nature portfolio

Peer Review File

Combined genome-wide association study of facial traits in Europeans increases explained variance and improves prediction

Corresponding Author: Professor Manfred Kayser

Parts of this Peer Review File have been redacted as indicated to remove third-party material.

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

Version 0:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
Xiong et al. report an extension of their previous efforts to catalog the genetic variants influencing facial variation in an expanded sample of ~12K European-ancestry participants utilizing a greater number of facial landmarks. The primary analysis is a C-GWAS that combines results from over 500 individual traits (distances between facial landmarks), which identified over 253 lead SNPs across 188 separate loci associated with facial variation. Among these, the authors state that 62 loci are novel. Trans-ethnic replication/comparisons were conducted on an independent sample of ~10K Chinese. Secondary analyses include: (1) condition and joint analysis to identify loci with multiple independent association signals; (2) heritability enrichment analyses; (3) functional annotation; (4) evaluation of a predictive model of nose features for reidentification; (5) measures of evolution suggesting nose shape underwent recent selection in Europeans; and (6) inference about facial variation in archaic human (including Neanderthal and Denisovan). Several results indicate the validity and utility of the GWAS results. First, a large number of signals were replicated or otherwise previously reported. Recapitulating known signals and replicating newly discovered associations lends credence to the study and validates the specific association signals. Second, the reidentification exercise suggests the utility of genetic facial prediction. Third, the inferences regarding facial shape of archaic humans are generally consistent with skull evidence and provide new insight into what facial shape might be like for specimens currently lacking skulls. The C-GWAS method used here was previously developed by the authors. This method scales well in the context of meta-GWAS consortia, as it does not requite the exchange of individual-level date.

Minor comments:

The primary analysis uses the authors' own C-GWAS method to synthesize results across many correlated facial traits to generate a single set of genome-wide results benchmarked to the traditional p-value threshold of 5E-8. C-GWAS is not widely used in the genetics field for multi-trait analysis. Therefore, the manuscript may benefit from a bit more guidance in interpreting the results and describing any limitations of the method. In particular, while it is clear that small p-values from the C-GWAS represent the alternative hypothesis that one or more traits is associated, it is unclear to me whether the C-GWAS p-value for a particular SNP can be lower (more significant) than the p-value for any individual trait. The comparison to minP would suggest this.

Along these lines, to help the reader best understand C-GWAS results, please provide the genomic inflation factor, lambda, for the C-GWAS. Also, please consider presenting a Q-Q plot and lambda for the C-GWAS results of the remainder of the genome after removing the 188 associated loci. The LDSC intercept shows the inflation is not due to population structure, which is very helpful to the reader. There may be other causes of inflation.

Did the Chinese replication dataset use the global-to-local facial segmentation method mentioned in the introduction? It seems so. But that detail is buried in this paper. Consider clarifying this in the Results section and pointing out any implications for the replication analysis.

There is some lack of clarity about the number of traits combined in the C-GWAS. The methods section mentions that among 518 traits, highly correlated traits were processed to reduce this number to 128. The manuscript would benefit from clarifying this process a bit more.

There is some lack of clarity about the numbers of SNPs in the novel regions. The abstract (Line 39) indicates “We identified 253 unlinked SNPs across 188 distinct loci significantly associated with facial variation, including 64 SNPs at 62 novel loci.” The 64 SNPs are not mentioned again in the paper. In the section on Multiple Neighboring SNPs, this text is a bit hard to reconcile with the abstract: (Line 162-165) “Overall, this analysis revealed a total of 253 independently associated SNPs... Notably, 97 (38.3%) of these SNPs were not in LD ($r2 < 0.1$) with any previously reported face-associated SNPs, underscoring their novelty.” Does this mean that 97 – 64 = 33 SNPs are located at established loci but in low LD with the previously reported top SNP? The manuscript would benefit from reconciling the abstract and results sections a bit to clarify this point. Also, it is not clear to me how the authors know these 33 are novel, as many GWASs only report the top SNP in a region, and do not provide COJO results.

(Line 126): It is not clear to me why normal distributions were included for RS and TwinsUK but not the other three cohorts.

There is a possible discrepancy in the number of SNPs indicated in the Nose Profile section. Lines 289-291 note “130 SNPs” and “383 SNPs”, but only 129 SNPs and 382 SNPs, respectively, were used in Supp. Data 13. Supp. Data 13 does note the fallout of one SNP with 0 effect on all traits, but main text indicates (Line 289) “130 SNPs were nominally significant ($p < 0.05$) in our C-GWAS”. The abstract and results section calls this a 383 SNPs PRS. It seems that it is actually a 382 SNPs PRS. Clarity is warranted.

Lines 303-306: Is there a reference that can be cited to support his statement? This may be helpful for the readers who are not familiar with the anthropological finding of SAS nose shape being more similar to EUR than to EAS.

Reviewer #4

(Remarks to the Author)
Xiong and colleagues present a genome-wide association study (GWAS) investigating genetic associations with facial features. GWAS were performed in 11,662 European individuals from five different cohorts and results combined across traits using the C-GWAS method to provide a final single GWAS result. This builds on a previous study that used similar cohorts but slightly different analysis methods. This represents an interesting area of research that could have some impact in a number of fields including facial reconstruction in both modern and archaic humans.

The study is extensive and comprehensively reported, and describes a great deal of work. Although the paper is generally clear and well written, it is very long making it a difficult read. There are “five” figures, each of which has up to nine parts and each one of those parts has up to eight actual figures (I lost count but there are actually about 80 – 100 figures in just the main text without including supplementary figures). I think the authors should be commended for the detailed amount of work but I think all of the interesting work gets lost in the sheer level of detail. The paper would greatly benefit from the authors identifying the key messages and building the manuscript around those, rather than including every possible analysis possible (some of which are of greater quality than others) and moving the rest of the information into a supplement.

I have the following comments:
Major:
• I think the results of the project are generally overstated. There are a lot of claims that I’m not sure all are justified by the results. For example:
o The interpretation on archaic humans is based on a tiny number of samples (8 Neanderthal, 1 Denisovan and 1 admixed). There is obviously very little that the authors can do about this, but the paper makes very strong conclusions based on these results that I don’t think I would have as much confidence in.
o The AUC=0.665 for the nose PRS is fairly modest (though it is above 0.5). The authors conclude that this “underscores the potential for DNA-based facial recognition”. From those results, I think you could also conclude that this shows the limitations in DNA-based facial recognition in how little predictiveness these models currently have.
o The authors use the PRS to predict facial profiles of archaic individuals and then make general conclusions about Neanderthals and Denisovans. I have two issues with this; firstly, as stated above, this is based on a tiny number of individuals so making any general conclusions about Neanderthals is impossible and secondly, there’s no way of checking that these profiles are even correct anyway. The authors have already shown that the AUC of the PRS is small so it is unlikely these profiles are close to the individual’s true facial profile. The authors try to validate the approach using features that could be determined from the skulls, but I feel the results from this section are not the most robust.
• I think the “Face-associated SNPs experience positive selection in Europeans” section has some major limitations. There are a number of mentions about how SNPs identified as associated with facial features have higher allele frequencies in Europeans compared to other ancestry groups. The more common the variant the more power the GWAS has to detect an association for that variant. If you perform a GWAS in a particular population, then you will tend to identify variants that are more common in that population. These will have a higher allele frequency on average than variants in other populations regardless of what trait you look at. I worry that if the results in this section are based on this, then there are some incorrect interpretations of the results.
• The lead SNP has been taken to be the variant of interest. Was there any fine-mapping of signals?
• It looks like the nose PRS was tested again for performance in the Rotterdam study, which had been used as part of the initial GWAS? If so, there is going to winner’s curse bias affecting these results and the results will be inflated. Even with this inflation the AUC was still fairly modest (0.665).

• I found the variant to gene mapping section a little hard to follow. It seems to be equally based on FUMA and closest gene, but from the results the closest gene seems to be the most commonly reported. The paper states this leads to 581 genes from the union of CG to FG which sounds like a lot of those genes are unlikely to be true genes. Is a more robust gene prioritisation strategy needed? These genes were then input into the gene enrichment analyses. As it's unclear how robust the variant to gene mapping was, I'm not sure how much I trust the gene enrichment analyses.

Minor:
• I think the title should reflect that this is in European populations (despite the look-up of results in Chinese individuals)
• Imputation methods have not been reported for all cohorts. Where it is, it appears that this has been largely done using the 1000 Genomes Reference Panel. There are much larger reference panels available that provide more genetic diversity than 100 Genomes. The current gold standard would usually be imputation to the TopMed reference panel.
• It appears that loci are defined as those being greater than 500kb apart. Were there any additional checks to see if the signals were in regions with large LD blocks that were maybe wider than 500kb that could be counted as two independent signals?
• Could be made clearer in the supplementary data what genetic build the data is on (looks like it is build 37).
• I'm not sure how well the Chinese GWAS acts as replication given differences in genetic profiles and (if I have understood correctly) the differences in phenotype definition. There are five European cohorts used which would give the authors opportunities to use methods that compare effect sizes across cohorts, such as MAMBA, to assess replicability of signals.

Reviewer #5

(Remarks to the Author)
Xiong et al. present a very comprehensive study of the non-pathological variation of the human face. Their discovery pipeline involved a statistically-appealing meta-analysis (C-GWAS) of various European cohorts and they test whether the findings are replicated in a Chinese cohort or not. Among other results, their study (i) substantially increases the knowledge about genetic determinism of facial variation (~1.5x the number of discovered loci; 2.25x the variance explained), (ii) discovers multiple independent signals in some of these loci, and (iii) shows that face-associated loci were a target of positive selection. Based on this large set of loci, they moved on towards facial prediction. To that end, they built polygenic risk scores (PRS) for the 23 traits explaining most of the variance. They used these PRS for facial re-identification as well as for predicting facial features of archaic humans, claiming an enhanced prediction of both modern and archaic faces.
However, this last claim raises two major concerns, about the validation procedure of the PRS and about its application to archaic humans (Neanderthal and Denisova).

1. Validation of the PRS
The Authors identified 383 prediction SNPs for 23 PRS models. These models were defined, selected and calibrated using a European cohort, the Rotterdam Study (RS), then applied to a diversity panel, the 1000 Genomes Project Panel and used for re-identification.
A. The Authors validated their approach by assessing how fit were the predicted nose PRS (nose PRS being the best in terms of variance explained) to so-called "anthropological knowledge". However, I have not found any reference supporting this "anthropological knowledge", although it was repeatedly used as a benchmark (in various results sections: L303-6, L367, L381-2). I would strongly advise the Authors to provide solid support for this anthropological knowledge. Indeed, this is not only relevant given the need to validate the PRS, but also given the potential misuses of scientific findings when it comes to ethnicity and the role of genetic variations on facial outlook.
B. The PRS have been applied to modern and archaic human groups whose genetic background is quite different than that of the RS, therefore raising the issue of their transferability to other ancestries. In addition, given the relatively large number of SNPs involved in these PRS, it is possible that the PRS partly captures an ancestry effect. To address these issues, I would recommend the Authors to perform the following analyses:
a. Assessing how fit are their PRS to observed phenotypes when applied to (1) a European cohort different than the RS, and (2) a non-European cohort, e.g. the Chinese cohort used for replication. This would at least tell how reliable are those PRS when applied to groups genetically closer to European than are Africans or archaic humans.
b. Training the PRSs with repeated random selections 383 (or 359) SNPs, e.g. the random selections of SNPs described in section "Genetic diversity and positive selection". Then, for each "random PRS", assessing the re-identification AUC yielded by that random selection and the result of the directionality analysis for Neanderthal samples. This would empirically provide a p-value for two reported scores, the re-identification AUC for nose PRSs (67%) and the score of 15 Neanderthal-specific characteristics predicted over 16 of them.
2. Application of the PRS to archaic humans
There is an apparent paradox in Figure 5F: whereas archaic introgression has occurred in non-African populations, the PRS of archaic samples (Neanderthals in particular) correlates more with PRS of Africans than with those from introgressed populations (EUR, EAS, SAS, AMR). It is even harder to understand when the Authors brought support for an enriched introgression in face-associated loci: 30 lead SNPs located in introgression segments in both EUR and EAS over 253 lead SNPs (11%) vs. the estimates of 1-2% of introgressed genome in these populations. In L383-6, the Authors explained this higher facial similarity by "African populations ancestral for all modern humans and therefore expected to be more similar to the common ancestors of all human groups". This claim is not supported and somewhat questionable: Africans population may also have experienced selection and drift since the split with archaic humans. In addition, as said above, they were not affected by introgression. I would recommend the Authors to comment on that apparent paradox in the light of the additional analyses proposed here above.
Also, although the Authors mentioned various limitations of their PRS approach to predict archaic faces, I would appreciate them to elaborate about: (1) the application of PRS trained on diploid samples to the pseudohaploid samples (as it is the

case of most archaic samples in the AADR) and (2) the use of the AADR itself, leading (if I correctly understood) to replace many prediction SNPs by their LD counterparts, instead of directly using the publicly available files (BAMs/VCFs). As material for this discussion, I recommend the following paper that addresses, among others, the transferability of PRS to archaic genomes: https://doi.org/10.1146/annurev-genom-111521-121903.

Minor comments
INTRODUCTION
In the first paragraph, many assertions lack literature support. Please provide references.

RESULTS
Nose PRS profiles are distributed according to anthropological knowledge. Please prune some of the admixed samples from the 1000 GP for this analysis. For instance, ACB and ASW, both part of the AFR continental group, contain individuals with substantial European ancestry. Also, PUR has a very different ancestry profile compared to the three other populations of the AMR group, which, in turn, contain samples with a wide variety of ancestry. You could either choose to select samples based on ancestry or simply drop these six populations from the analysis. The latter option might be the best in this case, since AMR predictions are not outlined in the main text and dropping ACB/ASW should yield more differentiated predictions for the AFR group.

Facial characteristics of archaic humans inferred from face-associated SNPs. The segmentation file results from three studies (Vernot & Akey 2014, Sankararaman et al 2014, Yuan et al. 2021) that are, to my knowledge, all based on only one high-coverage Neanderthal genome (the Altai Neanderthal genome). This genome may not be the best reference but I don't know any segmentation file based on the Vindija genome. Still, it would be good to bring a comment on that point. In particular, see Reilly et al. 2022 (ages of the high-coverage Neanderthal genomes and relation to introgressing population), Kuhlwilm et al. 2016 (ancient modern human introgressing in Eastern Neanderthals) and Prüfer et al. 2017 (the difference between the Altai VCFs produced in 2014 and 2017).

METHODS
L610. The usual metric for assessing imputation quality using IMPUTE2 is the INFO score? I understand you choose R2 instead of INFO score to be consistent with other cohorts (e.g. RS imputed with MaCH), however you should check that this filter correctly applies to only the so-called "type0" or "type1" SNPs (i.e. those present only in the highest density panel) and not all SNPs, otherwise your R2 may be inflated.
L629. I agree that you do not provide all details about data generation for the ALSPAC cohort, but at least please provide the genotyping chip and the imputation software.
L662. Please provide more details about the covariates: why were height and weight included in some cohorts and not in other? Please elaborate on how you verified that varying covariates between cohorts would not impact the GWAS results.
L666. Why were symmetric distances averaged in the RS cohort and not in other ones? Please provide details.
L672. You used fastGWA to deal with close-relatedness in the TwinsUK cohort. I understand kinship is quite high in that cohort compared to other ones, still spurious close-related pairs may be included in these cohorts. What did you implement to deal with such pairs?

Reviewer #6

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

Version 1:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
The authors were responsive to my concerns. I have no further comments.

Reviewer #3

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

Reviewer #4

(Remarks to the Author)
I would like to thank the authors for their detailed responses and clarifications. I think these largely address my concerns.

I still think the results relating to archaic individuals is severely limited given the low sample size but I think the text better reflects this.

My one remaining concern with the paper relates to the PRS analysis which contains overlapping individuals in the testing and training set. The authors have added a sensitivity analysis using Twins UK which has less overlap than the RS study. They have also shown the PRS performs better than a null model. My concerns are 1) the fact the AUC drops in Twins UK compared to RS would be the result you would expect if there was winner's curse bias bias, and 2) I don't think the null models show there is limited winners curse bias. If you select a small subset of random SNPs, then these are unlikely to show an association and are therefore expected to provide a poor PRS under winner's curse (a SNP not associated in the GWAS is guaranteed to be a poor predictor in the PRS if there is sample overlap).

Reviewer #5

(Remarks to the Author)
I'd like to first acknowledge the great amount of additional work that has been achieved by the Authors. On various aspects, the manuscript is now more robust than it was at first submission. Most of my major comments have been fully addressed. Still, I have two major comments regarding the interpretation of the high correlation of facial PRS between Neanderthals and Africans and the findings related to the effect of archaic introgression on the human face.

Interpretation of the high correlation between Neanderthals and Africans facial PRS
Your additional analyses are interesting but I suspect that what you have here is a case of bias from the discovery/training cohort. This cohort being European, the selected SNPs have a higher chance of being privative to Europeans, meaning that they appeared/were selected in Europeans (as supported by your observations in the section on positive selection). On this point, the HGDP sequence paper is relevant (Bergström et al., Science, 2020), including the finding that >99% of alleles privative to non-African populations are derived. Therefore, the proportion of ancestral alleles over the set of 357 SNPs (Fig. 7d) is a measure of how genetically distant are Neanderthals (and each continental population) from the study cohort on the basis of European-specific variants. In other words, a measure of how "non-European" are Neanderthals, Africans and co.
If the discovery GWAS and the PRS training were both performed on an African cohort (thus biasing towards African-specific variants), one may likely observe the reverse situation: Neanderthal PRS resembling more to European PRS. Note however that the trend may be less marked than the one you have reported because the alleles privative to African populations are ancestral in a higher proportion than non-African populations (i.e. Africans and Neanderthals do share a higher proportion of ancestral alleles, as seen in Bergström et al.).
Here an important point is the comparison to expectation. In Figure 7e and 7d, you have reported results drawn at random over 10,000 sets of 357 SNPs that matched the frequencies of your PRS SNP set. These results show the same pattern as for the PRS SNPs, namely that the mean frequency of ancestral allele in Africans is clearly distinct from that of non-Africans (higher by 5-7%), leading you to propose (line 421-2) that it may "extend to other polygenic traits". As mentioned above, African populations may well have higher frequencies of ancestral alleles (see also https://doi.org/10.1093/molbev/msm018). Nevertheless, the similarity between results from the random and PRS SNPs could be overestimated because the ransom sets have been matched to European frequencies. Thus, these sets would tend to capture SNPs that are more common in Europeans, reinforcing the observed pattern.
As a test, I have myself run comparisons using your SNP set and the HGDP continental populations (in order to have a less biased representation than with the 1000 Genomes data). For simplicity, I limited these comparisons to chromosome 2 (40 SNPs) and 1,500 fully-random draws (not MAF-stratified). I got a rather clear, positive relation between the AFR-EUR difference in mean frequency of the ancestral allele and the average European MAF of the drawn set. Extrapolating to the average European MAF of your set, it predicts a frequency delta around 9%, in line with your Figure 7e.
In conclusion, I find it important to be very cautious in your statements, so as to consider that cohort bias may partly drive the PRS results. For instance, in lines 421-2, when you state "the high PRS-derived similarity between AFR and archaic humans may extend to other polygenic traits", I would add "as long as the PRS for such traits are calibrated on European cohorts". More importantly, rather than "Facial PRS profiles of Neanderthals resemble Africans more than non-Africans", the interpretation should be something like "Facial PRS profiles of Neanderthals are even more different from facial PRS profiles of Europeans than are facial PRS profiles of Africans". These are only examples. I recommend to carefully go through the whole text and to nuance other similar statements.

Global effect of archaic introgression on the human face
The results outlined in lines 365-87 are solely based on the overlap between reported introgressed segments and GWAS leading SNPs. Ancestral/derived allelic states at these SNPs (in Neanderthals, Africans and non-Africans) are completely ignored. However, a leading SNP within an introgression segment could carry a derived allele postdating the introgression (thus remaining at the ancestral state in Neanderthals). This is actually what your Figure 7d suggests. In such a case, calling this an "introgressed face SNP" is an overstatetement
In consequence, I would recommend the Authors to nuance their claims, among other examples, in:
- Lines 373-87
- Lines 620-1: the study does not "link overall facial morphology with archaic introgression", but links it with genome segments where archaic introgression was detected.
- Lines 627-8: the study does not bring direct evidence as it does not establish the archaic origin of the variants located in the introgression segments.
NB: on this question, see also under which conditions archaic SNPs are defined, for instance in https://doi.org/10.1016/j.cub.2020.06.045: (i) 100% ancestral in Africans, (ii) homozygous derived in at least one archaic sample, (iii) derived allele present in at least one non-African population.

Minor issues
Lines 429-41: In this paragraph, when you refer to “concordant results”, please state well that the concordance is between (a) PRS applied to Neanderthals relatively to Europeans and (b) Neanderthal features reported in literature relatively to modern human features, rather than the between (a) Neanderthals and (b) EUR, which may be misleading. For instance, “The statistical significance of these findings suggests that the observed concordance between Neanderthals and EUR is unlikely to be due to random chance” should be “The statistical significance of these findings suggests that the observed concordance between European-based PRS applied to Neanderthals and reported facial features of Neanderthals is unlikely to be due to chance.”

Lines 638-53: There are actually 3 high-coverage Neanderthals, the third one being nicknamed Chagyrskaya (https://doi.org/10.1073/pnas.2004944117). Please rephrase accordingly.

Lines 650-1: You could add that the Vindija sample is expected to be the best representative of the introgressing population, as suggested by Reilly et al. 2022 (https://doi.org/10.1016/j.cub.2022.08.027), and that recent studies have proposed a narrow time range (~45-49 kya BP) for the introgression period (https://doi.org/10.1126/science.adq3010 and https://doi.org/10.1038/s41586-024-08420-x).

Reviewer #6

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

Version 2:

Reviewer comments:

Reviewer #5

(Remarks to the Author)
The authors have done an impressive work to address the two major points that remained from my last revision; I appreciate this very much.
Still I would consider it suitable for publication once the following comments regarding these two last points will be integrated to the current version of their manuscript.

(1) Interpretation of the high correlation between Neanderthal and African facial PRS
Although I appreciate very much the attempt to assess what the PRS comparisons would be in the case of an AFR-based GWAS, I find the approach pointless and misleading.
Pointless: unless running simulations (which is far beyond the scope of the current study), there is no way of fairly mimicking an AFR GWAS from a bunch of non-AFR GWASes. Indeed, in a regular GWAS setting, we can classify the set of tested SNPs between (A) those private to the tested population and (B) those shared with other populations. Note that a common trend is that the SNPs in (A) are: (i) more recent mutations, (ii) shared on longer segments (i.e. potentially confusing with selection sweeps), (iii) at a lower frequency, and, more importantly, (iv) with a stronger effect on the phenotype (higher beta). Although you put strict frequency filters, the AFR-selected SNP set that you construct in Supp. Note 10 is, as I understand it, just of a subset of the (B) set here above. Hence, it cannot be considered as a fair proxy for the SNP set that one would obtain through an AFR-based GWAS as it lacks the crucial part, the “true” AFR-private SNPs (part (A)). Therefore, it is not surprising that you get the same tendency in terms of PRS comparisons. Moreover, if your AFR-selected SNP set lacks the strong-beta, EUR- or EAS-private SNPs, this tendency in PRS comparisons will be attenuated, just as you have it now.
Misleading: as said above, this approach does not provide a proxy for the SNP set obtained in an AFR-based GWAS and is therefore not a fair comparison. The conclusions are thus overstated.

As a conclusion, I would suggest the Authors to:
(1) Drop this Supp. Note 10
(2) Relax the MAF filter when setting up the 10,000 random sets for Fig. 7e-d. As said in my previous revision, I have myself done the computation using HGDP’s chr2 data and did not observed any tendency towards higher ancestral alleles in AFR compared to other populations.
(3) Just mention in the Discussion that the similarity between AFR and NEA PRS can be due to a discovery cohort bias, i.e. due to a weak portability of the PRS to other populations.

(2) Global effect of archaic introgression on the human face
As for the first point, I appreciate very much the additional work carried by the Authors. With the corrections and nuances that were provided, the point is now almost closed to me. I would only suggest the Authors to stress that the definition of “archaic SNPs” as in Zeberg et al. (2020) does not imply that the presence of such variants in modern populations is necessarily due to hybridization between Neanderthals and Early Sapiens, it could also be due to incomplete lineage sorting. To disentangle this, it would be necessary to date the variants but this goes far beyond the scope of your study.

Reviewer #6

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

Version 3:

Reviewer comments:

Reviewer #5

(Remarks to the Author)
I acknowledge the work achieved by the Authors on understanding this complex issue of ancestral allele (AA) sharing between modern and archaic populations. However, I was rather surprised to see that your null distribution of mean frequency of ancestral alleles (Figure 7.e) among modern populations was quite different than the one I obtained myself using the chromosome 2 of the 1000 Genomes Project data (about 95% of AA in any modern population – see attached Fig1, left panel).
In my understanding, this difference stems from a SNP recruitment bias: when randomly sampling over the subset of 1000GP variants that are present in the AADR, the null distribution looks quite different (Fig1, right panel) and is actually very similar to yours (Figure 7.e), thus with substantially higher mean frequency of AA for African populations.
Thus, in my view, the higher AA proportion in AFR obtained in Figure 7.e should be interpreted as a consequence of the (EUR-oriented) recruitment bias present in the "1240K capture" of Mathieson et al. (2015), rather than by the hypothesis you made in L390-1, namely that "African populations retain a genetic lineage closer to the common ancestors of modern and archaic humans". Therefore, I recommend the Authors: (1) to verify that the dataset used for drawing their null distribution of mean AA frequency was not affected by such a recruitment bias, and (2) in case of such a bias, to withdraw biased null distributions and to rephrase the corresponding Results and Discussion sections accordingly.
Please also note that, for the sake of simplicity, I did not include archaic individuals in my analyses (computing frequencies in archaic individuals on SNPs drawn out of the AADR set involves to fetch data in the complete archaic genomes) but would also expect a very different null distribution of mean AA frequency for those.

Fig1. Attempts to reproduce Fig7.e (Mean frequency of ancestral alleles) over four modern populations. 1,000 sets of 40 SNPs were sampled randomly over: (left) all 1000G variants of the chromosome 2, or (right) the intersection of all 1000G variants and the AADR assay (the "1240K capture"– Mathieson et al. 2015). For each random set, I compute the per-population mean frequency of AA and report the distribution of this mean over 1,000 random sets on each panel. Black horizontal lines show the mean frequency of AA over the set of 40 predictor SNPs (= the subset of your PRS SNPs on chromosome 2).

Reviewer #6

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