nature portfolio

Peer Review File

A draft UAE-based Arab pangenome reference

Corresponding Author: Dr Mohammed Uddin

This manuscript has been previously reviewed at another journal. This document only contains information relating to versions considered at Nature Communications. Mentions of prior referee reports have been redacted.

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 #1

(Remarks to the Author)
The authors have done a good job of addressing my major concerns. I'm very glad to see the assemblies are now being openly shared. Given that the assemblies will be made openly available, I would only request that the publication be released with the GenBank accessions of the assemblies.

There are a few rough edges in the text - e.g., I noted CPC was not disambiguated when it was first used, and in one place, the text talks about 100 plus reference diploid Arab genome assemblies, when, in fact, it is talking about 100 plus haplotypes. These small issues should be carefully checked.

Altogether though, I think this paper represents a new and significant resource that will be of great value for the community.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
Reviewer comments to the authors

Genomic references from underrepresented populations including the Middle East are valuable to advance genomic research regionally and in the world. The current study aims to establish a pangenome reference for 'Arab populations' based on de novo assembled genomes form 53 individuals recruited from the United Arab Emirates (UAE). The authors use established methods to generate and process data, producing de novo assemblies which they incorporate to a pangenome graph and highlight some features from it. In principle the idea is valuable, however the implementation shows several issues that persist despite the multiple revisions. This includes lack of novel insights, weakness/inconsistency in population structure and other analyses, over statements on novelty, as well as ambiguity on ethical processes for data sharing. Notably, there is lack of purposeful study design to ensure the selection of representative ancestries that would be needed to cover wide Arab populations beyond UAE. As a result, this draft pan genome should be named UAE pan genome instead of Arab pangenome. The below elaborates on the main points.

Samples and study design

• The samples in this study were not targeted to be representative based on genetic ancestry. The 53 subjects were recruited from the UAE, whose population represent less than 1.9% of all Arab populations. The authors added a few expats from the region but still recruited form the UAE instead of engaging collaborators from other counties who know better about the local diversity. For example, two Moroccan nationals were but is well known that Morocco is predominantly native Berber (not Arab). See for example HPRC protocol that was established for sample selection to ensure added value in terms capturing diversity.

• In previous submissions authors indicated data was under restricted access due to IRB. Since authors now say informed written consent has been obtained, they should share these documents to ensure the protection of human subjects in this study.
• Authors mention study subjects were randomly selected from clinics in Dubai. Amongst the 53 individuals at least a few would have been expected to have some common complex disorder phenotypes. However none are which is contradictory.
• Data availably does not mention anything about HiC and raw PacBio and ONT read data.

Population structure

Persists to be severely undeveloped:
• The admixture analysis that was originally done had multiple inconsistencies with what has been previously established in the literature, and that persists here. Now, the admixture plots are omitted entirely from main and supplementary figures.
• The PCA shown in Figure 1e is still problematic showing East Asian clustered next to Americans. Notably, the UAE samples shown in red are shifted to the right in comparison to the reference Arab samples shown in pink. This reflects the lack of a priori proper selection of samples to ensure wide representation.

Data generation

• Given that various data types have been generated for the subjects in this study, there is no attempt to check for sample swaps and contamination

Assembly metrics

• There is mention of exclusion of “non-human eukaryotic pathogen genomes” in addition to 109 contigs from ChrM and exclusion of 4 contigs from other chromosomes based on sequence identify. What is the total length of what was excluded? From which individual assemblies? And what is the reason for the presence of pathogenic sequences? What is the possibility that some of what was labelled as mis-joining could be a genuine chromosomal rearrangement?
• In Suppl Table 11, authors report #Ns, #indels and #mismatches per 100kb. Should report the total numbers instead.
• In Suppl Table 11, the numbers of misassembled contigs relative to CHM13 as a fraction from number of contigs is on average 56%. This is too high.
• Authors make the statements “All samples surpassed the 40 Mb N50 of the HPRC reference genome”. However the N50 they report include the misassembled conigs shown in Supl Table 11.
• QV for base pair quality was it calculated using the long read data used for the assemblies? If yes, what bias this could constitute?
• Unclear how the coverage was computed exactly.
• The reported average size of 50/77 Mb of contigs that did not align to CHM13/GRCh38 does that include the misassembled contigs mentioned previously? Same question for the results reported regarding phasing quality etc

Gene duplication analysis

• This section shows several inconsistencies. Results were generated showing multiple copies of genes across the assemblies (Suppl Table 14a), but authors only discuss the duplications (Fig 2g). There is large variance in the number of genes with additional copies relative to GRCH38, ranging from 24 to 206 genes. The number of additional copies itself ranges between 1 and 46. Some samples have unrealistically high number of impacted genes such as APR_023_hap2 which has 206 genes with additional multiple copies relative to GRCH38, which is 4 times higher than the average, while hap 1 of the same sample has only 43 genes! Ultimately by looking at the trio, it is odd to see that the child has much lower number of genes with multiple copies in comparison to the father and the mother.
• In view of the above, I wonder how significant are the results of the comparison with HPRC.

Variant analysis

• No mention of what reference the variant calls were made against and if any filtering was applied including the low-quality regions of the assemblies.
• The authors report a high proportion of novel variants, including for SNVs and Indels where it is 15-19% per genome. This is odd because for this type of variation there is genomic completeness in the references used, and samples from Middle East have already been extensively sampled in the literature. The authors do not attempt to check in GNOMAD and other Middle Eastern cohorts such as Mineral et al. 2021 and Scott et al. 2016.
• For the SV calls, there is no attempt to QC the calls which are more prone to mis-assembly errors

Novel euchromatic sequences

• Authors report their assemblies contain 112 Mbs of novel sequence relative to CHM13, other pan genomes, and 1.4 kb of Mt DNA. This amounts to ~ 4% of the human genome which seems excessive and may not as striking. How much does that compare to the inverse situation i.e the amount of novel sequences in HPRC samples relative to the samples in this study?

Complex structural variation

• Strange the criteria used here for naming this category of SVs. Authors should just refer to SVs larger than certain length and frequent above a certain threshold.

• Again novelty claimed without assessing relevant datasets beyond HPRC+CPC.

Mitochondrial pangeome

• No mention which reference variants were called against
• The Mt assemblies are based on PacBio only. No mention of what particular QC was performed for these genomes nor what mis-assemblies were found

Performance gain from pangenome-aided

• Higher calling rate using Pan genomes has been shown before so that is not specific to the APR graph.
• The mapping rate is between the APR and HPRC is almost identical. Unclear what is the added value here

Discussion

To be revised based on the comments above.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

(Remarks on code availability)

Reviewer's Comments:

Reviewer #1 (Remarks to the Author)

The authors have done a good job of addressing my major concerns. I'm very glad to see the assemblies are now being openly shared. Given that the assemblies will be made openly available, I would only request that the publication be released with the GenBank accessions of the assemblies.

There are a few rough edges in the text - e.g., I noted CPC was not disambiguated when it was first used, and in one place, the text talks about 100 plus reference diploid Arab genome assemblies, when, in fact, it is talking about 100 plus haplotypes. These small issues should be carefully checked.

Altogether though, I think this paper represents a new and significant resource that will be of great value for the community.(Remarks on code availability)

Response: We thank the reviewer for their thoughtful feedback and constructive suggestions. We are pleased to have addressed the major concerns raised. As requested, we have included the GenBank accession numbers for the assemblies in the supplementary table (Supplementary Table 35). We have carefully reviewed the final manuscript to correct any typographical errors. The abbreviation "CPC" is clarified as "Chinese Pangenome Consortium" at its first mention in the main section. Furthermore, the reference to "106 assemblies" has been corrected to "106 haplotypes".

Reviewer #2 (Remarks to the Author)

Reviewer comments to the authors

Genomic references from underrepresented populations including the Middle East are valuable to advance genomic research regionally and in the world. The current study aims to establish a pangenome reference for 'Arab populations' based on de novo assembled genomes form 53 individuals recruited from the United Arab Emirates (UAE). The authors use established methods to generate and process data, producing de novo assemblies which they incorporate to a pangenome graph and highlight some features from it. In principle the idea is valuable, however the implementation shows several issues that persist despite the multiple revisions. This includes lack of novel insights, weakness/inconsistency in population structure and other analyses, over statements on novelty, as well as ambiguity on ethical processes for data sharing. Notably, there is lack of purposeful study design to ensure the selection of representative ancestries that would be needed to cover wide Arab populations beyond UAE. As a result, this draft pan genome should be named UAE pan genome instead of Arab pangenome. The below elaborates on the main points.

Samples and study design

• The samples in this study were not targeted to be representative based on genetic ancestry. The 53 subjects were recruited from the UAE, whose population represent less than 1.9% of all Arab populations. The authors added a few expats from the region but still recruited form the UAE instead of engaging collaborators from other counties who know better about the local diversity. For example, two Moroccan nationals were but is well known that Morocco is predominantly native Berber (not Arab). See for example HPRC protocol that was established for sample selection to ensure added value in terms capturing diversity.

Response: We thank the reviewer for raising this important point. However, we respectfully contend that the draft pangenome we constructed covers a significant gap in the global genomic literature despite the relatively small sample size. This is not very different from previous important work in other populations. For example, HPRC developed a draft global pangenome using 47 samples, including only 2% from Europe and just one sample from South Asia, a region with a population exceeding 2 billion. Similarly, the Chinese Pangenome includes one or two samples from most of its geographically distinct populations. While the sample sizes for these pangenomes remain small, the assemblies successfully capture common sequence diversity across a wide range of ethnicities and global populations. Likewise, the Arab Pangenome includes samples from eight Arab countries, representing most of the major populations within the MENA region.

The point regarding the two Moroccan nationals suggests a conflation of genetics, ethnicity and language. The population structure of Moroccans (Berber/Arab) has been published extensively, and they show the same ancestral components (Arauna et al., 2017), “the present analysis of additional Berber samples reinforces the idea of no strong genetic distinction between Arabs and most Berber groups.”). Second, in contrary to the statement, Morocco is not predominately “native Berber (not Arab)”, as two thirds of the population identify as Arab. Consequently, both Arab and Beber groups would benefit from our pangenome data, which is crucial as there is no high-quality pangenome including Moroccan samples published so far. We present below an ADMIXTURE plot of our Moroccan samples with published Moroccan samples genotyped on the Human Origins Array from the Allen Ancient DNA Resource (AADR) (Figure 1). As shown, the ADMIXTURE plots are very similar, illustrating that our samples capture the variation found in Moroccans.

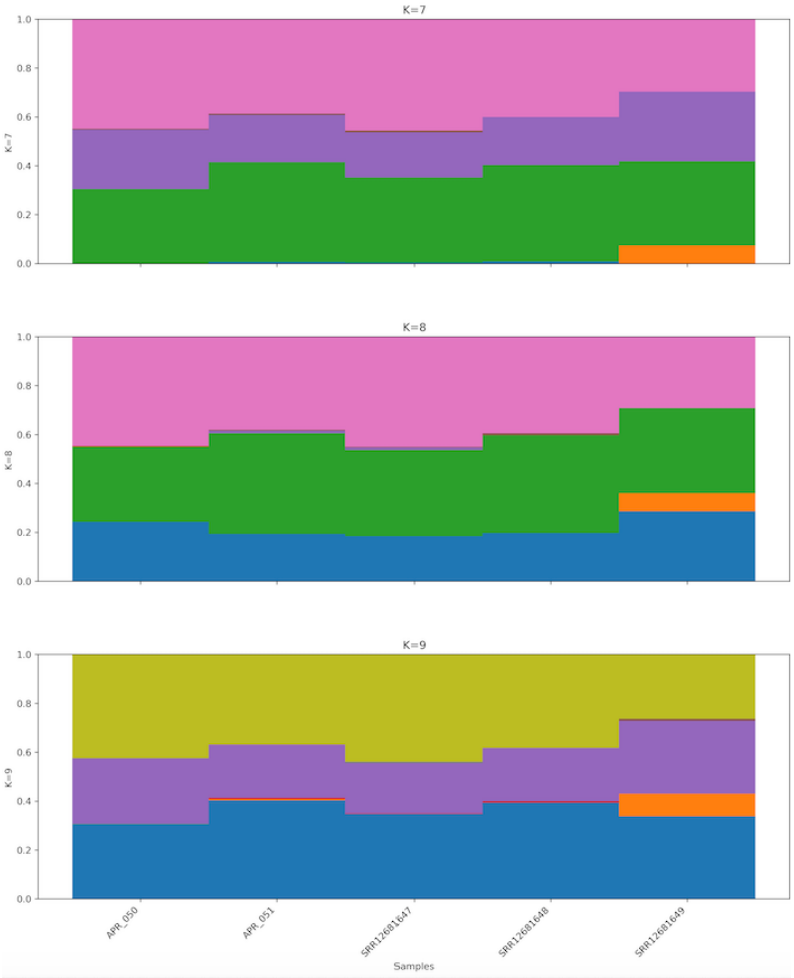


Figure 1: Population genetic ancestry inference for the Moroccan samples of the APR cohort. The ADMIXTURE plot shows ancestry components at K=7, K=8, and K=9 for two Moroccan samples from the Arab Pangenome Project alongside Moroccan samples from AADR dataset. Each bar represents an individual sample, with the proportions of ancestral components represented by different colors.

We would also like to emphasize that the ancestral ancient populations that gave rise to modern-day North African and Middle Eastern groups are similar, despite the geographic range. All modern-day Middle East samples (Egypt, Levant, Arabia etc) are a mixture of four ancient populations (Almarri et al., 2021), with current-day structure/cline resulting due slight differences of these source populations. Consequently, due to the ancestral history shared by all these groups, we believe that the pangenome will be relevant to all Middle Eastern and Arab populations.

• In previous submissions authors indicated data was under restricted access due to IRB. Since authors now say informed written consent has been obtained, they should share these documents to ensure the protection of human subjects in this study.

Response: We thank the reviewer for raising this point regarding data accessibility and consent. To clarify, at no point did we mention that the data was under restricted access. Our original intention was always to make the data publicly available upon the publication of the manuscript.

However, following the first review , we expedited this process and made the raw data publicly accessible at the review stage.

For reference, we have enclosed our earlier communication from the first rebuttal, which clearly outlines our actions to ensure data availability. Additionally, we confirm that copies of the informed consent documents will be provided to the editor to ensure transparency and compliance with ethical standards. The relevant part of the consent form reads:

“As part of this study, my de-identified data (i.e., data that does not include my name or any personal identifiers) will be uploaded to public research repositories. This data will be accessible to other researchers for future scientific studies aimed at advancing our understanding of genetics and related fields. My privacy will be protected, and no personally identifiable information will be shared.”

[REDACTED]

Snapshot to our previous response.

• Authors mention study subjects were randomly selected from clinics in Dubai. Amongst the 53 individuals at least a few would have been expected to have some common complex disorder phenotypes. However none are which is contradictory.

Response: We would like to clarify that random sampling was never mentioned in our study. Instead, our study employed convenience sampling with specific criteria to exclude individuals with chronic diseases. This approach ensured a cohort of apparently healthy individuals to establish a baseline for the Arab pangenome, as detailed in the methods and supplementary methods section (Page 4).

• Data availably does not mention anything about HiC and raw PacBio and ONT read data.

Response: All raw sequencing data, including Hi-C, PacBio, and ONT reads, have been uploaded to the global Sequence Read Archive (SRA) under the accession number PRJNA1108179.

The current data availability statement in the manuscript reads as follows:
"We have submitted our sequencing data (including PacBio, ONT) for all the samples to the global Sequence Read Archive (SRA) repository that can be openly accessed and downloaded under the accession no PRJNA1108179."

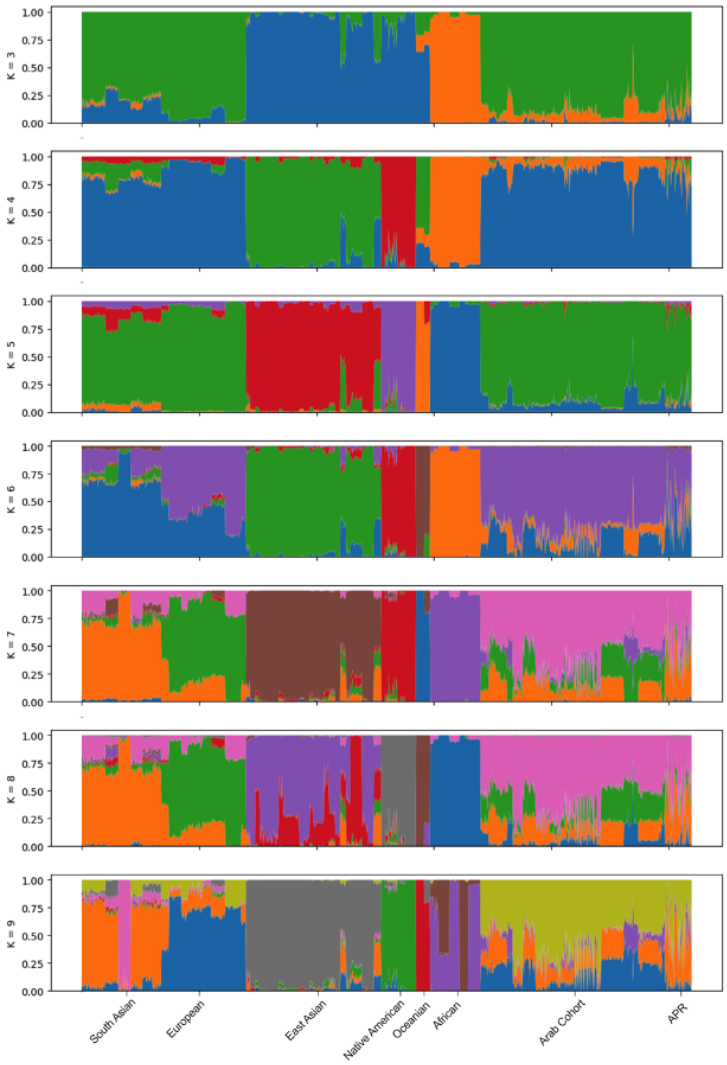
We revised this statement to include Hi-C data for completeness:
"We have submitted raw sequencing data (including PacBio, ONT, Hi-C) for all the samples to the global Sequence Read Archive (SRA) repository that can be openly accessed and downloaded under the accession no PRJNA1108179."

Additionally, details about the assemblies, pangenome, and VCF files can be accessed via the links provided in the data availability section.

Population structure

Persists to be severely undeveloped:
• The admixture analysis that was originally done had multiple inconsistencies with what has been previously established in the literature, and that persists here. Now, the admixture plots are omitted entirely from main and supplementary figures.

Response: We had updated the ADMIXTURE analysis and included the results in our previous submission, presented as Extended Data Fig. 1 (attached below). This analysis, consistent with the literature, highlights the APR cohort's genetic diversity in the context of global populations. Using linkage disequilibrium (LD) pruning (detailed in the methods section), we filtered SNPs with MAF > 0.05 and identified K=8 as the optimal value based on cross-validation.



Extended Data Fig. 1: Population genetic ancestry inference of the 53 APR samples using ADMIXTURE. Assuming ancestry components K ranging from 3 to 9, each plot shows independent ancestry fraction with its own color-coding representation labeled with short vertical lines. Samples included are from South Asian, European, East Asian, Native American, Oceanian, African and Arab ethnicities.

• The PCA shown in Figure 1e is still problematic showing East Asian clustered next to Americans. Notably, the UAE samples shown in red are shifted to the right in comparison to the reference Arab samples shown in pink. This reflects the lack of a priori proper selection of samples to ensure wide representation.

Response: We believe the PCA is correct, as East Asians will cluster near Americans. These samples are from the HGDP dataset which has been published in multiple studies (Bergstrom et al., 2020; Jakobsson et al., 2008; J. Z. Li et al., 2008) where the same patterns in the PCA emerge (Figure 2). East Asians and Americans will differentiate in subsequent principal components. We note that we have added other Middle Eastern samples to this PCA as we elaborate in the methods section. Please see below figures from the previous studies that support our statements above.

[REDACTED]

Figure 2.**a:** Principal component analysis (PCA) plot showing global population clustering (Figure 1E from Jakobsson et al., 2008, *Nature*). The PCA illustrates East Asians positioned near Americans, consistent with genetic proximity in subsequent principal components. **b:** PCA plot from (Figure S3B from Li et al., 2008, *Science*). The plot highlights similar clustering patterns, further validating the observed relationships among global populations, including East Asians and Americans.

Data generation

• Given that various data types have been generated for the subjects in this study, there is no attempt to check for sample swaps and contamination

Response: We appreciate your comment regarding sample validation. To ensure data integrity, we performed several quality checks, including manual verification of sample identities, sex chromosome checks to confirm consistency with metadata, and fineSTRUCTURE analysis to detect potential sample swaps or duplications. Additionally, relatedness analysis was conducted using identity-by-descent, and no anomalies were found from the analysis. Furthermore, we checked the number of variants in each sample as an additional quality control measure. Significant discrepancies in the number of variants would indicate potential sample contaminations or mix ups, but no such inconsistencies were observed. These measures provided confidence in the accuracy and contamination-free nature of the dataset.

Assembly metrics

• There is mention of exclusion of "non-human eukaryotic pathogen genomes" in addition to 109 contigs from ChrM and exclusion of 4 contigs from other chromosomes based on sequence identify. What is the total length of what was excluded? From which individual assemblies? And what is the reason for the presence of pathogenic sequences? What is the possibility that some of what was labelled as mis-joining could be a genuine chromosomal rearrangement?

Response: Thank you for your detailed comment. Non-human eukaryotic pathogen genomes were excluded to ensure the integrity of the assembly. Kraken, a widely used tool for read classification, was employed to identify and remove non-target reads that could have been naturally present in the individuals who donated samples or potentially introduced during DNA

extraction or sequencing (Cornet & Baurain, 2022; Q. Li, Yan, Lam, & Luo, 2022; Sherman et al., 2019). For example, in a previous African pangenome study based on short-read data, 29 individuals were found to have contigs related to malaria infections and 1 with human beta herpesvirus, which were excluded after screening (Sherman et al., 2019 Nature Genetics) On average, 7 HiFi reads (range: 0–15; representing ~0.00001% of total reads) with an average size of 61.8kb and 187 ONT reads (range: 11–1797; representing ~0.0003% of total reads) with an average size of 2.81 Mb per sample were removed. This represents a negligible fraction of the total sequencing data, and it is common to observe such levels of non-human eukaryotic reads in ONT data.

Regarding mitochondrial contigs, 109 mitochondrial contigs (average length: 16.29 kb) were removed and used to build a mitochondrial pangenome, which was separately analyzed.
Out of 10,997 total contigs across 53 assemblies, 4 contigs classified as misjoins were excluded. These were identified in APR031 hap1 (104.81 Mb), APR022 hap1 (98.41 Mb), APR003 hap1 (100.09 Mb), and APR043 hap2 (5.97 Mb). The total genome length of the respective samples was 2.95 Gb, 2.91 Gb, 2.95 Gb, and 3.04 Gb. Resolving misjoins is one of the most challenging aspects of genome assembly due to the involvement of rDNA and other complex repeat elements that can lead to inaccuracies in splitting the misjoined contigs.

• In Suppl Table 11, authors report #Ns, #indels and #mismatches per 100kb. Should report the total numbers instead.

Response: We report these metrics normalized per 100kb, as provided by QUAST, to allow comparisons with other studies that follow the same approach (see Table 1 below, where HPRC reported the same metric with values comparable to ours). This standardized format is widely used in assembly evaluations (Mikheenko, Prjibelski, Saveliev, Antipov, & Gurevich, 2018; Muñoz-Barrera et al., 2024; Zhang, Jia, & Wei, 2019) and provides a consistent framework for assessing assembly quality.

• In Suppl Table 11, the numbers of misassembled contigs relative to CHM13 as a fraction from number of contigs is on average 56%. This is too high.

Response: Thank you for your comment. We acknowledge the concerns regarding the reported fraction of misassembled contigs. To address this, we analyzed the Human Pangenome Reference Consortium (HPRC) assemblies using QUAST. While HPRC did not report the number of misassembled contigs in their manuscript, our analysis of their publicly available assemblies (https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/HPRC/HG00438/assemblies/), showed that the fraction of misassembled contigs in their dataset is comparable to the APR. Specifically, HPRC samples had an average of 147 misassembled contigs out of 371 total contigs, similar to the trends observed in the APR dataset (average 56 misassembled contigs out of 103 total contigs). For additional context, Table 1 presents the QUAST report for 10 HPRC samples, highlighting the similarities in assembly metrics.

Table 1: QUAST report of 10 HPRC samples

Sample	Total contigs	# misassembled contigs	N50 (bp)	# indels per 100 kbp	#mismatches per 100kb	Unaligned length (bp)
HG00438_Maternal	259	129	54936949	25.7	149.49	24947097
HG00438_Paternal	278	115	48061544	26.37	149.32	29891555
HG00735_Maternal	251	111	56474489	25.13	141.18	27588990
HG00735_Paternal	321	132	53422923	25.16	141.82	28779735
HG00741_Maternal	307	128	41001116	26	137.58	31296475
HG00741_Paternal	311	118	51040418	25.16	141.68	34453559
HG01123_Maternal	374	158	54362305	24.56	136.8	25535378
HG01123_Paternal	571	176	44719827	24.88	136.34	30607908
HG01175_Maternal	322	144	36535860	25.42	138.32	30713555
HG01175_Paternal	395	156	34803293	25.94	149.13	31661659
HG01358_Maternal	335	166	48753445	24.71	138.74	27266293
HG01358_Paternal	442	169	52599478	25.58	142.62	28700297
HG01361_Maternal	308	134	45122217	25.09	133.34	30143609
HG01361_Paternal	378	117	47178056	24.72	138.46	26966319
HG01891_Maternal	352	143	81112077	30.69	178.82	26657225
HG01891_Paternal	499	151	57096483	31.46	170.8	36542338
HG01952_Maternal	326	126	54639450	25.44	142.05	28299280
HG01952_Paternal	486	153	44250376	25.77	142.58	31846874
HG02148_Maternal	425	217	39938933	24.83	136.8	22557821
HG02148_Paternal	493	192	41874143	25.18	133.84	25156041

Moreover, it is important to consider that differences flagged as misassemblies by QUAST may represent true structural variations, such as rearrangements, large indels, gene duplication events, and variations in repeat copy numbers (Chawla, Kumar, & Shankar, 2016). These variations are an inherent feature of *de novo* genome assemblies, especially when dealing with complex and repetitive genomic regions.

• Authors make the statements “All samples surpassed the 40 Mb N50 of the HPRC reference genome”. However the N50 they report include the misassembled conigs shown in Supl Table 11.

Response: Table 1 demonstrates that the HPRC also reported N50 values within the 40 Mb range, which includes misassembled contigs. In our study, we have generated over 2,011 GB of sequencing data with reads exceeding 100 kb using Oxford Nanopore’s ultra-long read protocol. This significant amount of ultra-long read data has been pivotal in achieving our large N50 values. Moreover, the Japanese T2T group, as presented in the HPRC meeting, reported similar N50 values following a comparable methodology. We have ensured consistency in N50 computation by adhering to similar methods used by the HPRC, underscoring the robustness of our results.

• QV for base pair quality was it calculated using the long read data used for the assemblies? If yes, what bias this could constitute?

Response: Thank you for your comment. Multiple studies have demonstrated the use of Yak k-mer analysis on long-read data to estimate the base accuracy (QV) of contigs. This approach is becoming the new standard for assembly quality assessment, as long-read technologies provide high accuracy for assembly without requiring a hybrid approach. For instance, Benedict Paten's lab (Kolmogorov et al., 2023) recently applied Yak k-mer on long-read data for QV estimation, and no significant bias was observed in their analyses. Similarly, we utilized multiple long-read (PacBio 33x and ONT 54x ULK coverage) and Hi-C (65x coverage) data for building assembly and computed QV using PacBio reads. The Q scores achieved (33.1) using PCR free high fidelity PacBio long reads are comparable to or as reliable as those obtained using short reads, hence anticipated bias will be extremely minimum.

• Unclear how the coverage was computed exactly.

Response: Similar to HPRC and CPC, we computed coverage based on T2T-CHM13 and GRCh38 (excluding alt).

'Haploid assemblies with an X chromosome had an average total length of 3.02 Gb, representing 99.1% of the T2T-CHM13 (3.06 Gb). Conversely, haploid assemblies with a Y chromosome averaged a total length of 2.98 Gb, highlighting the inherent size difference between the sex chromosomes.'

• The reported average size of 50/77 Mb of contigs that did not align to CHM13/GRCh38 does that include the misassembled contigs mentioned previously? Same question for the results reported regarding phasing quality etc

Response: The unaligned length represents the total length of all segments from the assembled sequences that could not be aligned to the reference genomes (T2T-CHM13/GRCh38). This unaligned fraction likely includes individual-specific variations, insertion of new sequences, structural variations, gene duplication events or novel genomic regions that are absent or different in the reference genome.

Similar unmapped sequence data has been reported by the HPRC (refer to Table 1), further supporting the presence of novel regions or structural variations. Additionally, the unaligned sequences may encompass complex genomic rearrangements or repeats (i.e. rDNA), which remain challenging to map against reference genomes. This underscores the importance of comprehensive assemblies in capturing population-specific and individual-level genomic diversity.

Gene duplication analysis

• This section shows several inconsistencies. Results were generated showing multiple copies of genes across the assemblies (Suppl Table 14a), but authors only discuss the duplications (Fig

2g). There is large variance in the number of genes with additional copies relative to GRCH38, ranging from 24 to 206 genes. The number of additional copies itself ranges between 1 and 46. Some samples have unrealistically high number of impacted genes such as APR_023_hap2 which has 206 genes with additional multiple copies relative to GRCH38, which is 4 times higher than the average, while hap 1 of the same sample has only 43 genes! Ultimately by looking at the trio, it is odd to see that the child has much lower number of genes with multiple copies in comparison to the father and the mother.

Response: We appreciate your detailed comment regarding gene duplications and the observed variance. The APR cohort shows a gene duplication range of 24–206, with an outlier (a haplotype of APR_023 having 206 duplications) and only 4 haplotypes exceeding 100 duplications. The rest fall within the HPRC range (16–79). A similar pattern is observed in the HPRC dataset, where HG01358 has 79 duplicated genes on hap 2 and only 18 on hap 1. Such outliers are expected in these analyses due to natural biological variation, as outlier haplotypes may reflect individual-specific duplications. This pattern is not unique to our dataset and is part of the natural variation reported in genomic studies. For example, 60-80% of healthy individuals have a copy number variation (CNV) of at least 100 kb, while 5-10% have a CNV of at least 500 kb(Girirajan, Campbell, & Eichler, 2011).

In the HPRC dataset, additional gene copies range from 1 to 37 (Liao et al., 2023)(Supplementary Table 9), highlighting comparable observations across datasets. These findings underscore the inherent variability in gene duplication patterns across individuals and cohorts.

For the trio, the observed discrepancy in duplications may be due to various biological factors, including the variable inheritance of duplicated genes. Additionally, the high rate of loss of heterozygosity (LOH) in consanguineous populations, such as Arabs, may further contribute to these observed patterns. Such phenomena are indicative of the unique genetic dynamics within populations with high consanguinity, emphasizing the importance of considering population-specific genomic contexts in these analyses.

• In view of the above, I wonder how significant are the results of the comparison with HPRC.

Response: While the sample size for all three pangenomes remains relatively small, which may limit the ability to fully characterize the true distribution of duplicated genes per individual, our findings provide valuable insights. The number of unique duplicated genes observed in the APR (883 genes) and HPRC (801 genes) datasets is comparable, suggesting consistent trends in gene duplication between these populations. This consistency supports the validity of our comparison and highlights the utility of the APR dataset as a complementary resource to existing pangenome references. Further studies with larger cohorts will be necessary to refine these observations and fully capture the spectrum of gene duplication across diverse populations.

Variant analysis

• No mention of what reference the variant calls were made against and if any filtering was applied including the low-quality regions of the assemblies.

Response: Thank you for your observation. Variants from pangenome were called using T2T-CHM13 as the reference, which was also utilized as the backbone for graph construction in our study. Although this was previously mentioned in our manuscript, we explicitly clarified this in the methods and results section of the revised version.

We employed UniAligner, to assess mapping within highly complex genomic regions, including centromeric regions. We observed a mapping rate of 26.97% for APR samples, surpassing the 20.42% observed in the HPRC dataset. This highlights the better representation of complex genomic regions in our data, as detailed in Supplementary Fig. 11. These steps ensured robust variant calling, including low-quality regions. We have also included statistics on the coverage of subtelomeric and pericentric regions, which are enriched with rDNA (Fig. 1d).

• The authors report a high proportion of novel variants, including for SNVs and Indels where it is 15-19% per genome. This is odd because for this type of variation there is genomic completeness in the references used, and samples from Middle East have already been extensively sampled in the literature. The authors do not attempt to check in GNOMAD and other Middle Eastern cohorts such as Mineral et al. 2021 and Scott et al. 2016.

Response: Thank you for your comment. We have applied Inspector polishing algorithm to eliminate systematic errors, ensuring the reliability of our data. The reported number of small variants, including SNVs and indels, is consistent with findings from HPRC and CPC. For APR specific variants, we have conducted extensive filtering using dbSNP, gnomAD, the 1000 Genomes Project, and Greater Middle East (GME) - a comprehensive database that curates variants from numerous genome sequencing studies focused on Middle Eastern samples. The Arab representation is very limited in large genomic resources, including HPRC, 1000 Genomes Project and gnomAD (mostly exome). Our study will serve as a significant resource for understanding the genomic landscape of the MENA region and the broader global genomic community. We believe the reviewer is referencing Mineta et al, 2021 but this a genotyping array-based study that would not be useful to calculate novelty of variants.

• For the SV calls, there is no attempt to QC the calls which are more prone to mis-assembly errors

Response: Thank you for your comment. The number of structural variant (SV) calls per sample in our study is comparable to those reported in the HPRC dataset, indicating consistency in SV detection. We followed established methodologies aligned with HPRC QC standards, ensuring accurate and reliable SV calls. Furthermore, we built assemblies with >100X coverage (including 2011 GB of data with minimum >100kb stretch DNA with 12.53X coverage per sample) using long-read sequencing data, which are known for their high sensitivity in detecting structural variants. As the field of pangenome continues to evolve, these approaches contribute to standardizing SV detection processes, minimizing errors and maximizing data accuracy.

Novel euchromatic sequences

• Authors report their assemblies contain 112 Mbs of novel sequence relative to CHM13, other

pan genomes, and 1.4 kb of Mt DNA. This amounts to ~ 4% of the human genome which seems excessive and may not as striking. How much does that compare to the inverse situation i.e the amount of novel sequences in HPRC samples relative to the samples in this study?

Response: Our novel sequence estimation is based on SV insertions absent from GRCh38, T2T-CHM13, HPRC, CPC, and DGV. While HPRC and CPC reported novel sequences relative to the linear references GRCh38 and T2T-CHM13, our analysis incorporates all references, including both linear and graph pangenomes, to exclude previously reported sequences. Additionally, we filtered out insertions for variants present in the DGV database. The APR pangenome growth curve (Fig. 3h) and the per-sample contribution to novel sequences in APR (2.6 MB) is comparable to those observed in HPRC (1.9 MB) and CPC (2.8 MB).

Complex structural variation

• Strange the criteria used here for naming this category of SVs. Authors should just refer to SVs larger than certain length and frequent above a certain threshold.

Response: The term "Complex Structural Variation Site and Region" was used to describe regions in the graph genome containing multiple overlapping structural variations (SVs) from multiple haplotypes, including rearrangements, deletions, and insertions. This is illustrated in Figure 4, which highlights loci with multiallelic SVs in both HPRC and APR datasets. This aligns with terminology used in prior studies (e.g., "complex multiallelic SV loci"), as employed by HPRC, to denote such regions with high allelic diversity.

• Again novelty claimed without assessing relevant datasets beyond HPRC+CPC.

Response: The structural variants we reported underwent rigorous filtering, including comparisons against SVs in the reference genomes, HPRC, and CPC. Additionally, we excluded SVs present in the Database of Genomic Variants (DGV), which hosts one of the largest global datasets of structural variants.

Mitochondrial pangeome

• No mention which reference variants were called against

Response: Thanks for your comment. We used T2T-CHM13 mitochondrial genome sequence for mapping and variant calling. This information has now been explicitly stated in the revised manuscript.

• The Mt assemblies are based on PacBio only. No mention of what particular QC was performed for these genomes nor what mis-assemblies were found

Response: For Mt pangenome construction, we used PCR free high fidelity PacBio long reads with an average Q score of 33, a minimum length of 15kb, and at least 90% sequence identity. This stringent quality control allowed us to detect a set of high quality heteroplasmy variants from Arab populations.

As the Mt pangenome was directly built from high-fidelity heteroplasmy reads and not from assembly, concerns about potential misassemblies are not applicable in this context. Given that Minigraph-Cactus cannot handle construction of *de novo* cyclic genomes, we utilized PGGB algorithm to construct the Mt pangenome graph.

Performance gain from pangenome-aided

• Higher calling rate using Pan genomes has been shown before so that is not specific to the APR graph.

Response: Thank you for your comment. We agree that the higher calling rate using pangenome has been demonstrated previously, including the HPRC study. However, it is important to note that the HPRC dataset does not include Arab samples. In our analysis, we observed a higher genotype and structural variation calling rate in APR compared to HPRC (Fig. 5f and g). While this represents an overall improvement, it holds greater significance because some of these variants are located within APR-specific sequences that may not be adequately represented or mapped in HPRC. We illustrate this using trio data we generated: the APR pangenome variant calls exhibited lower errors in Mendelian concordance (with error rates of 0.0071 for SNPs and 0.0172 for indels), significantly lower than those observed by HPRC (with error rates of 0.0504 for SNPs and 0.0749 for indels).

• The mapping rate is between the APR and HPRC is almost identical. Unclear what is the added value here

Response: The mapping rate of APR is significantly higher compared to T2T-CHM13, confirming its advantage over the current linear references. While the mapping rate is comparable to HPRC, the sequence and haplotype structure in APR is specific to the Arab population, offering a distinct advantage in detecting structural variations and population-specific variants. See the Mendenlian concordance analysis we mentioned above. Our analysis using short read mapping demonstrates that APR achieves a superior structural variation detection rate compared to HPRC.

References:

Almarri, M. A., Haber, M., Lootah, R. A., Hallast, P., Al Turki, S., Martin, H. C., . . . Tyler-Smith, C. (2021). The genomic history of the Middle East. *Cell, 184*(18), 4612-4625 e4614. doi:10.1016/j.cell.2021.07.013

Arauna, L. R., Mendoza-Revilla, J., Mas-Sandoval, A., Izaabel, H., Bekada, A., Benhamamouch, S., . . . Comas, D. (2017). Recent Historical Migrations Have Shaped the Gene Pool of Arabs and Berbers in North Africa. *Mol Biol Evol, 34*(2), 318-329. doi:10.1093/molbev/msw218

Bergstrom, A., McCarthy, S. A., Hui, R., Almarri, M. A., Ayub, Q., Danecek, P., . . . Tyler-Smith, C. (2020). Insights into human genetic variation and population history from 929 diverse genomes. *Science, 367*(6484). doi:10.1126/science.aay5012

Chawla, V., Kumar, R., & Shankar, R. (2016). Identifying wrong assemblies in de novo short read primary sequence assembly contigs. *J Biosci, 41*(3), 455-474. doi:10.1007/s12038-016-9630-0

Cornet, L., & Baurain, D. (2022). Contamination detection in genomic data: more is not enough. *Genome Biol, 23*(1), 60. doi:10.1186/s13059-022-02619-9

Girirajan, S., Campbell, C. D., & Eichler, E. E. (2011). Human copy number variation and complex genetic disease. *Annu Rev Genet, 45*, 203-226. doi:10.1146/annurev-genet-102209-163544

Jakobsson, M., Scholz, S. W., Scheet, P., Gibbs, J. R., VanLiere, J. M., Fung, H. C., . . . Singleton, A. B. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature, 451*(7181), 998-1003. doi:10.1038/nature06742

Kolmogorov, M., Billingsley, K. J., Mastoras, M., Meredith, M., Monlong, J., Lorig-Roach, R., . . . Paten, B. (2023). Scalable Nanopore sequencing of human genomes provides a comprehensive view of haplotype-resolved variation and methylation. *bioRxiv*. doi:10.1101/2023.01.12.523790

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., . . . Myers, R. M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science, 319*(5866), 1100-1104. doi:10.1126/science.1153717

Li, Q., Yan, B., Lam, T. W., & Luo, R. (2022). Assembly-free discovery of human novel sequences using long reads. *DNA Res, 29*(6). doi:10.1093/dnares/dsac039

Liao, W. W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., . . . Paten, B. (2023). A draft human pangenome reference. *Nature, 617*(7960), 312-324. doi:10.1038/s41586-023-05896-x

Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D., & Gurevich, A. (2018). Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics, 34*(13), i142-i150. doi:10.1093/bioinformatics/bty266

Mineta, K., Goto, K., Gojobori, T., & Alkuraya, F. S. (2021). Population structure of indigenous inhabitants of Arabia. PLoS Genetics, 17(1), e1009210. https://doi.org/10.1371/journal.pgen.1009210

Muñoz-Barrera, A., Rubio-Rodríguez, L. A., Jáspez, D., Corrales, A., Marcelino-Rodriguez, I., Lorenzo-Salazar, J. M., . . . Flores, C. (2024). Benchmarking of bioinformatics tools for the hybrid de novo assembly of human whole-genome sequencing data. *bioRxiv*, 2024.2005.2028.595812. doi:10.1101/2024.05.28.595812

Sherman, R. M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., . . . Salzberg, S. L. (2019). Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat Genet, 51*(1), 30-35. doi:10.1038/s41588-018-0273-y

Zhang, W., Jia, B., & Wei, C. (2019). PaSS: a sequencing simulator for PacBio sequencing. BMC Bioinformatics, 20(1), 352. doi:10.1186/s12859-019-2901-7