

Near-complete Middle Eastern genomes refine autozygosity and enhance disease-causing and population-specific variant discovery

Received: 22 April 2024

Accepted: 18 March 2025

Published online: 5 May 2025

 Check for updates

Mohammadmersad Ghorbani¹, Shabir Moosa¹, Zenab Siddig¹, Radi Farhad¹, Haroon Naeem¹, William T. Harvey², Francesco Kumara Mastroso², Katherine M. Munson², Rozaimi Mohamad Razali³, Elbay Aliyev¹, Ilhame Diboun¹, Rawan Abouelhassan¹, Melissa Tauro¹, Sondoss Hassan¹, Rebecca Mathew¹, Muna Al Hashmi¹, Lisa S. Mathew¹, Kun Wang¹, Abdul Rahman Salhab¹, Fazulur Rehaman Vempalli¹, Ahmed El Khouly¹, Qatar Genome Program Research Consortium*, Iman Alazwani¹, Sara Tomei¹, Khalid A. Fakhro^{1,4,5}, Alia Satti¹, Ruba Benini^{1,4}, Arang Rhie⁶, Evan E. Eichler^{2,7} & Younes Mokrab^{1,3,4} ✉

Advances in long-read sequencing have enabled routine complete assembly of human genomes, but much remains to be done to represent broader populations and show impact on disease-gene discovery. Here, we report highly accurate, near-complete and phased genomes from six Middle Eastern (ME) family trios ($n = 18$) with neurodevelopmental conditions, representing ancestries from Sudan, Jordan, Syria, Qatar and Afghanistan. These genomes revealed 42.2 Mb of new sequence (13.8% impacting known genes), 75 new HLA/KIR alleles and strong signals of inbreeding, with ROH covering up to one-third of chromosomes 6 and 12 in one individual. Using assembly-based variant calling, we identified 23 de novo and recessive variants as strong candidates for causing previously unresolved symptoms in the probands. The ME genomes revealed unique variation relative to existing references, showing enhanced mappability and variant calling. These results underscore the value of de novo assembly for disease variant discovery and the need for sampled ME-specific references to better characterize population-relevant variation.

Advances in long-read sequencing and bioinformatics have enabled the generation of high-quality genome assemblies from various ancestries, ranging from Korea¹ to Puerto Rico^{2,3}. Recently, over a hundred near-complete and fully phased genomes have been produced by the Human Pangenome Research Consortium (HPRC)⁴ and the Chinese Pangenome Consortium (CPC)⁵, in addition to complete telomere-to-telomere (T2T) genomes for a haploid genome of primarily European ancestry (CHM13)⁶ and a diploid Han Chinese male

(CHN1)⁷. This has allowed major gaps in GRCh38 to be filled, especially in complex and repetitive regions^{6,8}, capturing previously inaccessible variation⁹ and uncovering genome mechanistic insights^{9,10}. Nevertheless, complete genome assembly is yet to be leveraged in disease-gene discovery, especially in under-represented populations where there is a particular need for understanding haplotype diversity and structure, which is important in assessing variant pathogenicity.

A full list of affiliations appears at the end of the paper. ✉ e-mail: ymokrab@sidra.org

The Middle Eastern (ME) populations lie at a historical intersection of human migration and civilization^{11–14}. They are generally characterized by isolated structures and high consanguinity, leading to increased incidence of genetic conditions impacting normal development^{15–17}, often due to founder mutations. Recent studies have identified peculiar ancestral groups such as Peninsular Arabs who were found to be the closest to the ancient population that migrated out of Africa^{11,14} and harbor a differential mutational spectrum of Mendelian disorders and cancer predisposition^{18,19}. Currently, ME populations are heavily under-represented in genomic literature, lacking high-quality assemblies^{20,21} needed to shed light on how inbreeding shapes genome structure and disease architecture^{20,21}.

In this study, we use a de novo genome assembly approach to characterize six family trios from diverse ME ancestries (Sudan, Jordan, Syria, Qatar and Afghanistan), involving probands with various unresolved neurodevelopmental conditions whereby we (1) assemble near-complete genomes and map their high-resolution features, (2) apply assembly approach to identify disease-causing variants and (3) assess the value of ME-specific genetic reference in improving read mapping and variant calling (Extended Data Fig. 1). We assemble near-complete phased genomes which we interrogate for various structural features, comparing against the latest genome references. Using trio analyses, we identify previously undetected deleterious variants. Finally, we demonstrate the value of population-specific references, emphasizing the need for a carefully sampled pangenome for ME ancestries.

Results

Study samples and sequencing

Our study samples consist of six parent–child trios ($n = 18$) from five different nationalities spanning the wide ME region: two trios from Qatar (Qatari 1 and Qatari 2) and single trios from Sudan, Jordan, Syria and Afghanistan (Fig. 1a). The children, male (XY) except for the Afghani female (XX), are aged 3–15 years and were ascertained with multiple symptoms of developmental disorders (Extended Data Fig. 2). Three of these families (Sudanese, Jordanian, Qatari 2) have undergone genetic testing previously based on microarrays and selected panels, returning negative results. Subsequent analysis using whole-genome sequencing from Illumina (Methods) did not identify plausible pathogenic variants except for Qatari 1, where we found a likely pathogenic variant (a de novo SNV in splice donor site) in the transcription factor nuclear factor IX (*NFIX*) reported to cause Malan syndrome²². This finding explained some of this patient's symptoms, namely global developmental delay and tall stature.

Given the lack of diagnosis of these diverse ME families, we used the latest technologies of long-read sequencing and de novo assembly methods to gain insight into high-resolution genomic structure and disease-causing genetic variants. For each participant, we generated high-quality PacBio HiFi long-read and Illumina short-read data and subsequently processed them using established pipelines (Methods). For the long-read data, the average read length is 19 kb (range = 15.1–21.3 kb), while the average coverage is 38.8 \times for the children and 37.2 \times for the parents (Supplementary Fig. 1). For the short-read data, the average coverage is 31–38 \times (Supplementary Fig. 2).

Genetic ancestry

To explore the detailed genetic ancestry of the ME trios, we conducted admixture inference of global and local ancestry based on a diverse population dataset as a reference panel based on 1000 Genomes (IKG) and Qatar Genome Program (QGP) phase 1 cohort (Methods). As expected, CHM13 is predominantly of European ancestry (Fig. 1b). The Sudanese trio has >99% African ancestry, while the Afghani, Syrian, Jordanian, Qatari 1 and Qatari 2 participants in this order showed a prominent trend of increasing General Arab (GAR) and Peninsular Arab (PAR) ancestries and decreasing European ancestry. The highest

proportion of GAR was shown in Qatari 2 participants (63%), the highest PAR was in Qatari 1 and Qatari 2 (15% and 16%, respectively), while the highest West Eurasian and Persian (WEP) ancestry was found in the Syrian, Jordanian and Qatari 1 (20–22%; Fig. 1b). The highest East Asian and South Asian ancestries were found in the Afghani participants (8% and 12% respectively). Consistent with this, an overlay of the dominant global ancestries on principal components showed that the family trios co-clustered with samples of similar ancestries, namely GAR, PAR, WEP¹⁴ (for Syrian, Jordanian, Qatari 1 and Qatari 2), Africans (for Sudanese), while the Afghanis are found between Europeans, West Eurasians and Asians (Fig. 1c). Next, we conducted local ancestry inference to assign ancestry components to individual chromosomal segments (Fig. 1d and Supplementary Fig. 3). We could see relatively high ancestral homogeneity for the Sudanese participants as well as CHM13, while more admixture is seen for the remaining participants. The average ancestry segment length across the samples is 570 kb. Notably, the longest non-African stretches were found in Qatari 2 (PAR), Syrian (GAR) and Qatari 1 (WEP).

Genome assembly and phasing

We generated de novo phased assemblies for the six ME trios using trio-Hifiasm²³, incorporating PacBio HiFi long reads and parental Illumina short reads, as per the best recommended workflow for phased genome assembly²³. Standard quality check (QC) metrics were calculated to evaluate assembly contiguity and accuracy (Fig. 2a, Supplementary Fig. 4 and Supplementary Table 1). The obtained assemblies were of high quality as indicated by the various metrics, the ranges and averages of which are as follows: coverage (30.4–49.1 \times , 37.7 \times), contig N50 (52.9–98.2 Mb, 77.3 Mb), contig count (146–372, 263.4), longest contig (133.5–232.6 Mb, 169.2 Mb) and per-base accuracy quality value (QV; 44.0–59.3, 53.7). Of note, the children were better than the parents for most metrics, reflecting the higher mean coverage (38.8 versus 34.9), higher average HiFi read length (19.9 kb versus 18.8 kb) and the complementary use of parental data for assembling and phasing. The assembly with the best contig N50 was for Qatari 1 child (94.6 Mb for haplotype 1 and 97.4 Mb for haplotype 2) which also had the best QV (59.3 for haplotype 1 and 58.8 for haplotype 2), the best coverage (49.1 \times) and second highest HiFi read length average (21.0 kb). The longest contig was 232.6 Mb, belonging to the Jordanian child participant.

Using the parental short reads, we evaluated the phasing accuracy for the child assemblies using haplotype-specific markers (hapmers)²⁴, whereby contigs are plotted as a function of hapmers assigned from maternal and paternal libraries (Fig. 2b and Supplementary Fig. 5). All assemblies show a clear segregation of the parental haplotypes indicating the high quality of phasing. Furthermore, N* plots indicate low levels of haplotype switch errors (Fig. 2c and Supplementary Fig. 6). In addition, inherited hapmer plots show balanced distributions of k -mers private to children, fathers, mothers and those shared between the parents (Supplementary Fig. 7). Furthermore, we performed an independent evaluation of the accuracy of the assemblies using the Flagger pipeline which assesses coverage and variants from the alignment of HiFi reads to the respective assemblies⁴. We found that >99% of the assemblies' sequences consisted of reliable (haploid) blocks (Fig. 2d).

Comparison with T2T assemblies and new sequences

Alignment of the child assemblies to CHM13 using Minimap2 (ref. 25) showed high mappability across various chromosomes, with an average of 91% of contigs with high alignment coverage mapping to a unique location (Fig. 3a and Supplementary Fig. 8). Using CHM13 as reference, the ratio of completeness per chromosome was 93% on average, with 43 chromosomes from all six participants having >99% completeness including 14 T2T chromosomes, which we verified confirming the presence of telomeric repeats (Fig. 3b). Qatari 2 shows the highest overall completeness of 92%. Notably, 14 chromosomes were spanned by a single contig covering the centromeric region, including

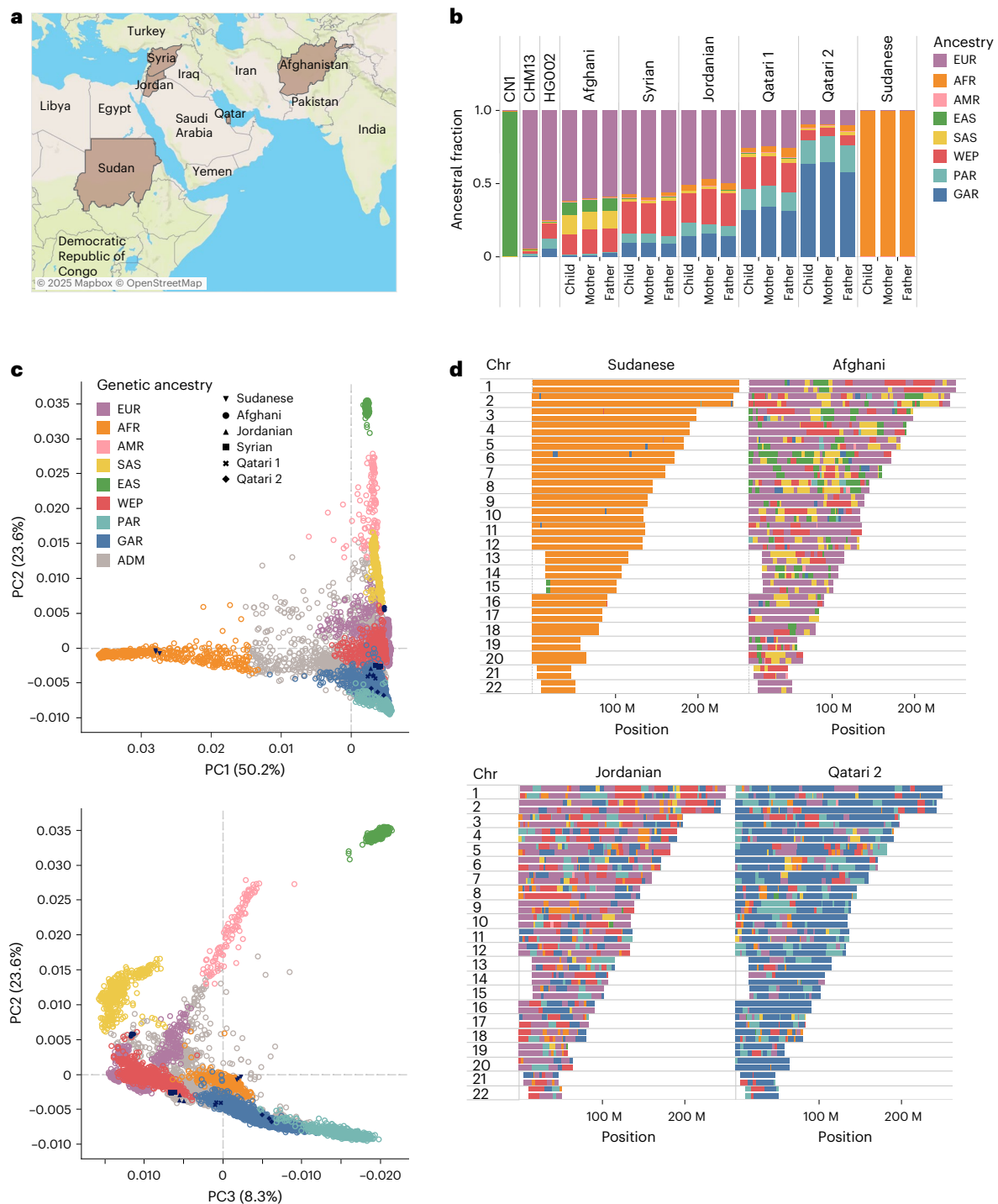


Fig. 1 | Genetic ancestry of study samples. **a**, Geographic location of the study cohort. **b**, Global ancestral composition of the individual participants alongside that of CN1, CHM13 and HG002. **c**, Principal component analysis showing the study samples and reference dataset from QGP and 1KG. **d**, Local ancestry analysis showing the genetic ancestral makeup of each chromosome for the

Sudanese, Afghani, Jordanian and Qatari 2 child participants. The displayed map is from Mapbox and OpenStreetMap, used under the ODbL. ADM, admixed; AFR, Africans; AMR, American; EAS, East Asian; EUR, Europeans; SAS, South Asian; ODbL, Open Database License.

a T2T single contig for chromosome 10 from a Qatari 2 child (Fig. 3c and Supplementary Fig. 9).

We checked the sequences of insertions identified from the alignments of the assemblies to CHM13 against a diverse set of highly curated calls of structural variations from HPRC⁸, HG002 (Ashkenazi Jews) and CN1 (Han Chinese)⁷, from which we excluded any regions identified with Flagger to be unreliable. This led to the identification of 42.2 Mb of new sequences across all the samples, ranging from 6.4 Mb for the Sudanese

to 10.6 Mb for Qatari 1 (Fig. 3d). Of these sequences, 10–40% were in the centromeres, highlighting the relatively higher divergence and mutation rate in these regions²⁶. By size, 86.2% of the new sequences are found in intergenic regions, while 12.6% are in intronic or untranslated regions (intronic/UTR) and 1.2% in exonic regions (Supplementary Fig. 10a). Also, 89.7% of the new sequences are found in repetitive regions, notably satellite (58.1%) and simple repeats (24.1%; Fig. 3d, middle). New sequences outside repetitive regions are enriched for

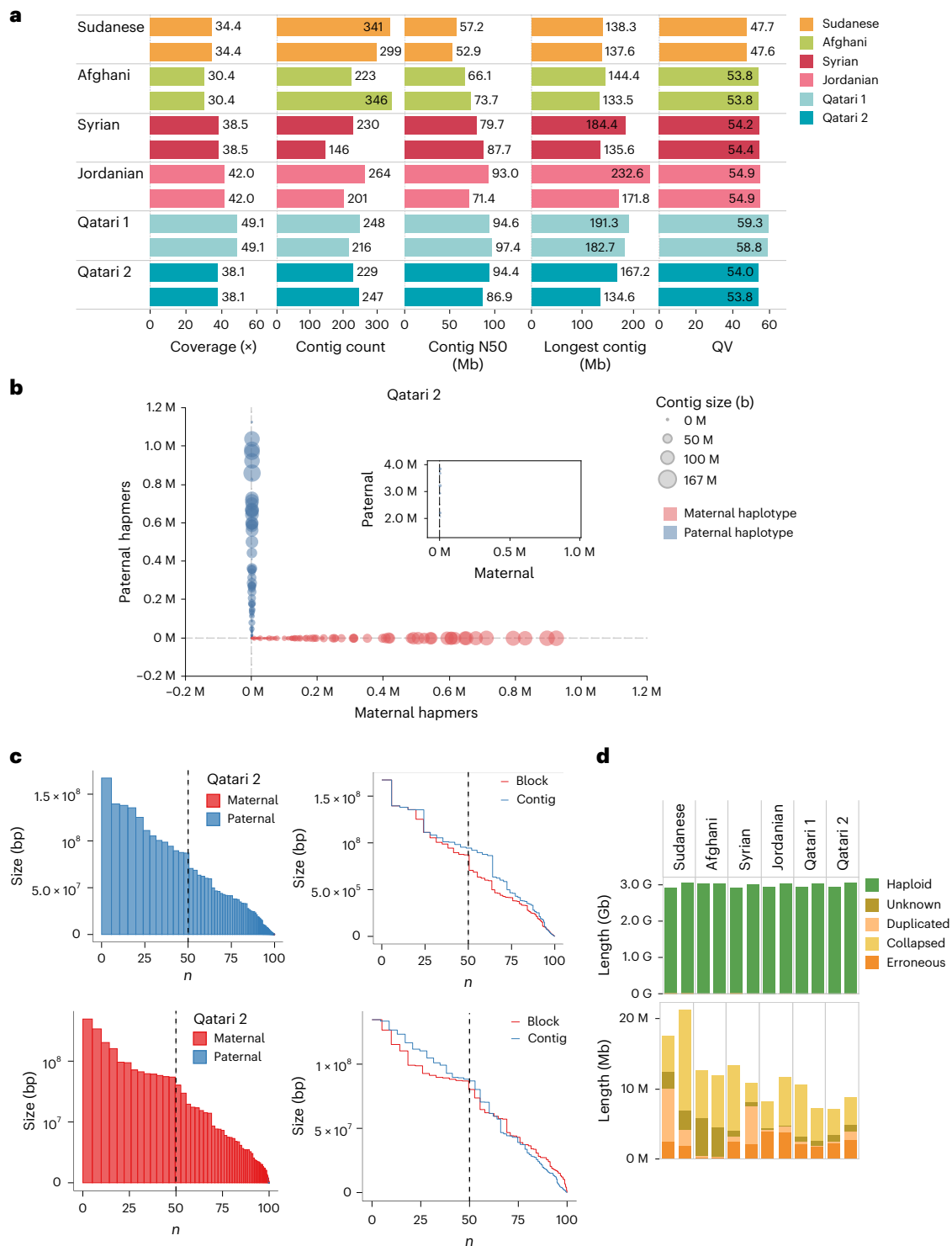


Fig. 2 | Assembly, phasing results and QC. a, QC metrics of each child assembly showing coverage, depth, total contigs, contig N50, maximum contig length and QV. **b**, Hapmer blob plot of Qatari 2 child sample showing a clear separation of maternal (red) and paternal (blue) haplotypes. Blob size is proportional to contig size, and each blob/contig is plotted according to the count of parental hapmers. **c**, Phase block NG plots (left) of haplotype-resolved assembly for paternal

(top) and maternal (bottom) contigs sorted by size. The x axis represents the percentage of genome size and y axis represents the block size. Incorrectly phased haplotype blocks are virtually absent. Phase block NG and contig NG plots (right) showing the phase block sizes being similar to contig sizes due to low switch error. **d**, Reliability of the assemblies using read mapping evaluation with Flagger.

exonic and intronic/UTR (Fig. 3d, right). We also noticed most new sequences (79.3%) occur outside segmental duplications (Supplementary Fig. 10b). For those inside segmental duplications, nearly a quarter overlap with intronic/UTR and 3.9% overlap with exonic regions (Supplementary Fig. 10b).

Consistent with the closer genetic distance to HG002 (ref. 27), the sequences from our assemblies, which were also absent from CHM13 and HPRC, have more shared sequences with HG002 than CN1 (Supplementary Fig. 10c). The new sequences are well spread across the chromosomes (Supplementary Fig. 10d,e).

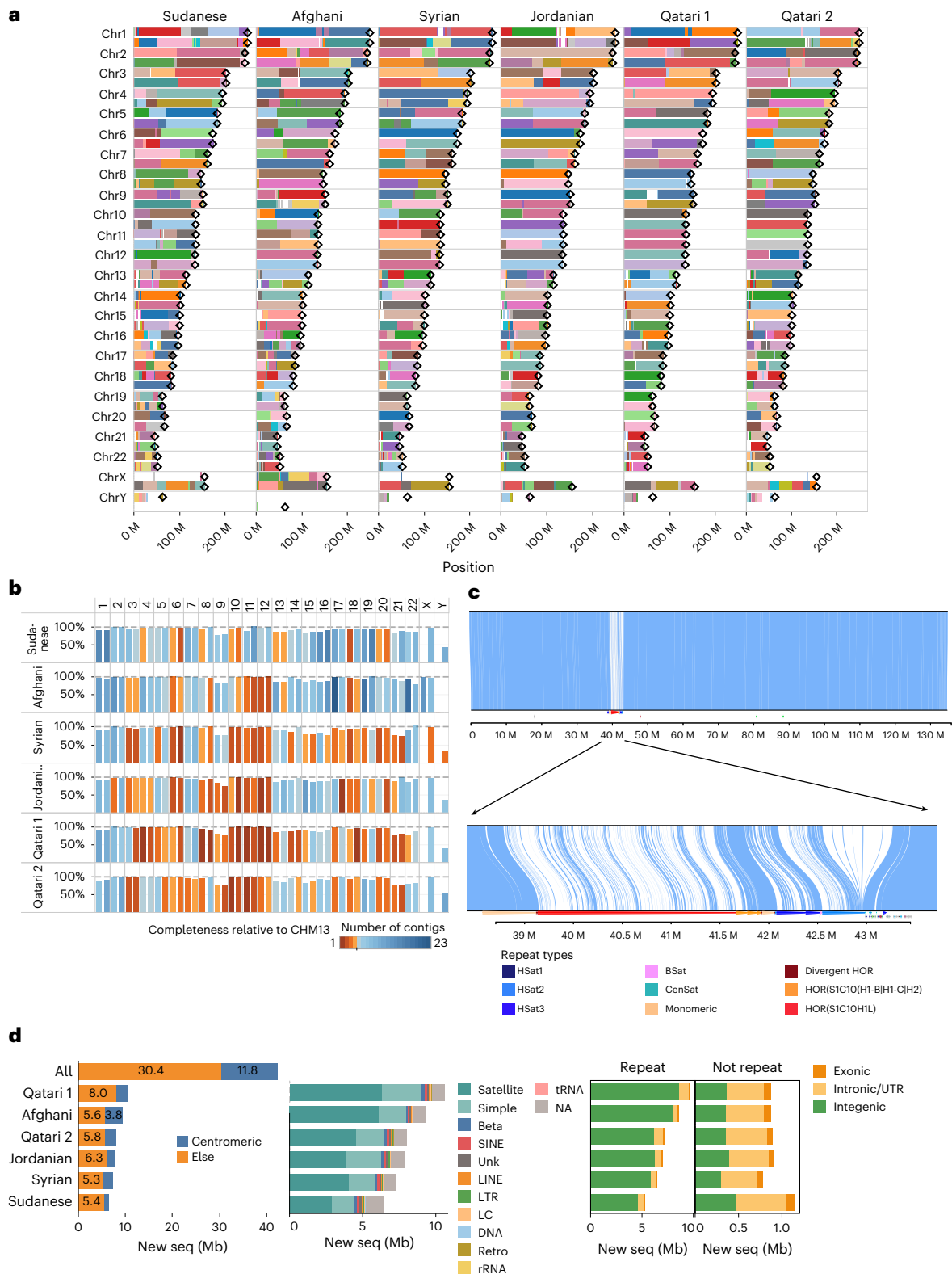


Fig. 3 | Contiguity, haplotype alignment to CHM13 and new sequences. **a**, Alignment of child assemblies to CHM13 for individual chromosomes. Individual contigs are delineated with distinct colors per chromosomal haplotype. Diamond symbols denote the end points of CHM13 haplotypes. **b**, Percentage of completeness (y axis) relative to CHM13 per chromosome (x axis), colored by the number of contigs in the alignment. **c**, Alignment of chromosome 10 of Qatari 2 (top) child assembly to CHM13 (bottom) showing

a singular contig spanning the entire chromosome with notable centromeric variation. **d**, Length of new sequences identified across samples, highlighting location in centromeric regions (left), repetitive regions (middle) and intergenic, intronic/UTR and exonic regions either inside or outside repetitive regions (right). LC, low complexity; LTR, long terminal repeats; SINE, short interspersed nuclear element; Unk, unknown.

Gene coverage

We annotated the assemblies with genes of various types based on the publicly available CHM13 annotation. The distribution of genes identified across the chromosomes was found to be proportionate and similar to CHM13 (Supplementary Fig. 11a). Stratifying per-gene category, the number of genes across the assemblies is relatively uniform and similar to CHM13 for all established 27 gene categories, except for rRNA genes known to vary in number of copies between individuals²⁸ (Fig. 4a). Specifically, the assemblies show 97–100% coverage for the six largest gene categories in the genome namely protein coding, lncRNA, pseudogenes, miRNA, transcribed pseudogenes and snoRNA (Fig. 4b). The number of genes observed for sex chromosome was consistent with the sample sex (Supplementary Fig. 11b,c). Per participant, Qatari 1 and 2 haplotypes demonstrate exceptionally high coverage, exceeding 99% for protein-coding genes.

HLA gene family and KIR gene family annotations

We annotated HLA and KIR genes in the assemblies using a recent whole-genome-assembly-based method and compared the annotated genes against a set of 220 pre-annotated assembly haplotypes, including HLA assemblies, HPRC, CPC, CHM13, CN-T2T and GRCh38 (ref. 29). The ME child assemblies had alleles in 47 genes, all of which had one copy except *C4B* which had two copies (Supplementary Fig. 12). The genes harbored 233 alleles, including previously reported alleles in the *HLA-A* locus, which are consistently found to be largely common or well documented in populations of ME ancestry³⁰ (Supplementary Fig. 13). Furthermore, amongst the 47 genes, we found 33 genes that contained 75 new alleles (that is absent from the Immuno Polymorphism Database (IPD)) with intact coding DNA sequence (CDS) region (Extended Data Fig. 3a). These alleles, of which there are 17–30 per assembly (Extended Data Fig. 1b), are found in regions of reliable assembly as determined using Flagger. They include 11 unique alleles, each having one to seven mutations in CDS regions (Fig. 4c and Supplementary Fig. 14a). In terms of impact on protein sequence, ten are new missense, five are known missense and ten are synonymous (Supplementary Fig. 14b,c). The DNA and protein sequence alignments highlighting these mutations are shown in Supplementary Fig. 15. Phylogenetic trees show the known alleles in the study participants are spread over multiple diverged lineages while some of the new alleles are found to be closely clustered, pointing to potential common founder lineages (Fig. 4d and Supplementary Fig. 16).

Genetic variation

To gain insight into the genetic variation in the ME assemblies, we compared them against the widely used genome references GRCh38 and CHM13, whereby we performed assembly-based alignment and called single-nucleotide variants (SNVs), short insertions and deletions (indels) and various types of structural variants (SVs). Assembly-based alignment relies on mapping assembled contigs against a given reference instead of individual reads in classical read-based alignment, and was shown previously to lead to better accuracy³¹. Consistent with this, using this approach, we observe more accurate calls than using read-based alignments (Supplementary Fig. 17). Overall, there were an average of 6.6% more variants against CHM13 than GRCh38, the largest being shown for Sudanese (10.3%; Fig. 5a). This is most prominent for SV deletions which are 61.5–87.2% higher. The exception to this trend is for inversions, whereby, on average, they are 43.2% less against CHM13 than GRCh38. Comparing the assemblies, the Sudanese has 18.6% more SNVs and indels on average than those of other participants, whereas SV counts are more uniform (Supplementary Fig. 18). We found that 1,177–1,568 deletions and 1,325–1,641 insertions across the six families were entirely absent from the HPRC dataset. Also, more than half of them (51–56%) fall in repetitive regions, notably simple repeats, while 19–22% overlap with segmental duplications (Fig. 5b).

Next, we overlaid the generic variants on local ancestry segments. The aggregated counts per ancestry were reflective of the

global ancestry of individual participants, rather than showing specific enrichment (Supplementary Fig. 19). When we plotted the number of variants per 1 Mb of ancestral segment as a function of chromosomal coordinate, we saw that the counts across segments fluctuated between ancestries, as illustrated for the near-complete chromosome 10 (Supplementary Fig. 20a). When we aggregated that, we observe 12–30% lower variant counts relative to African segments (Fig. 5c), in line with population distances (Supplementary Fig. 20b).

Runs of homozygosity (ROH)

In light of the background consanguinity known to characterize ME populations, we explored ROH in the assemblies, which we calculated from the assembly-based alignments to CHM13. We classified ROH to medium (1–3 Mb) and long (≥ 3 Mb), which respectively reflect distant and recent inbreeding^{32,33}. ROH segments are found genome-wide up to nine per chromosome, with variable cumulative ROH per chromosome (Extended Data Fig. 4). Cumulatively, the ME participants had an average 43.8 Mb medium ROH and 69.1 Mb long ROH, which is fivefold greater than a reference Yoruban trio (Fig. 5d). The father in the Jordanian trio exhibits the largest cumulative ROH (212.3 Mb), of which 174.6 Mb is of the long type which is remarkable given that he has not presented any clinical phenotypes. There is a high correlation between ROH segment size and gene content (Pearson correlation coefficient, $R = 0.8$, $P < 10^{-4}$; Supplementary Fig. 22). Various types of genes are found in these ROH segments, including protein coding, pseudogenes and lncRNA without any category being particularly enriched (Supplementary Fig. 21c). Notably, the Jordanian father, who is a healthy participant, harbors exceptionally long ROH segments on chromosomes 6 (52.9 Mb) and 12 (37.6 Mb), the longest amongst all participants, spanning 30% and 28% of the chromosomes and covering 794 and 620 genes, respectively (Fig. 5e). We confirmed these long ROHs by examining the alignments of HiFi reads against CHM13 which showed homozygous genotypes throughout the region and introduction of heterozygous sites at the ends (Extended Data Fig. 5). We note that of the 537 ROH segments identified in the study participants, 72% (386) encompass at least one protein-coding gene with high probability of loss-intolerance (pLI) score > 0.9 (ref. 34; Fig. 5f).

Genetic variants underlying disease phenotypes

We performed a trio-based analysis to search for disease-causing variants that explain the disease phenotypes in the child participants (Methods). We examined the SNVs, indels and SV variants obtained for families against CHM13. Given that only the children are affected, we considered recessive and de novo modes of inheritance. We prioritized variants that were rare in population databases (allele frequency, $AF < 0.01$), including gnomAD³⁵, IKG³⁶, HPRC⁸ and QGP phase 1 cohort^{14,37}, have high predicted functional scores and whereby the proband conditions are consistent with gene phenotypes reported in ClinGen³⁸, Gene2Phenotype (G2P) panel on Developmental Disorders³⁹ and DECIPHER⁴⁰.

Overall, we shortlisted 23 candidate variants relevant to the various phenotypes across the six probands, comprising 16 deletions, 6 insertions and 1 SNV, each participant having between 1 and 8 variants (Fig. 5g and Supplementary Fig. 23). In terms of inheritance mode, 12 are recessive while 11 are de novo. Regarding their location, four are shown to impact exons, whereas the rest impact intronic regions and overlapping regulatory features, therefore providing strong candidates as disease-causing (Supplementary Table 2). Based on American College of Medical Genetics and Genomics (ACMG) guidelines, two of the variants are classified as pathogenic and one as likely pathogenic, found in the well-established genes of *NFIX* (associated with Malan Overgrowth Syndrome) and *TSC2* (associated with Tuberous Sclerosis Complex), in addition to *ABO* which is less known to be implicated in neurodevelopment. Furthermore, 20 variants were classified as variant of uncertain significance (VUS) in genes reported to have strong evidence of

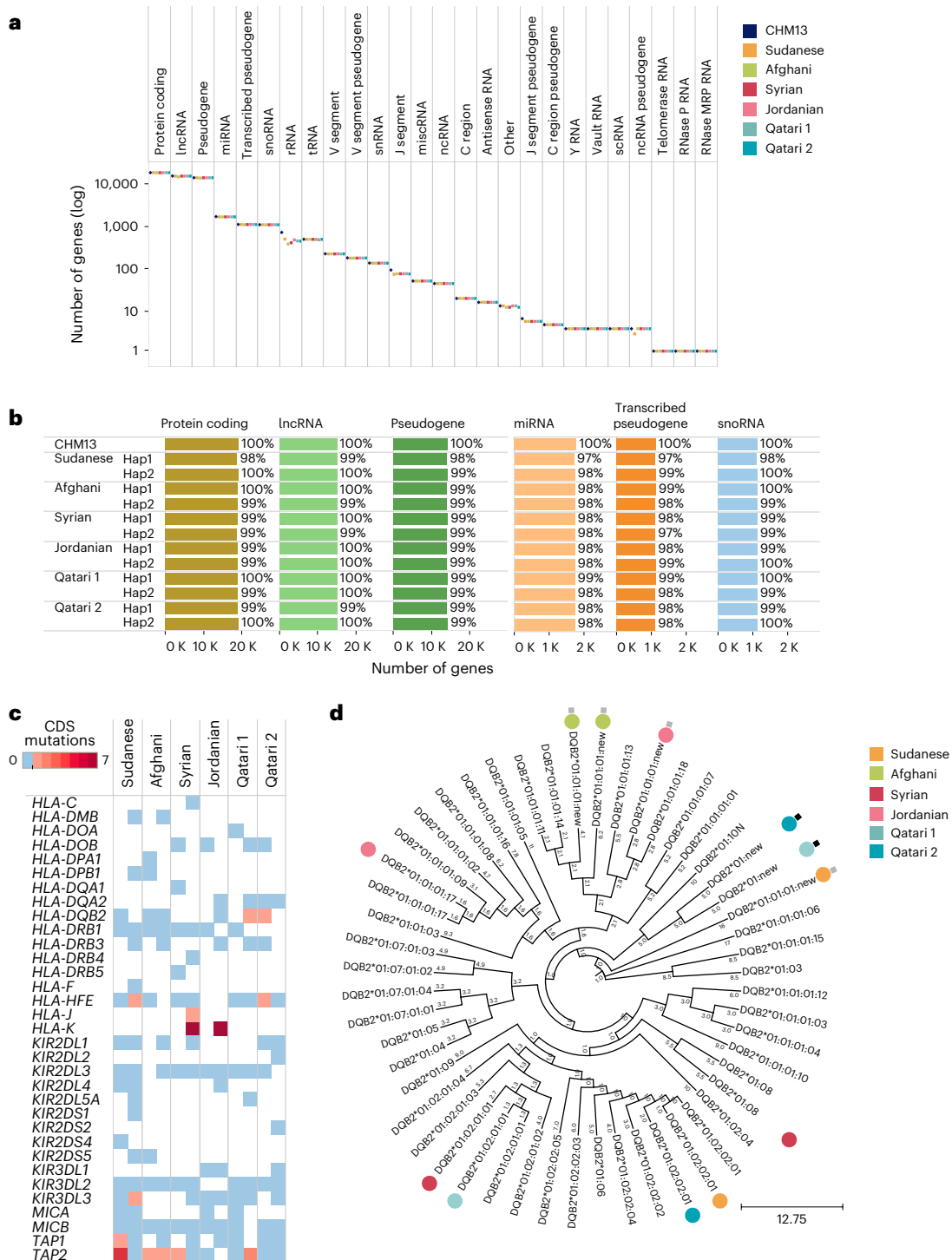


Fig. 4 | Gene coverage and HLA and KIR gene annotation. a, Gene counts across various gene categories. **b**, Coverage for the largest gene categories. **c**, New alleles in HLA and KIR genes in the child assembly haplotypes, highlighting the number of mutations in the CDS relative to a reference dataset of 220 pre-annotated reference haplotypes. **d**, Phylogenetic tree based on neighbor-joining method for the *HLA-DQB2* locus showing clustering pattern of the alleles

in the child assemblies. New alleles are labeled, showing those with mutations in the CDS region (black squares) and those with mutations in other part of the sequence (gray squares). Next to each leaf-node connecting branch, the corresponding evolutionary distance is marked. Length key is shown in the bottom right. CDS, coding DNA sequence.

association in the literature (ACMG total score ≥ 0.45). These include 10 that we classified as high interest VUS (ACMG total score = 0.60–0.75; Supplementary Table 2), such as *SOX5* (delayed speech and language development) and *BRSK2* (neurodevelopmental disorders). Notably, in

Qatari 1 proband who has global developmental delay, we identified a 234 bp Del in intron 5 of *KIF1A*, which encodes a motor protein essential for axonal transport of synaptic vesicle precursors. A de novo mutation in this gene has been previously reported to cause cerebellar atrophy

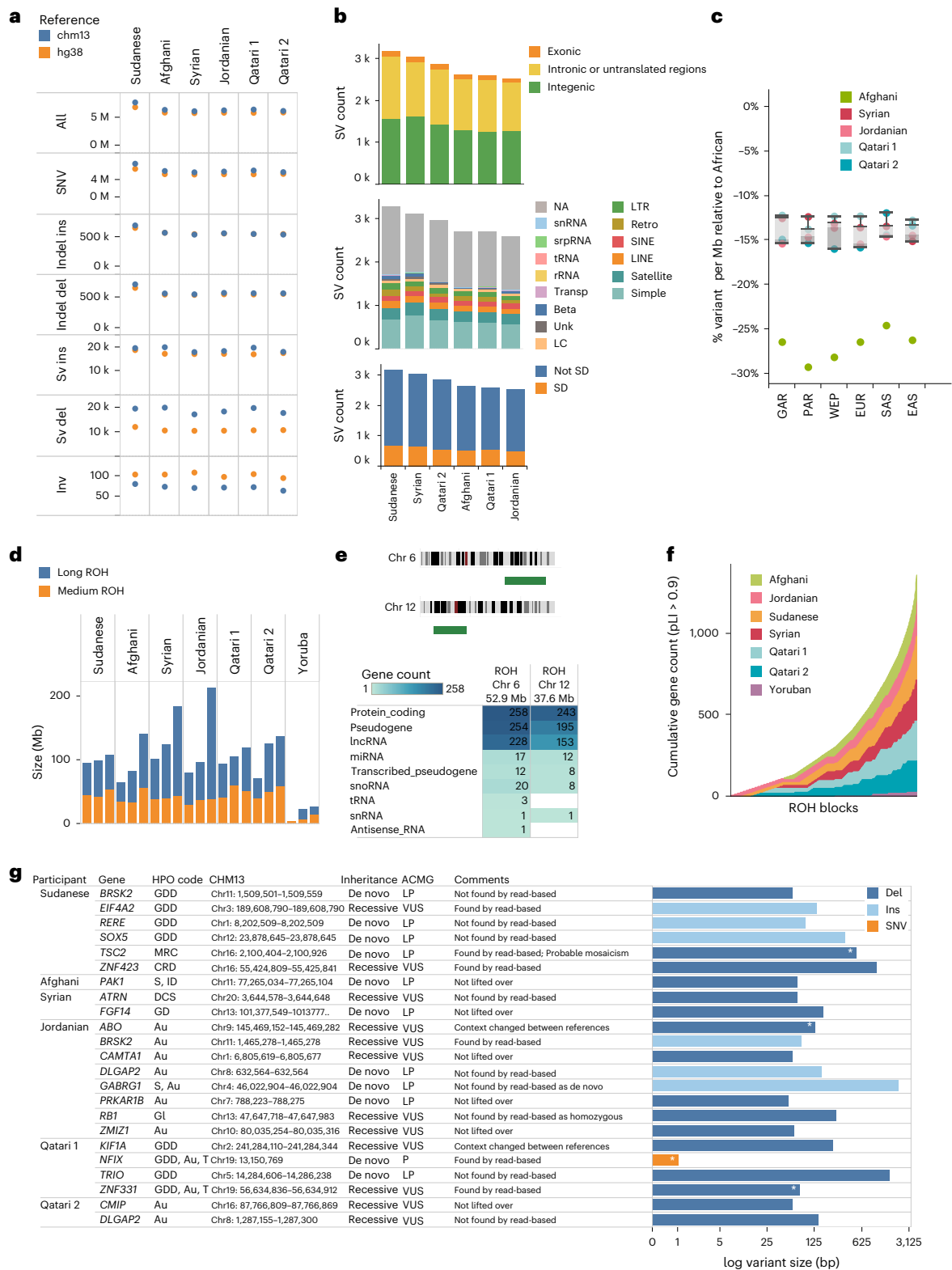


Fig. 5 | Genetic variation, ROH and candidate disease-causing variants. a, SV count against CHM13 and GRCh38 for each child assembly haplotype. **b**, Count of SV variants (deletions and insertions) in the family trios called against CHM13 and found to be absent from the HPRC dataset, highlighting their spread across intergenic, intronic/UTR and exonic regions (top), repetitive regions (middle) and segmental duplications (bottom). **c**, Box plot showing median counts of variants per MB relative to African segments in the same participants aggregated per family ($n = 15$), for various ancestries. **d**, Cumulative sizes of long and medium ROH of the ME assemblies and the Yoruba 1KG trio. **e**, Location and count of genes within the long ROH segments for chromosomes 6 and 12 of the Jordanian

father. **f**, Cumulative number of genes ($pLI > 0.9$) over contigs per child assembly. **g**, Candidate disease-causing variants in the probands. Shown are the variants, impacted genes, ascertained phenotypes in the child participants and associated details. The comments column indicates whether the variant was identified by read-based calling. Exonic deletions are denoted by an asterisk on the bars. SD, segmental duplication; HPO, Human Phenotype Ontology; Au, autism; CRD, cystic renal dysplasia; DCS, duplicated collecting system; GD, gait disturbance; GI, glaucoma; GDD, global developmental delay; ID, intellectual disability; MRC, multiple renal cysts; S, seizure; T, tall stature; P, pathogenic; LP, likely pathogenic.

and epilepsy in an Egyptian child from a consanguineous marriage⁴¹ and other variants in this gene have been classified by ClinGen and G2P to have a 'definitive' association with autosomal dominant syndromic intellectual disability. The largest shortlisted variant is a homozygous Del (2.2 kb) in the intronic region of *GABRG1* in the Jordanian proband who has seizures and autism. Variants in *GABRG1* are associated with seizures and global developmental delay⁴². This gene encodes the $\gamma 1$ subunit of the GABA-A receptor and has been associated with epileptic encephalopathy under 'limited' class on G2P. Therefore, our finding supports a more assertive classification for this gene.

We observed that 50% of candidate variants we shortlisted using the assembly-based alignment approach were not called based on the classical approach of aligning long reads, either entirely or by calling the wrong zygosity (Fig. 5g). We checked that the assembly-based variants occurred in regions of high-quality assembly as indicated by Flagger. Furthermore, when we repeated the exercise of trio analysis based on variants obtained using the classical read-based approach, we shortlisted seven de novo variants for the six participants (Supplementary Table 3). We found that two of those were already detected using an assembly-based approach, including the SNV variant for *NFIX*, while the others were not valid de novo upon further examination of the read alignments, since they also appeared in the parental genomes as illustrated for the variants in *TRPM3*, *AKT3* and *KMD4B* (Supplementary Fig. 24). Notably, that became clearer when the alignments were done against closer references, as illustrated by aligning Syrian participants against Jordanian father or HG002 (Supplementary Fig. 24).

Value of a ME population reference

The assemblies generated in this study allow us to assess the impact of a ME genome reference. First, to illustrate the relationship between genetic distance to the genome reference and the number of called variants, we aligned test samples with diverse ancestries from 1KG and ME populations against each of the ME child assemblies. For each assembly-ancestry combination pair, we used two test samples per population as follows: one closest and one furthest from the assembly, as identified based on principal component analysis (PCA). A significant linear correlation was observed between the number of variants and the Euclidean distance to the test samples for all assemblies except the Sudanese ($R = 0.18$; Fig. 6a).

Next, to specifically compare the performance of a ME reference relative to the published T2T references controlling for completeness, we examined the impact on read mappability and variant calling of SNVs and SVs from aligning 27 short-read test samples with diverse ME Arab ancestries encompassing PAR, GAR and WEP against GRCh38, CHM13, HG002 and MER, a hybrid assembly based on CHM13 in which we replaced six chromosomes with diverse T2T haplotypes from Qatari1 (chromosomes 4 and 18), Jordanian (chromosomes 10 and 12), Syrian (chromosome 6) and Qatari 2 (chromosome 11), which we denote as MER1. In terms of read mappability against the six chromosomes, MER1 shows a subtle but consistent trend of improved metrics relative to CHM13, HG002 and CN1 in this order (Fig. 6b). The difference in the normalized ratio of unmapped reads pairs over mapped relative to MER1 is highest for CN1 (median = 0.22) and lowest for CHM13 (median = 0.06). The number of singletons follows a similar trend with CN1, showing 3.6-fold more singletons relative to MER1 than CHM13. On the other hand, the metrics show a prominent improvement relative to GRCh38 (Supplementary Fig. 25). Regarding variant calls, more variants per MB relative to MER1 are obtained against other references, with HG002 being the closest. Respectively, GRCh38, CN1 and CHM13 show 7.1-, 2.3- and 1.3-fold more median variant counts in comparison to HG002 (Fig. 6c).

Discussion

With the increasing ability to produce high-quality genome assemblies at a larger scale, generating population genome references has become more compelling than ever. However, the challenge remains in the

strategy for sampling diverse admixture. While the HPRC pangenome included many East Asian genomes, the CPC pangenome revealed substantial diversity outside HPRC^{4,5}. ME populations, with their distinct genetic structures and history, require more refined sampling methods. Our findings show varied admixture patterns linked to geographical dispersal, even among populations sharing language and culture. Thus, future efforts to construct accurate population references must consider historical and anthropological factors, instead of solely relying on modern delineations of regions and territories.

The use of parental information has long been considered the gold standard for phasing genome assemblies⁴³; however, new methods relying on methylation, Hi-C and Strand-seq have substantially improved phasing quality for individual samples^{44–46}. Nevertheless, in endogamous/consanguineous populations, parental phasing remains essential in resolving the extensive homozygosity. Trio datasets like the one presented here will help in developing population-specific recombination maps facilitating larger-scale phasing.

Our assemblies are nearly complete compared to CHM13, revealing 7–11 Mb of new sequences per participant, consistent with previous reports⁴⁷. Some regions, such as centromeres and acrocentric p-arms, are not necessarily missing; however, they are difficult to align due to their divergence and ectopic recombination. Ninety percent of new sequences are in repetitive intergenic regions and the remaining are enriched for gene elements. Gene coverage is high with some decrease for rRNA genes, known to have multiple copies²⁸. We identified 75 new HLA/KIR alleles in the assemblies, including 11 with CDS mutations. Four alleles were unique to Sudanese ancestry, consistent with its greater African genetic divergence. We map these alleles' phylogenetic context, which will serve as a valuable addition to understanding HLA/KIR diversity and their role in immune-mediated processes across populations⁴⁸.

Soon, assembly-based methods are expected to replace read-based methods for alignment and variant calling, leading to higher accuracy^{49,50}, especially when the query is divergent from the genome reference. This is highlighted by our comparison of DeepVariant (read based)⁵¹ and phased assembly variant caller (PAV) (assembly based)⁵². The ME assemblies exhibited 6.6% more variants against CHM13 than GRCh38, consistent with CHM13 having 8% more sequence than GRCh38 (ref. 6). We observed fewer SV deletions and more inversions against GRCh38, reflecting known imbalances⁸ and misorientations in the latter^{9,53}. Nearly half of the new SVs in ME genomes span intronic and exonic regions, with a balanced distribution between repetitive and non-repetitive regions, suggesting a broad biological impact.

We show that assembly-based variant calling, especially with T2T references like CHM13, enhances variant detection and inheritance analysis in rare disease trios⁵⁴. However, while this improves the identification of disease-causing variants, it may not immediately increase genetic diagnosis yield due to limited functional annotation, especially in non-coding regions⁵⁵. Intronic variants are increasingly implicated in rare diseases⁵⁶ but are often classified as VUS under current ACMG criteria due to the lack of algorithms scoring complex non-coding variants⁵⁷. This makes functional validation a bottleneck for confirming many genetic diagnoses. As sequencing advances near-complete variant landscapes, large-scale genome annotation efforts—such as the IGVF consortium⁵⁸—are urgently needed for systematic interpretation and prioritization.

Phased genome assembly allows more accurate ROH calculation, essential when studying consanguineous populations. As expected, we observe high ROH levels in the ME trios, especially long ROH, indicative of recent inbreeding^{32,33}. This includes extended ROH in healthy individuals, leading to the autozygosity of many genes, uncovering extensive benign variation. Also, it harbors recessive deleterious variants contributing to disease phenotypes¹⁶. While de novo variants are a major cause of developmental disorders⁵⁹, recessive variants, as shown

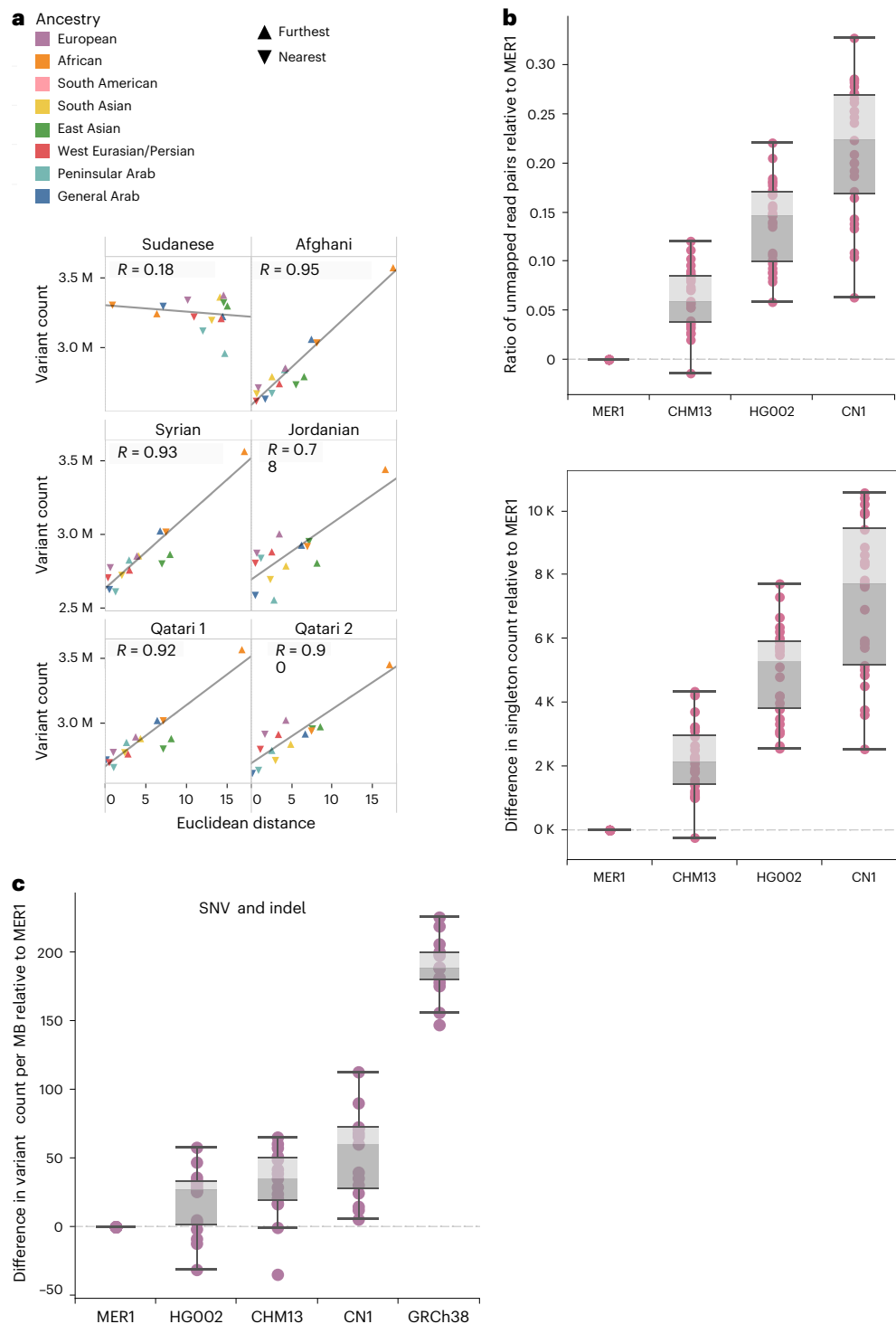


Fig. 6 | Variant calling and mappability against MER1 and other references. a, Euclidean distance versus variant count for each of the child assemblies. Color indicates the ancestry of the test samples. The shape of the markers differentiates the samples with maximum and minimum distance from our assembly for a given ancestry. Regression lines and coefficients of the Pearson correlations are shown. Corresponding P values are $<10^{-4}$ for all except Sudanese ($P = 0.54$), calculated

using two-sided t test. **b,** Ratio of unmapped read pairs over mapped and number of singletons relative to MER1 in the replacement chromosomes for ME query samples ($n = 15$) for various reference genomes. Values were calculated per 1 Mb regions and averaged over chromosomes. **c,** Differences in variant counts per 1 Mb for ME query samples ($n = 15$) from various ME ancestries against various reference genomes relative to MER1.

here, are relevant in consanguineous populations, and multiple variants can co-occur, explaining comorbidities and variable penetrance⁶⁰.

Finally, we demonstrated that using ME assemblies as references reduces variant calls and improves mappability compared to unmatched references, highlighting the need for population-specific

genome references to capture relevant variations, including pathogenic ones. As more disease cohorts undergo long-read sequencing and de novo assembly, these data could enhance pangenome efforts with informative haplotypes as long as they lack gross chromosomal abnormalities. While there is a need for a universal pangenome reference

cataloging global variation⁶¹, its wide adoption remains challenging. That is also compounded by the shift toward direct genome comparisons using multiple alignments. Nevertheless, integrating ME pangenomes and other regional datasets would greatly enrich the genomic and biomedical landscape.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41588-025-02173-7>.

References

- Seo, J. S. et al. De novo assembly and phasing of a Korean human genome. *Nature* **538**, 243–247 (2016).
- Shumate, A. et al. Assembly and annotation of an Ashkenazi human reference genome. *Genome Biol.* **21**, 129 (2020).
- O’Leary, N. A. et al. Exploring and retrieving sequence and metadata for species across the tree of life with NCBI datasets. *Sci. Data* **11**, 732 (2024).
- Liao, W.-W. et al. A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
- Gao, Y. et al. A pangenome reference of 36 Chinese populations. *Nature* **619**, 112–121 (2023).
- Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
- Yang, C. et al. The complete and fully-phased diploid genome of a male Han Chinese. *Cell Res.* **33**, 745–761 (2023).
- Aganezov, S. et al. A complete reference genome improves analysis of human genetic variation. *Science* **376**, eabl3533 (2022).
- Vollger, M. R. et al. Increased mutation and gene conversion within human segmental duplications. *Nature* **617**, 325–334 (2023).
- Guarracino, A. et al. Recombination between heterologous human acrocentric chromosomes. *Nature* **617**, 335–343 (2023).
- Rodriguez-Flores, J. L. et al. Indigenous Arabs are descendants of the earliest split from ancient Eurasian populations. *Genome Res.* **26**, 151–162 (2016).
- Almarri, M. A. et al. The genomic history of the Middle East. *Cell* **184**, 4612–4625 (2021).
- Mineta, K., Goto, K., Gojobori, T. & Alkuraya, F. S. Population structure of indigenous inhabitants of Arabia. *PLoS Genet.* **17**, e1009210 (2021).
- Razali, R. M. et al. Thousands of Qatari genomes inform human migration history and improve imputation of Arab haplotypes. *Nat. Commun.* **12**, 5929 (2021).
- Monies, D. et al. Lessons learned from large-scale, first-tier clinical exome sequencing in a highly consanguineous population. *Am. J. Hum. Genet.* **105**, 879 (2019).
- Temaj, G., Nuhii, N. & Sayer, J. A. The impact of consanguinity on human health and disease with an emphasis on rare diseases. *J. Rare Dis.* **1**, 2 (2022).
- El Naofal, M. et al. The genomic landscape of rare disorders in the Middle East. *Genome Med.* **15**, 5 (2023).
- Saad, M. et al. Genetic predisposition to cancer across people of different ancestries in Qatar: a population-based, cohort study. *Lancet Oncol.* **23**, 341–352 (2022).
- Aamer, W. et al. Burden of Mendelian disorders in a large Middle Eastern biobank. *Genome Med.* **16**, 46 (2024).
- Duncan, L. et al. Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* **10**, 3328 (2019).
- Martin, A. R. et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
- Malan, V. et al. Distinct effects of allelic NFIX mutations on nonsense-mediated mRNA decay engender either a Sotos-like or a Marshall-Smith syndrome. *Am. J. Hum. Genet.* **87**, 189–198 (2010).
- 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).
- 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).
- Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
- Logsdon, G. A. et al. The variation and evolution of complete human centromeres. *Nature* **629**, 136–145 (2024).
- Waldman, S. et al. Genome-wide data from medieval German Jews show that the Ashkenazi founder event pre-dated the 14th century. *Cell* **185**, 4703–4716 (2022).
- Parks, M. M. et al. Variant ribosomal RNA alleles are conserved and exhibit tissue-specific expression. *Sci. Adv.* **4**, eaao0665 (2018).
- Zhou, Y., Song, L. & Li, H. Full-resolution HLA and KIR gene annotations for human genome assemblies. *Genome Res.* **34**, 1931–1941 (2024).
- Hurley, C. K. et al. Common, intermediate and well-documented HLA alleles in world populations: CIWD version 3.0.0. *HLA* **95**, 516–531 (2020).
- Nd, O. et al. Variant calling and benchmarking in an era of complete human genome sequences. *Nat. Rev. Genet.* **24**, 464–483 (2023).
- Pemberton, T. J. et al. Genomic patterns of homozygosity in worldwide human populations. *Am. J. Hum. Genet.* **91**, 275–292 (2012).
- Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M. & Wilson, J. F. Runs of homozygosity: windows into population history and trait architecture. *Nat. Rev. Genet.* **19**, 220–234 (2018).
- Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
- Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
- 1000 Genomes Project Consortium. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
- Mbarek, H. et al. Qatar genome: insights on genomics from the Middle East. *Hum. Mutat.* **43**, 499–510 (2022).
- Rehm, H. L. et al. ClinGen—the clinical genome resource. *N. Engl. J. Med.* **372**, 2235–2242 (2015).
- Wright, C. F. et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* **385**, 1305–1314 (2015).
- Firth, H. V. et al. DECIPHER: database of chromosomal imbalance and phenotype in humans using ensembl resources. *Am. J. Hum. Genet.* **84**, 524–533 (2009).
- Megahed, H. et al. Utility of whole exome sequencing for the early diagnosis of pediatric-onset cerebellar atrophy associated with developmental delay in an inbred population. *Orphanet J. Rare Dis.* **11**, 57 (2016).
- Williams, A. et al. GABRG1 variant as a potential novel cause of epileptic encephalopathy, hypotonia, and global developmental delay. *Am. J. Med. Genet. A* **188**, 3546–3549 (2022).
- Jarvis, E. D. et al. Semi-automated assembly of high-quality diploid human reference genomes. *Nature* **611**, 519–531 (2022).

44. Kronenberg, Z. N. et al. Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C. *Nat. Commun.* **12**, 1935 (2021).
45. Fu, Y. et al. MethPhaser: methylation-based long-read haplotype phasing of human genomes. *Nat. Commun.* **15**, 5327 (2024).
46. Porubsky, D. et al. Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nat. Biotechnol.* **39**, 302–308 (2021).
47. Li, Q., Yan, B., Lam, T.-W. & Luo, R. Assembly-free discovery of human novel sequences using long reads. *DNA Res.* **29**, dsac039 (2022).
48. Barker, D. J. et al. The IPD-IMGT/HLA database. *Nucleic Acids Res.* **51**, D1053–D1060 (2023).
49. Heller, D. & Vingron, M. SVIM-asm: structural variant detection from haploid and diploid genome assemblies. *Bioinformatics* **36**, 5519–5521 (2021).
50. Porubsky, D. & Eichler, E. E. A 25-year odyssey of genomic technology advances and structural variant discovery. *Cell* **187**, 1024–1037 (2024).
51. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
52. Ebert, P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).
53. Porubsky, D. et al. Inversion polymorphism in a complete human genome assembly. *Genome Biol.* **24**, 100 (2023).
54. AlAbdi, L. et al. Diagnostic implications of pitfalls in causal variant identification based on 4577 molecularly characterized families. *Nat. Commun.* **14**, 5269 (2023).
55. Castro, C. P., Diehl, A. G. & Boyle, A. P. Challenges in screening for de novo noncoding variants contributing to genetically complex phenotypes. *HGG Adv.* **4**, 100210 (2023).
56. Lord, J. & Baralle, D. Splicing in the diagnosis of rare disease: advances and challenges. *Front. Genet.* **12**, 689892 (2021).
57. Ellingford, J. M. et al. Recommendations for clinical interpretation of variants found in non-coding regions of the genome. *Genome Med.* **14**, 73 (2022).
58. IGVF Consortium. Deciphering the impact of genomic variation on function. *Nature* **633**, 47–57 (2024).
59. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433–438 (2017).
60. Rahit, K. M. T. H. & Tarailo-Graovac, M. Genetic modifiers and rare Mendelian disease. *Genes (Basel)* **11**, 239 (2020).
61. Kaye, A. M. & Wasserman, W. W. The genome atlas: navigating a new era of reference genomes. *Trends Genet.* **37**, 807–818 (2021).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025

¹Sidra Medicine, Doha, Qatar. ²Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA. ³Department of Biomedical Science, College of Health Sciences, Qatar University, Doha, Qatar. ⁴Department of Genetic Medicine, Weill Cornell Medicine, Doha, Qatar. ⁵College of Health and Life Sciences, Hamad Bin Khalifa University, Doha, Qatar. ⁶National Human Genome Research Institute, Bethesda, MD, USA. ⁷Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. *A list of authors and their affiliations appears at the end of the paper.

✉ e-mail: ymokrab@sidra.org

Qatar Genome Program Research Consortium

Qatar Genome Project Management

Said I. Ismail⁸, Wadha Al-Muftah⁸, Radja Badji⁸, Hamdi Mbarek⁸, Dima Darwish⁸, Tasnim Fadl⁸, Heba Yasin⁸, Maryem Ennaifar⁸, Rania Abdellatif⁸, Fatima Alkuwari⁸, Muhammad Alvi⁸, Yasser Al-Sarraj⁸, Chadi Saad⁸ & Asmaa Althani⁸

Biobank and Sample Preparation

Eleni Fethnou⁹, Fatima Qafoud⁹, Eiman Alkhayat⁹ & Nahla Afifi⁹

Sequencing and Genotyping group

Sara Tomei¹, Wei Liu¹ & Stephan Lorenz¹

Applied Bioinformatics Core

Najeeb Syed¹, Hakeem Almabrazi¹, Fazalur Rehaman Vempalli¹ & Ramzi Temanni¹

Data Management and Computing Infrastructure group

Tariq Abu Saqri¹, Mohammedhusen Khatib¹, Mehshad Hamza¹, Tariq Abu Zaid¹, Ahmed El Khouly¹, Tushar Pathare¹, Shafeeq Poolat¹ & Rashid Al-Ali¹

Consortium Lead Principal Investigators (in alphabetical order)

Omar Albagha⁹, Souhaila Al-Khodor¹, Mashael Alshafai³, Ramin Badii¹⁰, Lotfi Chouchane⁴, Xavier Estivill¹, Khalid Fakhro^{1,4,5}, Hamdi Mbarek⁸, Younes Mokrab^{1,3,4}, Jithesh V. Puthen⁵, Karsten Suhre^{11,12} & Zohreh Tatari¹

⁸Qatar Genome Program, Qatar Foundation Research Development and Innovation, Qatar Foundation, Doha, Qatar. ⁹Qatar Biobank for Medical Research, Qatar Foundation, Doha, Qatar. ¹⁰Molecular Genetics Lab, Hamad Medical Corporation, Doha, Qatar. ¹¹Bioinformatics Core, Weill Cornell Medicine, Doha, Qatar. ¹²Department of Biophysics and Physiology, Weill Cornell Medicine, New York City, NY, USA.

Methods

Study samples

The study participants consist of six parent–child family trios ($n = 18$) from the local population in Qatar recruited at Sidra Medicine, whereby children were ascertained for neurodevelopmental disorders. Previously, three families (Sudanese, Jordanian and Qatari 2) had negative genetic diagnoses based on microarrays and selected panels.

The families are from various nationalities from the greater ME region—Sudan, Jordan, Syria, Qatar (2X) and Afghanistan. Whole peripheral blood samples were collected from the individuals, stored in EDTA tubes at -80°C and subsequently processed for DNA and RNA sequencing as described below.

Written informed consent and assent were obtained from all participants, and the study was approved by the Sidra Medicine Ethics Committee.

Sequencing data generation and quality assessment

Whole-genome Illumina short-read sequencing was performed at Sidra Medicine. First, genomic DNA was isolated from blood samples on the QIASymphony System (QIAGEN) using the DSP DNA Midi Kit, per the manufacturer's instructions. DNA quantity and quality were assessed using NanoDrop 8000 (Thermo Fisher Scientific); the absorbance at 260 and 280 nm wavelengths was used to check DNA purity. A fluorescence-based quantification was performed on FlexStation 3 (Molecular Devices) using Quant-iT PicoGreen dsDNA Assay (Thermo Fisher Scientific). DNA integrity was checked on LabChip GX (PerkinElmer). Next, 150 bp paired-end read libraries were constructed and sequencing was performed on a HiSeq X System at Sidra Medicine at $30\times$ coverage. QC was done using FastQC⁶² and MultiQC⁶³. All reads that had more than an average of 450 million unique paired-end reads were retained. SAMtools v1.17 (ref. 64) and Mosdepth v0.3.2 (ref. 65) were used to calculate coverage and number of mapped reads.

PacBio HiFi sequencing was performed at the University of Washington, as previously described⁶⁶ with some modifications. Briefly, high-molecular-weight DNA was extracted from the flash frozen peripheral blood samples using the QIAGEN MagAttract HMW DNA Kit (67563). After quality and quantity checks (Agilent FEMTO Pulse, M5330AA and FP-1002-0275; Denovix, DS-11 FX; Thermo Fisher Scientific Q32854), DNA was sheared on Megaruptor 3 (Diagenode, B06010003 and E07010003) twice using settings 29/31 to obtain a peak size around 22 kb. SMRTbell library preparation with the SMRTbell Express Template Prep Kit 2.0 (PacBio, 100-938-900) and size selection on Sage Science PippinHT (HPE7510) were performed as previously described in ref. 66. Next, sequencing was performed using 3–4 SMRT cells on the Sequel II platform with chemistry version 2.2, 30-h movie times and 2 h pre-extension (PacBio, 101-894-200), leading to coverage of $30\text{--}50\times$ per sample. Data quality was evaluated using SMRTlink v.10.1 QC reports from which we used the `awk v4.0.2 {print [field]}` command to extract the fields relating to sequencing quality, read length distribution and coverage. CCS reads with $\text{QV} \geq 20$ were retained for subsequent analysis.

Genome assembly and quality assessment

de novo genome assembly was conducted by running the trioHiFiiasm pipeline²³ (`hifiasm -o sample.asm -t32 -1 pat.yak -2 mat.yak HiFisample.fq.gz`), using the parental short reads to create k-mers generated using `yak/0.1 (yak count -k31 -b37 -t threads -o output.parental_yak)` to construct a pair of haplotype-resolved phased assemblies. The fasta files for the assembly pipeline was generated from gfa files using `awk v4.0.2` (<https://www.gnu.org/software/gawk/>).

Various assembly quality metrics were calculated on the haploid assemblies using `stats.sh` script in BBmap (v38.69)⁶⁷. Phasing quality was evaluated using Merqury v1.3 (ref. 24) by examining blob plots, switch error rates and phased blocks to ensure completeness, accuracy and reliability. The QV score was used to quantify phasing accuracy and completeness. Telomer count checks were made using `seqtk telo`⁶⁸,

confirming the presence of telomere repeats (TTAGGG) at chromosomal ends of the assemblies.

HiFi read-based evaluation of the assemblies was performed by running the Flagger v0.3.3 pipeline following the standard workflow as previously described⁴. For each assembly, it classifies segments of the assembly sequence into five categories, namely Haploid (reliable), unknown, duplicated, collapsed and erroneous.

Assembly alignment to GRCh38 and CHM13

Assembly-to-assembly alignment was conducted using Minimap2 (ref. 25) with options `-x asm5 -c`, which produced paf files with CIGAR string. To allow visualization with Saffire⁶⁹, the paf files were converted to Saffire format using `rustybam v0.1.31` (ref. 70) with the command `rb break-paf --max-size 5000 | rb orient | rb filter --paired-len 100000 | rb stats -paf`. The Saffire files were uploaded to an internally deployed Saffire server, which provided an interface to visualize the alignments of haplotypes to GRCh38 and CHM13 (v2.0).

Variant calling

For the variant calling in ME assemblies relative to GRCh38 and CHM13, we used two approaches: One based on HiFi read alignments and the other on assembly-based alignments. For the first approach, `pbmm2 1.13.0`, a wrapper for `minimap2` (ref. 25) was used to generate the alignment and `DeepVariant v1.5.0` (ref. 51) was used to call both SNVs and indels, while `PBSV` was used to call SVs⁷¹. For the second approach, `PAV v2.3.3` (ref. 52) was used to do both the alignment and calling of SNVs, indels and SVs including insertions, deletions (≥ 50 bp) and inversions.

For variant calling of various reference samples from the IKG and QGP against the assemblies, Illumina short reads were aligned using `BWA-MEM`⁷² and `DeepVariant v1.5.0` (ref. 51) was used to call SNVs and indels. Variants with the PASS filter were retained for downstream analysis.

PCA

PCA was performed using SNVs from a dataset combining the 18 study participants, 2,504 samples from the IKG phase 3 dataset³⁶ and 6,216 samples from the QGP phase 1 (refs. 14, 37) in addition to an internal diverse set of ME participants at Sidra Medicine ($n = 1,693$). For the published datasets, SNVs with the PASS filter were used. For the study participants, SNVs were obtained from the Illumina data by aligning to GRCh37 and calling the variants using the GATK best practice workflow⁷³, retaining only SNVs with the PASS filter. The combined dataset was generated using `BCFtools isec v1.17` (ref. 64) and contained 21.3 million SNPs. We applied a minor allele frequency (MAF) filter with a 0.01 threshold, resulting in 7.9 million variants from which principal components were calculated using `PLINK v2.00a2LM 64-bit Intel` (7 January 2019)⁷⁴.

Global and local genetic ancestry assignment

`RFMix v2.0.0` (ref. 75) was used to assign global and local ancestry for the genomes of each study participant based on VCF file containing Illumina-derived SNPs and indels, using as reference a diverse panel from the IKG and QGP phase 1 cohort which consists of 176 representative individuals identified as having a dominant ancestral fraction $>90\%$ for eight distinct ancestral fractions¹⁴. This was derived from Admixture analysis of 6,216 samples from QGP and 2,504 samples from IKG¹⁴, ensuring diversity encompasses various continental populations including various ME groups. Output from RFMix consisted of ancestry estimates for chromosomal painting as well as global ancestry estimates (.Q file). RFMix requires phased genotypes, therefore, these were generated using `EAGLE v2.4.1` (ref. 76), which was run with three iterations and by providing genetic map and default parameters.

New sequences

Insertions from PAV alignment for the assemblies against CHM13 (2.0) were collapsed per participant using `truvari`⁷⁷ (`truvari collapse -r 500 -p`

0.95-P0.95-s50-S100000). Unique sequences were identified by comparing with the cohort of 47 samples from HPRC for which SVs were previously called⁸ using `truvari (truvari bench -r 1000 -C 1000 -O 0.8 -p 0.8 -P 0.0 -s 50 -S15 --sizemax 100000)`. An 80% reciprocal overlap was used. In addition, insertions from various participants were also collapsed into a unique set of sequences using the same process starting from collapsed insertions per participant. Subsequently, the resulting sequences were checked similarly for overlap with HG002 (v. 1.0.1) and CN1.

SV annotation with gene elements, repetitive regions and segmental duplications

We used `bedtools v2.30` (ref. 78) with default options including overlap threshold of 1 bp to intersect SVs with bed files obtained from the CHM13 public resource, namely genetic features (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/009/914/755/GCF_009914755.1_T2T-CHM13v2.0/GCF_009914755.1_T2T-CHM13v2.0_genomic.gtf.gz), repetitive regions (https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/annotation/chm13v2.0_RepeatMasker_4.1.2p1.2022Apr14.bed), segmental duplications (https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/annotation/chm13v2.0_SD.full.bed), centromeric and satellite repeats (https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/annotation/chm13v2.0_censat_v2.1.bed).

Gene annotation

To annotate assemblies with known genes, we used `Liftoff v1.6.3` (ref. 79), using NCBI Refseq GTF file for CHM13 (<https://github.com/marbl/CHM13?tab=readme-ov-file>) with default parameters. The pLI scores for individual genes were obtained from `gnomAD`³⁵ (<https://gnomad.broadinstitute.org/downloads#v4-constraint>).

Typing of HLA and KIR alleles and generation of phylogenetic trees

We used `Immunoannot` to annotate the alleles of HLA and KIR genes per genome haplotype²⁹. We typed the alleles of these genes using default options and extracted CDS mutations and CDS distance for new gene alleles. The results were compared with previously published annotations of 220 pre-annotated reference haplotypes, including HLA assemblies and those from HPRC, CPC, CHM13, CN-T2T and GRCh38.p14 downloaded from <https://zenodo.org/records/8372992>.

Phylogenetic trees were generated by first extracting the DNA sequences of each immune gene from the study assemblies using `SAMtools v1.17` (ref. 64). These sequences were combined with the publicly available reference sequences into a single FASTA file per gene. This was used as input to generate a multiple-sequence alignment and a neighbor-joining phylogenetic tree using `Clustal Omega`⁸⁰. The trees were plotted using `TreeView`⁸¹. Sequences were retrieved from the IPD⁴⁸. Alignment figures were generated using `Jalview`⁸².

Allele frequencies for *HLA-A* alleles were obtained from the CIWD catalog v3.0 (ref. 30). Population ancestry of source samples for the closest alleles to the new alleles identified in the assemblies was obtained for the IPD⁴⁸.

ROH

ROH segments were identified per participant using `Automap v1.2` (ref. 83) using default parameters and variant calls from PAV for the assemblies against CHM13 as input. GTF file was also provided as input, which resulted in the annotation of ROH segments using genes. A selection of ROH regions was visually verified using `Integrative Genomics Viewer (IGV)`⁸⁴.

Family trio analysis and variant prioritization

Using SNVs, indels and SVs produced from the trios called against CHM13, we perform trio-based analysis for each family following the workflow of

`Exomiser v14` (ref. 85), which itself does not support the CHM13 reference. This method combines variant prioritization and phenotypic concordance⁸⁶. Given that the disease symptoms affect the children only, we considered the modes of inheritance of recessive and de novo. First, we shortlisted variants that met each mode of inheritance and then did a lift over to GRCh38 coordinates followed by running annotation using `Janovar`⁸⁷ and `annotSV v3.4.2` (ref. 88) to retrieve various functional scores and allele frequencies from various public databases including allele frequencies from various public databases including `gnomAD v4.0` (ref. 35), `dbVar` and `1KG`³⁶, to which we appended allele frequencies from Qatar Genome Project Phase 1 (refs. 14, 37). For the SV annotation with `annotSV v3.4.2` (ref. 88), we adopted minimum 80% reciprocal overlap to account for breakpoint imprecision. We selected variants with $AF < 0.01$ in various population databases and have genotype–phenotype association P value of < 0.05 and a gene combined score and genotype–phenotype score of > 0.8 . Further filtering was performed based on variants with a high functional consequence. Finally, we cross-checked in `ClinGen`³⁸, `DECIPHER`³⁸ and `Gene2Phenotype`³⁹ to confirm the relevance of the link obtained to the proband's condition. We checked that variants were in regions flagged as reliable by `Flagger v0.3.3` (ref. 4).

We performed variant classification for SNVs and Indels based on the guidelines recommended by ACMG⁸⁹, and for SVs based on the joint recommendations by ACMG and ClinGen⁹⁰ and as implemented in `Franklin` (<https://franklin.genoox.com/clinical-db/home>). For the latter, we classified variants as pathogenic, likely pathogenic and VUS if they have total ACMG scores ≥ 0.99 , $0.90-0.98$ and -0.89 to 0.89 , respectively.

Calculation of metrics of read mappability

To assess read mappability against various genome references, we selected 27 representative samples with diverse ME Arab ancestries encompassing PAR, GAR and WEP for which Illumina reads (30×) were mapped against each reference using `DRAGEN`⁹¹. Metrics of the ratio of unmapped read pairs over mapped, percentage of properly paired reads and number of singletons were calculated based on values obtained from the output of `DRAGEN` and were tabulated to generate the various plots.

Statistics and reproducibility

No statistical method was used to predetermine sample size, but our sample sizes are typical of those used in the field of study. No data were excluded from the analyses. The experiments were not randomized. Investigators were not blinded to allocation during experiments and outcome assessment. Pearson correlation with two-sided significance was calculated. The tools and packages used for the processing and analysis of data in this study have been reported in the Methods and Study summary, with the specific versions allowing the reproducibility of the results. In all box plots in this study, the boxes indicate the middle 50% of the data, while whiskers indicate 1.5 times the interquartile range. The lower quartiles are dark shaded and the upper ones are light shaded.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Sequencing data and assemblies generated in this study are available for general research use through controlled access at dbGAP to preserve patient confidentiality (accession ID `phs003917.v1.p1`; https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs003917.v1.p1).

Previously published data used in the study are accessible as follows: 1000 Genomes Project <https://hgdownload.cse.ucsc.edu/gbdb/hg19/1000Genomes/phase3> CHM13 assembly

https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/analysis_set/chm13v2.0.fa.gz
 CHM13 annotations https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/009/914/755/GCF_009914755.1_T2T-CHM13v2.0/GCF_009914755.1_T2T-CHM13v2.0_genomic.gtf.gz
https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/annotation/chm13v2.0_RepeatMasker_4.1.2p1.2022Apr14.bed
https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/annotation/chm13v2.0_SD.full.bed
https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/annotation/chm13v2.0_censat_v2.1.bed
 CIWD v3.0 catalog
<https://www.ihw18.org/component-immunogenetics/download-common-and-well-documented-alleles-3-0/>
 CN1 assembly
https://genome.zju.edu.cn/files/v1.0.1/CN1_pat.v1.0.1.fasta.gz
https://genome.zju.edu.cn/files/v1.0.1/CN1_mat.v1.0.1.fasta.gz
 gnomad v4.1.0
<https://gnomad.broadinstitute.org>
 gnomad gene constraint
https://storage.googleapis.com/gcp-public-data-gnomad/release/4.0/constraint/gnomad.v4.0.constraint_metrics.tsv
 GRCh38 assembly <https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/p13/hg38.p13.chromFa.tar.gz>
 HG002 assembly
<https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/HG002/assemblies/hg002v1.0.1.fasta.gz>
 immuannot IPD/KIR and CPC dataset
<https://zenodo.org/records/8372992/files/Data-2023Oct27.tar.gz?download=1>
 Qatar Biobank/Qatar Genome Project accession ID: QF-QGP-RES-PUB-007 (<https://www.qatarbiobank.org.qa/research/how-apply>). Source data are provided with this paper.

Code availability

The study used previously published analysis tools as described in the Methods. No custom tools were used in this study.

References

62. Andrews, S. *s-andrews/FastQC*. *GitHub* <https://github.com/s-andrews/FastQC> (2023).
63. Ewels, P., Magnusson, M., Lundin, S. & Kaller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
64. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
65. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).
66. Hallast, P. et al. Assembly of 43 human Y chromosomes reveals extensive complexity and variation. *Nature* **621**, 355–364 (2023).
67. SourceForge. BMap. <https://sourceforge.net/projects/bbmap/> (2023).
68. Li, H. et al. *lh3/seqtk*. *GitHub* <https://github.com/lh3/seqtk> (2024).
69. Vollger, M. R. et al. *mrVollger/SafFire* at v0.2. *GitHub* <https://github.com/mrvollger/SafFire> (2022).
70. Vollger, M. R. & Hulselmans, G. *mrVollger/rustybam*: v0.1.33 (v0.1.33). *Zenodo* <https://doi.org/10.5281/zenodo.8106233> (2023).
71. Töpfer, A. et al. *PacificBiosciences/pbsv*. *GitHub* <https://github.com/PacificBiosciences/pbsv> (2023).
72. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at <https://arxiv.org/abs/1303.3997> (2013).
73. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
74. Purcell, S. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
75. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).
76. Loh, P.-R. et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
77. English, A. C., Menon, V. K., Gibbs, R. A., Metcalf, G. A. & Sedlazeck, F. J. Truvari: refined structural variant comparison preserves allelic diversity. *Genome Biol.* **23**, 271 (2022).
78. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
79. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37**, 1639–1643 (2021).
80. Madeira, F. et al. Search and sequence analysis tools services from EMBL–EBI in 2022. *Nucleic Acids Res.* **50**, W276–W279 (2022).
81. Bianchini, G. & Sánchez-Baracaldo, P. TreeViewer: flexible, modular software to visualise and manipulate phylogenetic trees. *Ecol. Evol.* **14**, e10873 (2024).
82. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).
83. Quinodoz, M. et al. AutoMap is a high performance homozygosity mapping tool using next-generation sequencing data. *Nat. Commun.* **12**, 518 (2021).
84. Robinson, J. T. et al. Integrative Genomics Viewer. *Nat. Biotechnol.* **29**, 24 (2011).
85. Smedley, D. et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nat. Protoc.* **10**, 2004–2015 (2015).
86. Jacobsen, J. O. B. et al. Phenotype-driven approaches to enhance variant prioritization and diagnosis of rare disease. *Hum. Mutat.* **43**, 1071–1081 (2022).
87. Jäger, M. et al. Jannovar: a Java library for exome annotation. *Hum. Mutat.* **35**, 548–555 (2014).
88. Geoffroy, V. et al. AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics* **34**, 3572–3574 (2018).
89. Richards, S. et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).
90. Riggs, E. R. et al. Technical standards for the interpretation and reporting of constitutional copy number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). *Genet. Med.* **22**, 245–257 (2019).
91. Behera, S. et al. Comprehensive genome analysis and variant detection at scale using DRAGEN. *Nat. Biotechnol.* <https://doi.org/10.1038/s41587-024-02382-1> (2024).

Acknowledgements

We are grateful to the participants who provided the samples for this study. Sidra Medicine is supported by Qatar Foundation and the Ministry of Finance, Qatar. This study was funded in part by Qatar National Research Fund (QNRF awards ARG01-0426-230012, PPM1-1122-150008 and NPRP10-1219-160035) as well as Sidra internal funds. E.E.E. is an investigator of the Howard Hughes Medical Institute and is supported by US National Institutes of Health (NIH) grant R01MH101221. For public data used in this study, we would like to thank

the IKG Consortium (dataset). A.R. is supported by the Intramural Research Program of the National Human Genome Research Institute, NIH. We thank T. Brown from the University of Washington for assistance in editing the manuscript and other personnel from core and support services at Sidra Medicine.

Author contributions

Overall study was conceived and designed by Y.M. and E.E.E. Phenotypic ascertainment of probands was performed by A.S., R.B. and R.R. Sample collection and preparation was performed by R.A., M.T. and S.H. DNA extraction, library preparation and sequencing were done by K.M., S.T., R.M., L.S.M., K.W. and M.A. Assemblies were generated by W.H. and quality assessed by M.G., S.M., R.F. and Y.M. Data analysis and generation of results was performed by M.G. with assistance from S.M., Z.S., R.F., A.S., H.N., F.K.M., E.A. and I.D. Advice on method optimization and data interpretation was provided by A.R., E.E.E., M.G. and Y.M. Assistance with pipelines and scientific computing was provided by F.V. and A.E. Manuscript writing and figure preparation was performed by Y.M. with contributions from M.G., Z.S. and S.M. Input on scientific discussions and feedback on manuscript was provided by E.E.E., A.R., S.T., K.A.F. and I.A.

Competing interests

E.E.E. is a scientific advisory board member of Variant Bio. The other authors declare no competing interests.

Additional information

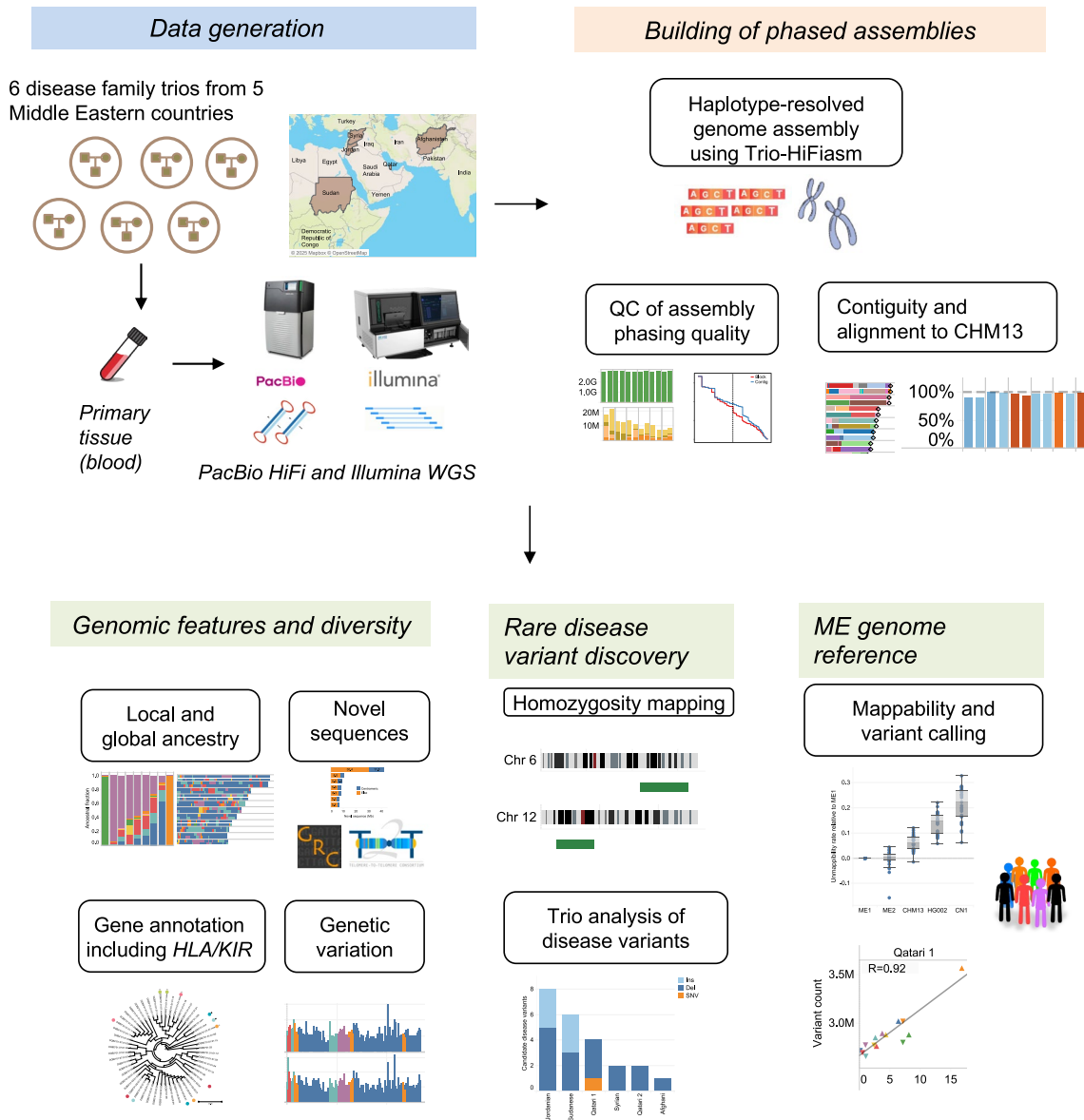
Extended data is available for this paper at <https://doi.org/10.1038/s41588-025-02173-7>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41588-025-02173-7>.

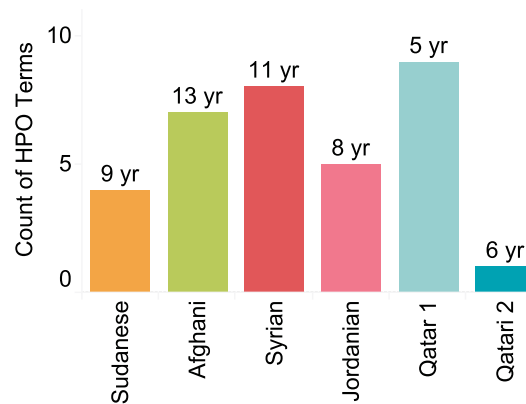
Correspondence and requests for materials should be addressed to Younes Mokrab.

Peer review information *Nature Genetics* thanks Mohamed Elmonem, Jue Ruan and Vagheesh Narasimhan for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.



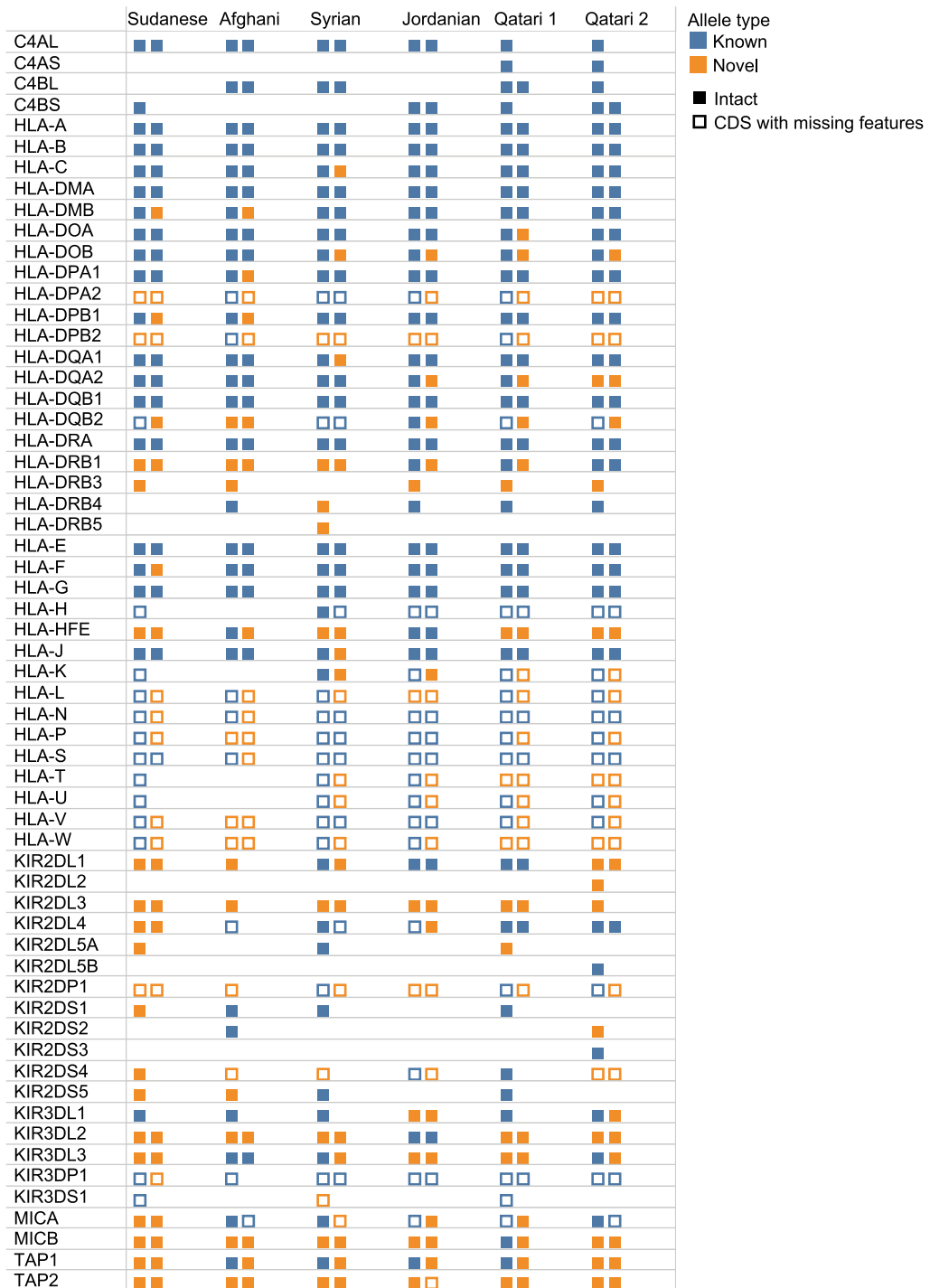
Extended Data Fig. 1 | Study overview. The displayed map is from Mapbox and OpenStreetMap, used under the Open Database License (ODbL).

a**b**

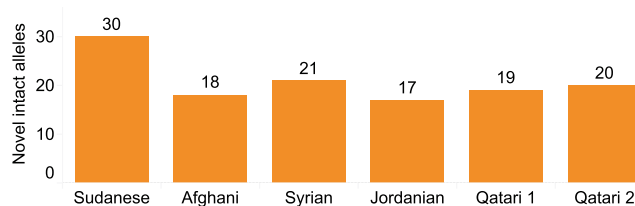
Child	HPO Term	HPO code	HPO Term ID	
Sudanese	Cystic renal dysplasia	CRD	HP:0000800	
	Global developmental delay	GDD	HP:0001263	
	Multiple renal cysts	MRC	HP:0005562	
	Proteinuria	Pr	HP:0000093	
Afghani	Cerebral palsy	CP	HP:0100021	
	Dysphagia	Df	HP:0002015	
	Gastroesophageal reflux	Ga	HP:0002020	
	Global developmental delay	GDD	HP:0001263	
	Intellectual disability	ID	HP:0001249	
	Polymicrogyria	Po	HP:0002126	
	Seizure	S	HP:0001250	
	Syrian	Abnormal heart morphology	AHM	HP:0001627
		Cholelithiasis	Ch	HP:0001081
Delayed speech and language development		DSL	HP:0000750	
Duplicated collecting system		DCS	HP:0000081	
Dysuria		Ds	HP:0100518	
Gait disturbance		GD	HP:0001288	
Intellectual disability		ID	HP:0001249	
Motor delay		MD	HP:0001270	
Jordanian	Autism	Au	HP:0000717	
	Glaucoma	GI	HP:0000501	
	Increased intraocular pressure	IIP	HP:0007906	
	Nevus flammeus	N	HP:0001052	
	Seizure	S	HP:0001250	
Qatar 1	Alternating exotropia	AE	HP:0031717	
	Astigmatism	As	HP:0000483	
	Concomitant strabismus	Cs	HP:0025069	
	Exophoria	E	HP:0025313	
	Global developmental delay	GDD	HP:0001263	
	Heterophoria	Het	HP:0032011	
	Hypermetropia	Hyp	HP:0000540	
	Tall stature	T	HP:0000098	
Qatari 2	Autism	Au	HP:0000717	

Extended Data Fig. 2 | Clinical phenotypes of the children from the family trios in this study. (a) Age and count of Human Phenotype Ontology (HPO) terms per child. **(b)** Detailed list of HPO terms and HPO IDs.

a

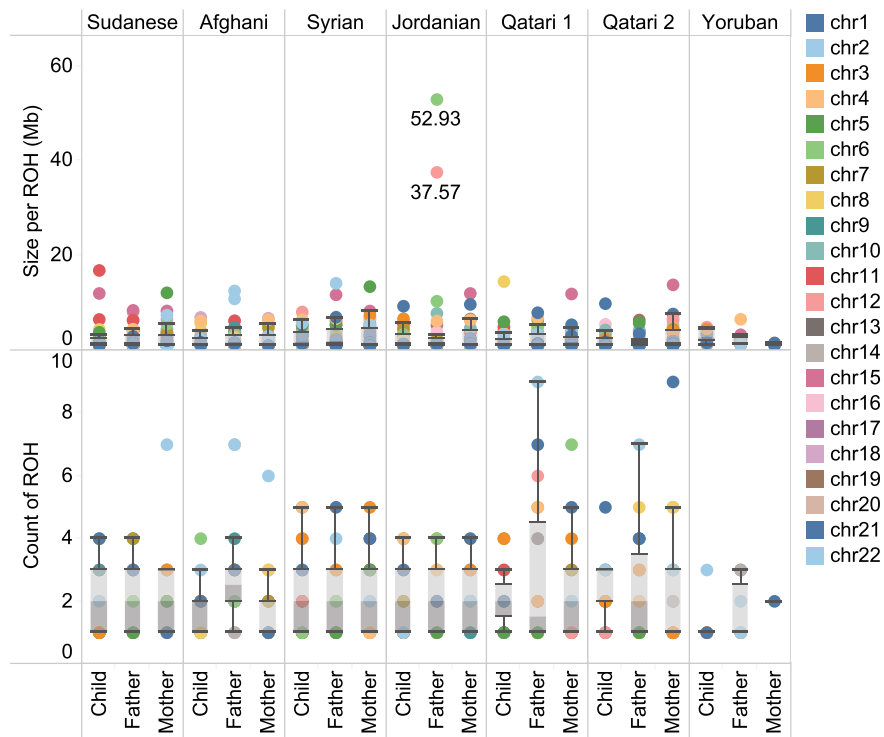


b

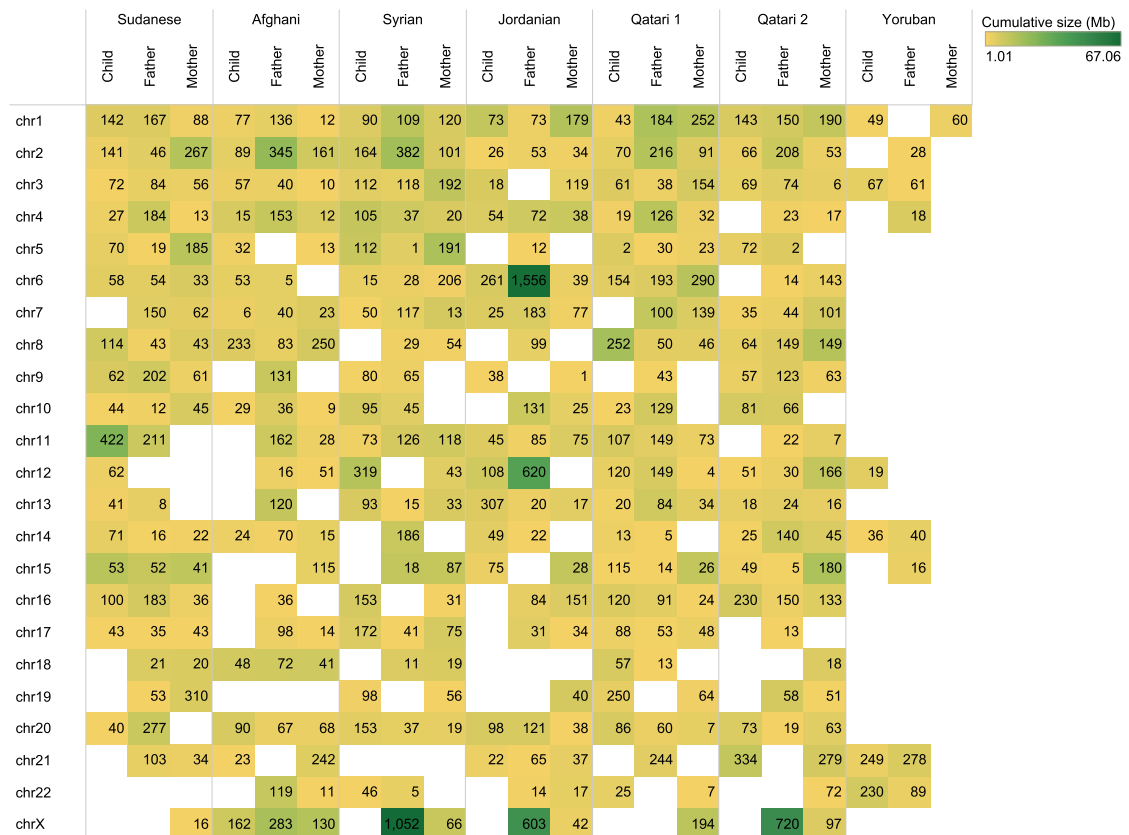


Extended Data Fig. 3 | HLA and KIR gene annotations. a. Known and new alleles in HLA and KIR genes in the child assemblies' haplotypes, highlighting intact sequences and those having CDS with missing features. **b.** Count of new alleles with intact CDS regions in each child assembly.

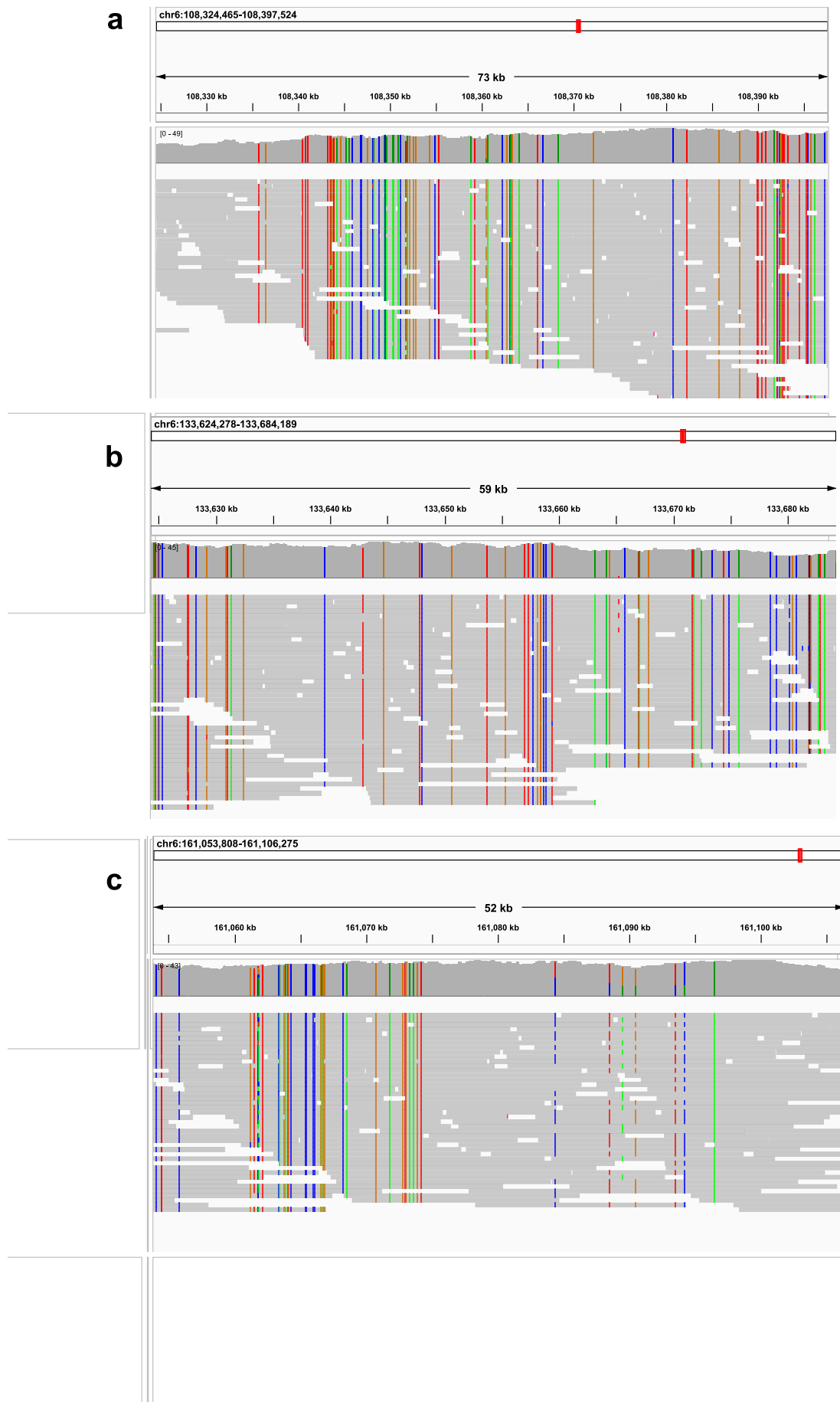
a



b



Extended Data Fig. 4 | Runs of homozygosity (ROH) per chromosome. a, Size and count of ROH across chromosomes in all participants compared to YRI trios. **b**, Heatmap of ROH by chromosome and size, with color intensities reflecting the cumulative ROH size on each chromosome.



Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Long ROH in chr6 of the Jordanian father. a–c, IGV visualization of the HiFi read alignments against CHM13 showing the (a) start, (b) middle and (c) the end of the ROH region, revealing uniform homozygous region and introduction of heterozygous sites at the end of ROH.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	ILLUMINA HiSeq X Software v3, SMRT Link v10.1
Data analysis	SAMtools v1.17, mosdepth v0.3.2, FastQC v0.11.9, MultiQC v1.21, SMRTlink v. 10.1, trioHiFiiasm 0.16.1, BBmap v38.69, Merqury v1.3, seqtk v1.4, Flagger v0.3.3, Minimap2 v2.9, Saffire v0.2, rustybam v0.1.31, pbmm2 v1.13.0, DeepVariant v1.5.0, PAV v2.3.3, BWA-MEM v 0.7.15, GATK v4.0.9.0, bcftools v1.17, PLINK v2.00a2LM 64-bit Intel (7 Jan 2019), RFMix v2.0.0, EAGLE v2.4.1, truvari v4.1.0, Liftoff v1.6.3, Immuannot (feb2024), Clustal Omega v1.2.4, TreeViewer v2.2.0, Automap v1.2, IGV v2.9.4, awk v4.0.2, bedtools v2.3, Jalview 2.11.3, Exomiser v14, DRAGEN 4.2. annotSV v3.4.2, Jannovar v0.41, yak v0.1

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

The assemblies and sequencing data generated in this study is deposited to dbGAP (Accession ID phs003917.v1.p1). Previously published data used in the study is accessible as follows: 1000 Genomes Project <https://hgdownload.cse.ucsc.edu/gbdb/hg19/1000Genomes/phase3>, Qatar Biobank/Qatar Genome Project Accession ID: QF-QGP-RES-PUB-007 <https://www.qatarbiobank.org.qa/research/how-apply>, CHM13 assembly https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/analysis_set/chm13v2.0.fa.gz, CHM13 annotation https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/009/914/755/GCF_009914755.1_T2T-CHM13v2.0/GCF_009914755.1_T2T-CHM13v2.0_genomic.gtf.gz, https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/annotation/chm13v2.0_RepeatMasker_4.1.2p1.2022Apr14.bed, https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/annotation/chm13v2.0_SD.full.bed, https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/annotation/chm13v2.0_censat_v2.1.bed, GRCh38 assembly <https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/p13/hg38.p13.chromFa.tar.gz>, <https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/HG002/assemblies/hg002v1.0.1.fasta.gz>, immuannot IPD/KIR dataset <https://zenodo.org/records/8372992/files/Data-2023Oct27.tar.gz?download=1>, CIWD v3.0 catalogue <https://www.ihw18.org/component-immunogenetics/download-common-and-well-documented-alleles-3-0/>, gnomad v4.1.0 <https://gnomad.broadinstitute.org>, CN1 assembly https://genome.zju.edu.cn/files/v1.0.1/CN1_pat.v1.0.1.fasta.gz, https://genome.zju.edu.cn/files/v1.0.1/CN1_mat.v1.0.1.fasta.gz, https://storage.googleapis.com/gcp-public-data--gnomad/release/4.0/constraint/gnomad.v4.0.constraint_metrics.tsv

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

Sex of the study subjects was reported as commonly practiced in genomic research

Reporting on race, ethnicity, or other socially relevant groupings

Nationality and genetic ancestry based on admixture analysis and principal component analysis was reported.

Population characteristics

The main study subjects consist of six parent-child family trios (n=18) from the local population in Qatar whereby children were ascertained for neuro developmental disorders and their genetic diagnosis was unresolved. The families are from various nationalities from the greater Middle Eastern region: Sudan, Jordan, Syria, Qatar and Afghanistan. Furthermore, previously published datasets from Qatar Genome Project/Qatar Biobank and 1000 Genomes project and internal Middle Eastern subjects were used. The former consists of Predominantly healthy individuals recruited from the general population at Qatar, aged 18 or over and nationals. 56.7 % of the participants were females whereas 43.3% were males. The latter consists of participants from 26 world populations from the five world continents.

Recruitment

The six families were recruited at Sidra Medicine. Any bias in the selection of the families does not impact the results because each family is analyzed as an independent unit.

Ethics oversight

Informed consent and assent were obtained for all subjects and the study was approved by The Sidra Medicine Ethics Committee.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

18 subjects were recruited in this study. As comparison reference datasets, we used phase 1 of Qatar Genome Project consisting of 6,216 subjects, 2,504 subjects from 1000 Genomes project phase 3 and 1693 Middle Eastern subjects from internal dataset at Sidra Medicine. No statistical method was used to predetermine the number of the families because it is a family based study to do genome assembly not a cohort based. Our sample sizes are typical of what is used in the field.

Data exclusions

Standard thresholds for quality measures were used when running various tools for data processing and downstream analyses as mentioned in the manuscript, unless indicated otherwise.

Replication	Experiments were computational so replication is not applicable. For reproducibility describe all of the data sets codes/workflows in the manuscript.
Randomization	Randomization is not applicable to this study because we did not perform any experiments with treatment or control groups that would necessitate randomization between the subjects
Blinding	Blinding is not applicable to this study because we did not perform any experiments with treatment or control groups that would necessitate blinding

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Included in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern
<input checked="" type="checkbox"/>	<input type="checkbox"/> Plants

Methods

n/a	Included in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration	N/A
Study protocol	N/A
Data collection	N/A
Outcomes	N/A

Plants

Seed stocks	<i>Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.</i>
Novel plant genotypes	<i>Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.</i>
Authentication	<i>Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.</i>