Cell Genomics Transparent Peer Review Record

Performance and accuracy evaluation of reference panels for genotype imputation in sub-Saharan African populations

Dhriti Sengupta[1], Gerrit Botha[2], Ayton Meintjes[2], Mamana Mbiyavanga[2], AWI-Gen study, H3Africa consortium, Scott Hazelhurst[1,3], Nicola Mulder[2,#], Michèle Ramsay[1,4,#], Ananyo Choudhury[1,#]

Summary

Initial submission:	Received : 8/21/2022
	Scientific editor: Laura Zahn
First round of review:	Number of reviewers: 2 Revision invited : 11/10/2022 Revision received : 2/11/2023
Second round of review:	Number of reviewers: 2 Accepted : 5/2/2023
Data freely available:	Yes
Code freely available:	Yes

Referees' reports, first round of review

Reviewer #1: The authors have examined the accuracy of imputation using multiple publicly available panels into geographically diverse African populations to examine which reference panel provides greater accuracy and coverage with imputation into genotype data. Accuracy is determined using the info and r2 metrics, as well as with 95 high coverage genomes with both genotype and whole genome sequence data. The authors suggest that diversity and ancestry match of the reference panel to the target panel is likely more important than size. These are striking and important findings with important implications for African genomics - future GWAS and fine mapping. However there are a few methodological issues that may require greater discussion or assessment:

1) While the impact of Afroasiatic and Nilo-Saharan ancestry appears to have been examined, the assessment does not seem to have specifically ascertained or assessed the impact of Eurasian ancestry (which forms a significant proportion of many East African populations), and can impact genetic diversity, and imputation accuracy depending on how these haplotypes are represented in the reference panel. The authors speak about lower level of diversity in Ethiopian populations, which is almost certainly related to the very high levels of Eurasian ancestry in these population groups, rather than other aspects of population divergence, and population history. The ADMIXTURE plot doesn't seem to differentiate Eurasian ancestry from Afro-Asiatic and Nilo-Saharan ancestry specifically. Similarly, the impact of rain-forest hunter-gatherer ancestry in East African populations is also not measured, although this is likely to be less substantive overall.
2) Is taking an intersection of all SNPs when only using a variant dataset for high coverage data to assess the NDR the best approach? The ideal approach (as has been used before for NDR) is to recall all sites (even if ref/ref) - in terms of gvcf format for high coverage data, to ensure that all false positives are included in the numerator and denominator of the NDR calculation. This would mean using gvcf data where specific individual sample quality metrics were met rather than combining gvcfs, and doing variant calibration.
3) Given the methods for imputation used on the SIS and Michigan server are different, how can we rule out differences in methodology rather than panel contributing to differences? Table 1 shows vastly different number of SNPs for the 1KG panel hosted on SIS and Michigan server - why is this? Perhaps the same panel can be compared on both to address the differences likely arising from different imputation methodology? This seems to be mentioned in the discussion, but could further analysis clarify this? If not, perhaps the message should be that this is a real-world imputation scenario as reference panels are available on different servers, but discrepancies due to algorithm cannot be fully ruled out
4) Is it possible that AGR performed better on accuracy because relatively rare SNPs (mono-allelic SNPs) were removed from the 95 high coverage genomes as part of QC, so many rare variants were dropped? What is the rationale for doing this is such a small sample of high coverage genomes? I can see why this would be useful to do for the large imputation panels, but not for the high-coverage 95 genomes, where this would lead to low frequency SNPs being disproportionately discarded. Rather than this, quality metrics can be used so that private and rare SNPs are also included, and the coverage, and accuracy with respect to these can be assessed in the imputed data. It does appear as if the biggest difference between imputation with the AGR and TopMed panel is imputation of rare variants.
5) The authors don't show NDR by population - as a comparison between AGR and TopMed- this would be useful to see to examine whether AGR provides a specific advantage in some populations over others.
6) It would be informative to see SNP coverage (compared to WGS data) by allele frequency bin for AGR and TopMed to see if coverage diverges at specific allele frequencies. This is also an important point for discussion, in terms of implications for GWAS and fine mapping

Reviewer #2: Major Revisions:

1. The introduction and especially the discussion should briefly tie back to the major motivation - the huge underrepresentation of SSA populations in GWAS. Evaluating the accuracy of imputation in SSA populations is a vital step towards improving their inclusion in GWAS. It would be great to cite more sources for underrepresentation of African and especially non-Western African ancestry populations in GWAS, e.g. only 2% of individuals included in GWAS are individuals of 'African ancestry' and vast majority of African ancestry populations in genetic studies are western African e.g. African Americans or Afro-Caribbeans (72%-93% in the GWAS catalog and ≥90% in gnomAD) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6494470/.
2. I recommend including a plot that shows NDR by INFO/R^2 score bins above 0.6. Researchers, such as ourselves, often need to use a much more stringent quality score cut-off than employed in this study. It will be helpful to include the NDR rates across the entire range of quality score bins to evaluate the accuracy gained versus number of variants filtered out with more stringent cut-offs. Figure 5c is a striking and a key finding, but would be improved by showing NDR rates per quality score bins across the different reference imputation panels.
3. It would be insightful to show the ancestry-specific NDR rates for the WGS samples. Is imputation accuracy worse for Khoe-San ancestry segments than Niger-Congo segments in the WGS samples or has the inclusion of 84 Nama (Khoe-San) samples in AGR been sufficient to impute Khoe-San ancestry at comparable accuracies? Your conclusions about the presence of Khoe-San genomes in AGR positively impacting imputation in South African samples could be strengthened if you incorporated this analysis.

Suggested Revisions:
1. More details on samples in AWI - at least include brief breakdown of the ethnic groups in each region, as there is strong population structure in places like Kenya and South Africa.
2. Provide an explanation for your determination of the 3 Khoe-San ancestry and 3 non-Niger Congo ancestry proportion bins. How did you choose to split the data before testing for group mean differences needs to have clear a priori reasoning? Why not just run a regression between ancestry fraction and imputed SNP count by individual?
3. Related to the prior comment: you see a strong decline in the number of imputed SNPs per individual for genomes with greater AA or NS ancestry. I'm not totally convinced this is due to AA and NS genomes having lower SNP diversity than Bantu-speaking populations (page). I'll note that there is only 1 NS- speaking population in AGR which is the Gumuz (n=24), and even the other Ethiopians are not especially ancestrally closely related to Kenyans (Gopalan et al. 2022, Current Biology). I wonder rather if this is a mismatch between the ancestry of the AWI Kenyans and Eastern African references in AGR?
4. Were ambiguous A/T C/G SNPs filtered prior to imputation? https://www.frontiersin.org/articles/10.3389/fgene.2019.00034/full Schurz et al., 2019 found a big difference in imputation accuracy when excluding these SNPs because of strand bias on arrays.
5. It would be helpful to show the complete distribution of INFO/R^2 scores in addition to Figure 2a) showing proportion variants over 0.6. A figure showing the overlapping distributions of INFO/R^2 scores across the different imputation panels would be informative. It is especially interesting that at high allele frequencies the mean INFO score in AGR out performs TopMed (Fig 2b), although perhaps this is a function of the Sanger IS vs MIS.

Minor Comments:
1. Methods: NDR equation format does not display correctly in my pdf
2. Include a more precise definition of Khoe-San under "Influence of ancestry and admixture". For example, "Khoe-San is a collective term for populations across Southern Africa that predate the expansion of Bantu-speakers; these populations descend from the earliest human population divergence and thus harbor greater genetic diversity and higher SNP content…"
3. Caption for S figure 6 is reversed: red circles are quality score < 0.6
4. Footnote in Table 1, the Nama here are from South Africa - not Namibia (can cite van Eeden et al. Genome Biology 2022 for the data deposition).

Authors' response to the first round of review

Response: We sincerely thank both the reviewers for their careful reading of the manuscript and the thoughtful comments. These have provided us with directions for more effective analysis, representation and interpretation of the results. We have added a line in the acknowledgements to express our gratitude. Summarized below are our response to the comments (in green), the amendments made (in blue) and supporting data added.

Reviewer #1

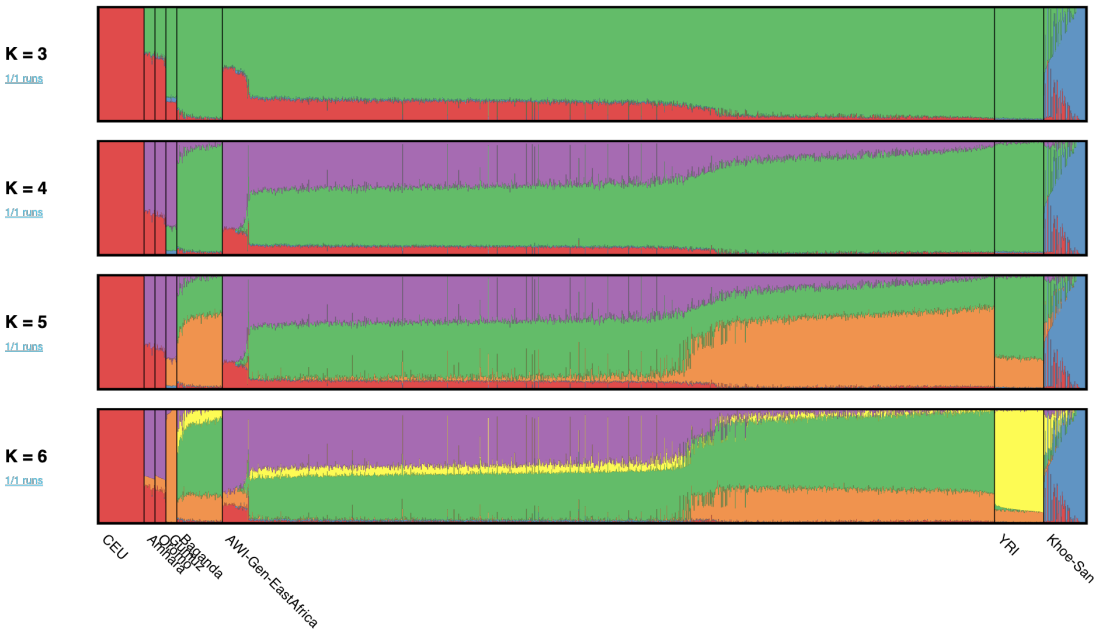
The authors have examined the accuracy of imputation using multiple publicly available panels into geographically diverse African populations to examine which reference panel provides greater accuracy and coverage with imputation into genotype data. Accuracy is determined using the info and r2 metrics, as well as with 95 high coverage genomes with both genotype and whole genome sequence data. The authors suggest that diversity and ancestry match of the reference panel to the target panel is likely more important than size. These are striking and important findings with important implications for African genomics - future GWAS and fine mapping. However there are a few methodological issues that may require greater discussion or assessment:

1) While the impact of Afroasiatic and Nilo-Saharan ancestry appears to have been examined, the assessment does not seem to have specifically ascertained or assessed the impact of Eurasian ancestry specifically. Similarly, the impact of rain-forest hunter-gatherer ancestry in East African populations is also not measured, although this is likely to be less substantive overall.

Response: We are grateful for this and a related comment by Reviewer 2 regarding how the ancestral composition of the East African participants has been considered in our study. Indeed a part of non-Niger-Congo ancestry in East African populations corresponds to Eurasian gene flow as demonstrated by their characteristic separation by Principal component (PC) 1 in African PC plots (such as in Gurdasani et al 2015, Choudhury et al. 2022). The admixture plot based on our data (please see below) shows a red component corresponding to Eurasian ancestry to separate out in some of the participants at higher values of K.

As the aim of our analysis was to broadly assess the impact of non-Niger-Congo ancestry on imputation performance, instead of going into complex ancestry deconvolutions (which as the admixture plot shows would vary widely depending on the value of K used as well as reference datasets included), we have used ancestry proportions at K=3 which partitions the dataset into a composite non-Niger-Congo (shown in red, we have referred to this as the EA-non NC ancestry in the text), Niger-Congo (green) and Khoe-San (blue) related ancestries. We have amended the figure (included the full Admixture plot in S Figure 5B), the figure legend and the text to clarify this. We have also added a caveat stating that the results might differ if a more complex ancestry deconvolution is employed.

We agree with the reviewer's point on the rain forest forager (RFF) ancestry. As our preliminary analysis of the East African data did not detect strong RFF ancestry in any of our study participants, we have not included this in our analysis. However, the presence of a trace level of RFF ancestry cannot be ruled out. We have added a line in the results section of the to indicate this.



2) Is taking an intersection of all SNPs when only using a variant dataset for high coverage data to assess the NDR the best approach? The ideal approach (as has been used before for NDR) is to recall all sites (even if ref/ref) - in terms of gvcf format for high coverage data, to ensure that all false positives are included in the numerator and denominator of the NDR calculation. This would mean using gvcf data where specific individual sample quality metrics were met rather than combining gvcfs, and doing variant calibration.

Response: We thank the reviewer for this valuable comment. The reason that motivated the use of our particular approach was the underlying large-scale (almost 20 fold) difference in the panel size of AGR and TOPMed. This difference, especially when considering Ref/Ref sites, might over penalize the larger panel, as just by chance, the propensity of mismatch with WGS would be higher for a dataset imputed using TOPMed panel.

Instead to provide the readers with an estimate of the extent of possible mismatches for the Ref/Ref sites in the WGS, we have presented in the Venn Diagram of Figure 5B, the number of sites that are potentially Ref/Ref in the WGS data but are non-ref/ref (in one or more individuals) in the imputed dataset. The figure shows that there are 1.45 million such sites in the TOPMed imputed dataset and 0.93 million in the AGR imputed dataset.

We completely agree that a gVCF approach could have provided a more accurate and appropriate estimate. Unfortunately, the analyst who had performed the core sequence data curation is not currently active in the project, consequently the curation of individual level gVCFs files would have considerably delayed this revision. Therefore, we have added the following lines in the Discussion section to highlight that there are alternative approaches to calculate NDR and also that the use of these approaches might lead to different NDR estimates:

"Due to the large-scale differences in size of the panels compared, which could intrinsically bias the NDR estimates against larger panels, we did not include sites that were Reference/Reference across the WGS dataset in our comparisons. Therefore, NDR estimates derived using alternative approaches, such as those based on comparison of individual level

gVCF files, might differ from the results presented here. As an indirect estimate of the level of difference that might be observed if Reference/Reference sites in the WGS were included in NDR estimation, we have noted the number of sites (Figure 5B) that were Reference/Reference in the WGS dataset but had at least one non-Reference allele in each of the imputed datasets. The observation of a considerably higher number of such sites for the TOPMed panel (1.45 million) compared to the AGR panel (0.93 million) hints that the consideration of such sites in NDR estimates could further increase the difference in NDR with respect to WGS for these panels. ”

3) Given the methods for imputation used on the SIS and Michigan server are different, how can we rule out differences in methodology rather than panel contributing to differences? Table 1 shows vastly different number of SNPs for the 1KG panel hosted on SIS and Michigan server - why is this? Perhaps the same panel can be compared on both to address the differences likely arising from different imputation methodology? This seems to be mentioned in the discussion, but could further analysis clarify this? If not, perhaps the message should be that this is a real world imputation scenario as reference panels are available on different servers, but discrepancies due to algorithm cannot be fully ruled out

Response: We agree with the reviewer that more information on this would have been helpful. The inaccessibility of the actual panels hosted at these services did not allow us to perform any direct comparison. However, based on the information provided in the respective websites, it is clear that very different curation thresholds were used for generating the panels that represent KGP dataset at the SIS and MIS servers. While the 1000 Genomes panel at SIS includes 85 million SNPs, which is closer to the size of the original KGP dataset, the KGP panel at MIS is much smaller and only contains 45 million SNPs.

Based on the reviewer’s suggestion we have added the following lines to the main text:

“These differences could be driven by panel size variation (the KGP_M panel is almost half the size of the KGP_M panel) as well as the use of different imputation algorithms (SIS uses PBWT while MIS employs MiniMac4). Due to the unavailability of the actual panels for evaluation, we were unable to assess the relative contribution of these factors. However, as observed in previous studies 27,28 both these factors probably contribute to the differential imputation performance of the KGP panel hosted at the SIS and MIS.”

4) Is it possible that AGR performed better on accuracy because relatively rare SNPs (monoallelic SNPs) were removed from the 95 high coverage genomes as part of QC, so many rare variants were dropped? What is the rationale for doing this is such a small sample of high coverage genomes? I can see why this would be useful to do for the large imputation panels, but not for the high-coverage 95 genomes, where this would lead to low frequency SNPs being disproportionately discarded. Rather than this, quality metrics can be used so that private and rare SNPs are also included, and the coverage, and accuracy with respect to these can be assessed in the imputed data. It does appear as if the biggest difference between imputation with the AGR and TopMed panel is imputation of rare variants.

Response: We acknowledge the reviewer’s concern about the small sample size of the sequence dataset that was employed for evaluating accuracy. However, given the current status of African genomics, where the largest high-coverage sequence datasets still contain only a couple of hundred genomes, it might take us years to have WGS datasets that are large enough to capture ultra-rare variants and enable a thorough comparison. Moreover, the sample size of

most continental African GWASs are still modest (N<5000) and consequently they are intrinsically under-powered to perform association testing on ultra-rare variants. In many cases, these SNPs are removed on the basis of minor allele frequency/count filtering. Despite the limited sample size we expect our high-coverage sequence datasets to provide reasonable estimates of imputation accuracy of relatively common SNPs, that small to moderate sized African GWASs are powered to test.

The major insight from our analysis (as indicated in the figure below, which has been added as Figure 5D) is that the NDR with WGS, observed for the two panels, increases with the level of Khoe-San ancestry in a genome. Moreover, the line showing the extent of increase is steeper for TOPMed compared to AGR. Therefore, for a GWAS based on a population that has high Khoe-San admixture, the AGR, due to the inclusion of Khoe-San ancestry genomes, would probably be a far better choice compared to TOPMed. It can also be extrapolated that similar differences would be observed due to presence/absence or differential representation of other African ancestries (such as RFF) in any two panels. We expect our results to inform future panel design to be mindful of the importance of the inclusion of understudied African ancestries to enable the panels to be truly representative. We have added a following lines to the results section to describe this data and its relevance:

"To investigate whether differences in ancestral composition lead to the large-scale differences in observed NDR rates among the 95 individuals, we compared the level of NDR to the level of Khoe-San ancestry in each individual. For all the imputed datasets, we observed an increase in NDR with an increase in the level of overall Khoe-San ancestry in an individual (Figure 5D). The dataset imputed using the AGR panel not only showed the overall lowest NDR but also the lowest rise in NDR with the increase in Khoe-San ancestry. Based on these results, we postulate that the inclusion of 84 Nama genomes in this panel to be the source of the considerably better performance of AGR, especially in individuals with substantial Khoe-San ancestry."

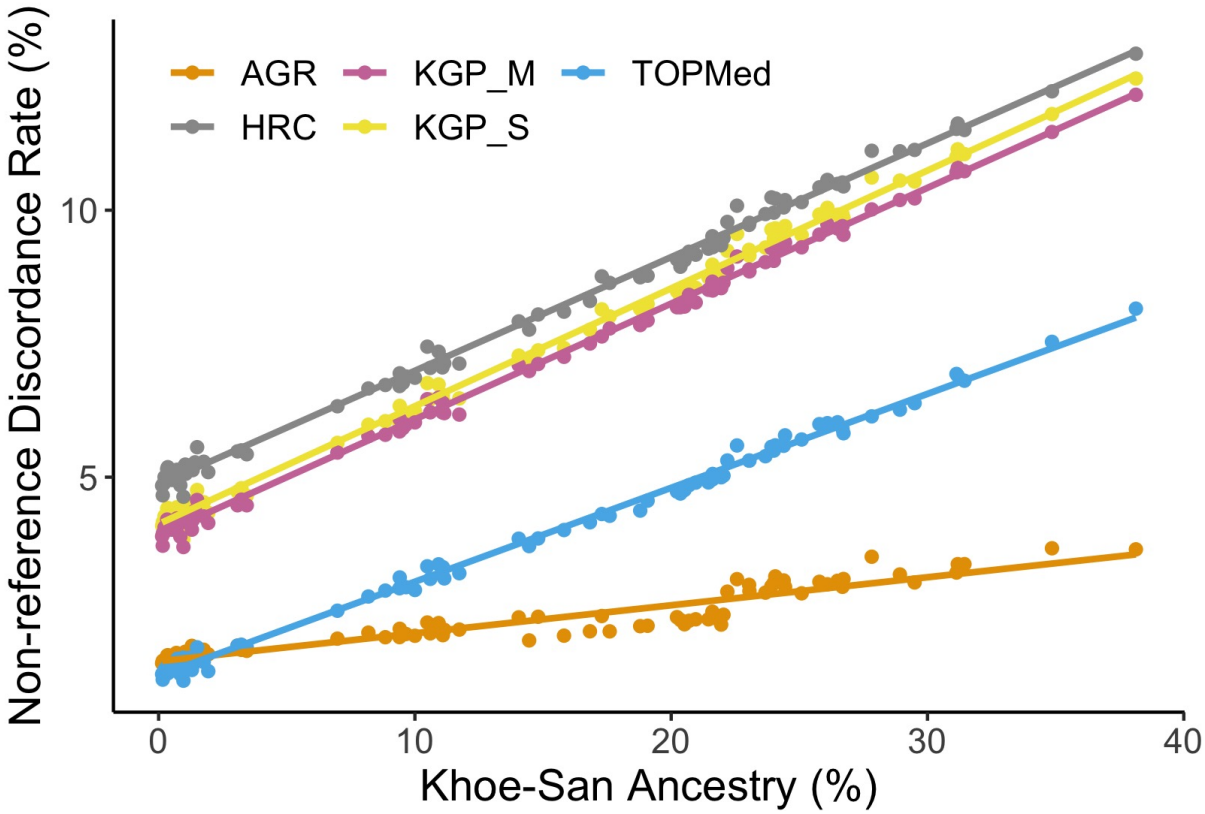


Figure 5D. Correlation between the overall genotype discordance (estimated by NDR) and the level of Khoe-San ancestry in the five imputed datasets. The inclusion of the representative Khoe-San population probably leads to a much lower discordance and a gentler slope in the AGR compared to other panels.

5) The authors don't show NDR by population - as a comparison between AGR and TopMedthis would be useful to see to examine whether AGR provides a specific advantage in some populations over others.

Response: We thank the reviewer for this thoughtful suggestion which has provided us an additional avenue to explore the regional differences. We have added a figure which compares the NDR of AGR vs TOPMed in the East, West and South African populations. As can be observed in the Figure 4D (please see below), for West Africa (shown in yellow) the NDR is almost constant across the dataset, while the NDR shows huge variation among the South African participants (green). Furthermore, S Figure 7a shows that the difference in NDR corresponds to the level of Khoe-San ancestry. Similarly, the East African participants also show clear variation in NDR that correlates with the level of non-Niger-Congo ancestry in an individual (S Figure 7b). We have added the following text to the results section to introduce and explain this data:

"We also investigated the level of discordance of genotypes imputed by AGR and TOPMed for East, West and South African populations. This was done by estimating non-reference discordance rates (NDR) between AGR and TOPMed for individuals from each geographic region separately. Data for the South African participants showed the highest NDR between

the two panels (NDR= 4.89± 1.26%), followed by East Africans (NDR=3.61± 0.74%). The West African participants showed the lowest discordance (NDR= 2.71± 0.14%) (Figure 4D). To identify the possible source of systematic differences in NDR between the two panels, we assessed the relationship between the level of non-Niger Congo ancestry in the individuals from East and South Africa and their respective NDR scores. The observation of very strong correlation in both South African ($R = 0.99$, $p<2.2e-16$; S Figure 7a) and East African ($r2 = 0.93$, $p<2.2e-16$; S Figure 7b) datasets suggests that genotype discordance between imputed datasets generally increase with increase in non-Niger Congo ancestry in a genome. Therefore, the same panel might perform differently in populations from different geographic regions of the continent."

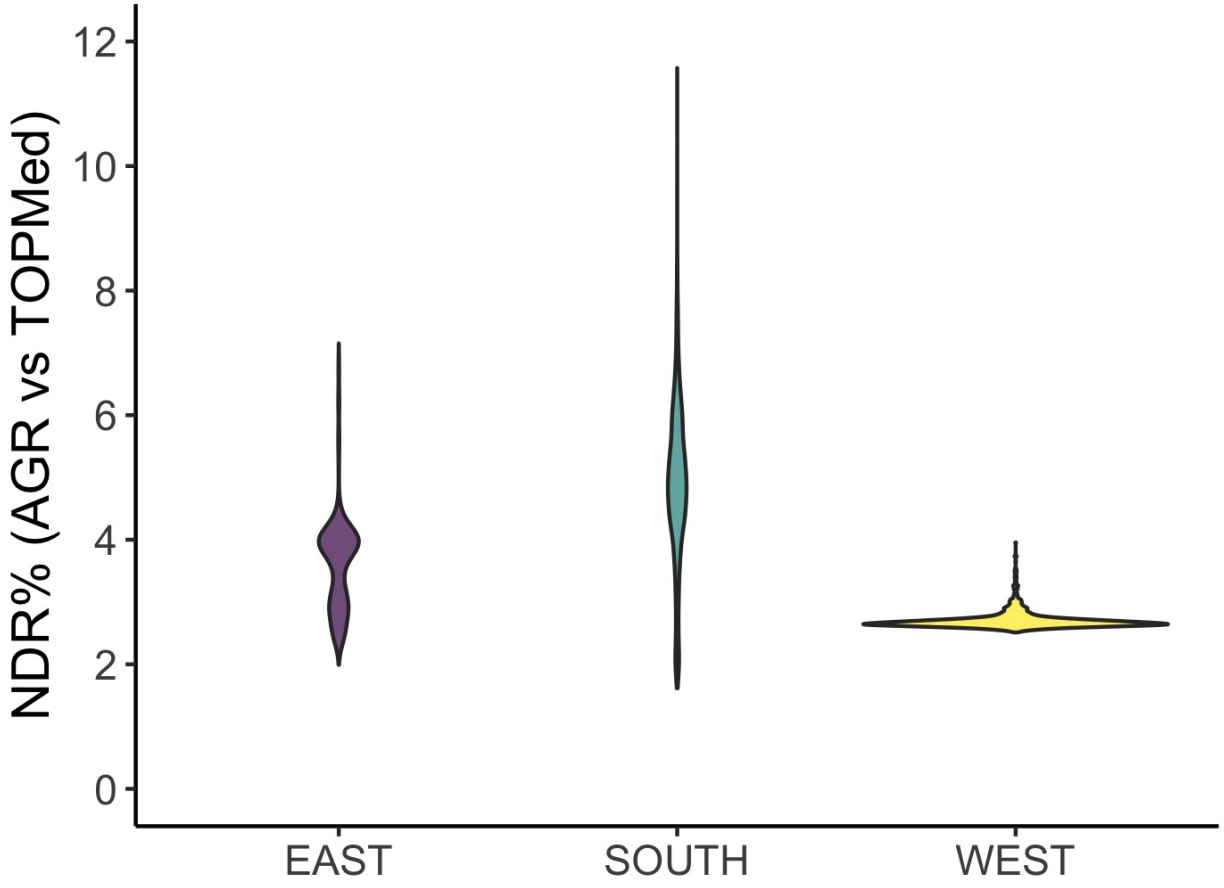
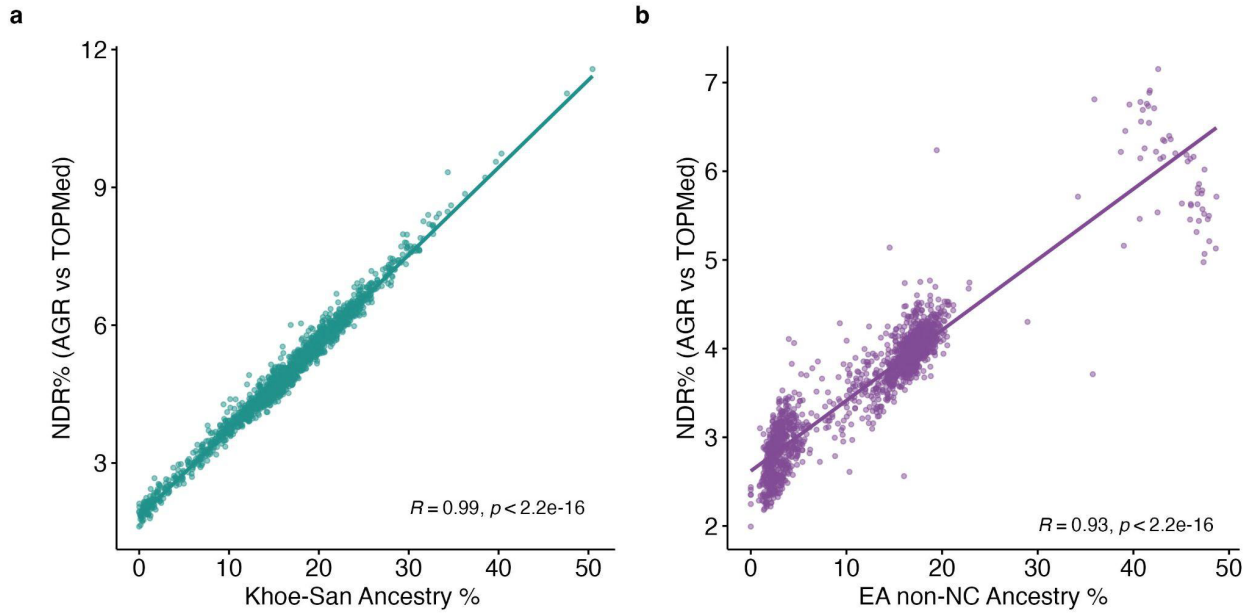


Figure 4D. Plot comparing the distribution of non-reference discordance rate (NDR) between genotypes imputed using AGR vs TOPMed in the East, West and South African populations. The NDR is almost constant across the dataset for West African participants, while the NDR shows substantial variation among the South African participants.



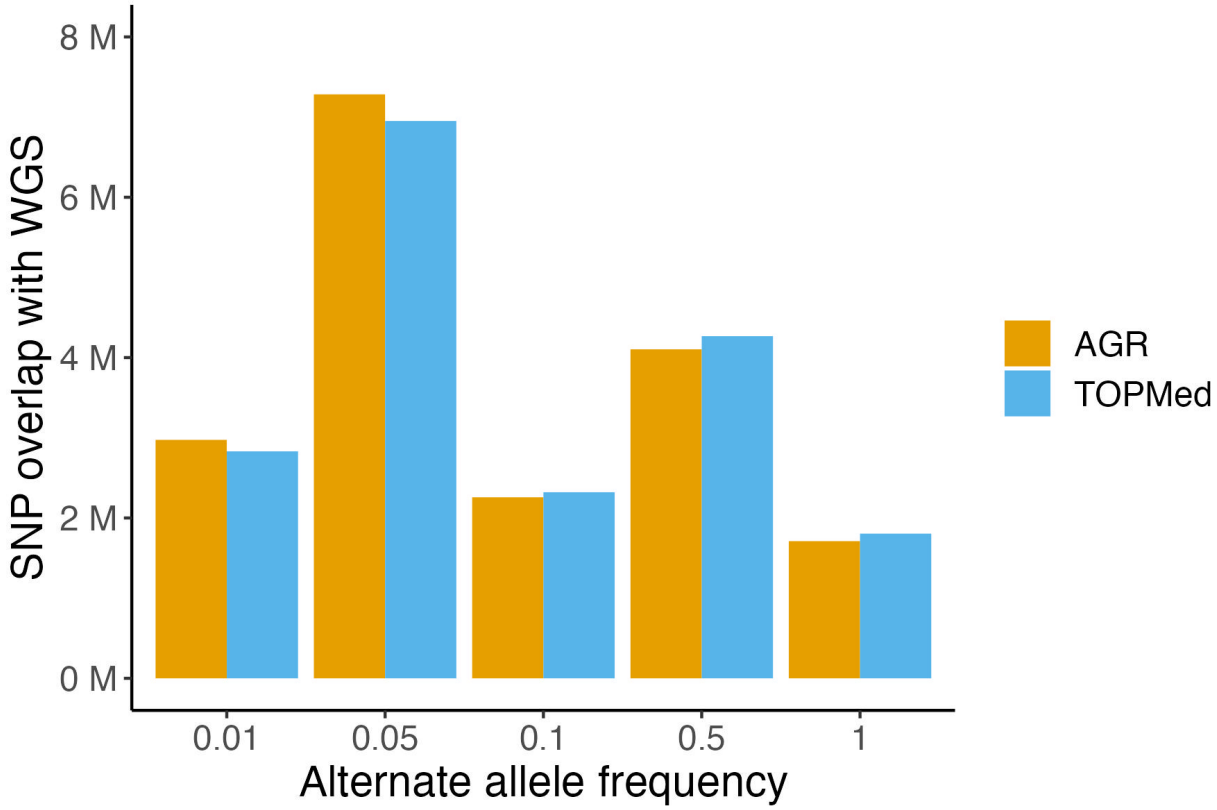
S Figure 7. Impact of ancestry on non-reference discordance rate (NDR) between genotypes imputed using AGR and TOPMed. Correlation between NDR and (a) the level of Khoe-San ancestry in South African participants (b) the level of east African non-Niger-Congo (EA non-NC) ancestry (Afro-Asiatic/Nilo-Saharan/Eurasian ancestry) in the east African participants. The ancestry proportions were inferred using ADMIXTURE (see Figure 5).

6) It would be informative to see SNP coverage (compared to WGS data) by allele frequency bin for AGR and TopMed to see if coverage diverges at specific allele frequencies. This is also an important point for discussion, in terms of implications for GWAS and fine mapping

Response: Based on the suggestion we have added a plot (S Figure 8) comparing the number of SNPs in the TOPMed and AGR imputed datasets that matched the WGS data, by allele frequency bins. The plot (please see below) shows the AGR to have better coverage for rare to moderate frequency SNPs (<0.05).

We have added the following lines in the results section to introduce this figure and describe the observation.

"The partitioning of SNPs that were shared by the imputed dataset and the WGS, by allele frequency bins (S Figure 8) further shows the AGR to better represent rare to moderate frequency SNPs (<0.05) compared to TOPMed, while TOPMed marginally better represents more common SNPs."



S Figure 8. Number of SNPs imputed by AGR and TOPMed that overlap with WGS data across allele frequency bins.

Reviewer #2
Major Revisions
1. The introduction and especially the discussion should briefly tie back to the major motivation - the huge underrepresentation of SSA populations in GWAS. Evaluating the accuracy of imputation in SSA populations is a vital step towards improving their inclusion in GWAS. It would be great to cite more sources for underrepresentation of African and especially non-Western African ancestry populations in GWAS, e.g. only 2% of individuals included in GWAS are individuals of 'African ancestry' and vast majority of African ancestry populations in genetic studies are western African e.g. African Americans or Afro-Caribbeans (72%-93% in the GWAS catalog and ≥90% in gnomAD) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6494470/.

Response: We sincerely thank the reviewer for this suggestion. Based on the advice we have added the following text to the introduction and discussions and also added new references including the one mentioned in the comment.

Introduction:
"The sequence datasets included in global reference panels are predominantly based on West African populations and diaspora populations originating from this geographic region 2,4. Due to this geographic bias, these panels lack the representation of diversity from eastern, central and southern Africa, which might impact the imputation performance in populations from these regions more adversely compared to populations with predominantly West African ancestry."

Discussion

Text has been added to three different paragraphs:

"The under-representation of African ancestry populations in current GWASs has been flagged as a major concern and the need to improve this representation has been highlighted repeatedly in recent literature 15,22. Currently, only 2% of individuals included in GWAS are individuals of 'African ancestry' and vast majority of these African ancestry populations in genetic studies are related to western African e.g. African Americans or Afro-Caribbeans (72%-93% in the GWAS catalog and ≥90% in gnomAD) 23,24. Even among the limited SSA GWAS datasets that are around, many have been imputed using panels that are not -optimal for these populations. Therefore, evaluating the accuracy of imputation in SSA populations is a vital step towards utilizing the full-potential of these datasets and improving their inclusion in global GWASs."

"Despite the increase in overall representation of African ancestry in some of the more recent reference panels, there is a bias toward West African origin populations and lack of the representation of other African regions and ancestries in them."

"Moreover, the complex pattern flow of African ancestry to Caribbean Islands and different parts of North and South America 28 suggests that this lack of representation could also impact the imputation performance and accuracy for some of the diaspora populations"

2. I recommend including a plot that shows NDR by INFO/R^2 score bins above 0.6. Researchers, such as ourselves, often need to use a much more stringent quality score cut-off than employed in this study. It will be helpful to include the NDR rates across the entire range of quality score bins to evaluate the accuracy gained versus number of variants filtered out with more stringent cut-offs. Figure 5c is a striking and a key finding, but would be improved by showing NDR rates per quality score bins across the different reference imputation panels.

Response: Based on the suggestion, we have added a new table (Table 3) that shows the NDR rates as well as the number of SNPs across the full range of INFO/R2 score cut-offs.

INFO score	Non-Reference Discordance rate (NDR)					SNP Count (in Millions)				
	AGR	TOPMed	KGP_S	KGP_M	HRC	AGR	TOPMed	KGP_S	KGP_M	HRC
>0	2.23 ± 0.58	3.57 ± 1.88	7.01 ± 2.37	6.74 ± 2.32	7.64 ± 2.28	18.34	18.19	16.25	16.07	14.09
> 0.3	2.23 ± 0.58	3.59 ± 1.90	7.01 ± 2.37	6.71 ± 2.34	7.64 ± 2.28	18.34	18.15	16.24	16.01	14.08
> 0.4	2.23 ± 0.58	3.59 ± 1.90	6.99 ± 2.37	6.67 ± 2.35	7.63 ± 2.29	18.33	18.13	16.22	15.97	14.05
> 0.5	2.21 ± 0.58	3.58 ± 1.90	6.93 ± 2.36	6.61 ± 2.35	7.55 ± 2.28	18.30	18.11	16.15	15.91	13.97
> 0.6	2.18 ± 0.58	3.56 ± 1.90	6.80 ± 2.35	6.52 ± 2.34	7.37 ± 2.27	18.22	18.05	15.99	15.79	13.79
> 0.7	2.11 ± 0.57	3.50 ± 1.88	6.54 ± 2.31	6.23 ± 2.33	7.00 ± 2.24	18.02	17.93	15.65	15.51	13.41
> 0.8	1.95 ± 0.53	3.39 ± 1.82	6.01 ± 2.20	5.91 ± 2.24	6.30 ± 2.12	17.47	17.60	14.86	14.83	12.60
> 0.9	1.55 ± 0.43	2.97 ± 1.63	4.70 ± 1.86	4.66 ± 1.94	4.73 ± 1.76	15.59	16.43	12.68	12.79	10.54

3. It would be insightful to show the ancestry-specific NDR rates for the WGS samples. Is imputation accuracy worse for Khoe-San ancestry segments than Niger-Congo segments in the WGS samples or has the inclusion of 84 Nama (Khoe-San) samples in AGR been sufficient to impute Khoe-San ancestry at comparable accuracies? Your conclusions about the presence of Khoe-San genomes in AGR positively impacting imputation in South African samples could be strengthened if you incorporated this analysis.

Response: We are grateful for this wonderful suggestion. Based on the reviewer's advice we

have included Figure 5D (Please see below) that compares global Khoe-San ancestry levels to the NDR between WGS and imputed data. As evident here, the NDR in the South African populations is highly correlated with the level of Khoe-San ancestry. The inclusion of the Nama which has substantial Khoe-San ancestry probably leads to a much lower discordance and a gentler slope in the AGR compared to all other panels. We have added a section in the results to introduce this figure and point out its relevance.

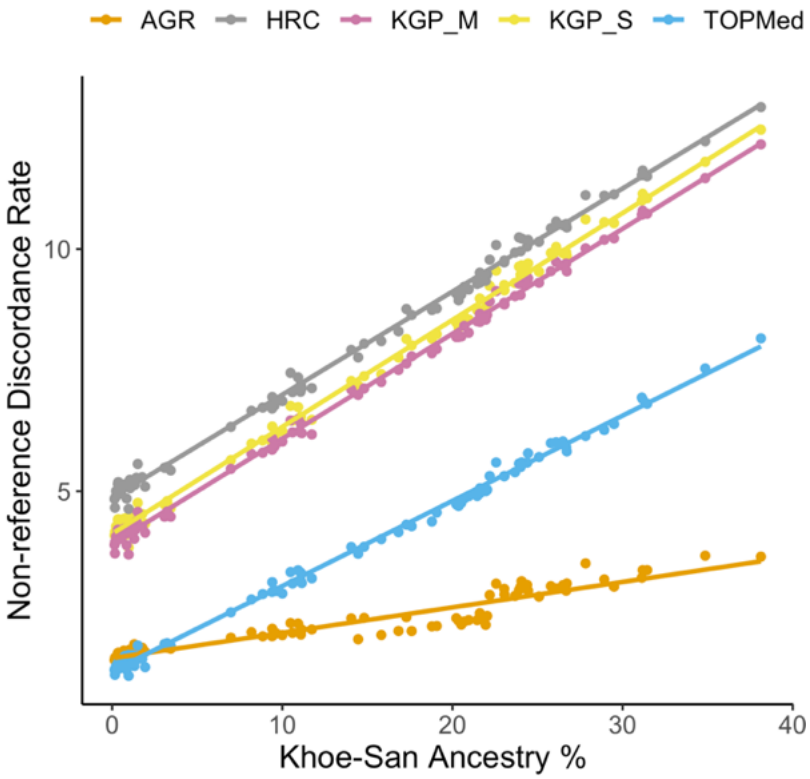


Figure 5D. Correlation between the overall genotype discordance (estimated by NDR) and the level of Khoe-San ancestry in the five imputed datasets. The inclusion of the representative Khoe-San population probably leads to a much lower discordance and a gentler slope in the AGR compared to other panels.

Suggested Revisions:
1. More details on samples in AWI - at least include brief breakdown of the ethnic groups in each region, as there is strong population structure in places like Kenya and South Africa.

Response: This is a great suggestion. Although this information was provided in our previous publication, it is not straightforward for the reader to link. Therefore, we have referenced the table from our previous publication and added the full list of ethnolinguistic groups included in the AWI-Gen study in S Table 1. A line has been added to the introduction to convey this : "The list of ethnolinguistic groups in the AWI-Gen dataset is provided in S Table 1, more details can be found in Ramsay et al.9 "

S Table 1. Self-reported ethnic distribution of AWI-Gen participants across the four countries

Country (study centre)	Ethnolinguistic group
South Africa (Agincourt, Dikgale and Soweto)	Tsonga, BaPedi, Zulu, Sotho, Tswana, Xhosa, Swati, Venda, Ndebele, Other[a], Unknown[b]
Burkina Faso (Nanoro)	Mossi, Gourounsi, Peulh, Dagara, Dioula, Samo, Gourmatche, Other[a], Unknown[b]
Ghana (Novrongo)	Kassena, Nankana, Bulsa, Mampruga, Frafra, Kantosi, Mossi, Other[a], Unknown[b]
Kenya (Nairobi)	Kikuyu, Kamba, Luo, Luhya, Kisii, Somali, Meru, Embu, Borana, Gari, Kalenjin, Maasai, Other[a]

[a] Only one or two individuals in a specific ethnic category.
[b] Person did not provide information on ethnicity.

2. Provide an explanation for your determination of the 3 Khoe-San ancestry and 3 non-Niger Congo ancestry proportion bins. How did you choose to split the data before testing for group mean differences needs to have clear a priori reasoning? Why not just run a regression between ancestry fraction and imputed SNP count by individual?

Response: We thank the reviewer for raising this. Although the classification was based on the distribution of ancestry levels in each dataset, we agree that a regression plot is a more accessible and intuitive approach to represent the data. Therefore, we have amended Figure 4 (please see below) accordingly.

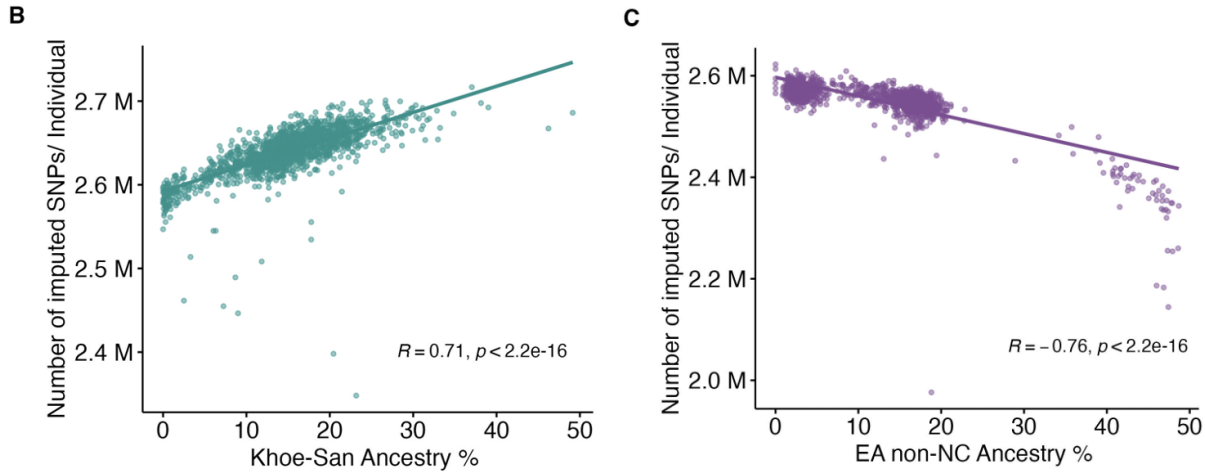


Figure 4B and 4C. Correlation between the number of SNPs imputed per individual by AGRand (B) the level of Khoe-San ancestry in South African participants. (C) the level of East African non-Niger-Congo (EA non-NC) ancestry (Afro-Asiatic/Nilo-Saharan/Eurasian

ancestry) in the East African participants. The ancestry proportions were inferred using ADMIXTURE (see S Figure 5).

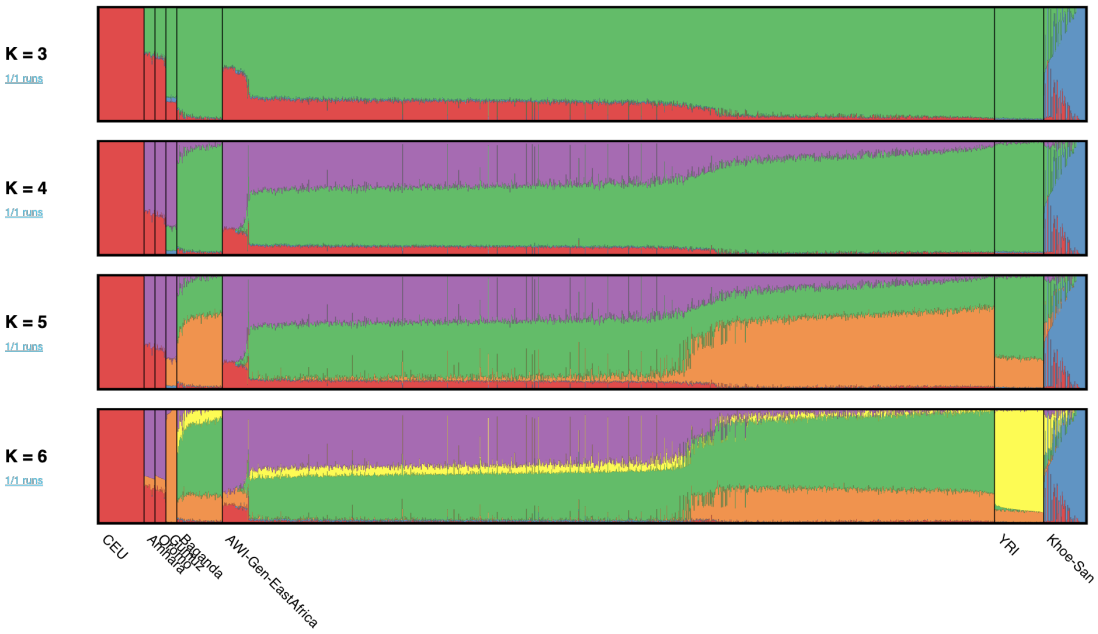
3. Related to the prior comment: you see a strong decline in the number of imputed SNPs per individual for genomes with greater AA or NS ancestry. I'm not totally convinced this is due to AA and NS genomes having lower SNP diversity than Bantu-speaking populations (page). I'll note that there is only 1 NS- speaking population in AGR which is the Gumuz (n=24), and even the other Ethiopians are not especially ancestrally closely related to Kenyans (Gopalan et al. 2022, Current Biology). I wonder rather if this is a mismatch between the ancestry of the AWI Kenyans and Eastern African references in AGR?

Response: We are thankful to the reviewer for this important observation. Indeed, a proper categorization of non-Niger-Congo ancestry in East Africans participants is complex and beyond the scope of the current study. A separate in-depth study from our group, that is currently underway, is examining this and exploring similarity between our Kenyan groups to corresponding East African groups included in previous studies.

As the objective of this particular analysis was only to test whether differential non-Niger-Congo ancestry leads to differences in imputation of the East African participants, we have considered ancestry composition at K=3 (Admixture plot below) to delineate a composite non-Niger-Congo ancestry (red) in addition to Niger-Congo (green) and Khoe-San (blue) related ancestries. We have amended the figure, the figure legend and main text to reflect this. We have also added a caveat in the text stating that the extent of impact of the non-Niger-Congo ancestry presented here might be altered if more complex ancestry deconvolution is employed.

We also agree that the connection between SNP counts in the Ethiopian and Baganda genomes and that between the imputed dataset with high and low Niger-Congo ancestry from Kenya is speculative. The following line has been added to the section to clarify this:

“However, large pan-African high-coverage WGS datasets will be required to assess if there are systematic differences in SNP count between East African ancestral groups and also whether the differences observed in our data corresponds to inherent genomic features or the varied ability of existing panels to impute genotype for such ancestral groups.”



4. Were ambiguous A/T C/G SNPs filtered prior to imputation? https://www.frontiersin.org/articles/10.3389/fgene.2019.00034/full Schurz et al., 2019 found a big difference in imputation accuracy when excluding these SNPs because of strand bias on arrays.

Response: In order to keep the pre-imputation QC to as minimal as possible, we did not filter out A/T/ C/G SNPs prior to imputation.

Based on the reviewer's suggestion we have imputed the 95 samples (that have WGS data) using the AGR panel after filtering for A/T G/C SNPs. The tables below compare the imputed datasets that were generated using genotype datasets with and without this filtering. The results do not show any clear difference in imputation either in terms of number of SNPs imputed, number of imputed SNPs with info score >0.6 or average info score per frequency bin .

Table A. Imputation evaluation without filtering for palindromic SNPs

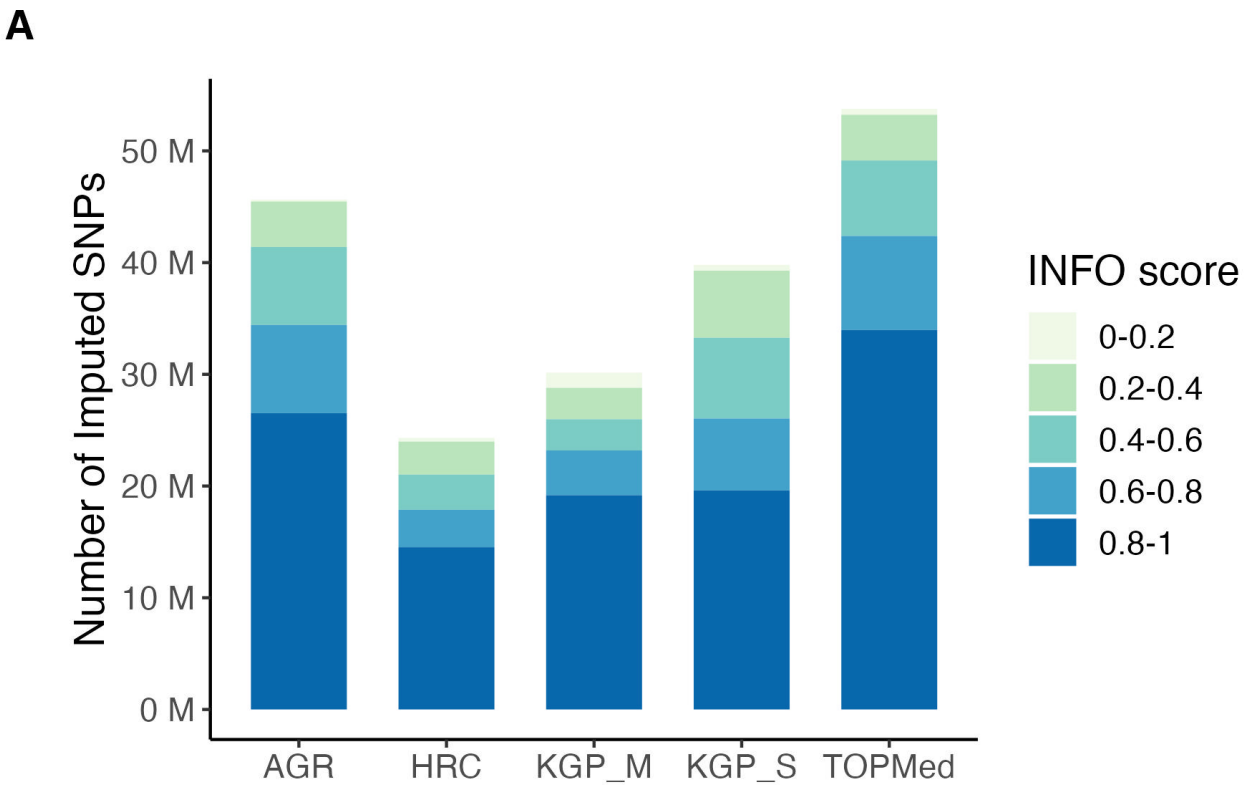
Allele Freq Bins	SNP Count	Avg INFO/R2	SNP Count (INFO > 0.6)
0.005-0.01	3751638	0.884066	3407415
0.01-0.05	7585581	0.948447	7519606
0.05-0.1	2311526	0.978196	2309856
0.1-0.5	4203341	0.986878	4200875
0.5-1	1691243	0.984693	1687073

Table B. Imputation evaluation with palindromic SNPs filtered during Pre-Imputation QC

Allele Freq Bins	SNP Count	Avg INFO/R2	SNP Count (INFO > 0.6)
0.005-0.01	3752839	0.883059	3405105
0.01-0.05	7581167	0.947697	7513727
0.05-0.1	2311099	0.977642	2309306
0.1-0.5	4202862	0.986392	4200354
0.5-1	1691269	0.984132	1687018

5. It would be helpful to show the complete distribution of INFO/R^2 scores in addition to Figure 2a) showing proportion variants over 0.6. A figure showing the overlapping distributions of INFO/R^2 scores across the different imputation panels would be informative. It is especially interesting that at high allele frequencies the mean INFO score in AGR out performs TopMed (Fig 2b), although perhaps this is a function of the Sanger IS vs MIS.

Response: Based on the reviewer's advice we have updated Figure 2A (updated version below) to show the number of imputed SNPs across all R2 or INFO score bins. However, we were not very sure about the second part of the comment. We hope our amended figure addresses the figure suggestion.



Minor Comments:

1. Methods: NDR equation format does not display correctly in my pdf
2.

Response: We thank the reviewer for pointing this out. We have updated the equation, hopefully, it will be displayed correctly in the pdf file now.

2. Include a more precise definition of Khoe-San under "Influence of ancestry and admixture". For example, "Khoe-San is a collective term for populations across Southern Africa that predate the expansion of Bantu-speakers; these populations descend from the earliest human population divergence and thus harbor greater genetic diversity and higher SNP content…"

Response: We appreciate this comment and the text suggestion. We have added the following description to the main text (Result section 3, Paragraph 2).

"Khoe-San is a collective term for populations across Southern Africa that predate the expansion of Bantu-speakers; these populations descend from the earliest human population divergence and thus harbor greater genetic diversity and higher SNP content 17–20."

3. Caption for S figure 6 is reversed: red circles are quality score < 0.6

Response: We thank the reviewer for the careful observation. We have amended the caption for the figure (now S figure 10) accordingly.

4. Footnote in Table 1, the Nama here are from South Africa - not Namibia (can cite van Eeden et al. Genome Biology 2022 for the data deposition).

Response: Thanks for the details and the citation. We have corrected the text and added the

citation.

Referees' report, second round of review

Reviewer #1: I commend the authors for this important and thorough assessment of reference panels for African genomics and comprehensive response. I think this is a valuable piece of work with important messages that are relevant to African and global medical genomics research. That representation of ancestry even in a smaller number of genomes can markedly improve imputation accuracy, even for rare SNPs compared to a much larger reference panel with poorer representation is an important observation. It's vitally important to capture diversity over increasing sample sizes of relatively genetically homogeneous populations.

A few minor points that would be good to clarify in the text:

1. The authors suggest that non-Niger Congo ancestry correlates inversely with SNP diversity. And they find that Ethiopian genetic diversity is lower than Ugandan diversity. It is very likely that the non-Niger Congo ancestral component that is associated with lower diversity is the Eurasian component. There's a breadth of evidence that suggests that genetic diversity in many populations within Africa inversely correlates with the amount of Eurasian ancestry. E.g. Ethiopian populations are an admixture of Afro-asiatic and Eurasian ancestry, with 40-50% Eurasian ancestry, which is associated with reduced heterozygosity. It's perhaps important to make this clear, as it's very likely that Afro-Asiatic and Nilo-Saharan components may be genomically diverse, and the factor driving the correlation here is the level of Eurasian ancestry. I suspect the correlation with Eurasian ancestry would be even stronger than the correlation with non-Niger congo ancestry as that's the component driving the reduced diversity (unsurprising given genomic diversity is lower in Eurasia due to the out-of-Africa bottleneck)

2. I understand the difficulty with recalling all ref/ref sites - I think it's helpful that the potential bias introduced by the method has been clarified in the text, but looking at the numbers presented, it does appear that if ref/ref sites had been included, it would have increased the NRD of TopMed imputation greater than AGR, and further tilted the results in favour of the AGR panel in terms of accuracy (given TopMed appears to have a higher false positive rate). I think it's worth making this more explicit in the text included- that using an NDR with ref/ref sites would very likely tilt further in favour of AGR, given the higher ref discordance for TOPMed.

3. Should Non-reference discordance be abbreviated as NRD rather than NDR throughout the manuscript?

Reviewer #2: The authors have satisfactorily answered my questions and suggested revisions. I think the results presented here will be a very useful benchmark for imputation in African genomics.

Authors' response to the second round of review

Reviewer #1: I commend the authors for this important and thorough assessment of reference panels for African genomics and comprehensive response. I think this is a valuable piece of work with important messages that are relevant to African and global medical genomics research. That representation of ancestry even in a smaller number of genomes can markedly improve imputation accuracy, even for rare SNPs compared to a much larger reference panel with poorer representation is an important observation. It's vitally important to capture diversity over increasing sample sizes of relatively genetically homogeneous populations.

We sincerely thank the reviewer for the follow-up comments. The inputs of both the reviewers have been key to the development of the manuscript, and we hope the additional clarifications would be helpful in improving this further.

A few minor points that would be good to clarify in the text:

1. The authors suggest that non-Niger Congo ancestry correlates inversely with SNP diversity. And they find that Ethiopian genetic diversity is lower than Ugandan diversity. It is very likely that the non-Niger Congo ancestral component that is associated with lower diversity is the Eurasian component. There's a breadth of evidence that suggests that genetic diversity in many populations within Africa inversely correlates with the amount of Eurasian ancestry. E.g. Ethiopian populations are an admixture of Afro-asiatic and Eurasian ancestry, with 40-50% Eurasian ancestry, which is associated with reduced heterozygosity. It's perhaps important to make this clear, as it's very likely that Afro-Asiatic and Nilo-Saharan components may be genomically diverse, and the factor driving the correlation here is the level of Eurasian ancestry. I suspect the correlation with Eurasian ancestry would be even stronger than the correlation with non-Niger congo ancestry as that's the component driving the reduced diversity (unsurprising given genomic diversity is lower in Eurasia due to the out-of-Africa bottleneck)

Response: We completely agree with the reviewer about the need for specifying that the differences could be driven by Eurasian ancestry and not the African ancestries that are clubbed together under non-Niger Congo. We have added the following lines to the main text to address this concern:

≈ However, large pan-African high-coverage WGS datasets will be required to assess if there are systematic differences in SNP count between East African ancestral groups and also whether the differences observed in our data corresponds to inherent genomic features or the varied ability of existing panels to impute genotype for such ancestral groups. Nevertheless, the much higher genomic diversity of African populations compared to the neighbouring Eurasian populations, lead us to speculate that the observed differences among the East African participants could be primarily driven by underlying differences in levels of Eurasian component that is included under the non-Niger Congo ancestry.≈

2. I understand the difficulty with recalling all ref/ref sites - I think it's helpful that the potential bias introduced by the method has been clarified in the text, but looking at the numbers presented, it does appear that if ref/ref sites had been included, it would have increased the NRD of TopMed imputation greater than AGR, and further tilted the results in favour of the AGR panel in terms of accuracy (given TopMed appears to have a higher false positive rate). I think it's worth making this more explicit in the text included- that using an NDR with ref/ref sites would very likely tilt further in favour of AGR, given the higher ref discordance for TOPMed.

Response: We thank the reviewer for the suggestion. We too anticipated the use ref/ref would further tilt the results towards AGR and had therefore included a few lines in the discussion to clarify this. Based on the reviewer˜ s advice we have amended the lines which now reads:

≈ As an indirect estimate of the level of difference that might be observed if Reference/Reference sites in the WGS were included in NDR estimation, we have noted the number of sites (Figure 5B) that were Reference/Reference in the WGS dataset but had at least one non-Reference allele in each of the imputed datasets. The observation of a considerably higher number of such sites for the TOPMed panel (1.45 million) compared to the AGR panel (0.93 million) hints that the consideration of Reference/Reference sites in the comparisons could further augment the difference in discordance between these panels and the WGS, as the discordance estimates for TOPMed would increase by much larger values compared to AGR.≈

3. Should Non-reference discordance be abbreviated as NRD rather than NDR throughout the manuscript?

Response: The acronym NDR for non-reference discordance rates has been used in several previous publications. While we agree with the reviewer˙ s suggestion that the use of NRD or perhaps NRDR would perhaps make this more explicit, amending an established acronym might also create confusion for some readers, therefore, we decided to adhere to this.