nature communications

Article

https://doi.org/10.1038/s41467-026-73461-x

Accessing medically relevant complex regions with a pangenome graph of 20 near-complete Japanese haplotypes

Received: 27 January 2025

Accepted: 11 May 2026

Published online: 27 May 2026

Check for updates

Yoshihiko Suzuki[1,3] ✉, Chie Owa[1], Haruka Kobayashi[1], Ryo Nakabayashi[1], Brandy McNulty[2], Ivo Violich[2], Benedict Paten[2], Karen H. Miga[2] & Shinichi Morishita[1,3] ✉

Pangenome projects have enhanced our understanding of human genomic and genetic diversity, but repetitive regions are still challenging to assemble and yet medically important. Here we generate 20 near-complete haplotypes from 10 Japanese male individuals using three complementary long-read and long-range datasets and construct a pangenome graph from these haplotype-resolved assemblies. All haplotypes achieve an N50 value exceeding 100 Mbp for gapless contigs. We substantially improve the average reconstruction rate of complete haplotypes from 46.8% and 52.8% in two previous pangenome graphs to 91.2% within 30 segmentally duplicated complex regions. Furthermore, we identify complete minor haplotypes in the *KIR* and *SMN* regions that are absent in those previous pangenome graphs. We find putatively biased gene conversion events occurring in only one direction around the *SMN* and beta-defensin (*DEFB*) genes, implying non-random evolution in these regions. Our study contributes to the growing body of pangenomic data, offering a more refined view of human genomic diversity involving complex segmental duplications.

A complete reference human genome sequence was released in 2022 by the Telomere-to-Telomere (T2T) Consortium[1]. This ultimate degree of human genome completeness has both helped reveal the full spectrum of genomic diversity and led to a race to create additional complete or near-complete genomes using long reads. Groups including the 1000 Genomes Project (1KGP)[2,3], the Human Genome Structural Variation Consortium (HGSVC)[4–6], and the Human Pangenome Reference Consortium (HPRC)[7], as well as the Human Pangenome Project (HPP)[8]—an international partnership community encompassing the HPRC—have expanded genomic resources to improve coverage of the global population. A pangenome graph is a reference representation in which genome sequences from multiple individuals are encoded as paths through a graph, allowing shared sequences and genetic or structural variants to be represented in a single framework. In 2023, the HPRC released a Year-1 draft human pangenome graph reference that incorporates 47 phased, diploid assemblies from genetically diverse individuals[7]. This dataset contains eight East Asian haplotypes, including six Southern Han Chinese (CHS) and two Kinh Vietnamese (KHV) haplotypes, but no Japanese haplotypes were included. Another recent long-read pangenome project for East Asian populations is the Chinese Pangenome Project (CPC)[9], which sequenced 58 samples from 36 populations, including minority populations in China.

Most of the haplotypes in the HPRC Year-1 pangenome and the Chinese pangenome were constructed from PacBio HiFi long reads together with trio Illumina short reads or Hi-C reads. HiFi reads enabled accurate haplotype-resolved sequences for most genomic regions[10–12], but highly repetitive regions such as centromeres and

[1]Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Chiba, Japan. [2]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, CA, USA. [3]Present address: Research Organization of Information and Systems (ROIS), Tokyo, Japan. ✉e-mail: yszkshk@gmail.com; moris@edu.k.u-tokyo.ac.jp

large segmental duplications remained unassembled in a complete form. As a result, the average contig NG50 metrics for the haplotypes in these pangenome projects was limited to 40 Mbp and 36 Mbp, respectively. More recently, a hybrid approach combining the two complementary long-read data types of PacBio HiFi reads, which have an average accuracy greater than 99.9%, and Oxford Nanopore Technology (ONT) ultra-long reads of >100 kbp length, along with supplementary short reads, has been verified to be effective for reconstructing near-complete diploid assemblies[13] through algorithmic advancements within the verkko[14] and hifiasm[15] assemblers. In this study, we deep-sequenced ONT ultra-long reads, as well as PacBio HiFi and Omni-C reads, for ten Japanese samples, achieving the contig NG50 value of >100 Mbp for every assembled Japanese haplotype.

Many clinically important genomic regions are structurally complex and exhibit high sequence diversity, partly due to rapid evolution mediated by large segmental duplications subject to nonallelic homologous recombination and gene conversions[16–18]. Notable examples include the Human Leukocyte Antigen (*HLA*) and Killer-cell Immunoglobulin-like Receptor (*KIR*) regions, which are related to immune system function[19–21], and the Survival Motor Neuron (*SMN*) region associated with spinal muscular atrophy (SMA)[22–24]. Another example is the *DEFB* (beta-defensin) gene cluster, whose increase and decrease in copy number are associated with psoriasis and Crohn's disease, respectively[25–28]. Moreover, using large cohort datasets from the UK BioBank[29] and BioBank Japan[30], associations have been found between the *RASA4* gene cluster encoding the Ras GTPase-activating protein 4 on chromosome 7 and type II diabetes and chronotype, and between the autoimmunity-related *FCGR3* (Fc Gamma Receptor III) gene on chromosome 1 and the count of basophils in the blood[31,32]. Despite their recognized importance, these regions have remained technically challenging for complete sequence reconstruction, impeding more detailed analysis, although the demand for complete haplotypes, including novel ones, continues to grow[7,18,33,34]. Here we present near-complete diploid assemblies for ten Japanese individuals, aiming to capture more complete sequences of medically important complex regions in the Japanese population. Our analysis of the 20 assembled haplotypes demonstrates a substantial improvement in assembly completeness compared to previous pangenome projects and reveals novel haplotype structures and putatively biased gene conversions, including those known to be medically relevant regions.

Results

Reconstruction of 20 near-complete Japanese haplotypes and construction of a Japanese pangenome

We sequenced PacBio HiFi, ONT UL, and Omni-C reads from ten healthy Japanese lymphoblastoid cell lines. All samples were those collected in the 1000 Genomes Project as Japanese in Tokyo (JPT) samples and distributed by the Coriell Institute[2,3]. We obtained 35–55× HiFi reads, 50–90× Omni-C reads, and 45–65× ONT UL reads (only >100 kbp) per sample (Supplementary Table 1). On top of these reads, we collected additional 25–35× ONT UL reads used for validation of assembly (Supplementary Table 2). All the ten samples are male, and thus we could obtain ten haplotype sequences for each of chromosome X and Y as well as the mitochondrial genome, in addition to 20 haplotypes for each of the autosomal chromosomes.

We reconstructed a chromosome-scale, near-complete diploid assembly for each of the ten samples using hifiasm[15] and RagTag[35], along with manual curation performed mainly for correcting haplotype misplacement of some relatively short contigs in extremely heterochromatic regions of chromosome Y ("Methods" and Supplementary Table 3). By virtue of the high-coverage dataset of ONT UL reads longer than 100 kbp and up to 1.8 Mbp, an average of 63.7 Mbp of segmental duplication regions per haplotype were resolved without gaps. The NG50 values of the contigs alone (excluding scaffolds containing sequence gaps) were greater than 100 Mbp in all the 20 Japanese haplotypes, and the mean contig auN (area under the Nx curve) value per haplotype was 127.8 Mbp (Fig. 1a). On average, 7.1 chromosomes were gapless, complete sequences spanning from telomere to telomere per haplotype among the 20 haplotypes. Every assembled haplotype showed an assembly quality value of >60 based on the k-mer completeness calculated with Merqury[36] (Fig. 1b). Assembly statistics are summarized in Supplementary Table 4.

The assembly size per chromosome was generally consistent with the T2T-CHM13 reference genome, with several exceptions (Fig. 1c). Chromosomes containing large HSat2 and HSat3 arrays—chromosomes 1, 9, 16, and Y as well as the acrocentric chromosomes (13, 14, 15, 21, and 22)—exhibited greater assembly size variability, with standard deviations exceeding 2 Mbp across Japanese assemblies (Fig. 1d and Supplementary Fig. 1a). Among these chromosomes with large size variability, we observed reduced assembly sizes more than one standard deviation than the T2T-CHM13 assembly size for chromosomes 9, 13, 15, 16, and Y in Japanese haplotypes (Supplementary Fig. 1b). For chromosomes 1, 9, 16, and Y, the differences in total chromosome size compared to T2T-CHM13 were largely explained by differences in the lengths of alpha satellite and HSat1/2/3 arrays[37,38], as indicated by regression slopes exceeding 0.8 between total chromosome size differences and array length differences (Supplementary Fig. 1c). Note that both the acrocentric chromosomes and chromosome Y showed relatively low T2T assembly rates below 30% (Supplementary Table 4), implying that assembly difficulties also contribute to the assembly size variability observed. In addition, a known huge recurrent inversion of ~4 Mbp in the p-arm of chromosome 8 was identified in 15 (75%) out of 20 haplotypes (Fig. 1e and Supplementary Data 1)[39–41]. This inversion is located around polymorphic gene families such as *DEFB* (beta-defensin) and *FAM90A* (primate-specific family with sequence similarity 90), previously difficult to assemble.

Segmental duplications (SDs) were identified using Biser[42] after masking repeats with RepeatMasker (http://www.repeatmasker.org) in each Japanese haplotype, and these SDs were lifted over to the T2T-CHM13 genome to enable comparison with other haplotypes (Fig. 1f and "Methods"). We merged the lifted-over Japanese SD regions and excluded those within centromeric satellite regions, as these regions are difficult to compare across different haplotypes. This resulted in a total of 46.3 Mbp of SD regions, of which 44. 2 Mbp (95%) showed complete or partial overlap with SDs annotated in T2T-CHM13, while the remaining 2.1 Mbp (5%) did not. Within the Japanese haplotypes, 40.6 Mbp (88%) of the lifted-over SD regions were found in two or more haplotypes, suggesting that these SDs are common in the Japanese population. It should be noted, however, that confirming their recurrence at a population scale will require a larger cohort study. The majority of SDs identified in the Japanese haplotypes (42.3 Mbp, 91%) completely or partially overlapped with SD regions in the non-Japanese East Asian (EAS) haplotypes of the HPRC Year-1 pangenome, and an additional 2.2 Mbp (5%) overlapped with those in the haplotypes of the Chinese pangenome. Among the remaining non-overlapping, Japanese-specific SD regions (1.8 Mbp, 4%), only 54 kbp (0.1%) overlapped with SDs in non-EAS haplotypes of HPRC Year-1 pangenome. Although 154 out of the 155 Japanese-specific SD regions did not contain protein-coding genes, one region on chromosome 19 present in a single haplotype contained four protein-coding genes, *MAP3K10*, *TTC9B*, *CCNP*, and *AKT2* (Supplementary Data 2).

To build and understand the distribution of SNVs and SVs unique to individual populations, we constructed two types of pangenome graph using minigraph-cactus[43] and pggb[44] from the 20 near-complete Japanese haplotypes plus two haplotypes of T2T-CHM13 and GRCh38 ("Methods"). SNVs and SVs, defined as variants of >50 bp, for each haplotype were identified from the pangenome graph by detecting branching nodes that diverge from the reference genome path of either T2T-CHM13 or GRCh38. To reduce the redundancy of detected SVs, we performed base-level normalization of SVs called from the

https://doi.org/10.1038/s41467-026-73461-x

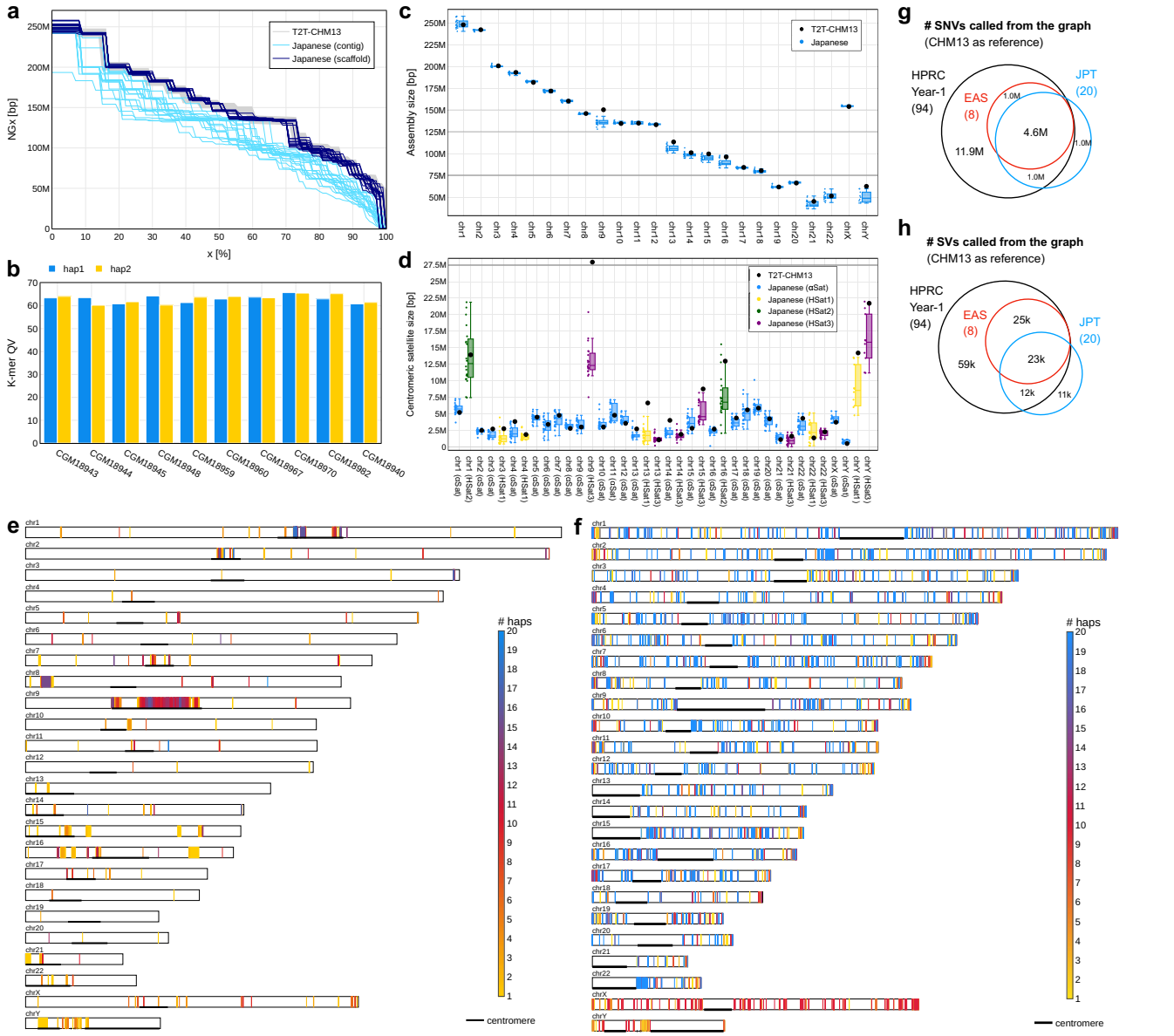

Fig. 1 | Assembly status of the 20 near-complete Japanese haplotypes. **a**, Contiguity of contigs and scaffolds of 20 haplotypes assembled from 10 Japanese samples (Japanese in Tokyo, JPT). T2T-CHM13 is shown in thick light gray as a practical upper bound. **b**, Quality values of the assembled haplotypes. **c**, Assembly size per chromosome among assembled haplotypes. **d**, Total size of large centromeric satellite arrays (alpha satellites and HSat1/2/3) per chromosome. For HSat1/2/3, only chromosomes with >1 Mbp arrays in T2T-CHM13 are shown. In **c**,**d**, box plots show assembled haplotypes from 10 biological samples ($n = 20$ for autosomes and $n = 10$ for sex chromosomes); boxes indicate the median and 25th–75th percentiles, whiskers extend to 1.5 × IQR, and all points are shown. **e**–**f**, Distributions of inversions (**e**) and SDs (**f**) found in the 20 Japanese haplotypes, plotted together on the corresponding regions in T2T-CHM13. SDs found within centromeres are excluded due to its highly repetitive nature. Colors indicate the number of Japanese haplotypes in which the region is in an inversion or SD. **g**, **h**, Comparison between the numbers of SNVs (**g**) and SVs (**h**) identified from the pangenome graph of the 20 Japanese haplotypes and those from the HPRC Year-1 graph. T2T-CHM13 was used as the reference for variant calling, and variants on chromosome Y and private SVs with allele count 1 were excluded. Numbers in parentheses indicate haplotype counts; the eight HPRC Year-1 East Asian (EAS) haplotypes comprise six Southern Han Chinese (CHS) and two Kinh Vietnamese (KHV) haplotypes.

pangenome graph using vcfbub[45] and vcfwave[46]. Each method yielded a comparable pattern of identified SVs, with peaks in the SV length distribution corresponding to insertions and deletions of the *Alu* element of ~340 bp and LINE elements of ~6 kbp (Supplementary Fig. 2b,c), although the total number of SVs differed between minigraph-cactus (127k) and pggb (115k). Notably, 96% of the difference in SV counts between the two methods (12k) was attributable to centromeric regions, where minigraph-cactus reported more SVs (2k) than pggb (14k). Outside centromeres, SVs identified by the two methods showed a 74% F1 score using Truvari[47]. For subsequent analyses, we used the pggb graph containing fewer unique SVs, and excluded variants located within centromeres.

Among the SNVs reported in the 1KGP[6], which were called using Illumina short reads aligned to the GRCh38 reference genome, 97% were also identified in our graph-based analysis for the same ten Japanese samples. We next compared SNVs identified in the Japanese graph with those called from the HPRC Year-1 graph using the same T2T-CHM13 reference genome. In the Japanese pggb graph, we identified 6,659,402 SNVs relative to T2T-CHM13. Of these, 5,637,761 SNVs (85%) were shared with the HPRC Year-1 graph, primarily with the eight

non-Japanese EAS haplotypes (4,590,109 SNVs, 69%), whereas 1,021,641 SNVs (15%) were not observed in the HPRC Year-1 haplotypes and were therefore classified as Japanese-specific in this comparison (Fig. 1g)[7]. Among the 6,659,402 Japanese SNVs, 5,064,963 SNVs (76%) were present in two or more Japanese haplotypes, suggesting their commonality in the Japanese population. Of these 5,064,963 SNVs observed in multiple Japanese haplotypes, 4,817,785 SNVs (95%) were also found in haplotypes of other populations within the HPRC dataset, whereas only 247,178 SNVs (5%) were found specific to the Japanese haplotypes.

For SV analysis, we first merged similar SVs across all haplotypes using Truvari with 80% sequence overlap, and then retained only SVs detected in two or more haplotypes within either Japanese or HPRC haplotypes. Among 46,173 SVs (11.0 Mbp in total) identified in the Japanese haplotypes, 35,168 SVs (76%, 8.0 Mbp) were also observed in the HPRC Year-1 graph, primarily in EAS haplotypes, whereas 11,005 SVs (24%, 3.0 Mbp) were not observed in the HPRC Year-1 haplotypes (Fig. 1h). However, the majority of these 11,005 SVs were located in complex genomic regions, where it is difficult to distinguish actual uniqueness from assembly artifacts or representation differences. Specifically, among these 11,005 SVs, 2064 SVs (19%) were located in subtelomeric regions, which comprise only 1.5% of the genome, and 4520 SVs (41%) were located in segmental duplications plus 100-kbp flanking regions, which occupy only 6% of the genome (Supplementary Fig. 3a). Moreover, of the remaining 4421 (40%) SVs not in subtelomeric regions nor segmental duplications, 4040 SVs (91%) contained tandem repeat sequences.

To further characterize these 11,005 SVs (5437 insertions and 5568 deletions) observed only in the Japanese haplotypes, we aligned their inserted sequences to the pangenome graph of the Chinese and HPRC haplotypes and confirmed that 4666 (86%) of the 5437 insertions exhibited high (>95%) sequence similarity, as expected for SVs that typically arise from duplications of existing sequences, whereas 771 (14%) showed <95% sequence similarity (Supplementary Fig. 3b).

Gene copy number analysis

We then verified the quality of our assemblies at the gene level. On average, 98.3% of the single-copy genes in GRCh38 with alternative sequences were fully reconstructed in the Japanese haplotypes using asmgene[48] (Supplementary Table 4), which was comparable to the values of the East Asian haplotypes in the HPRC Year-1 dataset (98.3%) and the T2T-CHM13 reference genome (98.5%). In contrast to single-copy genes, large segmental duplications frequently contribute to variations in gene copy number and increased gene diversity[16,17]. While Japanese haplotypes are expected to generally share similar gene copy number distributions with non-Japanese East Asian haplotypes, if Japanese haplotypes have large discrepancies from the reference genome especially in complex regions, the difference is likely to cause misassignments of reads and assemblies. Such a reference bias often confounds population-specific analyses of polymorphic genes. To assess the extent of these differences, we conducted a comparative analysis of gene copy numbers between the T2T-CHM13, non-Japanese East Asian haplotypes in the HPRC dataset, and our Japanese haplotypes. We calculated the copy number for each gene by lifting over the GENCODE Comprehensive annotation[49], which includes genes in alternate loci of complex regions, to each of the haplotypes. The percentage of protein-coding genes whose copy number was identical to T2T-CHM13 averaged 96.7% across the 20 Japanese haplotypes and 96.6% across the East Asian haplotypes. These results support the high quality of both the HPRC and our assemblies in general.

We investigated genes whose copy numbers were consistently different from T2T-CHM13 in all the Japanese haplotypes, and there existed 23 such genes (Fig. 2a). Although accurate annotation of paralogous genes in complex regions is difficult, all of those 23 genes were either gene families in segmental duplications or genes within alternate loci of complex regions whose names start with ENSG, implying the presence of systematic sequence variants in these regions in the Japanese population compared with the reference genome. These gene families include *KIR2DL2*, a member of the *KIR* gene family that is known to be rare in the Japanese population, whereas T2T-CHM13 contains one copy[21]. Other gene families on autosomal chromosomes include *DEFB* and *FAM90A* near the large recurrent inversion on chromosome 8, *ELOA3* (Elongin A3, a primate-specific RNA polymerase II elongation factor[50]) on chromosome 18, *TBCD* (Tubulin Folding Cofactor D, a candidate causal gene of an atypical infantile SMA accompanying progressive neurodegenerative encephalopathy[51]) and *B3DNTL1* (Beta-1,3-N-Acetylglucosaminyl-transferase-like 1) on chromosome 17, and *SPDYE8* near *RASA4* on chromosome 7. In the next section, we will show that most of the Japanese haplotypes were completely assembled in these complex medically important regions, while many of the East Asian haplotypes in the HPRC Year-1 and Chinese pangenomes were not. On sex chromosomes, cancer/testis antigen families of *GAGE* and *CT47* on chromosome X, testis-specific genes of *TSPY* and *MAFIP* on chromosome Y, and *CSF2RA* (granulocyte/macrophage colony-stimulating factor alpha) gene on the pseudoautosomal region (PAR) had consistently different copy number annotations among the Japanese haplotypes than T2T-CHM13.

We then compared the Japanese haplotypes with the East Asian haplotypes in the HPRC pangenome rather than T2T-CHM13. There were 43 genes with different copy numbers between those two datasets (two-sided Wilcoxon rank-sum test, $P < 0.05$; Supplementary Table 5). We found that several genes located in segmentally duplicated regions showed an average copy number difference greater than 1 between the East Asian and Japanese haplotypes, specifically within the *TSPY* and *RBMY* gene families (Fig. 2b). Considering that complete assembly was challenging in the HPRC Year-1 haplotypes for these segmentally duplicated complex regions, although their assemblies are of high quality in relatively unique regions, the observed copy number differences may be due to assembly artifacts rather than true biological variation. Indeed, in the following section, we will demonstrate that our Japanese haplotypes achieved a substantially higher rate of complete sequence reconstruction than previous pangenome projects in such complex yet medically important genomic regions, making them more reliable for accurate downstream analyses.

Complete haplotype reconstruction in complex regions of medically important genes

Based on the literature analyzing challenging complex genomic regions in the assemblies of T2T-CHM13, GRCh38, and previous pangenome efforts[7,9,16–18,50,52–54], we selected 30 genomic regions that contain medically and biologically important genes but are difficult to assemble due to segmental duplications, of which 23 are autosomal, five are in chromosome X, and two in Y (see Supplementary Table 6 and Supplementary Fig. 4 for the coordinates and sequence structure in T2T-CHM13, along with the names of key genes for each region). These include genes such as *FAM90A*, *ELOA3*, *DEFB*, *SPDYE*, *KIR*, *GAGE*, *TSPY,* and *RBMY*, which were noted in our copy number analysis. We examined the sequence contiguity of haplotypes in these medically important regions across the following three pangenomes: the HPRC pangenome of their Year-1 dataset, the Chinese pangenome of the CPC, and our Japanese pangenome (Fig. 2c). Compared to the other two datasets, the Japanese haplotypes had much fewer sequence gaps. Notably, all Japanese haplotypes were completely gapless in 20 regions, whereas at least one haplotype in each of the other datasets had gaps in every region.

Misassemblies can exist even when an assembled sequence has no sequence gaps. In such cases, reads from the original location of one haplotype will not map to the misassembled segment of that location, or reads of a distant location instead may map to that location, making inconsistent nucleotide sequences between the reads and the

https://doi.org/10.1038/s41467-026-73461-x

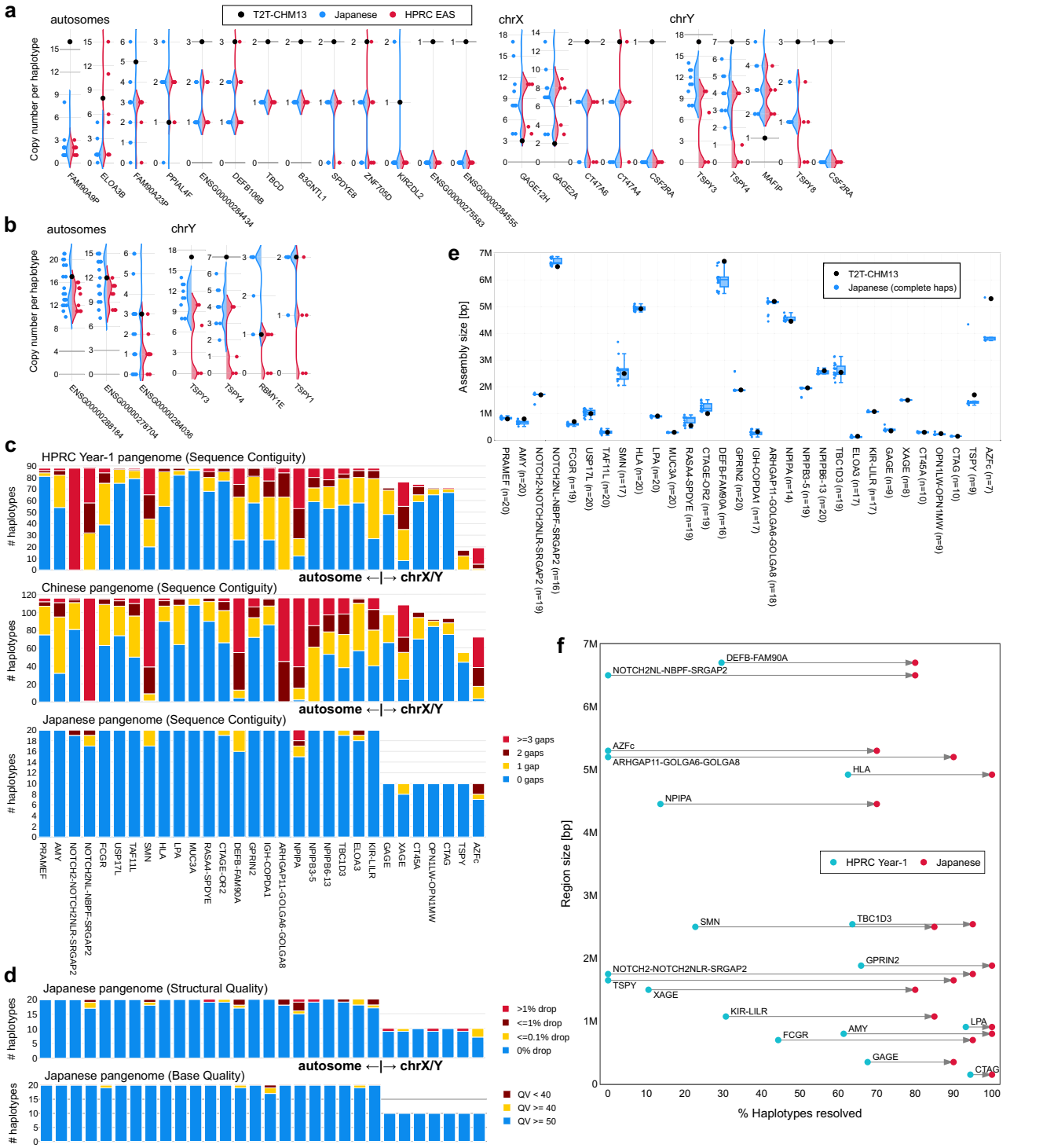

Fig. 2 | Haplotype reconstruction in medically important complex regions. **a**, Copy number distributions of genes differing from T2T-CHM13 (black) in all Japanese haplotypes (JPT; blue), along with those in HPRC Year-1 EAS haplotypes (red). Genes whose copy numbers in T2T-CHM13 are within the 95% confidence interval of the JPT and EAS means were excluded. **b**, Copy number distributions of genes differing by at least one copy between JPT and EAS haplotypes. Genes with differences smaller than the standard deviations of both distributions were excluded. **c**, Sequence contiguity of haplotypes from the HPRC Year-1, Chinese, and Japanese pangenomes across 30 selected medically and biologically important regions. Colors indicate the number of gaps. **d**, Structural and base-quality metrics for the 20 Japanese haplotypes. **e**, Assembly size distribution of complete Japanese haplotypes in the selected regions. Box plots show complete haplotypes from ten biological samples (exact *n* values shown in the plot). Boxes indicate median and 25th–75th percentiles, whiskers extend to 1.5 × IQR, and all points are shown. Outliers, including two *SMN* and one *TSPY* haplotypes, represent large structural rearrangements. **f**, Haplotype reconstruction rates between HPRC Year-1 and Japanese assemblies in 18 representative regions containing challenging medically relevant genes.

assembled haplotype. To detect such deficiencies, along with the basic contiguity metrics, we assessed two additional quality metrics, structural quality and base quality, of the Japanese haplotypes (Fig. 2d). Structural quality is the percentage of regions where sequencing coverage dropped below three, calculated from a pileup of ONT ultra-long reads of the same sample. Base quality is the number of SNVs detected from the read pileup. A high-quality assembly is expected to have no regions with coverage drops and minimal SNVs, indicating inconsistency between reads and the assembly. We accepted up to one SNV per 100 kbp (i.e., QV ≥ 50) to account for systematic errors in sequencing and variant detection, while no sequence gaps and no coverage drops were accepted as the best category, as shown in Fig. 2c, d. In this regard, both metrics were as good as contiguity among the 20 haplotypes, indicating the remarkably high quality of our Japanese assemblies even in segmentally duplicated genomic regions.

We then defined complete haplotypes as haplotypes that belong to the best category in all three criteria above (sequence contiguity, structural quality, and base quality). For 24 (80%) of the complex regions, the size in T2T-CHM13 was not an outlier in the distribution of the complete Japanese haplotypes, indicating overall consistency between Japanese haplotypes and T2T-CHM13 (Fig. 2e). In contrast, we identified considerable size deviations in the remaining six regions: the *NOTCH2-NOTCH2NLR-SRGAP2*, *FCGR*, *DEFB-FAM90A*, *OPN1LW-OPN1MW*, *TSPY*, and *AZFc* regions (Fig. 2e). Our Japanese pangenome achieved an average complete haplotype reconstruction rate of 91.2% across the 30 examined regions. In comparison, the HPRC Year-1 and Chinese pangenomes, when measured solely by sequence contiguity, showed lower average complete reconstruction rates of 52.8% and 46.8%, respectively. The most challenging regions to completely assemble in the Japanese haplotypes were the 5.3 Mbp *AZFc* (Azoospermia Factor c) region on chromosome Y and the 4.5 Mbp *NPIPA* gene family region (Supplementary Fig. 5). Nevertheless, 70% of Japanese haplotypes were completely assembled in these two regions. This is considerable progress from the HPRC and Chinese pangenomes, whose complete assembly rates were 0% and 4.2% for the *AZFc* region and 13.6% and 1.7% for the *NPIPA* region, respectively. The *AZFc* region consists of *RBMY, TSPY, DAZ, BPY2*, and *CDY1* gene families expressed specifically in testis, and deletions in the *AZFc* region can lead to spermatogenic failure[52,55]. The *NPIPA* gene family is a highly duplicated hominoid gene family associated with the nuclear pore complex on chromosome 16[16,56]. In addition to these two regions, our Japanese assemblies showed substantial improvement on complete haplotype reconstruction in complex regions around *NOTCH2, NOTCH2NL, ARHGAP11*, *GOLGA6/8*, and *TSPY* gene families (Fig. 2f). The sizes of the assembled sequences without gaps in these complex regions were comparable between HPRC Year-1 and our assemblies, with no overall tendency toward shorter sequences in Japanese haplotypes (Supplementary Fig. 6). This suggests that these complex regions were successfully assembled in our samples not because of reduced intrinsic sequence complexity but due to advancements in sequencing technologies, particularly ONT UL reads, and assembly algorithms. In brief, we uncovered complete haplotype structures in the Japanese population for many medically relevant complex regions.

Complete haplotypes of medically important regions unidentified in previous pangenome efforts

Among the 20 complete Japanese haplotype sequences of medically important regions along with gene annotations (Supplementary Fig. 7), we found some Japanese haplotypes harboring a gene structure that was previously not identified in the HPRC Year-1 and Chinese haplotypes. One example of such a region is *KIR*, which has two categories, the major A-haplotypes and minor B-haplotypes. Previous studies reported that about 75% of the haplotypes are A-haplotypes in Japan[21,57], while minor B-haplotypes are defined by the presence of at least one copy of the *KIR2DL5*, *KIR2DS1*, *KIR2DS2*, *KIR2DS3*, *KIR2DS5*, or *KIR3DS1* genes. We performed annotation of the detailed and accurate *KIR* genes in the Japanese haplotypes using Immuannot[58], which is based on the public IPD-KIR database comprehensively cataloging *KIR* genes[19]. Of the 20 Japanese haplotypes we assembled, 16 (80%) were A-haplotypes and the remaining four were B-haplotypes (Supplementary Fig. 8). The whole *KIR* gene array is often divided into two subarrays, the centromeric *KIR* array and telomeric *KIR* array, with a small gap region containing no genes between them. Three of the four B-haplotypes had multiple B-haplotype-specific genes in the telomeric *KIR* array, while the remaining one was classified into B-haplotype due to the presence of one gene, *KIR2DS2*, in the centromeric *KIR* array. In total, all the six genes that define B-haplotypes appeared at least once in the 20 haplotypes, demonstrating that our Japanese pangenome catalogs minor B-haplotypic *KIR* genes as well.

More importantly, we found one B-haplotype that has a unique tandem duplication of an array of three *KIR* genes (Fig. 3a and Supplementary Fig. 8), and this haplotype was not observed in the analysis of the HPRC Year-1 and Chinese pangenomes[58]. The duplicated segment of this haplotype contains two copies of the *KIR2DL5B*, *KIR2DS5*, and *KIR2DS1* genes. We confirmed that in the novel haplotype there are no indications of misassemblies in the entire *KIR* gene array, including the expanded part, based on manual inspection of ONT ultra-long reads mapped to the assembled haplotype and by checking read depths, soft-clipped read mappings, and nucleotide consistency between the assembly and the reads (Fig. 3b). Furthermore, we found two ONT ultra-long reads that span the duplicated segment of the haplotype and support the assembled sequence (Supplementary Fig. 9).

Another important example of a more refined view of complex regions is *SMN*. In the *SMN* gene cluster region, a single-nucleotide variant, c.840C>T, located in exon 7 defines the characteristics of transcribed mRNA and distinguishes the two paralogous copies of *SMN*, *SMN1,* and *SMN2*, where *SMN1* is associated with a higher risk of SMA[59,60]. An increased copy number of *SMN2* reduces the severity of SMA, and gene conversion from *SMN1* to *SMN2* can result in a modifier gene that alters the symptoms in some patients[24,61,62]. Approximately 90% of the individuals have exactly one copy of *SMN1* gene in all populations except African populations, whereas the copy number of *SMN2* gene is more variable[23]. There were two Japanese complete haplotypes that possess three copies of *SMN* genes (Fig. 3c), whereas no haplotypes in HPRC and Chinese pangenomes contain three *SMN* gene copies. Since liftover programs do not consider the difference at a specific single nucleotide as an important and discriminative marker between paralogous genes, we manually distinguished *SMN1* and *SMN2* based on the functionally discriminative single-nucleotide variant on exon 7, c.840C>T[24]. As a result, we found that both of the two Japanese haplotypes contained one *SMN1* gene and two *SMN2* genes. We did not identify any haplotypes with multiple copies of *SMN1*, since all samples in this study were healthy controls. The largest *SMN* haplotype in the Japanese pangenome spans a 2.5 Mbp region featuring a complex pattern of inversions and tandem segmental duplications; nonetheless, the assembly of this region was confirmed to be accurate in the same manner as the *KIR* region, using two ONT ultra-long read datasets (Fig. 3d).

Note that the HGSVC has recently reported their project phase 3, analyzing 65 near-complete diploid assemblies of worldwide individuals[63]. Their assemblies achieved a level of haplotype completeness comparable to ours. For instance, in the *SMN* region, they reported 64% completeness, while we estimated 85% for our Japanese haplotypes using our definition of completeness based on sequence contiguity, structural quality, and base quality. HGSVC3 evaluated assembly quality by integrating multiple orthogonal tools, including Flagger[7], Merqury[36], NucFreq[64], and Inspector[65] that collectively assess both structural and base-level errors, requiring that at least two tools consider the assembly reliable (see HGSVC3 Supplementary Methods).

https://doi.org/10.1038/s41467-026-73461-x

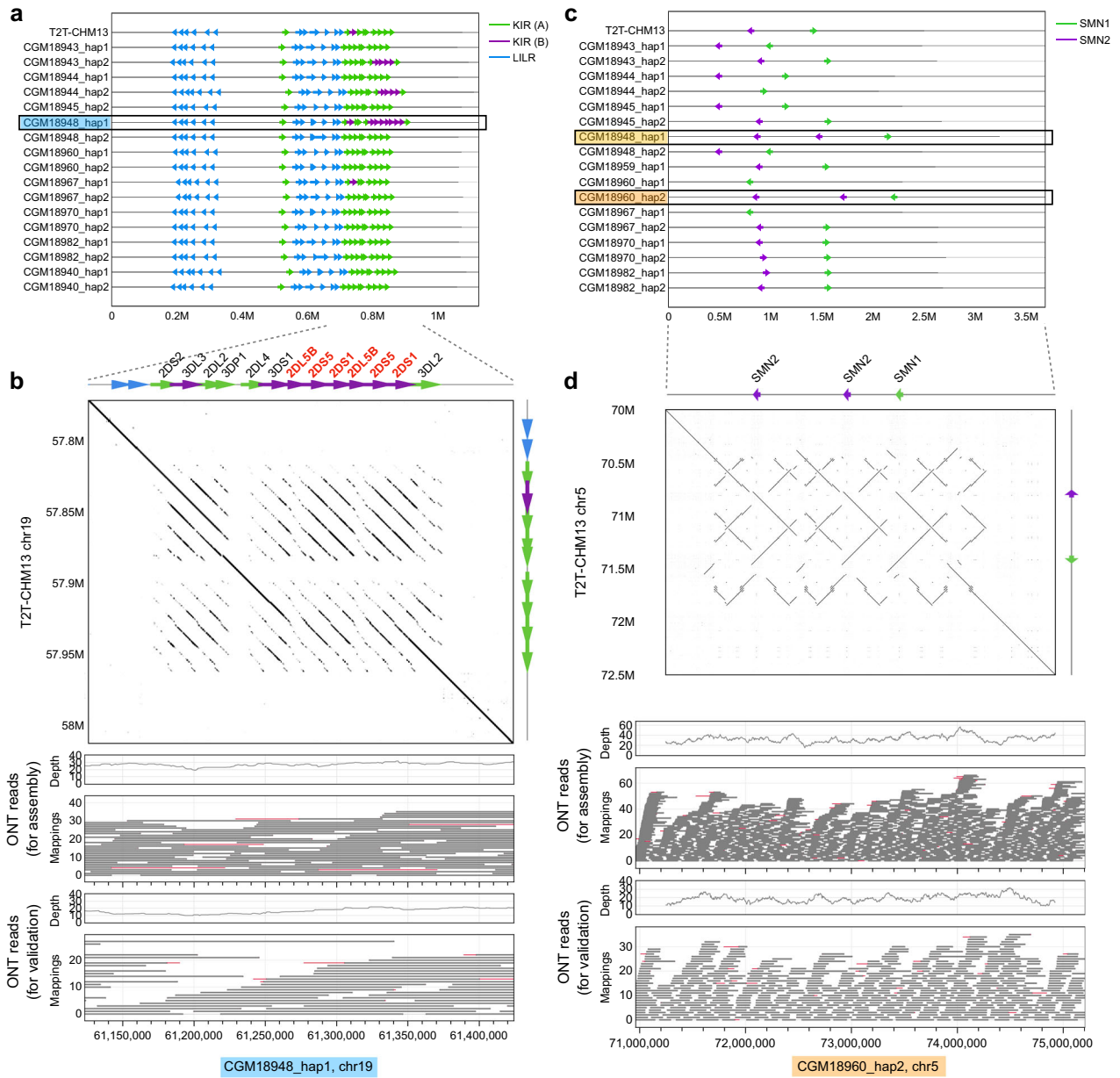

Fig. 3 | Haplotypes of medically important regions that were not identified in previous pangenome efforts. a, Complete Japanese haplotypes and T2T-CHM13 in the *KIR-LILR* region, showing only *KIR* and *LILR* genes. *KIR* B-haplotype genes (*KIR* (B)) are purple and the other *KIR* genes (*KIR* (A)) are green. Arrows indicate gene orientation (forward: right arrow, reverse complement: left arrow). The blue-highlighted Japanese haplotype (CGM18948_hap1) is absent from both HPRC Year-1 and Chinese haplotypes. **b**, Dot plot against T2T-CHM13 with coverage and pileup of ONT ultra-long reads for CGM18948_hap1 around the expanded *KIR* gene array. Gene names are shown at the top, and red genes indicate duplication of three B-haplotype genes, *KIR2DL5B*, *KIR2DS5*, and *KIR2DS1*. Two ONT ultra-long read datasets, used for assembly and validation, show no abnormal read depth shifts or clustered soft-clippings of alignments (red bars in mappings), implying that there are no large structural misassemblies. In addition, there were no SNVs in this region. **c**, Complete Japanese haplotypes plus T2T-CHM13 in the *SMN* region, showing only *SMN* genes. Two Japanese haplotypes (highlighted in yellow and orange, respectively) harbor three paralogous copies of *SMN*, whereas there were no such haplotypes present in the HPRC Year-1 and Chinese assemblies. **d**, Dot plot against T2T-CHM13 with ONT ultra-long read information for the longest, orange Japanese *SMN* haplotype (CGM18960_hap2). Both read datasets support the assembly as in **b**, with no SNVs detected. Anti-diagonal lines in the dot plot represent reverse complement matches between the two sequences.

Both studies share the fundamental goal of evaluating haplotype assemblies in terms of base-level accuracy and structural integrity. Consistent with this, both studies estimated the completeness of the HPRC Year-1 pangenome as approximately 20%. Among 127 haplotypes of HGSVC3 in which *KIR* genes were contained in a single contig, 73 (57.5%) were A-haplotypes and 54 (42.5%) were B-haplotypes (Supplementary Data 3), compared to 16 (80%) A-haplotypes and 4 (20%) B-haplotypes among our Japanese haplotypes. Although one B-haplotype identified in our Japanese pangenome was also present in the HGSVC3 assemblies, the novel B-haplotype with two tandem copies of the *KIR2DL5B*, *KIR2DS5*, and *KIR2DS1* genes (Fig. 3a, b), as well as the remaining two B-haplotypes, was absent from the HGSVC3 dataset. Additionally, just as we observed two haplotypes with three copies of *SMN* genes among the 20 Japanese haplotypes (Fig. 3c, d), they also found two such haplotypes from one African and one admixed American among the 130 haplotypes.

We also observed a large contribution of Japanese haplotypes to worldwide haplotypes by checking a pangenome graph of

https://doi.org/10.1038/s41467-026-73461-x

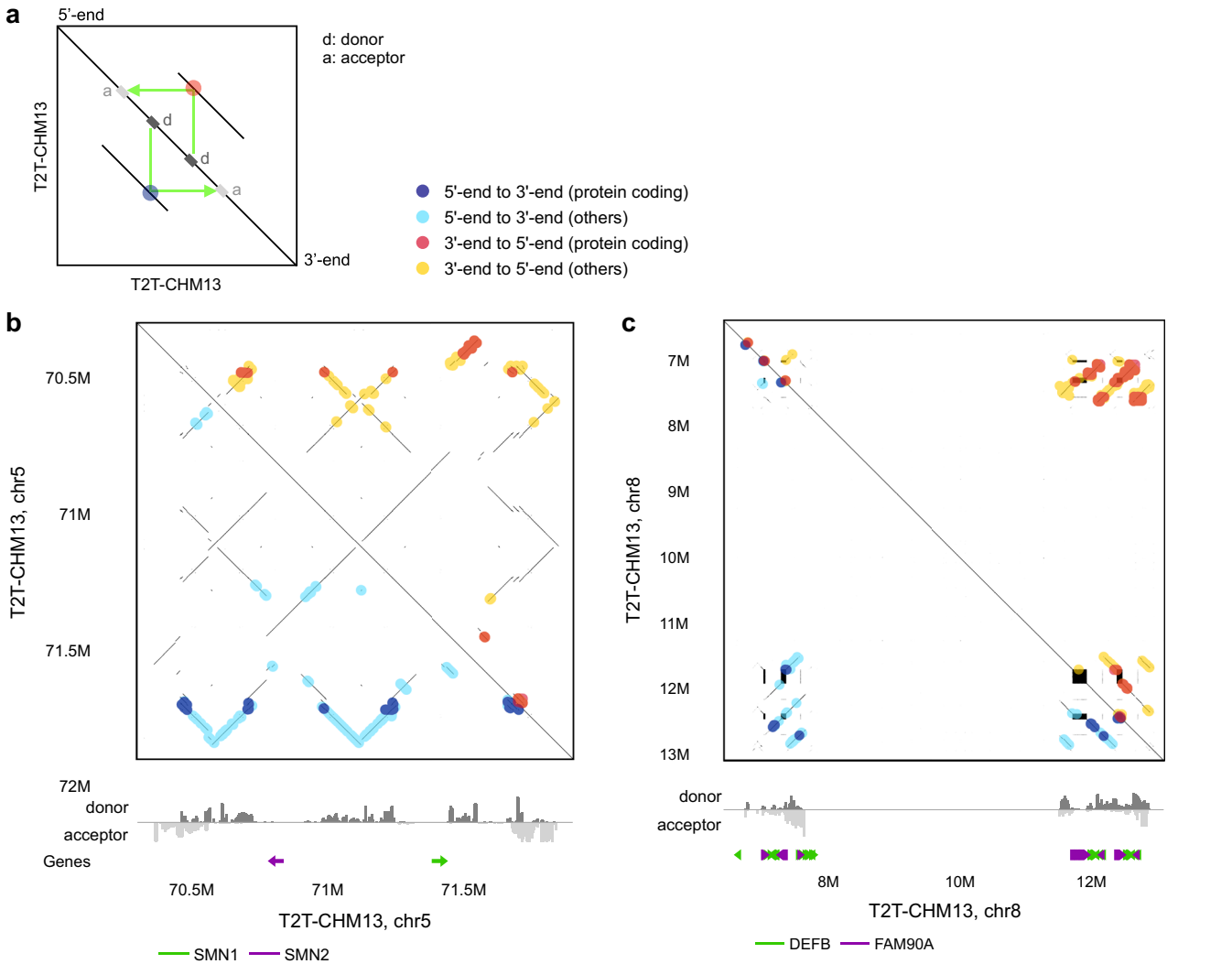

Fig. 4 | Candidates of biased gene conversions in medically important complex regions. a, Schematic description of the two-dimensional representation of putative interlocus gene conversion events. In the self dot plot of T2T-CHM13, off-diagonal lines indicate segmental duplications or other types of repeats. Putative gene conversion events found in each of the 20 Japanese haplotypes were mapped onto T2T-CHM13 and combined into one dot plot. Colored dots indicate donor-to-acceptor gene conversion events, with green arrows showing direction. Events in the 5′-end to 3′-end direction are plotted below the diagonal, with dark blue for protein-coding genes and light blue for other sequences. Those in the 3′-end to 5′-end direction are plotted above the diagonal, with red for protein-coding genes and yellow for other sequences. **b**, **c**, Putatively biased gene conversions identified in two regions containing the *SMN1* and *SMN2* genes (**b**) and the *DEFB* and *FAM90A* genes (**c**). The word size of dot plots is 100 bp. Bar plots display aggregated frequencies of donor (dark gray, upper) and acceptor (light gray, lower) sites, corresponding to vertical and horizontal aggregations of colored dots, respectively. Only *SMN*, *DEFB*, and *FAM90A* genes are annotated.

chromosome Y. Although the extremely repetitive heterochromatic region of cytoband Yq12 is still difficult to fully resolve in the pangenome graph, we obtained many complete haplotypes in other regions with protein-coding genes. For the second-largest gene cluster in the pangenome graph, the *TSPY* gene family, and the *AZFc* region, we obtained nine and seven complete haplotypes out of ten, respectively, with a large polymorphism of copy numbers of the genes (Supplementary Fig. 10). We found one complete Japanese haplotype in the *TSPY* region that has a huge inversion of ~4 Mbp in length. Similar inversions in complete haplotypes were previously reported by the HGSVC, but not in the six samples of the East Asian population in their dataset[5]. In the *AZFc* region, the pangenome graph made from HPRC and Japanese haplotypes contained two large nodes specific to Japanese whose sizes were 24,774 bp and 14,986 bp (Supplementary Fig. 10), highlighting the potential to enrich an existing pangenome graph and provide a more comprehensive representation of genomic diversity.

Putatively biased gene conversions

Medically important, segmentally duplicated regions are thought to be partly shaped by gene conversions that are non-reciprocal sequence transfer between duplicated, similar sequences that result in replacement of a local sequence of one allele (acceptor) with another allele (donor). We searched for putative interlocus gene conversion events on segmentally duplicated sequences in the medically important complex regions above by comparing local and syntenic sequence similarities between each Japanese haplotype and T2T-CHM13 (Fig. 4a, and "Methods")[17]. From the segmentally duplicated region around *SMN* genes, we identified 480 putative gene conversions in total from the 20 Japanese haplotypes, where 145 overlapped protein-coding genes. Although we did not detect gene conversions between the gene bodies of *SMN* genes, which are often seen in SMA patients[24], we observed that putatively biased gene conversions in the *SMN* region occurred in only one direction between specific duplicated segments (Fig. 4b). We revealed that 13% (96 kbp) of the gene conversion

locations were potentially biased based on a position-wise two-sided binomial test on the frequency of donor and acceptor events with a significance level of $P < 0.01$ (Supplementary Data 4). These putatively biased gene conversions include the *NAIP*, *GTF2H2*, *OCLN*, and *GUSBP* (pseudo-)genes.

The region containing the *DEFB* and *FAM90A* genes showed a similar trend of putatively biased gene conversions (Fig. 4c), where 8.8% (149 kbp) of the gene conversion locations were potentially biased (Supplementary Data 5). The frequencies of donors and acceptors identified on the *DEFB* genes were 60 and 54, compared to 72 and 71 of the *FAM90A* genes. Although the *DEFB* gene family has 28 different paralogous copies on T2T-CHM13, the skew in *DEFB* genes was found only on a single pseudogene, *DEFB109E*, at two locations of chr8:7,532,138-7,539,216 and chr8:12,180,513-12,187,611 (chi-squared test, $\chi^2 = 11.285$, d.f. = 1, $P = 7.8 \times 10^{-4}$).

Discussion

In this study, we constructed near-complete diploid assemblies for ten Japanese male individuals, achieving both high contiguity and base quality for every assembled haplotype. The contig NG50 contiguity metric exceeded 100 Mbp, and the quality value was over Q60 for all assembled haplotypes according to a k-mer statistical analysis. On average, we achieved 7.1 telomere-to-telomere chromosomal contigs per haplotype among the 20 haplotypes. The completeness of our Japanese haplotypes surpasses that of the current best Japanese reference genome, JG3, which features two telomere-to-telomere chromosomes (chromosomes 16 and X)[66]. This was made possible by combining deep-sequenced ONT ultra-long reads with PacBio HiFi long reads and Omni-C long-range reads. Note that, however, current assemblers still face challenges in certain regions. For instance, manual curation was necessary for assigning sex chromosomes into single separated haplotypes for several samples, particularly in highly heterochromatic HSat regions of chromosome Y. Moreover, accurate haplotype phasing of acrocentric chromosomes' p-arms remains unresolved.

By selecting only male samples, we were able to obtain the maximum number of chromosome Y sequences from ten samples. Since chromosome Y is transferred across generations without recombination except within the PARs, accumulating complete haplotypes from many populations will help trace the worldwide history of large-scale rearrangements[5,67]. A previous study has reported 43 nearly complete chromosome Y sequences[5], and our Japanese haplotypes will further expand the evolutionary and comparative analyses.

Our goal was to unveil the hidden genomic diversity within the Japanese population, with a particular focus on complete haplotypes in medically relevant complex regions characterized by large segmental duplications. These complete haplotype-resolved assemblies provide a deeper view of genomic variation, including small-scale polymorphisms and larger structural variants. This comprehensive view is crucial for understanding the full spectrum of genetic diversity and its implications for disease susceptibility, drug response, and human evolution while avoiding reference biases and biases due to unassembled or misassembled sequences[4,16,17,68,69]. When mapping (ultra-)long reads that span multiple genes to a pangenome, the comprehensiveness of the haplotype structure in the underlying population becomes important for mapping quality[7,33]. For instance, in this study, we observed that the copy numbers of the *TSPY3* and *TSPY4* genes were zero or one in some East Asian haplotypes in the HPRC Year-1 dataset, compared to 8–14 and 2–6, respectively, in our Japanese haplotypes (Fig. 2b). However, in the near-complete Y chromosomes study by HGSVC[5], the copy numbers in East Asian haplotypes were 8–9 and 4–5, respectively, aligning our observations and improvement in assembly quality concerning segmental duplications. Together, we aim to expand the accurate global pangenome.

Notably, we observed minor but novel haplotypes in the *KIR* and *SMN* regions that were not identified in previous pangenome projects. In the *KIR* region, we fully assembled one non-canonical B-haplotype carrying a previously unreported combination of genes: two copies of the *KIR2DL5B, KIR2DS5*, and *KIR2DS1* genes. B-haplotypes in the *KIR* regions have been reported to exist only in ~25% of Japanese haplotypes[21], and our findings highlight the utility of long-read sequencing in uncovering novel B-haplotypes even in control samples. In the *SMN* region, we obtained two complete haplotypes carrying two copies of the *SMN2* gene, which is a copy number profile reported in 2.4% of individuals in the East Asian population[23]. Note that the absence of haplotypes with three-copy *SMN* genes in previous HPRC Year-1 and Chinese pangenomes likely reflects assembly challenges rather than their true absence. This suggests that highly complex haplotypes may be under-represented in current pangenomes and emphasizes the importance of complete assemblies of complex genomic regions. Expanding (near-)complete assemblies across diverse populations, including well-characterized populations, will be essential for accurately uncovering hidden genomic diversity in medically important complex regions[8]. International consortiums such as HGSVC, HPRC, and HPP have been making tremendous efforts in this direction.

Several studies have suggested that variations in the *SMN1* and *SMN2* genes alone cannot fully account for the severity of SMA symptoms, and that the broader genomic environment, including nearby (pseudo-)genes such as *NAIP*, *SERF1*, *GTF2H2*, *OCLN*, and *GUSBP,* may play a role in explaining the discrepancy[24,62,70]. Our complete *SMN* haplotypes and findings on putatively biased gene conversions among these genes would help facilitate a better understanding of associations between sequence composition and SMA symptoms, particularly in a Japanese population. We also found putatively biased gene conversions in the region around *DEFB* and *FAM90A* gene families. This region has also been challenging to assemble despite the importance of beta-defensin and primate-specific gene expansion[26,39]. These observations on putatively biased gene conversion events suggest a potentially non-random evolutionary process in Japanese haplotypes within these highly complex segmental duplications, and will be further validated as more complete haplotypes from diverse global populations become available in the near future. Our findings will contribute to a better understanding of the evolutionary history of both the Japanese population and the worldwide populations, as well as the genetic factors that have shaped them.

A key future challenge lies in the accurate annotation of paralogous copies of genes in complex segmental duplications. Current liftover tools for gene annotation are reference-dependent and do not consider the relevance of specific nucleotide variants when performing assignments of paralogous gene copies. While using a comprehensive annotation of a reference genome including alternative sequences can alleviate this, it is not sufficient. We performed manual annotation for the *KIR* and *SMN* regions; specifically, we utilized the Immuannot program using the IPD-KIR database for the *KIR* region and the single variant, c.840C>T, on exon 7 for the *SMN* region. In addition, some specialized methods for specific regions have been reported[5,23,58,71,72]. However, a more unified method generalized for a wide range of complex regions is ultimately needed for accurate analyses of medically relevant genes in segmental duplications.

Methods

Ethical statement

This study was conducted in accordance with the Ethical Guidelines for Medical and Biological Research Involving Human Subjects and was approved by the Research Ethics Committee of the Graduate School of Frontier Sciences, The University of Tokyo (review no. 22-454). This study used publicly available and de-identified samples and data from the 1000 Genomes Project. No new human participants were

recruited, and no new human specimens were collected. The original 1000 Genomes Project samples were collected with informed consent and ethical review as part of the 1000 Genomes Project.

Selection of ten Japanese samples

For Pacific Biosciences HiFi (PacBio HiFi), Oxford Nanopore Technologies ultra-long (ONT-UL), and Dovetail Omni-C (Omni-C) DNA sequencing, ten lymphoblastoid cell lines from healthy Japanese male donors were used. These cell lines were obtained from the Coriell Institute for Medical Research in Camden, NJ, USA. The IDs of the cell lines were GM18940, GM18943, GM18944, GM18945, GM18948, GM18959, GM18960, GM18967, GM18970, and GM18982. After the samples were distributed, we added the letter C to the prefix of each sample ID: e.g., CGM18940 and so on. Cells were cultured in RPMI-1640 medium supplemented with 10% fetal bovine serum (FBS) at UNITECH Co., Ltd. in Japan, and karyotyping by Trans Chromosomics, Inc. (Japan) confirmed that these samples had no abnormalities. All expansions for sequencing were derived from the original expansion culture lot to ensure the lowest possible number of passages. This resulted in a total of three or four passages, except for GM18944, which had five passages. Cells were divided into vials with production-specific sizes: PacBio HiFi (3×10^6 cells), ONT-UL (6×10^6 cells), and Omni-C (1×10^6 cells). Cells for ONT-UL were frozen in RPMI-1640 supplemented with 20% FBS and 6% DMSO. Cells for PacBio HiFi and Omni-C were washed in PBS and flash-frozen as cell pellets. Following Cantata Bio's (Scotts Valley, CA, USA) shipping manual, ten cell lines' frozen pellets were sent for Omni-C. The Omni-C library was prepared and sequenced using the Dovetail Omni-C Kit at Cantata Bio.

DNA isolation and sequencing for PacBio HiFi reads

High-molecular-weight DNA was extracted from frozen cell pellets using the Blood & Cell Culture DNA Mini Kit (QIAGEN) with Genomic-tip 20/G, following the QIAGEN Genomic DNA Handbook. The DNA was fragmented using the Diagenode Megaruptor 3 system. Throughout the process, the quality of the DNA was checked using various instruments such as the Qubit 4 Fluorometer with a dsDNA HS Assay Kit (Thermo Fisher Scientific), NanoDrop spectrophotometer (Thermo Fisher Scientific), and FEMTO Pulse system with Genomic DNA 165 kb analysis kit (Agilent Technologies) to examine the sizes. SMRTbell libraries were prepared using SMRTbell prep kit 3.0 (PacBio). The libraries were size-selected with the AMPure PB bead size selection step, which removed DNA fragments shorter than 5 kb. The selected libraries were bound with Sequel II Primer 3.2 and Sequel II DNA Polymerase 2.2 using Binding Kit 3.2 from PacBio. The sequencing was performed on Sequel II instruments (PacBio) on SMRT Cell 8M with Sequel II sequencing kit 2.0. The process included adaptive loading, 2 h of pre-extension, and 30 h of movie time.

DNA isolation and sequencing for ONT ultra-long reads at The University of Tokyo

We aimed to extract gDNA with extremely low fragmentation and ultra-high molecular weight for ONT-UL DNA sequencing. According to the Ultra-Long DNA Sequencing protocol for SQK-ULK114 (ONT), the Monarch HMW DNA Extraction Kit for Tissue (New England BioLabs) was used. A total of six million frozen cells, pre-washed twice with cold PBS, was required for DNA extraction, which contained sufficient DNA for one flow cell 72-h run. The quality of the eluted ultra-long DNA was measured using the Qubit 3 Fluorometer with a dsDNA BR Assay Kit (Thermo Fisher Scientific). The sizes of the DNA fragments were examined using the Pippin Pulse system (Sage Science) with a 5–430 kb protocol developed by Sage Science using a 0.75% agarose gel made of SeaKem Gold (Lonza). We used the same protocol for creating ultra-long libraries using an Ultra-Long DNA Sequencing Kit V14 (SQK-ULK114, ONT). The protocol revision F was used for CGM18943, CGM18944, CGM18945, CGM18948, CGM18959, and CGM18960, while revision K was used for the other samples. After the ultra-long DNA tagmentation and clean-up steps, a glassy white precipitate may form when added with the Precipitation Buffer (PTB, ONT). The resulting ultra-long library (glassy white mass) was then eluted in 300 μl EB (ONT) overnight at room temperature. The ultra-long library in the Eppendorf DNA LoBind tube could be stored at 4°C for a short term until ready to load.

We followed ONT's instructions to prime the R10.4.1 version FLO-PRO114M flow cell with Flow Cell Flush/Flush Tether UL mixture before loading the ultra-long library onto the PromethION sequencer (P2 Solo, ONT). The sequencing experiment set 1.5 h as the time between pore scans, and the run limit was 72 h as default. To increase data output, a second library (at 24 h after the run was started) and a third library (at 48 h after the run was started) were reloaded straight after flushing a flow cell using Flow Cell Wash Kit (EXP-WSH004-XL, ONT). Flow cells with high numbers of active pores remaining after the 72-h run were sequenced for an additional 24 h by nuclease flushing the flow cells with the Flow Cell Wash Kit and reloading the library. Basecalling was performed using Guppy (v4.2.0), with default parameters and the high-accuracy (HAC) basecalling model.

DNA isolation and sequencing for ONT ultra-long reads at UCSC

Lymphoblastoid cell lines were purchased from the Coriell Institute and cultured in RPMI-1640 medium with 2 mM L-glutamine and 15% FBS at 37°C, 5% CO_2. Cells were counted by Automated Cell Counter (BioRad, TC20), placed into 18 million cell aliquots, washed two times with phosphate-buffered saline (Gibco, 10010023), flash-frozen in liquid nitrogen, and stored at −80°C until extraction. High-molecular-weight (HMW) DNA was extracted from approximately 18 million frozen cells per sample using the NEB HMW DNA Extraction Kit for Tissues (NEB #T3060), following the ONT recommended gDNA Extraction Protocol (v114_revL_27Nov2022). The protocol was modified to increase the elution volume due to the threefold increase in cell input, resulting in a total elution volume of 2250 μL. All other steps in the protocol remained unchanged. Library preparation was performed using the ONT Ultra-Long DNA Sequencing Kit (SQK-ULK114). Given the increased DNA input (2250 μL), all reagent volumes were scaled by a factor of 3. Samples were eluted in a final volume of 820 μL EB to accommodate four library loads across three flow cells per sample, resulting in a total of 12 libraries per sample. All samples were run on the PromethION 48 sequencer (PRO-SEQ048) using R10.4.1 flow cells (FLO-PRO114M). Flow cells were washed using the ONT wash kit (EXP-WSH004) and reloaded with fresh libraries every 24 h, for a total runtime of 96 h. Basecalling of the raw signal data was performed using Dorado (v0.5.3, commit d9af343) with default parameters. The super high-accuracy (SUP) basecalling model dna_r10.4.1_e8.2_400bps_sup@v4.3.0 was used. Modified base detection was enabled for 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC) using the model dna_r10.4.1_e8.2_400bps_sup@v4.3.0_5mCG_5hmCG@v1, called with the option --modified-bases 5mCG_5hmCG. Basecalled reads were output in BAM format, with headers annotated to include metadata for sequencing runs, platform details, and software versions.

DNA sequencing for Dovetail Omni-C reads

Two Illumina libraries were prepared from each sample (a total of 20 libraries from 10 samples). All libraries were then sequenced in paired-end mode at Cantata Bio (Scotts Valley, CA, US). Each Omni-C library was deep-sequenced via Illumina sequencers ~300 million read pairs (2 × 150 bp).

Genome assembly

For each of the ten Japanese samples, we assembled the HiFi, ONT-UL, and Omni-C reads together using hifiasm (v0.19.8) in the ultra-long read mode[15]. The hifiasm assembler separates the two haplotypes of a given

sample using the long-range information of Omni-C reads and generates each haplotype as hap1 and hap2, where, for each chromosome, the maternal and paternal haplotypes are arbitrarily assigned to either hap1 or hap2 in the case of using Omni-C reads instead of trio Illumina reads. At this stage, the correctness of haplotype assignment of the contigs was inspected for each sample based on the alignment plot between each haplotype and the T2T-CHM13 v2.0 reference genome, including chromosome Y, generated using MUMmer (v4.0.0)[73]. Since all the samples in this study are male samples, chromosome X and Y must belong to only one of the two haplotypes, respectively, although how to organize the sex chromosomes in each of the two haplotypes, hap1 and hap2, is arbitrary. In this study, we define hap1 contains chromosome X (and the mitochondrial genome) and hap2 contains chromosome Y. In all samples except CGM18943, most contigs of chromosome X already belonged to hap1 and those of chromosome Y belonged to hap2. For the remaining one sample, CGM18943, we switched hap1 and hap2. After this, several contigs misassigned to a wrong haplotype (typically in a sex chromosome) were manually transferred to the other haplotype as listed in the Supplementary Table 3. For one sample, CGM18940, we observed short and redundant contigs with small sequencing coverages and filtered out such extra sequences using purge_dups (v1.2.5)[74] before scaffolding.

After the manual curation of contigs, the curated contigs of each haplotype were then scaffolded into chromosomal sequences using RagTag (v2.1.0)[35], with the T2T-CHM13 reference genome as the guide sequence. We skipped de novo scaffolding using long-range information of Omni-C reads because most of the contigs were already very long and scaffolding them into chromosomal sequences was obvious for the reference-based scaffolding method, and the short contigs were full of repetitive sequences for which Omni-C short reads cannot be effective when scaffolding. For one sample, CGM18943, one large contig was still unlocalized and thus manually scaffolded into chromosome Y based on the MUMmer alignment plot with T2T-CHM13. Several scaffold sequences were manually curated for CGM18940. The list of contigs and scaffolds manually curated is in the Supplementary Table 3.

Assembly cleanup and quality assessment

The assembled haplotypes were post-processed using the hpp_production_workflows pipeline (https://github.com/human-pangenomics/hpp_production_workflows; GitHub commit ID: a4e5a15). First, we ran the assembly_cleanup.wdl workflow to remove short contaminated contigs from the assembled scaffolds using NCBI's Foreign Contamination Screen (FCS) program[75] and to identify a consensus mitochondrial chromosome sequence using the mitoHiFi program[76]. A single complete sequence of the mitochondrial chromosome was successfully determined for all the ten samples. Moreover, unlocalized sequences shorter than 50 kbp were filtered out during the workflow. Then, we ran the comparison_qc.wdl workflow for quality assessment of the cleaned haplotypes. It calculates the auN metrics of the assembled haplotypes, as well as the single-copy gene completeness with asmgene[48] and compleasm[77] and the quality value with yak (https://github.com/lh3/yak). We downloaded the Illumina reads of the ten samples required for the yak QV calculation from the International Genome Sample Resource (IGSR) website (https://www.internationalgenome.org). These datasets, NA18940.final.cram to NA18982.final.cram, are the same files previously collected by the 1000 Genomes Project. In addition to the pipeline above, the N50 length of each haplotype assembly and the size of each chromosomal scaffold were calculated with seqkit (v2.0.0)[78]. The quality of the assembled scaffolds was also evaluated with Merqury (v1.3)[36] for each sample using the 21-mer table of the HiFi reads.

Analysis on the assembled haplotypes

Pairwise genome comparison between each haplotype and T2T-CHM13 along with detection of structural rearrangements was performed with SyRI[79]. Regions of inversions were extracted from SyRI's results, and the overlapping inversion regions were calculated with bedtools (v2.30.0)[80]. Annotations of centromeric alpha satellite and HSat1 arrays were derived from the output of RepeatMasker (v4.1.5). Centromeric HSat2 and HSat3 sequences were difficult to distinguish and were therefore annotated using the Assembly_HSat2and3_v3.pl script (https://github.com/altemose/chm13_hsat, v3).

Segmental duplication analysis

Segmental duplications (SDs) were identified from each haplotype using Biser (v1.4)[42] after masking repeats with RepeatMasker using the following parameters: -s -xsmall -e ncbi -species human. Only SDs longer than 1 kbp and with sequence similarity greater than 90% were retained. The resulting SDs were lifted over to T2T-CHM13 coordinates for comparison with other haplotypes using minimap2 (v2.24)[48], and merged across haplotypes using the bedtools merge command for the Japanese, HPRC Year-1 (EAS and non-EAS), and Chinese pangenomes. SDs found in centromeric regions were excluded due to their highly repetitive nature. Comparison of the lifted-over Japanese SD regions with SDs in T2T-CHM13 and other pangenomes was performed using the bedtools intersect -v command, which returns completely non-overlapping SDs. The number of haplotypes covering the SD regions was calculated using the bedtools multiinter command, which takes SD regions of multiple haplotypes and reports the sub-regions shared by the same set of haplotypes.

Pangenome graph construction

Two types of Japanese pangenome graph were constructed: minigraph-cactus (v2.9.0)[43] and pggb (nf-core/pangenome v1.1.2)[44,81]. Two haploid reference genomes, T2T-CHM13 and GRCh38, were incorporated into each graph as well as the 20 Japanese haplotypes. The pggb graph was generated for each chromosome with the nextflow pipeline, and all the graphs were optimized and merged into a single graph using odgi (v0.8.6)[82]. The minigraph-cactus graph was constructed with the cactus-pangenome pipeline command. In addition, local pggb graphs of medically important complex regions were constructed from the sequences of the specific region extracted from Japanese haplotypes, HPRC haplotypes, and the two reference genomes. These graphs are available at https://doi.org/10.5281/zenodo.14160240.

Variants in each haplotype with respect to T2T-CHM13 (or GRCh38) were called from the graph using vg (v1.57.0)[83], vcfbub (v0.1.1)[45], and vcflib (v1.0.9)[46]. Specifically, we first generated a vcf file containing all variants inferred from the pangenome graph using the following command: vg deconstruct -P CHM13 -e -a <pggb_graph.gfa>. Then, the resulting variants were normalized with the following command: vcfbub -l 0 -a 100000 --input <pggb_graph.vcf.gz> | vcfwave -I 1000. The variants were further normalized and sorted using bcftools (v1.18)[84], and variants within centromeric regions were excluded. SNVs were extracted using the following command: bcftools view -v snps, and SVs were defined as variants for which the length difference between the reference sequence and the alternative sequence was greater than 50 bp. The identified SVs of the Japanese haplotypes were compared between minigraph-cactus and pggb with Truvari (v5.3.0)[47] using the following command: truvari bench -r 1000 --sizemax 100000. For the comparison involving other haplotypes, similar SVs were merged using the following command: truvari collapse -r 1000 -p 0.8 -P 0.0 -O 0.8 -s 50 -S 100000, following a previous study[54]. For the stratification of SVs, subtelomeric regions on T2T-CHM13 were defined as the 1-Mbp segments at each chromosome end, and regions surrounding segmental duplications were defined using annotated segmental duplications (https://github.com/marbl/CHM13) with 100-kbp flanking regions. Tandem repeats within SV sequences were annotated with Tandem Repeats Finder[85] using the following command: trf 2 7 7

80 10 50 2000. Inserted sequences of SVs unique to Japanese haplotypes were aligned to the pangenome graph of the Chinese and HPRC Year-1 haplotypes (https://pog.fudan.edu.cn/cpc) using GraphAligner[86] with the default settings. The sequence identity of each inserted sequence with the graph was calculated as the number of matches divided by its sequence length.

We downloaded the small-variant dataset from the 1000 Genomes Project, which was generated using high-coverage whole-genome Illumina reads, from its FTP site (folder name: 20220422_3202_phased_SNV_INDEL_SV/)[6]. The variants in this dataset are defined on the GRCh38 reference genome and include the ten Japanese samples analyzed in this study. We extracted these ten Japanese samples from the vcf file for comparison with the variants calculated with respect to GRCh38 from the pangenome graph constructed in this study. Low-complexity regions and tandem repeat regions were excluded from the comparison.

Gene annotation and gene copy number calculation

For comprehensiveness, instead of the gene annotation on T2T-CHM13, we lifted over the GENCODE v38 comprehensive gene annotation of GRCh38, including the alternative haplotypes to each Japanese haplotype using Liftoff (v1.6.3)[49,87]. For compatibility, the same annotation was lifted over to the T2T-CHM13 reference genome, as well as the EAS haplotypes of the HPRC Year-1 dataset and the haplotypes of the CPC dataset. Gene family names were retrieved from the HGNC database[88]. For several medically important gene cluster regions such as *HLA*, *KIR*, and *SMN*, a manual gene annotation was performed. Specifically, *HLA* and *KIR* genes were annotated with the help of public databases using Immuannot (GitHub commit ID: bca45ef)[58]. The two paralogous copies of the *SMN* gene, i.e., *SMN1* and *SMN2*, are known to be distinguished by a paralogous single-nucleotide variant, c.840C>T[24]. Liftoff's annotations were updated based on this important paralogous SNV using a custom script.

Quality assessment of haplotypes in medically relevant regions

ONT ultra-long reads of >100 kbp in length were used for quality assessment of assemblies in segmentally duplicated complex regions. The ONT ultra-long reads were mapped to the chromosomal scaffolds with winnowmap2 (v2.03)[89]. The average sequencing coverage for every non-overlapping 1 kbp window on the scaffolds was calculated with mosdepth (v0.3.1)[90]. Windows with an average coverage of less than three were regarded as a coverage drop. Small variants were detected with DeepVariant (v1.5.0)[91]. Variants with a genotype quality less than 25 were filtered out according to a previous study[92]. A stretch of consecutive N characters in the nucleotide sequence of a scaffold was counted as one sequence gap. For each of the 30 selected medically relevant complex genomic regions, we assessed sequence contiguity based on the number of sequence gaps, structural quality based on the number of coverage-dropping windows, and base quality based on the number of small variants, within the region.

Gene conversion analysis

Putative gene conversions were detected using the asm-to-reference-alignment workflow (https://github.com/mrvollger/asm-to-reference-alignment, v0.1)[17]. Specifically, we first calculated long, syntenic sequence alignments between each of the Japanese complete haplotypes and T2T-CHM13 using minimap2 with the options -x asm20 --secondary=no -s 25000. Within each syntenic alignment, the Japanese haplotype sequence was then divided into short, windowed subsequences of 1 kbp, and each subsequence was realigned to T2T-CHM13 to search for the best-matching homologous sequence across the reference genome. For each 1 kbp subsequence, we compared the sequence similarity to the expected syntenic position with the sequence similarity to other duplicated loci in T2T-CHM13. If a subsequence within a syntenic alignment showed higher sequence similarity to a different duplicated sequence at a distant location than to the syntenic reference region, the subsequence was considered a potential acceptor sequence, and the corresponding best-matching region in the distant duplicated sequence was designated as its donor sequence. The coordinates of candidate gene conversion events were defined based on T2T-CHM13.

Reporting summary

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

Data availability

The PacBio HiFi, ONT ultra-long, and Omni-C reads generated in this study at the University of Tokyo have been deposited in the NCBI SRA database under accession code PRJNA1459655. ONT ultra-long reads generated at UCSC have also been deposited in the NCBI SRA database under accession code PRJNA731524. The 20 assembled haplotype sequences have been deposited in the NCBI GenBank database under accession codes PRJNA1184421–PRJNA1184440. Sequence names in the fasta files of the assembled haplotype sequences follow the format NA<sample_ID>#<hap_ID>#<scaffold_name>; e.g., NA18940#1#chr1_hap2, where 18940 indicates the sample ID and chr1_hap2 denotes the chromosomal scaffold sequence name used in this paper. Reads and assemblies are also available at https://humanpangenome.org/hprc-data-release-2/. The pangenome graphs, segmental duplications, and Immuannot results generated in this study are available at https://doi.org/10.5281/zenodo.14160240. The underlying data for all plots generated in this study are provided in the Source Data file. Previously published data used in this study are available as follows: HPRC pangenome graphs under accession code PRJNA850430; CPC pangenome graphs from the CPC website [https://pog.fudan.edu.cn/cpc/#/data]; Illumina reads for the ten Japanese samples from the IGSR website [https://www.internationalgenome.org]; small variants called from Illumina reads for the ten Japanese samples from the 1000 Genomes Project website [https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20220422_3202_phased_SNV_INDEL_SV]; and T2T-CHM13 genome sequence and annotations from the CHM13 GitHub repository [https://github.com/marbl/CHM13]. Source data are provided with this paper.

Code availability

Custom scripts and Jupyter Notebooks for data analysis used in this study are available at https://doi.org/10.5281/zenodo.14160240.

References

1. Nurk, S. et al. The complete sequence of a human genome. *Science (1979)* **376**, 44–53 (2022).
2. The 1000 Genomes Project Consortium An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
3. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
4. Ebert, P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, 1979 (2021).
5. Hallast, P. et al. Assembly of 43 human Y chromosomes reveals extensive complexity and variation. *Nature* **621**, 355–364 (2023).
6. 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).
7. Liao, W.-W. et al. A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
8. Wang, T. et al. The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604**, 437–446 (2022).
9. Gao, Y. et al. A pangenome reference of 36 Chinese populations. *Nature* **619**, 112–121 (2023).

10. Wenger, A. M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
11. 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).
12. Nurk, S. et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **30**, 1291–1305 (2020).
13. Jarvis, E. D. et al. Semi-automated assembly of high-quality diploid human reference genomes. *Nature* **611**, 519–531 (2022).
14. Rautiainen, M. et al. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat. Biotechnol.* **41**, 1474–1482 (2023).
15. Cheng, H., Asri, M., Lucas, J., Koren, S. & Li, H. Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. *Nat. Methods* **21**, 967–970 (2024).
16. Vollger, M. R. et al. Segmental duplications and their variation in a complete human genome. *Science* **376**, 1979 (2022).
17. Vollger, M. R. et al. Increased mutation and gene conversion within human segmental duplications. *Nature* **617**, 325–334 (2023).
18. Jeong, H. et al. Structural polymorphism and diversity of human segmental duplications. *Nat. Genet.* **57**, 390–401 (2025).
19. Robinson, J., Mistry, K., McWilliam, H., Lopez, R. & Marsh, S. G. E. IPD—the Immuno Polymorphism Database. *Nucleic Acids Res.* **38**, D863–D869 (2010).
20. Hirata, J. et al. Genetic and phenotypic landscape of the major histocompatibilty complex region in the Japanese population. *Nat. Genet.* **51**, 470–480 (2019).
21. Sakaue, S. et al. Decoding the diversity of killer immunoglobulin-like receptors by deep sequencing and a high-resolution imputation method. *Cell Genomics* **2**, 100101 (2022).
22. Mercuri, E., Sumner, C. J., Muntoni, F., Darras, B. T. & Finkel, R. S. Spinal muscular atrophy. *Nat. Rev. Dis. Primers* **8**, 52 (2022).
23. Chen, X. et al. Comprehensive SMN1 and SMN2 profiling for spinal muscular atrophy analysis using long-read PacBio HiFi sequencing. *Am. J. Hum. Genet.* **110**, 240–250 (2023).
24. Zwartkruis, M. M. et al. Long-read sequencing identifies copy-specific markers of SMN gene conversion in spinal muscular atrophy. *Genome Med.* **17**, 26 (2025).
25. Bosch, N. et al. Characterization and evolution of the novel gene family FAM90A in primates originated by multiple duplication and rearrangement events. *Hum. Mol. Genet.* **16**, 2572–2582 (2007).
26. Fellermann, K. et al. A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. *Am. J. Hum. Genet.* **79**, 439–448 (2006).
27. Hollox, E. J. et al. Psoriasis is associated with increased β-defensin genomic copy number. *Nat. Genet.* **40**, 23–25 (2008).
28. Stuart, P. E. et al. Association of β-defensin copy number and psoriasis in three cohorts of European origin. *J. Invest. Dermat.* **132**, 2407–2413 (2012).
29. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
30. Nagai, A. et al. Overview of the BioBank Japan Project: study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).
31. Hujoel, M. L. A. et al. Protein-altering variants at copy number-variable regions influence diverse human phenotypes. *Nat. Genet.* **56**, 569–578 (2024).
32. Fanciulli, M. et al. FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat. Genet.* **39**, 721–723 (2007).
33. Sirén, J. et al. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science (1979).* **374**, abg8871 (2021).
34. Groza, C. et al. Pangenome graphs improve the analysis of structural variants in rare genetic diseases. *Nat. Commun.* **15**, 657 (2024).
35. Alonge, M. et al. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol.* **23**, 258 (2022).
36. 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).
37. Craig-Holmes, A. P. & Shaw, M. W. Polymorphism of human constitutive heterochromatin. *Science (1979)* **174**, 702–704 (1971).
38. Miga, K. H. Centromeric satellite DNAs: hidden sequence variation in the human population. *Genes (Basel)* **10**, 352 (2019).
39. Logsdon, G. A. et al. The structure, function and evolution of a complete human chromosome 8. *Nature* **593**, 101–107 (2021).
40. Porubsky, D. et al. Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell* **185**, 1986–2005.e26 (2022).
41. Mohajeri, K. et al. Interchromosomal core duplicons drive both evolutionary instability and disease susceptibility of the Chromosome 8p23.1 region. *Genome Res.* **26**, 1453–1467 (2016).
42. Išerić, H., Alkan, C., Hach, F. & Numanagić, I. Fast characterization of segmental duplication structure in multiple genome assemblies. *Algorithms Mol. Biol.* **17**, 4 (2022).
43. Hickey, G. et al. Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat. Biotechnol.* **42**, 663–673 (2024).
44. Garrison, E. et al. Building pangenome graphs. *Nat. Methods* **21**, 2008–2012 (2024).
45. Garrison, E., Guarracino, A. & Hickey, G. pangenome/vcfbub: v0.1.1. *Zenodo*. https://doi.org/10.5281/zenodo.12206906 (2024).
46. Garrison, E., Kronenberg, Z. N., Dawson, E. T., Pedersen, B. S. & Prins, P. A spectrum of free software tools for processing the VCF variant call format: vcflib, bio-vcf, cyvcf2, hts-nim and slivar. *PLoS Comput. Biol.* **18**, e1009123 (2022).
47. 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).
48. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
49. Frankish, A. et al. GENCODE 2021. *Nucleic Acids Res.* **49**, D916–D923 (2021).
50. Morgan, M. A. J. et al. *ELOA3*: a primate-specific RNA polymerase II elongation factor encoded by a tandem repeat gene cluster. *Sci. Adv.* **9**, eadj1261 (2023).
51. Ikeda, T. et al. TBCD may be a causal gene in progressive neurodegenerative encephalopathy with atypical infantile spinal muscular atrophy. *J. Hum. Genet.* **62**, 473–480 (2017).
52. Rhie, A. et al. The complete sequence of a human Y chromosome. *Nature* **621**, 344–354 (2023).
53. Nie, J., Tellier, J., Tarasova, I., Nutt, S. L. & Smyth, G. K. T2T-CHM13 versus hg38: accurate identification of immunoglobulin isotypes from scRNA-seq requires a genome reference matched for ancestry. *NAR Genom. Bioinform.* **7**, lqaf074 (2025).
54. Nassir, N. et al. A draft UAE-based Arab pangenome reference. *Nat. Commun.* **16**, 6747 (2025).
55. Kuroda-Kawaguchi, T. et al. The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. *Nat. Genet.* **29**, 279–286 (2001).
56. Johnson, M. E. et al. Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413**, 514–519 (2001).
57. Yawata, M. et al. Roles for HLA and KIR polymorphisms in natural killer cell repertoire selection and modulation of effector function. *J. Exp. Med.* **203**, 633–645 (2006).
58. Zhou, Y., Song, L. & Li, H. Full-resolution HLA and KIR gene annotations for human genome assemblies. *Genome Res.* **34**, 1931–1941 (2024).

59. Lorson, C. L., Hahnen, E., Androphy, E. J. & Wirth, B. A single nucleotide in the *SMN* gene regulates splicing and is responsible for spinal muscular atrophy. *Proc. Natl. Acad. Sci.* **96**, 6307–6311 (1999).
60. Kashima, T. & Manley, J. L. A negative element in SMN2 exon 7 inhibits splicing in spinal muscular atrophy. *Nat. Genet.* **34**, 460–463 (2003).
61. Lefebvre, S. et al. Correlation between severity and SMN protein level in spinal muscular atrophy. *Nat. Genet.* **16**, 265–269 (1997).
62. Calucho, M. et al. Correlation between SMA type and SMN2 copy number revisited: an analysis of 625 unrelated Spanish patients and a compilation of 2834 reported cases. *Neuromuscul. Disord.* **28**, 208–215 (2018).
63. Logsdon, G. A. et al. Complex genetic variation in nearly complete human genomes. *Nature* **644**, 430–441 (2025).
64. Vollger, M. R. et al. Long-read sequence and assembly of segmental duplications. *Nat. Methods* **16**, 88–94 (2019).
65. Chen, Y., Zhang, Y., Wang, A. Y., Gao, M. & Chong, Z. Accurate long-read de novo assembly evaluation with Inspector. *Genome Biol.* **22**, 312 (2021).
66. Takayama, J. et al. Construction and integration of three de novo Japanese human genome assemblies toward a population-specific reference. *Nat. Commun.* **12**, 226 (2021).
67. Poznik, G. D. et al. Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. *Nat. Genet.* **48**, 593–599 (2016).
68. Höps, W. et al. Impact and characterization of serial structural variations across humans and great apes. *Nat. Commun.* **15**, 8007 (2024).
69. Makova, K. D. et al. The complete sequence and comparative analysis of ape sex chromosomes. *Nature* **630**, 401–411 (2024).
70. Wadman, R. I. et al. Intragenic and structural variation in the SMN locus and clinical variability in spinal muscular atrophy. *Brain Commun.* **2**, fcaa075 (2020).
71. Chen, X. et al. Genome-wide profiling of highly similar paralogous genes using HiFi sequencing. *Nat. Commun.* **16**, 2340 (2025).
72. Prodanov, T. et al. Locityper enables targeted genotyping of complex polymorphic genes. *Nat. Genet.* **57**, 2901–2908 (2025).
73. Marçais, G. et al. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
74. Guan, D. et al. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
75. Astashyn, A. et al. Rapid and sensitive detection of genome contamination at scale with FCS-GX. *Genome Biol.* **25**, 60 (2024).
76. Uliano-Silva, M. et al. MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads. *BMC Bioinformatics* **24**, 288 (2023).
77. Huang, N. & Li, H. compleasm: a faster and more accurate reimplementation of BUSCO. *Bioinformatics* **39**, btad595 (2023).
78. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One* **11**, e0163962 (2016).
79. Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 277 (2019).
80. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
81. Heumos, S. et al. Cluster-efficient pangenome graph construction with nf-core/pangenome. *Bioinformatics* **40**, btae609 (2024).
82. Guarracino, A., Heumos, S., Nahnsen, S., Prins, P. & Garrison, E. ODGI: understanding pangenome graphs. *Bioinformatics* **38**, 3319–3326 (2022).
83. Garrison, E. et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–879 (2018).
84. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
85. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
86. Rautiainen, M. & Marschall, T. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biol.* **21**, 253 (2020).
87. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37**, 1639–1643 (2021).
88. Seal, R. L. et al. Genenames.org: the HGNC resources in 2023. *Nucleic Acids Res.* **51**, D1003–D1009 (2023).
89. Jain, C., Rhie, A., Hansen, N. F., Koren, S. & Phillippy, A. M. Long-read mapping to repetitive reference sequences using Winnowmap2. *Nat. Methods* **19**, 705–710 (2022).
90. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).
91. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
92. Mc Cartney, A. M. et al. Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. *Nat. Methods* **19**, 687–695 (2022).

Acknowledgments

The authors are grateful to Julian Lucas and Andrew Blair for the support on running assembly cleanup and quality assessment workflows and on uploading data to the HPRC Amazon S3 bucket, and to Heng Li for valuable comments on assembly registration. They are thankful to Milinn Kremitzki and Sarah Cody for coordinating with the HPRC community, and to Heather Lawson for coordinating with the HPP community. They thank Yuta Suzuki for basecalling PacBio and ONT reads, and Information Technology Center, The University of Tokyo for permission to use their supercomputing facilities.

Author contributions

S.M. designed the research. Y.S. performed most of the informatics analysis, and R.N. helped Y.S. examine medically important duplicated regions. C.O. and H.K. sequenced DNA samples and generated PacBio HiFi and Nanopore ultra-long reads for assembly. B.P. and K.M. conducted Nanopore ultra-long read sequencing for validation and managed the coordination with the HPRC. B.M. and I.V. generated nanopore ultra-long reads for validation. Y.S. and S.M. wrote the manuscript with input from all authors.

Funding

This research was supported by the Japan Agency for Medical Research and Development (AMED) Grant Number 24tm0424219h0004 and the JSPS KAKENHI Grant Number JP22H04925 (PAGS) to S.M., the JSPS KAKENHI Grant Number JP24K18091 to Y.S., and NIH/NHGRI R01HG011274 and UM1HG010971 to K.H.M.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-73461-x.

Correspondence and requests for materials should be addressed to Yoshihiko Suzuki or Shinichi Morishita.

Peer review information *Nature Communications* thanks Giuliana Giannuzzi, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.