SUPPLEMENTARY INFORMATION

A Draft Arab Pangenome Reference

Correspondence and requests for materials should be addressed to Mohammed Uddin (mohammed.uddin@mbru.ac.ae) or Alawi Alsheikh-Ali (alawi.alsheikhali@mbru.ac.ae)

Contents

Supplementary Methods

1. APR cohort and sample phenotype information

We have recruited all samples from different outpatient clinics within Dubai Academic Healthcare Corporation (DAHC). The trio were recruited from Al Jalila Children's Hospital, Dubai, UAE. The unrelated 40 Emirati individuals were recruited from DAHC healthcare facilities who were clinically assessed for multiple years by family physicians and identified as healthy. For all adult healthy 40 samples, we have obtained multiyear clinical health records to identify various physiological parameters and their distribution. The project research genetic counselor independently interviewed regarding the absence and presence of most of the common chronic conditions.

2. Pacific Bioscience High Fidelity (HiFi) Sequencing

We constructed whole genome sequencing (WGS) libraries using the long-read PacBio SMRTbell prep kit 3.0 protocol (102-166-600) as detailed below.

a. DNA quality control

To evaluate and ensure suitable quality and size of DNA for the use of this protocol, we began by homogenizing the DNA in the solution by pulse vortexing and gently pipetting our samples. Quick spinning the samples, we took 1 µL aliquot from each sample and diluted them with 9 µL of elution buffer or water. Next we measured the DNA concentration using the 1X dsDNA HS kit and the Qubit fluorometer. Each aliquot was diluted in Tape Station dilution buffer to a total concentration of 250 pg/µL, based on the Qubit readings. DNA size was measured using the Femto Pulse system and the gDNA 165Kb analysis kit.

b. DNA shearing and cleanup

Low TE buffer was added to bring the volume of DNA samples to 130 µL, with a target concentration of 30 ng/µL. The DNA was sheared on the Megaruptor 3 system as per the

recommended settings. Sheared DNA was transferred into a tube, and 1 X v/v of resuspended, room-temperature SMRTbell cleanup beads were added. The beads were pipette mixed until evenly distributed. Tubes were then quick-spinned and incubated at room temperature for 10 min. After incubation, tubes were placed in a magnetic separation rack to allow the beads to separate fully and the supernatant was discarded. The beads were slowly covered by adding 200 µL of fresh 80% ethanol. After 30 seconds, the ethanol was discarded, and the step was repeated.

To remove the residual ethanol, the tubes were removed from the magnetic rack, spun, placed again in the rack until the beads separated, and the remaining solution was aspirated. Immediately after removing the tubes from the magnetic rack, 47 µL of low TE buffer were added to the tubes and the beads were resuspended by pipetting 10 times. The tubes were quick-spinned and incubated at room temperature for 5 min for DNA elution. Next, the tubes were placed again in the magnetic rack, beads separated, and the supernatant transferred to a new tube. To verify sample quality per its concentration and size distribution, we aliquoted 1 µL from each sample and diluted with 9 µL elution buffer or water. Next, the Qubit fluorometer was used to measure DNA concentration using the 1X dsDNA HS kit. Each aliquot was diluted to 250 pg/µL in Tape Station dilution buffer and DNA size was measured using the Tape Station system.

c. Repair and A-tailing

Into a centrifuge tube, we created reaction mix 1 by adding repair buffer (35.2 µL for 4 libraries, 80 µL for 8 libraries), end repair mix (17.6 µL for 4 libraries, 40 µL for 8 libraries), and DNA repair mix (8.8 µL for 4 libraries, 30 µL for 8 libraries). After pipetting and spinning the reaction mix, we added 14 µL of reaction mix 1 to each of our samples, bringing the total reaction volume to 60 µL. Then we pipetted and spinned our sample reactions, followed by running the repair and A-tailing thermocycling program, 37°C for 30 min, 65°C for 5 min, and a 4°C hold.

d. Adapter ligation

Since we barcoded our samples, we added 4 µL of barcoded adapters from the SMRTbell barcoded adapter plate 3.0 to each sample. In a centrifuge tube, we created reaction mix 2 by adding ligation mix (132 µL for 4 libraries, 300 µL for 8 libraries), and ligation enhancer (4.4 µL for 4 libraries, 10 µL for 8 libraries). After pipetting and spinning the reaction mix, we added 31 µL of reaction mix 2 to each of our samples, bringing the total reaction volume to 95 µL. Then we pipetted and spinned our sample reactions, followed by running the adapter ligation thermocycling program, 20°C for 30 min and a 4°C hold.

e. Cleanup

95 µL of resuspended SMRTbell cleanup beads were added to each sample and pipetted for even distribution. The tubes were spun and incubated at room temperature for 10 min. Tubes were then placed in a magnetic separation rack, when the beads separated from the solution the supernatant was discarded. The beads were slowly covered by adding 200 µL of fresh 80%

ethanol. After 30 seconds, the ethanol was discarded, and the step was repeated. To remove the residual ethanol, the tubes were removed from the magnetic rack, spun, placed again in the rack until the beads separated, and the remaining solution was aspirated. Immediately after removing the tubes from the magnetic rack, 40 µL of elution buffer were added to the tubes and the beads were resuspended. The tubes were quick-spinned and incubated at room temperature for 5 min for DNA elution. Next, the tubes were placed again in the magnetic rack, beads separated, and the supernatant transferred to a new tube.

f. Nuclease treatment

In a centrifuge tube, we created reaction mix 3 by adding nuclease buffer (22 µL for 4 libraries, 50 µL for 8 libraries), and nuclease mix (22 µL for 4 libraries, 50 µL for 8 libraries). After pipetting and spinning the reaction mix, we added 10 µL of reaction mix 3 to each of our samples, bringing the total reaction volume to 50 µL. Then we pipetted and spinned our sample reactions, followed by running the nuclease treatment thermocycling program, 37°C for 15 min and a 4°C hold.

g. AMPure PB bead size selection

By adding 1.75 mL of resuspended AMPure PB beads to 3.25 mL of elution buffer, we made a 35% v/v dilution. Then we added 155 µL (3.1 X v/v) of resuspended 35% AMPure PB beads to each of our samples.

h. Binding, washing, and eluting

The samples were pipette mixed and spun, then incubated at room temperature for 20 minutes. Tubes were then placed in a magnetic separation rack, when the beads separated from the solution the supernatant was discarded. The beads were slowly covered by adding 200 µL of fresh 80% ethanol. After 30 seconds, the ethanol was discarded, and the step was repeated. To remove the residual ethanol, the tubes were removed from the magnetic rack, spun, placed again in the rack until the beads separated, and the remaining solution was aspirated. Immediately after removing the tubes from the magnetic rack, 15 µL of elution buffer were added to the tubes and the beads were resuspended by pipetting 10 times. The tubes were quick-spinned and incubated at room temperature for 5 min for DNA elution. Next, the tubes were placed again in the magnetic rack, beads separated, and the supernatant transferred to a new tube. From each tube, a 1 µL aliquot is taken and diluted with 9 µL of elution buffer or water. DNA concentration was measured using a Qubit fluorometer and the 1X dsDNA HS kit.

i. ABC Steps

After completing library preparation, we performed ABC sequencing preparation as per the (102-739-700) kit protocol, detailed below.

A. Annealing Sequencing primer

Following the DNA quantification, we diluted our samples to a concentration of 30 ng/μL. In a new Lo-bind tube, we added 11 μL of sample, 5/5 μL of annealing buffer, and 5.5 μL of sequencing primer, bringing the total volume to 22 μL. After pipetting well, the tubes were incubated at room temperature for 15 min.

B. Sequencing polymerase

We began by diluting 11.5 μL of sequencing polymerase in 563.5 μL of polymerase buffer in a new lo-bind tube, bringing the total volume to 575 μL diluted polymerase master mix to be used immediately. Next, for each of our samples, we added 22 μL annealed sample and 22 μL diluted polymerase, bringing the total volume to 44 μL. The tubes were then incubated at room temperature for 15 min.

C. Purification of Polymerase bound SMRTbell complexes

SMRTbell cleanup beads and loading buffer were brought to room temperature. For each sample, 44 μL of Binding reaction and 56 μL of dilution buffer were added, bringing the total reaction to 100 μL. Next, 120 μL of SMRTbell cleanup beads were added to the samples, gently pipetted, and incubated at room temperature for 10 min. Tubes were then placed in a magnetic bead rack until the beads completely separated from the solution. The supernatant was discarded and the beads were immediately resuspended in 50 μL loading buffer and gently pipetted. For elution of the polymerase-bound complexes, the samples were incubated at room temperature for at least 5 min. Tubes were placed again on the magnetic rack until the beads separated. Then the eluted solution was transferred to new Lo-Bind tubes and placed on ice, protected from the light.

D. Sequencing control dilution

The control underwent three dilution steps, producing a master mix that is enough for all our samples. In a new Lo-bind tube, 19 μL dilution buffer and 1 μL sequencing control were mixed by flicking the tube and pulse spinning, then kept on ice. This process was repeated again to achieve dilution 2. For the third and final dilution, 76 μL dilution buffer and 4 μL diluted sequencing control from the second dilution were mixed well by flicking and spinning, and kept on ice.

E. Loading dilution

For each of the samples loaded on the Revio system, Revio Polymerase Kit was used (Serial Number: 102-739-700). We combined 50 μL of the prepared sample, 47 μL loading buffer, and 3 μL of the diluted sequencing control from the previous step, bringing the total volume to 100

µL. 95 µL of each sample were loaded per well. For samples loaded on Sequel, the starting library concentration was 40 ng/µL. Sample dilution and loading followed the same protocol but differed in concentrations dependent on the starting sample concentration.

3. Nanopore ultra-long sequencing

We used Oxford Nanopore Technologies ultra-long read sequencing kit (SQK-ULK114) protocol. Each sample was aliquoted at the time of collection into 1.8 ml Cryovials and stored at -80°C. 3 aliquots of each sample was used for preparing ultra-long libraries generating about 50× coverage of unsheared sequencing from 3 PromethION flow cells (R10.4.1) and a N50 value of around 60 kb. We have used additional flow cells (see supplementary table 2) for samples that did not yield adequate data.
Ultra-long libraries were prepared using the protocol outlined below:

a. PBMC isolation from frozen blood

1.6 ml frozen blood was thawed in 3X volume of cold NEB RBC lysis buffer in a 15 ml Falcon tube. Tube was gently inverted 10 times and kept at 4°C for 5 minutes. The tube was centrifuged at 2000 X g at 4°C for 2 min, after which the supernatant was discarded. The pellet was resuspended in 1.6 ml 1X PBS buffer by gently flicking and pipetting. The above steps were repeated for a total of 3 washes with NEB RBC lysis buffer. The final pellet of around 60 million cells was resuspended in 40 µl of 1X PBS and carried forward for ultra-long DNA extraction.

b. DNA extraction

Resuspended cells in 40ul of PBS were transferred to a fresh 5 ml Lo-Bind Eppendorf tube. In a separate tube, 1.8 ml NEB Monarch tissue lysis buffer and 40ul of NEB Monarch Proteinase K was mixed, then added to the resuspended PBMCs in the Lo-Bind tube. The reaction was mixed by slow pipetting five times using a 1 ml wide-bore pipette tip and incubated at 56°C on a thermomixer for 10 min. 15ul of NEB Monarch RNase A was added to the reaction and mixed by slow pipetting five times using a 1 ml wide-bore pipette tip. The reaction was incubated at 56°C on a thermomixer for 10 min at 650 rpm. Next, 900 µl of NEB Monarch protein separation solution was added to the reaction and the tube was placed in a Hula mixer for 10 mins at 3 rpm. The 5 ml tubes were then centrifuged for phase separation at 16,000 X g at 4°C for 10 min. Upper phase containing ultra-long DNA fragments was carefully collected into a fresh 5 ml Lo-Bind Eppendorf tube using a 1 ml wide-bore pipette tip. To this, 3 glass beads were added along with 2.5 ml isopropanol. The tube was placed in a Hula mixer for 20 mins at 3 rpm.

The ultra-long DNA fragments were precipitated around the 2 glass beads by the end of the incubation. The supernatant was discarded and the glass beads were washed two times using 2ml

NEB Monarch wash buffer per wash. The glass beads were then introduced to a Monarch bead retainer inserted into a Monarch collection tube, and centrifuged at 1000 X g for 1 minute to remove any remaining wash buffer. The glass beads were immediately moved into a fresh 2 ml Lo-Bind Eppendorf tube containing 560 µl of ONT extraction elution buffer. The tube was incubated at 56°C for 10 min. The beads and the elution buffer were introduced into a clean bead retainer inserted into a Monarch collection tube, and centrifuged at 1000 X g for 1 minute. The eluate was collected in the collection tube. The beads were discarded. To the collection tube, 200 µl of Oxford Nanopore Technologies extraction elution buffer (EEB) was added and the total volume of 760 µl of eluant was transferred to a fresh 1.5 ml Lo-Bind Eppendorf tube. The reaction was incubated at 56°C for 10 min. The final eluate of UHMW DNA was mixed by slow pipetting five times using a 1ml wide-bore pipette tip and stored overnight at room temperature.

c. DNA Tagmentation

In a 1.5 ml Lo-Bind Eppendorf tube, 5 µl of ONT ULK fragmentation mix (FRA) and 245 µl of FRA dilution buffer (FDB) were mixed by pipetting. 250 µl of the diluted fragmentation mix was added to 760 µl of UHMW DNA. The reaction was immediately mixed by slow pipetting ten times using a 1 ml wide-bore pipette tip. The reaction was incubated at room temperature for 10 min, followed by incubation on a thermomixer at 75°C for 10 min, and finally cooled on ice for 15 min.

d. Adapter attachment

5ul of rapid adapter (RA) was added to the DNA reaction and gently mixed by slow pipetting five times using a 1 ml wide-bore pipette tip. The reaction was incubated at room temperature for 30 min.

e. Clean-up

A metal precipitation star (PS) was added to the adapted DNA for clean-up. 500 µl of precipitation buffer (PTB) was added to the reaction and mixed by rotating on a Hula mixer for 20 min at 3 rpm. The adapted ultra-long DNA was precipitated around the precipitation star and the supernatant was discarded. The tube was briefly spun to remove any remaining supernatant. 300 µl of elution buffer (EB) was added to the tube containing the metal star and DNA. The tube was stored overnight at room temperature. Using a 1ml wide-bore pipette tip the ultra-long DNA library was removed and retained in a fresh 1.5 ml Lo-Bind Eppendorf tube. The tube containing the precipitation star was briefly spun and any remaining eluate was transferred to the tube containing the final ultra-long DNA library. The final library was gently mixed by slow pipetting five times using a 1 ml wide-bore pipette tip.

f. Flow cell loading and sequencing

ONT sequencing buffer (SQB) (100 µl) and ONT loading solution (10 µl) were added to 90 µl of the ultra-long DNA library from above. The mixture was gently mixed by slow pipetting five times using a wide-bore pipette tip. Libraries were then incubated at room temperature for 30 min. Next, the libraries were gently mixed by slow pipetting with a wide-bore tip to ensure homogeneity. Before loading the library, the flow cell was primed with flush buffer/flush tether mixture per ONT directions. The library was then added to the flow cell. The mixture was viscous and loaded in a drop-wise manner using a 1 ml wide-bore tip. The sequencing run had a pore scan time set for every 1.5h and a minimum read length set to 1000 bp. Live base calling using the high accuracy (HAC) basecalling model with modified base detection for 5mC was performed on MinKNOW.

4. *de novo* genome assembly

High-quality *de novo* assemblies were generated for each sample using a hybrid assembly approach, combining PacBio HiFi and ONT ultralong reads. For each of the samples we ran HiFiAdapterFilt on the ubams to clean up any reads with adapters still attached. The resulting fastq files were combined into one file per sample using the zcat utility. We also combined all of the ultra-long DNA sequencing kit (ULK) reads (in fastq format) into one fastq file before using it along with the HiFi reads to build assemblies. Further, Hifiasm v0.19.5-r593 (ref.[1]) was used with the following arguments --ul APR029.ul.fastq and APR029.hifi.fastq.

For the trio, Verkko and Hifiasm was run to produce completely phased assemblies. To run Verkko, kmer databases were built using Meryl with the following command using the paternal HiFi reads

```
meryl count compress k=30 threads=XX memory=YY maternal.*fastq.gz output
maternal_compress.k30.meryl
meryl count compress k=30 threads=XX memory=YY paternal.*fastq.gz output
paternal_compress.k30.meryl
$MERQURY/trio/hapmers.sh maternal_compress.k30.meryl paternal_compress.k30.meryl
```

Verkko pipeline was run using the following command:

```
verkko -d asm  --hifi son.hifi.fastq  --nano  son.ul.fastq  --hap-kmers
maternal_compress.k30.hapmer.meryl paternal_compress.k30.hapmer.meryl  trio
```

To run Hifiasm, parental kmer databases were built using the following commands:

yak count -k31 -b37 -t16 -o pat.yak paternal.fq.gz
yak count -k31 -b37 -t16 -o mat.yak maternal.fq.gz

Hifiasm pipeline was run using the following sample command

hifiasm -o apr_son.asm -t 32 -1 pat.yak -2 mat.yak --ul son.ul.fastq son.hifi.fastq

5. Genome assembly polishing

Genome assemblies were screened for contaminant sequences and polished to remove artifacts prior to analysis. Kraken was used to taxonomically classify assembled contigs. The following sample command was used to first classify each contig:
kraken2 --memory-mapping --unclassified-out APR043.1.unclassified.fastq --classified-out APR043.1.classified.fastq --output APR043.1.out.bed --threads 32 --db kraken_db APR043.bp.hap1.p_ctg.fa

Once the classification was completed only the contigs that were classified as taxid 9606 were retained.

Mitochondrial read pairs were identified by mapping chrM to the contigs using minimap2 v2.26 using the following example command:
minimap2 -a -t 10 APR027.2.classified.fasta chrM.hg38.fasta.

Contigs that chrM mapped to were removed from the assemblies.

6. Assembly quality assessment

QUAST v5.2.0 (ref.[2]) was run with the following command: quast.py -o APR043.chm13v2.0 -r chm13v2.0.fa -t 16 APR043.bp.hap1.p_ctg.fa APR043.bp.hap2.p_ctg.fa --large -e

Quast was, however, not used to assess the correctness of the assemblies as it has been known to mischaracterize SV's as misassemblies. The yak suite v0.1-r66 of tools[1], including yak count and yak qv, were employed for genome quality validation.

yak count -t16 -b37 -o APR043.pb.yak APR043.pb.fastq.gz was first used to build kmer database and yak qv was used to calculate the QV score with the following parameters -t32 -p -K 3.2g -l 100k APR043.pb.yak APR043.bp.hap1.p_ctg.fa > APR043.hap1.pb.yak.qv.txt

Mapping rate was assessed using minimap2 v2.26 by mapping the HiFi reads to the assemblies. Potential misjoins in the assembly were identified using the minigraph and paftools.js. The specific commands initiated were -cxasm chm13v2.0.fa APR043.bp.hap1.p_ctg.fa and paftools misjoin -e APR043.1.misjoins.paf, respectively.

7. Evaluation of short read and long read whole genome mapping to APR graph

We used 21 WGS data from Arab descent[3] to quantify the mappability accuracy of short read sequences into APR pangenome and CHM13 reference. To calculate the mapping rate of short reads, we first aligned the paired end reads to CHM13v2 using bwa-mem and to APR graph reference using vg Giraffe. The following commands were used:

bwa mem -K 100000000 -Y -R '@RG\tID:APPG7555863\tLB:APPG7555863\tPL:ILLUMINA\tPM:HISEQ\tSM:APPG7555863' Filtered-APPG7555863_R1.fastq Filtered-APPG7555863_R2.fastq -t 64 > APPG7555863.sam

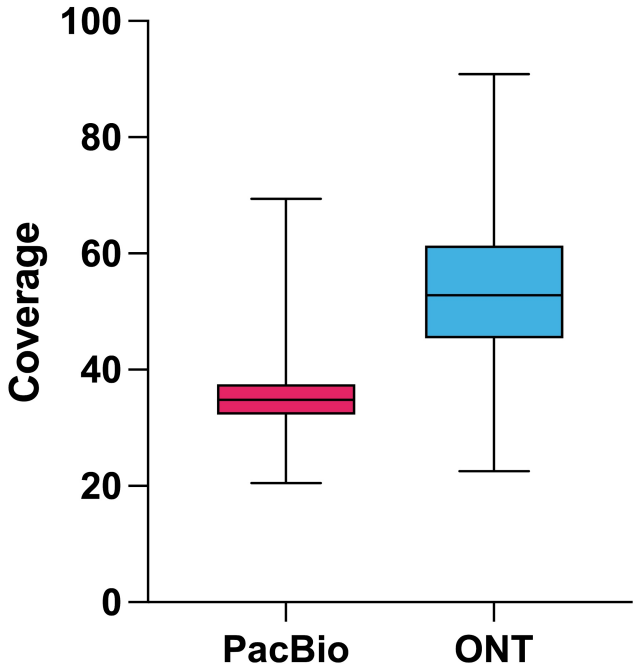
vg giraffe -p -t 48 -Z apr-v2.1-mc-chm13.gbz -d apr-v2.1-mc-chm13.dist -m apr-v2.1-mc-chm13.min -x /apr-v1.1-mc-chm38-without-hap.xg -f Filtered-APPG7555866_R1.fastq -f Filtered-APPG7555866_R2.fastq > APPG7555866.gam

We then used the samtools flagstat and vg stats on the resulting bam and gam files to produce stats for the properly paired reads for linear and graph references respectively. The following are examples of commands used:

samtools view -@ 64 -bS APPG7555866.sam | samtools flagstat -@ 64 - > APPG7555866_Flagstat.txt

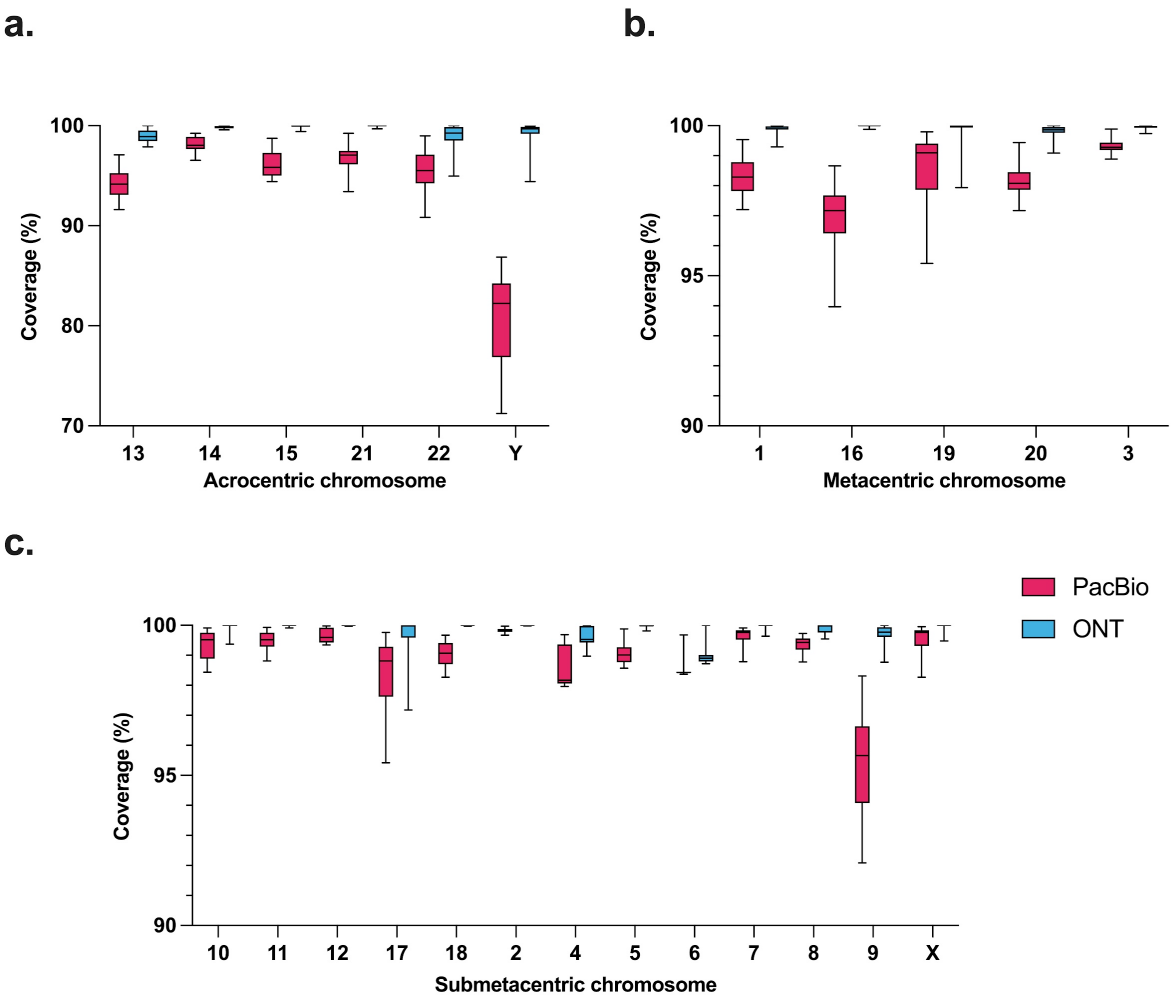
vg stats -p 32 -a APPG7555866.gam

Supplementary Figures



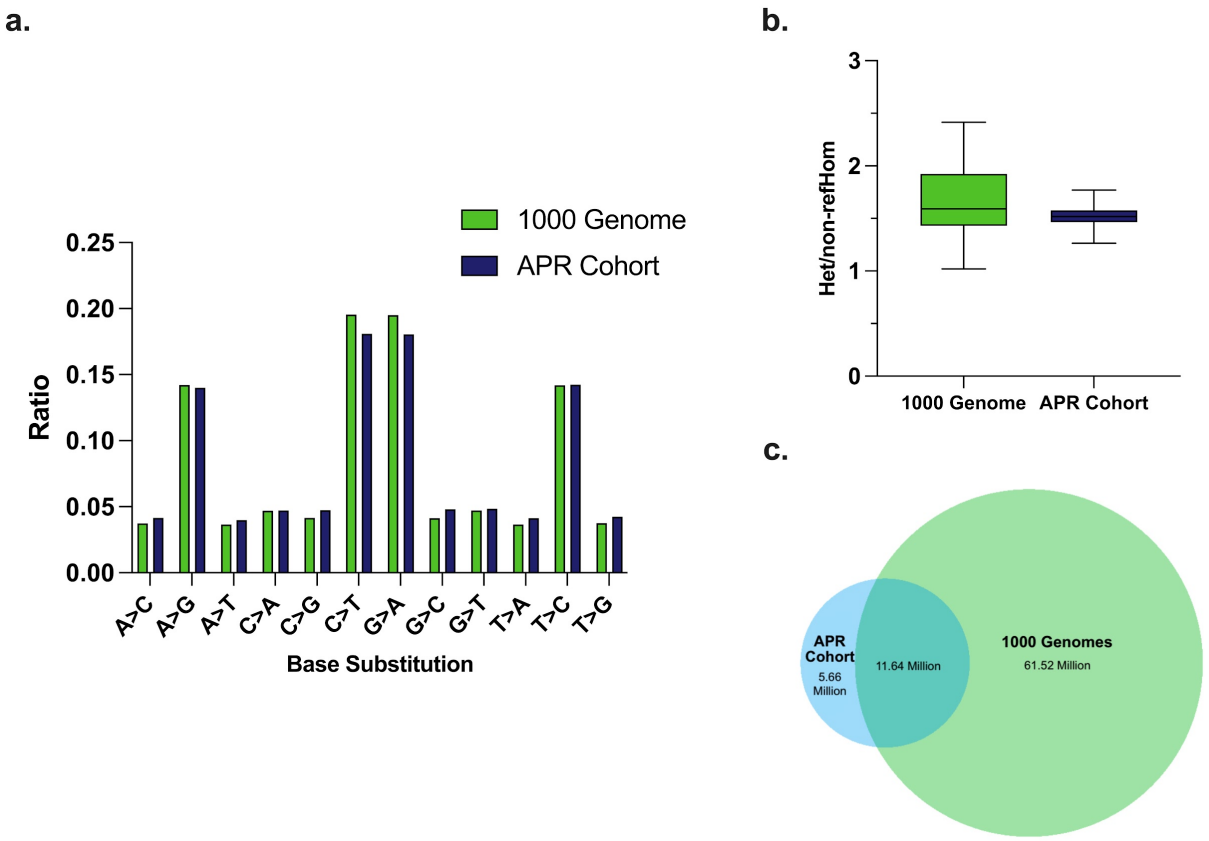
Supplementary Figure 1: APR long read sequencing coverage.

The y-axis represents the genome coverage for the entire data. The neon pink box represents data from PacBio high fidelity reads and the teal blue box represents ONT ultra long reads.



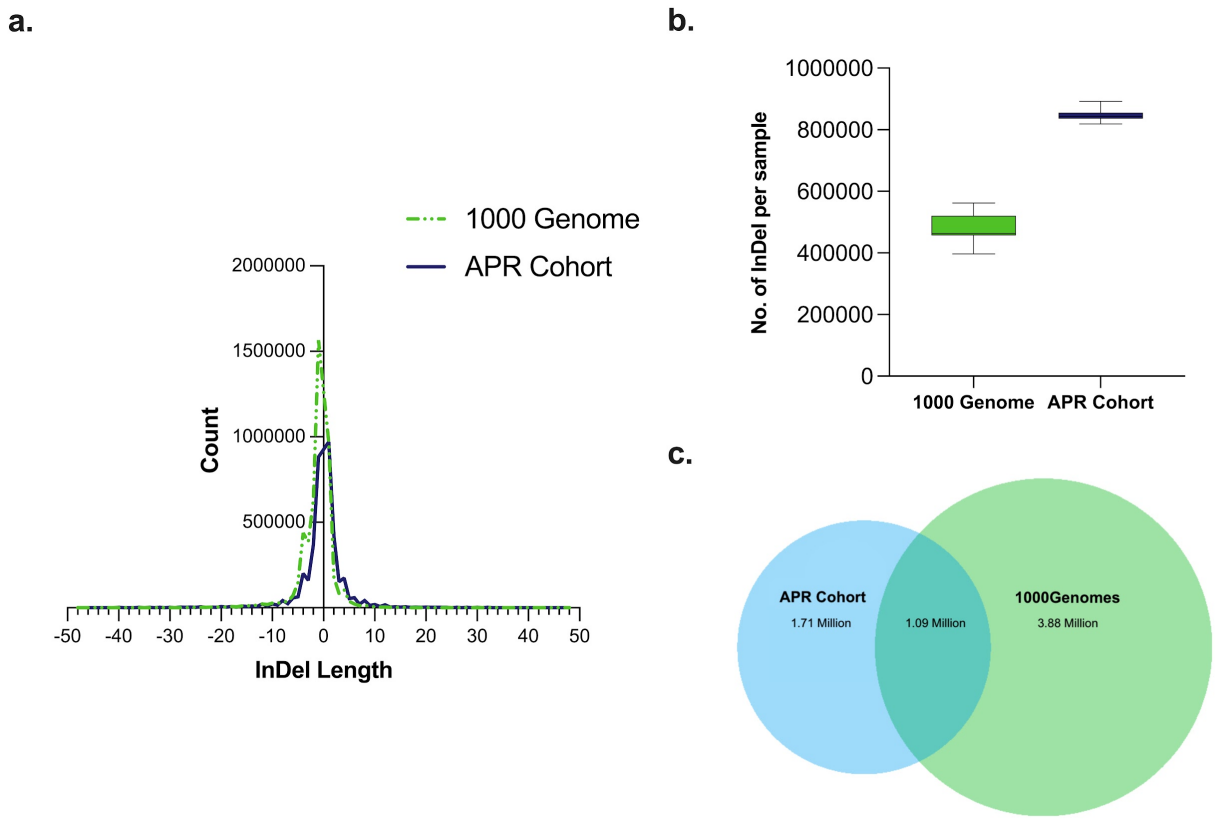
Supplementary Figure 2: Coverage of chromosomes from APR cohort.

Sample wise alignment to CHM13 across each chromosome comparing Hifi (red) and ONT (blue) reads. The box plot represents the distribution of coverage from all samples for **a**. acrocentric, **b**. metacentric and **c**. submetacentric chromosomes. Note the different Y-axis scales in acrocentric chromosomes.



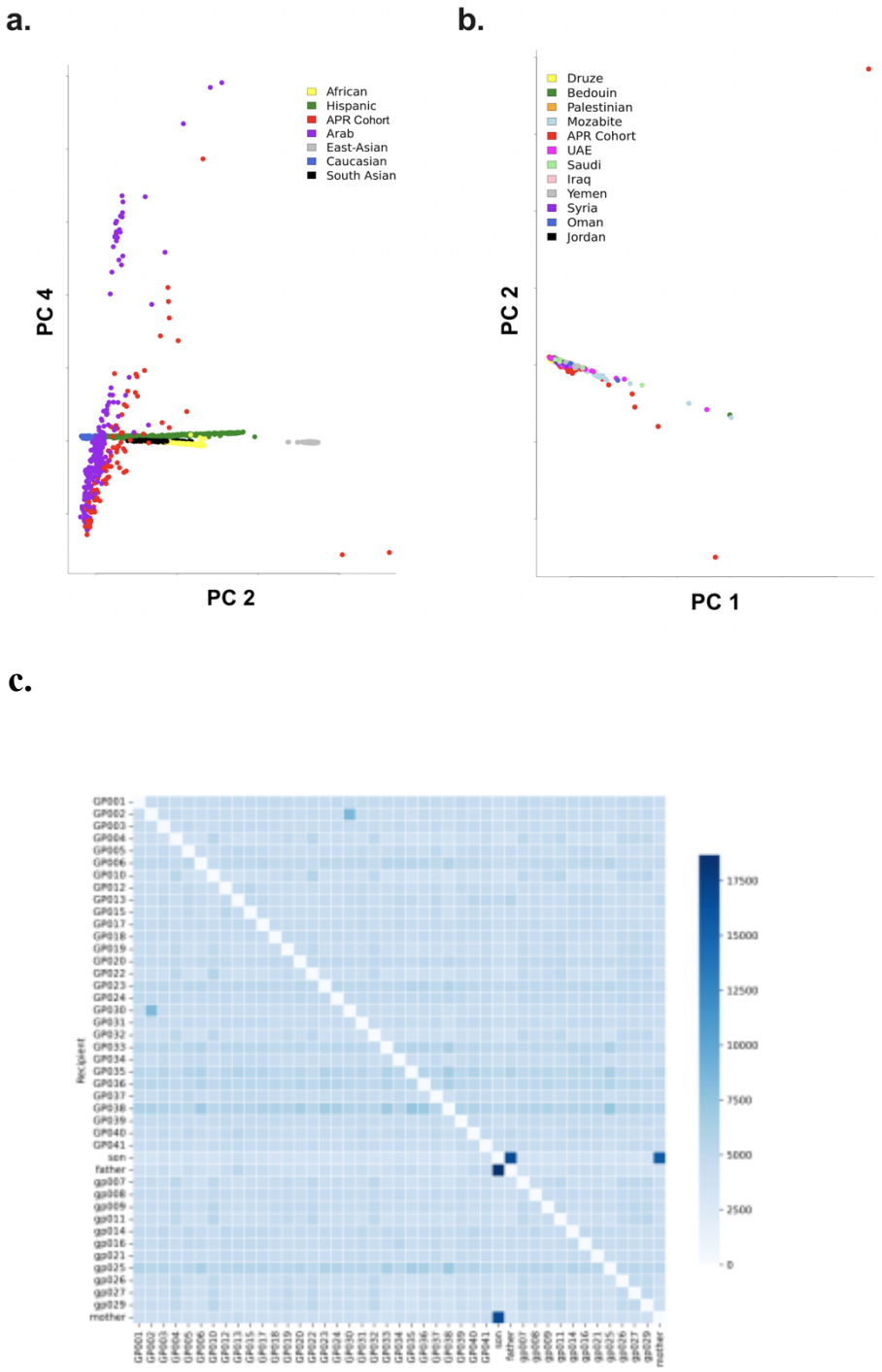
Supplementary Figure 3: APR SNV variant distribution from HiFi data in comparison with 1000 genomes.

a. The transition to transversion substitution ratio (y-axis) base between (x-axis) APR cohort and 1000G high coverage dataset[4]. **b.** The box plot distribution of het/non reference homozygous distribution between the two color coded cohorts. **c.** The venn diagram depicts the overlapped and unique variants from APR cohort (blue circle) and 1000 genomes (green circle). The size of the circles are proportional to the number of variants.



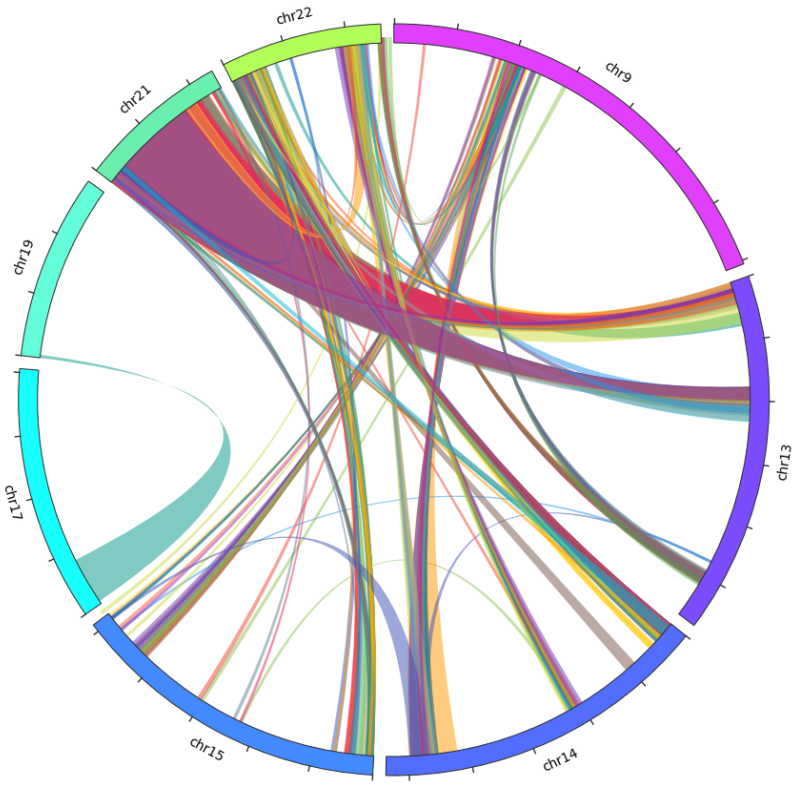
Supplementary Figure 4: Indel comparison with 1000 genomes.

a. The indel length (x-axis) and count (y-axis) distribution from APR (blue) and 1000 genome cohorts (green line) that has no whole genome data from Middle Eastern population. **b.** Total number of indel distribution, and **c.** the venn diagram depicting the overlapped and unique indels from two color (blue and green) coded cohorts. The size of the circles are proportional to the number of variants.



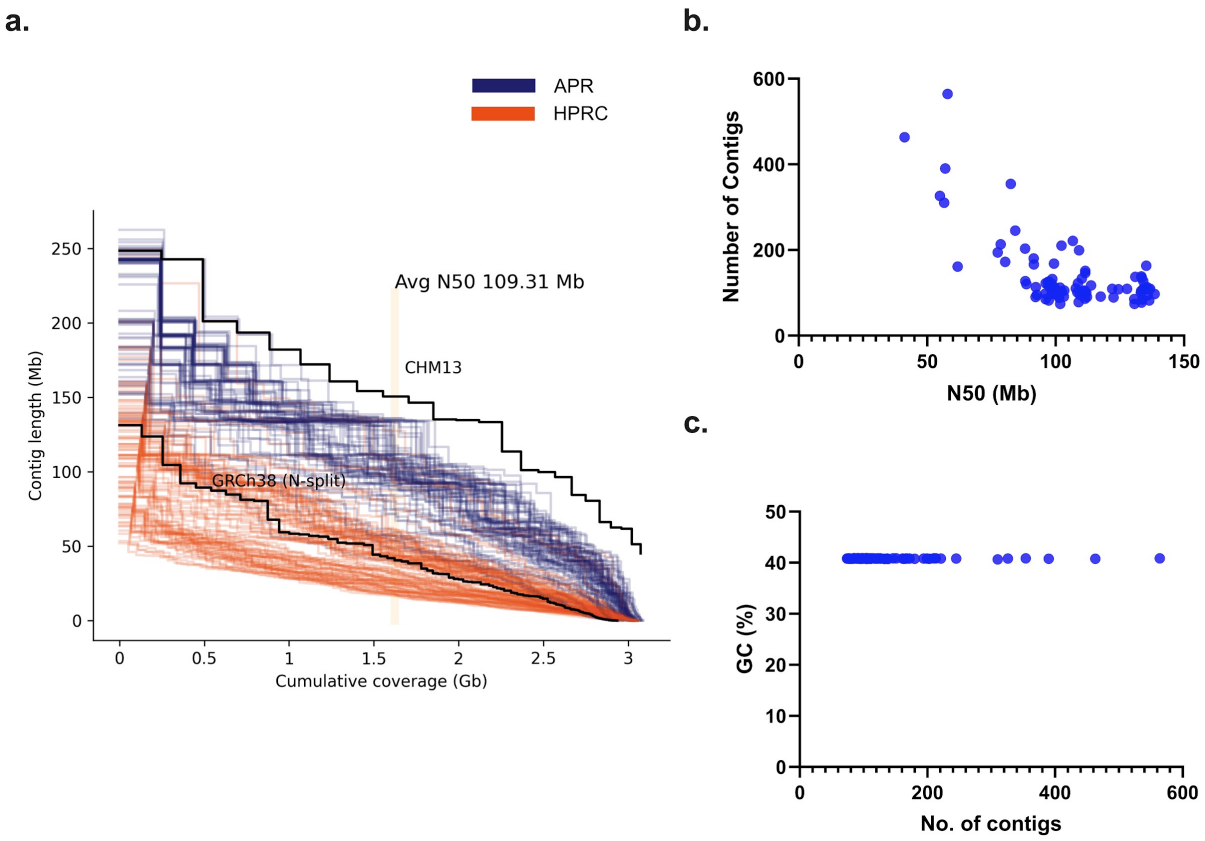
Supplementary Figure 5: Principal Component Analysis scatter plot.

Each dot, representing a sample, is color-coded to signify its association with APR cohort, Arab, or the 1000 Genomes Project. **a,** The rotated view indicate the variance captured by principal components (PC2) and 2 (PC4) of APR cohort alignment with global ethnicities. **b,** Scatter plot emphasizing the first PCA component clustering for APR cohort and other Arab ethnicities. The distribution is plotted using PC2 on the y-axis against PC1 on the x-axis. **c,** The matrix shows relatedness between APR samples and the trio (dark blue square) from fineSTRUCTURE haplotype sharing information.



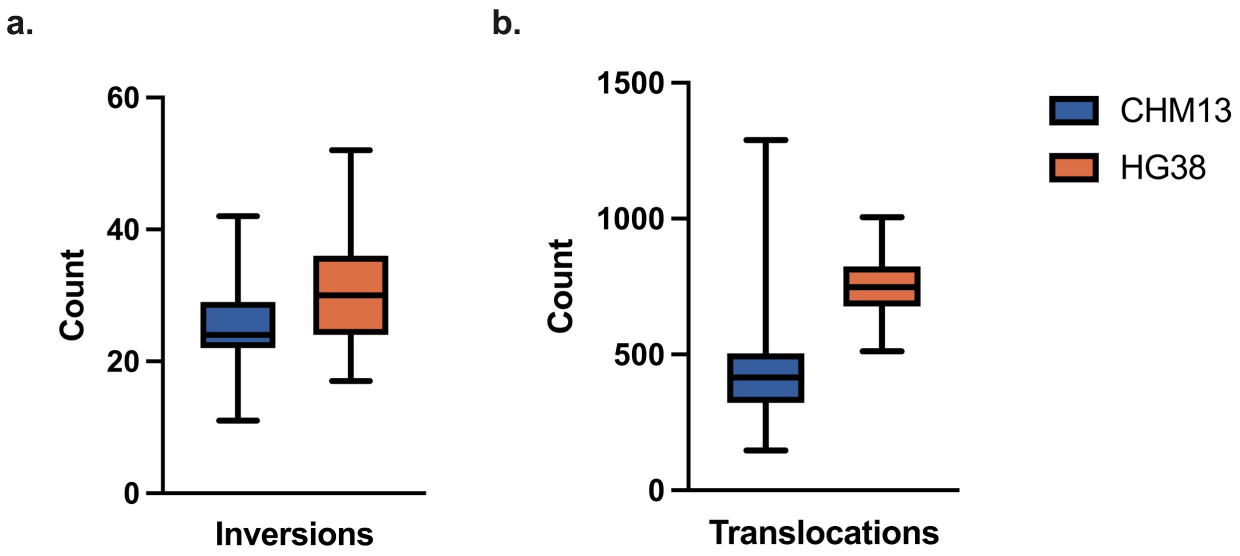
Supplementary Figure 6: Visualization of interchromosomal misjoins in APR assemblies.

The circular plot represents different chromosomes labeled as ‘chr’ followed by their respective numbers. Each connecting line signifies an interchromosomal misjoin, with the width of the line corresponding to the length of the misjoin. The most prominent misjoin, observed between chr21 and chr13, is also found in the HPRC dataset.



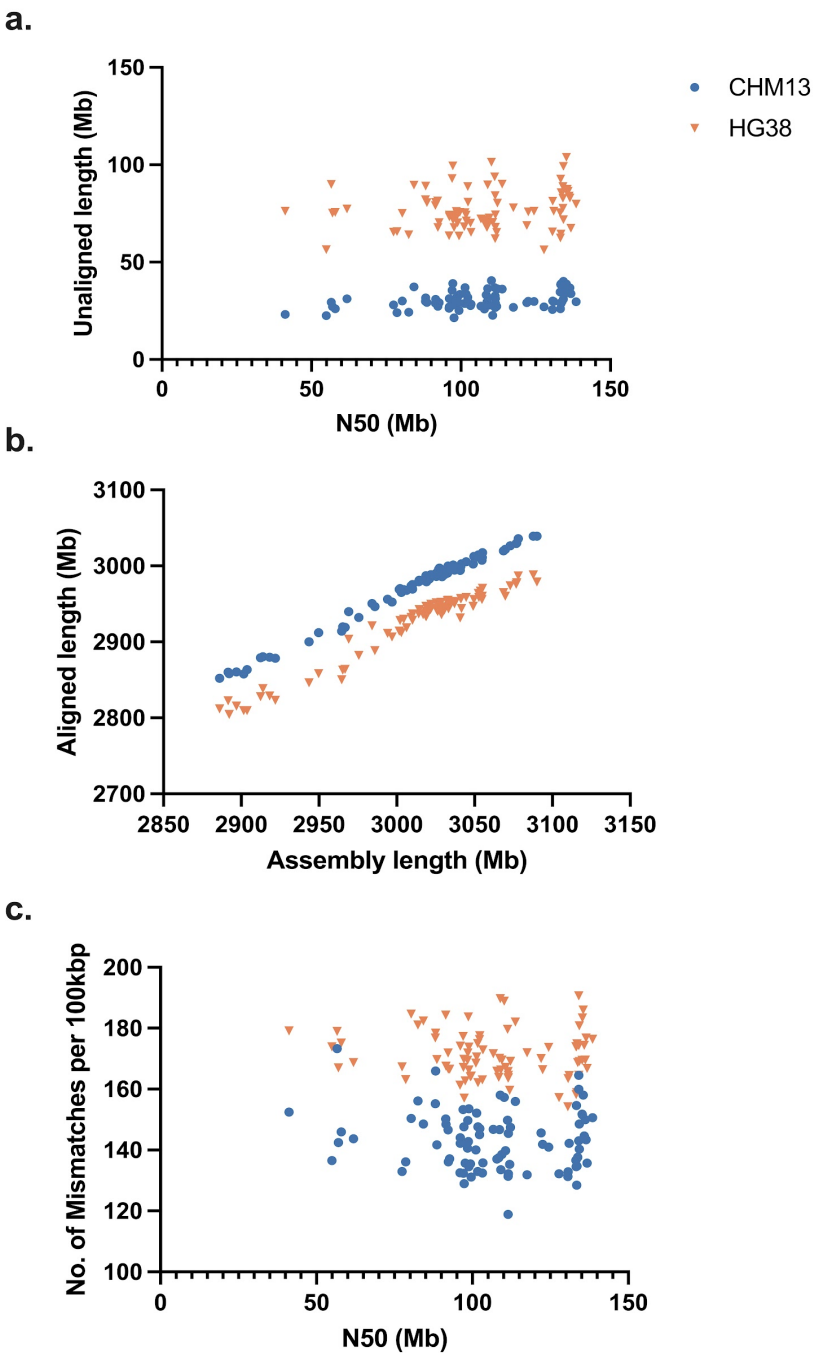
Supplementary Figure 7: Contig length distribution.

a, The contig length (y-axis) and cumulative coverage (x-axis) comparison between HPRC (orange lines) and APR (blue lines). **b,** The total number of contigs (y-axis) and their corresponding N50 length per sample (blue circles). **c,** The GC content (y-axis) for each sample and the contig numbers (x-axis) per sample.



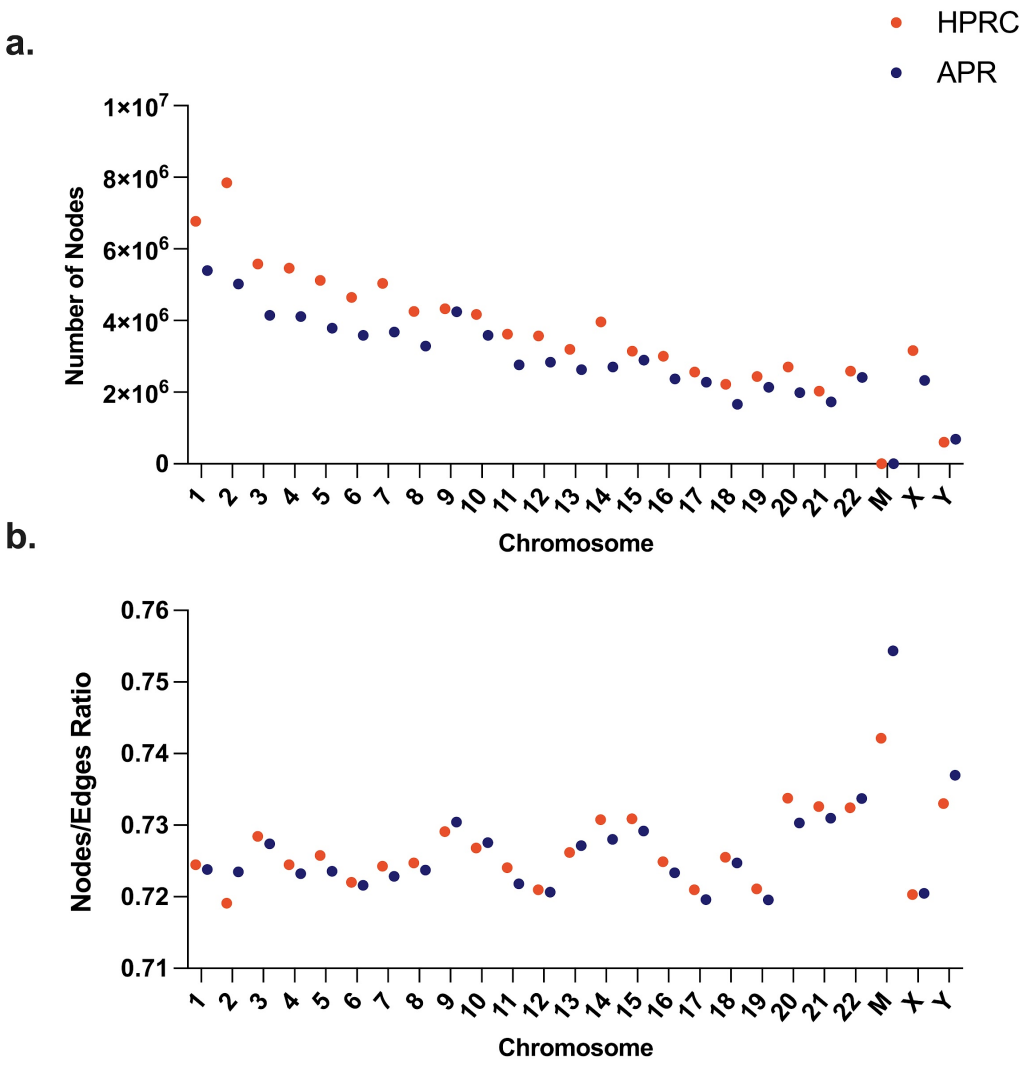
Supplementary Figure 8: Distribution of misassembled contigs identified as inversions and translocations.

a, Number of inversions detected with reference to CHM13 and GRCh38. **b,** Number of translocations detected with reference to CHM13 and GRCh38.



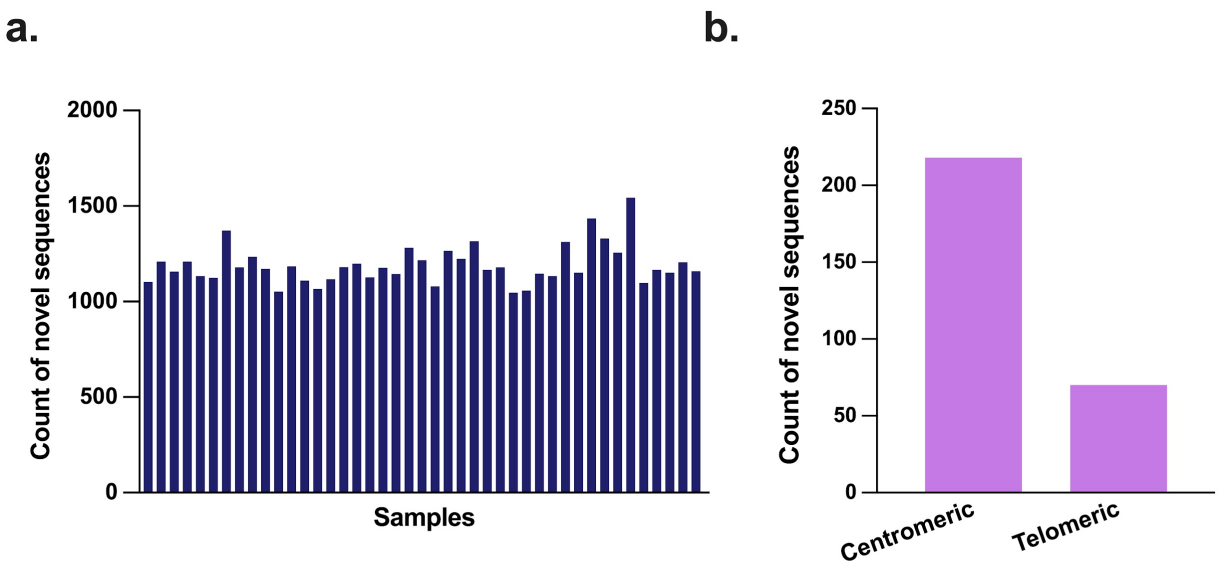
Supplementary Figure 9: Aligned and unaligned contigs.

The data points depicted in triangles represent hg38 and blue circles represent CHM13 genome reference. **a,** The x- and y-axis refer to unaligned contig length in mega bases (Mb) and their contig N50 length, respectively, for two references. **b,** The aligned bases (y-axis) and the assembly length (x-axis) for each APR sample. **c,** The mismatch in every 100,000 bases (y-axis) and the contig length (x-axis).



Supplementary Figure 10: Pangenome graph characteristics.

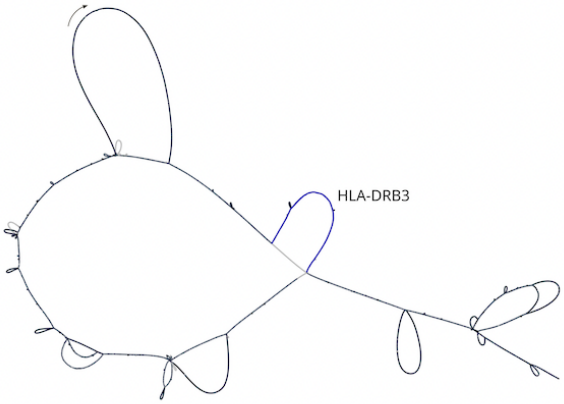
a. Each dot represents the number of nodes per chromosome. The blue and orange dot represents HPRC and APR pangenome graph node numbers, respectively. **b.** The Node/edge ratio comparison of to HPRC and APR pangenome for each chromosome.



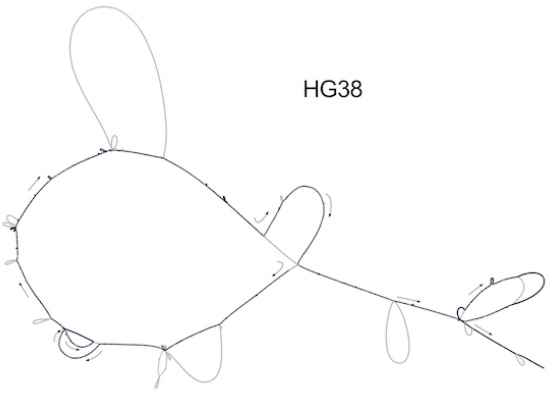
Supplementary Figure 11: Novel sequences with respect to T2T-CHM13 identified in the APR assemblies.

a, Bar graph indicating the count of newly discovered sequences for every sample, demonstrating the abundance of such sequences in the samples. **b,** Bar graph indicating counts of novel sequences across centromeric and telomeric regions.

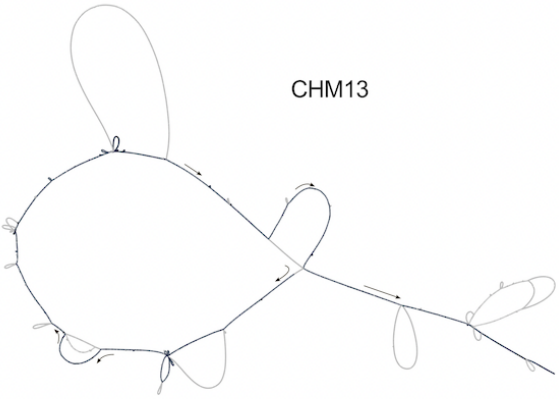
a.



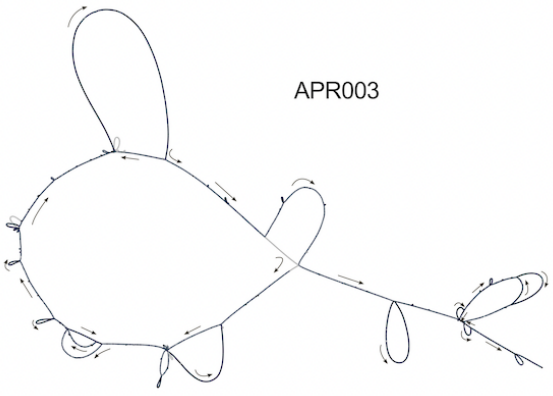
b.



c.

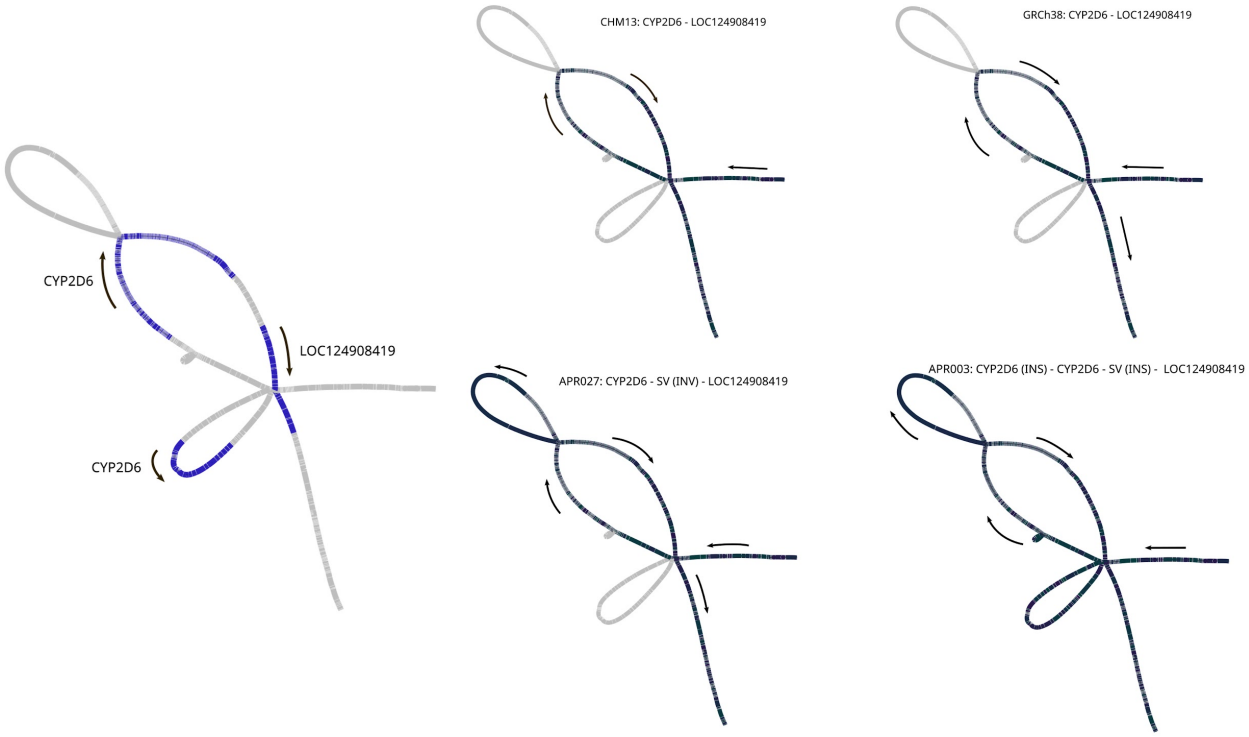


d.



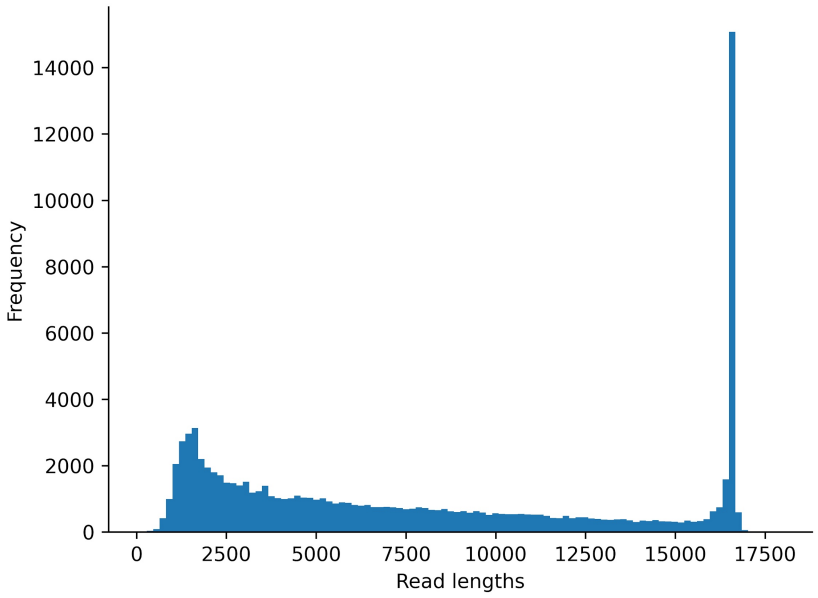
Supplementary Figure 12: Visualizing complex pangenome loci.

HLA region encompassing HLA-DRB3 gene showing highly varying paths taken by APR sample which is absent in references (HG38 and CHM13).



Supplementary Figure 13: Visualizing complex pangenome loci.

CYP2D6 region encompassing SVs showing absence of haplotypes in references, GRCh38 and CHM13. Inversions and insertions within CYP2D6 in APR samples are mentioned.



Supplementary Figure 14: Mitochondrial read distribution.

Read length (x-axis) distribution of mitochondrial reads (ONT and Hifireads), where y-axis represents read number.

Supplementary Tables

Supplementary Table 1 - Phenotypic details of APR samples.

Sample ID	Gender	Age	Height (cm)	Weight (kg)	Blood Pressure
APR001	Female	54	162	62	123/65
APR002	Female	54	146	49	139/87
APR003	Male	47	187	87	116/74
APR004	Female	51	168	95	130/86
APR005	Male	54	177	81	110/69
APR006	Male	55	-	73.5	131/90
APR007	Female	57	165	79	124/75
APR008	Female	57	162	74.5	123/75
APR009	Female	58	155	57	132/83
APR010	Male	59	176	92	122/85
APR011	Female	52	158	93	121/74
APR012	Male	53	183	78	113/68
APR013	Male	58	176	84.3	115/67
APR014	Female	52	154	72	134/84
APR015	Female	70	154	60	100/72
APR016	Female	51	162	79	109/67
APR018	Female	49	162	85.9	105/69
APR019	Female	50	161	78	102/85
APR020	Female	54	153	89	112/60
APR021	Female	60	157	72	110/75
APR022	Female	63	-	-	-
APR023	Male	54	165	69.4	115/76
APR024	Male	42	-	-	123/74
APR025	Male	29	183	73	117/75
APR026	Male	22	171	84	118/75
APR027	Female	33	164	85	100/70

APR029	Female	60	165	92	100/56
APR030	Female	61	154	80	114/69
APR031	Female	40	147	67.2	104/67
APR032	Female	48	163	82	115/67
APR033	Male	39	-	113	129/71
APR034	Female	27	160	82.9	119/75
APR035	Female	36	150	56.8	113/61
APR036	Male	52	-	104	128/71
APR037	Female	50	-	71	109/70
APR038	Male	20	177	81.5	123/67
APR039	Male	20	171	77	117/77
APR040	Female	22	-	-	115/74
APR041	Male	32	177	90	122/77

Supplementary Table 3 - Sample-wise ultra long read (>100kb) count and coverage.

Sample ID	Coverage	Total sequenced bases with greater than 100kb (bp) length
APR001	13.5235435	40570630476
APR002	17.5960556	52788166730
APR003	20.7247738	62174321299
APR004	19.3800634	58140190072
APR005	18.9823212	56946963462
APR006	10.4140496	31242148874
APR007	19.1582528	57474758357
APR008	15.2495923	45748777027
APR009	22.655163	67965489146
APR010	13.112767	39338300903
APR011	18.7574225	56272267470
APR012	13.6684289	41005286553
APR013	15.5452517	46635755104
APR014	12.9466975	38840092357
APR015	19.2821444	57846433278
APR016	14.0907649	42272294655
APR017	3.45649985	10369499554
APR018	13.7969021	41390706346
APR019	13.873482	41620445880
APR020	12.8439449	38531834769
APR021	14.8527837	44558350970
APR022	11.9297603	35789281042

APR023	22.0962993	66288897930
APR024	13.7655261	41296578258
APR025	16.751073	50253219117
APR026	18.9713638	56914091460
APR027	12.7115674	38134702289
APR029	12.3809919	37142975789
APR030	6.35730861	19071925836
APR031	11.9854624	35956387064
APR032	5.99528527	17985855812
APR033	5.63591098	16907732932
APR034	9.7725032	29317509592
APR035	10.2176266	30652879897
APR036	6.94947899	20848436972
APR037	11.627317	34881950914
APR038	6.43130418	19293912528
APR039	7.30018519	21900555565
APR040	5.7482963	17244888893
APR041	8.32129083	24963872482
APR-F	14.4466994	43340098317
APR-M	14.4447902	43334370646
APR-S	11.80888	35426639991

Supplementary Table 7 - Contiguity of Verkko and Hifiasm assemblies.

N50 (Mb)	Verkko - sample	Verkko - sample trio		Hifiasm - sample		Hifiasm - sample trio	
		hap1	hap2	hap1	hap2	hap1	hap2
APR-S	9.16	142.81	149.22	-	-	135.26	107.84
APR-M	3.02	-	-	111.54	97.72	-	-
APR-F	9.52	-	-	102.31	88.08	-	-

Supplementary Table 8 - Mitochondrial contigs obtained for each sample by mapping to chrM.

Sample ID	Haplotype	Contig ID	Aligned length	Contig length	Percentage alignment
APR-F	1	h1tg000180l	16135	17071	0.945170172
APR-F	2	h2tg000123c	9751	16570	0.588473144
APR-M	2	h2tg000085l	16568	32061	0.516764917
APR-M	1	h1tg000107l	16568	32061	0.516764917
APR-S	2	h2tg000079c	16344	16570	0.986360893
APR-S	1	h1tg000111c	16344	16570	0.986360893
APR001	1	h1tg000072c	16569	33138	0.5
APR001	2	h2tg000064c	16569	33138	0.5
APR002	1	h1tg000068l	16569	98424	0.168343087
APR002	2	h2tg000062l	16569	81858	0.202411493
APR003	1	h1tg000080l	16566	32490	0.509879963
APR003	2	h2tg000067l	16566	32490	0.509879963
APR004	1	h1tg000057l	16569	46115	0.359297409

APR004	2	h2tg000069l	16569	31650	0.523507109
APR005	1	h1tg000042c	16569	49710	0.333313217
APR005	2	h2tg000079c	15985	16570	0.964695232
APR006	1	h1tg000088l	16569	49502	0.334713749
APR006	2	h2tg000086l	16569	49502	0.334713749
APR007	2	h2tg000056l	16569	32867	0.504122676
APR007	1	h1tg000050l	16569	32867	0.504122676
APR008	1	h1tg000060l	16569	49209	0.3367067
APR008	2	h2tg000069l	16569	49209	0.3367067
APR009	1	h1tg000054l	16569	32938	0.503036007
APR009	2	h2tg000064l	16569	32938	0.503036007
APR010	2	h2tg000094l	16569	33094	0.500664773
APR010	1	h1tg000115l	16569	33094	0.500664773
APR011	1	h1tg000063l	16569	32939	0.503020735
APR011	2	h2tg000058l	16569	32939	0.503020735
APR012	1	h1tg000060l	16566	32299	0.512895136
APR012	2	h2tg000064l	16566	32299	0.512895136
APR013	2	h2tg000079c	15468	16568	0.933606953
APR013	1	h1tg000074c	15468	16568	0.933606953
APR014	1	h1tg000051c	9969	16569	0.601665761
APR014	2	h2tg000046c	9969	16569	0.601665761
APR015	2	h2tg000065c	9723	16578	0.586500181
APR015	1	h1tg000072c	9723	16578	0.586500181
APR016	2	h2tg000041c	10338	16575	0.623710407

APR016	1	h1tg000036c	10338	16575	0.623710407
APR017	2	h2tg000141l	16558	40011	0.413836195
APR017	1	h1tg000125l	16558	40011	0.413836195
APR018	2	h2tg000058l	16569	32879	0.503938684
APR018	1	h1tg000055l	16569	32879	0.503938684
APR019	2	h2tg000071c	12783	16569	0.771500996
APR019	1	h1tg000067c	12783	16569	0.771500996
APR020	1	h1tg000087l	16569	60428	0.274194082
APR020	2	h2tg000089l	16569	45635	0.363076586
APR021	1	h1tg000082c	14412	16568	0.869869628
APR021	2	h2tg000087c	14412	16568	0.869869628
APR022	1	h1tg000074l	16569	32955	0.502776513
APR022	2	h2tg000091l	16569	32955	0.502776513
APR023	2	h2tg000057l	16567	49436	0.335120155
APR023	1	h1tg000053l	16567	49436	0.335120155
APR024	2	h2tg000047l	16569	38699	0.428150598
APR024	1	h1tg000052l	16569	38699	0.428150598
APR025	2	h2tg000092l	16567	33044	0.501361821
APR025	1	h1tg000087l	16567	33044	0.501361821
APR026	1	h1tg000066c	15910	16571	0.960111037
APR026	2	h2tg000074c	15910	16571	0.960111037
APR027	1	h1tg000067c	16569	33144	0.499909486
APR029	2	h2tg000089l	16569	49485	0.334828736
APR029	1	h1tg000090l	16569	49485	0.334828736

APR030	1	h1tg000059c	16569	49707	0.333333333
APR030	2	h2tg000071c	12163	16569	0.734081719
APR031	2	h2tg000088l	16569	32866	0.504138015
APR031	1	h1tg000075l	16569	32866	0.504138015
APR032	1	h1tg000059l	16569	49192	0.336823061
APR032	2	h2tg000056l	16569	49192	0.336823061
APR033	2	h2tg000081c	16569	33138	0.5
APR033	1	h1tg000074c	16569	33138	0.5
APR034	2	h2tg000113l	16569	32457	0.510490803
APR034	1	h1tg000094l	16569	32457	0.510490803
APR035	1	h1tg000059l	16567	44087	0.375779708
APR035	2	h2tg000067l	16567	44087	0.375779708
APR036	2	h2tg000229l	16558	30953	0.53494007
APR036	1	h1tg000236l	16558	30953	0.53494007
APR037	2	h2tg000100l	16569	40629	0.407812154
APR037	1	h1tg000069l	16569	40629	0.407812154
APR038	1	h1tg000246l	16564	32794	0.505092395
APR038	2	h2tg000249l	16564	32794	0.505092395
APR039	2	h2tg000123l	16569	32587	0.508454292
APR039	1	h1tg000143l	16569	32587	0.508454292
APR040	1	h1tg000123l	16569	32818	0.504875373
APR040	2	h2tg000131l	16569	32818	0.504875373
APR041	2	h2tg000373c	16569	33140	0.499969825
APR041	1	h1tg000400c	16569	33140	0.499969825

Supplementary Table 15: Pan ethnic disease autosomal recessive of X-linked genes within APR specific duplicated genes.

Gene	Career Screening
CASQ2	Catecholaminergic polymorphic ventricular tachycardia (CASQ2-related); Autosomal Recessive; CASQ2
DNAI1	Primary ciliary dyskinesia (DNAI1-related); Autosomal Recessive; DNAI1
EXOSC3	Pontocerebellar hypoplasia type 1B; Autosomal Recessive; EXOSC3
FANCG	Fanconi anemia type G; Autosomal Recessive; FANCG
GALT	Galactosemia (GALT-related); Autosomal Recessive; GALT
GAMT	Guanidinoacetate methyltransferase deficiency; Autosomal Recessive; GAMT
GRHPR	Primary hyperoxaluria type 2; Autosomal Recessive; GRHPR
HSD3B2	Congenital adrenal hyperplasia due to 3-beta- hydroxysteroid dehydrogenase deficiency; Autosomal Recessive; HSD3B2
LAMA3	LAMA3-related conditions; Autosomal Recessive; LAMA3
LPL	Familial chylomicronemia syndrome; Autosomal Recessive; LPL
NDUFS7	Mitochondrial complex I deficiency 3; Autosomal Recessive; NDUFS7
NPC1	Niemann-Pick disease type C (NPC1-related); Autosomal Recessive; NPC1
OCA2	Oculocutaneous albinism type 2; Autosomal Recessive; OCA2
PDHA1	Pyruvate dehydrogenase complex deficiency (PDHA1-related); X-Linked; PDHA1
PHGDH	Phosphoglycerate dehydrogenase deficiency; Autosomal Recessive; PHGDH
PHKG2	Glycogen storage disease type IXc; Autosomal Recessive; PHKG2
SLC25A15	Hyperornithinemia-hyperammonemia-homocitrullinuria syndrome; Autosomal Recessive; SLC25A15
ST3GAL5	GM3 synthase deficiency; Autosomal Recessive; ST3GAL5

Supplementary Table 17. Comparison on size of APR HPRC graph reference and HPRC, CPC and APR graph reference.

Genome	No. of Nodes	No. of Edges	Complexity	Length (bp)
APR	72,256,745	99,652,737	1.3761	3,307,861,124
HPRC	92,019,585	126,863,170	1.3760	3,328,787,872
CPC	64,474,746	89,583,031	1.389	3,284,609,818

*Complexity=Ratio of Edges to Nodes

Supplementary Table 23. Repeat elements within APR specific novel sequences.

Repeat Elements	Number of Elements	Length	Percentage
ALUs (SINE)	21886	4958828	4.91
MIRs (SINE)	3416	474855	0.47
LINE1	8633	5465252	5.41
LINE2	2474	576783	0.57
L3/CR1 (LINE)	206	50497	0.05
ERVL (LTR)	1513	687192	0.68
ERVL-MaLRs (LTR)	3211	1431290	1.42
ERV_classI (LTR)	2279	1930489	1.91
ERV_classII (LTR)	237	323743	0.32
hAT-Charlie (DNA)	2116	438494	0.43
TcMar-Tigger (DNA)	845	242894	0.24
Unclassified	21979	5509762	5.46
Interspersed		22329369	22.12
Small RNA	612	59617	0.06
Satellites:	7995	32022812	31.73
Simple Repeats	20407	22529584	22.32
Low Complexity	1849	494198	0.49

Supplementary Table 28 - Structural haplotypes present in APR, HPRC+CPC

Region	Haplotype (Genic)	APR Count	APR Sample ID	HPRC-CPC Count
chr1	PRAMEF19(1)-PRAMEF14(1)-PRAMEF15(1)-PRAMEF8(1)-PRAMEF33(1)	38	APR-F APR001 APR002 APR003 APR004 APR005 APR006 APR007 APR008 APR010 APR011 APR012 APR013 APR014 APR015 APR016 APR017 APR018 APR019 APR020 APR022 APR024 APR026 APR027 APR029 APR030 APR032 APR033 APR034 APR035 APR036 APR037 APR038 APR039	105

			APR040 APR041 APR-M APR-S	
	PRAMEF19(1)-PRAMEF14(1)-PRAMEF15(2)-PRAMEF8(1)-PRAMEF33(1)	4	APR009 APR023 APR025 APR031	0
	PRAMEF19(1)-PRAMEF14(2)-PRAMEF15(2)-PRAMEF8(1)-PRAMEF33(1)	1	APR021	0
chrX	CT45A3-CT45A7-CT45A2-CT45A9	31	APR-F APR002 APR004 APR005 APR008 APR009 APR010 APR011 APR012 APR013 APR016 APR017 APR018 APR020 APR021 APR022 APR023 APR026 APR027 APR029 APR030 APR032 APR034 APR035 APR036 APR037 APR038 APR039	0

			APR040 APR-M APR-S	
	CT45A3-CT45A7-CT45A2	12	APR001 APR003 APR006 APR007 APR014 APR015 APR019 APR024 APR025 APR031 APR033 APR041	0
	CT45A2-CT45A9-CT45A9	0		91
	CT45A9-CT45A9	0		9
	CT45A9	0		4

References

1. 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).
2. Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D. & Gurevich, A. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* **34**, i142–i150 (2018).
3. Almarri, M. A. *et al.* The genomic history of the Middle East. *Cell* **184**, 4612-4625.e14 (2021).
4. Byrska-Bishop, M. *et al.* High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426-3440.e19 (2022).