An Emirati pangenome incorporating a diploid telomere-to-telomere reference

Michael Olbrich[1]*, Mira Mousa[1,2]*, Inken Wohlers[3,4], Amira Al Aamri[1], Halima Alnaqbi[1,5], Aisha Hanaya Alsuwaidi[1,5], Hima Vadakkeveettil Manoharan[1], Nour al-dain Marzouka[1], Sanjay Erathodi Ramachandran[1], Anju Annie Thomas[1], Mohammed Alameri[6], Guan K Tay[7,8,9], Rifat Hamoudi[10], Saleh Ibrahim[1,7], Noura Al Ghaithi[6], Habiba Alsafar[1,5]**

Affiliations:

1. Center for Biotechnology, Khalifa University for Science and Technology, Abu Dhabi, United Arab Emirates
2. Department of Public Health and Epidemiology, Khalifa University for Science and Technology, Abu Dhabi, United Arab Emirates
3. Biomolecular Data Science in Pneumology, Research Center Borstel - Leibniz Lung Center, Borstel, Germany
4. University of Lübeck, Lübeck, Germany
5. Department of Biomedical Engineering and Biotechnology, Khalifa University for Science and Technology, Abu Dhabi, United Arab Emirates
6. Abu Dhabi Department of Health, Abu Dhabi, United Arab Emirates
7. Department of Medical Sciences, Khalifa University for Science and Technology, Abu Dhabi, United Arab Emirates
8. Psychiatry, UWA Medical School, University of Western Australia, Perth, Western Australia, Australia
9. School of Medical and Health Sciences, Edith Cowan University, Joondalup, Western Australia.
10. College of Medicine, University of Sharjah, Sharjah, United Arab Emirates

Keywords: Human genetic variation, United Arab Emirates, Middle East, Genome reference, Diploid assembly, Telomere-to-telomere, Pangenome, Population genetics

*These authors contributed equally to this work

**Correspondence: habiba.alsafar@ku.ac.ae

Abstract

Reference data on genomic variation forms the basis of genetics research. Limitations in identifying genetic variation from single reference sequences have recently been addressed through improvements in sequencing technologies, allowing the generation of pangenomic references from multiple accurate chromosome-level *de novo* assemblies. Nevertheless, global pangenomes to date have yet to include genomes from the populations of the Middle Eastern Region. To address this shortcoming, this study provides an Emirati genome reference. Its core is a diploid assembly with a Quality Value (QV) of 60 that includes ten telomere-to-telomere chromosomes. This assembly is incorporated into a pangenome graph constructed of 52 additional high-quality assemblies, half of which are trio-based. This Emirati pangenome reveals a similar level of genomic variation as the one compiled by the Human Pangenome Reference Consortium, underscoring its utility for the identification of both global and population-centered genomic variation, even in genome regions that have been traditionally challenging to assemble but are covered by the Emirati telomere-to-telomere assembly. As such, the Emirati genome reference significantly contributes to genomic research globally and is an essential resource for genomics-based personalized medicine in the United Arab Emirates and other parts of the Middle East.

Introduction

Over the past two decades, the human reference genome has been the cornerstone of genetic research and biomedical advancements, offering a centralized framework for interpreting human genetic variation. Landmark initiatives, such as the Human Pangenome Reference Consortium (HPRC)[1] and Telomere-to-Telomere (T2T) consortium[2] have expanded this foundation. The HPRC introduced a draft pangenome comprising 47 phased, diploid assemblies from a cohort of genetically diverse individuals[1], while the T2T consortium completed the first contiguous sequence of a haploid human genome, the T2T-CHM13[2]. These milestones have provided open-access genomic resources critical for variant detection and interpretation, functional annotation, population genetics, and epigenetic analyses.

Despite these advances, the HPRC and T2T-CHM13 references capture only a narrow subset of global human genetic diversity. The absence of variation from underrepresented populations introduces reference bias, limiting the ability to accurately characterize the full spectrum of human genetic variation and its implications for health and disease. This gap has prompted a shift toward population-specific pangenomic references integrating genetic diversity. Regional efforts, such as the Chinese Pangenome Project, which includes genomic data from over 50 individuals of Chinese ancestry[3], the Arab pangenome of 53 individuals of diverse Arab Ancestry[4], and the African Pangenome project[5], have aimed to construct population-specific pangenomes unique to these regions. Such initiatives have improved variant detection and functional understanding and emphasized the critical need for diverse genomic resources to drive precision medicine and population health research. However, significant gaps remain, particularly for populations in the Middle East, whose unique genetic landscapes are shaped by indigenous ancestry, historic trade routes, and extensive global connectivity.

The United Arab Emirates (UAE) is leading one of the largest population genome projects, the Emirati Genome Program (EGP), with over 600,000 genomic samples collected to date, with the aim of covering the entire citizen population of one million. In response to the need for more population-specific genomic resources, this study presents the Emirati pangenome, which incorporates a diploid telomere-to-telomere reference. This resource offers unprecedented genomic completeness and specificity for the UAE population, addressing gaps left by previous global pangenomes. The telomere-to-telomere assembly described in this study complemented the pangenome, enhancing the representation of genetic diversity within the UAE and establishing a model for how population-specific genomic data can complement global resources.

Results

Cohort Selection for the Reference Assembly

A total of 53 genomes were included in the Emirati genome reference, comprising one female Emirati T2T assembly and 52 pangenome assemblies, which included 26 trio family members and 26 single samples. These individuals were carefully selected to capture the UAE's genetic diversity, adhering to the following criteria: (1) all pangenome individuals were chosen from distinct cohort stratifications based on the global ancestral components of the Emirati population, and the T2T prime sample was selected to represent the highest admixture proportion of the main ancestral component; (2) all individuals were free from congenital disorders, providing a healthy baseline for genomic analysis; and (3) trio samples were restricted to individuals with unrelated parents, defined by a kinship threshold up to the fourth degree, ensuring accurate genetic phasing. Figure 1 illustrates the results of the population genetic admixture analysis and the sex distribution of the cohort, emphasizing the extensive genetic diversity captured within this dataset.

Telomere-to-Telomere Assembly

This study generated a diploid, trio-based T2T assembly of the genome of a female Emirati individual, called T2T-ERG_XX, using a combination of Illumina (Supplementary Table 1), PacBio HiFi (Supplementary Table 2), and ONT ultra-long sequencing data (Supplementary Table 2). The sequencing effort included 93X coverage of Illumina reads for the child, 127X coverage from PacBio HiFi reads for the child (Supplementary Table 2), and 65X and 63X coverage of PacBio HiFi reads for the mother and father, respectively (Supplementary Table 3). The data were complemented by 3.7X coverage from ONT ultra-long reads generated from 21 ONT PromethION sequencing runs, with a median N50 of 67.8 kb (Supplementary Table 2). PacBio HiFi data were derived from nine sequencing runs, producing a median N50 of 16.1 kb and a median coverage of 31.4X.

As outlined in Supplementary Figure 1, the assembly process involved multiple stages of read trimming, polishing, scaffolding, and assembly refinement using various tools. The quality of the assemblies was evaluated after the assembly and polishing phases, respectively, using a balance of contingency (number of contigs and auN/N50-values), base accuracy (Quality Value (QV) score), and completeness (percentage of N-values) (Supplementary Table 4). The final assembly was selected based on its optimal trade-off between these parameters and was derived from an initial assembly with Verkko[6], scaffolding with ntLink[7], followed by CHM13 reference-guided scaffolding using RagTag[8], and gap-filling via two alternative hifiasm[9] assemblies using quarTeT[10] (Supplementary Figure 1).

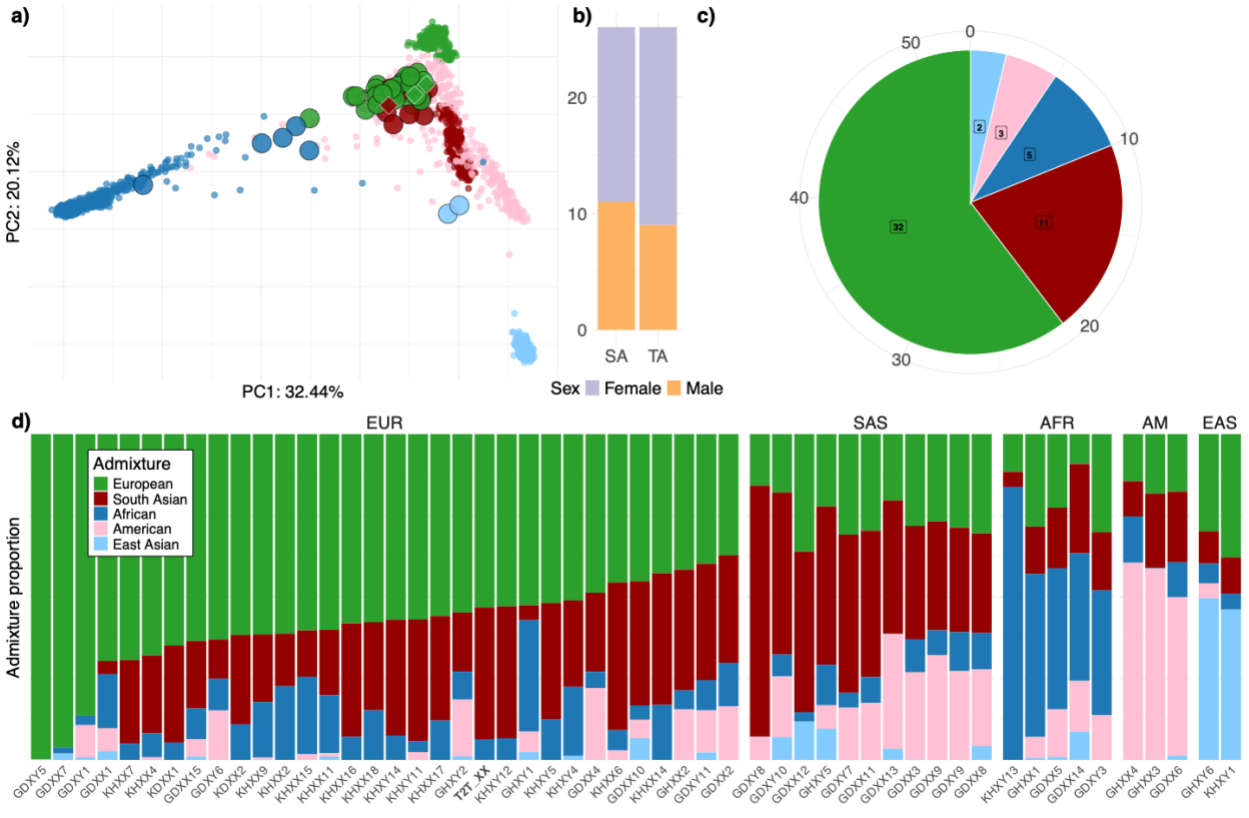


Figure 1 Basic and ancestry-based characteristics of the 53 genomes included in the Emirati genome reference. (a) The genomes are projected onto the genotype principal components computed by the 1000 Genomes phase 3 samples. Each Emirati genome is colored according to its 1000 Genomes-based global ancestry component with maximum value. The T2T-ERG_XX trio is denoted by diamond-shaped symbols, with the T2T-ERG_XX genome located between the two parent genomes. (b) The number of pangenome single (SA) and trio (TA) assemblies stratified by sex. The female trio-based T2T assembly is not included here. (c) The distribution of the genomes' main ancestry components across the cohort. (d) For each Emirati genome, the contributing ancestry components after supervised admixture analysis based on 1000 Genomes global populations, i.e., European, South Asian, African, American, and East Asian, c.f. Supplementary Table 17.

The final assembly comprises chromosome-level haplotypes with an overall length of 3.03 Gb. The maternal haplotype has an adjusted QV score of 60.8, and the paternal haplotype has an adjusted QV score of 59.8, with 2% and 3.2% switch errors, respectively (Supplementary Table 5). The assembly contains 101 gaps, of which 37 are found in the maternal haplotype and 64 in the paternal haplotype (Supplementary Table 6). The corresponding N-bases per 100 kb are 2.21 for the maternal and 7.10 for the paternal haplotype (Supplementary Table 5). Notably, chromosomes 4, 6, 12, 14, 15, 18, 21, and 22 in the maternal haplotype and chromosomes 6 and 8 in the paternal haplotype are gap-free (Supplementary Table 6), indicating that these chromosomes were fully assembled from telomere to telomere. This study generated the mitochondrial sequence from Illumina short-read-based mitochondrial variant calls using the reference rCRS and confirmed the lack of mitochondrial structural variation using the Integrative Genomics Viewer (IGV). Our assembly covered 94.4% of the sequence of CHM13, with 32.5 and 29.2 Mb of the maternal

and paternal haplotypes, respectively, not aligning to CHM13, corresponding to approximately 1% of the genome. The total CHM13 aligned length was 2.99 Gb.

Regarding gene coverage, the assembly included 96.6% of the genes reported in CHM13 (62,060 of 64,000 genes), with 258 genes detected only partially. Of the 1,550 multi-copy genes in CHM13, 1,219 were also present with multiple copies in this assembly (78.6%), amounting to an overall 4,746 genes (Supplementary Table 5). Interestingly, these numbers were identical for both the maternal and paternal haplotypes. After manually fixing a misjoin between chromosomes 14 and 15 (also reported and corrected for CHM13), this study identified five and four intrachromosomal misjoins in the maternal and paternal haplotypes, respectively, compared to CHM13.

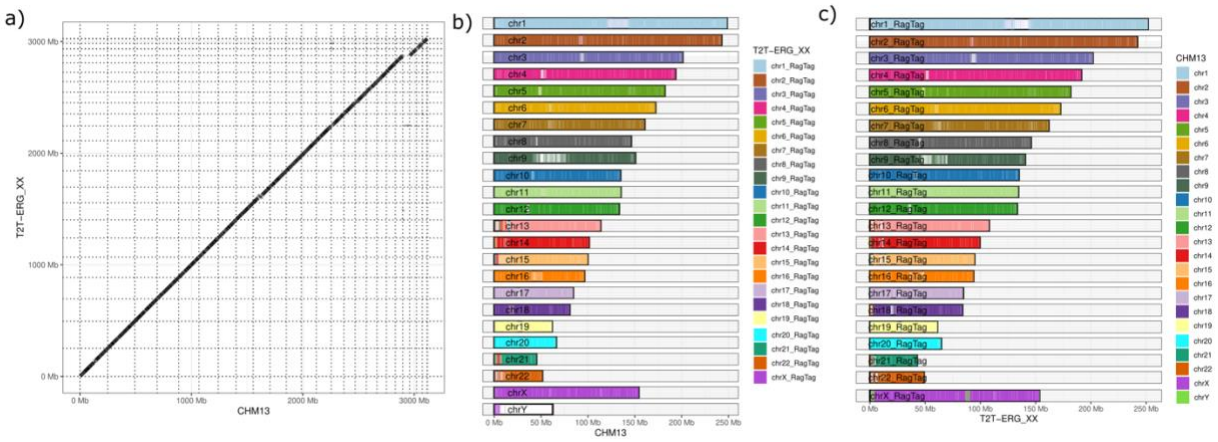


Figure 2 Alignment of T2T-ERG_XX maternal haplotype with CHM13v2.0. All primary alignments are shown. (a) Dotplot. (b) The coverage of T2T-ERG_XX chromosomes with CHM13 chromosomes. (c) The coverage of CHM13 chromosomes with T2T-ERG_XX chromosomes. In (b) and (c), all assembly sequences with a length of more than 5 Mb are shown.

The alignment of this assembly with CHM13 showed that all chromosomes were assembled accurately (Figure 2a maternal and Supplementary Figure 2a paternal haplotype), but it revealed that centromeric regions were partially not aligned. Several larger centromeric regions in chr13, chr14, chr15, chr18, chr21, and chr22 aligned between non-homologous chromosomes, highlighting differences that relate to assembly challenges that were reported for CHM13 previously (Figure 2b-c maternal and Supplementary Figure 2b-c paternal haplotype). Reads from chromosome X were also aligned to the pseudoautosomal regions of the Y chromosome in CHM13 despite the assembly not containing a Y chromosome.

Fifty-Two High-Quality Diploid Assemblies

The T2T assembly of the genome of an Emirati individual was complemented with the assemblies of 52 representative genomes. The underlying data was generated using one PacBio Revio flow cell per sample, with nine samples receiving an additional sequencing top-up. The median coverage per sample was 30.3X,

with a range of 25.5X to 40.1X and a median read N50 of 16.7 kb (Supplementary Table 7 and Supplementary Figure 3). Adapter trimming was performed for HiFi data, but no adapters were detected. Contamination screening was conducted with Kraken2[11]. The assembly was performed using hifiasm[9], considering parental HiFi data for the 26 trio assemblies to generate fully phased haplotypes. Scaffolding was then performed using ntLink, followed by correction and polishing with Inspektor[12]. Supplementary Figure 4 summarizes the pangenome processing strategy.

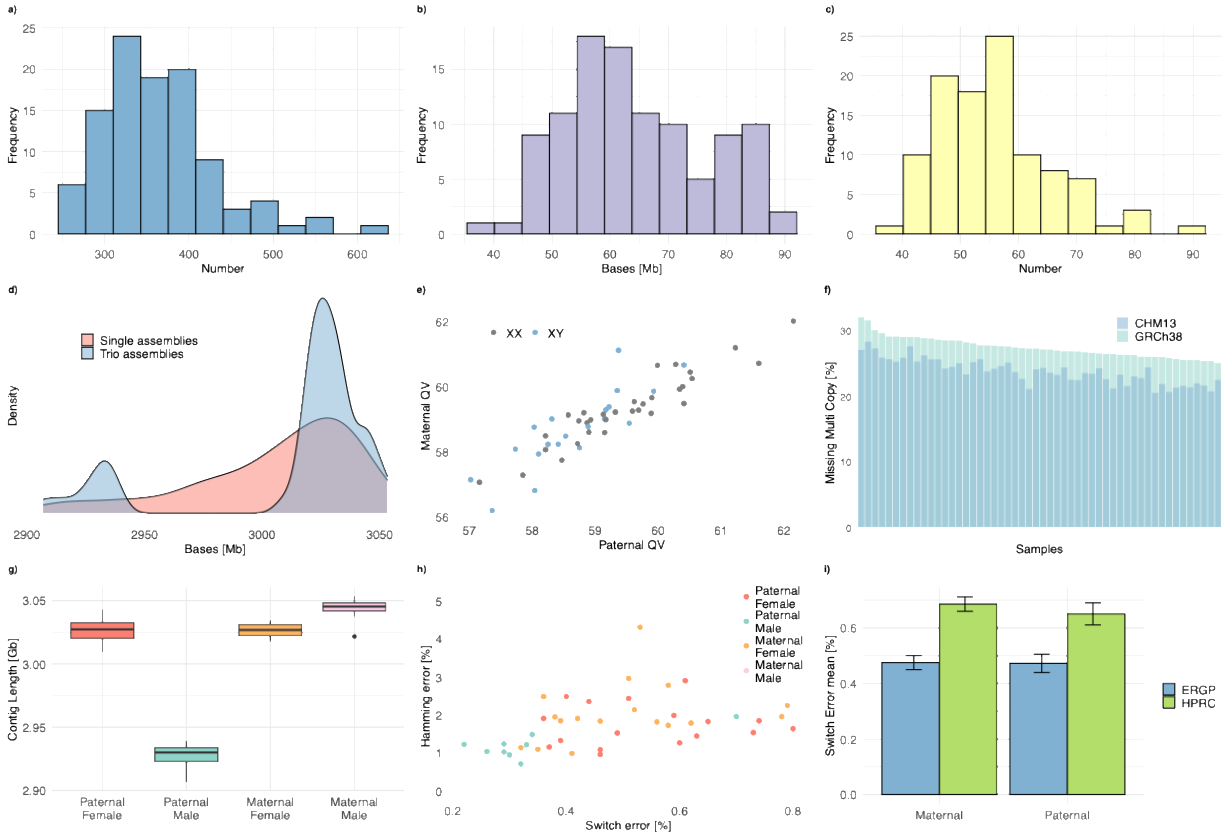


Figure 3: Assembly characteristics of 52 HiFi assemblies after post-processing. (a) Contig number (b) N50 (c) L90. (d) The overall sequence length for 104 haplotypes stratified by single sample and trio-assembly approach. (e) Paternal versus maternal haplotype QV. (f) Missing multi-copy genes in percent. (g) Overall contig lengths stratified by sex and haplotype. (h) For the 26 trio assemblies, Hamming error versus Switch error in percent. Note that n=9 male paternal haplotypes are not shown, since they have a substantial Hamming error, c.f. main text. (i) Comparison of mean switch error for all haplotypes of the 26 trio assemblies with the mean switch error of assemblies reported by the human pangenome reference consortium (HPRC).

The quality and characteristics of the assembly were evaluated at the draft hifiasm haplotypes stage (Supplementary Table 8) and the final haplotype stage (Supplementary Table 9 and Supplementary Figure 5), with quality generally improving through post-processing. For the final assemblies' haplotypes, the median number of contigs was 358.5 (see Figure 3a), the N50 was 36.28 Mb (see Figure 3b), and the L90 was 55 (see Figure 3c). The median overall length across all haplotypes was 3.02 Gb, and for the 64 haplotypes from female genomes (which contain no Y chromosome and diploid X chromosomes), the

median length was 3.03 Gb. The paternal haplotypes of nine male individuals from trio assemblies had a median length of 2.93 Gb (see Figure 3d), confirming the accurate phasing of diploid genomes containing a Y chromosome.

The adjusted QV score of maternal and paternal haplotypes was highly correlated, with values ranging from 56.18 to 62.15 and a median of 59.13 (Figure 3). Notably, 22 haplotypes (21.15%) exhibited no intrachromosomal misjoins, while the remaining haplotypes had a median of 1 misjoin, with a maximum of 6 misjoins. On average, 89.49% of these misjoins were located in centromeric regions. A median of 61,875 of the 64,214 genes reported in CHM13 were detected across the haplotypes, representing 96.34% of the total, with a median of 362 genes detected only partially (Supplementary Table 9). For the 1,550 multi-copy genes reported in CHM13, a median of 1,181 (76%) was detected in this assembly.

When comparing the assembly in this study to the GRCh38 reference, the percentage of missing multi-copy genes was higher and less variable across haplotypes compared to the CHM13 reference, as shown in Figure 3f. Differences in haplotype lengths for the male genome, with the paternal haplotype being shorter due to the presence of the Y chromosome instead of an X chromosome, were observed in this study. The maternal haplotypes of males tended to be longer when compared to those of females, although both covered the same chromosomes (Figure 3g).

Switch and Hamming errors were assessed across the 52 haplotypes of the 26 trio assemblies. The paternal haplotypes from males exhibited low Hamming error rates (Figure 3h), while the nine maternal haplotypes from males exhibited exceptionally high Hamming errors (ranging from 16.11 to 26.50% - not shown in Figure 3h). The remaining 43 haplotypes had Hamming errors of less than 4.3%, with a median of 1.78%. Switch errors across all haplotypes were generally low, with a median of 0.44%, and no haplotype had a switch error rate higher than 0.80% (Figure 3h), closely matching the switch error rate of the HPRC pangenome assemblies (Figure 3i).

Emirati Pangenomic Variation

This Emirati assemblies were used in this study to generate three distinct pangenome graphs using the Minigraph-Cactus[13] pipeline under different settings, each incorporating CHM13 as the backbone and GRCh38 as an additional, unclipped reference (Supplementary Figure 6). The three settings investigated were the: (i) inclusion of only the two haplotypes of the T2T-ERG_XX assembly from this study, (ii) use of 104 haplotypes from the 52 HiFi-based assemblies, and (iii) inclusion of the maternal T2T-ERG_XX haplotype (which demonstrated better quality than the paternal one) as an additional, unclipped reference

alongside the 104 haplotypes. The number of nodes, edges, and corresponding sequence lengths for each graph are summarized in Supplementary Table 10.

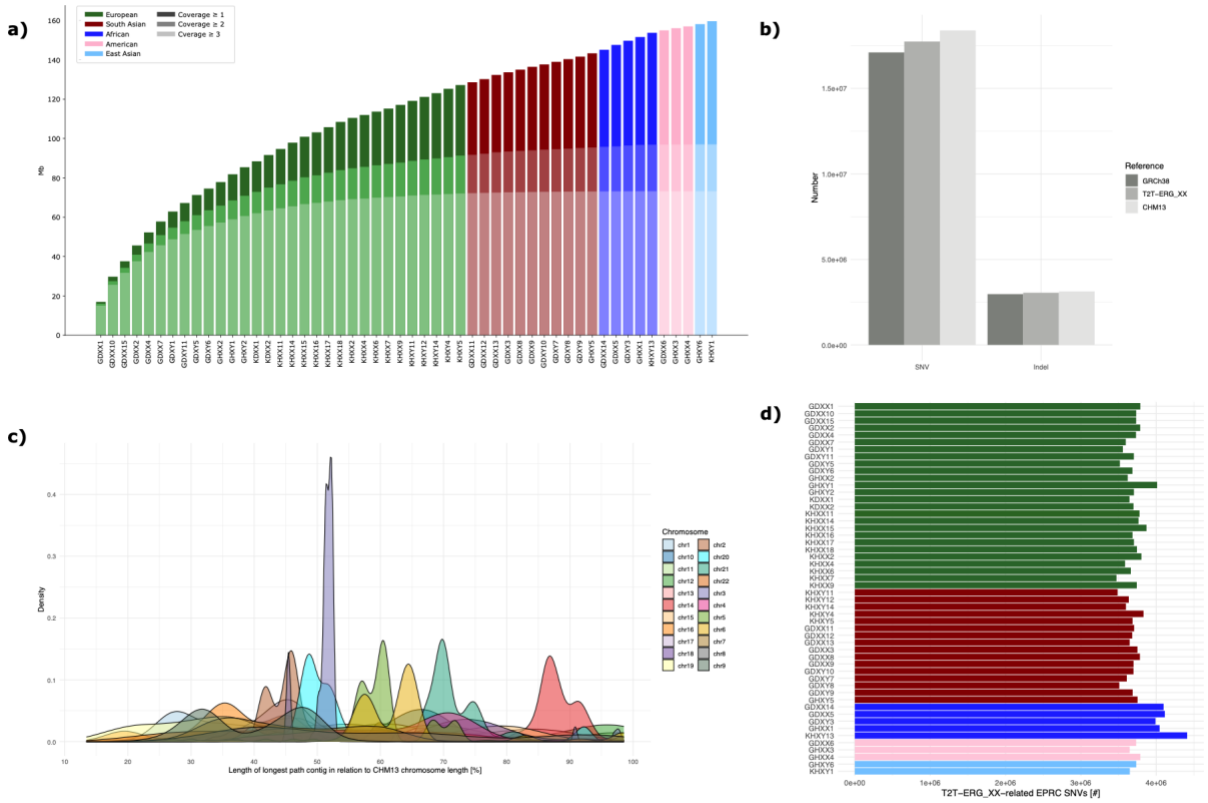


Figure 4: Characteristics of the pangenome graph. (a) The pangenome growth curve shows the cumulative novel sequence added per genome. (b) The overall number of single-nucleotide variants (SNVs) and insertions and deletions (summarized as indels) covered by the pangenome with respect to the unclipped GRCh38, T2T-ERG_XX, and CHM13 sequences as respective references. (c) For each chromosome, the distribution of the longest CHM13 chromosomal path that is covered by the same assembly contig. (d) For each sample, the number of pangenome-reported SNVs with respect to CHM13. Coloring and sorting are according to the main ancestry component and then lexicographically.

The inclusion of T2T-ERG_XX haplotypes contributed 16.60 Mb of novel sequence. In the graph of the 52 further Emirati assemblies, there were 81.59 million nodes spanning 3.31 Gb of sequence, connected by 112.27 million edges. The total novel sequence not included in CHM13 and GRCh38, across the 52 Emirati genomes, amounted to 179.57 Mb (Supplementary Table 11). Including the maternal T2T-ERG_XX haplotype as an additional reference slightly reduced the number of nodes to 81.15 million (98.23%) and the sequence length to 3.30 Gb, connected by 110.28 million edges (98.22%). The graph containing only the two T2T-ERG_XX haplotypes represented 95% of the length of the graph with 52 pangenome assemblies, with and without the maternal ERG haplotype, but its node and edge counts were approximately 26% of those in the graphs covering more than one genome (Supplementary Table 10). The total added sequence from the 52 Emirati genomes that is not captured by CHM13, GRCh38, or T2T-ERG_XX was

159.63 Mb (Figure 4, Supplementary Table 11), with the per-genome contribution detailed in Figure 4a. Furthermore, 380,655 novel bases, shared by at least 50 Emirati genomes, were identified in the graph based on CHM13 and GRCh38. This decreased to 44,707 bases when the Emirati T2T reference was incorporated (Supplementary Table 11).

Variant counts were calculated relative to GRCh38 and CHM13. The number of variant records in relation to GRCh38 was 20.10 million, compared to 21.80 million relatives to CHM13 (Supplementary Table 12). For the graph with the maternal T2T-ERG_XX haplotype as a reference, the variant count slightly increased to 20.18 million for GRCh38 and decreased to 21.70 million with respect to CHM13. The variant record count related to the maternal T2T-ERG_XX haplotype was 20.84 million, falling between the variant counts for GRCh38 and CHM13 (Supplementary Table 12). Between 98.6 and 99.8% of variant records are SNVs or insertions and deletions (summarized as indels), with their overall number in relation to the three references in the pangenome graph shown in Figure 4b.

The pangenome graph paths relating to the CHM13 chromosomes were investigated. As expected, the longest scaffold in each CHM13 chromosomal path corresponds to the matching T2T-ERG_XX chromosome. While the lengths of the CHM13 and T2T chromosomes were similar, the CHM13 path covered between 75.82% (maternal chromosome 9) and 98.86% (maternal chrX), with a median coverage of 94.99% across both haplotypes (Supplementary Table 13). When considering the 52 pangenome assemblies, the length of the CHM13 chromosome path associated with the longest haplotype scaffold was typically shorter, as many of these assemblies did not reach the chromosomal level (Supplementary Table 14). Overall, 194 pangenome assembly scaffolds achieved >90% coverage of the CHM13 chromosome path, including 48 mitochondrial contigs. Of these, 146 scaffolds can be considered chromosome-level, including chromosomes chr3 (5x), chr4 (5x), chr5 (2x), chr6 (8x), chr7 (1x), chr8 (16x), chr10 (10x), chr11 (19x), chr12 (44x), chr14 (25x), chr18 (3x), and chr20 (8x). The path lengths of the longest scaffolds in relation to CHM13 chromosome lengths are shown in Figure 4c. The observed variations in chromosome-level scaffolds and peak scaffold sizes for certain chromosomes were correlated with the lengths of the respective p and/or q arms.

The number of SNVs per genome identified through the pangenome assembly (Figure 4d) was compared with those obtained via traditional read mapping-based variant calling using Sentieon, with the same HiFi data that was used for assembly construction. SNV counts ranged from 4.08 million to 5.68 million, with a median of 4.59 million (Supplementary Table 15). Notably, genomes with the largest African ancestry component, along with one genome with a main European ancestry component, exceeded 5 million SNVs. In contrast, SNV counts from the pangenome graph were significantly lower, with a median of 3.66 million

SNVs when aligned to the CHM13 backbone. SNV counts for the graph relative to GRCh38 totaled 3.66 million, representing a 3.89% reduction compared to CHM13. The median number of SNVs relative to the maternal T2T-ERG_XX haplotype was 3.70 million, closely matching the CHM13-based count (Supplementary Table 15). While two African ancestry genomes showed slight decreases (0.24% and 0.78%), the number of SNVs in relation to the maternal T2T-ERG_XX haplotype was, on average, 1.27% higher than for CHM13, with a maximum difference of 2.74%, corresponding to a median difference of 43,924 SNVs and a maximal difference of 98,294 SNVs in a genome with a European ancestry component.

Emirati Reference-Based Variant Detection

The performance of short-read mapping and corresponding variant calling using the Telomere-to-Telomere Emirati Reference Genome (ERG; maternal T2T-ERG_XX haplotype) and the Emirati Pangenome Graph (EPR; 42 genomes) was assessed with sequencing data from an independent cohort of 10 male and 10 female Emirati individuals (Supplementary Figure 7). These samples were from a previous study, which established the first Emirati reference genome.[14] Illumina sequencing coverages ranged from 35X to 77X. Mapping rates varied between 93.78% and 98.89% for reads and between 90.71% and 98.28% for bases. The differences in mapping rates between the linear GRCh38, CHM13, and ERG references were minimal. Specifically, GRCh38 mapped 0.22% fewer bases on average compared to CHM13, while ERG mapped 0.15% fewer bases than CHM13. In most samples (85%), the ERG mapping rate was higher than that of GRCh38, with a median difference of 0.07% (Supplementary Table 16).

Pangenome graph references generally exhibited lower mapping rates than individual linear reference genomes. Using the HPRC pangenome with GRCh38 as the internal reference, the median mapping rate was 93.60%, with a range of 90.71% to 95.56%. The EPR pangenome, using CHM13 as the internal reference, achieved a slightly higher median mapping rate of 94.19%, ranging from 92.98% to 98.27% (Supplementary Table 16).

We also compared the number of single nucleotide variants (SNVs) and insertions/deletions (indels) called by Sentieon [REF] (Supplementary Table 16). When comparing GRCh38 to the HPRC pangenome using GRCh38 as the internal reference, the median number of SNVs increased by 2.73%, from 4,046,201 to 4,156,524, and the median number of indels increased by 1.93%, from 971,792 to 990,502. Similarly, with the EPR pangenome using CHM13 as the internal reference, the median number of SNVs compared to linear CHM13 increased by 3.90%, from 3,866,878 to 4,017,770, while the median number of indels increased by 2.46%, from 921,915 to 944,572.

Discussion

The Emirati telomere-to-telomere (T2T) high-quality diploid reference assembly, complemented by a pangenome graph, provides a comprehensive representation of the genomic diversity within the UAE population, a cornerstone of this study. The inclusion of 52 genomes in the pangenome graph adds approximately 159.63 Mb of novel sequence, offering valuable insights into the region's genetic landscape. Compared to other pangenomes, this study provides a focused and detailed portrayal of the Emirati population's unique genetic structure.

This study presents a diploid, trio-based T2T assembly of the genome of a female Emirati individual, which provides a marked contribution to the understanding of the Emirati genome and its variability. The assembly achieved a high level of completeness, with 94.4% of the CHM13 reference sequence covered, and revealed a substantial portion of novel sequences not captured by previous reference genomes like CHM13 or GRCh38. This assembly included 96.6% of the genes from CHM13, with an overlap in multi-copy genes, demonstrating the accuracy and richness of the sequence data. The maternal and paternal haplotypes were well-phased, with minor discrepancies in centromeric regions and intrachromosomal misjoins affecting previously reported regions, which are common challenges in genome assembly due to their repetitive nature. The maternal haplotype showed higher quality, and chromosomes 4, 6, 12, 14, 15, 18, 21, and 22 were gap-free. Notably, this assembly also revealed some structural variations, particularly in the centromeric regions, contributing valuable insights into human genomic diversity in regions that are currently only covered by very few and recent telomere-to-telomere assemblies.

The cohort of 52 pangenome individuals was selected based on the computationally assigned primary genomic global ancestry. This approach offers a practical framework for selecting samples that are covering the Emirati genomic spectrum while considering abundances within this spectrum. It is, however, essential to recognize that the Emirati genomes in this study are typically composed of several components at high frequencies, constituting a genetic continuum roughly spanned by the global components. Further, the global components used here (European, South Asian, African, American, East Asian) were chosen to maximize the human diversity covered by the Emirati cohort. They do not necessarily reflect significant shared ancestry with these global populations and are likely proxies for more specific ancestry components described earlier to originate within the Arabian Peninsula and geographically neighboring regions. However, in the context of the population's high heterogeneity, this approach facilitated the selection of a cohort that robustly encapsulates the spectrum of genetic diversity, ensuring the reference genome reflects the complexity of the population's genomic landscape.

The presented assembly quality metrics demonstrate the competitiveness of the genome assembly approach taken in this study within the context of pangenome development efforts by the Human Pangenome Reference Consortium (HPRC), the Chinese Pangenome Consortium (CPC), and the Arab Reference Pangenome Project. The cohort of 52 samples in this study provides a representative framework for capturing the genetic diversity of the Emirati population, which, while small, is highly heterogeneous due to its unique history of migration and admixture. The Emirati pangenome, comprising 53 samples, lies between the HPRC's 47 samples for a global draft and the CPC's 58 samples, representing a population of 1.4 billion. Since the approximately one million Emirati population is significantly smaller, the sampling ensures a robust representation of its genetic structure. In terms of assembly contiguity, the average N50 of 64.25 Mb from this study surpasses both the HPRC (40 Mb) and CPC (>35.63 MB), highlighting the strength of the assembly pipeline used. The Arab pangenome reference reports a higher N50 of 124.25 Mb, likely reflecting the inclusion of ONT ultra-long sequencing data into their pangenome assemblies, which also in this study was essential to reach chromosome-level scaffolds. Notably, the N50 that was achieved for pangenome assemblies in this study utilized solely HiFi reads at an average sequencing coverage of 30.9X. Similarly, the QV score of 59.13 exceeds those of the HPRC (53.57), CPC (48–60), and the Arab reference (57.53), indicating a high degree of base-calling accuracy and assembly fidelity. Furthermore, the switch error rate of 0.47% from this study is the lowest among the pangenomes evaluated, outperforming the HPRC (0.67%), CPC, and Arab reference (0.64%) and demonstrating superior phasing and haplotype accuracy.

The variant record count for the maternal T2T-EGR_XX haplotype in this Emirati pangenome was 20.84 million, falling between the variant counts observed for GRCh38 and CHM13, the previous human genome references. Conventional, linear reference-based SNV counts for pangenome reference samples ranged from 4.08 million to 5.68 million, with a median of 4.59 million, showing significant diversity, particularly in genomes with the largest African ancestry components. In contrast, the pangenome graph itself showed a reduction in SNV counts, with a median of 3.66 million SNVs in relation to the CHM13 backbone, representing a 3.89% reduction compared to the GRCh38 reference. These results align closely with previous findings in pangenome studies, where different populations' genomic variation has been mapped to reference genomes with varying degrees of resolution.

The pangenome described here fills an important gap by focusing on the genetic variation of a population of the Middle Eastern and North African (MENA) region. This adds to the global representation of the spectrum of human genetic diversity captured in other pangenomes, such as the HPRC, which is crucial for understanding both population health and the unique genetic markers that may influence disease susceptibility and drug response. Moreover, the pangenome graph created through this study allows for

exploring common and rare variants and population-specific traits that contribute to personalized medicine. Integrating these diverse genomic data provides a more comprehensive resource for future research in genomic medicine, as it enables more precise mapping of genetic variants that could have clinical relevance. This pangenomic approach will be crucial in identifying novel disease-associated variants, understanding complex traits, and informing public health strategies tailored to Emirati and broader Middle Eastern populations. Overall, this study represents a significant step forward in building a more complete, high-quality reference genome for the Emirati population and paves the way for more inclusive and precise genomic research in the region.

Methods

Sample Selection

We identified parent-child trios and single samples from the Emirati Genome Program (EGP) and local recruitment sites to capture the genetic diversity of the UAE population. The EGP dataset was first screened for congenital disorders to exclude affected individuals, thereby ensuring that the cohort represented genetically unaffected individuals. Familial trios were then identified, with a focus on selecting unrelated parents, as this distance in kinship (up to the fourth degree) represents the algorithmic cutoff for reliable genomic analysis (Supplementary Table 17).

To assess familial relationships, KING[15] software was used to infer relatedness up to the fourth degree, employing algorithms for kinship estimation and identity-by-descent (IBD) segment analysis. Following this, a graph-based approach in R was applied to visualize and further refine familial relationships. The analysis began with the construction of a subgraph focusing on parent-offspring relationships, enabling the identification of familial clusters. Clusters containing fewer than three individuals were excluded, and adjustments were made to account for varying cluster sizes. Additionally, these clusters were examined for consanguinity, with parent relationships inferred up to the third degree. Sex assignments, derived from DRAGEN[16] ploidy information, were used to verify the correct parent-child assignments.

Due to the complexity of classifying familial relationships, certain groups were excluded in specific cases, including (1) parents with first- or second-degree relationships; (2) full-sibling relationships, particularly when a trio involved two brothers and the offspring of one brother; and (3) cases involving full- or half-sibling relationships, where trios resembled consanguineous relationships, such as a child born between two half-siblings. Additionally, multigenerational configurations, such as a trio consisting of a grandfather, father, and grandson, were excluded due to ambiguities in kinship, such as a "half-sibling edge" between

the grandfather and grandson, which complicated the accurate classification of parent-offspring connections. The resulting 24 trios derived from the EGP cohort were complemented with two recruited in-house.

Admixture Analysis

Given the genetic diversity of the UAE population, we performed a population genetic admixture analysis using Admixture[17] to stratify the population into distinct genetic subgroups. This stratification formed the basis for selecting familial trios that accurately represented the population's genetic diversity. To further enhance the representation, a complementary sampling approach was employed to include 26 single-sample assemblies alongside the 26 familial trios. These single samples underwent the same admixture stratification to ensure their genetic background was consistent with the broader population structure.

We used a set of global population informative variants from 12 studies curated by Wohlers *et al*.[18]. This dataset encompassed 127,261 variants observed in 5,429 individuals across 144 diverse populations selected for admixture analyses. The cohort in this study was stratified according to continental origins based on five global populations of 3,502 individuals as part of the 1000 Genomes Project[19]. Each of the five components can be considered to have distinct global genetic ancestry, namely, European, South Asian, African, South American, and East Asian ancestry. Employing a supervised mode of admixture, we quantified the proportions of genetic ancestry that these five distinct admixture components contribute to the Emirati population's genomes. This analysis allowed us to determine ancestral ratios, revealing the underlying global genetic substructure of the Emirati population. Additionally, principal component analysis (PCA) was performed using Plink[20] on population-specific variants from the reference samples. The PCA enabled the projection of the Emirati population within the context of these principal components, facilitating ordination within the broader genetic framework of the global reference populations.

Sequencing

Illumina Sequencing

The genomic DNA was extracted from 200 µL of EDTA-whole blood using QIAamp DNA Blood Mini Kit (Qiagen, Germany) according to the manufacturer's instructions. The quality of the extracted genomic DNA (gDNA) was examined using a Nanodrop 1000 spectrophotometer (Thermo Fisher Scientific, USA) to determine A260/230 and A260/280 ratios. The integrity of the gDNA was evaluated through gel electrophoresis methods, while samples were quantified using the Qubit dsDNA BR assay kit on Qubit 4.0 Fluorometer (Thermo Fisher Scientific, USA).

Subsequently, the gDNA samples were normalized to 200 ng in 30 µL and proceeded with library preparation using Illumina DNA PCR-Free Library Prep kit and IDT® for Illumina® DNA/RNA UD Indexes Set A for library barcoding. Libraries were quantified using KAPA library quantification assay (KAPA Biosystems, USA) on ViiA 7 qPCR System (Applied Biosystems, USA), were normalized to 1.1 nM and then pooled at equal volumes, followed by denaturation and further dilution before machine loading and sequencing. The S4 flow cells generated 150 bp paired-end reads in a standard workflow. Short-read sequencing of the final libraries was carried out using S4 300 cycles Reagent Kit v1.5 on a NovaSeq 6000 Sequencer (Illumina, USA).

A 30X read coverage was estimated for the average human genome size (~3.2 Gb) to perform sequencing in NovaSeq 6000. All the samples were sequenced to achieve 96 Gb of raw data (FASTQ), corresponding to 30X coverage of the whole genome. We allowed some tolerance and accepted WGS with a yield of >87 Gb. Samples falling short of the 96 Gb and 30X autosomal sequencing coverage threshold underwent an additional step of top-up that included repeating the library preparation and barcoding. The number of samples selected for pooling in the top-up process was determined by the additional data required and the throughput capacity of the NovaSeq 6000 S4 Reagent Kit v1.5 (300 cycles). Following sequencing, all sample BCL files were generated and demultiplexed utilizing bcl2fastq on the DRAGEN platform (Illumina).

High-Fidelity (HiFi) Sequencing

PacBio Nanobind CBB kit (Pacific Biosciences, 102-301-900) was used to extract high molecular weight (HMW) DNA from 200 µl peripheral whole blood samples. This extraction yields DNA fragments ranging from 50 to 500+ Kb, making it ideal for HiFi long-read WGS sequencing on the PacBio system. The concentration of the extracted gDNA samples was examined using Qubit dsDNA BR kit fluorometric assay on Qubit 4.0 Fluorometer (Thermo Fisher Scientific, USA), and the fragment size was assessed using gDNA 165 Kb kit on a Femto Pulse System (Agilent Technologies, USA). All DNA samples achieved the recommended Genome Quality Number (GQN) of 9.0 or above at 10 kb and 5.0 or above at 30 kb. The samples were then processed with a PacBio Short Read Eliminator kit (SRE), following the manufacturer's guidelines, to remove DNA fragments under 10 kb in length selectively. The concentration and GQN of the gDNA were determined using Qubit 1x dsDNA HS kit fluorometric assay and Femto Pulse - gDNA 165Kb method, respectively. Subsequently, sample shearing was executed using the Megaruptor 3 system (Diagenode s.a, Belgium) at a speed of 29-31 to achieve target insert lengths of 15-18 Kb, followed by sample clean-up using 1.0 X SMRTbell cleanup beads. The concentration and size distribution of the fragmented DNA was assessed using the Qubit 1x dsDNA HS kit fluorometric method and Femto Pulse -

gDNA 165Kb assay, respectively. Libraries of 12-20 kb size were then constructed using the PacBio SMRTbell prep kit 3.0 protocol (Pacific Biosciences, USA). Library preparation was performed using the manufacturer's protocol, including DNA repair, A-tailing, adaptor ligation, and cleanup, followed by nuclease treatment to eliminate unligated DNA fragments and residual SMRTbell adapters. AMPure PB beads with a 3.1X ratio of beads (35%) were used to remove the shorter fragments (<5 kb) and concentrate the DNA sample. The concentration of the SMRTbell libraries was assessed using the Qubit 1x dsDNA HS fluorometric assay, followed by the Femto Pulse - gDNA 165Kb method to determine the fragment size distribution.

SMRTLink (v12.0) Sample Setup software was used to obtain instructions for WGS HiFi Reads Sequencing preparation (ABC), including primer annealing and polymerase binding. The final libraries were sequenced on a single 8 M SMRT Cell on the PacBio Revio platform.

Oxford Nanopore (ONT) Ultra-long Sequencing

To extract ultra-high molecular weight (UHMW) DNA for Oxford nanopore ultra-long sequencing, a buffy coat from a fresh 5-ml blood sample tube was aliquoted for peripheral blood mononuclear cells (PBMCs) isolation using RBC Lysis Solution (QIAGEN, 158904). The purified PBMCs were counted, and approximately 60 million cells per sample were carried forward for UHMW DNA extraction using the NEB Monarch Cells & Blood DNA extraction kit (New England Biolabs, T3050), with modifications as recommended by the manufacturer for ONT Ultra long sequencing application. The gDNA was eluted in the ONT Extraction EB (EEB) buffer.

The extracted gDNA was quantified using a method developed by Paul A. 'Giron' Koetsier & Eric J Cantor, 2021. Subsequently, 30 to 40 µg of DNA in 750 µl of EEB buffer was taken to the library preparation step using the ONT Ultra-Long DNA Sequencing Kit V14 (SQK-ULK114) in accordance with the supplied protocol. The library preparation began with an initial tagmentation step for 10 minutes at room temperature, then for 10 minutes at 75°C and finally kept in ice for 10 minutes. The rapid adapters were added and incubated for 30 minutes at room temperature; then, the adapter-ligated DNA was cleaned using a precipitation method. The final libraries were quantified, and 90 ul of the library was mixed with the sequencing buffer and the loading solution, which were kept at RT for 30 min. To proceed with the loading, the flow cells were primed with the priming mix as per protocol. After adding the libraries, they were incubated in the flow cell for 10 minutes. The sequencing setup was done using the MinKNOW software. The pores were scanned every 2 hours for optimal loading, and the live base calling was conducted using the high accuracy (HAC) base-calling model with base detection for 5mC on the MinKNOW software. The

flow cells were reloaded two more times after following the flow cell wash protocol as described in the protocol.

Pangenome Graph Construction

The construction of the pangenome graph complemented with T2T-ERG_XX assembly followed a systematic, multi-step approach to ensure high data quality and accuracy. Supplementary Figure 6 depicts this process, which involved assembling 26 single samples using the standard hifiasm[9] mode, the phased assembly of 26 family trios in 'trio-dual' mode, and the generation of a primary T2T assembly using a dedicated strategy. Details of the assembly strategies are provided in the respective sections. The graph was constructed using the current version of the Minigraph-Cactus[13] pipeline. Three distinct pangenome graphs were generated to evaluate the Emirati pangenome. These graphs incorporated CHM13 as the primary backbone, GRCh38 as an additional unclipped reference, and haplotypes from the 52 Emirati assemblies and the T2T-ERG_XX assembly. The graph configurations included (i) both T2T-ERG_XX haplotypes (ii) 104 haplotypes from the 52 Emirati assemblies (called EPR) and (iii) the maternal T2T-ERG_XX haplotype as an unclipped reference alongside the 104 haplotypes (called EPRC).

Telomere-to-Telomere Assembly

The T2T-ERG_XX assembly strategy is outlined in Supplementary Figure 1. Briefly, the process began with the removal of adapters from the HiFi data using HiFiAdapterFilt[21], with no adapters detected, and from ONT data using Porechop_ABI[22]. ONT reads longer than 20 kb were kept and then combined with PacBio HiFi data to construct a diploid, trio-based draft assembly using Verkko[6]. Parental k-mers, generated from the HiFi data with Meryl[23], were employed in this process. The initial draft assembly was then scaffolded using ntLink[7], followed by reference-guided scaffolding using RagTag[8] with the CHM13 version 2 reference genome. To address gaps in the assembly, gap-filling was performed using two alternative hifiasm-based assemblies. These assemblies were generated from yak[24]-preprocessed parental HiFi data. The first hifiasm assembly was constructed using the same ONT and HiFi data as the Verkko assembly. For the second assembly, the ONT reads were polished with Ratatosk[25], using Illumina data that had been down-sampled to 50% with BBMap[26] before performing the assembly. Finally, contigs from both hifiasm draft assemblies were utilized to polish the scaffolded Verkko assembly using QuarTeT[10], resulting in the final assembly. Key assembly quality measures for selecting the optimal assembly strategy at different stages, as well as for the selected alternative assembly strategies, are provided in Supplementary Table 4. Further assembly characteristics of the final assembly can be found in Supplementary Table 5.

Alignment of T2T-ERG_XX with CHM13 was performed using minimap2[27] alignment, followed by filtering to keep primary alignments and visualization using the R package pafR. Dot plots and coverage plots were generated for all primary alignments. Coverage plots display only contigs with a length of more than 5 Mb.

Pangenome Assembly

The core pangenome assemblies constructed using PacBio HiFi sequencing data comprise 26 phased trio assemblies (TAs) and 26 single assemblies (SAs). Single-sample assemblies were generated using hifiasm[9] in standard mode, while trio-based assemblies were phased using hifiasm trio-dual mode. Sequencing data were preprocessed to remove adapters using HiFiAdapterFilt[21], which confirmed the absence of adapter sequences across all datasets. Parental k-mer databases were created using yak[24] with recommended settings for trio assemblies to ensure precise phasing and haplotype assignment.

Initial quality control assessments of the generated assemblies were conducted using Quast[28] and yak[24]. Post-assembly scaffolding was performed with NTLink[7] to optimize assembly continuity. Kraken2[11] screening was carried out against human and non-human reference databases to identify and remove potential contamination. Chromosomal misjoins were detected using 'paftools misjoin' leveraging the CHM13v2 reference genome. This process identified three intrachromosomal misjoins in the maternal haplotypes of samples GDXX5, GHXX3, and GDXX9, specifically: a misjoin between chromosomes 1 and 3 in GDXX5 (ntLink_4), chromosomes 10 and 12 in GHXX3 (ntLink_15), and chromosomes 1 and 13 in GDXX9 (ntLink_10). Each misjoin was resolved by splitting the affected contigs at the midpoint (with padding of 2 to 1,500 bp) and annotating the resulting fragments with suffixes'_part1' and '_part2' to maintain traceability.

Gene duplication analysis was conducted using 'paftools asmgene' against GRCh38 and CHM13v2 references, revealing no significant outliers. Further error correction at the base pair and structural variant levels was performed using Inspector, utilizing the original HiFi sequencing data as input.

A systematic renaming scheme was introduced to maintain consistency with pangenome standards. Each sample was encoded with identifiers reflecting its source (Genome Project or extended sampling), assembly type (haploid or diploid), inferred sex (XX or XY), and a sequential numerical ID. Contig names were updated to include the new sample ID and haplotype information, enabling seamless integration into the pangenome graph.

Finally, secondary quality control checks using Quast[28] and yak[24] confirmed the accuracy of the polished assemblies.

Pangenome Graph Assessment

The pangenome graphs' structural properties and genomic content were evaluated using Panacus[29], a tool specifically designed to assess pangenome graph structure. Key graph metrics were computed for each configuration, including the number of nodes, edges, and sequence lengths. To characterize the graphs further, the contribution of novel sequences from the T2T-ERG_XX assembly and the 52 Emirati genomes was quantified. Additionally, the overlap of graph paths with CHM13v2.0 and GRCh38 chromosomes was measured, allowing for an evaluation of coverage and single nucleotide and structural variation representation.

Variant calling was conducted using both linear reference-based and graph-based approaches. For linear references, Sentieon[30]'s variant-calling pipeline processed short-read Illumina sequencing data aligned to CHM13 and GRCh38, generating detailed variant statistics, including counts of single nucleotide variants (SNVs) and small insertions and deletions (indels). This workflow ensured a consistent framework for comparing variants detected through graph-based mapping.

The VG Giraffe[31] aligner maps short-read sequencing data to the constructed pangenome graphs. This alignment enabled Sentieon to call variants directly from the pangenome, producing variant records relative to the CHM13 backbone. Additionally, the variant call outputs from the Minigraph-Cactus[13] pipeline were integrated into the analysis to evaluate graph-derived variants comprehensively. Finally, the sequence coverage of CHM13 chromosome paths by graph paths was measured to assess the contiguity of scaffolds within the pangenome.

Read-mapping Based Evaluations

The novel linear and pangenome references were applied to a distinct cohort of 20 Emirati individuals with Illumina whole genome sequencing data described previously.[14] For the comparison of read and base mapping rates across graphs and linear references, ‘samtools[32] stats’ was used and for compiling variant summaries from VCF files ‘bcftools[32] stats’. Linear reference mapping and mapping-based variant calling were performed using the Sentieon[30] software. Variant numbers detected from graph alignments were cross-compared with those from linear reference alignments to investigate variation numbers that are representative of Emiratis and identify differences. Read-based variant calling evaluation ensured that the constructed linear reference and pangenome graph effectively captured the Emirati population's genomic variation.

Code Availability

The list of all tools, including versions, used for this study is available in Supplementary Table 18.

References

1. Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).

2. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).

3. Gao, Y. *et al.* A pangenome reference of 36 Chinese populations. *Nature* **619**, 112–121 (2023).

4. Nassir, N. *et al.* A draft Arab pangenome reference. 2024.07.09.602638 Preprint at https://doi.org/10.1101/2024.07.09.602638 (2024).

5. Tetikol, H. S. *et al.* Pan-African genome demonstrates how population-specific genome graphs improve high-throughput sequencing data analysis. *Nat. Commun*. **13**, 4384 (2022).

6. Rautiainen, M. *et al.* Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat. Biotechnol*. **41**, 1474–1482 (2023).

7. Coombe, L., Warren, R. L., Wong, J., Nikolic, V. & Birol, I. ntLink: A Toolkit for De Novo Genome Assembly Scaffolding and Mapping Using Long Reads. *Curr. Protoc*. **3**, e733 (2023).

8. Alonge, M. *et al.* Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol*. **23**, 258 (2022).

9. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).

10. Lin, Y. *et al.* quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Hortic. Res*. **10**, uhad127 (2023).

11. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol*. **20**, 257 (2019).

12. Chen, Y., Zhang, Y., Wang, A. Y., Gao, M. & Chong, Z. Accurate long-read de novo assembly evaluation with Inspector. Genome Biol. **22**, 312 (2021).

13. Hickey, G. *et al.* Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat. Biotechnol*. **42**, 663–673 (2024).

14. Daw Elbait, G., Henschel, A., Tay, G. K. & Al Safar, H. S. A Population-Specific Major Allele Reference Genome from The United Arab Emirates Population. *Front. Genet*. **12**, 660428 (2021).

15. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinforma. Oxf. Engl*. **26**, 2867–2873 (2010).

16. Behera, S. *et al.* Comprehensive genome analysis and variant detection at scale using DRAGEN. *Nat. Biotechnol*. (2024) doi:10.1038/s41587-024-02382-1.

17. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).

18. Wohlers, I. *et al.* An integrated personal and population-based Egyptian genome reference. *Nat. Commun.* **11**, 4719 (2020).

19. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

20. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

21. Sim, S. B., Corpuz, R. L., Simmonds, T. J. & Geib, S. M. HiFiAdapterFilt, a memory efficient read processing pipeline, prevents occurrence of adapter sequence in PacBio HiFi reads and their negative impacts on genome assembly. *BMC Genomics* **23**, 157 (2022).

22. Bonenfant, Q., Noé, L. & Touzet, H. Porechop_ABI: discovering unknown adapters in ONT sequencing reads for downstream trimming. Preprint at https://doi.org/10.1101/2022.07.07.499093 (2022).

23. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol*. **21**, 245 (2020).

24. Li, H. yak: Yet Another K-mer toolkit.

25. Holley, G. *et al.* Ratatosk: hybrid error correction of long reads enables accurate variant calling and assembly. *Genome Biol*. **22**, 28 (2021).

26. Bushnell, B. BBMap: A Fast, Accurate, Splice-Aware Aligner. (2014).

27. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

28. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).

29. Parmigiani, L., Garrison, E., Stoye, J., Marschall, T. & Doerr, D. Panacus: fast and exact pangenome growth and core size estimation. *bioRxiv* 2024.06.11.598418 (2024) doi:10.1101/2024.06.11.598418.

30. Kendig, K. I. *et al.* Sentieon DNASeq Variant Calling Workflow Demonstrates Strong Computational Performance and Accuracy. *Front. Genet*. **10**, 736 (2019).

31. Sirén, J. *et al.* Genotyping common, large structural variations in 5,202 genomes using pangenomes, the Giraffe mapper, and the vg toolkit. *bioRxiv* 2020.12.04.412486 (2021) doi:10.1101/2020.12.04.412486.

32. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).

Acknowledgments

We would like to extend our sincere gratitude to all the individuals who generously donated their samples to the EGP and to the local recruitment sites. Their willingness to participate in this study is invaluable and forms the foundation of this research, contributing to our understanding of the genetic landscape of the UAE population. Special thanks to Mubarak Almehairbi and Mauricio Paton Gasso for their contributions to quality control of the pangenome datasets, Salsabeel A.S. Juneidi for recruitment and DNA extraction, Sarah Chehadeh for ONT protocol follow-up, and Khacey Juma for recruitment. Their expertise and dedication have been instrumental in ensuring the success of this project.

Author Contributions

MO, IW, and HVM developed the assembly strategy, with MO and HVM carrying out T2T assembly construction. MO, AA, and AHA constructed the pangenome assemblies, while AA led the post-processing and QC of the assemblies. MO, IW, AA, and HVM examined the characteristics of the assemblies. MO and AHA were responsible for creating the pangenome graph, while AHA led the post-processing and visualization of the pangenome, with IW and AHA analyzing its characteristics. MO, MM, HAN, and AA

handled data coordination and management. NaM, MA, and MPG conducted the graph variant calling evaluation. The original draft was written by MO and MM, with IW contributing, while MA, GKT, RH, SI, and NaG undertook the review and editing of the manuscript. Admixture analysis was conducted by MO and NaM. Sample selection was carried out by MO and MM. HAN, SAAS, AAT, and SER performed sequencing. HS served as the principal investigator and oversaw the laboratory organization that HAN managed. The overall conceptualization, funding acquisition, and supervision of the project were led by HS. All authors reviewed and approved the final draft of the manuscript.

Ethics Declarations

The study procedures and protocols adhered to the ethical guidelines and regulations established by the Research Ethics Scientific Committee in the UAE (DOH: DOH/CVDC/2022/1701), Khalifa University (KU: H21-045), and the principles outlined in the Declaration of Helsinki.

Competing Interests

The authors declare that they have no competing interests.

Supplementary Information

Supplementary Tables

Supplementary Table 1: Summary of Illumina short-read sequencing data generated using the NovaSeq 6000 platform. Sequencing was performed with a read length of 150 bp, yielding coverage exceeding 90× for all individuals. Coverage was calculated based on the CHM13 genome length.

Supplementary Table 2: Summary of long-read sequencing data generated for ERG_FATHER, ERG_MOTHER, and ERG_XX using PacBio HiFi and Oxford Nanopore Technologies (ONT) platforms. Coverage is computed for the genome length of CHM13. ERG_FATHER has two HiFi and one ONT run, ERG_MOTHER has three HiFi and one ONT, and ERG_XX has four HiFi and 19 ONT runs. HiFi sequencing data demonstrated high coverage across all individuals, with coverage ranging from 29.52x to 34.54x. ONT sequencing data included multiple files per sample, with varying sequencing depths and coverage ranging from 0.05x to 31.83x.

Supplementary Table 3: Summary of processed and merged sequencing data for ERG_XX, ERG_FATHER, and ERG_MOTHER across Illumina, Oxford Nanopore Technologies (ONT), and PacBio HiFi platforms. The coverage is computed for the genome length of CHM13. Illumina data has been down

sampled by 50% for each sequencing file in Supplementary Table 1 and achieved a coverage of 46.73×. ONT and HiFi data are combined from long-read sequencing data listed in Supplementary Table 2. ONT data for ERG_XX amounts to coverage of 111.81x with long-read lengths, as shown by a maximum read length of 824,775 bp and an N50 of 16,928 bp. HiFi sequencing provided high-accuracy data with coverage values ranging from 63.01x to 126.64x and N50 values between 15,167 bp and 17,934 bp.

Supplementary Table 4: Summary of assembly strategies and performance metrics for maternal and paternal haplotypes calculated before manually fixing one assembly error in the final maternal and paternal assembly. The strategy incorporated combinations of assembly tools, including Verkko, ntLink (HiFi and ONT), RagTag, hifiasm, and quarTeT, with and without Ratatosk correction. The number of contigs ranged from 25 to 195, with total base counts exceeding 3.02 Gb for all assemblies. Maternal haplotypes achieved N50 values up to 153.4 Mb and QV scores over 60, while paternal haplotypes exhibited N50 values up to 162.2 Mb. Hamming error rates ranged from 7.2962 to 18.1856, highlighting variability across strategies.

Supplementary Table 5: Analysis of maternal and paternal ERG haplotypes derived from a single individual's genome using combined Verkko and hifiasm assemblies. Metrics from Quast, yak, and Inspector (see Methods) include contiguity, assembly size, alignment quality, and structural correctness. The maternal haplotype contains 80 contigs (largest: 251.6 Mb, N50: 153.4 Mb), while the paternal haplotype has 100 contigs (largest: 247.6 Mb, N50: 162.2 Mb). Both haplotypes achieved high genome coverage (94.38% maternal, 94.40% paternal) and minimal duplication rates (1.017–1.018). Gene coverage, based on CHM13v2.0 annotation, and alignment metrics (NA50, NGA50) indicate high completeness and alignment quality.

Supplementary Table 6: Analysis of maternal and paternal ERG haplotypes, detailing the start and end positions of gaps across chromosomes and their respective gap sizes. The table provides information on gaps identified in both haplotypes, spanning autosomal chromosomes, sex chromosomes, and mitochondrial DNA. Gap lengths range from 1 base (e.g., T2TXXmatChrM) to 100,000 bases (e.g., chr19_RagTag for the paternal haplotype). Maternal gaps are predominantly small, with most ranging between 100 bp and 7,500 bp, and are distributed across chromosomes such as chr1_RagTag, chr2_RagTag, and chrX_RagTag. In contrast, paternal haplotypes show a similar distribution but include a larger gap, with the 100,000 bases gap on chr19_RagTag.

Supplementary Table 7: Sequencing data quality metrics the HiFi data underlying pangenome assemblies across trio (TA) and single (SA) assemblies for male (XY) and female (XX) genomes. Coverage values range from 25.53x to 40.06x, with multiple samples marked for sequencing top-up to reach the 25x coverage cutoff.

Supplementary Table 8: The assembly characteristics of 104 haplotypes from 52 genomes were categorized by sample ID, sex, and type (TA: trio assembly or SA: single assembly). Metrics were generated using Quast, yak, and Inspector, as described in the methods. The table includes contiguity metrics such as the number of contigs per haplotype, which range from 317 to 836, and N50 values spanning 25.6 Mb to 175.0 Mb. Total genome lengths (≥0 bp) are consistent across assemblies, ranging from 2.91 Gb to 3.05 Gb, with GC content uniformly reported between 40.76% and 40.86%. The largest contigs vary between approximately 99.3 Mb and 199.0 Mb. Misassemblies and mismatches per 100 kbp remain low, with misassembly counts between 0 and 9,669, local misassembly counts between 0 and 502, and mismatches ranging from 127.6 to 168.7 per 100 kb. Genome alignment metrics are high, with genome fractions exceeding 90.56% and the largest alignments spanning 175.5 Mb. Duplication ratios are stable around 1.017–1.027, and structural error metrics such as switch error rates and Hamming error rates remain below 0.003%. Gene coverage was assessed using all 64,214 genes (including 20,067 protein-coding) annotated for CHM13v2 downloaded from https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/annotation/chm13.draft_v2.0.gene_annotation.gff3. Note that two individuals share a family ID but are unrelated parents of a child not represented in this dataset.

Supplementary Table 9: The assembly characteristics of 104 haplotypes from 52 genomes were categorized by sample ID, sex, and type (TA: trio assembly or SA: single assembly). Metrics were generated using Quast, yak, and Inspector after scaffolding, polishing, and error correction. The table includes contiguity metrics such as the number of contigs per haplotype, which range from 260 to 619, and N50 values spanning 40 Mb to 90 Mb. Total genome lengths (≥0 bp) are consistent across assemblies, ranging from 2.9 Gb to 3.1 Gb, with GC content uniformly reported between 40.75% and 40.86%. The largest contigs vary between approximately 120 Mb and 200 Mb. Misassemblies and mismatches per 100 kb remain low, with misassembly counts between 0 and 12, local misassembly counts between 0 and 17, and mismatches ranging from 130 to 160 per 100 kbp. Genome alignment metrics are high, with genome fractions exceeding 94% and the largest alignments spanning up to 150 Mb. Duplication ratios are stable around 1.02, and structural error metrics such as switch error rates and Hamming error rates remain below 0.01%. Gene coverage was assessed using all 64,214 genes (including 20,067 protein-coding) annotated for CHM13v2 downloaded from https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/annotation/chm13.draft_v2.0.gene_annotation.gff3. Note that two individuals share a family ID but are unrelated parents of a child not represented in this dataset.

Supplementary Table 10: Comparison of graph assembly metrics across chromosomes for the ERG_XX sample, a cohort of 52 samples, and the combination of ERG_XX with 52 samples. The table details the number of edges, nodes, and graph lengths for each dataset. Edges range from 22 (chrM, ERG_XX sample)

to 9.4 million (chr2, 52 samples), with the total graph containing 28.6 million edges for ERG_XX alone, 112.3 million for 52 samples, and 110.3 million for the combined dataset. Nodes range from 21 (chrM, ERG_XX sample) to 6.8 million (chr2, 52 samples), with total node counts of 21.4 million (ERG_XX), 81.6 million (52 samples), and 80.1 million (combined dataset). Graph lengths span from 16,571 bases (chrM, ERG_XX sample) to 263.4 Mb (chr1, combined dataset), with total graph lengths of 3.15 Gb (ERG_XX), 3.31 Gb (52 samples), and 3.30 Gb for the combined dataset.

Supplementary Table 11: Pangenome graph growth curve data underlying Figure 3 illustrates the ordered growth of the graph in base pairs (bp) for multiple samples. Metrics are shown for single-sample growth, incremental growth with additional samples (e.g., 1, 2, 3, up to 50 samples), and specific growth comparisons with the ERG_XX sample. The data includes growth curves for panacus-based ordered-growth increments, with values for ERG_XX and combined datasets. Growth metrics reveal initial rapid expansions with fewer samples and a plateau as additional samples contribute less to the graph's total base pairs.

Supplementary Table 12: The number of variants in the pangenome graph was calculated for 52 Emirati genomes, incorporating the maternal Emirati T2T haplotype (T2T-ERG_XX) as a novel reference alongside the CHM13 and GRCh38 unclipped references. The table reports the total number of variant records, categorized into SNPs, MNPs, indels, other variant types, multiallelic sites, and multiallelic SNP sites. Comparisons between graphs with and without the maternal T2T-ERG_XX haplotype highlight the impact of integrating the novel reference. Variants supported by T2T-ERG_XX are particularly notable for their contribution to multiallelic sites and overall variant counts. CHM13 consistently shows the highest number of total variants across categories, reflecting its extensive variant representation.

Supplementary Table 13: Comparison of CHM13 and T2T-ERG_XX chromosome assembly metrics for maternal and paternal T2T haplotypes. The table includes the mapping between CHM13 chromosomes and corresponding T2T-ERG_XX chromosomes (RagTag paths), along with path length, percentage of CHM13 coverage, T2T-ERG_XX and CHM13 chromosome lengths, and the number of gaps in each T2T-ERG_XX chromosome. Path lengths vary from approximately 35 Mb (chr21) to 239 Mb (chr2), with longer path lengths aligning to larger chromosomes such as chr1 and chr2. CHM13 coverage ranges from 77.78% (chr21, ERG_XX#2) to 98.86% (chrX, ERG_XX#2), with most chromosomes exhibiting high coverage (>90%), indicating strong alignment between CHM13 and T2T-ERG_XX assemblies. T2T chromosome lengths range from 38 Mb (chr21) to 251 Mb (chr1), closely matching CHM13 chromosome lengths, which span from 45 Mb (chr21) to 248 Mb (chr1). The number of gaps within T2T chromosomes ranges from 0

to 13, with most chromosomes containing fewer than 5 gaps, demonstrating high assembly continuity overall.

Supplementary Table 14: Contig with the longest path in the pangenome graph in relation to CHM13 chromosomes jointly across all 104 EPR haplotypes. The percentage conveys the length of the contig assigned to the respective CHM13 chromosome to the overall size of the chromosome.

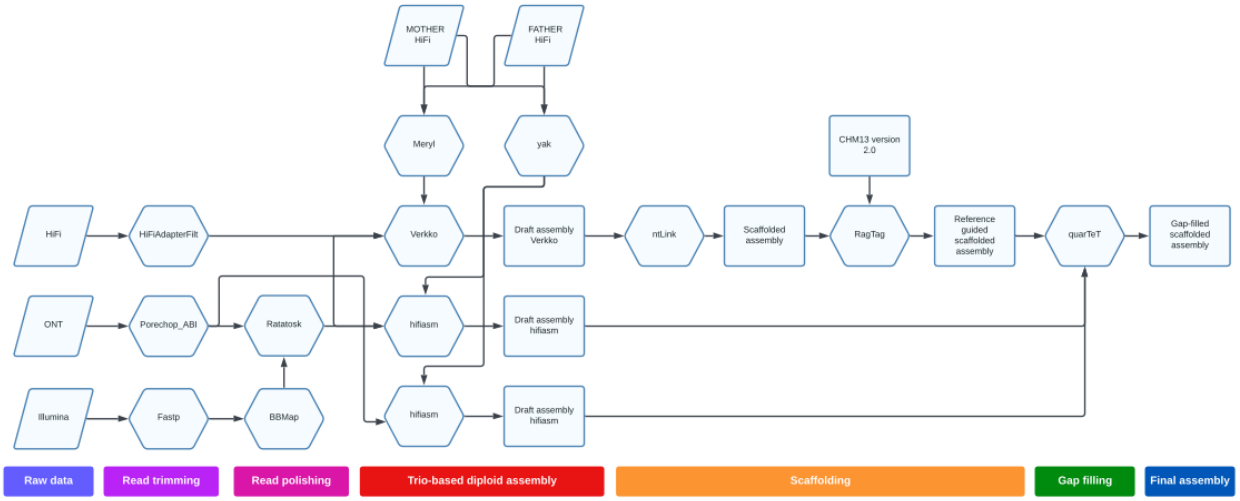
Supplementary Table 15: SNP counts for 52 pangenome assemblies as reported by the Minigraph-Cactus pipeline for the CHM13-based EPR graph with respect to CHM13, GRCh38, and T2T-ERG_XX are provided. For comparison, the number of SNP calls obtained by HiFi sequencing data-based SNP calling with Sentieon with respect to CHM13 and GRCh38 is provided. Ordering is according to the growth curve in Figure 3, i.e., the main ancestry component, and then lexicographically.

Supplementary Table 16: Assessment of different linear and graph-based genomic references using Illumina short-read whole-genome sequencing data from n=20 Emirati individuals generated using the NovaSeq 6000 platform. The table evaluates mapping, coverage, duplication rates, and variant calling metrics across multiple references: GRCh38, CHM13, T2T-ERG_XX, and graph-based assemblies (EPR graph with CHM13 backbone and HPRC graph with GRCh38 backbone). Coverage was calculated using the CHM13 genome size. Mapping rates were high across all references, with mean mapped read percentages ranging from 94.55% (HPRC graph GRCh38) to 96.64% (CHM13), and duplication rates varied by reference. Mapped base percentages ranged from 93.36% (HPRC graph GRCh38) to 95.95% (CHM13), indicating consistent alignment performance. The variant analysis identified mean counts of 4.1 million SNVs and 964,796 indels across references.

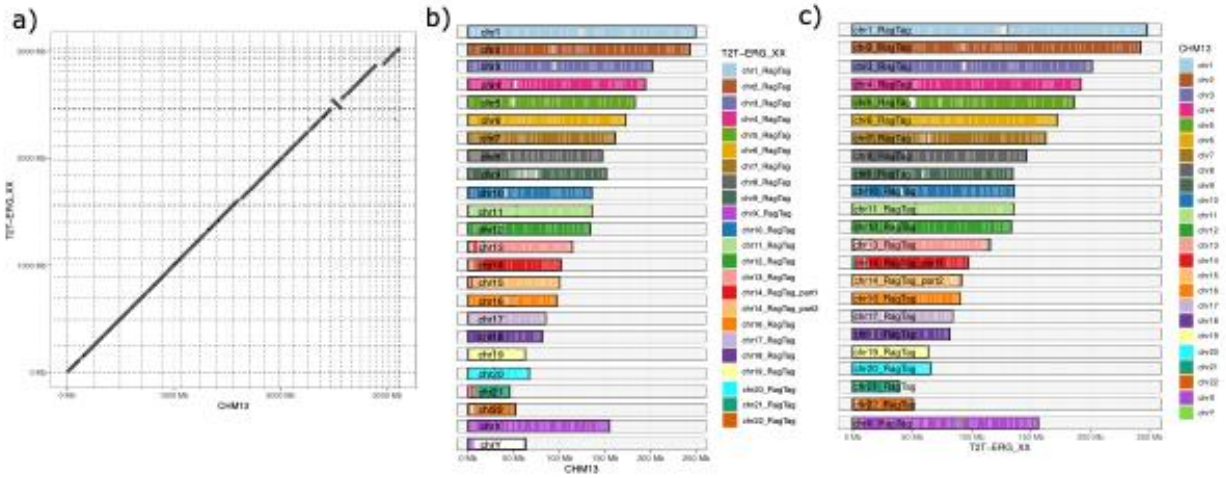
Supplementary Table 17: Overview of sample metadata, including sex, assembly type (single assembly (SA) or trio assembly (TA)), ancestry proportions (European (EUR), South Asian (SAS), American (AMR), African (AFR), and East Asian (EAS)), and the main ancestry component for each sample, defined as the ancestry with the highest proportional value. The table summarizes data for 52 samples from two sources (EGP cohort and KU), with half of the assemblies classified as trio assemblies. European ancestry is the most common main global ancestry component across the samples. The ancestry proportions vary widely, with European contributions reaching as high as 99.7% in GDXY5 and African contributions as high as 83.8% in KHXY13.

Supplementary Table 18: Tools and versions used in the respective analysis steps.

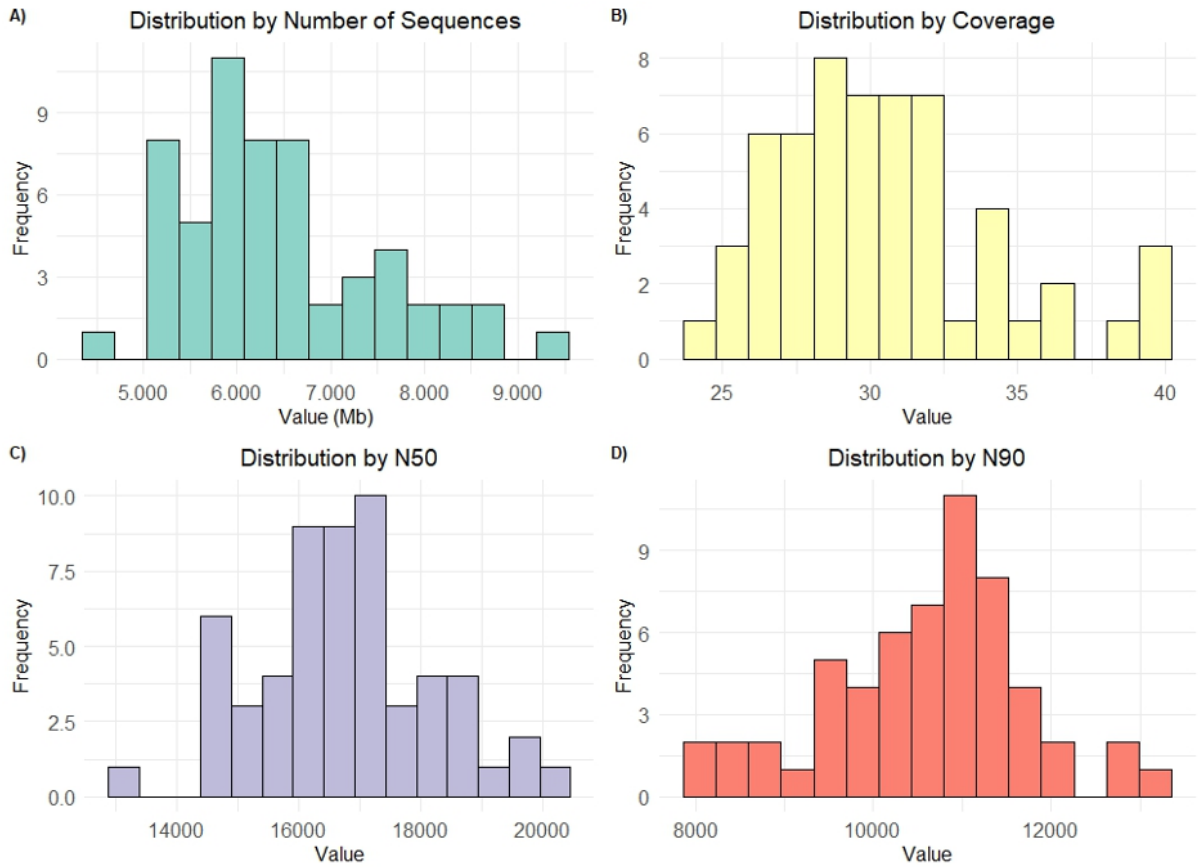
Supplementary Figures



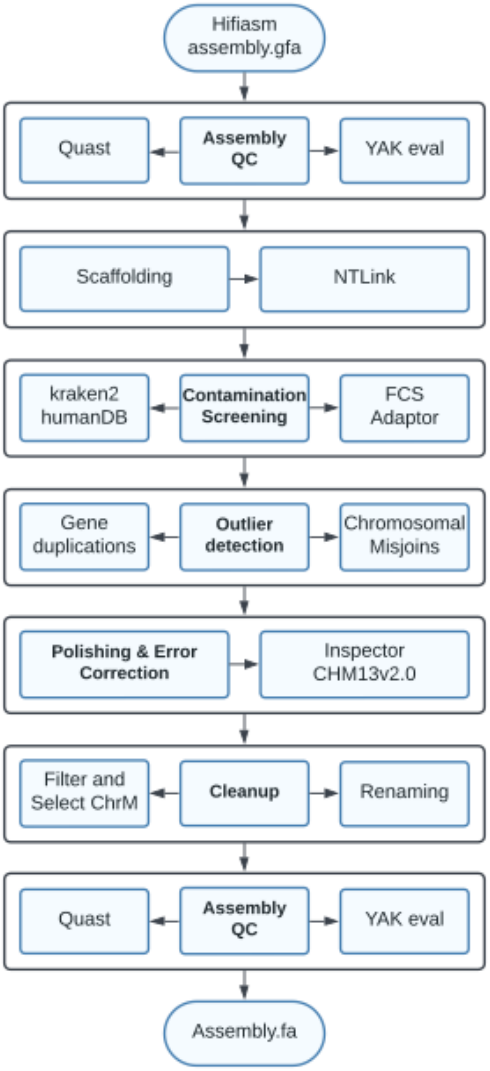
Supplementary Figure 1: Processing workflow for human chromosome-level genome assemblies. The diagram outlines steps for integrating sequencing data from multiple platforms (HiFi, ONT, and Illumina) to create a comprehensive trio-based diploid assembly. Following read trimming (purple) and long-read polishing (pink), the assembly stage (red) starts by generating Kmer databases of parental reads using Meryl and yak for the assembly pipeline Verkko and hifiasm, respectively. The generated draft assemblies are scaffolded (orange) using ntLink and further improved using RagTag for reference-guided assembly based on the CHM13v2.0 reference genome. The final assembly refinement involves gap-filling (green) using quarTeT, culminating in the gap-filled scaffolded assembly (dark blue).



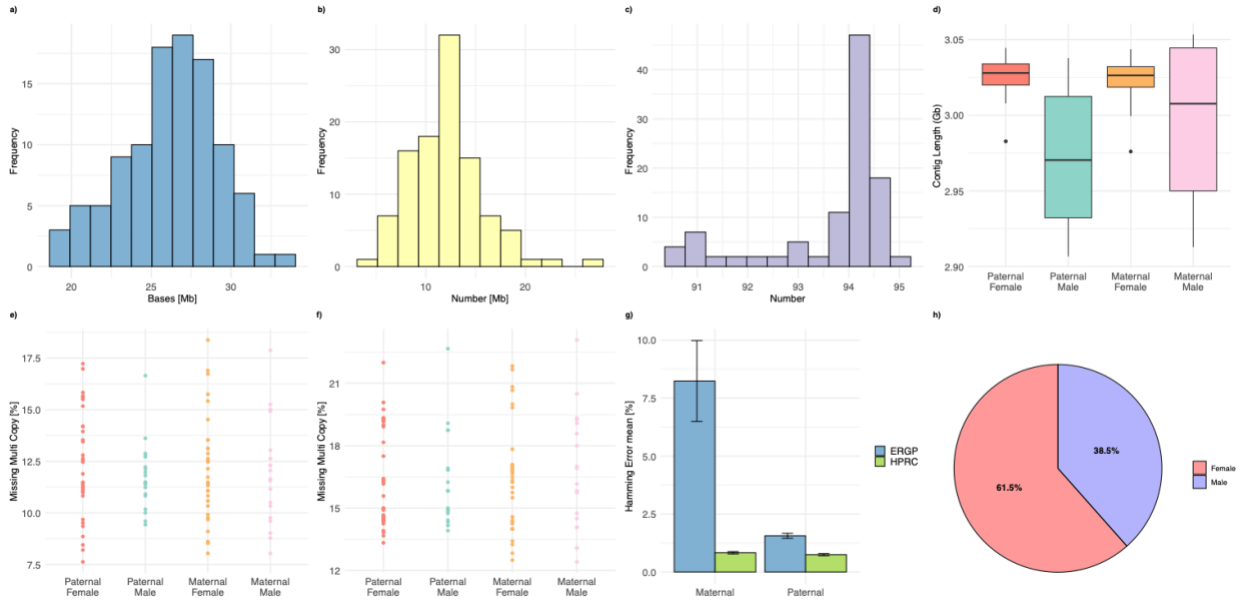
Supplementary Figure 2: Alignment of T2T-ERG_XX paternal haplotype with CHM13v2.0. All primary alignments are shown. (a) Dotplot. (b) The coverage of T2T-ERG_XX chromosomes with CHM13 chromosomes. (c) The coverage of CHM13 chromosomes with T2T-ERG_XX chromosomes. In (b) and (c), all assembly sequences with a length of more than 5 Mb are shown.



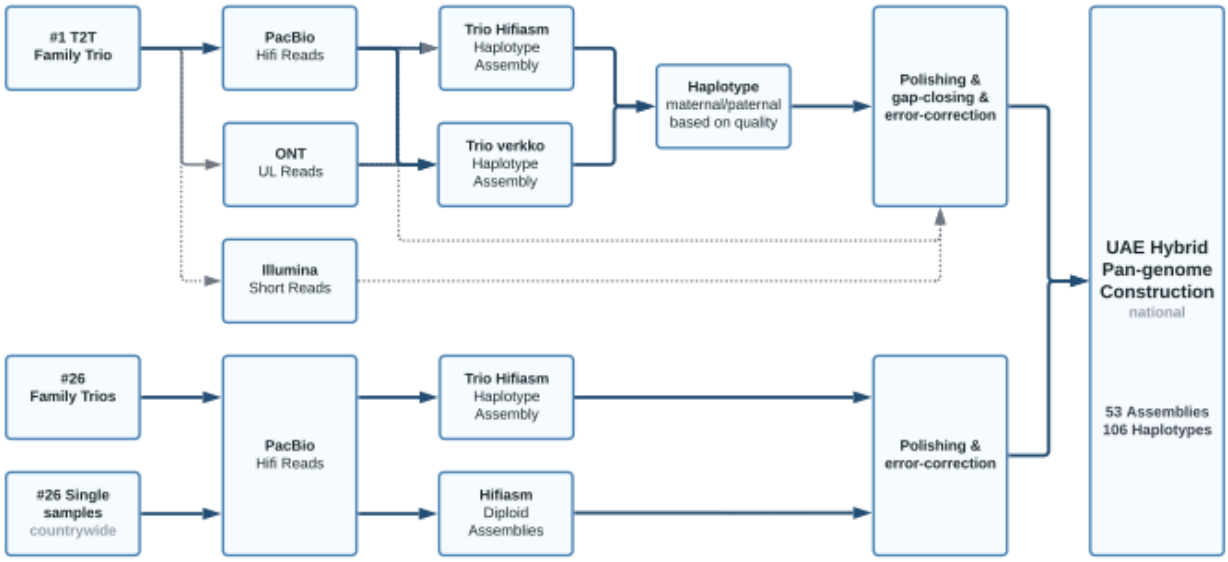
Supplementary Figure 3: The Figure depicts quality metrics calculated on the PacBio HiFi reads generated using Revio for the pangenome samples.: (A) The distribution by the number of sequences, demonstrating a normal distribution with most sequence counts clustered around the mean=6,443,018. (B) The distribution by coverage illustrates how well the sequence represents the genome reads with a coverage mean of 30.5. (C) The distribution by N50 indicates the length at which 50% of the genome is contained in reads of at least that length (mean=16,743). (D) The distribution by N90 indicates the length at which 90% of the genome is contained in reads of at least that length (mean=10,591).



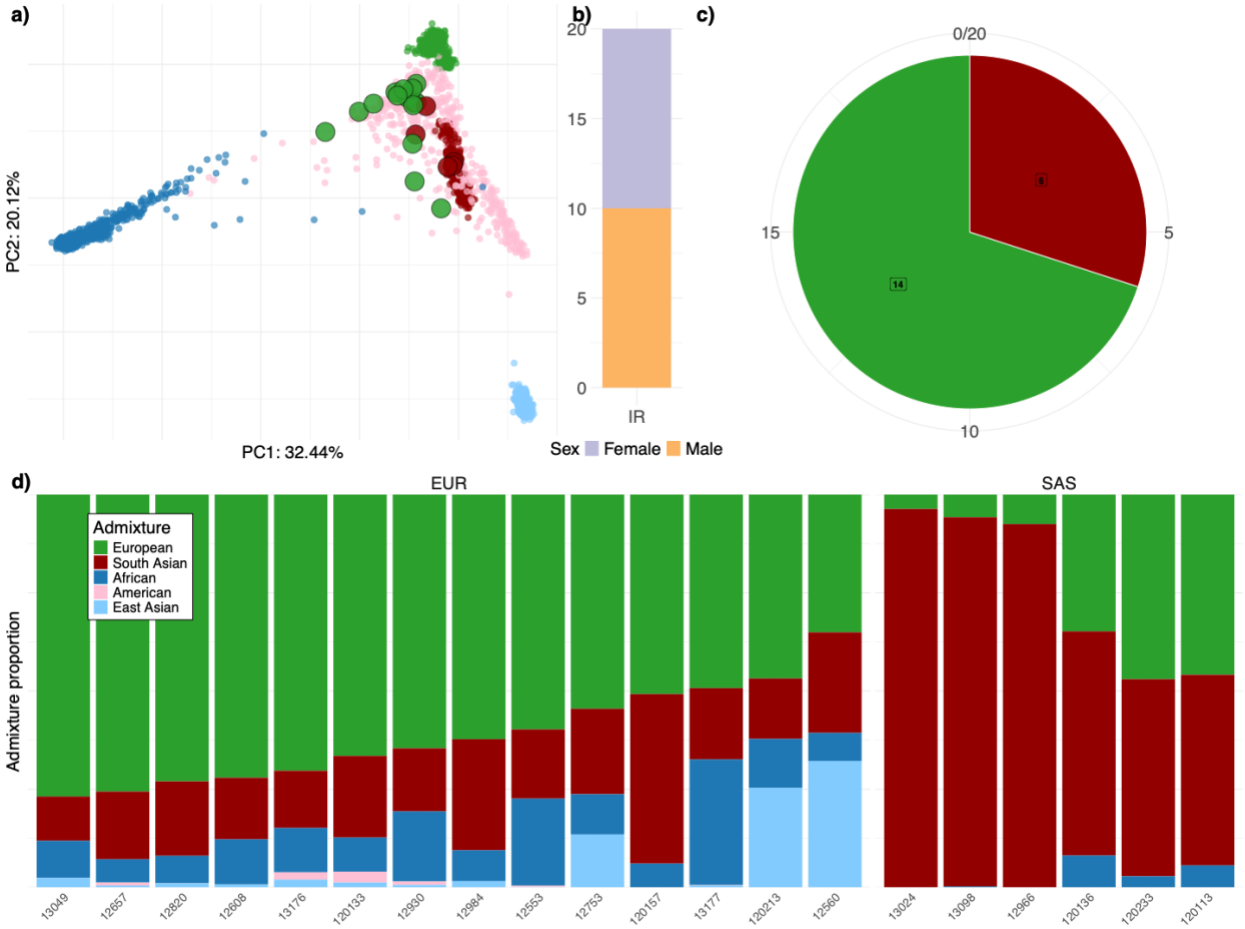
Supplementary Figure 4: Workflow for post-assembly processing of hifiasm-generated assemblies in pan-genome construction. The initial assembly undergoes quality control (QC) using Quast and yak evaluation. Scaffolding is performed with NTLink, followed by contamination screening using Kraken2 with the human database and FCS Adaptor to identify contaminants. Outlier detection assesses chromosomal misjoins and gene duplications. Polishing and error correction are executed with reference to CHM13v2.0 using Inspector. Cleanup steps include filtering and selecting mitochondrial chromosomes (ChrM) and renaming. The final assembly is subjected to another round of QC with Quast and yak evaluation to ensure quality and accuracy.



Supplementary Figure 5: The figure shows genome assembly and quality assessment evaluations across various panels. Panel (a) presents the distribution of unaligned contig lengths megabases (Mb). Panel (b) depicts the N90 statistic for contig sizes (Mb). Panel (c) shows the histogram of genome coverage for all 104 haplotypes. Panel (d) compares the assembly length (Gb) across maternal and paternal haplotypes and TA and SA assembly strategies. Panel (e) and Panel (f) display percentages of missing multi-copy regions with respect to GRCh38 and CHM13, respectively. Panel (g) contrasts Hamming error rates (mean percentage) between maternal and paternal genomes, as evaluated by the ERGP and HPRC datasets. Panel (h) summarizes the sex composition of the analyzed dataset in percentage values, indicating the proportions of male and female genomes.



Supplementary Figure 6: Workflow for the Emirati hybrid pangenome construction, integrating sequencing and assembly of a T2T family trio, 26 family trios, and 26 single samples using PacBio HiFi, ONT ultra-long, and Illumina reads. Assemblies are generated with trio hifiasm and trio Verkko, while hifiasm also produces diploid assemblies for single samples. Maternal and paternal haplotypes are selected based on quality, followed by polishing, gap-closing, and error correction. The final Emirati hybrid pangenome comprises 53 assemblies and 106 haplotypes, capturing the genetic diversity of the Emirati population.



Supplementary Figure 7: Basic and ancestry-based characteristics for a cohort of 20 Emirati individuals with Illumina whole genome sequencing data used for T2T-ERG_XX- and EPR graph-based variant calling evaluation. (a) The genomes are projected onto the genotype principal components computed by the 1000 Genomes phase 3 samples. Each sample is colored according to its 1000 Genomes-based global ancestry component with maximum value. (b) Shows the cohort stratified by sex. (c) The distribution of the genomes' main ancestry components across the cohort. (d) The contributing ancestry components for each sample after supervised admixture analysis based on 1000 Genomes global populations, i.e., European, South Asian, African, American, and East Asian.