wuzhuangguoliang explains the N-S differentiation?

We can model almost all present-day Han Chinese with 77-93% of ancestry related to Yellow River Basin groups like the Wuzhuangguoliang individuals, and the north/south cline is evident in the figure. We make this clearer in the revision.

8. I am interested in the finding of West Eurasian ancestry in Han_N_China sample (Line 423-425). We and others have reported the potential admixture signals in modern Han Chinese (Liu et al. Cell 2018 [10.1016/j.cell.2018.08.016]; Chiang et al. MBE 2018 [10.1093/molbev/msy170]). Chiang et al. estimated similar admixture proportions and dates. Is there more information on the geographical origin of the Han_N_China samples? Both Liu et al. and Chiang et al. are much more recent and updated investigations of population structure and admixture history of Han Chinese that should be cited along with the two current citations from over a decade ago (Chen et al. 2009 and Xu et al. 2009).

In our revised manuscript we cite the two mentioned papers: Liu et al. Cell 2018 and Chiang et al. MBE 2018. With the methods available to us we have been unable to trace the geographical origin of West Eurasian-related ancestry in Han_N_China samples since the proportion of this ancestry is low (with so little ancestry, we cannot distinguish several alternative models with f-statistics). Our analysis does, however, document multiple sources of West Eurasian-related ancestry in East Asia in the Middle to Late Bronze Age and Iron Age, so these could be plausible sources for the ancestry

9. Line 443-453: In the analysis using the Jomon sample, why not use Wuzhuangguoliang sample? Particularly since the narrative left an open question whether a mainland ancestry in modern Japanese is due to an ancient ancestry that contributed significantly to both Han Chinese and Koreans (Line 451-453).

We did not use Wuzhuangguoliang as a source for modeling admixture in present-day Japanese due to the limited number of SNPs in the Wuzhuangguoliang data which compromises statistical power. However, in our revised paper we have added an analysis that uses as sources the Neolithic to Iron Age West Liao River populations in northern China from Ning et al. Nature Communications. We use this along which we shown as sources together with Jomon to model the formation of Japanese and Koreans.

Specific presentation comments:

1. Figure 1: It would be helpful to include the sample size in the legend. Also, to this colorblind referee I cannot distinguish the two different colors for the Mongolian aDNA samples, please use two colors with more contrast than the current "orange".

We have included the sample size in the legend and also changed the colors.

2. Throughout the paper, multiple geographical locations are referred to without any explicit indication where they are situated. If locations such as the Tibetan plateau, Amur River Basin, Yellow River Basin, Tarim Basin, etc. could be indicated on Figure 1 will help orient the readers tremendously.

We have indicated the Tibetan plateau, Amur River, Yellow River, Yangtze River, Tarim Basin in the revised Figure 1.

3. I will respect the editorial / journal policy on this, but I always felt that it would be a nice gesture to include the archaeological site name in the native language, at least in parenthetical statements at first usage in the supplement, in deference to the local culture.

We have added the archaeological site name in every case where the archaeologist co-authors agreed it was appropriate.

Additional minor comments:

1. Line 166-167: how exactly were individuals "clustered"?

The clustering described in line 166-167 was mainly based on qualitative Principal Component Analysis and ADMIXTURE analysis, but for some additional analyses we used other metrics. We have now added a new section to the Methods discussing our clustering strategy as follows:

*"**Clustering of ancient individuals.** We clustered ancient individuals based on chronology and archaeological association, and then further based on both qualitative similarity (in PCA and ADMIXTURE and outgroup f_3-statistics) and quantitative homogeneity (based on f_4-statistics, and qpAdm results)."*

2. Line 354-356: similarly, how was this clustering conducted? By hierarchical clustering? Or by visual inspection?

We grouped 17 present-day Tibetan populations from the highlands into three categories based on genetic clustering patterns observed from Principal Component Analysis, ADMIXTURE analysis, and outgroup f_3 statistics (visual inspection), and f_4 statistics, qpAdm (quantitative testing for homogeneity). We clarify this in the revision.

3. Line 178-180: I do not think ALDER estimates multiple episodes of admixture; it assumes a single pulse event. I think authors meant multiple populations each having a signature consistent with past admixture. Should clarify the writing to avoid confusion.

We agree and to clarify have revised the writing as follows: "Groups in Northwest China, Nepal, and Siberia deviate toward West Eurasians in the PCA (Supplementary Information section 2, Figure 2), reflecting West Eurasian-related admixture in many populations that averaged 5 to 70 generations ago based on the decay of linkage disequilibrium. The method[34] we use cannot distinguish between admixture over a small number of generations or admixture over a long period driven by gene flow among neighbors (Online Table 6A and Online Table 6B)."

4. Line 425: I cannot find results that estimate 2-4% admixture in Northern Han Chinese from Tables S14 or S15.

The results are in Table 6B in the revision..

5. Can the MSMC results be shown in the supplement?

We showed the MSMC results in Supplementary Information section 3: Admixture graph modeling. We have added this as an Extended Data Figure as well.

6. In general, the authors can refer more specifically to supplemental or online tables for the reader. Some of these tables are extensive and a generic reference is not helpful. For example: at Line 207, I believe the relevant reference is Online Table 5B

We now refer to specific to supplemental or extended data information in our revision.

7. Details of the statistical analyses are generally short in the Methods. In contrast, the authors seem to pack quite a bit of the descriptions in the supplemental Table titles and legends (e.g. Table S14 and S15). I think it makes more sense to centrally describe the statistical analyses in the Methods, although perhaps the authors are limited by length constraints. Regardless, I do think the authors can do a bit more explaining of the interpretation of various statistics. For example, in qpWave, what does rank 1 imply, what does rank 2 imply? In a particular formulation of f4, what does positive values imply, what does negative values imply?

We have moved substantial method descriptions from supplemental table titles and legends to the Methods section. We have also added the following clarification about rank and f-statistics into the methods:

"In qpWave, a test for rank N means that we are evaluating whether the test populations are consistent with descending from as few as N+1 sources of ancestry."

"We use "outgroup-f_3" statistics of the form f_3(African_outgroup; Test, Comparison) to measure allele sharing between a Test population a Comparison panel. If we detect a significantly negative value for an "admixture-f_3" statistic of the form f_3(Test; Source1, Source2) we have evidence that a Test population is mixed between at least two ancestral populations differentially related (perhaps anciently) to Source1 and Source2. If we detect a significantly non-zero value of a statistic of the form f_4(A,B;C,D) we can be confident that populations A and B (or C and D) are not consistent with being descended from a homogeneous ancestral population that split earlier in time from the ancestors of the other two groups. A significantly positive value of an f_4-statistic of the form f_4(A,B;C,D) implies an excess allele sharing between populations A and C or B and D, while a negative value implies sharing between populations B and C, or A and D."

8. Line 204: Mongolia_East_N is labeled Monglia_N_East in Figure 2.

We have systematized to a uniform nomenclature in our revision, and in general reduced the use of abbreviated names.

9. Line 257: "showing"

This is fixed.

10. Line 393: I believe it is Table S5, not Table S8.

This is fixed.

Referee #2:

This is one of the most important papers on the human population history of Asia to appear in recent years. 191 ancient and 343 modern genomes are analysed, with the oldest going back as much as 8000 years. The results are overwhelming in the detail, although many of the conclusions that are drawn equate closely with current understanding of population history in the general region. The paper does not produce any new controversies, but it does reinforce a number of observations made previously, for instance about the importance of the Yangtze and Yellow river valleys in China, links between China and Southeast Asia, and relationships of the Jomon of Japan.

Since I am not a geneticist I cannot comment on the accuracy of the genomic calculations and comparisons presented. The text is very dense, and it is unlikely that anyone unfamiliar with the genomics of East Asian populations and their history and archaeology would be able to understand it easily. Paragraphs are very long, and embedded with many site and people names, statistical terms, and genetic categories. As I read it, I found myself wishing for more aids to understanding.

We have restructured the revised manuscript to make it easier to understand for a broad multidisciplinary audience.

One would be a list of all the abbreviations used in the text. For instance, I could discover that WSHG means West Siberian hunter-gatherer, but I still do not know what SG means.

We have passed through the manuscript reducing the use of abbreviations.

Individuals with an ".SG" suffix refer to individuals for whom the data derive from shotgun sequencing rather than in-solution enrichment data. We no longer use this abbreviation anywhere in the main text as it is unnecessarily confusing to a general reader..

Another would be to reduce the wonderful and highly informative figure S2 to something that can be more easily read. I needed a magnifying glass, since if this figure is magnified sufficiently to read, it becomes too large on a screen to navigate easily. Would it be possible to produce a simplified second version in which related populations are grouped in some way into many fewer columns, for instance, living Tibetans, living Europeans, Bronze Age Central Asians, prehistoric Southeast Asians, and so forth? Do we need to have every single population labelled and put in a separate column? Yes, maybe we do in the main chart, but information of this density could hopefully be summarised for presentation purposes.

We have grouped the populations regionally to fewer columns and replotted the figure. We have also made a second version of the figure that focuses on one particularly informative cluster-value, making it possible for readers to study the clustering at higher resolution (this is our new Extended Data Figure 3).

Another very useful addition for this table would be some idea of what the different ancestry component colours actually mean in terms of real populations. I can see that blue is West Eurasian, as in Anatolian Neolithic and derived populations, whereas orange presumably has something to do with the Neolithic in East Asia. Red is obviously Neolithic and Bronze Age steppes, green is Iran Neolithic, yellow is sub-Saharan African, and so forth. There is a discussion about each K number in the supplementary information, but it is hard to relate to the figure when the information is presented on this huge scale.

We present a guide to ancestry component colors—that is, we specify the populations in which the components are maximized—in Supplementary Information section 2: Overview of genetic substructure, as well as in a legend for the figure.

I noticed a few typos and I expect these will be picked up later in proof reading. However, line 128 appears to misspell Yukaghiric and Kamchatkan, and the so-called 'Altaic' listed here, together with Koreanic and Japonic, is referred to as 'Transeurasian' by the only author on the list that I perceive to be a linguist, namely Martine Robbeets. In the supplementary section around line 183 there are many (wrongly) lower case genus names.

We have corrected the spellings. In our revision, we no longer use "Transurasian" to refer to a larger grouping of languages in recognition of the fact that the existence of this grouping remains controversial.

I think my overall reaction to this paper is that it is a statement of the enormous power of modern genomic research into both ancient and modern populations. But presenting the information to people who are not geneticists or statisticians is a daunting task. Some of the paragraphs in this paper are almost 2 pages long, and the sheer density of information is very difficult to absorb. The paper has no discussion or conclusion, and indeed no subheadings at all. The abstract is a clear statement, however.

In our revision we have sought to make the manuscript more accessible by radically revising its structure:

1. *We have added a substantial Introduction that outlines questions we address and summarizes the archaeological and linguistic and genetic issues with which we engage*
2. *We have changed the order of presentation of results in a way that makes the logic easy to follow, and lay out this logic in the new expanded Introduction*
3. *We have added sub-headings*
4. *We have broken up paragraphs within each sub-heading.*

We think that the revised manuscript is much easier to read for a general multi-disciplinary audience.

In terms of the population sampling for ancient DNA, I did notice the complete absence of samples from Southern China (excluding Taiwan). The potential importance of the Yangzi River population is mentioned in the article from time to time, yet there is no ancient DNA from there.

We agree that the absence of Neolithic, Bronze Age, and Iron Age Yangtze River ancient samples is a limitation, and we are eager to see this data gap addressed in future research. While we unfortunately do not have such data, we feel that the inferences we are able to make without such data are robust.

This is mentioned in the very last sentence, which also contains the strange statement '… in particular to understand if dispersals of people in Southeast Asia do or do not correlate to ancient movements of people.' But what is it is a dispersal, if it does not involve a movement?

We meant to write "to understand if dispersals of languages in Southeast Asia do or do not correlate to ancient movements of people." However, we have now actually removed the sentence altogether.

In this regard, I notice that one of the authors on this paper is Hirofumi Matsumura, who in my view has presented clear craniometric information to suggest that Neolithic populations in SE Asia were part of a major movement, just as they were in Europe. I know that a paper such as this cannot easily discuss findings made by other disciplines, but it would be good to have some author opinions apart from those of the geneticists alone.

We now cite the Matsumura work – the genetic findings are entirely concordant with these craniometric results.

I strongly support publication, but it would be good if the article could be more user-friendly for non-geneticåist readerships.

We are optimistic that the re-writing and signposting in our revision will make the manuscript more accessible to all readers.

Referee #3:

The manuscript of Wang, Yeh, Popov, Zhang et al. brings to the published record ancient data from East Asia, a region where the amount of available ancient DNA data is very limited. Yet such data are crucial for our understanding of population demography of this and adjacent areas. Several sub-regions were analyzed: 52 samples come from Taiwan, 89 from Mongolia, 20 from a Neolithic site in China (Yellow River region) and 30 from other regions throughout East Asia. The authors also present genotype data from 383 present day individuals for comparison purposes. This data will be a major addition to the data already produced from populations in East Asia and this dataset should be published.

The analyses (laboratory and computational) are performed very well and the manuscript was easy to read and carefully written. The data will be likely used in the future by a number of researchers and the conclusions of this manuscript will be of interest to a large number of archaeologists and anthropologists working in different parts of East Asia.

While the data from East Asia are limited, it should be noted that some of the populations, regions and even sites have been analyzed for ancient DNA before (e.g. Jomon individuals, individuals from Amur River Basin, Neolithic individuals from Devil's Gate, individuals from Mongolia).

This does not invalidate the importance of the data as such a vast, anthropologically complex region requires very dense sampling. The scarcity of the data in the region makes many conclusions speculative, especially when based on qualitative descriptions and linked to linguistic theories (see below) but I particularly enjoyed the contextualized results (especially the case of the Afanasievo boy).

We are glad that the referee appreciates the finding of the non-Afanasievo-derived ancestry coupled with Afanasievo-derived material culture. Although Referee #1 suggested moving this result to the supplement we agree it is of high interest to archaeologists who are an important readership and have chosen to keep this finding as a highlighted result in our manuscript.

The diverse nature of the samples analyzed in this study is however one of the major weaknesses of the manuscript. The connection between different regions and periods that their dataset covers is very loose. No hypothesis or results seem to justify analyzing these particular samples together, except that they all are located in some region of East Asia. This negatively influences many aspects of the manuscript and it is difficult to easily fix. This also heavily undervalues the actual results of the study and their understanding for the readers. My suggestion would be to either rewrite the manuscript considerably (addressing the issues described in detail below) or to split this manuscript to two or more self-contained studies. For the benefit of the readers, I would prefer the latter solution but I understand that the authors may have their reasons to describe the data together in this manner. I have also some additional concerns regarding clustering of the individuals to groups, contamination and presentation of the results.

We agree that the original presentation of our manuscript was disconnected. To address this, we have heavily rewritten the manuscript as the referee suggests, making the following revisions to tie together our paper and clarify our approach:

(a) We start the revised manuscript by reviewing the results of previous studies and highlighting open debates related to population history in East Asia in all regions. We end the Introduction with an outline of our approach and a list of the questions we address.

(b) We next present the dataset and then a qualitative description of the dataset based on Principal Component Analysis and other methodologies

(c) We re-order the manuscript to start with the findings about the deep lineages in East Asia including the global admixture graph model. This provides a big-picture framing. We then point out how with high-resolution data from a couple of regions we are next able to go into detail about Holocene and Late Holocene population changes.

(d) We organize the part of the manuscript on Holocene population history of East Asia into three main sets of findings associated with large-scale expansions of Amur River Basin-related groups, expansions of Yellow River Basin groups, and expansions of southern groups.

(e) We next have a major section on East / West Eurasian contacts, where the Mongolian data are primarily described.

(f) We finally conclude

The new sections are all clearly marked out with headings. We think the resulting revision is much easier to read.

Major points

Temporal and geographical diversity of the samples negatively affects most parts of the manuscript:

See above for how we have re-ordered the manuscript to better sign-post the results and present them coherently. We think that the revised presentation provides readers with an effective way to navigate the findings of the study.

Title: Given the ambiguity of the some results and unequal sampling over vast regions and chronological periods, it is not appropriate to title the manuscript as being able to comment on the formation of the populations in the whole region. Yet I see that the lack of a result connecting the samples makes finding a better title difficult.

We have changed the title to a more careful "Insights From Ancient and Modern DNA into Human Population Formation in Eastern Asia".

Abstract: The abstract, as the manuscript itself, contains rather disconnected statements. Some of these statements are very vague and therefore hard to grasp or disapprove. Specifically connections to the spread and branching of language groups are quite weak and are not sufficiently discussed in the manuscript itself: there is no description of different theories regarding the linguistic variability, no hypotheses are formed and there is no discussion how different aspects of the analysis could influence them. The authors might be right that some qualitative aspects of their analysis are not inconsistent with some of the linguistic hypotheses mentioned. However without providing details of the linguistic discussion on the topic, alternative scenarios or discussion on the role of their results on the ability to reject these scenarios, it is not possible to make these claims, however carefully they are written. Especially not in the abstract where there is a little space for nuance. This is crucial as such statements might be used to support claims in other fields and scholars not familiar with genetics might interpret these narrative connections as scientific proofs (since they are so prominent in the abstract). This could be quite likely due to a lack of space in the manuscript because of the large number of different topics that need to be covered (as the context variability is so high).

We have completely rewritten the manuscript and believe it is much less easy to misinterpret.

If the authors want to make these claims, they should provide at least hypothesis descriptions and discussion on alternative scenarios in the manuscript and add a section regarding linguistics and its connection to their results to the supplementary material.

To provide much better context, we have a substantial Introduction summarizing the results of previous studies, and highlighting the archaeological and linguistic theories to which our genetic data could potentially speak. We conclude with a paragraphs listing the questions and open debates we address.

We do think that some of the most important implications of our manuscript lie in their connection to linguistic theories. For example, our analyses are almost impossible to reconcile with the hypothesis that a proposed "Transeurasian" language macrofamily (comprised of the widely accepted Mongolic, Turkic, Tungusic, Koreanic, and Japonic families) was propelled by a spread of people from the West Liao River center of agriculture, because the unique combination of ancestries that characterized the West Liao River Neolithic is not present in all these groups (even while we show that it is present in mixed form in Japan and Korea so a West Liao River spread to those regions is plausible). Similarly, our data provide extremely strong evidence of a recent connection between Tibetans, Han Chinese, and Upper Yellow River Neolithic people, highlighting this ancestry spread as a likely vector for the spread of Sino-Tibetan languages to both these regions. Finally, our data provide strong evidence of a connection of Taiwan Iron Age ancestry to ancestry across a broad region of southeast Asia. We thus continue to mention these linguistic theories in our abstract and main text, and have tried to motivate them better.

Introduction: I find the beginning of the manuscript difficult for the reader to follow: the first paragraph gives some generic information about East Asia and then it continues with description of the data acquired with the emphasis on their counts. Then there are some methodological details that should be (and actually sufficiently are, no need to reiterate them in this section) placed in the methods, especially since no new methods have been introduced in this study. Then there are results justifying the grouping of individuals and the manuscript further continues with more results with embedded interpretations that are rather divided by regions, each starting with a mini-introduction of the sub-region of their own (see lines 348, 381, 401 especially). What is seriously lacking is the comprehensive (even if short) description of the state of the art, the results of previous studies of ancient and modern genetic variability that comment on the forming of populations in these regions. Therefore the manuscript does not provide any hypotheses in the beginning or identify gaps in the knowledge that need to be filled. It is then hard for the reader to follow the flow of the results and the authors' reasoning.

We agree that the original presentation was disconnected. We have fully restructured the manuscript to provide an Introduction with pointers to current genetic understandings as well as open debates. The structure of the manuscript itself also now provides a much more compelling path through the results of our study.

Disjoint parts: There are parts in the manuscript that are completely disconnected to others (see paragraphs starting with lines 348, 381, 401, 443). Those could be deleted or placed elsewhere without any damage to the article. Again, as the readers go through the article, this easily creates confusion. The authors should describe in the beginning what they intend to tell the readers and use subheadings.

Following the referee's suggestions (and also the suggestions of other referees), we have intensively edited the manuscript, re-ordered content, and also added sub-headings to make the manuscript's findings easier to follow and to underscore how each set of findings relates to the broader picture.

Contamination issues:
I do not understand the use of “questionable” samples. The authors made a very detailed and clear Online Table 1 with information about the samples and they mark samples with problematic date, contamination estimate or other issue as “questionable”. For contamination filtering, they used the upper limit of contamMix estimation (of the amount of endogenous sequences) and they should definitely switch to median or to be conservative (as many other researchers) to the lower bound. What I find even more surprising is that some samples that were marked as questionable continue to be used in various further analysis. For example in Figure 3, several of the samples that were marked as questionable are displayed (among others, sample I13958 with contamination estimated to 16.9-40.8% as can be seen in Online Table 1). This should be at least mentioned in the caption of the figure and the contamination estimate (lower bound or the median) or the reason for doubting this sample should be specified. In general, the questionable samples should be excluded from all the analysis and if an exception is necessary, the reader should be made aware. If problematic samples are further used alongside samples without any issues, I do not see a point of any contamination analysis or dating evaluation. The readers expect that after the contamination analysis was performed, problematic samples are excluded and the data are hence believable. If I understand correctly, the problematic samples were also included in the counts of samples throughout the manuscript and abstract and in the analysis where groups were required (e.g. qpAdm analysis), potentially putting all those in question. Out of 191 ancient samples analyzed, 52 are marked as “questionable” and 9 even as “questinable_critical”: this puts 31.9% of their samples into serious doubt. Similarly, the samples with high contamination estimates were used to support genetic contextualization of other samples (that is especially problematic for groups with low number of samples: e.g. Mongolia_N_East sample I7031 with a contamination estimated to 10.9-21% as can be seen in Online Table 1). This is circular and problematic because there could have been a cross-contamination in the lab, the site or at the storage facility (not unlikely and understandable given the number of samples, as evidenced for example by misdating of some samples due to storage issues that is very well described and discussed in Online Table 1). Thus the contaminated samples have to be excluded from this analysis as well. If the authors want to use these samples anywhere e.g. because there are no other samples available from this particular site, the readers need to be warned of the contamination issues repeatedly in the main text and the implications for the results need to be thoroughly discussed. Contamination estimation for Wuzhuangguoliang female samples has not been performed at all even though there are methods to do so (e.g. based on LD or PMD patterns).

We completely agree that we had not addressed the contamination issues systematically and that some samples that had evidence of contamination had entered the analysis in our originally submitted manuscript. We have fully addressed all the issues the referee identified, and repeated our analysis on the appropriately revised dataset. Specifically:

- We have computed LD contamination estimates for all samples with sufficient coverage following the bioRxiv preprint:

Nakatsuka N, Harney É, Mallick S, Mah M, Patterson N, Reich D (2020) ContamLD: Estimation of Ancient Nuclear DNA Contamination Using Breakdown of Linkage Disequilibrium. bioRxiv, doi.org/10.1101/2020.02.06.938126.

- For all three contamination estimates, we rigorously use the conservative bounds of the 95% confidence intervals to flag “QUESTIONABLE” samples.

(a) For mtDNA contamination we conservatively set to an upper bound of match to the consensus sequence of <0.98 for “QUESTIONABLE and <0.9 for “FAIL / QUESTIONABLE_CRITICAL” (we no longer list this latter category of samples as newly reported).

(b) For the X chromosome contamination this is conservatively set to a lower bound of >0.01 for “QUESTIONABLE and >0.02 for “FAIL / QUESTIONABLE_CRITICAL”

(c) For the contamLD estimates applying the damage correction methodology this is conservatively set to a lower bound of >0.01 for “QUESTIONABLE and >0.02 for “FAIL / QUESTIONABLE_CRITICAL”

- We are now consistent about <u>never</u> including samples in our main population genetic analysis if they are inferred to have contamination of >0.01 (lower bound of 95% confidence interval) in nuclear data. Moreover, we only analyze samples with the mtDNA contamination estimates with upper bounds of <0.98 match to the consensus sequence if the nuclear contamination estimates rescue them by showing upper bounds in nuclear contamination of <0.02.

- We also apply a further test for contamination based on the ratio of Y to X+Y sequences, marking individuals with ratios in the range 0.03-0.35 as “QUESTIONABLE” (0.03-0.1 and 0.3-0.35) or “FAIL / QUESTIONABLE_CRITICAL” (0.1-0.3) as we have found empirically that these ranges are characteristic of data that is a mixture of male and female sequences.

- We do not remove 16 “QUESTIONABLE” or 10 additional low coverage samples with 5000-15000 SNPs on the nuclear targets from the list of reported individuals paper, as they are valuable data that generally have high proportions of authentic ancient DNA and provide reliable information such as sex assignment or mitochondrial haplogroup. However, we do not use them in any of them in our main population genetic analyses. The column “Included in population genetic analysis (also indicated if included in Figure 6)” indicates which samples are included in the main population genetic analysis; in addition to the 26 mentioned above, we also exclude from our main population genetic analyses 10 individuals who are detected genetically as first degree relatives of another higher coverage individual in the dataset.

- We have now also performed intensive analysis of the Wuzhuangguoliang samples and perform custom filtering on them on a sample-by-sample basis guided by mtDNA contamination estimates (which we have now generated based on the off-target sequences that do not map to the exome) as well as X chromosome contamination estimates for the males, using the rules described above to indicate which samples “PASS”. For all the Wuzhuangguoliang samples, we are now restricting to damaged sequences supplemented by additional undamaged sequences with lengths up to a threshold we infer from our analysis, and only include in the paper individuals that pass our rigorous threshholds for high quality data by all contamination metrics we can compute (and at least one of these metrics) and that yield at least 5000 SNPs on target. We have added a new Online Table 9 describing the filtering process. After removing individuals that do not yield sufficient data to analyze after damage restriction or that we detect as duplicates of other samples, we have reduced our Wuzhuangguoliang newly reported sample size from 20 to 11.

- After applying our more rigorous filters, the newly reported samples drop from 191 to 168. All of these 168 are valuable data and of these 132 are included in our main analyses as they give consistent evidence of high coverage (>15000 SNPs minimum) and authentic ancient DNA.

Cluster assignment and number of samples per group
The number of samples in the study is large, however given the size of the region and the chronological variability, this means that at times, the authors used only a few samples to represent whole populations at various time points. That, by itself, is standard when there is no other data available. However, this might result in an increased number of alternative interpretations of their analysis and the authors do not discuss this in sufficient detail. This is even exaggerated by the use of low coverage hence pseudohaploid data that further limit the amount of the data available for the analysis. While they mention the scarcity of their data regarding the interpretation of qpGraph results, it is not mentioned elsewhere. For example, the Mongolian clusters and the interpretation of the different waves of the Yamnaya-related admixture is dependent on the grouping of populations composed of a few individuals. While I do trust their qpAdm analysis (the Afanasievo being in the outgroups when Sintashta is the source indeed points to some later gene flow), I am hesitant to agree that composing the groups based on genetic similarity and then treating them as populations even though they rather overlap chronologically is a good course of action. This analysis (Figure 3) should instead be performed on an individual basis. A similar point is actually argued by the authors themselves in another part of the manuscript: in the case of Heishui_Mohe samples (see Supplementary Section 2), they refuse to make claims regarding the West Euroasian admixture in this population because the two samples they have from this period fall at different locations in PCA plot. Yet, if they would sample only one of these individuals, they would be confident (as the Mongolian example above suggests) to make completely opposite claims depending on which individual was sequenced. That demonstrates the dangers of overinterpreting results from a very low number of individuals and this needs to be addressed in the manuscript whenever this is the case. It should be also noted that the authors generally disregard a hypothesis that individuals that do show different genetic patterns (on PCA for example) still could be from the same group but that such a group is in a process of admixture (they form groups based on genetic similarity). Additionally, they repeatedly favour hypotheses of demographic events as migrations (“waves”) over continuous gene flow that are actually quite likely.

We agree that with a small number of individuals from any ancient group, it is impossible to make a confident assessment about the full spectrum of genetic variation in that region at that time. That said, a single individual or a few individuals can be very important, as they document the presence of a particular kind of ancestry in a place at a given time (proof-by-example). We have revised our manuscript everywhere to not overgeneralize from the very small numbers of individuals we have from some archaeological contexts; in contrast, when we have substantial numbers of individuals from particular contexts and observe either homogeneity (as in ancient Taiwan) or significant variation (as in the Afanasievo culture of Mongolia), we highlight it.

We agree with the referee that the individuals showing different genetic patterns still could be from a group that is in a process of admixture. For example, this is likely the case at the Kurgak Govi site where the individuals have different proportions of Western Siberian Hunter Gatherer-related and East Mongolian Neolithic-related ancestry and are not consistent with forming a clade but instead are on a cline. We now make this point explicitly in our revised manuscript.

In our revised manuscript, we no longer discuss heterogeneity of the Heishui Mohe samples because we removed one of the two from the dataset due to evidence of contamination.

In Online Table 8, we now have reported qpAdm results for each genetically homogeneous cluster (where the genetic homogeneity is assessed formally as described in the new Online Table 8G), as well as on individuals that cannot be clustered.

We have now also passed through the manuscript making sure the language does not favor pulse migration scenarios over continuous gene flow when both are plausible.

Our grouping of the individuals is not only based on genetic similarity, but based on time period and cultural associations, then further by genetic cluster which in the Mongolian samples we designated by number (our group names thus have the format "<Country>_<Additional Geographic Detail If Any>_<Time Period>_<Cultural Association If Any>_<Genetic Cluster>).We think that this naming scheme is in fact state-of-the-art and reflective of suggestions in the recent literature (e.g. the Eisenmann et al. Scientific Reports 2018 paper discussing naming of genetic clusters) and we continue to use this approach.

Presentation of the results
The graphical material added to the manuscript has some issues. That is especially difficult for evaluation of the clusters the authors defined. This is a major point because many other analysis and conclusions are based on this grouping.
Figure 1: the colours used in this map should be of more variety than red, orange and brown, it would then be easier to distinguish the points, also the authors could add some additional graphical elements to this map to better illustrate what is known and what they try to uncover (e.g. arrows).

We have replotted the map with better-contrasting colors as well as labels of geographic features mentioned in the text.

Figure 2: this is a very crucial figure for the text: among others, it should clearly demonstrate the grouping of the individuals. This is however very difficult to see at the moment. It would help if the authors added another figure with simplified legend and visually marked the individuals belonging to their clusters (or make other adjustments to the figure). I also suggest that more PCAs containing only samples from similar periods would be very insightful (instead of projecting all ancient samples to so many modern samples).

We have simplified the legend to make the figure more readable, and included polygons highlighting the clusters of ancient individuals. We have added PCAs containing only samples from similar periods into Supplementary Information section 2: Overview of genetic substructure.

Figure 3: This figure has some issues I mentioned elsewhere (contaminated samples and it should be done on the individual basis), I also think the ancestry proportions along the time axis are a bit difficult to compare (their size depends on the date) and it would be better to use for example the other axis to this purpose. Graphically, it seems to me of low quality (the figure resolution).

We have removed the contaminated samples. With respect, we have chosen to keep the display with bars of variable size as we found it informative and an efficient display of data; it was also highlighted as an excellent figure by referee #1.

Figure 4: I fail to see an apparent cline on this figure as stated in the text (line 432). For example, the most upper point is actually with more orange than those below; also the differences observable on the pie charts seem to be small. But otherwise it is an informative figure and the authors should consider if they do not want to add more figures of this type to illustrate other results of qpAdm that are otherwise not so easy to see.

The uppermost point is CHB (Han Chinese from Beijing, 1000 Genomes Project). Those individuals were collected from the residential community at Beijing Normal University which recruits people country-wide and thus is not particularly representative of the Beijing geographic region as described at the 1000-Genomes-Collections website:

https://www.coriell.org/1/NHGRI/Collections/1000-Genomes-Collections/Han-Chinese-in-Beijing-China-CHB

Since the residents of the Beijing Normal University come from many different parts of China, it is reasonable that they have a higher proportion of southern-associated ancestry (in orange) than other northern Han Chinese groups. We have removed this data-point from the plot as it is a cosmopolitan sample and mention this in the revised figure legend.

We agree with the referee that the genetic differences among Han Chinese groups are small but measurable as is shown in this figure and as we also document formally in Online Table 6M.

Figure 5: I must say I quite agree with the authors that "this admixture graph is an oversimplification" (line 489) and while I think it is a good addition to the paper, the authors might think about if it needs to discuss it in such detail in the text.

We have moved the admixture graph discussion forward in the paper as a way of framing the subsequent analysis, and in our revision are clear about the caveats even while highlighting it for the same reason that the referee feels it should be highlighted.

Figure 6: This figure lacks any labels on the x axis and I have some issues with the median of the first group, it rather looks like that the Fst is only a bit higher than in the Neolithic but one population is quite different. The caption should also mention in which table are the values underlying this figure placed.

The revised caption specifies explicitly that the x axis represents pairwise Fst.

In the first group "Before Neolithic Farming", the Fst values between Japan_Jomon and other populations are quite large. We have provided the Fst values in tabular form in a new Online Table 7.

Figure S1: This figure of Fst results is also crucial for the clustering and the clusters should be marked very clearly. I must say I am a bit worried that the authors use this figure for interpretation of some results while they themselves notice discrepancies for other populations (Papuans, as mentioned in the SI).

We have added cluster names onto the figure. We want to keep this supplementary figure because it shows a striking pattern that genetic clustering generally corresponds to linguistic and geographic classifications, but we also caution in the supplementary sections that some of the clustering can be deceptive. We discuss the Papuan and Australian case explicitly and also highlight other cases in which the position in the tree does not correspond perfectly to linguistic clusters or geography.

Figure S2: This admixture result is very hard to examine. Since this is again crucial for the group assignments, important findings should be presented in subfigures made from this main figure. However, I very much like the textual description of differences between analysis under different K in the SI.

We have added a separate supplementary figure for the ADMIXTURE plot which focuses on cluster number (k=15) and groups the populations regionally within each cluster.

We also use this regional clustering in a revised version of the main figure.

We separated important findings into subfigures in Supplementary Information section 2: Overview of genetic substructure.

Other points

Ethical concerns
I have concerns regarding ethics and sampling of the ancient individuals. While there is no formal requirement for the authors to contact local communities and get their approval for the sampling of the individuals discovered at archaeological excavations, it is certainly highly desirable and especially so if the ancient DNA studies (e.g. from adjacent regions and from some of the same authors) can be considered problematic in this regard. In my opinion, a publication aspiring to be at a high-level journal has to address this in the supplementary material and in the ethical statement (required by Nature) to make sure there are no ethical issues with the material (both from the research community and from the public). It should be noted however that the authors got proper ethical approvals for contemporary samples and with the ancient samples they did not breach any standard and that authors of many other publications do not go through the process of getting approval from the

communities etiher. But it would result in avoiding any dangers of bringing further distrust to the field of human population genomics.

We have adding an Ethics Statement covering both the modern and ancient samples to the beginning of the Method section:

"The modern sample collection was carried out in 2014 in accordance with the human ethical research principles of The Ministry of Science and Technology of the People's Republic of China (Interim Measures for the Administration of Human Genetic Resources, June 10, 1998) and genotyping was reviewed and approved by the Ethics Committee of the School of Life Sciences, Fudan University. Study staff informed potential participants about the goals of the project, and individuals who chose to participate gave informed consent consistent with broad studies of population history.

The ancient samples were collected with the permission of the custodians of the samples, who are the archaeologists or museums in each of the countries for which we analyzed the data. We applied a case-by-case approach to obtaining permissions for each set of samples depending on the local expectations as these vary by region and cultural context. Every newly reported ancient sample in this study has permission for analysis from custodians of the samples who are co-authors and who affirm that ancient DNA analysis of these samples is appropriate. For most samples, we prepared formal collaboration agreements to explicitly list the ancient DNA work being performed by our team. In other instances sample custodians who are co-authors determined that generation and publication of ancient DNA data was covered under their existing permissions for sample analysis, and so determined that new sampling agreements were not required."

Archaeological information
Given the high number of regions and periods covered, the archaeological supplementary material is rather long, containing information essential to the manuscript. I especially like the care taken to prepare very informative Online Tables (I applaud the Online Tables 1 and 2 especially). However, the quality level of SI is inconsistent: some archaeological sites are described with a lot of attention to detail, some are barely mentioned (e.g. compare the information about 18 individuals from Boisman-2 site and one individual from Nevelsk 2 site). The information about the archaeological context of the ancient individuals should be added to all sites (any graphical material would also be welcome). Also there is quite some lack of references (e.g. no reference at all for the information provided for Slab Grave culture graves, Mongol graves or one reference for Xiongnu burials). I understand that it is difficult to provide information for all the sites and individuals because of the number of samples and the diverse nature of the region and periods. But since this is essential for the grouping of the individuals (a basis of most of their analysis), this cannot be taken lightly. The authors should provide clear references for all the sites, in case some sites are not published, then the authors should provide enough information for other scholars to evaluate them (archaeological documentation of the finds and the site itself). Otherwise, the samples can be used incorrectly in future analysis of researchers using this dataset and it might be also hard to evaluate the current analysis.

I also have some reservations about the anthropometric descriptions used for interpretations of similarities of populations in this part of SI, they are rather dated (e.g. Slab Grave culture: "anthropological typing of this culture suggests they are 'Mongoloid' or similarly in the description Boisman site).

We have revised the manuscript to enrich a number of the archaeological discussions that were previously thin. We found it difficult to achieve homogeneous detail in the discussion of the archaeological content—in particular because of the limited archaeological information available from some sites especially in Mongolia—but the revision is better in this respect.

We have reduced the morphological descriptions of skeletons as they are not critical for our grouping of samples, and we no longer use terms like "Mongoloid". We note that in the countries where the physical anthropological analyses are done (Mongolia and Russia especially) these types of descriptions are generally viewed as valuable and meaningful and are widely used; in fact, the experience of the genetics community is that in Central Asia and Siberia and East Asia these physical anthropological classifications often correlate strongly genetic findings. Thus, fully excluding them from the descriptions completely would devalue the research of local scholars many of whom are co-authors, and so in some cases we do cite and briefly discuss observations based on these kind of data without using terms that are dated.

Different pipelines
Samples have been analyzed with 3 different pipelines. Already a part of the data has been processed differently in the lab and I wonder why the authors additionally bias the analysis by using different pipelines. It has been shown previously how even a small amount of bias can severely impact genetic inferences (Günther & Nettelblad 2019 PLOS Genetics) and here they treat the data with different versions of the same tool or even completely different tools. While the authors claim that

the difference in results between the pipelines is not large and cite the analysis of Fernandez et al. that showed that some of the conclusions in that paper are not biased while using different pipelines, I strongly disagree. The f4 statistics analysis of Fernandez et al. between the pipelines can hardly be generalized to other studies and samples and even in that study, it is hard to generalize for all the results. Furthermore, the Wuzhuangguoliang samples were treated with a third, completely different pipeline that has not been compared to the other pipelines at all..
Reanalysis of the data with the same pipeline should not be a major issue for the authors: the pipeline is scripted and the scripts are available and the data are quite low coverage and hence rather small. Therefore, all samples have to be analyzed as similarly as possible.

In our revised manuscript, we have addressed these issues in two major ways:

1. We have completely reprocessed the Wuzhuangguoliang samples using pipeline 1 to reduce the three pipelines to two comparable pipelines. As part of the Wuzhuangguoliang reprocessing, we have also applied sample-specific filters which have successfully removed evidence of contamination in these samples. This is a major improvement for these samples, and a new Online Table 9 describes this work and processing.

2. We have added an entirely new analysis in which we systematically compare pipeline 1 and pipeline 2 on sets of samples where we have individuals processed in both ways. Consistent with the findings in Fernandes et al. Nat. Ecol. Evol. 2020, we observe no systematic biases in symmetry-f_4 statistics between samples from the same context processed using pipeline 1 compared to pipeline 2. As an example, the Z-score of f_4(Boisman_MN_pipeline1, Boisman_MN_pipeline2; Taiwan_Hanben_pipeline1, Taiwan_Hanben_pipeline2)=0.348. We mention these analyses in the revised methods and in Online Table 6Q.

On reflection, we have decided not to reprocess all of our dataset using a single pipeline. There is heterogeneity in ancient DNA data in many ways, including

1. *between shotgun data and in-solution enrichment data*
2. *between samples processed with profoundly different library preparation methods (e.g. with and without UDG-treatment)*
3. *between samples sequenced on different sequencing instruments*
4. *between data generated at different laboratories*
5. *between data processed with different bioinformation pipelines (the issue the referee raises)*

For cases 1-4—which together are much more problematic than issue 5 in our experience—it is impossible to achieve homogeneity. But the truth is, we cannot even reprocess all samples using the same bioinformatic pipeline (issue 5), because the data relevant to this study is not just our newly reported data but also the previously published data with which we re-analyze it which inevitably was processed using different bioinformatic pipelines. While it would be a useful project to reprocess all the world's ancient DNA data from multiple laboratories from scratch using a uniform bioinformatic pipeline (and indeed we are currently hiring a person who will have this as a task), this would be a uear-long project that is not the focus of our current work, and we have demonstrated here that it is not necessary to obtain robust results for this particular set of samples.

Genetic results correlate with linguistic and geographic patterns
There are mentions of correlation between genetic, linguistic and geographic patterns in the region in the abstract, the manuscript and in the SI. The authors use only "qualitative" (as they state) assessments to make this case (mainly a PCA plot and Fst distance tree). But correlation cannot be mentioned when no statistical test was performed. While the authors might be right, it is very difficult to assess, given the figures provided: the plots contain too many groups and are not informative about linguistic or geographical assignments. The authors have to support these claims by one of many tests that would allow them to compare geographical, linguistic and genetic diversity. I would suggest using a simple Mantel test on the respective distances.

We have now added Mantel tests and report them in Online Table 5. Correlations to linguistics and geography are both significant at $P<0.0001$.

Incorporation of previously published results

Previous studies on the topic are mostly mentioned only when their data were analyzed together with the presented dataset. Or together with the results of the authors when they reach similar or the same conclusion without explicitly describing what new insight is brought even when there is one (e.g. line 432). The authors should be more precise in distinguishing their (indisputably important) findings from previous work. Additionally, some sections are missing mentions of previously published work. For example in lines 443-453 the authors discuss the formation of the modern Japanese population without mentioning any previous study that exists on the topic (there are many), not even the studies from which they use the data from (and are cited in the manuscript elsewhere).

We agree that we did not cite previously reported findings that were relevant in every case where we should have done so. We extend the content on the formation of Japanese by referencing the following previous studies that are related to ours:

Nakagome S, Sato T, Ishida H, et al. Model-based verification of hypotheses on the origin of modern Japanese revisited by Bayesian inference based on genome-wide SNP data. Mol Biol Evol. 2015;32(6):1533–1543.

Hammer MF, Karafet TM, Park H, et al. Dual origins of the Japanese: common ground for hunter-gatherer and farmer Y chromosomes. J Hum Genet. 2006;51(1):47–58.

Japanese Archipelago Human Population Genetics Consortium, Jinam T, Nishida N, et al. The history of human populations in the Japanese Archipelago inferred from genome-wide SNP data with a special reference to the Ainu and the Ryukyuan populations. J Hum Genet. 2012;57(12):787–795.

Kanzawa-Kiriyama, H., et al. A partial nuclear genome of the Jomons who lived 3000 years ago in Fukushima, Japan. J. Hum. Genet 62, 213–221 (2016).

Kanzawa-Kiriyama H., et al. Late Jomon male and female genome sequences from the Funadomari site in Hokkaido, Japan. Anthropol Sci. 127, 83-108 (2019).

We also have enriched the referencing elsewhere in the study.

Minor points and typos

- Supplementary section 3, 2nd and 3rd paragraphs: several typos ("implanted" instead of "implemented"; "to further differentiated" instead of "to further differentiate"; "constrained the model used MCMC")

We have made these corrections.

- Line 263: derived

We have made this correction.

- Data merging - it might be renamed to something like "reference panel preparation" (merging usually suggests merging of reads or data per sample).

We agree this term was confusing. We now use the header "Procedure for combining new Affymetrix Human Origins genotyping data with previously published data" in the Revised Methods.

- The authors sometimes use their codes instead of the labels they decided to use, e.g. in lines 239 and 248 where the reader needs to consult supplementary tables to identify the individuals at Figure 3 that are being referenced.

Everywhere in the text the population names are now consistent with the figures.

- Table S1: the caption should mention these are the modern individuals

We have changed the caption to add "present-day" before "individuals".

- Line 227: qpAdm mentioned for the first time, even though stated that it was mentioned "again"

We mentioned qpAdm in the previous paragraph describing the genetic modeling of Neolithic individuals from the cis-Baikal region; therefore this was the second mentioning of qpAdm.

- Line 249: individual instead of individuals

We have made this correction.

- Line 257: showing instead of shownning

We have made this correction.

- The authors treat all Han as one group while there is very well described variability between Southern and Northern Han (among many others, see Liu et al. 2018 Cell)

In the qpAdm modeling of Han-related influence in Mongolian ancient groups, we used CHB (Han from Beijing in 1000 genomes project) as an outgroup and Han.DG (Han Chinese samples from SGDP) as a potential source. We have added this detail into Online Table 5 to make it clear.

We fully agree about the existence of variability between southern and northern Han and indeed providing new insights into this is a result of our study (summarized for example in Figure 5). In modeling the two-way admixture of Han Chinese, we classified Han Chinese samples into subgroups according to their region and calculate the admixture proportions related to Upper_YR_LN and Liangdao2. We find a statistically significant north-south genetic cline and highlight this in our manuscript.

- Lines 256-260: all in one sentence, it is a bit hard to read

We have shortened this sentence as follows: "This model fits even when ancient European farmers are included in the outgroups, showing that if the long-distance transfer of West European megalithic cultural traditions to people of the Chemurchek culture[65] occurred, it was through spread of ideas rather than through movement of people."

- Line 321: the authors use term genetic continuity very loosely: if they want to use it they should statistically test for continuity. Similar Y and mtDNA haplogroups and approximately the same spot on the PCA is highly insufficient to claim this.

Thanks for catching this loose terminology. We did not actually mean to use continuity here, but instead just to state that the individuals were consistent with deriving entirely from the pre-Afanasievo components of ancestry in Mongolia (albeit with significantly different proportions of Western Siberian Hunter Gatherer and East Mongolian Neolithic related ancestry), without any Afanasievo/Yamnaya-related mixture. We have clarified this in our revision.

- Bwa versions used are not stated

We state in the methods section that we used bwa version 0.6.1 for pipeline 1, and version 0.7.15 for pipeline 2.

- There is a mistake in the Online Table 1 or 2 for the individual I6365. In one of them, the contamination on X chromosome is estimated as 20% and this is not mentioned in the other table. Probably the contamination method was used for a female while this works only on males and the result was not deleted after sexing.

The contamination estimate for individual I6365 was correct, and this individual was removed from the dataset as part of our new data curation. It is correct that the individual I3358 was a female and we forgot to delete the X-chromosome contamination estimate for this sample after sexing. We have fixed this in our revision.

Referee #4 (Remarks to the Author):

1. Title: “The Genomic Formation of Human Populations in East Asia”. This title does not accurately reflect the content of the manuscript, which aims to reconstruct the demographic history of East Asian populations through the lens of the genome, rather than to trace the formation of the genomes themselves. The title also does not accurately reflect the scope of the paper, which addresses particular aspects of the genetic formation of particular Asian populations; much of the existing genetic variation in East Asia is not treated at all. A large part of the paper (lines 194-315) deals with the eastward expansion of people associated with steppe cultures such as the Yamanya/Afanasievo and Sintastha/Andronovo. This aspect of the manuscript, although very interesting, is not central to the genetic history of most East Asians.

We have changed the title to more accurately capture the scope of the paper: “Insights from Ancient and Modern DNA into the Demographic History of Human Populations in Eastern Eurasia”.

2. Concerning the abstract in general: it is difficult to discern a unifying thread connecting the many disparate pieces of evidence presented throughout the text. It would be helpful if the hypotheses motivating the study were clearly stated at the outset.

We have rewritten the abstract and introductory section to provide much clearer guidance about the structure of the study. Our approach to this is described in our Response to Referees #1 and #3.

3. Line 94: “We document how 6000-3600 BCE people of Mongolia and the Amur River Basin were from populations that expanded over Northeast Asia, likely dispersing the ancestors of Mongolic and Tungusic languages“.

The significance of this statement is not clear: does the observation that ancient people from Mongolia and North China were spreading over Northeast Asia violate a commonly-held expectation? Second, it would be appropriate to state why the authors believe that these historical genomes represented the carriers of the two language groups. Perhaps a short statement motivating this inference would be helpful, along the lines of "...as inferred from the chronological/geographical appearance/distribution of..."? The same for the appearance of genetic signals associated with the Afanasievo culture in Mongolia “... plausibly acting as the source of the early- splitting Tocharian branch of Indo-European languages”. While the authors have clearly established that genetic signals from further west appear as far east as Mongolia, there are several other conceivable explanations for the spread of Trocharian into this region, and the origin of the language in that region could be considerably older—or younger. The claim is based on the assumptions that people associated with the Yamnaya and Afanasievov cultures spoke Indo-European languages, language transfer is primarily associated with the spread of groups of people, and an unbroken line of gene and language transfer has taken place from the European steppes to the eastern parts of Eurasia. While such tantalizingly simple models are not excluded by the present evidence, neither are they the sole obvious inferences in light of the current analyses. These claims are more appropriately understood as hypotheses requiring further testing.

The same is true for: “Analyzing 20 Yellow River Basin farmers dating to ~3000 BCE, we document a population that was a plausible vector for the spread of Sino-Tibetan languages both to the Tibetan Plateau and to the central plain ...” I would suggest reformulating this as one possible hypothesis while also addressing other models that could explain the data. A brief mention of the limitations of this analysis would be appropriate here.

We agree that we had inadequately contextualized the implications of the genetic data for linguistic theories; at the same time, our data do substantially increase the weight of evidence in favor of some scenarios and against others. In our revised manuscript, we are now much clearer about cases where our data provide evidence against an existing hypothesis (in some cases quite strongly) or for a hypothesis. In every case of the latter, the genetic data are of course not definitive so in our revision we always make it clear that our data increase the weight of evidence in favor of some hypotheses but do not prove them. Here we briefly discuss the three-major linguistic implications of this paper in turn.

(1) <u>Strong evidence against a prediction of the “Transeurasian hypothesis,” which helps to settle an ongoing debate.</u> Some linguists have proposed a Transeurasian macro-language family Mongolic, Turkic, Tungusic, Koreanic, and Japonic, based on reconstructed shared words for agricultural items and argued that its dispersal owes its origin to the spread of farmers from the West Liao River Region in northeast China. However, all the West Liao River farming individuals in the dataset we analyze harbor a mixture of Amur River Basin-related and Upper Yellow River Neolithic-river ancestry, a characteristic mixture that while present in modern Koreans and Japanese, is absent in ancient Mongolians and Tungusic speakers who

instead show high degrees of continuity from the pre-farming populations of each region. We have added discussion of this into the main text.

(2) Increasing the plausibility of the theory that Tocharian languages were spread by the Yamnaya expansion. To date, ancient DNA analyses have shown an extremely strong correlation of ancestry from Yamnaya steppe pastoralists with the distribution of spoken Indo-European languages. Along with historical linguistic reconstructions of a shared vocabulary for wheels and carts across these languages which set an upper bound on their spread to a time shortly before the Yamnaya, this makes it likely that the Yamnaya played an important role in the spread of Late Proto-Indo European languages. However, there has been more of a mystery around the earlier splitting lineage Tocharian, known only from inscriptions found in the Xinjiang region of eastern China; was it plausibly spread by the Yamnaya, too, and is there evidence of persistence of ancestry from the Yamnaya expansion into Xinjiang where it was spoken? Here we show not only that people from Afanasievo culture of the Altai but also people in Iron Age Xinjiang carried ancestry directly derived from the Yamnaya unadmixed with later waves of Yamnaya-derived ancestry spread from Andronovo/Sintashta-associated people. We are able to firmly reject recently proposed alternative models for the origin of the distinctive ancestry signal in Iron Age China (e.g. as mixtures of pre-Yamnaya western Siberian pastoralist ancestry and local Tibetan groups – we now discuss this in the text), and clearly document a specifically Yamnaya signal. Thus, our results document a chain-of-transmission of Yamnaya ancestry over time into the Tarim Basin of eastern China affirm that Yamnaya ancestry spread is an entirely plausible source for Tocharian languages just as it is for all spoken Indo-European languages. We clarify this in our revision, while also being careful not to state that our findings are a proof. For example, we have added the caveat “an important caveat is that the Tocharian texts date to a millennium after the Shirenzigou individuals so it is an open question how much of the ancestry of the writers of the Tocharian texts themselves derived from the Afanasievo.”

(3) Increasing the plausibility of a Yellow River basin farmer origin for the spread of Sino-Tibetan languages. There is a long-running debate on the origin of the linguistic link between Tibetan and Sinitic languages. Many linguists prefer the hypothesis that Neolithic farmers from the Upper and Middle Yellow River basin, speaking proto-Sino-Tibetan languages, expanded between 6,000–4,000 years BP both to the Tibetan Plateau becoming the main ancestors of present-day Tibeto-Burman speaking populations and to the central plain ultimately giving rise to groups including Han Chinese. By contrast, the other major hypothesis suggests that the origin and expansion of the Sino-Tibetan languages occurred at approximately 10,000–9,000 years BP from the southwest region of East Asia. Our finding that Han Chinese are the non-Tibetan group in East Asian most strongly related to Tibetans—and that Wuzhuangguoliang and other Upper and Middle Yellow River Basin Neolithic people are a uniquely good ancestry source for both groups—adds to the evidence of recent shared ancestry for both groups within the last 5000 years based on common Y chromosome to support the Yellow River basin hypothesis, while providing no support for the alternative theory of a specific link between indigenous groups in the Yangtze River Valley jointly to both highland and ancient Tibetans, and to Han Chinese). We have added this hypothesis testing into the discussion of the main text.

(4) Increasing the plausibility of a Yangtze River basin farmer origin for Austronesian, Tai-Kadai, and Austroasiatic languages. A common theory that has gained support from previous ancient DNA publications (especially in Southeast Asia and the Southwest Pacific, for example Skoglund et al. Nature 2016, Lipson et al. Science 2018, McColl et al. Science 2018) is that the expansion of these three languages families was propelled by the expansion of rice farmers from the Yangtze River valley as farming technology spread out of this region. However, there was little evidence for this from the core region where these languages were hypothesized to have originated. By sampling many indigenous modern populations from this region of China as well as documenting genetic continuity for >3500 years in Neolithic Taiwan, we show that indeed the ancestry associated with people speaking these languages is likely to have been anciently established in exactly this region. In our manuscript we are careful to caveat this – making clear that ancient DNA data from the Yangtze River region itself will provide additional insights – but we make it clear that our findings increase the weight of the evidence.

4. I am particularly surprised by the statement: “Yangtze Valley first farmers who likely spread Austronesian, Tai-Kadai and Austroasiatic languages across Southeast and South Asia. The basis for this claim (genomic or other evidence? Previously published work?) is not clear, and it is not an obvious inference from the data presented in the manuscript.

The inference is indirect and the weakest of our findings about language spread—since of course there is no available ancient DNA data from Yangtze River farmers--and we are clear about this in our revision. For example, in our abstract, we now no longer claim that this ancestry was associated with this language spread, writing now: “Third, we document a continuous population in Taiwan from >1400 BCE that derived ~75% of its ancestry from a group plausibly corresponding to Yangtze River Valley farmers and also matches southeast Asians speaking Austronesian, Tai-Kadai and Austroasiatic languages”

That said, our data and analysis increase the weight of evidence in favor of this theory by documenting a homogeneous ancestry type that is common and maximized in modern indigenous groups across the Yangtze River Valley, and that is also present in our >3500 year time transect in Taiwan and that has been shown in previous ancient DNA papers to have spread into Southeast Asia and the Southwest Pacific along with the spread of farming often admixing with previously established groups (Skoglund, M., et al. Genomic insights into the peopling of the Southwest Pacific. Nature 538, 510-3, Lipson, M., et al. Ancient genomes document multiple waves of migration in Southeast Asian prehistory. Science 361, 92-95 (2018), McColl, H., et al. The prehistoric peopling of Southeast Asia. Science. 361, 88-92 (2018)). Taking these findings together increases the weight of evidence in favor of a Yangtze River origin for this ancestry, and we hypothesize this clearly in our revision, while emphasizing that it is a hypothesis that can only be fully tested with future ancient DNA work from the Yangtze River Valley itself.

5. I don´t think that the following sentence will be clear to all readers because it uses nomenclature particular to archaeological specialists: “In a time transect of 89 Mongolians, we reveal how Yamnaya steppe pastoralist spread from the west by 3300-2900 BCE in association with the Afanasievo culture”. The key findings seem to be that the graves from “Shatar Chuluu kurgan” in Mongolia belong culturally to Afanasievo, and the genomes found there are identical to those in the core Afanasievo region further west. This is exciting, but I think that a broader readership will not be familiar with these particular archaeological terms. I suggest describing this finding in more accessible terms that better convey its significance, at least in the abstract.

We have retained the term “Yamnaya” as it has been revealed by ancient DNA beginning in 2015 that this archaeological culture is of critical importance for understanding transformations of culture in Eurasia and spread of culture and languages from Hungary to eastern China. To make the abstract more accessible, we have removed the reference to Afanasievo.

6. Given the geographic and chronological gap, is it appropriate to speak of the appearance of the Yamnaya Steppe pastoralists in Mongolia? How is the phylogenetic “Yamnaya” signal related to the actual processes under study?

Indeed, the Yamnaya material culture complex per se never reached Mongolia. Instead, ancestry derived from the Yamnaya steppe pastoralist arrived via successor cultures that received this ancestry from Yamnaya. This is why we always carefully use the phrase “Yamnaya-derived”. Pointing out that the ancestry is Yamnaya-derived is important because as far as we know the Yamnaya genetic profile existed in unmixed form only for a window in time between around 3300-2500 BCE; after the Yamnaya cultural phenomenon ended, this ancestry profile seems to have become mixed with ancestry from other sources. Thus, by studying whether people in East Asia have ancestry derived from the Yamnaya in this period, we can tell whether they trace ancestors directly to the Yamnaya spread itself. This is important in Mongolia and Xinjiang, as such an analysis shows which of the Yamnaya-derived ancestry in the region came from the initial Yamnaya-associated expansion from the west (via the contemporary Afanasievo culture who were genetically almost identical to Yamnaya and shared many cultural traits with them), or later expansions from the west.

I have a similar question concerning the following statement: “The second spread of Yamnaya-derived ancestry came via groups that harbored about a third of their ancestry from European farmers, which nearly completely displaced unmixed Yamnaya-related lineages in Mongolia in the second millennium BCE, but did not replace Afanasievo lineages”. It is interesting that the authors can detect several waves of gene flow from populations originating in Western Eurasia: the fact that these newcomers bear signs of deeper ancestry from southern Russia— or even from prehistoric populations in Europe— can help to better understand later population developments . But they are not the message itself. What could it mean that signals from European farmers and the western steppe are appearing in Mongolia? The fact that human populations are mobile and connected through a band of shared ancestry is interesting, but unenlightening without an understanding of the historical process that generated it.

We have revised our manuscript to provide archaeological background that helps the reader to understand the context of each of these ancestry expansions.

7. I find the following statement too anecdotal to appear in the abstract: “we also document a boy buried in an Afanasievo barrow with ancestry entirely from local Mongolian hunter-gatherers, representing a unique case of someone of entirely non-Yamnaya ancestry interred in this way“

The fact that an individual from a local population was buried in a kurgan from groups of immigrants is not entirely surprising.

While referee #3 thought this finding was of great interest we agree on reflection with referees #1 and #4 that it is of more peripheral interest to the general reading and so we removed it from the abstract, while still discussing it and contextualizing it in the body of the manuscript. We have retained the discussion in the main text because the finding is of importance for understanding the nature of Yamnaya and linked Afanasievo cultures, as to date, all individuals associated with Yamnaya culture burials (or associated with burials of their eastern counterpart Afanasievo) have been genetically homogeneous. Indeed, based on the dozens of samples to date from these cultures it seemed like there was a perfect correspondence between genetics and archaeological culture (unlike some other ancient archaeological cultures characterized with ancient DNA), which was consistent with the hypothesis that the Yamnaya and Afanasievo did not incorporate foreigners into their communities which was an observation that provided important social insight. The fact that we see an exception to this pattern is thus important.

8. "Groups in Northwest China, Nepal, and Siberia deviate towards West Eurasians in the PCA ..., reflecting multiple episodes of West Eurasian- related admixture "The fact that populations further west show more west Eurasian signals may simply reflect the natural spatial variation: how do the authors distinguish between an "admixture" event and other explanations for spatial variation? Are these "western signals" absent in earlier populations of the same area?

The same is true for "The other seven Neolithic hunter-gatherers from northern Mongolia ... can be modeled as having 5.4% ± 1.1% ancestry from a source related to previously reported West Siberian Hunter-gatherers" The authors write that they form "part of an east-west Neolithic admixture cline in Eurasia with increasing proximity to West Eurasians in groups further west". It would be helpful if the authors provided their criteria for distinguishing "ancestry complexity" from clinal variation under a null model of genetic variation.

We did not mean to make a distinction between admixture and clinal variation due to isolation by distance and gene flow among neighbors. We have edited the manuscript to remove any suggestion that we might be arguing for a process that is not driven by gene flow among neighbors setting up these clines.

9. Lines 182 to 189, "The "Amur Basin Cluster correlates geographically with .. .. southern parts of China speaking Austroasiatic, Tai-Kadai and Austronesian languages": The authors use terms like "correlates", "is most strongly represented", and "is maximized". Would it be possible to attach some numbers to these terms so that the reader can see how significant these associations are? The same comment applies to the following: "one falls closer to ancient individuals from the Amur Basin Cluster ('East' based on their geography), and the second clusters toward ancient individuals of the Afanasievo culture ('West'), while a few individuals take intermediate positions between the two": please clarify how many genomes and individuals are involved in each of these clustering analyses.

Every qualitative statement that we make in describing the Principal Component Analysis and ADMIXTURE plots is supported by formal statistical testing as reflected in the f_4-symmetry tests presented in the supplementary tables. In our revised manuscript we are now quantitative in our descriptions of correlations to linguistics and geography, and have added Mantel tests to support these inferences in Online Table 5.

10. The paragraph starting in line 317 contains a lot of fascinating results. I think it would be helpful if the essential questions and hypotheses discussed here were presented right at the beginning. As it stands, it seems a bit disconnected from the section of the paper immediately preceding it.

We have added the following sentences at the beginning of this paragraph:

"Previous studies have hypothesized local population continuity in the Amur River Basin based on the similarity of early Neolithic individuals from Devil's Gate at ~6000 BCE to present-day Ulchi[32]. However, the degree to which this population has been truly continuous has been difficult to determine due to the lack of ancient data from intermediate times."

We have also revised the manuscript with an Introductory section framing the questions, and with section headers, which help to guide the reader through the findings.

11. Regarding the sentences that start with line 340: "Some present-day populations...": the referent of these statements is not clear; perhaps the authors can clarify and re-phrase slightly?

We have removed this sentence in the revised manuscript.

12. Regarding line 368: “We estimate that the mixture occurred 60-80 generations ago ...” In my opinion, the single admixture event model on which this estimate is based is unrealistic, which has important consequences for the putative admixture event and further arguments following from this estimate. Can the authors test other models and compare them to their “model of a single pulse of admixture“?

Unfortunately the problem of distinguishing between a pulse admixture event that occurred over a relatively small number of generations and a long drawn out mixture process is very difficult, as discussed for example in Moorjani et al. PLoS Genetics 2011 and Loh, Lipson et al. Genetics 2013, due to the mathematical difficulty of distinguishing a mixture of exponential functions with similar decay constants with a single exponential. We have tried to distinguish these scenarios, but have not been able to. We have therefore simply revised our manuscript to be clear that we cannot distinguish these scenarios.

13. Regarding the “formation” of the Japanese (line 443): Can the authors please explain the actual significance of the admixture proportions estimated herein for Japanese demographic history? As the authors state, it is a signal from the deeper past, "that it is from an ancestral population related to those that contributed in large proportion to Han Chinese as well as to Korea". What can be learned from this, except that the Japanese had other East Asian ancestors? What was the initial hypothesis? Is it somehow confirmed or disproved by the admixture estimates?

The dual origins of present-day Japanese have been widely discussed (Hammer MF, et al. J Hum Genet. 2006, 51(1):47–58; Nakagome S, et al. Mol Biol Evol. 2015;32(6):1533–1543; Japanese Archipelago Human Population Genetics Consortium, J Hum Genet. 2012, 57(12):787–795; Kanzawa-Kiriyama, H., et al. 2016, J. Hum. Genet 62, 213–221; Kanzawa-Kiriyama H., et al. Anthropol Sci. 2019, 127, 83-108), but the admixture proportion from Jomon has been controversial and not resolved due to the deficiency in high-quality ancient Jomon genomes. We here reported seven whole-genome datasets from Jomon hunter-gatherers and used qpAdm modeling to propose ~8% Jomon deep ancestry in present-day Japanese. We have added the above background and discussion in the section of modeling Japanese. We also show that mixture of an anCient group (West Liao River farmers) with Jomon works as a model.

14. Line 455: Maybe the authors can explain right at the outset of this paragraph what they used qpGraph for? (I infer that it is to “identify a parsimonious working model for the deep history of key lineages“). I assume the authors use “lineage history” to infer population history, but it would be helpful if the authors stated what features of human population history they expect to be illuminated by reconstructing lineage histories. The genetic composition of human populations is uninteresting to most readers unless it can be used to elucidate real historical processes.

We have followed the referee’s suggestion to explain at the outset of this paragraph: “We used qpGraph[36] to explore models with population splits and gene flow to identify a parsimonious working model for the deep history of key lineages that contribute to populations representing ancestry extremes in our PCA (Supplementary Information section 3: qpGraph Modeling).” The admixture graph modeling is not just of esoteric interest but is likely to be of deep interest to the reader of this paper. For example, it is the admixture graph inference that allows us to confidently show recent shared ancestry between Tibetans and Yellow River farmers which is a key inference of the study, and also to reveal the linkages associated with the hypothesized Late Pleistocene coastal route expansion.

Reviewer Reports on the First Revision:

Referee #1 (Remarks to the Author):

In my view, the revision from Wang et al. has sufficiently addressed all my comments. The new framing of the narrative made the text flow more smoothly. The authors also did a good job incorporating recent aDNA data from East Asia that were published since their original submission. I am also glad to see that removal of low-quality samples from the previous version did not qualitatively change the results. I have only minor comments as described below.

1. More geographical locations should still be labeled on Figure 1. For example, Central Plains, Tibetan-Yi Corridor, Inner Asian Mountain Corridor.

2. In Figure 3, should the West Liao River farmers have greater proportion of the blue ancestry, having derived 67% ancestry?

3. Line 372-383 / Section C.3: the conflation of reduced heterogeneity in West Eurasia and the reduction in East Eurasia is a bit more complicated. The authors' claim rests on a point estimate of Fst over time transect of samples. However, PCA maps appear to give a different story by comparing the pattern of variation in modern West Eurasians (e.g. Figure 2 from Haak et al. Nature 2015) to the pattern seen here (Extended Figure 1). In West Eurasia, ancient samples generally lie outside of the present samples, suggesting a homogenization due to migration and mixture, resulting in low Fst in modern West Eurasians. In East Eurasia, on the other hand, many modern East Eurasian samples remain at the vertices of the PCA plot co-occupied by ancient individuals. I suppose these could be due to more recent population-specific drifts, but I do not think it would be fair to claim a parallel process in East Eurasia leading to lower Fst. The authors should further qualify this claim.

4. Line 529: please clarify what is referred to as the "Han Period". Throughout the paper Han is referring to an ethnic group, not a time period.

Referee #2 (Remarks to the Author):

This submission is a major improvement over the original. It is now broken down into shorter paragraphs and it follows a logical sequence in terms of time periods, regions, and important questions to be asked. I certainly found it much easier to follow than its predecessor, and for me as a non-geneticist it was much easier to understand, unburdened with large numbers of abbreviations and technical data.

On reading through I noted a number of potential issues.

1. Line 112 states: In a surprise, the remaining ~25% of Taiwan ancestry is from a distinctive lineage related to northern farmers, which may be linked to the earlier presence in Taiwan of foxtail millet than elsewhere in southern China.

On this, see Z. Deng et al., https://doi.org/10.1177/0959683617714603, who here date millet in Fujian to 4000 BP. But in communication with Deng I gather there is millet (unpublished) in the Guodishan site in Jiangxi at 5000 BP. The Taiwan millet, at 4800-4500 BP, is not really older than elsewhere in south China, and there is millet in Neolithic sites in the middle Yangzi, but not the lower. I suggest care be taken with millet – the grains are much smaller than rice and it may be that archaeologists have so far overlooked them.

It is later stated: A speculative possibility is that this ancestry was carried by cultivators of foxtail millet which was domesticated in the north by ~8000 BCE62,63, and which in the south appears earliest in the Taiwan Neolithic Dabenkeng culture (~3000-2500 BCE), but does not appear in mainland southern China until post-Neolithic times. This is incorrect – there is Neolithic millet in southern China, and mainland Southeast - Vietnam and Thailand.

2. Line 116 states: Yamnaya Steppe pastoralist ancestry, which was likely a vector for spreading late-proto-Indo-European languages
Linguistically, there can be no such thing as a late, as opposed to an early, proto-language on the scale of a whole language family. PIE was the first split that can be reconstructed (Anatolian vs the rest, presumably), and all else is later than PIE, including no doubt the Yamnaya.

3. Line 132 seems to suggest that the IE language family originated in East Asia. Surely not?

4. Lines 151 to 162 suggest that Tibetans have a separate Pleistocene genetic ancestry from Yellow River Neolithic farmers. So how does this square with a Yellow River origin for Sino-Tibetan languages? Lines 196-197 go on to state 'Our finding that the shared genetic signal between

Tibeto-Burman and Han Chinese is associated with Upper and Middle Yellow River farmers who lived ~3000 BCE supports the ‘northern origins hypothesis’; we estimate this group contributed ~84% of the ancestry to Tibetans and ~59-84% to modern Han Chinese.’
I am not suggesting anything is wrong here, but I find the wording confusing.

5. Lines 241-245 state: we have no other evidence of Turan-associated ancestry in the region until two millennia later in the Sagly culture of the 6th to 3rd centuries BCE who have cultural connections to the Scythians/Saka who are known to derive some of their ancestry from Turan. Why cannot this Turan ancestry be related to the Tocharians? If it was Scythian (Iranian?), then the Sagly culture was relatively close to the Tocharians in date and language, even if not in space. Tocharians have no archaeological or genetic identity, so relating them genetically to Yamnaya just because they had an Indo-European language seems to me to be jumping the gun. I do not doubt that Afanasievo people came from the western steppes, but they lived more than 1000 km north of the Tarim Basin, and 3000 years before the existence of the Tocharian Buddhist and commercial documents. If there is Turan ancestry in the general region of the Tarim Basin, why does the Afanasievo ancestry have to trump it?

5. References to indigenous peoples in Taiwan should call them Formosans where possible, rather than Taiwanese, who are of course Chinese both in language and in their recent mainland origin.

6. I was gratified to see that this paper supports the general view that major movements of early farming populations were very significant in human prehistory, particularly in the case of the language families indigenous to China, both northern and southern. I was a little surprised to see the Transeurasian hypothesis being demolished, especially given that linguist Martine Robbeets is one of the authors. What is her view on this? The paper does not say.
My own view on Transeurasian is that Neolithic Liao Valley population expansions around 6000 years ago, as with the Yellow River, were confined to areas where millet farming was significant, as in present day northeastern China, including inner Mongolia, Korea, and the Liao Valley. Existing Transeurasian subgroups are of much more recent origin than Proto-Transeurasian (if it existed) and may reflect language shift on the part of pastoralist populations on the fringes of the most productive regions of crop agriculture.

If you want my honest opinion, this is also how the Yamnaya originated!

Referee #3 (Remarks to the Author):

Response of Reviewer 3 to “Point by point response to referees for manuscript 2020-03-04062A”

As a a referee 3, I would like to react to the point by point response of the authors of 2020-03-04062A and the resubmitted manuscript itself. The authors have considerably improved the manuscript. I am especially happy that the contamination issues (a major flaw) has been remedied by removing problematic samples from the analysis (but not always from the counts, see below). I had however other major concerns regarding the analysis and not all of them have been addressed satisfyingly. In the following text I am commenting on the responses of the authors where a comment of mine is needed. All the other responses of authors I find fully or mostly satisfying and resolved.

Before I comment on specific changes, I would like to mention the rearrangement of the manuscript. This indeed helped a lot. The manuscript is much easier to follow, the major findings are highlighted and questions are posed and grounded in the previous literature. I really think this will help readers considerably. It should be noted that as a whole the article is still rather disconnected but given the variation of data, in all probability this is the best that could have been achieved. I find it peculiar that the introduction contains not only questions but also answers and that the state of the art of ancient genomics in the region is not introduced separately to give credit to authors of previous studies and clearly distinguish the findings but these references are intertwined with the discussions throughout the article and it might be better for readability this way. The only major thing that was not covered in the resubmission is to discuss potential alternative explanations and pitfalls – the manuscript is mainly a summary of the results. The rewriting of the manuscript suggests that the authors invested considerable effort into this resubmission and it is just a pity that some parts of the analysis were not treated the same way.

The specific comments:
(1)
My original review: Title: Given the ambiguity of the some results and unequal sampling over vast regions and chronological periods, it is not appropriate to title the manuscript as being able to comment on the formation of the populations in the whole region. Yet I see that the lack of a result connecting the samples makes finding a better title difficult.

The response of authors: We have changed the title to a more careful "Insights From Ancient and Modern DNA into Human Population Formation in Eastern Asia".

My comment to the response: This is only a small cosmetic change, just slightly weakening the previously over-stated title. The point still stands. Given the relatively small and unevenly distributed sample size (132 usable samples covering thousands of years and scattered over the whole East Asia) and disconnected results, the authors should be more specific.

(2)
My original review mentioned the severe contamination issues.
Among others authors responded: - We do not remove 16 "QUESTIONABLE" or 10 additional low coverage samples with 5000-15000 SNPs on the nuclear targets from the list of reported individuals paper, as they are valuable data that generally have high proportions of authentic ancient DNA and provide reliable information such as sex assignment or mitochondrial haplogroup. However, we do not use them in any of them in our main population genetic analyses. The column "Included in population genetic analysis (also indicated if included in Figure 6)" indicates which samples are included in the main population genetic analysis; in addition to the 26 mentioned above, we also exclude from our main population genetic analyses 10 individuals who are detected genetically as first degree relatives of another higher coverage individual in the dataset.
Reviewer 3: If the authors want to additionally report the samples that they call questionable (I agree with the reasoning but I worry if these will not be by mistake used by others), they should remove them from the sample counts in the abstract and manuscript so these numbers are not inflated (as they could mislead potential users of the dataset, or, very unfortunately, be used as some mark of quality and support for the conclusions).

(3)
The original review: Cluster assignment and number of samples per group
The number of samples in the study is large, however given the size of the region and the chronological variability, this means that at times, the authors used only a few samples to represent whole populations at various time points. That, by itself, is standard when there is no other data available. However, this might result in an increased number of alternative interpretations of their analysis and the authors do not discuss this in sufficient detail. This is even exaggerated by the use of low coverage hence pseudohaploid data that further limit the amount of the data available for the analysis. While they mention the scarcity of their data regarding the interpretation of qpGraph results, it is not mentioned elsewhere. For example, the Mongolian clusters and the interpretation of the different waves of the Yamnaya-related admixture is dependent on the grouping of populations composed of a few individuals. While I do trust their qpAdm analysis (the Afanasievo being in the outgroups when Sintashta is the source indeed points to some later gene flow), I am hesitant to agree that composing the groups based on genetic similarity and then treating them as populations even though they rather overlap chronologically is a good course of action. This analysis (Figure 3) should instead be performed on an individual

basis. A similar point is actually argued by the authors themselves in another part of the manuscript: in the case of Heishui_Mohe samples (see Supplementary Section 2), they refuse to make claims regarding the West Euroasian admixture in this population because the two samples they have from this period fall at different locations in PCA plot. Yet, if they would sample only one of these individuals, they would be confident (as the Mongolian example above suggests) to make completely opposite claims depending on which individual was sequenced. That demonstrates the dangers of overinterpreting results from a very low number of individuals and this needs to be addressed in the manuscript whenever this is the case. It should be also noted that the authors generally disregard a hypothesis that individuals that do show different genetic patterns (on PCA for example) still could be from the same group but that such a group is in a process of admixture (they form groups based on genetic similarity). Additionally, they repeatedly favour hypotheses of demographic events as migrations ("waves") over continuous gene flow that are actually quite likely.

The response of the authors: We agree that with a small number of individuals from any ancient group, it is impossible to make a confident assessment about the full spectrum of genetic variation in that region at that time. That said, a single individual or a few individuals can be very important, as they document the presence of a particular kind of ancestry in a place at a given time (proof-by-example). We have revised our manuscript everywhere to not overgeneralize from the very small numbers of individuals we have from some archaeological contexts; in contrast, when we have substantial numbers of individuals from particular contexts and observe either homogeneity (as in ancient Taiwan) or significant variation (as in the Afanasievo culture of Mongolia), we highlight it.

We agree with the referee that the individuals showing different genetic patterns still could be from a group that is in a process of admixture. For example, this is likely the case at the Kurgak Govi site where the individuals have different proportions of Western Siberian Hunter Gatherer-related and East Mongolian Neolithic-related ancestry and are not consistent with forming a clade but instead are on a cline. We now make this point explicitly in our revised manuscript.

In our revised manuscript, we no longer discuss heterogeneity of the Heishui Mohe samples because we removed one of the two from the dataset due to evidence of contamination.

In Online Table 8, we now have reported qpAdm results for each genetically homogeneous cluster (where the genetic homogeneity is assessed formally as described in the new Online Table 8G), as well as on individuals that cannot be clustered.

We have now also passed through the manuscript making sure the language does not favor pulse migration scenarios over continuous gene flow when both are plausible.

Our grouping of the individuals is not only based on genetic similarity, but based on time period and cultural associations, then further by genetic cluster which in the Mongolian samples we designated by number (our group names thus have the format "<Country>_<Additional Geographic Detail If Any>_<Time Period>_<Cultural Association If Any>_<Genetic Cluster>).We think that this naming scheme is in fact state-of-the-art and reflective of suggestions in the recent literature (e.g. the Eisenmann et al. Scientific Reports 2018 paper discussing naming of genetic clusters) and we continue to use this approach.
Reviewer 3: I agree that the grouping is not based only on genetic clustering but it is ALSO based on genetic clustering which is a highly problematic issue when this is then a basis of another analysis. Of course, it is really perfect to use this naming in the text as a label for certain samples that where shown to be similar. However, then applying an analysis according to these groups, especially if the groups are so so very small (1-3 individuals) is a bit pointless. Of course that these groups differ, that is how they were made! And it does not say anything about overall composition of the population at one point of time and space. Either the authors should group by geography+chronology+culture or they should analyze the samples on the individual basis. The

latter might be superior to show the variability in the populations (the individual mixtures are not dissimilar to ADMIXTURE graph), the former can be used to comment on the overall makeup of the populations that in turn is interpretable as formation of these populations. An analysis by clusters may be of interests if clusters are large and/or there is some suggestion of a genetic barrier at the site.

Additionally, the authors even remove outliers that do not fit with their expectations for population genetic inference: “We manually curated the data using ADMIXTURE and EIGENSOFT to identify samples that were outliers compared with other samples from their own populations. We removed seven individuals from subsequent analysis in this way; the population IDs for these individuals are prefixed by the string “Ignore_” in the dataset we release.” (as added in the newly resubmitted text). A few individuals that are found on the same place and dated to the same time are not necessarily “outliers” or a cluster to be analyzed as a population, they are just from a heterogeneous population and that is not a bad thing, it is actually something interesting and worth discussing. I do not see the need to create these uniform clusters, especially when they contain only 1-3 samples.

(4)
The original review: Figure 2: this is a very crucial figure for the text: among others, it should clearly demonstrate the grouping of the individuals. This is however very difficult to see at the moment. It would help if the authors added another figure with simplified legend and visually marked the individuals belonging to their clusters (or make other adjustments to the figure). I also suggest that more PCAs containing only samples from similar periods would be very insightful (instead of projecting all ancient samples to so many modern samples).
The response of the authors: We have simplified the legend to make the figure more readable, and included polygons highlighting the clusters of ancient individuals. We have added PCAs containing only samples from similar periods into Supplementary Information section 2: Overview of genetic substructure.
Reviewer 3: I thank the authors that they added the Supplementary Figures and improved the main Figure (that changed due to removal of the contaminated samples I presume). However, I was suggesting to attempt to do PCA without projecting (as stated in the original review). On reflection, it is likely that the authors cannot do that due to the low coverage of their data. It should be noted that the period PCA figures with projecting are not much of use (and if the authors want to remove them I fully support that decision).

(5)
The original review:Figure 6: This figure lacks any labels on the x axis and I have some issues with the median of the first group, it rather looks like that the Fst is only a bit higher than in the Neolithic but one population is quite different. The caption should also mention in which table are the values underlying this figure placed.

The response of the authors:The revised caption specifies explicitly that the x axis represents pairwise Fst.

In the first group “Before Neolithic Farming”, the Fst values between Japan_Jomon and other populations are quite large. We have provided the Fst values in tabular form in a new Online Table 7.
Reviewer 3: I am sorry I mentioned the first group in the my comment. I meant the difference between the medians of the second and the third group that might be driven by an outlier value and not informative of a general trend. This was however in the original Figure 6 and the newly resubmitted Figure contains much more points for some reason and it does not seem to be affected by this.

(6)
The original review: Figure S1: This figure of Fst results is also crucial for the clustering and the

clusters should be marked very clearly. I must say I am a bit worried that the authors use this figure for interpretation of some results while they themselves notice discrepancies for other populations (Papuans, as mentioned in the SI).
The response of the authors: We have added cluster names onto the figure. We want to keep this supplementary figure because it shows a striking pattern that genetic clustering generally corresponds to linguistic and geographic classifications, but we also caution in the supplementary sections that some of the clustering can be deceptive. We discuss the Papuan and Australian case explicitly and also highlight other cases in which the position in the tree does not correspond perfectly to linguistic clusters or geography.
Reviewer 3: I am a bit confused by the reasoning here. I of course noticed that some populations (Papuans) are discussed as an example of the weaknesses of this plot but that is exactly my point – when it agrees with what the authors expect, they believe the results, if it does not, they do not believe them. Either this analysis is solid or it is not – if it is affected by admixture here, maybe it is affected elsewhere and actual position on the tree of some other samples should be different too. Also, in this part authors continue to mention (as observed in my original review) that they have correlations with language and geography and no correlation test was performed between the tree and languages and geography (Mantel was performed – great!) . That is just a loose terminology though.

(7)
My original review: Ethical concerns
I have concerns regarding ethics and sampling of the ancient individuals. While there is no formal requirement for the authors to contact local communities and get their approval for the sampling of the individuals discovered at archaeological excavations, it is certainly highly desirable and especially so if the ancient DNA studies (e.g. from adjacent regions and from some of the same authors) can be considered problematic in this regard. In my opinion, a publication aspiring to be at a high-level journal has to address this in the supplementary material and in the ethical statement (required by Nature) to make sure there are no ethical issues with the material (both from the research community and from the public). It should be noted however that the authors got proper ethical approvals for contemporary samples and with the ancient samples they did not breach any standard and that authors of many other publications do not go through the process of getting approval from the communities etiher. But it would result in avoiding any dangers of bringing further distrust to the field of human population genomics.

The response of the authors: We have adding an Ethics Statement covering both the modern and ancient samples to the beginning of the Method section:

"The modern sample collection was carried out in 2014 in accordance with the human ethical research principles of The Ministry of Science and Technology of the People’s Republic of China (Interim Measures for the Administration of Human Genetic Resources, June 10, 1998) and genotyping was reviewed and approved by the Ethics Committee of the School of Life Sciences, Fudan University. Study staff informed potential participants about the goals of the project, and individuals who chose to participate gave informed consent consistent with broad studies of population history.

The ancient samples were collected with the permission of the custodians of the samples, who are the archaeologists or museums in each of the countries for which we analyzed the data. We applied a case-by-case approach to obtaining permissions for each set of samples depending on the local expectations as these vary by region and cultural context. Every newly reported ancient sample in this study has permission for analysis from custodians of the samples who are co-authors and who affirm that ancient DNA analysis of these samples is appropriate. For most samples, we prepared formal collaboration agreements to explicitly list the ancient DNA work being performed by our team. In other instances sample custodians who are co-authors determined that generation and publication of ancient DNA data was covered under their existing permissions for sample analysis, and so determined that new sampling agreements were not

required."
Reviewer 3: As stated, the authors followed the standard procedures and now added statements about that to the text which is great as it provides an example to others. It would have been nice to go beyond that and really involve the communities concerned not just officials (a ministry), though. I see that it is difficult to do so post quem and I will just keep hoping that no communities or their diasporas would find the results or the discussion in this study damaging or hurtful (I am not aware of any such issues but then again I am not, for example, a Tibetan).

(8)
My original review:Archaeological information
Given the high number of regions and periods covered, the archaeological supplementary material is rather long, containing information essential to the manuscript. I especially like the care taken to prepare very informative Online Tables (I applaud the Online Tables 1 and 2 especially). However, the quality level of SI is inconsistent: some archaeological sites are described with a lot of attention to detail, some are barely mentioned (e.g. compare the information about 18 individuals from Boisman-2 site and one individual from Nevelsk 2 site). The information about the archaeological context of the ancient individuals should be added to all sites (any graphical material would also be welcome). Also there is quite some lack of references (e.g. no reference at all for the information provided for Slab Grave culture graves, Mongol graves or one reference for Xiongnu burials). I understand that it is difficult to provide information for all the sites and individuals because of the number of samples and the diverse nature of the region and periods. But since this is essential for the grouping of the individuals (a basis of most of their analysis), this cannot be taken lightly. The authors should provide clear references for all the sites, in case some sites are not published, then the authors should provide enough information for other scholars to evaluate them (archaeological documentation of the finds and the site itself). Otherwise, the samples can be used incorrectly in future analysis of researchers using this dataset and it might be also hard to evaluate the current analysis.

I also have some reservations about the anthropometric descriptions used for interpretations of similarities of populations in this part of SI, they are rather dated (e.g. Slab Grave culture: "anthropological typing of this culture suggests they are 'Mongoloid' or similarly in the description Boisman site).

The response of the authors: We have revised the manuscript to enrich a number of the archaeological discussions that were previously thin. We found it difficult to achieve homogeneous detail in the discussion of the archaeological content—in particular because of the limited archaeological information available from some sites especially in Mongolia—but the revision is better in this respect.

We have reduced the morphological descriptions of skeletons as they are not critical for our grouping of samples, and we no longer use terms like "Mongoloid". We note that in the countries where the physical anthropological analyses are done (Mongolia and Russia especially) these types of descriptions are generally viewed as valuable and meaningful and are widely used; in fact, the experience of the genetics community is that in Central Asia and Siberia and East Asia these physical anthropological classifications often correlate strongly genetic findings. Thus, fully excluding them from the descriptions completely would devalue the research of local scholars many of whom are co-authors, and so in some cases we do cite and briefly discuss observations based on these kind of data without using terms that are dated.
Reviewer 3: The term "Mongoloid" and "Caucassian" is still used in the Supplementary information at line 455. It is worrying that the authors do not realize the dangers of such terminology. As was shown previously in anthropological literature, many of the skeletal traits used for such classifications are distributed along geographical clines and thus such "classes" obviously correlate with genetics. However, the act of classifying individuals on the basis of these continuous traits into groups is just not valuable in anyway (yes, it is done today but in all countries where it is done, it is only for continuity with past studies that were written in different times). What is more,

the classification was misused in the past for racial theories and continues to be problematic and hurtful for some communities. If the authors want to correlate skeletal measures with genetics, they should do so properly – using the actual measurements. Even then I do not really see the point if the goal is to study population formation and movement as these skeletal measures are just (weak) proxies for genetics. But why not. Otherwise, the SI did improve.

(9)
The original review: Different pipelines
Samples have been analyzed with 3 different pipelines. Already a part of the data has been processed differently in the lab and I wonder why the authors additionally bias the analysis by using different pipelines. It has been shown previously how even a small amount of bias can severely impact genetic inferences (Günther & Nettelblad 2019 PLOS Genetics) and here they treat the data with different versions of the same tool or even completely different tools. While the authors claim that the difference in results between the pipelines is not large and cite the analysis of Fernandez et al. that showed that some of the conclusions in that paper are not biased while using different pipelines, I strongly disagree. The f4 statistics analysis of Fernandez et al. between the pipelines can hardly be generalized to other studies and samples and even in that study, it is hard to generalize for all the results. Furthermore, the Wuzhuangguoliang samples were treated with a third, completely different pipeline that has not been compared to the other pipelines at all.. Reanalysis of the data with the same pipeline should not be a major issue for the authors: the pipeline is scripted and the scripts are available and the data are quite low coverage and hence rather small. Therefore, all samples have to be analyzed as similarly as possible.

The response of the authors: In our revised manuscript, we have addressed these issues in two major ways:

1. We have completely reprocessed the Wuzhuangguoliang samples using pipeline 1 to reduce the three pipelines to two comparable pipelines. As part of the Wuzhuangguoliang reprocessing, we have also applied sample-specific filters which have successfully removed evidence of contamination in these samples. This is a major improvement for these samples, and a new Online Table 9 describes this work and processing.

2. We have added an entirely new analysis in which we systematically compare pipeline 1 and pipeline 2 on sets of samples where we have individuals processed in both ways. Consistent with the findings in Fernandes et al. Nat. Ecol. Evol. 2020, we observe no systematic biases in symmetry-f4 statistics between samples from the same context processed using pipeline 1 compared to pipeline 2. As an example, the Z-score of f4(Boisman_MN_pipeline1, Boisman_MN_pipeline2; Taiwan_Hanben_pipeline1, Taiwan_Hanben_pipeline2)=0.348. We mention these analyses in the revised methods and in Online Table 6Q.

On reflection, we have decided not to reprocess all of our dataset using a single pipeline. There is heterogeneity in ancient DNA data in many ways, including
1. between shotgun data and in-solution enrichment data
2. between samples processed with profoundly different library preparation methods (e.g. with and without UDG-treatment)
3. between samples sequenced on different sequencing instruments
4. between data generated at different laboratories
5. between data processed with different bioinformation pipelines (the issue the referee raises)

For cases 1-4—which together are much more problematic than issue 5 in our experience—it is impossible to achieve homogeneity. But the truth is, we cannot even reprocess all samples using the same bioinformatic pipeline (issue 5), because the data relevant to this study is not just our newly reported data but also the previously published data with which we re-analyze it which inevitably was processed using different bioinformatic pipelines. While it would be a useful project to reprocess all the world's ancient DNA data from multiple laboratories from scratch using a uniform bioinformatic pipeline (and indeed we are currently hiring a person who will have this as a task), this would be a uear-long project that is not the focus of our current work, and we have demonstrated here that it is not necessary to obtain robust results for this particular set of samples.

Reviewer 3: I am glad that the authors decided to reduce the amount of artificial bias by lowering the number of pipelines used. It is also nice that they added f4-statistics to assess the bias. I however do not agree that this is conclusive in any way about the non-existence of such bias as the test is limited to only some samples and some comparisons. While the presence of a significant systematic results in this test (Online Table 25 I think – not 6Q as mentioned in the rebuttal letter) would indeed be a proof of a bias, the absence of such significant results is definitely not a proof of the absence of a bias. This is a very important distinction. The only conclusion that can be drawn from this analysis is that a bias could not be found between those exact samples in this exact setting. This does not exclude a potential structural biases resulting in differing results between other samples or other groups of samples. The only way the authors could prove the absence of a bias generally would be to reanalyze all the samples and compare the results, especially the analyses that matter for their conclusions (which amounts to essentially reanalyzing the data as they should be done in the first place). All kinds of analysis (even those based on f-statistics, especially sensitive might be qpGraph) might have been influenced by this in an unknown manner. I was worried when the f-statistics test was interpreted this way in literature (Fernandez et al. 2020) and further propagating this misinterpretation (over-generalization) would be most unfortunate and it can be even further misused so very easily (e.g. by selecting to show only those comparisons that are indeed insignificant - which I trust the authors did not do in this case but can be easily done to support any differences in pipelines following the lead of this manuscript and Fernandez et al.). If the authors decide to keep this “bias analysis”, they should definitely avoid any suggestions that the test proves the absence of a bias, especially expressions like “Indistinguishability of population genetic results” (it is the absence of the artificial shared drift between some particular configurations of some of the samples) and “We verified that results from the two pipelines were indistinguishable from a population genetic point of view” (again it is only from a point of a view of a few particular comparisons). This is not to say that such a test is without merit, I can imagine that there could be no way for the samples to be made comparable (e.g. different laboratory procedures of destructive nature) and such a test could alleviate some concerns over systematic biases (it really does). However, even then a potential bias would have to be still brought up in a discussion of the result. However, the authors do not discuss what portion of their samples was analyzed with what pipeline in relation to their results or in discussion (the distribution of samples between different pipelines is not random among the groups and populations and some results might be affected more than others). And this all while the heterogeneity of the bioinformatic analysis is just for convenience, not a must.

Additionally, I reject in principle one argument the authors make in order to justify keeping the different pipelines. The fact that there is some heterogeinity in the aDNA data production is by no means an argument to produce more bias in the bioinformatic analysis. And it does not matter that the authors claim that in their experience the different pipelines make usually less of a difference, in my experience it matters. That is connected to another point - there might have been some misunderstanding: I did not suggest that the authors should “reprocess all the world’s ancient DNA data from multiple laboratories from scratch using a uniform bioinformatic pipeline” (as stated in the rebuttal letter), though I strongly agree with the authors that this would be the best course of action and quite a few laboratories now proceed in this manner (exactly for the reasons of producing sound high quality results). I asked in my review that the data produced in this one study are analyzed the same way (it would have been nice if they included a few published samples from the same region but this is not a minimal quality standard). Given the scarcity of the aDNA data in this region, that means that many of the results will then be produced without this potential bias (as the authors themselves state several times, the amount of other data from the region is limited). I do not think that asking for this is unreasonable and I am not convinced that some results comparing populations analyzed with different pipelines are not affected and hence are questionable.

It should be noted that in aDNA literature, it often happens that new samples and published reference samples from different laboratories and publications are compared together without re-analysis. But a level of caution is often expressed in discussions in these cases and most readers from the field do implicitly understand the issue of problematic comparability between different studies. And as the field moves forward more and more attention is given to reducing such biases (see Martiniano et al 2020, Cassidy et al. 2020, Gunther et al 2018). Analyzing samples in the same study differently is a step in a completely opposite direction to this trend and a validation of such an approach by publishing a study with such a bad practice in a high profile journal could set the field back. The statement in the rebuttal letter that suggests that it requires effort to produce comparable data is true but such an absence of full comparability of data coming from the same study undermines the effort of other groups that do spend the time and effort to avoid such issues even between different studies from different laboratories. It even signals that the field of ancient population genomics itself is still not mature and it is embarassing (it would be just unthinkable that controls and cases in clinical studies would be analyzed with different pipelines as some populations in this study are) .

(10)
The original review: The authors sometimes use their codes instead of the labels they decided to use, e.g. in lines 239 and 248 where the reader needs to consult supplementary tables to identify the individuals at Figure 3 that are being referenced.
The response of the authors:Everywhere in the text the population names are now consistent with the figures.
Reviewer 3: This is a detail but in both of the lines mentioned above (line 239, now 611 and line 248 now 621), the individuals are described by the skeletal code and the lab code only and cannot be identified in Figure 6 easily without looking into the tables for the population label used. In the revised text, the population labels are mentioned some time before (line 606) but I think it would just be really more clear to have the code and the population label somewhere mention together outside of SI for those samples that are mentioned idividually (maybe in the figure caption or in the same brackets on lines 611 and 621). It was quite confusing to me.

(11)
The original review: Line 321: the authors use term genetic continuity very loosely: if they want to use it they should statistically test for continuity. Similar Y and mtDNA haplogroups and approximately the same spot on the PCA is highly insufficient to claim this.
The response of the authors: Thanks for catching this loose terminology. We did not actually mean to use continuity here, but instead just to state that the individuals were consistent with deriving entirely from the pre-Afanasievo components of ancestry in Mongolia (albeit with significantly different proportions of Western Siberian Hunter Gatherer and East Mongolian Neolithic related ancestry), without any Afanasievo/Yamnaya-related mixture. We have clarified this in our revision.
Reviewer 3: While the authors exchange the expression “The genetic continuity is also evident” (previously line 321) to “The genetic similarity is also evident” (now line 400), immediately in the following sentence and concerning exactly the same populations they bring the term “continuity” back (line 400-401): “The continuity in the Amur River Basin was disrupted..”. The text has therefore not been clarified for this issue at all. What is more this whole paragraph in the resubmitted manuscript now centers around the question of continuity in the region, as a question the authors are addressing (a question stemming from previous works). But the authors in this study actually do not show continuity in their analysis. Again, there are dedicated methods for that and PCA (Figure 2) and ADMIXTURE (Extended Figure 3) they mention to support the statement is not one of them. For PCA, obviously any change in a population not captured by the principal components would not be observable and could completely derail the qualitative assessment of “continuity” – reasons for such an occurrence could vary, scarcity of the sampling and affinities to unsampled populations immediately come to mind. ADMIXTURE is arguably better but still a failure to observe substructure for these few individuals when the analysis was performed on a large data

set would not surprise me and hence I am not convinced. In sum, the resubmitted text is in this regard even worse than the initial submission. And this is not only wrong use of terminology but over-interpretation of results (language would have to be improved and potential pitfalls mentioned) and an incorrect method for a question asked.

Referee #4 made no comments to the authors.

Author Rebuttals to First Revision:

Referee #1 (Remarks to the Author):

In my view, the revision from Wang et al. has sufficiently addressed all my comments. The new framing of the narrative made the text flow more smoothly. The authors also did a good job incorporating recent aDNA data from East Asia that were published since their original submission. I am also glad to see that removal of low-quality samples from the previous version did not qualitatively change the results. I have only minor comments as described below.

1. More geographical locations should still be labeled on Figure 1. For example, Central Plains, Tibetan-Yi Corridor, Inner Asian Mountain Corridor.
Reply：We have added labels for the Central Plains and the Tibetan-Yi Corridor as suggested. Since we cropped the PCA in the east-west direction to reduce space, the Inner Asian Mountain Corridor is no longer visible on the map.

2. In Figure 3, should the West Liao River farmers have greater proportion of the blue ancestry, having derived 67% ancestry?

Reply：We made a mistake in the coloring for West Liao River farmers on this figure and we have corrected this in the revision.

3. Line 372-383 / Section C.3: the conflation of reduced heterogeneity in West Eurasia and the reduction in East Eurasia is a bit more complicated. The authors' claim rests on a point estimate of Fst over time transect of samples. However, PCA maps appear to give a different story by comparing the pattern of variation in modern West Eurasians (e.g. Figure 2 from Haak et al. Nature 2015) to the pattern seen here (Extended Figure 1). In West Eurasia, ancient samples generally lie outside of the present samples, suggesting a homogenization due to migration and mixture, resulting in low Fst in modern West Eurasians. In East Eurasia, on the other hand, many modern East Eurasian samples remain at the vertices of the PCA plot co-occupied by ancient individuals. I suppose these could be due to more recent population-specific drifts, but I do not think it would be fair to claim a parallel process in East Eurasia leading to lower Fst. The authors should further qualify this claim.
Reply：On reflection we agree that the analogy to the West Eurasian case was forced, and we feel that at present there is not enough sampling of East Asians to make a truly region-wide assessment about how population differentiation has changed over time (in particular, there is not enough sampling of hunter-gatherer and early farmer populations in diverse places in China and Tibet). We have therefore removed Figure 4 and discussion of this analogy, and now simply mention how admixture reduced population differentiation in the Holocene in three instances that we track in this study, pointing to Online Table 8 for the F_{ST} results.

4. Line 529: please clarify what is referred to as the “Han Period”. Throughout the paper Han is referring to an ethnic group, not a time period.
Reply：We agree our terminology was inconsistent. We now write: “The cline of increasing Liangdao-related ancestry in southern Han today is plausibly due to expanding Han mixing with southern groups as they spread into southern China as recorded in the historical literature.[35]”

Referee #2 (Remarks to the Author):

This submission is a major improvement over the original. It is now broken down into shorter paragraphs and it follows a logical sequence in terms of time periods, regions, and important questions to be asked. I certainly found it much easier to follow than its predecessor, and for me as a non-geneticist it was much easier to understand, unburdened with large numbers of abbreviations and technical data.

On reading through I noted a number of potential issues.

1. Line 112 states: In a surprise, the remaining ~25% of Taiwan ancestry is from a distinctive lineage related to northern farmers, which may be linked to the earlier presence in Taiwan of foxtail millet than elsewhere in southern China.

On this, see Z. Deng et al., https://doi.org/10.1177/0959683617714603, who here date millet in Fujian to 4000 BP. But in communication with Deng I gather there is millet (unpublished) in the Guodishan site in Jiangxi at 5000 BP. The Taiwan millet, at 4800-4500 BP, is not really older than elsewhere in south China, and there is millet in Neolithic sites in the middle Yangzi, but not the lower. I suggest care be taken with millet – the grains are much smaller than rice and it may be that archaeologists have so far overlooked them.

It is later stated: A speculative possibility is that this ancestry was carried by cultivators of foxtail millet which was domesticated in the north by ~8000 BCE62,63, and which in the south appears earliest in the Taiwan Neolithic Dabenkeng culture (~3000-2500 BCE), but does not appear in mainland southern China until post-Neolithic times. This is incorrect – there is Neolithic millet in southern China, and mainland Southeast - Vietnam and Thailand.
Reply：We thank the referee for making us aware of this archaeological context. We have removed the mention of millet from the Abstract. We have rewritten the later statement to be more conservative and to be fully consistent with current understandings about the dates of arrival of millet in the south as follows: “A speculative possibility is that this ancestry was carried by cultivators of foxtail millet which was domesticated in the north by ~8000 BCE[42], and which in the south appears relatively early in the Taiwan Neolithic Dabenkeng culture (~3000-2500 BCE).”

2. Line 116 states: Yamnaya Steppe pastoralist ancestry, which was likely a vector for spreading late-proto-Indo-European languages
Linguistically, there can be no such thing as a late, as opposed to an early, proto-language on the scale of a whole language family. PIE was the first split that can be reconstructed (Anatolian vs the rest, presumably), and all else is later than PIE, including no doubt the Yamnaya.

Reply：We appreciate that the term "late-proto-Indo-European" is a concept that not all readers accept. Since the term is also not necessary for our narrative, we no longer mention it.

3. Line 132 seems to suggest that the IE language family originated in East Asia. Surely not?
Reply：The revised sentence reads: "East Asia was one of the earliest centres of animal and plant domestication, and harbours an extraordinary diversity of language families including Sino-Tibetan, Tai-Kadai, Austronesian, Austroasiatic, Hmong-Mien, Indo-European, Mongolic, Turkic, Tungusic, Koreanic, Japonic, Yukaghiric, and Chukotko-Kamchatkan[1]."

4. Lines 151 to 162 suggest that Tibetans have a separate Pleistocene genetic ancestry from Yellow River Neolithic farmers. So how does this square with a Yellow River origin for Sino-Tibetan languages? Lines 196-197 go on to state 'Our finding that the shared genetic signal between Tibeto-Burman and Han Chinese is associated with Upper and Middle Yellow River farmers who lived ~3000 BCE supports the 'northern origins hypothesis'; we estimate this group contributed ~84% of the ancestry to Tibetans and ~59-84% to modern Han Chinese.'
I am not suggesting anything is wrong here, but I find the wording confusing.
Reply：What we were trying to say was that Tibetans are a mixture of one lineage that is closely related (within the Neolithic time frame) to Yellow River farmers and tracks the spread of Sino-Tibetan languages, and another lineage that is deeply divergent and likely reflects the pre-farming population of Tibet. In our revision, we removed the Introductory section in which this confusing statement appeared; nothing is lost as the content reappeared later on.

5. Lines 241-245 state: we have no other evidence of Turan-associated ancestry in the region until two millennia later in the Sagly culture of the 6th to 3rd centuries BCE who have cultural connections to the Scythians/Saka who are known to derive some of their ancestry from Turan.
Why cannot this Turan ancestry be related to the Tocharians? If it was Scythian (Iranian?), then the Sagly culture was relatively close to the Tocharians in date and language, even if not in space. Tocharians have no archaeological or genetic identity, so relating them genetically to Yamnaya just because they had an Indo-European language seems to me to be jumping the gun. I do not doubt that Afanasievo people came from the western steppes, but they lived more than 1000 km north of the Tarim Basin, and 3000 years before the existence of the Tocharian Buddhist and commercial documents. If there is Turan ancestry in the general region of the Tarim Basin, why does the Afanasievo ancestry have to trump it?
Reply：Thanks for raising this possibility about Turan ancestry being the source of Tocharian languages; this hypothesis is in fact contradicted by the data, and in our revised manuscript we explain clearly why this so, writing: "for the two individuals with the most West Eurasian-related ancestry (Xinjiang_EIA_Shirenzigou_1C) all fitting three-way models include Russian Afanasievo (71-77%) (Figure 3, Online Table 24), and the total ancestry from the two other West Eurasian-related groups that can fit in small proportions in such models is always <9% (Online Table 24). Languages usually spread through movements of people[47], and thus these result adds weight to the theory that the Tocharian languages of the Tarim Basin spread through the migration of Yamnaya descendants to the Altai Mountains and Mongolia (in the guise of the Afanasievo culture), from whence they spread further to Xinjiang[4,5,6,46,48,49]. These results are significant for theories of Indo-European language diversification, as they increase the evidence in favour of the hypothesis that the split of the second-oldest branch in the Indo-European language tree occurred at the end of the fourth millennium BCE[46,48,49]."

5. References to indigenous peoples in Taiwan should call them Formosans where possible, rather than Taiwanese, who are of course Chinese both in language and in their recent mainland origin.
Reply：We have changed “Taiwanese” to “Formosans” as suggested.

6. I was gratified to see that this paper supports the general view that major movements of early farming populations were very significant in human prehistory, particularly in the case of the language families indigenous to China, both northern and southern. I was a little surprised to see the Transeurasian hypothesis being demolished, especially given that linguist Martine Robbeets is one of the authors. What is her view on this? The paper does not say.
My own view on Transeurasian is that Neolithic Liao Valley population expansions around 6000 years ago, as with the Yellow River, were confined to areas where millet farming was significant, as in present day northeastern China, including inner Mongolia, Korea, and the Liao Valley. Existing Transeurasian subgroups are of much more recent origin than Proto-Transeurasian (if it existed) and may reflect language shift on the part of pastoralist populations on the fringes of the most productive regions of crop agriculture.
If you want my honest opinion, this is also how the Yamnaya originated!
Reply: Co-author Martine Robbeets has reviewed our revised manuscript, has added some references, and supports publication.

Referee #3 (Remarks to the Author):

Response of Reviewer 3 to “Point by point response to referees for manuscript 2020-03-04062A”

As a a referee 3, I would like to react to the point by point response of the authors of 2020-03-04062A and the resubmitted manuscript itself. The authors have considerably improved the manuscript. I am especially happy that the contamination issues (a major flaw) has been remedied by removing problematic samples from the analysis (but not always from the counts, see below). I had however other major concerns regarding the analysis and not all of them have been addressed satisfyingly. In the following text I am commenting on the responses of the authors where a comment of mine is needed. All the other responses of authors I find fully or mostly satisfying and resolved.

Before I comment on specific changes, I would like to mention the rearrangement of the manuscript. This indeed helped a lot. The manuscript is much easier to follow, the major findings are highlighted and questions are posed and grounded in the previous literature. I really think this will help readers considerably. It should be noted that as a whole the article is still rather disconnected but given the variation of data, in all probability this is the best that could have been achieved. I find it peculiar that the introduction contains not only questions but also answers and that the state of the art of ancient genomics in the region is not introduced separately to give credit to authors of previous studies and clearly distinguish the findings but these references are intertwined with the discussions throughout the article and it might be better for readability this way. The only major thing that was not covered in the resubmission is to discuss potential alternative explanations and pitfalls – the manuscript is mainly a summary of the results. The rewriting of the manuscript suggests that the authors invested considerable effort into this resubmission and it is just a pity that some parts of the analysis were not treated the same way.

The specific comments:

(1)

My original review: Title: Given the ambiguity of the some results and unequal sampling over vast regions and chronological periods, it is not appropriate to title the manuscript as being able to comment on the formation of the populations in the whole region. Yet I see that the lack of a result connecting the samples makes finding a better title difficult.

The response of authors: We have changed the title to a more careful “Insights From Ancient and Modern DNA into Human Population Formation in Eastern Asia”.

My comment to the response: This is only a small cosmetic change, just slightly weakening the previously over-stated title. The point still stands. Given the relatively small and unevenly distributed sample size (132 usable samples covering thousands of years and scattered over the whole East Asia) and disconnected results, the authors should be more specific.

Reply：We revised the title to “Genomic Insights into the Formation of Human Populations in East Asia” which we feel is a maximally effective summary given the 75 character limit.

(2)

My original review mentioned the severe contamination issues.

Among others authors responded: - We do not remove 16 “QUESTIONABLE” or 10 additional low coverage samples with 5000-15000 SNPs on the nuclear targets from the list of reported individuals paper, as they are valuable data that generally have high proportions of authentic ancient DNA and provide reliable information such as sex assignment or mitochondrial haplogroup. However, we do not use them in any of them in our main population genetic analyses. The column “Included in population genetic analysis (also indicated if included in Figure 6)” indicates which samples are included in the main population genetic analysis; in addition to the 26 mentioned above, we also exclude from our main population genetic analyses 10 individuals who are detected genetically as first degree relatives of another higher coverage individual in the dataset.

Reviewer 3: If the authors want to additionally report the samples that they call questionable (I agree with the reasoning but I worry if these will not be by mistake used by others), they should remove them from the sample counts in the abstract and manuscript so these numbers are not inflated (as they could mislead potential users of the dataset, or, very unfortunately, be used as some mark of quality and support for the conclusions).

Reply：Following the referee’s first round of comments, we filtered out and do not report at all samples with major evidence of contamination.

Our use of the term “Questionable” for 16 of the remaining samples probably raises too high a level of concern for readers about the data quality for these samples. The samples that we are calling “Questionable” actually all have ancient DNA authenticity metrics that would have caused them to be included in the main analysis dataset in many other published ancient DNA studies; thus, it would have been reasonable to include them in our main analyses and we are just being maximally conservative here. For this reason, we think that it is appropriate to state that we are newly reporting 167 individuals and indeed that it would be inaccurate to state that we are reporting fewer samples. Nevertheless, to reduce confusion, we have revised the section to:

“We newly report data from 167 individuals (Figure 1, Online Table 1): from Mongolia 83 between ~6000 BCE to ~1000 CE, from China 11 from a ~3000 BCE site in the Yellow River Basin, from

Japan 7 Jomon hunter-gatherers from 2500-800 BCE, from the Russian Far East 18 from the Boisman-2 cemetery at 5400-3600 BCE as well as an individual at ~1000 BCE and another at ~1000 CE, and from two sites in Taiwan 46 individuals spanning 1400 BCE - 800 CE (Online Table 1). For analysis we focused on 130 individuals after excluding 16 with evidence of low but non-zero contamination, 10 with 5000-15000 SNPs covered, and 11 that are close relatives of another higher coverage individual in the dataset (Extended Data Table 2).”

(3)
The original review: Cluster assignment and number of samples per group
The number of samples in the study is large, however given the size of the region and the chronological variability, this means that at times, the authors used only a few samples to represent whole populations at various time points. That, by itself, is standard when there is no other data available. However, this might result in an increased number of alternative interpretations of their analysis and the authors do not discuss this in sufficient detail. This is even exaggerated by the use of low coverage hence pseudohaploid data that further limit the amount of the data available for the analysis. While they mention the scarcity of their data regarding the interpretation of qpGraph results, it is not mentioned elsewhere. For example, the Mongolian clusters and the interpretation of the different waves of the Yamnaya-related admixture is dependent on the grouping of populations composed of a few individuals. While I do trust their qpAdm analysis (the Afanasievo being in the outgroups when Sintashta is the source indeed points to some later gene flow), I am hesitant to agree that composing the groups based on genetic similarity and then treating them as populations even though they rather overlap chronologically is a good course of action. This analysis (Figure 3) should instead be performed on an individual basis. A similar point is actually argued by the authors themselves in another part of the manuscript: in the case of Heishui_Mohe samples (see Supplementary Section 2), they refuse to make claims regarding the West Euroasian admixture in this population because the two samples they have from this period fall at different locations in PCA plot. Yet, if they would sample only one of these individuals, they would be confident (as the Mongolian example above suggests) to make completely opposite claims depending on which individual was sequenced. That demonstrates the dangers of overinterpreting results from a very low number of individuals and this needs to be addressed in the manuscript whenever this is the case. It should be also noted that the authors generally disregard a hypothesis that individuals that do show different genetic patterns (on PCA for example) still could be from the same group but that such a group is in a process of admixture (they form groups based on genetic similarity). Additionally, they repeatedly favour hypotheses of demographic events as migrations (“waves”) over continuous gene flow that are actually quite likely.

The response of the authors: We agree that with a small number of individuals from any ancient group, it is impossible to make a confident assessment about the full spectrum of genetic variation in that region at that time. That said, a single individual or a few individuals can be very important, as they document the presence of a particular kind of ancestry in a place at a given time (proof-by-example). We have revised our manuscript everywhere to not overgeneralize from the very small numbers of individuals we have from some archaeological contexts; in contrast, when we have substantial numbers of individuals from particular contexts and observe either homogeneity (as in ancient Taiwan) or significant variation (as in the Afanasievo culture of Mongolia), we highlight it.

We agree with the referee that the individuals showing different genetic patterns still could be from a group that is in a process of admixture. For example, this is likely the case at the Kurgak Govi site where the individuals have different proportions of Western Siberian Hunter Gatherer-related and East Mongolian Neolithic-related ancestry and are not consistent with forming a clade but instead are on a cline. We now make this point explicitly in our revised manuscript.

In our revised manuscript, we no longer discuss heterogeneity of the Heishui Mohe samples because we removed one of the two from the dataset due to evidence of contamination.

In Online Table 8, we now have reported qpAdm results for each genetically homogeneous cluster (where the genetic homogeneity is assessed formally as described in the new Online Table 8G), as well as on individuals that cannot be clustered.

We have now also passed through the manuscript making sure the language does not favor pulse migration scenarios over continuous gene flow when both are plausible.

Our grouping of the individuals is not only based on genetic similarity, but based on time period and cultural associations, then further by genetic cluster which in the Mongolian samples we designated by number (our group names thus have the format "<Country>__<Time Period>_<Cultural Association If Any>_<Genetic Cluster>). We think that this naming scheme is in fact state-of-the-art and reflective of suggestions in the recent literature (e.g. the Eisenmann et al. Scientific Reports 2018 paper discussing naming of genetic clusters) and we continue to use this approach.

Reviewer 3: I agree that the grouping is not based only on genetic clustering but it is ALSO based on genetic clustering which is a highly problematic issue when this is then a basis of another analysis. Of course, it is really perfect to use this naming in the text as a label for certain samples that where shown to be similar. However, then applying an analysis according to these groups, especially if the groups are so so very small (1-3 individuals) is a bit pointless. Of course that these groups differ, that is how they were made! And it does not say anything about overall composition of the population at one point of time and space. Either the authors should group by geography+chronology+culture or they should analyze the samples on the individual basis. The latter might be superior to show the variability in the populations (the individual mixtures are not dissimilar to ADMIXTURE graph), the former can be used to comment on the overall makeup of the populations that in turn is interpretable as formation of these populations. An analysis by clusters may be of interests if clusters are large and/or there is some suggestion of a genetic barrier at the site.

Additionally, the authors even remove outliers that do not fit with their expectations for population genetic inference: "We manually curated the data using ADMIXTURE and EIGENSOFT to identify samples that were outliers compared with other samples from their own populations. We removed seven individuals from subsequent analysis in this way; the population IDs for these individuals are prefixed by the string "Ignore_" in the dataset we release." (as added in the newly resubmitted text). A few individuals that are found on the same place and dated to the same time are not necessarily "outliers" or a cluster to be analyzed as a population, they are just from a heterogeneous population and that is not a bad thing, it is actually something interesting and worth discussing. I do not see the need to create these uniform clusters, especially when they contain only 1-3 samples.

*Reply：We have plotted the results by "**geography+chronology+culture+genetic**"cluster as our philosophy is that when two individuals are temporally, geographically, culturally, and genetically consistent with being from a homogenous group the data should be combined. However, we agree* *that it is important to study the results by individual, and also to present analyses in which samples are considered individually, for example in PCA and ADMIXTURE..*

We think the referee may have misunderstood the section in the Methods where we discuss removing outliers. The outlier removal is for modern individuals (not ancient individuals). We have clarified this in the subtitle for this section, which now reads "Procedure for combining new Affymetrix Human Origins genotyping data on modern individuals with previously published data on modern individuals." We have not removed any ancient sample due to their being outliers. We clarify this in the revised sentence: "We removed seven present-day individuals as outliers from subsequent analysis; the population IDs for these individuals are prefixed by the string "Ignore_" in the dataset we release (for analyses of ancient individuals, we do not remove outliers)."

(4)
The original review: Figure 2: this is a very crucial figure for the text: among others, it should clearly demonstrate the grouping of the individuals. This is however very difficult to see at the moment. It would help if the authors added another figure with simplified legend and visually marked the individuals belonging to their clusters (or make other adjustments to the figure). I also suggest that more PCAs containing only samples from similar periods would be very insightful (instead of projecting all ancient samples to so many modern samples).
The response of the authors: We have simplified the legend to make the figure more readable, and included polygons highlighting the clusters of ancient individuals. We have added PCAs containing only samples from similar periods into Supplementary Information section 2:

Overview of genetic substructure.
Reviewer 3: I thank the authors that they added the Supplementary Figures and improved the main Figure (that changed due to removal of the contaminated samples I presume). However, I was suggesting to attempt to do PCA without projecting (as stated in the original review). On reflection, it is likely that the authors cannot do that due to the low coverage of their data. It should be noted that the period PCA figures with projecting are not much of use (and if the authors want to remove them I fully support that decision).
Reply：We have rerun the PCA without projecting the ancient samples by removing the"lsqproject: YES" and "poplistname" options from the parameter settings in smartpca. We show the resulting plots in Supplementary Information section 2: Overview of genetic substructure, and have removed the figures with projection from that section.

(5)
The original review:Figure 6: This figure lacks any labels on the x axis and I have some issues with the median of the first group, it rather looks like that the Fst is only a bit higher than in the Neolithic but one population is quite different. The caption should also mention in which table are the values underlying this figure placed.

The response of the authors:The revised caption specifies explicitly that the x axis represents pairwise Fst.

In the first group "Before Neolithic Farming", the Fst values between Japan_Jomon and other populations are quite large. We have provided the Fst values in tabular form in a new Online Table 7.
Reviewer 3: I am sorry I mentioned the first group in the my comment. I meant the difference

between the medians of the second and the third group that might be driven by an outlier value and not informative of a general trend. This was however in the original Figure 6 and the newly resubmitted Figure contains much more points for some reason and it does not seem to be affected by this.

Reply：On reflection, we decided to remove this figure from the manuscript following the comments of both Referee #1 and #3.

(6)

The original review: Figure S1: This figure of Fst results is also crucial for the clustering and the clusters should be marked very clearly. I must say I am a bit worried that the authors use this figure for interpretation of some results while they themselves notice discrepancies for other populations (Papuans, as mentioned in the SI).

The response of the authors: We have added cluster names onto the figure. We want to keep this supplementary figure because it shows a striking pattern that genetic clustering generally corresponds to linguistic and geographic classifications, but we also caution in the supplementary sections that some of the clustering can be deceptive. We discuss the Papuan and Australian case explicitly and also highlight other cases in which the position in the tree does not correspond perfectly to linguistic clusters or geography.

Reviewer 3: I am a bit confused by the reasoning here. I of course noticed that some populations (Papuans) are discussed as an example of the weaknesses of this plot but that is exactly my point – when it agrees with what the authors expect, they believe the results, if it does not, they do not believe them. Either this analysis is solid or it is not – if it is affected by admixture here, maybe it is affected elsewhere and actual position on the tree of some other samples should be different too. Also, in this part authors continue to mention (as observed in my original review) that they have correlations with language and geography and no correlation test was performed between the tree and languages and geography (Mantel was performed – great!) . That is just a loose terminology though.

Reply：We are glad to have added the Mantel tests after the referee’s first round of comments, which provide a statistical basis for our statement about correlation of language and geography.

(7)

My original review: Ethical concerns

I have concerns regarding ethics and sampling of the ancient individuals. While there is no formal requirement for the authors to contact local communities and get their approval for the sampling of the individuals discovered at archaeological excavations, it is certainly highly desirable and especially so if the ancient DNA studies (e.g. from adjacent regions and from some of the same authors) can be considered problematic in this regard. In my opinion, a publication aspiring to be at a high-level journal has to address this in the supplementary material and in the ethical statement (required by Nature) to make sure there are no ethical issues with the material (both from the research community and from the public). It should be noted however that the authors got proper ethical approvals for contemporary samples and with the ancient samples they did not breach any standard and that authors of many other publications do not go through the process of getting approval from the communities etiher. But it would result in avoiding any dangers of bringing further distrust to the field of human population genomics.

The response of the authors: We have adding an Ethics Statement covering both the modern and ancient samples to the beginning of the Method section:

"The modern sample collection was carried out in 2014 in accordance with the human ethical research principles of The Ministry of Science and Technology of the People's Republic of China (Interim Measures for the Administration of Human Genetic Resources, June 10, 1998) and genotyping was reviewed and approved by the Ethics Committee of the School of Life Sciences, Fudan University. Study staff informed potential participants about the goals of the project, and individuals who chose to participate gave informed consent consistent with broad studies of population history.

The ancient samples were collected with the permission of the custodians of the samples, who are the archaeologists or museums in each of the countries for which we analyzed the data. We applied a case-by-case approach to obtaining permissions for each set of samples depending on the local expectations as these vary by region and cultural context. Every newly reported ancient sample in this study has permission for analysis from custodians of the samples who are co-authors and who affirm that ancient DNA analysis of these samples is appropriate. For most samples, we prepared formal collaboration agreements to explicitly list the ancient DNA work being performed by our team. In other instances sample custodians who are co-authors determined that generation and publication of ancient DNA data was covered under their existing permissions for sample analysis, and so determined that new sampling agreements were not required."

Reviewer 3: As stated, the authors followed the standard procedures and now added statements about that to the text which is great as it provides an example to others. It would have been nice to go beyond that and really involve the communities concerned not just officials (a ministry), though. I see that it is difficult to do so post quem and I will just keep hoping that no communities or their diasporas would find the results or the discussion in this study damaging or hurtful (I am not aware of any such issues but then again I am not, for example, a Tibetan).

Reply：We affirm confidently that in the course of the sample collection and manuscript writing we proactively took into account perspectives of Indigenous communities and concerns about how our work might affect community self-perception; this approach was not clear from the previous submissions of our manuscript. We discuss this in more detail in what follows.

(1) For the analysis DNA samples from modern populations, we sought to make our work sensitive to Indigenous community perspectives in two ways:

(a) First, a community consultation process was an integral part of sample collection. This process is described in our revised manuscript but that we did not describe in the last version. For each minority group, community representatives affirmed community support for the study through a signature or thumbprint on a form summarizing the community consultation process after discussions with the scientific team about the goals of the study (these forms were completed between November 10 2014 and December 10 2014).

(b) Second, multiple Indigenous people were involved in the collection of modern DNA samples and are co-authors of the manuscript, which was an important component of our approach for making the manuscript sensitive to and respectful of Indigenous perspectives. For example, two co-authors are people of Indigenous Tibetan ancestry, including two at the Tibetan University of Nationalities: Longli Kang (who has one parent who is Tibetan) and Nini (who co-organized the Tibetan DNA sample collection). The co-authors also include Indigenous co-authors from several other minority communities sampled in this study.

(2) For the ancient DNA samples, we also sought to make our work sensitive to Indigenous perspectives when these perspectives were relevant.

For some regions for which we obtained DNA such as the southern islands of Japan and the Russian Far East sites we are not aware of modern communities with traditions of biological or cultural connection to the ancient remains. For other regions such as the Upper Yellow River Chinese or Mongolia, the modern nation-states in which the ancient individuals lived are plausible modern inheritors of the cultural and genetic heritage of the ancient people. In contrast, in Taiwan where the Han Chinese have exerted political control over the island since the 17th century, it is not as clear that Indigenous perspectives are fully represented by the formal permission we received to sample. In our manuscript, we sought to represent Indigenous Taiwan / Formosan perspectives through review by two co-authors with Indigenous Formosan ancestry. Co-first author Hui-Yuan Yeh whose excavation yielded the largest single source of ancient samples in this study has one parent from the Paiwan Indigenous group. Hana Looh was the excavation leader for the Bilhun Hanben site, and is the local community leader for the Indigenous Ami group.

In the revised manuscript we have addressed points (1) and (2) both in short statements in the main manuscript and in a detailed Ethics statement in the Methods. Here are the sections in the manuscript that address these topics:

1. Newly revised first sentence of the paragraph that describes the modern DNA sampling:

"For modern individuals, we collected DNA from 383 individuals from 46 populations from China (n=337) and Nepal (n=46) who provided informed consent for broad studies of human population history; we also carried out community consultation with minority group leaders as an integral part of the consent process (see Ethics Statement)."

2. Newly revised first sentence of the paragraph that describes the ancient DNA sampling:

"For ancient individuals, we obtained permission for analysis from sample custodians, following protocols to minimize damage to skeletal material and including members of local minority groups as part of our study team when there was a plausible cultural connection between modern communities and ancient individuals (see Ethics Statement)."

3. Greatly enhanced Ethics Statement at the beginning of the ethics section:

"Ethics Statement

The modern sample collection was carried out in 2014 in strict accordance with the ethical research principles of The Ministry of Science and Technology of the People's Republic of China (Interim Measures for the Administration of Human Genetic Resources, June 10, 1998). Our sample collection and genotyping was further reviewed and approved by the Ethics Committee of the School of Life Sciences, Fudan University (October 22, 2014). Study staff informed potential participants about the goals of the project, and individuals who chose to participate gave informed consent consistent with broad studies of population history and human variation and public posting of anonymized data. There were no rewards for participating and no negative consequences for not participating; all participants signed or affixed a thumbprint to the consent form reviewed by Fudan University. An important principle of our study was to ensure that the research was underpinned not only by individual informed consent, but also support from community representatives sensitive to local perspectives, and thus we carried out community consultation with minority group leaders or village *leaders as an integral part of the consent process. For each minority group, community representatives affirmed community support for the study through a signature or thumbprint on a form summarizing the Community Consultation process (these forms were completed between November 10 2014 and December 10 2014). Co-authors of the manuscript who were culturally Indigenous and in some cases were legally registered as members of minority groups specifically reviewed the manuscript's discussion of population history to increase sensitivity to local perspectives. Specifically, co-author L.W. is a Tai-Kadai speaking Zhuang person from Guangxi in southwest China; R.S. is from Nepal; and L.K. and N. are based at the Tibet University for Nationalities, and N. is an Indigenous Tibetan. We emphasize that Indigenous and community narratives co-exist with scientific ones and may or may not align with them. Indigenous ancestry should not be confused with identity, which is about self-perception and culture and cannot be defined by genetics alone.*

The ancient samples newly reported in this study were collected with the permission of the custodians of the samples, who are the archaeologists or museums in each of the countries for which we analyzed the data. We applied a case-by-case approach to obtaining permissions for each set of samples depending on the local expectations as these vary by region and cultural context. Every newly reported ancient sample in this study has permission for analysis from custodians of the samples who are co-authors and who affirm that ancient DNA analysis of these samples is appropriate.
For most samples, we prepared formal collaboration agreements to explicitly list the ancient DNA work being performed by our team. In other instances, sample custodians who are co-authors determined that generation and publication of ancient DNA data was covered under their existing permissions for sample analysis, and so new sampling agreements were not required. Going beyond what was formally required, we also ensured that the presentation of the scientific findings was sensitive to local perspectives from the regions from which the skeletons were excavated. For some regions for which we obtained DNA such as the southern islands of Japan and the Russian Far East sites we are not aware of modern communities with traditions of biological or cultural connection to the ancient remains. For other regions such as the Upper Yellow River Chinese or Mongolia the modern nation-states in which the ancient individuals lived are modern inheritors of the cultural and genetic heritage of the ancient groups. In Taiwan, in addition to obtaining formal permission for sampling from government institutions, we sought to ensure that the presentation of our results was sensitive to the perspectives of Indigenous Formosans who plausibly descend thousands of years ago from groups related to those from which we report data. The existence of at least sixteen non-Han Chinese Indigenous groups in Taiwan makes it difficult to connect particular sites to specific modern ethnic groups for prehistoric sites older than four hundred years, and it is rare for local communities to express connections with prehistoric sites. Nevertheless, two co-authors with Indigenous Formosan ancestry or cultural affiliation to these groups specifically reviewed the discussion of the Taiwan results to increase the sensitivity of our study to Indigenous group perspectives. H.-Y.Y. who is co-first author of the study has ancestry from the Paiwan Indigenous group. H.L. was the excavation leader for the Bilhun Hanben site and is the local community leader for the Ami group, whose present-day culture shows some similarities to the material culture of the site."

(8)
My original review: Archaeological information
Given the high number of regions and periods covered, the archaeological supplementary material is rather long, containing information essential to the manuscript. I especially like the care taken to prepare very informative Online Tables (I applaud the Online Tables 1 and 2 especially). However,

the quality level of SI is inconsistent: some archaeological sites are described with a lot of attention to detail, some are barely mentioned (e.g. compare the information about 18 individuals from Boisman-2 site and one individual from Nevelsk 2 site). The information about the archaeological context of the ancient individuals should be added to all sites (any graphical material would also be welcome). Also there is quite some lack of references (e.g. no reference at all for the information provided for Slab Grave culture graves, Mongol graves or one reference for Xiongnu burials). I understand that it is difficult to provide information for all the sites and individuals because of the number of samples and the diverse nature of the region and periods. But since this is essential for the grouping of the individuals (a basis of most of their analysis), this cannot be taken lightly. The authors should provide clear references for all the sites, in case some sites are not published, then the authors should provide enough information for other scholars to evaluate them (archaeological documentation of the finds and the site itself). Otherwise, the samples can be used incorrectly in future analysis of researchers using this dataset and it might be also hard to evaluate the current analysis.

I also have some reservations about the anthropometric descriptions used for interpretations of similarities of populations in this part of SI, they are rather dated (e.g. Slab Grave culture: “anthropological typing of this culture suggests they are ‘Mongoloid’ or similarly in the description Boisman site).

The response of the authors: We have revised the manuscript to enrich a number of the archaeological discussions that were previously thin. We found it difficult to achieve homogeneous detail in the discussion of the archaeological content—in particular because of the limited archaeological information available from some sites especially in Mongolia—but the revision is better in this respect.

We have reduced the morphological descriptions of skeletons as they are not critical for our grouping of samples, and we no longer use terms like “Mongoloid”. We note that in the countries where the physical anthropological analyses are done (Mongolia and Russia especially) these types of descriptions are generally viewed as valuable and meaningful and are widely used; in fact, the experience of the genetics community is that in Central Asia and Siberia and East Asia these physical anthropological classifications often correlate strongly genetic findings. Thus, fully excluding them from the descriptions completely would devalue the research of local scholars many of whom are co-authors, and so in some cases we do cite and briefly discuss observations based on these kind of data without using terms that are dated.
Reviewer 3: The term “Mongoloid” and “Caucassian” is still used in the Supplementary information at line 455. It is worrying that the authors do not realize the dangers of such terminology. As was shown previously in anthropological literature, many of the skeletal traits used for such classifications are distributed along geographical clines and thus such “classes” obviously correlate with genetics. However, the act of classifying individuals on the basis of these continuous traits into groups is just not valuable in anyway (yes, it is done today but in all countries where it is done, it is only for continuity with past studies that were written in different times). What is more, the classification was misused in the past for racial theories and continues to be problematic and hurtful for some communities. If the authors want to correlate skeletal measures with genetics, they should do so properly – using the actual measurements. Even then I do not really see the point if the goal is to study population formation and movement as these skeletal measures are just (weak) proxies for genetics. But why not. Otherwise, the SI did improve.

Reply：We removed the terms “Mongoloid” and “Caucasian” from the supplement. The two relevant *sentences now read: "Physical anthropological studies have identified the people buried in these graves as having more morphological affinity to West Eurasian populations in the west and East Asian populations in the east, a suggestion that is qualitatively supported from a genetic point of view as ancestry related to Sintashta and Andronovo people is more evident in the west than in the east. Russian anthropologist I. Gohman identified similarities between these populations and Afanasievo people of the Russian Altai and Minusinsk Basin."*

(9)

The original review: Different pipelines

Samples have been analyzed with 3 different pipelines. Already a part of the data has been processed differently in the lab and I wonder why the authors additionally bias the analysis by using different pipelines. It has been shown previously how even a small amount of bias can severely impact genetic inferences (Günther & Nettelblad 2019 PLOS Genetics) and here they treat the data with different versions of the same tool or even completely different tools. While the authors claim that the difference in results between the pipelines is not large and cite the analysis of Fernandez et al. that showed that some of the conclusions in that paper are not biased while using different pipelines, I strongly disagree. The f4 statistics analysis of Fernandez et al. between the pipelines can hardly be generalized to other studies and samples and even in that study, it is hard to generalize for all the results. Furthermore, the Wuzhuangguoliang samples were treated with a third, completely different pipeline that has not been compared to the other pipelines at all..

Reanalysis of the data with the same pipeline should not be a major issue for the authors: the pipeline is scripted and the scripts are available and the data are quite low coverage and hence rather small. Therefore, all samples have to be analyzed as similarly as possible.

The response of the authors: In our revised manuscript, we have addressed these issues in two major ways:

1. We have completely reprocessed the Wuzhuangguoliang samples using pipeline 1 to reduce the three pipelines to two comparable pipelines. As part of the Wuzhuangguoliang reprocessing, we have also applied sample-specific filters which have successfully removed evidence of contamination in these samples. This is a major improvement for these samples, and a new Online Table 9 describes this work and processing.

2. We have added an entirely new analysis in which we systematically compare pipeline 1 and pipeline 2 on sets of samples where we have individuals processed in both ways. Consistent with the findings in Fernandes et al. Nat. Ecol. Evol. 2020, we observe no systematic biases in symmetry-f4 statistics between samples from the same context processed using pipeline 1 compared to pipeline 2. As an example, the Z-score of f4(Boisman_MN_pipeline1, Boisman_MN_pipeline2; Taiwan_Hanben_pipeline1, Taiwan_Hanben_pipeline2)=0.348. We mention these analyses in the revised methods and in Online Table 6Q.

On reflection, we have decided not to reprocess all of our dataset using a single pipeline. There is heterogeneity in ancient DNA data in many ways, including

1. between shotgun data and in-solution enrichment data
2. between samples processed with profoundly different library preparation methods (e.g. with and without UDG-treatment)
3. between samples sequenced on different sequencing instruments

4. between data generated at different laboratories
5. between data processed with different bioinformation pipelines (the issue the referee raises)

For cases 1-4—which together are much more problematic than issue 5 in our experience—it is impossible to achieve homogeneity. But the truth is, we cannot even reprocess all samples using the same bioinformatic pipeline (issue 5), because the data relevant to this study is not just our newly reported data but also the previously published data with which we re-analyze it which inevitably was processed using different bioinformatic pipelines. While it would be a useful project to reprocess all the world's ancient DNA data from multiple laboratories from scratch using a uniform bioinformatic pipeline (and indeed we are currently hiring a person who will have this as a task), this would be a uear-long project that is not the focus of our current work, and we have demonstrated here that it is not necessary to obtain robust results for this particular set of samples.

Reviewer 3: I am glad that the authors decided to reduce the amount of artificial bias by lowering the number of pipelines used. It is also nice that they added f4-statistics to assess the bias. I however do not agree that this is conclusive in any way about the non-existence of such bias as the test is limited to only some samples and some comparisons. While the presence of a significant systematic results in this test (Online Table 25 I think – not 6Q as mentioned in the rebuttal letter) would indeed be a proof of a bias, the absence of such significant results is definitely not a proof of the absence of a bias. This is a very important distinction. The only conclusion that can be drawn from this analysis is that a bias could not be found between those exact samples in this exact setting. This does not exclude a potential structural biases resulting in differing results between other samples or other groups of samples. The only way the authors could prove the absence of a bias generally would be to reanalyze all the samples and compare the results, especially the analyses that matter for their conclusions (which amounts to essentially reanalyzing the data as they should be done in the first place). All kinds of analysis (even those based on f-statistics, especially sensitive might be qpGraph) might have been influenced by this in an unknown manner. I was worried when the f-statistics test was interpreted this way in literature (Fernandez et al. 2020) and further propagating this misinterpretation (over-generalization) would be most unfortunate and it can be even further misused so very easily (e.g. by selecting to show only those comparisons that are indeed insignificant - which I trust the authors did not do in this case but can be easily done to support any differences in pipelines following the lead of this manuscript and Fernandez et al.). If the authors decide to keep this "bias analysis", they should definitely avoid any suggestions that the test proves the absence of a bias, especially expressions like "Indistinguishability of population genetic results" (it is the absence of the artificial shared drift between some particular configurations of some of the samples) and "We verified that results from the two pipelines were indistinguishable from a population genetic point of view" (again it is only from a point of a view of a few particular comparisons).

Reply：We have revised the way we discuss these evaluations, now writing: "To evaluate whether there was evidence that ancient DNA data processed using the same bioinformatic pipeline was artifactually biased to appear similar to each other in f-statistic analysis, we computed statistics of the form f_4(Group1Pipeline1, Group1Pipeline2; Group2Pipeline1, Group2Pipeline2) for all groups for which we had individuals in our main analysis dataset processed by both pipelines (Mongolia_EIA_Sagly_4, Mongolia_EIA_SlabGrave_1, Mongolia_LBA_CenterWest_4, Mongolia_LBA_MongunTaiga_3, Russia_MN_Boisman, and Taiwan_Hanben). For all 15 possible pairwise comparisons, the Z-scores for deviation from zero had magnitude < |2.7| (Online Table 25), which is not significant after correcting for the 15 tests we performed (P=0.11 after applying a Bonferroni correction)."

This is not to say that such a test is without merit, I can imagine that there could be no way for the samples to be made comparable (e.g. different laboratory procedures of destructive nature) and such a test could alleviate some concerns over systematic biases (it really does). However, even then a potential bias would have to be still brought up in a discussion of the result. However, the authors do not discuss what portion of their samples was analyzed with what pipeline in relation to their results or in discussion (the distribution of samples between different pipelines is not random among the groups and populations and some results might be affected more than others). And this all while the heterogeneity of the bioinformatic analysis is just for convenience, not a must.

Additionally, I reject in principle one argument the authors make in order to justify keeping the different pipelines. The fact that there is some heterogeinity in the aDNA data production is by no means an argument to produce more bias in the bioinformatic analysis. And it does not matter that the authors claim that in their experience the different pipelines make usually less of a difference, in my experience it matters. That is connected to another point - there might have been some misunderstanding: I did not suggest that the authors should “reprocess all the world’s ancient DNA data from multiple laboratories from scratch using a uniform bioinformatic pipeline” (as stated in the rebuttal letter), though I strongly agree with the authors that this would be the best course of action and quite a few laboratories now proceed in this manner (exactly for the reasons of producing sound high quality results). I asked in my review that the data produced in this one study are analyzed the same way (it would have been nice if they included a few published samples from the same region but this is not a minimal quality standard). Given the scarcity of the aDNA data in this region, that means that many of the results will then be produced without this potential bias (as the authors themselves state several times, the amount of other data from the region is limited). I do not think that asking for this is unreasonable and I am not convinced that some results comparing populations analyzed with different pipelines are not affected and hence are questionable.

It should be noted that in aDNA literature, it often happens that new samples and published reference samples from different laboratories and publications are compared together without re-analysis. But a level of caution is often expressed in discussions in these cases and most readers from the field do implicitly understand the issue of problematic comparability between different studies. And as the field moves forward more and more attention is given to reducing such biases (see Martiniano et al 2020, Cassidy et al. 2020, Gunther et al 2018). Analyzing samples in the same study differently is a step in a completely opposite direction to this trend and a validation of such an approach by publishing a study with such a bad practice in a high profile journal could set the field back. The statement in the rebuttal letter that suggests that it requires effort to produce comparable data is true but such an absence of full comparability of data coming from the same study undermines the effort of other groups that do spend the time and effort to avoid such issues even between different studies from different laboratories. It even signals that the field of ancient population genomics itself is still not mature and it is embarassing (it would be just unthinkable that controls and cases in clinical studies would be analyzed with different pipelines as some populations in this study are) .
Reply: We agree about the importance of reducing biases in ancient DNA analysis. Nevertheless, the comparison to medical genetic case-control studies is not valid as it implies that it would be possible to achieve a degree of homogeneity in data generation that can never in fact be achieved with ancient DNA. Ancient DNA data are fundamentally inhomogeneous, and we believe that the best approach is to evaluate which sources of inhomogeneity have evidence of biasing the types of analyses we are doing—which we have done in this study (Online Table 25)—rather than carrying out an analysis that

superficially seems to be imposing homogeneity on the data but in fact has no evidence of being a meaningful improvement. Thus, while in principle we could follow the referee's suggestion and re-analyze all the data newly reported in this study using the same bioinformatics processing script, it would take literally hundreds of person-hours to repeat all the analyses, and we think it would not meaningfully increase the value of the dataset for the community while creating an illusion that a significant source of bias had been addressed.

To elaborate on this point, the field of ancient DNA has to contend with major sources of data inhomogeneity not just across studies but within studies. This includes: (a) unavoidable systematic differences in the chemical properties and preservation conditions of DNA from different archaeological sites (for example petrous bones versus other types of skeletal remains, and different degrees of sample preservation resulting in different fragmentation patterns and different levels of contamination that can of course never completely be filtered out); (b) differences in wet laboratory protocols that can be quite profound (for example UDG-treatment and no UDG-treatment, single-stranded versus double-stranded library preparation); (c) differences in data generation after the library is processed (for example, in-solution enrichment versus shotgun sequencing, or single-end or paired-end sequencing, or differences in sequencing instruments); and (d) differences in wet laboratory and bioinformatic processing protocols across laboratories that produced the data being co-analyzed. Next to these issues, slight variations on the computer programs to implement essentially the same bioinformatic procedure (which is the difference between pipelines 1 and 2) are small. In our manuscript and in Online Tables 1 and 2, we aimed to provide a transparent documentation of how processing of samples in this study evolved over time both from the wet laboratory and bioinformatic point of view, and in retrospect we feel that this attempt to be maximally transparent might have made the differences in bioinformatic processing protocols seem larger than in fact they were (in our view the wet laboratory inhomogeneities are almost certainly likely to be more important and from a scientific point of view and raise more concerns about biases due to inhomogeneity and yet the reviewer has not asked us to repeat all wet lab experiments on all samples using the exact same set of protocols). In our revision, we have tried to provide better context to the reader for these issues, writing:

"While these analyses reduce concerns about systematic differences in population genetic analysis driven by changes over time in the software we used to carry out our bioinformatic processing steps, we caution that there are other inhomogeneities in our ancient DNA dataset that also have the potential to affect inferences. Other sources of inhomogeneity include systematic differences in the chemical properties and preservation conditions of DNA from different archaeological sites, (b) differences in wet laboratory protocols including differences between data from in-solution enrichment and direct shotgun sequencing, and (c) differences in wet laboratory and bioinformatic processing protocols across research groups that published the various datasets co-analyzed in our study. The fact that we can obtain fitting models of population history through admixture graph analysis (Figure 2) even in the presence of these differences, and that the admixture graph model also fits when restricting to transversion polymorphisms (Supplementary Information section 3), and finally that our f_4-symmetry tests reveal no significant differences between data generated for this study using wet laboratory and bioinformatic protocols that changed over time (Online Table 25), increases confidence that our inferences are valid even in the presence of inhomogeneities (Gunther and Nettlebad PLoS Genetics 2019)."

In our revision we have chosen not to entirely redo the entire bioinformatic processing of our dataset followed by redoing all population genetic analysis, since we are confident that this will not meaningfully increase the value of our dataset and the reliability of our scientific inferences. To properly address concerns about bioinformatic inhomogeneity would require a far more ambitious re-analysis than the referee suggests, that is, a bioinformatic reprocessing from raw sequences not only of our own data, but also all the other data from published studies of other groups that we co-analyze with our own data. We appreciate the referee highlighting these important issues, and we hope that the better contextualization we provide in our revision and the explicit warnings we now give about the existence of inhomogeneity in ancient DNA data including ours one will be useful to the reader by providing caveats they should take away while interpreting our results. We think the addition of this content and these caveats makes the manuscript better.

(10)
The original review: The authors sometimes use their codes instead of the labels they decided to use, e.g. in lines 239 and 248 where the reader needs to consult supplementary tables to identify the individuals at Figure 3 that are being referenced.
The response of the authors: Everywhere in the text the population names are now consistent with the figures.
Reviewer 3: This is a detail but in both of the lines mentioned above (line 239, now 611 and line 248 now 621), the individuals are described by the skeletal code and the lab code only and cannot be identified in Figure 6 easily without looking into the tables for the population label used. In the revised text, the population labels are mentioned some time before (line 606) but I think it would just be really more clear to have the code and the population label somewhere mention together outside of SI for those samples that are mentioned idividually (maybe in the figure caption or in the same brackets on lines 611 and 621). It was quite confusing to me.
Reply：In our revision, we have eliminated mentions of specific individuals as a way of further increasing accessibility of the manuscript (it also helped us to decreasing the length).

(11)
The original review: Line 321: the authors use term genetic continuity very loosely: if they want to use it they should statistically test for continuity. Similar Y and mtDNA haplogroups and approximately the same spot on the PCA is highly insufficient to claim this.
The response of the authors: Thanks for catching this loose terminology. We did not actually mean to use continuity here, but instead just to state that the individuals were consistent with deriving entirely from the pre-Afanasievo components of ancestry in Mongolia (albeit with significantly different proportions of Western Siberian Hunter Gatherer and East Mongolian Neolithic related ancestry), without any Afanasievo/Yamnaya-related mixture. We have clarified this in our revision.
Reviewer 3: While the authors exchange the expression “The genetic continuity is also evident” (previously line 321) to “The genetic similarity is also evident” (now line 400), immediately in the following sentence and concerning exactly the same populations they bring the term “continuity” back (line 400-401): “The continuity in the Amur River Basin was disrupted..”. The text has therefore not been clarified for this issue at all. What is more this whole paragraph in the resubmitted manuscript now centers around the question of continuity in the region, as a question the authors are addressing (a question stemming from previous works). But the authors in this study actually do not show continuity in their analysis. Again, there are dedicated methods for that and PCA (Figure 2) and ADMIXTURE (Extended Figure 3) they mention to support the statement is not one of them. For PCA, obviously any change in a population not captured by the principal components would not be observable and could completely derail the qualitative assessment of “continuity” – reasons for such

an occurrence could vary, scarcity of the sampling and affinities to unsampled populations immediately come to mind. ADMIXTURE is arguably better but still a failure to observe substructure for these few individuals when the analysis was performed on a large data set would not surprise me and hence I am not convinced. In sum, the resubmitted text is in this regard even worse than the initial submission. And this is not only wrong use of terminology but over-interpretation of results (language would have to be improved and potential pitfalls mentioned) and an incorrect method for a question asked.

Reply: We agree that we needed to provide statistical support for this statement, and we have now carried out a formal statistical test using qpWave to detect if the Amur River populations can be modeled as deriving from a single source as a test for whether there was genetic continuity in this region. We use Amur River populations from the early Neolithic to the Iron Age (AR_EN, Boisman_MN, Yankovsky_IA, and AR_Xianbei_IA) as left pops and use worldwide representative ancient and modern populations (Mbuti.DG, Onge.DG, Russia_MA1_HG.SG, Germany_EN_LBK, Germany_CordedWare, CHG, Iran_GanjDareh_N, Miaozigou_MN, Upper_YR_LN, Russia_Afanasievo, Russia_MLBA_Sintashta, and WSHG) as right populations. The p-value for rank 0 is 0.568928134, supporting Amur River populations from the early Neolithic to Iron Age (AR_EN, Boisman_MN, Yankovsky_IA, and AR_Xianbei_IA) as consistent with deriving ancestry from a same source. We have added this result into Online Table 9.

We have rewritten this section as follows:

"we began by studying our time transect in the Amur River Basin[23]. From the ~5500 BCE early Neolithic individuals and ~5000 BCE Boisman individuals until the ~900 BCE Iron Age Yankovsky culture and 50-250 CE Xianbei culture, those Amur River Basin individuals are consistent with being a clade according to qpWave (Online Table 9). This locally continuous population also contributed to later populations, as reflected in the Y chromosomal haplogroup C2b-F1396 and mitochondrial haplogroups D4 and C5 of Boisman, which are predominant in present-day Tungusic, Mongolic, and some Turkic-speakers, and also in a Heishui Mohe culture individual at ~1100 CE who had an estimated 43±15% Amur River Basin Neolithic ancestry (the remainder modelled by Han Chinese documenting migrations from the south) (Extended Data Fig. 3 and Online Table 9)."

Referee #4 (made no additional comments to the authors)

Reviewer Reports on the Second Revision:

<u>Referee #1 (Remarks to the Author):</u>

I think the authors have sufficiently addressed all of my concerns.

The only thing I would add is I would like to push back a bit on the use of Formosan to describe the indigenous people of Taiwan. Formosa is a name coined by the Portuguese explorers in the 1500s (https://www.taiwan.gov.tw/content_3.php). The explorers and colonialists forcibly named multiple geographical locations throughout East Asia with a Western name, disregarding the native name (e.g. Cape St. Jacques in Vietnam, https://en.wikipedia.org/wiki/V%C5%A9ng_T%C3%A0u). Continue use of the term Formosan to describe the indigenous Taiwanese, without consultation to the community, could be unwelcomed. Instead, the name Taiwan is an internationally recognized geographical location, and has its origin (I believe) through phonetic transliteration from an Austronesian language. Therefore, in light of the ethic concerns raised during the review process, and short of consulting the community for an appropriate term to describe themselves, I think the more conservative approach would be referring to the indigenous people as aboriginal Taiwanese or indigenous Taiwanese. The "indigenous people of Taiwan" could also be OK, though a bit longer.

Referee #2 (Remarks to the Author):

I have no further comment, except for one possible source of confusion:

Line 229: The heading 'Farming and Language Expansions: No Evidence that Speakers of Transeurasian Languages Share Ancestry from West Liao River Farmers' is puzzling, since it is stated later that Japanese and Koreans do share this ancestry (~92% Bronze Age West Liao in the case of the Japanese). Should this heading be changed to read 'Farming and Language Expansions: No Evidence that Speakers of Mongolic and Tungusic Languages Share Ancestry from West Liao River Farmers'?

Apart from this I think the article should be published.

Peter Bellwood

Referee #3 (Remarks to the Author):

Response of Reviewer 3 to "Point-by-point response for second round reviews of submission 2020-03-04062B":

Most issues raised in the second round of reviews were answered to my satisfaction and I will only concern myself with those that did not or not completely. In general, I do believe that manuscript improved over the course for the rounds of reviews considerably.

I am particularly glad that the authors have thought about the effects their study might have for local communities and asked the permission of such communities prior to sampling (though note that the results have not been discussed with them or their representatives). It is also reassuring that there were co-authors of the study that came from some of the relevant Indigenous backgrounds (I am less sure about publishing their minority status publicly and I hope they were not pressured to consent to this in any way whatsoever). Hopefully, these co-authors were empowered to represent their and other communities and considered all ramifications of misuse of genetic inferences in the political sphere. Myself, I would require a sociological/politological consultation to be sure, even if my own ethnic background was concerned. However, this should be left to the autonomy of the relevant co-authors and to the responsibility of leaders of the study that the autonomy of these co-authors was protected from outside pressures.

A few issues that I have raised and with which the authors disagree, I could nevertheless accept despite the difference of opinion. In particular, they insist on using the definition of clusters also with genetic components for further analysis which I find circular but since they also present the individualized analysis in most cases this can be accepted.

There is however one issue that I have still deep concerns about: the authors' refusal to analyze data consistently with the same pipeline. The authors rest their arguments on the facts that 1) there is already some heterogeneity in the data, 2) it takes too long and 3) it might not bring a meaningful difference. I strongly disagree with 1). The authors suggest that since I did not ask them to redo lab experiments, I have no right to ask for data analysis consistency. This logic I do not follow. Of course, it would be great if they also analyzed the samples in the lab consistently but this would require them to invest into redoing a half of the whole study and it might not be even possible for many samples due to the destructive nature of aDNA analysis. They also suggest that to be fully consistent they would have to reanalyze the previously published data – indeed that is what many labs in the field do and the authors must be aware of this. But I did not ask to follow the lead of such labs, I have only asked that they analyze the data newly reported in this study with the same pipeline, using to their advantage the fact that most of their relevant results are in comparisons amongst their own newly reported samples. I could see that the samples produced as whole genomes might be in the initial steps analyzed differently to SNP captures but this is not the case here – samples differ in the pipelines used just for historic reasons. Simply, the fact that there are inconsistencies in the lab analysis of the samples should not mean that researchers should give up on any attempts at consistency whatsoever. As to the point 2), the one claiming that it would take them too much effort. I am quite concerned what kind of analysis practices are in place when rerunning not so many samples (most of them captures, hence quite small files) would take so much time. Computer time will be less then a week on any decent cluster and if the scripting was done reproducibly, running the pipeline should be basically automatic. If other analysis were not done to the same reproducible standard, some might take a bit of time to regenerate but I just do not see how this could take the time the authors suggest. When I first suggested this, it did not occur to me that this was such a big issue, rerunning of the analysis of all samples when there is an improvement of the pipeline is a standard practice. What is however possible and even likely is the last point, 3). The reanalysis indeed might not bring any big differences and quite likely will not change conclusions. Me and the authors can both expect this, especially given that the types of analysis used are usually quite robust. However, none of us can be sure about this. To be honest, I would probably disagree with the authors but not insist on the reanalysis if they wanted to publish in a small journal. However, they aspire to be published in a journal that expects them to follow the highest standard and as such they should in my opinion follow at least the basic bioinformatic practices.

Author Rebuttals to Second Revision:

Referee #1:
Remarks to the Author:
I think the authors have sufficiently addressed all of my concerns.

The only thing I would add is I would like to push back a bit on the use of Formosan to describe the indigenous people of Taiwan. Formosa is a name coined by the Portuguese explorers in the 1500s (https://www.taiwan.gov.tw/content_3.php). The explorers and colonialists forcibly named multiple geographical locations throughout East Asia with a Western name, disregarding the native name (e.g. Cape St. Jacques in Vietnam, https://en.wikipedia.org/wiki/V%C5%A9ng_T%C3%A0u). Continue use of the term Formosan to describe the indigenous Taiwanese, without consultation to the community, could be unwelcomed. Instead, the name Taiwan is an internationally recognized geographical location, and has its origin (I believe) through phonetic transliteration from an Austronesian language. Therefore, in light of the ethic concerns raised during the review process, and short of consulting the community for an appropriate term to describe themselves, I think the more conservative approach would be referring to the indigenous people as aboriginal Taiwanese or indigenous Taiwanese. The "indigenous people of Taiwan" could also be OK, though a bit longer.
Reply：We no longer use the term "Formosan" and use "Taiwan" in its place everywhere.

Referee #2:
Remarks to the Author:
I have no further comment, except for one possible source of confusion:

Line 229: The heading ‘Farming and Language Expansions: No Evidence that Speakers of Transeurasian Languages Share Ancestry from West Liao River Farmers’ is puzzling, since it is stated later that Japanese and Koreans do share this ancestry (~92% Bronze Age West Liao in the case of the Japanese). Should this heading be changed to read ‘Farming and Language Expansions: No Evidence that Speakers of Mongolic and Tungusic Languages Share Ancestry from West Liao River Farmers’? Apart from this I think the article should be published.
Peter Bellwood

Reply：We revised the heading to “Refining the Transeurasian Hypothesis” which addresses the referee’s suggestion and also the formatting requirement to keep headings to 40 characters or fewer.

Referee #3:
Remarks to the Author:
Response of Reviewer 3 to “Point-by-point response for second round reviews of submission 2020-03-04062B":

Most issues raised in the second round of reviews were answered to my satisfaction and I will only concern myself with those that did not or not completely. In general, I do believe that manuscript improved over the course for the rounds of reviews considerably.

I am particularly glad that the authors have thought about the effects their study might have for local communities and asked the permission of such communities prior to sampling (though note that the results have not been discussed with them or their representatives). It is also reassuring that there were co-authors of the study that came from some of the relevant Indigenous backgrounds (I am less sure about publishing their minority status publicly and I hope they were not pressured to consent to this in any way whatsoever). Hopefully, these co-authors were empowered to represent their and other communities and considered all ramifications of misuse of genetic inferences in the political sphere. Myself, I would require a sociological/politological consultation to be sure, even if my own ethnic background was concerned. However, this should be left to the autonomy of the relevant co-authors and to the responsibility of leaders of the study that the autonomy of these co-authors was protected from outside pressures.

A few issues that I have raised and with which the authors disagree, I could nevertheless accept despite the difference of opinion. In particular, they insist on using the definition of clusters also with genetic components for further analysis which I find circular but since they also present the individualized analysis in most cases this can be accepted.

There is however one issue that I have still deep concerns about: the authors’ refusal to analyze data consistently with the same pipeline. The authors rest their arguments on the facts that 1) there is already some heterogeneity in the data, 2) it takes too long and 3) it might not bring a meaningful difference. I strongly disagree with 1). The authors suggest that since I did not ask them to redo lab experiments, I have no right to ask for data analysis consistency. This logic I do not follow. Of course, it would be great if they also analyzed the samples in the lab consistently but this would require them to invest into redoing a half of the whole study and it might not be even possible for many samples due to the destructive nature of aDNA analysis. They also suggest that to be fully consistent they would have to reanalyze the previously published data – indeed that is what many labs in the field do and the authors must be aware of this. But I did not ask to follow the lead of such labs, I have only asked that they analyze the data newly reported in this study with the same pipeline, using to their advantage the fact that most of their relevant results are in comparisons amongst their own newly reported samples. I could see that the samples produced as whole genomes might be in the initial steps analyzed differently to SNP captures but this is not the case here – samples differ in the pipelines used just for historic reasons. Simply, the fact that there are inconsistencies in the lab analysis of the samples should not mean that researchers should give up on any attempts at consistency whatsoever. As to the point 2), the one claiming that it would take them too much effort. I am quite concerned what kind of analysis practices are in place when rerunning not so many samples (most of them captures, hence quite small files) would take so much time. Computer time will be less then a week on any decent cluster and if the scripting was done reproducibly, running the pipeline should be basically automatic. If other analysis were not done to the same reproducible standard, some might take a bit of time to regenerate but I just do not see how this could take the time the authors suggest. When I first suggested this, it did not occur to me that this was such a big issue, rerunning of the analysis of all samples when there is an improvement of the pipeline is a standard practice. What is however possible and even likely is the last point, 3). The reanalysis indeed might not bring any big differences and quite likely will not change conclusions. Me and the authors can both expect this, especially given that the types of analysis used are usually quite robust. However, none of us can be sure about this. To be honest, I would probably disagree with the authors but not insist on the reanalysis if they wanted to publish in a small journal. However, they aspire to be published in a journal that expects them to follow the highest standard and as such they should in my opinion follow at least the basic bioinformatic practices.

Reply： We have implemented all of Referee #3's suggestions in our revision, <u>except</u> for their suggestion to reanalyze all our newly reported data (but not previously reported data) using an identical bioinformatic pipeline. We respectfully disagree with them on the necessity of this reanalysis as we have discussed in our previous responses, and believe that this is a point on which we will simply disagree with them. As we have written in our response to their previous comments, we are confident that the suggested analysis will not change the conclusions of the study, and indeed have provided evidence for this through Online Table 3) which shows that the extremely minor differences between the bioinformatic pipelines have no evidence of biasing our inferences. We refer to this analysis three times in the final paper:

- *In the main text, we write: "We sequenced the DNA, and processed the data using one of two nearly identical bioinformatic procedures (Methods, Online Table 2) that we found gave indistinguishable results from the perspective of analyses of population history (Online Table 3)."*

- *In the Bioinformatics Processing section of the methods, we write: "To evaluate whether there was evidence that ancient DNA data processed using the same bioinformatic pipeline was artifactually biased to appear similar to each other in f-statistic analysis, we computed statistics of the form f_4(Group1Pipeline1, Group1Pipeline2; Group2Pipeline1, Group2Pipeline2) for all groups for which we had individuals in our main analysis dataset processed by both pipelines (Mongolia_EIA_Sagly_4, Mongolia_EIA_SlabGrave_1, Mongolia_LBA_CenterWest_4, Mongolia_LBA_MongunTaiga_3, Russia_MN_Boisman, and Taiwan_Hanben). For all 15 possible pairwise comparisons, the Z-scores for deviation from zero as computed based on a a Block Jackknife standard error had magnitude < |2.7| (Online Table 3), which is not significant after correcting for the 15 tests we performed (P=0.11 after applying a Bonferroni correction). While these analyses reduce concerns about systematic differences in population genetic analysis driven by changes over time in the*

> *software we used to carry out our bioinformatic processing steps, we caution that there are other inhomogeneities in our ancient DNA dataset that have the potential to affect inferences. Other sources of inhomogeneity include systematic differences in the chemical properties and preservation conditions of DNA from different archaeological sites, (b) differences in wet laboratory protocols including differences between data from in-solution enrichment and direct shotgun sequencing, and (c) differences in wet laboratory and bioinformatic processing protocols across research groups that published the various datasets co-analyzed in our study. The fact that we can obtain fitting models of population history through admixture graph analysis (Figure 2) even in the presence of these differences, and that the admixture graph model also fits when restricting to transversion polymorphisms (Supplementary Information section 3), and finally that our f_4-symmetry tests reveal no significant differences between data generated for this study using wet laboratory and bioinformatic protocols that changed over time (Online Table 3), increases confidence that our inferences are valid even in the presence of inhomogeneities.*[64]

That said, we agree with the general value of re-analyzing ancient DNA data—both published by ourselves and by others—with a uniform bioinformatic process. We were already planning to carry such a re-analysis as part of a separate project for which we are currently laying the groundwork, in which we will uniformly co-analyze all studies from our group and others published to date. The exchange with this reviewer has prompted us to prioritize this homogeneous re-analysis even more.