A Complete Telomere-to-Telomere Diploid Reference Genome for Indian Population

Prasad Sarashetti[1#], Josipa Lipovac[2#], Qi Jia[1], Ling Wang[1], Zhe Li[1], Lovro Vrcek[1], Fei Yao[1], YY Sia[1], Dehui Lin[1], Xinyi Zhang[1], Daniel Muliaditan[1], Hwee Meng Low[1], See Ting Leong[1], Chiea Chuen Khor[1], Eleanor Wong[1], Weiling Zheng[3], Kresimir Krizanovic[2], John C Chambers[3,4], William Y K Hwang[5,6,7,8], Patrick Tan[1,3,9,10,], Mile Sikic[1,2*], Jianjun Liu[1,11*]

[1]Genome Institute of Singapore (GIS), Agency for Science, Technology and Research (A*STAR), 60 Biopolis Street, Genome, Singapore 138672, Singapore

[2]Laboratory for Bioinformatics and Computational Biology, Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia

[3]Precision Health Research, Singapore, 139234, Singapore

[4]Population and Global Health, Lee Kong Chian School of Medicine, Nanyang Technological University, 308232, Singapore

[5]Department of Hematology, Singapore General Hospital, Singapore 169856, Singapore

[6]Cancer & Stem Cell Biology, Duke-NUS Medical School, Singapore, Singapore

[7]Department of Hematology, National Cancer Centre Singapore, Singapore 168583, Singapore

[8]Regenerative Medicine Institute of Singapore, SingHealth Duke-NUS Academic Medical Centre, Singapore 169857, Singapore

[9]SingHealth Duke-NUS Institute of Precision Medicine, 169609, Singapore

[10]Cancer & Stem Cell Biology Program, Duke-NUS Medical School, 169857, Singapore

[11]Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117597, Singapore

[#] Co-first authors; * Corresponding authors.

Email: miles@gis.a-star.edu.sg, liuj3@gis.a-star.edu.sg

Abstract

Human reference genomes have been instrumental in advancing genomic and biomedical research, but South and Southeast Asian populations are underrepresented, despite accounting for a large proportion of world population. As a part of effort on generating reference genomes for these populations, we present the first gapless, telomere-to-telomere (T2T) diploid genome assembly created by using a trio sample set of Indian ancestry (I002C), with NG50 of 154.89 Mb and 146.27 Mb for the maternal and paternal haplotypes, including the fully assembled rDNA array for the maternal chromosome 21 and Y chromosome. With the Merqury QVs of 82.05, 83.08 and 82.64 for the maternal, paternal and haploid assemblies respectively, I002C represents the highest-quality human genome assembled in both diploid and haploid forms to date. Compared to CHM13, the I002C genome displays substantial sequence diversity, resulting in 14,943 structural variants, including 3,236 novel variants absent from public databases. Analysis of trio-phased haplotypes further revealed elevated inter-haplotype divergence within centromeric and subtelomeric regions, along with identification of differentially methylated regions (DMRs) as candidates for novel imprinting loci. As a result of substantial SVs between them, I002C is a more suitable reference than CHM13 for the genomic analysis of South Asian samples with less reference bias and better performance in mapping and variant calling, particularly for long read sequencing data. As the first high-quality T2T diploid reference genome for Indian, the largest world's population, I002C contributes to the growing set of population-specific reference genomes and helps to overcome a significant gap in human genome diversity.

Introduction

A human reference genome serves as the foundation for genomic analyses, such as gene and regulatory sequence annotations and variation analysis, making it an essential resource for biomedical and clinical research to understand human genetic variation, disease mechanisms, and evolutionary biology, and enable precision medicine. The most widely used reference genome, GRCh38, was assembled from a few donor individuals, primarily of African and European descent[1]. Despite its widespread use, GRCh38 still contains large unresolved regions, including ribosomal DNA (rDNA) arrays, as well as pericentromeric and subtelomeric regions[2], which are critical for understanding fundamental cellular processes. Recognizing these limitations, there have been several efforts on generating new human reference genomes with better completeness and quality, such as CHM13[2], HG002[3] and CN1[4], by taking the advantages of long read sequencing technologies and new algorithms for de novo assembling analysis. More recently, recognizing the limitation of current references in capturing the genetic diversity information, efforts have gone beyond the generation of complete, telomere-to-telomere (T2T) and phased reference genomes to build pangenome reference consisting of a panel of multiple reference genomes, such as the project led by the Human Pangenome Reference consortium (HPRC)[5,6].

These efforts are, however, mainly focused on analyzing individuals of European and East Asian ancestries, with over 85% of studies involving populations of European descent[7,8]. Despite representing around 34% of the world population, South and Southeast Asian populations are very much under-represented in these efforts, with only very few studies being carried out[9–17], and currently no complete T2T reference genome available. It has been increasingly recognized that alignment biases, errors in variant calling, misinterpretation of pathogenic mutations, and incorrect inferences may arise when the reference genome for analysis is very different from the samples under study[4,6,18]. These limitations of reference bias hamper the translation of genetic research to clinical applications in underrepresented populations and call for the generation of ancestry-specific T2T reference genomes.

Here we present the first high-quality, gapless, telomere-to-telomere, haplotype-resolved reference genome (I002C) with 44 autosomes and XY chromosomes for South Asia, using a biologically normal Indian sample, from a father-mother-child trio samples set recruited in Singapore. We employed state-of-the-art sequencing technologies, including PacBio HiFi circular consensus long-read sequencing and Oxford Nanopore ultralong sequencing, and advanced algorithms for de novo assembling, including Verkko[19] and Hifiasm[20,21] to construct chromosome-level contigs, including the long and highly repetitive centromeric regions and the short arms of five acrocentric chromosomes. We utilized both parental sequencing data information for phasing and generating the paternal and maternal haploid genomes. We further demonstrated that with this reference genome, Indian-specific novel variants can be identified with greater accuracy, and sequencing results from Indian samples can be analyzed with higher mapping rates and lower clipping rates. This effort has helped to overcome a significant gap in the presentation of human genome diversity by providing the first reference genome for the Indian population, the largest one in the world that accounts for 72% of South Asian populations.

Results

Sample and Sequencing Data

To assemble a diploid, telomere-to-telomere (T2T) reference genome, a healthy Singaporean male and corresponding parental samples with self-reported Indian ancestry for all the grandparents were recruited. Karyotyping at a resolution of 450 bands confirmed the diploid nature of the child sample (I002C) [Additional file 1: Figure S1]. Using the DNAs of the child sample cell line, we generated 318.62 Gb (103x) of Pacbio HiFi, 193.66 Gb (62x) of Nanopore duplex, and 669.94 Gb (216x) Nanopore ultra-long reads. In addition, the short-reads dataset included 297.61 Gb (96x) from MGI, 371.18 Gb (119x) OmniC and 22 Gb total RNASeq from the Illumina platform (Additional file 2: Table S1). PacBio HiFi and Nanopore duplex reads were also generated for the two parental samples (Additional file 2: Table S1). The reference genome generated from this child sample was intended to represent one of the most underrepresented populations in global genomics—South Asians (SAS) (Figure 1). Local ancestry analysis based on data from the 1000 Genomes Project[22] revealed that over 90% of the I002C genome is of SAS origin, while the second-largest contribution is from European populations, along with scattered markers from East Asian and American, consistent with previous findings[23] (Additional file 1: Figure S2). Further comparison of ancestry markers between the two haplotypes shows diversity, highlighting inter-haplotype variation. For example, on chromosome 15, the region distal to the centromere is inferred as American in origin on the maternal haplotype, whereas the corresponding paternal region is of European ancestry (Figure 1).

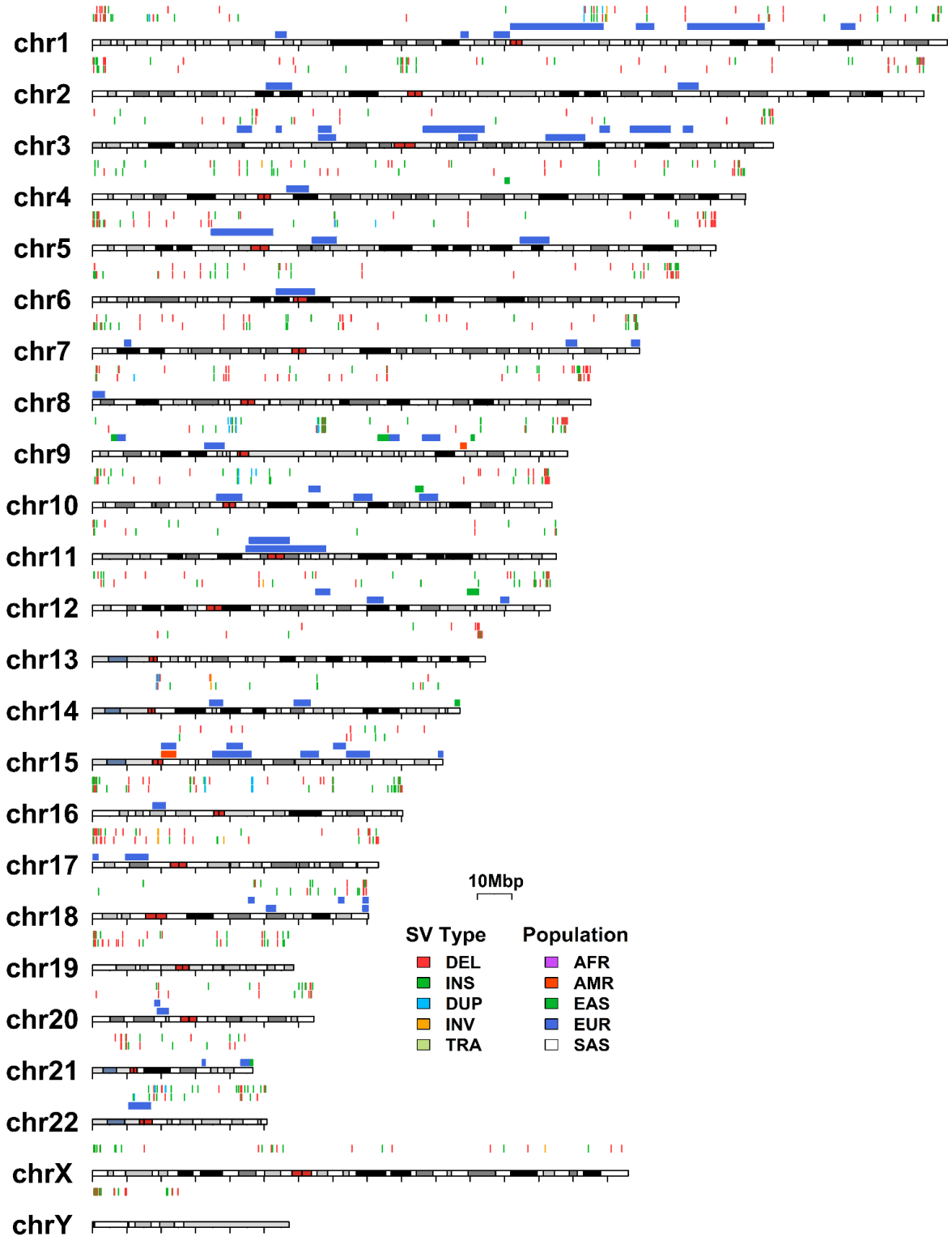

Figure 1: Ancestral composition and unique structural variants of I002C. Overview of ancestry markers and structural variants across I002C haplotypes. Chromosomal

regions are coloured based on predicted ancestral marker and structural variant types. For each chromosome (tracks from bottom to top), (1) ancestral markers for the maternal haplotype; (2) ancestral markers for the paternal haplotype; (3) structural variants unique to the maternal haplotype and (4) structural variants unique to the paternal haplotype, identified relative to CHM13 and absent from gnomAD, HGSVC, and HPRC. **Populations:** AFR, African; AMR, American; EAS, East Asian; EUR, European; SAS, South Asian. **Structural variant types:** DEL, deletion; INS, insertion; DUP, duplication; INV, inversion; TRA, translocation.

A telomere-to-telomere diploid assembly, polishing, and assessment

Initial assemblies were generated using Verkko[19] and Hifiasm[20,21] in trio mode. For each haplotype, we selected chromosome-scale sequences from either assembler based on quality, contiguity and presence of telomere motif on both ends. This resulted in 20 maternal chromosomes (excluding the p-arms of chromosomes 13, 15 and 22) with 39 gaps, and 19 paternal chromosomes (excluding the p-arms of chromosomes 4, 13, 14 and 15) with 14 gaps. The gaps were closed using combinatorial methods, including TGS-Gapcloser[24], contigs from local reassemblies of binned reads and ultra-long read alignments (Additional methods: **Gap-filling**). The remaining telomeres were patched using subtelomere ultra-long read mapping and local assemblies (Additional methods: **Telomere patching and extension**) (Additional file 2: Table S2). The rDNA arrays were verified and corrected by local re-assembly and manual curation (Additional methods: **rDNA**). The final assembly includes full T2T sequences for all 22 autosome pairs, both sex chromosomes (X and Y), and the mitochondrial genome.

To further improve assembly quality, we performed multiple rounds of polishing using binned HiFi, ONT reads (duplex, corrected and uncorrected ultra-long), and MGI short reads (Additional methods: Read binning, Additional file 2: Table S3-S4). The polishing workflow included two rounds of automated polishing[25], one round of manual curation (Additional file 1: Figure S3), and one round of polishing using assemblies from corrected ultra-long reads. This process increased the assembly quality (QV) from an initial 68.05 to 82.65 (Figure 2a), as assessed by Merqury[26] using hybrid k-mers obtained from MGI

short reads and HiFi reads. For more than two-thirds of the chromosomes in both maternal and paternal haplotypes, the Merqury QV exceeded the genome-wide average of 83, with 25 chromosomes reaching a QV of 100 or infinity, indicating full k-mer support from the input reads (Additional file 1: Figure S4). A range of evaluation tools confirmed the high quality of the final I002C assembly (Additional file 1: Figure S5-S6). Notably, 99.24% (5.93 Gbp) of the genome was flagged as haploid and error-free, while 44 Mbp (0.75%) were identified as duplicated or collapsed regions, and only 0.93 Mbp (0.02%) showed coverage-based errors detected by Flagger[6] (Figure 2b). In addition, we curated platform-specific (HiFi, Duplex and ultra-long) coverage-bias and clipped-read regions from mapping of binned reads, and annotated them following previously described method[25] (Additional file 1: Figure S7). The final polished I002C genome achieved an NG50 of 154.89 Mb for the maternal haplotype and 146.27 Mb for the paternal haplotype, comparable to previous high-quality human assemblies (Figure 2c). With Merqury QV scores of 82.05 for the maternal and 83.08 for the paternal assemblies, I002C represents the highest-quality diploid human genome to date. A haploid version of the I002C genome was also constructed by selecting the higher-quality assembly for each autosome and both sex chromosomes (X and Y) (Additional file 2: Table S5), resulting in a QV of 82.64. To our knowledge, I002C is the highest-quality diploid and haploid human genome currently available.

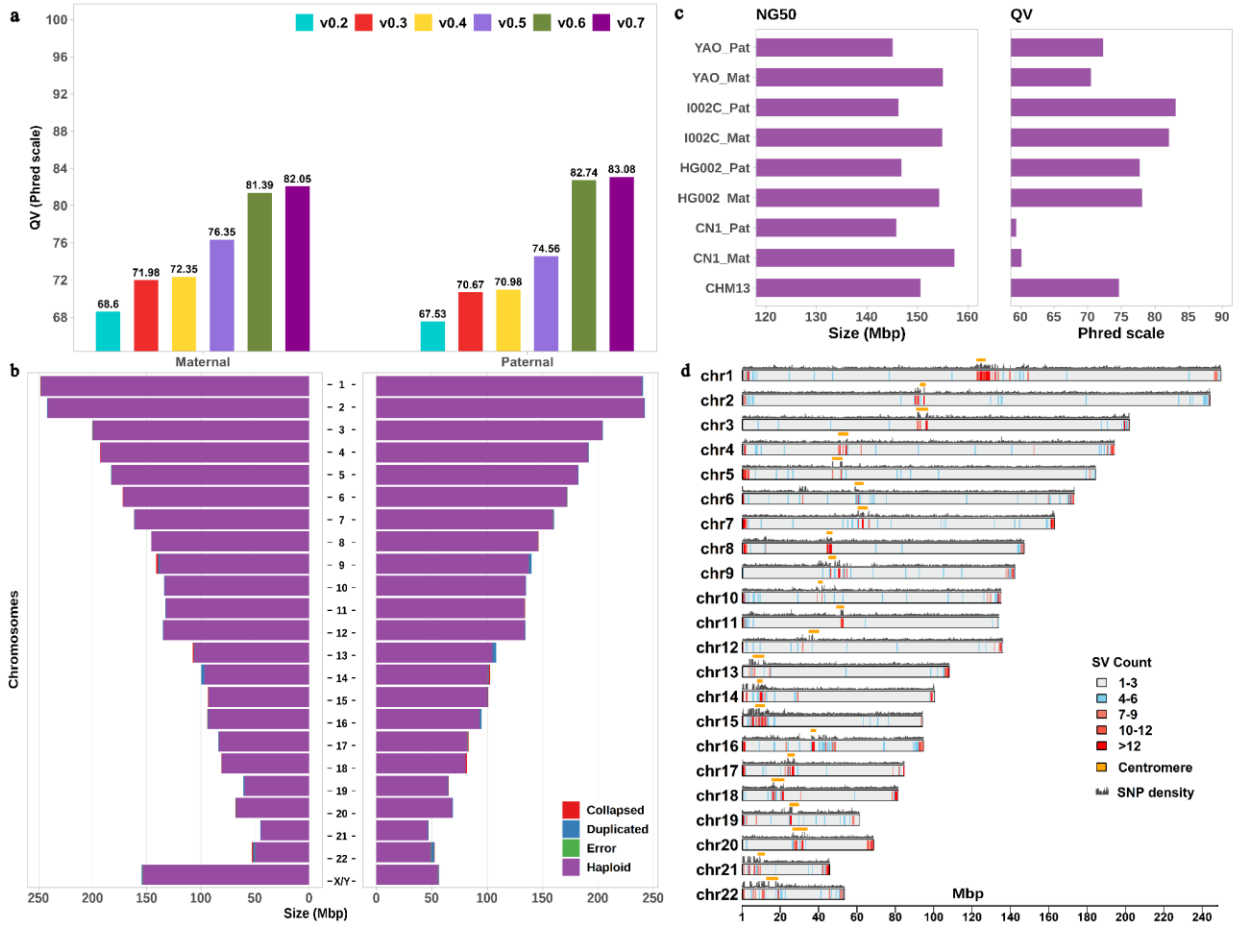

Figure 2: Overview of the haplotype-resolved I002C genome. a) Progression of I002C genome versions with polishing, showing improvements in QV scores assessed using hybrid k-mers. b) Assessment of haplotype reliability based on read mapping. For each chromosome, the left bar represents the maternal haplotype and the right bar the paternal haplotype. Regions marked as haploid (purple) are considered reliable, comprising over 99% of each haplotype on average. c) Comparison of NG50 and QV values for I002C haplotypes with other high-quality human genomes including YAO, HG002, CHM13 and CN1. d) Genome-wide heterozygosity profile of I002C based on structural variant counts in 500 kb windows, overlaid with SNP density between maternal and paternal haplotypes. Higher diversity is observed near centromeric regions (highlighted by gold bars above the SNP density track).

Annotation of Repeats, Centromeres and rDNAs

The haplotype-resolved genome enabled the annotation and analysis of repeat variation between the two haplotypes and the comparison with CHM13[2]. Genome-wide repeat content in I002C (including 22 autosomes and XY) was comparable to CHM13, with 53.29% of sequences being repetitive in the maternal and 53.43% in the paternal genomes, relative to 54.2% in CHM13 (Additional file 2: Table S6). We focused on the repeat families from peri/centromeric regions, where alpha-satellite (αSat), beta-satellite (βSat), gamma-satellite (γSat), and human satellites 1–3 (HSat1–3) together accounted for 5.09% and 5.96% of the maternal and paternal genomes, respectively, compared to 6.45% in CHM13 (Additional file 2: Table S7-S8). The annotated αSat arrays ranged from 1.40 to 6 Mbp in the maternal and 0.49 to 7.80 Mbp in the paternal genomes. These lengths, when compared to 165 samples from previous studies[27,28], highlighted variation in both haplotypes and across individuals (Figure 3a).

To assess the integrity and completeness of the annotated centromeric regions, we first mapped the 10 kb peri-centromeric regions from CHM13 to both I002C haplotypes and confirmed the presence of corresponding loci on each chromosome (Additional file 2: Table S9-S10). Next, all long-read datasets (HiFi, duplex, and ultra-long ONT) were aligned to the combined I002C reference, revealing uniform coverage across the annotated centromeric regions, consistent with accurate representation and assembly continuity (Additional file1: Figure S8-S9). Centromere repeat annotation showed that the overall repeat composition in I002C is similar to CHM13, mainly consisting of αSat made up of high-order repeats (HORs) classified as active/live, inactive, divergent, and monomeric, along with βSat, γSat, and HSat1–3. At the individual chromosome level, we observed the variations in repeat families and repeat array lengths (Figure 3b). On average, the active αSat arrays in I002C are 1.33 times longer than in CHM13. In particular, nine autosomes from both haplotypes and the X and Y chromosomes of I002C are more than 1.8 times longer, with chromosome 21 showing the largest difference of 6.4-fold (Additional file 2: Table S11). However, these differences fall within the known range of human variation[23] (Figure 3a). While satellite length generally differed, the largest difference was observed in the I002C paternal chromosome 14 HSat1, which

spans 1.8 Mbp compared to 143.8 Kbp in CHM13. In contrast, the maternal chromosome 14 monomeric αSat measured 6.1 Kbp, substantially shorter than the 91.6 Kbp in CHM13. Despite these length differences, most peri/centromeric satellites are conserved between the I002C and CHM13 genomes, although some are uniquely present or absent. For example, maternal chromosome 14 and paternal chromosome 21 carry around ~2 Kbp of HSat2 not found in CHM13, and both I002C haplotypes also include about ~2 Kbp of HSat2 on chromosome 15 and 55 bp of γSat on chromosome 16 not present in CHM13 (Figure 3b, Additional file 1: Figure S10-S13).

To further explore the αSat HOR array structure between genomes, we compared centromeres of chromosome pairs. This analysis showed differences in HOR organization between CHM13 and I002C genomes that deviated from canonical patterns[29,30], specifically HOR SVs (Additional file 2: Table S12). Notably, some HOR monomers found in I002C were completely absent from CHM13. For instance, CHM13 chromosome 1 was primarily composed of 6-mer HORs such as S1C1/5/19H1L.1-5(_6/4_5){n}-6, S1C1/5/19H1L.1-6 while I002C haplotypes contained distinct 2-mer variants like S1C1/5/19H1L.5_6/4 and S1C1/5/19H1L.5-6, which were missing from CHM13 (Figure 3c). The active HOR array of chromosome 17 in CHM13 mainly consists of the canonical 16-mer, whereas I002C haplotypes contain a mix of variant 16-mers along with 15-, 14-, and 13-mers. Notably, a novel 13-mer HOR variant (S3C17H1L.1-10_15_15-16) was found in both I002C haplotypes but not in CHM13 (Figure 3d). In contrast, chromosome 21 in both I002C haplotypes was primarily made up of 5-mer variants, compared to the canonical 11-mer in CHM13. A 5-mer variant (S2C13/21H1L.1-5) observed in I002C was completely absent in CHM13, and additional HORs unique to the paternal haplotype were not found in either the maternal genome or CHM13 (Figure 3e). Despite these differences, monomers within these HOR variants remain sequence-identical, suggesting structural polymorphism without extensive sequence divergence.

Similar differences were observed among the I002C maternal and paternal haplotypes (Additional file 1: Figure S14). Overall, 52% of the unique HOR SVs were specific to either the maternal or paternal haplotype. The dominant αSat HORs also differed between

haplotypes—for instance, chromosome 5 had an 8-mer and a 4-mer; chromosome 10 showed a 6-mer and an 8-mer; and chromosome 17 carried a 16-mer and 13-mers in maternal and paternal haplotypes, respectively (Figure 3b). Together, these findings highlight the polymorphic nature of centromeric components among different human genomes and across haplotypes within the same genome.

Taking advantage of single-molecule long-read sequencing, which preserves native DNA methylation signals, we investigated methylation patterns across centromeric regions of both haplotypes. To identify putative kinetochore regions within the centromeres, we analyzed methylation patterns for signatures associated with centromere function. A characteristic centromeric methylation dip—marked by a sharp decrease in methylation levels within the active HOR subregions relative to their flanking sequences—was observed consistently across all chromosomes (Figure 3f, Additional file 1: Figure S15). The size of the methylation dip ranged from 0.1 to 0.4 Mb, consistent with previous reports in CHM1 and CHM13[31] (Additional file 1: Figure S16).

Ribosomal DNA (rDNA) arrays located on the short arms of acrocentric chromosomes are among the most challenging regions to assemble due to their high repeat content, extensive segmental duplications, and complex satellite structures. The total rDNA copy number (CN) in I002C was estimated to be 416 for the diploid genome using digital PCR, which was consistent with estimations from the k-mer-based approach[2]. Based on the coverage of binned rDNA reads, we estimated 157 and 259 CNs for maternal and paternal haplotypes, respectively. Chromosome-level rDNA CNs were inferred by integrating read clustering with coverage-based and fluorescence in situ hybridization (FISH)-based estimates, and further cross-validated using independent FISH results (Additional methods: rDNA). Comparison of the estimated CN with the CN observed in the assembled genome revealed that rDNA arrays were underrepresented for most of the chromosomes, except for maternal chromosomes 13 and 21. The maternal chromosome 21 rDNA sequence was found to be completely assembled in the initial assembly, supported by coverage based estimates, FISH estimates, and bridging ultra-long reads. As for maternal chromosome 13, the reassembled rDNA sequence was nearly identical

to the initial assembly in terms of both sequences and CNs, despite the absence of bridging ultra-long reads and a higher CN estimate than indicated by FISH. The rDNA sequences from the initial assembly were retained for maternal chromosomes 13 and 21. As for the remaining acrocentric chromosomes, due to the difficulty of resolving base-level variation among highly similar rDNA units, rDNA morphs, defined as representative sequences of one or more nearly identical units, were identified through further read clustering (Additional methods: rDNA). Arrays were reconstructed using consensus morph sequences scaled by the estimated CN. These reconstructed arrays were then incorporated into the final assembly by replacing the originally assembled rDNA segments (Figure 3g).

In general, the typical rDNA array structure consisted of consecutive ~45 kb rDNA units, flanked by a ~4 kb rDNA fragment upstream and a ~6 kb rDNA fragment downstream, and further bordered by the distal and proximal junctions (DJ, PJ). In total, we identified 19 major rDNA morphs, including 7 maternal and 12 paternal morphs in I002C. Comparison with the 8 major morphs in CHM13 revealed sequence diversity, with all major morphs in I002C showing >200 edit distance[32] from their CHM13 counterparts (Additional file 1: Figure S17). Despite this divergence, several large SVs were shared between I002C and CHM13 major morphs when compared to the canonical rDNA reference (Additional file 1: Figure S18). For example, a 1,530 bp deletion in the tandem repeat block 1 (TR1) is shared among I002C Mat_chr22_A, chm13-model15c and chm13-model22a morphs. In contrast, a 988 bp deletion in the long repeat (LR) region present in chm13-model15b and chm13-model15c was not observed in I002C major morphs. Additionally, we identified a ~900 kbp deletion in the TR2 region of I002C Pat_chr22_B1 and Pat_chr22_B2 that is not present in CHM13 major morphs.

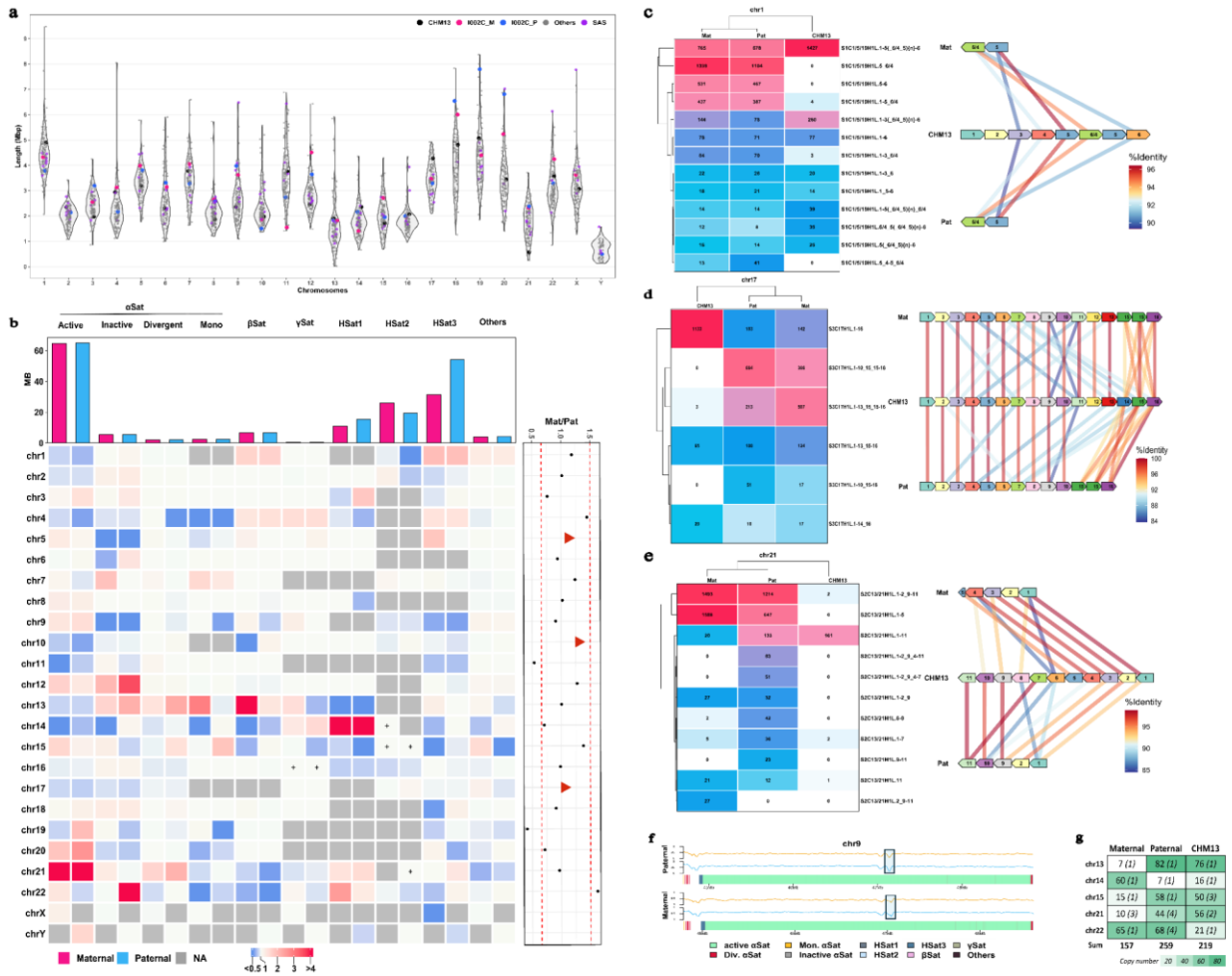

Figure 3: Peri/Centromeric satellite diversity, methylation, and rDNA variation among the I002C maternal, paternal and CHM13. a) Comparison of HOR lengths across I002C haplotypes and 165 human genomes. **b)** Heatmap showing the ratio of peri/centromeric satellite lengths in I002C haplotypes relative to CHM13. "+" denotes the presence of a satellite repeat in I002C but absence in CHM13. A dot plot compares active α-satellite (αSat) lengths between maternal and paternal haplotypes of I002C; chromosomes with >1.5-fold difference are marked with red dotted lines, and those with distinct HOR structures are indicated with triangles. **c-e)** Heatmaps of the top 10 most abundant HORs that differ structurally between CHM13 and I002C haplotypes. Structural variation of the most abundant HORs and their sequence similarity are shown. **f)** DNA methylation profiles across centromeric regions, measured from PacBio HiFi (blue) and Oxford Nanopore (yellow) data. Methylation levels (range 0–1) are averaged in 10 kb windows and normalized using adjacent windows. Centromere dip regions (CDRs) are

indicated by shaded boxes. **g)** Heatmap of ribosomal DNA (rDNA) copy number and major morph count [CN (morph count)] across I002C haplotypes and CHM13.

Chromosome Y analysis

Assembling the human Y chromosome to a T2T chromosome remains challenging due to its highly repetitive architecture, especially within large heterochromatic regions and was only recently completed[18]. The I002C assembly provides a T2T, gapless Y chromosome from a South Asian individual, representing a valuable addition to the current catalogue of Y chromosome diversity[33]. The total length of I002C-Y is 56.6 Mb, consistent with other recent Y chromosome assemblies ranging from 45 to 85 Mb[33] (Additional file 1: Figure S19). Notably, it is comparable in size to other South Asian individuals, including HG03009 (58.8 Mb), HG02492 (51.66 Mb), HG03732 (51.5 Mb), and HG03492 (51.2 Mb).

Comparative analysis of I002C-Y and T2T-HG002-Y showed concordance in overall structure, with the largest differences observed in ampliconic and heterochromatic subregions (Additional file 2: Table S13). Specifically, the AMPL2 subregion in I002C-Y is 11.7% (375 kb) shorter, the centromeric HET1_centro2 is 18.5% (327 kb) longer, HET2_DYZ19 is 52.7% (140 kb) longer, and HET3_Yq12 is 18.1% (6.25 Mb) shorter than their counterparts in T2T-HG002-Y. The multi-megabase variation observed in Yq12 is expected, as this region is composed of alternating repeat arrays known to be the largest heterochromatic block in the human genome and exhibits extensive inter-individual variability[33]. Sequence identity between subregions of I002C-Y and T2T-HG002-Y exceeds 97% for most regions. The most divergent region is HET2_DYZ19, with 89.4% identity, followed by centromeric regions HET1_centroDYZ3 and HET1_centro2, both below 95%.

Finally, haplogroup assignment based on I002C-Y sequence places this individual in the R-Y6 lineage, a Y-chromosome haplogroup prevalent in South Asian populations and

estimated to have emerged ~4,000–4,500 years ago, underscoring the deep-rooted paternal ancestry in the region.

Structural Variation in I002C Relative to CHM13 and Between the Paternal and Maternal Haplotypes

In addition to variations in peri/centromeric repeats, whole-genome comparison between I002C haplotypes and CHM13 identified 14,943 structural variants (SVs; ≥50 bp) spanning ~19.50 Mbp in the maternal and 14,243 SVs spanning ~18.74 Mbp in the paternal genome, with ~95% of these being indels (Additional file 2: Table S14–S15). Comparing I002C SVs against the combined SVs from public databases, including gnomAD[34], HGSVC[27,35], and HPRC[6], showed that approximately 13.63% (3,236) of SVs were found to be unique to I002C (Additional file 2: Table S16). These I002C-specific SVs were primarily duplications and deletions (~34% each), followed by insertions (30%) (Figure 1; Additional file 2: Table S16). Further repeat annotation showed that 72% of these unique SV regions overlapped with various repeat types (Additional file 2: Table S17).

Comparative analysis of the two complete haplotypes revealed approximately 2.90 million SNVs, 329,328 small indels (<50 bp), and 16,308 SVs. The heterozygosity rate, calculated as the number of SVs per 500 kb, showed that highly divergent regions are enriched in centromeric and subtelomeric regions (Figure 2d). In these regions, the SV density was more than four fold higher, and the SNV density more than two fold higher, compared to the genome-wide average (Additional file 1: Figure S20).

Population-specific Reference Reduces Variant Calling Bias

To evaluate the impact of reference bias on read mapping and structural variant (SV) calling, we compared three complete haploid reference genomes: two I002C haplotypes (maternal and paternal, including mitochondrial DNA) and CHM13, using both short-read (NGS) and long-read (ONT) sequencing data. We randomly selected a subset of publicly

available NGS data from the 1000 Genomes Project (1KGP)[22], comprising 50 individuals from each of five South Asian (SAS) subpopulations (BEB, GIH, ITU, PJL, STU). Additionally, we used all available ONT data from the Singapore HELIOS study[36], including 26 Chinese, 31 Indian, and 37 Malay individuals. Raw sequencing reads from these individuals were mapped to each of the three reference genomes, and SVs were identified from each alignment.

We observed that the rate of uniquely mapped reads was higher on CHM13 for NGS data in SAS populations compared to either the maternal (median difference ranged from 0.06%-0.12%) or paternal I002C haplotype (median difference ranged from 0.01%-0.07%), but significantly lower on CHM13 for ONT data from Singaporean samples (median difference 0.05%-0.12%) (Additional file 1: Figure S21). Interestingly, in the same SAS samples with short-read data, the number of detected SVs was generally lower when using the I002C haplotypes compared to CHM13 (Additional file 1: Figure S22). An exception was observed in the PJL subgroup, where a slightly higher number of SVs was identified using the paternal genome. A similar trend was observed in long-read data across the three Singaporean populations, although the number of SVs was slightly higher in Chinese samples mapped to the maternal genome (Additional file 1: Figure S22).

We then assessed CHM13-specific SVs after projecting their coordinates to GRCh38. SVs were merged across all SAS samples (183,811 SVs) and all Singaporean samples (138,909 SVs), respectively. Of these, 21.2% (39,051 SVs) in SAS and 15.4% (21,353 SVs) in Singaporeans were unique to CHM13. Of these CHM13-specific SVs, 5.5% in SAS and 18.8% in Singaporeans were shared across at least 50% of the population. These additional SVs identified using CHM13 as the reference are likely variants that are either fixed or highly prevalent in the SAS or Singaporean populations and thus also present in the I002C haplotype genomes, while others may also be unique to CHM13. Notably, 71 and 152 medically relevant genes[37] were affected by fixed or common SVs

observed in over 90% of SAS or Singaporean individuals, respectively, indicating potential population-specific relevance in clinical interpretation. Notably, the gene TSC2—included in the ACMG SF v3.2[38] list of actionable genes recommended for opportunistic screening was affected by an SV present in 99% of the SAS population or 94% of Singaporean individuals. Additionally, two other ACMG genes, TRDN and STK11, were affected by SVs present in 100% of Singaporean individuals. Further, OPRM1 and DPYD, both included in the Clinical Pharmacogenetics Implementation Consortium (CPIC) gene panel, were associated with SVs commonly found in 87% of SAS samples and 91.5% of Singaporeans. These findings suggest that the I002C haplotypes more accurately represent the genomic architecture, particularly the SVs, of the SAS population or even EAS population and help mitigate reference bias in SV analyses.

Genome-wide DNA Methylation Analysis

Besides the analysis of methylation patterns within the centromere regions, we also analyzed the genome-wide methylation pattern in the I002C genomes. Single-base resolution DNA methylation patterns were extracted from both Nanopore R10.4.1 and PacBio Revio data. The results showed high concordance, with a Pearson correlation coefficient of 0.953 when analyzed using 1000 bp sliding windows with 100 bp steps (Additional file 2: Table S18).

To leverage the haplotype-resolved methylation data, we identified 71,020 single-CpG differentially methylated loci (DMLs) between the paternal and maternal haplotypes. These DMLs were not randomly distributed across the genome; rather, they often clustered into regions with consistent patterns of haplotype-specific methylation. Such uniformly differentially methylated regions (DMRs) may have functional significance, as observed in imprinting regions associated with well-characterized imprinting disorders, including Prader-Willi syndrome (PWS)[39,40], Angelman syndrome (AS)[40], Beckwith-Wiedemann syndrome (BWS)[41], and Russell-Silver syndrome (SRS)[42]. Using these known imprinting regions as references, we calibrated our thresholds and identified 980 DMRs (Additional file 1: Figure S23). To evaluate the robustness of our approach, we

intersected these DMRs with 93 previously reported imprinting regions supported by multiple independent studies[43–48]. Reassuringly, all 53 overlapping regions (Additional file 1: Figure S24, Additional file 2: Table S19) exhibited the same methylated allele as reported, demonstrating strong concordance, including nine regions linked to the four well-characterized imprinting disorders. The remaining 40 reported regions were not detected, likely due to the stringency of our criteria, which required a methylation difference ≥97.5% at each DML and a minimum of five DMLs per DMR[48].

To identify novel candidate imprinting regions, we intersected the 980 DMRs with experimentally validated core promoter regions of all transcripts in GRCh38, where DNA methylation is known to strongly influence gene expression. This approach resulted in 193 candidate transcripts corresponding to 127 unique genes whose expressions would potentially be regulated by imprinting in the I002C cell line (Additional file 1: Figure S25, Additional file 2: Table S20). Of these 127 genes, 12 have been previously reported as imprinted[43–48].

Gene Annotation

The LiftOver[49] successfully transferred 96% of CHM13 v2 gene model (RefSeq/Liftoff annotation v5.1) to the I002C maternal and 93% to paternal assembly (Additional file 2: Table S21-23). To complement this, a protein-to-genome mapping analysis using UniProt[50] reference proteins was performed. Twenty-two proteins failed to align to any locus in the I002C assemblies (Additional file 2: Table S24). Of these, 20 were also unmapped in CHM13, suggesting that these proteins may correspond to sequences absent from both references or to alternatively spliced isoforms not represented in current annotations. An additional 16 proteins mapped to the expected loci in I002C but showed reduced sequence identity (<80%) (Additional file 2: Table S24). In these cases, alignments to CHM13 confirmed the presence of the corresponding sequences, while the I002C assemblies contained small indels in the affected regions. Manual inspection

confirmed that these indels are supported by long-read data (Additional file1: Figure S26), indicating they likely represent genuine haplotype-specific variants rather than assembly errors.

To facilitate the gene annotation analysis, we also generated a total RNA-seq data for I002C using cultured cells. Combining the TSEBRA[51] annotated gene predictions, supported by both protein and RNA-seq evidence with the LiftOver annotations, contributed 1,249 and 1,366 additional known protein-coding genes that can be mapped to the paternal and maternal haplotypes, respectively. These numbers represent genes that were successfully annotated in the assemblies but may not be entirely novel, as they could correspond to known genes that were previously unannotated due to haplotype-specific variation or sequence divergence. This integrated approach improved annotation completeness while minimizing redundancy, leading to the final numbers of 179784 transcript counts and 44496 genes for maternal haplotype, of which there are 20213 coding and 23478 non_coding ones; and 174847 transcripts and 43691 genes for paternal haplotype, of which there are 21062 coding and 23434 non-coding ones (Additional file 2: Table S23,S25).

Discussion

In this study, we present the first telomere-to-telomere (T2T), fully phased diploid genome assembly of an Indian individual using a trio sample set of Indian ethnicity from Singapore. The majority of Singaporean Indians descend from migrants during the British colonial period, primarily of Telugu and Tamil origin, with smaller contributions from Northern groups such as Sikhs and Pathans[11]. As such, the I002C genome reflects broader Indian subcontinental ancestry and offers value to individuals of similar backgrounds in India and worldwide. Additionally, it contributes to improving representation for these populations in human genomic datasets.

Using high-quality long read data, trio-aware assembly, ultra long-read gap filling, and multi-platform polishing, we achieved complete, gapless assemblies of all the autosomes,

sex chromosomes, and a fully resolved circular mitochondrial genome representative of the individual. The final assemblies reached NG50 values of over 154 Mb for maternal 146 Mb for paternal and Merqury QV scores exceeding 82 for both haplotypes, establishing I002C as the highest-quality diploid and haploid human genome assembled to date.

To the best of our knowledge, the maternal chromosome21 rDNA sequence of the I002C was the first fully assembled rDNA cluster sequence that has been published. Furthermore, I002C also contributes one of the few complete Y chromosome assemblies from a South Asian individual, expanding current knowledge of Y-linked diversity and structural variation.

In addition, the fully resolved diploid genome enabled the generation of haplotype-resolved methylome. This allowed validation of known imprinting-associated DMRs and the identification of novel candidates, supported by allele-specific expression and promoter methylation patterns. Together, these findings highlight the value of fully phased diploid genomes for exploring complex and polymorphic regions.

As the first reference genome for the South Asian population, I002C highlights important features that remain incompletely characterized in many reference genomes. For example, in addition to providing the first fully assembled rDNA cluster on Chr21 (maternal), I002C also revealed 19 major rDNA morphs, compared to the 8 major rDNA morphs for CHM13. Comparison of I002C genome to CHM13 revealed that over 13% of SVs in I002C are novel, with many falling in repeat-rich regions, which are often sources of structural complexity in the genome. When comparing two I002C haplotypes, the analysis revealed 2.9 million SNVs, more than 329 thousands small indels and more than 16 thousands SVs. Many of these variants were concentrated in centromeric and subtelomeric regions, identifying them as hotspots of structural divergence.

Our analysis has demonstrated that the I002C is a better reference than CHM13 for the genomic analysis of Indian samples as well as other Asian samples by reducing population bias. For example, population-specific analysis, based on 314 randomly selected individuals from the 1000 Genomes Project, showed comparable rates of

uniquely mapped reads across references. However, the number of called SVs was reduced when mapping SAS samples to I002C instead of CHM13, reflecting the presence of population-specific sequences in I002C. The additional SVs identified with CHM13 included variants in medically relevant genes, such as TSC2, which was affected in 99% of SAS samples.

By providing a complete, ancestry-relevant reference genome, I002C improves the representation of South Asians in genomic research and enhances the resolutions of both structural and regulatory genomic variation. In addition to supporting more accurate variant interpretation, it refines epigenetic and imprinting analyses. As a first step of our efforts on generating reference genomes for South and Southeast Asian populations, I002C will serve as a linear backbone for constructing locally and Asian relevant pangenomes and contribute to the growing set of population-specific references that are critical for reducing global reference bias and advancing equitable precision medicine.

Methods

Sample selection and sequencing:

The details of sample selection, Pacbio HiFi, Oxford Nanopore duplex, ultra-long and Omni-C data generation are described in the previous study[52].

Karyotyping

Cells were cultured in RPMI 1640 medium (#11875-093, Gibco, USA) supplemented with 10% fetal bovine serum (#A5670801, Gibco) and 1% penicillin–streptomycin (#15140-122, Gibco), and maintained in suspension at 37 °C in a humidified incubator with 5% CO_2. Approximately 5 million cultured cells were treated with KaryoMAX™ Colcemid™ Solution in HBSS (#15210040, Gibco) for 2 hours to induce mitotic arrest. Following standard protocols from *The AGT Cytogenetics Laboratory Manual* (4th Edition), routine G-banded metaphase preparations were performed[53].

Karyotype analysis of 20 cells was conducted using the Ikaros karyotyping software (Metasystem) at a 450-band resolution, following the International System for Human Cytogenomic Nomenclature (ISCN 2020) guidelines.

RNA Sequencing

RNA was extracted using the QIAGEN PAXgene Blood RNA Kit according to the manufacturer's instructions. Total RNA was used as input material for the RNA sample preparations. rRNA was removed from total RNA by using TIANSeq rRNA Depletion Kit (Animal). Fragmentation was carried out using divalent cations under elevated temperature in the First Strand Synthesis Reaction Buffer. First strand cDNA was synthesized using random hexamer primer and reverse transcriptase, followed by the synthesis of the second strand by adding buffer and dNTPs (with dTTP replaced by dUTP). The synthesized double-stranded cDNA was treated by terminal repair, A-tailing, adaptor ligation, and PCR enrichment after fragment selection. The strand-specific library was obtained by selecting cDNA fragments of 370-420bp by purification with AMPure XP beads. The qualified library was pooled and sequenced on the Illumina NovaSeq X Plus platform with a 150 × 2 paired-end configuration.

WGS Analysis by Next Generation Sequencing (NGS)

WGS PCR-free libraries for the gDNA samples were prepared using MGIEasy FS PCR-free DNA library prep set (MGI; Cat. no. 1000013455) as per manufacturer’s protocol. The WGS libraries were quality checked by the HSD 1000 Screentape and reagent (Agilent Technologies; cat. no. 5067-5585 and 5067-5584) on the Tapestation 4200 (Agilent Technologies). The single strand circularized DNA product of the library was measured with Qubit™ ssDNA HS Assay Kit (Thermo Fisher Scientific; Cat. no. Q10212) on the Qubit fluorometer 4 (Thermo Fisher Scientific). DNA nanoball (DNB) was made for each library from the single strand circularized DNA using the DNBSEQ DNB Rapid Make Reagent Kit (FCL PE150) (MGI; 1000028453) and concentration of DNB were determined by the Qubit™ ssDNA HS Assay Kit (Thermo Fisher Scientific; Cat. no. Q10212) on the Qubit fluorometer 4 (Thermo Fisher Scientific). DNBs were pooled based on the

sequencing data requirement, and the pooled DNBs were sequenced 150bp paired-end (PE150) on the MGI DNBSEQ-T7 sequencer using the DNBSEQ-T7RS High-throughput Sequencing Kit (FCL PE150) V1.0 sequencing reagent (MGI; Cat. no. 1000016106) and DNBSEQ-T7RS Sequencing Flow Cell (T7 FCL) (MGI; Cat. no. 1000016269).

Genome Assembly and Polishing

The initial assemblies were constructed using PacBio HiFi + uncorrected ONT ultra-long reads in trio mode with Verkko (v1.4.1)[19] and Hifiasm (v0.19.6-r595)[20,21] algorithms. Contigs from acrocentric chromosomes were further assembled into full-length chromosomes through manual curation with 3D-DNA[54] and a template-based approach using the CHM13 genome to determine the correct contig order for joining.
To support gap filling and manual curation, ONT ultra-long reads were also corrected using HERRO[55] and assembled in trio mode with Verkko. In addition, all long reads were haplotype-binned and reassembled using both Hifiasm and LJA (v0.2)[56].
Some gaps were falsely introduced during the scaffolding process, as confirmed by long-read and Hifiasm and LJA reassembled contig alignments to the assembly gap regions. These false gaps appeared as large deletions in the alignments and were subsequently removed. We considered all the remaining gaps to be real gaps. Real gaps were first closed using TGS-GapCloser (v1.2.1)[24]. The remaining gaps were manually filled by aligning corrected ultra-long reads, Verkko HERRO reads assembly, as well as Hifiasm and LJA reassemblies to the 10 kb flanking regions around each gap.
Alignments were manually examined in IGV (v2.15.2)[57], and only reciprocal best hits were selected for gap filling. Telomeres were patched and extended using ultra-long, duplex and HiFi reads containing the (TTAGGG)n motif at either end, identified using Tel_Sequences[58]. These reads were aligned to the first or last 250 kb of each chromosome, and consensus sequences were generated using SPOA (v4.1.4)[59] for patching, resulting in 23 gapless telomere-to-telomere chromosome level assemblies for each haplotype.

The assembly was polished through two rounds of automated processes using the T2T polishing pipeline[25], followed by one round of manual curation and an additional round of

ONT-based assembly patching. In brief, haplotype-specific binned long reads (HiFi, duplex and ultra-long) reads were mapped to their respective haplotypes using winnowmap2 (v2.03)[60] with default platform-specific parameters. PCR-free MGI short reads were aligned using BWA MEM (v0.7.17-r1188)[61], and the resulting sorted HiFi and short alignment files were merged using SAMtools (v1.18)[62]. SNV + indel (<50 bps) were called using DeepVariant (v1.6.1)[63] in "HYBRID_PACBIO_ILLUMINA" for merged HiFi, short reads and in “ONT_R104” mode for duplex reads, while ultra-long reads were processed with PEPPER-Margin-DeepVariant (v0.8)[64]. SVs were identified using Sniffles2 (v2.2)[65] and subsequently refined and merged using the Iris (v1.0.5) - Jasmin (v1.1.5)[66]. These variants were filtered using merfin (v1.0)[67] and used for genome polishing with bcftools (v1.10.2)[68]. During the manual curation round, variants were identified as previously described and carefully examined in IGV for genome polishing. Subsequently, remaining erroneous regions within 1 kb window were merged using BEDTools (v2.30)[69], and 1 kb flanking sequences, including the erroneous regions, were extracted. These sequences were aligned to the corresponding haplotype assemblies from corrected ultra-long assemblies using Minimap2 (v2.28)[70] to identify reciprocal best matches. Only error-free regions from the ultra-long assemblies were used for patching (Additional methods: **Genome assembly**).

Evaluation of Assembly

Assembly quality (QV) was assessed using hybrid k-mers derived from HiFi and MGI reads, with completeness and phasing statistics calculated using Merqury (v1.3)[26]. Assembly errors and overall quality were also evaluated using CRAQ (v1.0.9-alpha)[71], which utilizes read clipping information, and Inspector (v1.3)[72], which detects both large (SV) and small-scale (SNV) errors based on reads-to-assembly alignments. Additionally, coverage-based tools such as Flagger (v0.3.3)[6] and GCI (v1.0)[73] were employed to identify errors based on coverage inconsistencies. Haplotype assemblies were further validated using Omni-C contact maps, which were visualized in Juicebox (v1.11.08)[74].

Repeat Annotation and Centromere Analysis

Repeat annotation for both haplotypes was performed using RepeatMasker (v4.1.5)[75] with a curated repeat library from Dfam (v3.7)[76]. Centromere alpha satellite (αSat) monomers were identified using Hum-AS-HMMER[77], and high-order repeats (HORs) were classified as active/live, inactive, divergent and monomeric based on previous studies[29,30]. HSat2 and HSat3 were annotated following established methods[30] while other centromeric satellites were identified by parsing RepeatMasker output. Structural variants within active centromeric HORs were predicted using Stv[78].

Genome Comparison Between I002C and CHM13

To minimize the identification of mapping-biased false structural variants (SVs), all centromeric regions, heterochromatin, and repeats ≥10 kb were masked with 'N'. Each I002C chromosome from the maternal and paternal haplotypes was aligned to the corresponding CHM13 chromosome using minimap2 with the '-x asm5 -r2k' options for the contiguous alignments. SVs were identified using Syri (v1.6.3)[79] and svim-asm (v1.0.3)[80]. A consensus SV set for each haplotype was generated by merging SVs detected by both tools using Jasmine with 'max_dist=1000 min_support=2". The final combined SV set for the I002C genome was generated by merging both haplotypes using jasmine 'max_dist=1000 min_support=1'.

To identify I002C-specific SVs, they were compared against publicly available SV datasets from gnomAD (v4.1)[34], HGSVC (v1–3)[27,35], and HPRC[6]. Since SV coordinates for I002C and HPRC were based on CHM13, whereas gnomAD and HGSVC SVs were based on hg38, the latter were lifted over to CHM13 using Picard LiftOverVCF (v2.26.9)[81]. Finally, SVs from public datasets were merged using Jasmine with the parameters max_dist=1000 min_support=1.

Assessing Reference Genome Suitability

Read mapping performance We compared mapping performance for both short-read NGS data and long-read ONT data across three complete haploid reference genomes: I002C maternal haploid (includes chrY and mitochondrial genome), I002C paternal haploid (includes chrX and mitochondrial genome), and CHM13 (v2.0).

For short-read analysis, we randomly selected a subset of South Asian (SAS) individuals from the 1000 Genomes Project (1KGP), comprising 50 individuals from each SAS subpopulation: BEB, GIH, ITU, PJL, and STU. The corresponding CRAM files were converted to FASTQ format using samtools (v1.18). Paired-end reads from each sample were aligned separately to the three reference genomes using BWA-MEM (v0.7.17). Duplicate reads were marked using GATK MarkDuplicates (v4.1.8.1)[82].

For long-read analysis, we used ONT data from the Singapore HELIOS cohort. The reads were aligned to each reference using minimap2 (v2.28). Mapping statistics were collected from each BAM file using samtools, and the percentage of uniquely mapped reads was calculated using the formula $(N_{\text{Mapped reads}} - N_{MQ0}) / N_{\text{Mapped reads}}$.

Assembly-based variant calling Haplotypes from I002C and the CHM13 reference were aligned to GRCh38 using minimap2 (v2.28) with the assembly-to-assembly setting (-x asm10). Variant calling was performed with paftools, and the resulting variants were annotated using the ClinVar database[83] to assess clinical significance. Variants classified as pathogenic, of uncertain significance (VUS), risk-associated, or with conflicting interpretations of pathogenicity were further cross-referenced against gnomAD (v4.1)[84].

Structural variant calling To further assess reference genome performance, structural variants (SVs) were called from the NGS alignments using Manta (v1.6.0)[85], and from ONT alignments using Sniffles (v2.6.2). To evaluate the reference specific SVs, we first projected SVs on HG38 coordinates using CrossMap (v0.7.0)[86]. The chain files from I002C to HG38 were generated following the nf-LO pipeline[87].

A combined SV set for each haplotype was generated by merging SV across all samples using Jasmine (v1.1.5) with 'max_dist=1000'. The same merging strategy was used to merge three reference-based SV files to identify the reference-specific SVs.

Gene Annotation

To annotate genes on the I002C maternal and paternal genome assemblies, as part of the gene annotation pipeline (Additional file 1: Figure S27; Additional methods: **Gene annotation**), we first applied a liftOver (conda version 377)[49] approach using CHM13 as a reference. Chain files for both haplotypes were generated from CHM13 to the I002C assemblies using the nf-LO pipeline. This pipeline uses minimap2 for alignment and paf2chain to convert alignments into UCSC-style chain files. The resulting chain files were applied with liftOver to project CHM13 gene annotations (v2.0) onto the I002C assemblies.

To independently validate and complement the annotation, we also performed protein-to-genome mapping. Human protein sequences were obtained from UniProt (release 2024_06)[50] and mapped to both I002C haplotypes and CHM13 using miniprot (v.0.13-

r262-dirty)[88]. To complement the liftover-based annotation and ensure detection of potential haplotype-specific genes, we additionally performed protein-supported gene prediction using TSEBRA (v.1.1.2.5)[51]. RNA-seq reads were aligned to each haplotype-specific assembly using HISAT2 (v.2.2.1)[89], and transcript models were constructed using StringTie (v2.2.3)[90]. TransDecoder (v5.7.1)[91] was then applied to extract coding regions, yielding genome-aligned protein-coding gene models. Protein hints were generated by aligning previously mentioned UniProt sequences to the assemblies using miniprot. TSEBRA was used to select transcript models supported by these protein hints, using TransDecoder predictions as input. It was run separately for each haplotype using a custom priority configuration. The final annotations were obtained by merging the filtered TSEBRA predictions with the liftOver annotations, retaining all liftover genes and incorporating only non-overlapping genes from TSEBRA.

Ancestry Analysis

Local ancestry inference for the I002C genome was performed by calling variants relative to the GRCh38 reference using dipcall (v0.3)[92]. The resulting variant set was merged with biallelic SNPs from 1000 Genomes Project (1KGP) samples representing African (AFR), American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS) super-populations. Ancestry was inferred at the super-population level using RFMix (2.03-r0)[93].

Methylation

Genome-wide and centromeric DNA methylation profiling for I002C genomes was performed using PacBio HiFi and ONT simplex reads with platform specific pipelines. To ensure reliable quantification, methylation ratios were calculated only for CpG sites supported by at least four methylation-informative reads (containing MM/ML tags). To evaluate platform concordance, Pearson correlation coefficients were computed for methylation levels across common CpG sites and predefined genomic windows. CDR for each chromosome was manually examined and annotated as described in the previous study[30]. Haplotype-specific differentially methylated loci (DMLs) were identified by

aligning parental reads (excluding homozygous reads) to the GRCh38 reference, requiring a minimum methylation difference of 97.5% and at least five informative reads per site. Differentially methylated regions (DMRs) were defined as regions containing at least five DMLs and majority of DMLs within this region to have consistent methylated allele. To identify candidate genes controlled by imprinting, DMRs were intersected with core promoter regions of experimentally validated transcripts (Additional methods: **Methylation**).

ChrY analysis

The subregions of I002C-Y were annotated using subregion definitions from HG002-Y. These regions were mapped onto I002C-Y using minimap2 (v2.28) with the parameters "-x asm20 -Y -N1 --cs -c --eqx --paf-no-hit." To assess sequence conservation, we computed average sequence identity for each subregion based on pairwise alignments between I002C-Y and HG002-Y using nucmer (v3.1)[94] with "--maxmatch -l 100" parameters. For each subregion, identity was calculated as a length-weighted average across all alignments.

To further characterize the paternal lineage, we determined the Y chromosome haplogroup using short read data aligned to the I002C-Y assembly. Haplogroup assignment was performed using Yleaf (v3.2.1)[95] by comparing SNPs in the non-recombining region of the Y chromosome to those in the YFull phylogenetic reference tree (v10.01)[96].

Data availability

The sequencing reads have been submitted to NCBI under Project ID PRJNA1150503. The I002C assemblies are available at https://github.com/LHG-GG/I002C. Additionally, the publicly available data resources used in this study are listed in Additional file 2: Table S28, along with their corresponding access URLs.

References

1. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res* **27**, 849–864 (2017).
2. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
3. HG002. MarBL (2025).
4. Yang, C. *et al.* The complete and fully-phased diploid genome of a male Han Chinese. *Cell Res* **33**, 745–761 (2023).
5. Wang, T. *et al.* The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604**, 437–446 (2022).
6. Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
7. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164 (2016).
8. Fatumo, S. *et al.* Diversity in Genomic Studies: A Roadmap to Address the Imbalance. *Nat Med* **28**, 243–250 (2022).
9. Yew, C.-W. *et al.* Genomic structure of the native inhabitants of Peninsular Malaysia and North Borneo suggests complex human population history in Southeast Asia. *Human genetics* **137**, 161–173 (2018).
10. Wall, J. D. *et al.* The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature* **576**, 106–111 (2019).
11. Wu, D. *et al.* Large-Scale Whole-Genome Sequencing of Three Diverse Asian Populations in Singapore. *Cell* **179**, 736-749.e15 (2019).
12. Le, V. S. *et al.* A Vietnamese human genetic variation database. *Human mutation* **40**, 1664–1675 (2019).
13. Deng, L. *et al.* Analysis of five deep-sequenced trio-genomes of the Peninsular Malaysia Orang Asli and North Borneo populations. *BMC Genomics* **20**, 842 (2019).

14. Piriyapongsa, J. *et al.* Pharmacogenomic landscape of the Thai population from genome sequencing of 949 individuals. *Sci Rep* **14**, 30683 (2024).
15. Azmi, M., Chen, L., Idris, A. & Lu, Z. H. Genomic Insights of Bruneian Malays. 2022.06.01.492266 Preprint at https://doi.org/10.1101/2022.06.01.492266 (2022).
16. Bhattacharyya, C. *et al.* Mapping genetic diversity with the GenomeIndia project. *Nat Genet* **57**, 767–773 (2025).
17. Ardiansyah, E. *et al.* Sequencing whole genomes of the West Javanese population in Indonesia reveals novel variants and improves imputation accuracy. *Front. Genet.* **15**, (2025).
18. Rhie, A. *et al.* The complete sequence of a human Y chromosome. *Nature* **621**, 344–354 (2023).
19. Rautiainen, M. *et al.* Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nature Biotechnology* **41**, 1474–1482 (2023).
20. 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).
21. Cheng, H., Asri, M., Lucas, J., Koren, S. & Li, H. Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. *Nature Methods* **21**, 967–970 (2024).
22. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
23. Samuels, D. C. *et al.* Heterozygosity Ratio, a Robust Global Genomic Measure of Autozygosity and Its Association with Height and Disease Risk. *Genetics* **204**, 893–904 (2016).
24. Xu, M. *et al.* TGS-GapCloser: A fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *GigaScience* **9**, giaa094 (2020).
25. Mc Cartney, A. M. *et al.* Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. *Nature Methods* **19**, 687–695 (2022).
26. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality,

completeness, and phasing assessment for genome assemblies. *Genome Biology* **21**, 245 (2020).

27. Logsdon, G. A. *et al.* Complex genetic variation in nearly complete human genomes. 2024.09.24.614721 Preprint at https://doi.org/10.1101/2024.09.24.614721 (2024).
28. Gao, S. *et al.* Centromere Landscapes Resolved from Hundreds of Human Genomes. *Genomics Proteomics Bioinformatics* **22**, qzae071 (2024).
29. Miga, K. H. & Alexandrov, I. A. Variation and Evolution of Human Centromeres: A Field Guide and Perspective. *Annual Review of Genetics* vol. 55 583–602 (2021).
30. Altemose, N. *et al.* Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178.
31. Logsdon, G. A. *et al.* The variation and evolution of complete human centromeres. *Nature* **629**, 136–145 (2024).
32. Šošić, M. & Šikić, M. Edlib: a C/C ++ library for fast, exact sequence alignment using edit distance. *Bioinformatics* **33**, 1394–1395 (2017).
33. Hallast, P. *et al.* Assembly of 43 human Y chromosomes reveals extensive complexity and variation. *Nature* **621**, 355–364 (2023).
34. Collins, R. L. *et al.* A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).
35. Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).
36. Wang, X. *et al.* The Health for Life in Singapore (HELIOS) Study: delivering Precision Medicine research for Asian populations. *medRxiv* 2024.05. 14.24307259 (2024).
37. Wagner, J. *et al.* Curated variation benchmarks for challenging medically-relevant autosomal genes. *Nat Biotechnol* **40**, 672–680 (2022).
38. Miller, D. T. *et al.* ACMG SF v3.2 list for reporting of secondary findings in clinical exome and genome sequencing: A policy statement of the American College of Medical Genetics

and Genomics (ACMG). *Genet Med* **25**, 100866 (2023).

39. Butler, M. G. Prader-Willi Syndrome: Obesity due to Genomic Imprinting. *Curr Genomics* **12**, 204–215 (2011).
40. Buiting, K. *et al.* Epimutations in Prader-Willi and Angelman Syndromes: A Molecular Study of 136 Patients with an Imprinting Defect. *Am J Hum Genet* **72**, 571–577 (2003).
41. Viljoen, D. & Ramesar, R. Evidence for paternal imprinting in familial Beckwith-Wiedemann syndrome. *Journal of medical genetics* **29**, 221–225 (1992).
42. Ishida, M. New developments in Silver–Russell syndrome and implications for clinical practice. *Epigenomics* **8**, 563–580 (2016).
43. Court, F. *et al.* Genome-wide parent-of-origin DNA methylation analysis reveals the intricacies of human imprinting and suggests a germline methylation-independent mechanism of establishment. *Genome Res* **24**, 554–569 (2014).
44. Joshi, R. S. *et al.* DNA Methylation Profiling of Uniparental Disomy Subjects Provides a Map of Parental Epigenetic Bias in the Human Genome. *The American Journal of Human Genetics* **99**, 555–566 (2016).
45. Hernandez Mora, J. R. *et al.* Characterization of parent-of-origin methylation using the Illumina Infinium MethylationEPIC array platform. *Epigenomics* **10**, 941–954 (2018).
46. Zink, F. *et al.* Insights into imprinting from parent-of-origin phased methylomes and transcriptomes. *Nature genetics* **50**, 1542–1552 (2018).
47. Akbari, V. *et al.* Genome-wide detection of imprinted differentially methylated regions using nanopore sequencing. *eLife* **11**, e77898.
48. Akbari, V. *et al.* Parent-of-origin detection and chromosome-scale haplotyping using long-read DNA methylation sequencing and Strand-seq. *Cell Genomics* **3**, 100233 (2023).
49. Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. *Briefings in Bioinformatics* **14**, 144–161 (2013).
50. UniProt: a hub for protein information. *Nucleic Acids Res* **43**, D204–D212 (2015).

51. Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M. & Stanke, M. TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics* **22**, 566 (2021).

52. Sarashetti, P., Lipovac, J., Tomas, F., Šikić, M. & Liu, J. Evaluating data requirements for high-quality haplotype-resolved genomes for creating robust pangenome references. *Genome Biology* **25**, 312 (2024).

53. Lawce, H. J. & Brown, M. G. Cytogenetics: an overview. *The AGT cytogenetics laboratory manual* 25–85 (2017).

54. Dudchenko, O. *et al.* De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).

55. Stanojević, D., Lin, D., Nurk, S., Florez de Sessions, P. & Šikić, M. Telomere-to-Telomere Phased Genome Assembly Using HERRO-Corrected Simplex Nanopore Reads. *bioRxiv* 2024.05.18.594796 (2024) doi:10.1101/2024.05.18.594796.

56. Bankevich, A., Bzikadze, A. V., Kolmogorov, M., Antipov, D. & Pevzner, P. A. Multiplex de Bruijn graphs enable genome assembly from long, high-fidelity reads. *Nat Biotechnol* **40**, 1075–1081 (2022).

57. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* **14**, 178–192 (2013).

58. Sarashetti, P. T2T Sequences.

59. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome research* **27**, 737–746 (2017).

60. Jain, C., Rhie, A., Hansen, N. F., Koren, S. & Phillippy, A. M. Long-read mapping to repetitive reference sequences using Winnowmap2. *Nature Methods* **19**, 705–710 (2022).

61. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://doi.org/10.48550/arXiv.1303.3997 (2013).

62. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–

2079 (2009).

63. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology* **36**, 983–987 (2018).
64. Shafin, K. *et al.* Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nature Methods* **18**, 1322–1332 (2021).
65. Smolka, M. *et al.* Detection of mosaic and population-level structural variants with Sniffles2. *Nature Biotechnology* **42**, 1571–1580 (2024).
66. Kirsche, M. *et al.* Jasmine and Iris: population-scale structural variant comparison and analysis. *Nature Methods* **20**, 408–417 (2023).
67. Formenti, G. *et al.* Merfin: improved variant filtering, assembly evaluation and polishing via k-mer validation. *Nature Methods* **19**, 696–704 (2022).
68. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
69. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
70. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
71. Li, K., Xu, P., Wang, J., Yi, X. & Jiao, Y. Identification of errors in draft genome assemblies at single-nucleotide resolution for quality assessment and improvement. *Nature Communications* **14**, 6556 (2023).
72. Chen, Y., Zhang, Y., Wang, A. Y., Gao, M. & Chong, Z. Accurate long-read de novo assembly evaluation with Inspector. *Genome Biology* **22**, 312 (2021).
73. Chen, Q., Yang, C., Zhang, G. & Wu, D. GCI: a continuity inspector for complete genome assembly. *Bioinformatics* **40**, btae633 (2024).
74. Durand, N. C. *et al.* Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Systems* **3**, 99–101 (2016).

75. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0. 2013–2015. (2015).
76. Storer, J., Hubley, R., Rosen, J., Wheeler, T. J. & Smit, A. F. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA* **12**, 2 (2021).
77. Ryabov, F. HumAS-HMMER_for_AnVIL. (2024).
78. Ryabov, F. stv. (2024).
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. Heller, D. & Vingron, M. SVIM-asm: structural variant detection from haploid and diploid genome assemblies. *Bioinformatics* **36**, 5519–5521 (2020).
81. Broad Institute. Picard Tools.
82. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297–1303 (2010).
83. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* **42**, D980-985 (2014).
84. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
85. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
86. Zhao, H. *et al.* CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006–1007 (2014).
87. Chen, N.-C. & Vollger, M. nfLO.
88. Li, H. Protein-to-genome alignment with miniprot. *Bioinformatics* **39**, btad014 (2023).
89. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907–915

(2019).

90. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290–295 (2015).
91. Haas, B. TransDecoder.
92. Li, H. *et al.* A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat Methods* **15**, 595–597 (2018).
93. 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).
94. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol* **5**, R12 (2004).
95. Ralf, A., Montiel González, D., Zhong, K. & Kayser, M. Yleaf: Software for Human Y-Chromosomal Haplogroup Inference from Next-Generation Sequencing Data. *Molecular Biology and Evolution* **35**, 1291–1294 (2018).
96. Yfull. Haplogroup YTree v10.01.00. https://www.yfull.com/arch-10.01/tree/ (2022).
97. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research* **27**, 722–736 (2017).
98. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nature methods* **9**, 676–682 (2012).
99. Akalin, A. *et al.* methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol* **13**, R87 (2012).

Acknowledgement

We want to thank the anonymous volunteers who contributed their blood samples to this study. We would also like to thank the A*STAR Microscopy Platform (AMP) and the AMP staff, especially Dr Goh Wah Ing, for the consultation, advice, and technical support. AMP

is funded by Singapore's Agency for Science, Technology & Research (A*STAR) through core funding. We also want to thank Oxford Nanopore and PacBio for their long read sequencing support. This work was supported by the funding from the A*STAR and PRECISE and the high-performance computing resources provided by the A*STAR Computational Resource Centre (A*CRC).

Funding

This study was supported by the Agency for Science, Technology and Research (A*STAR) and the Singapore National Precision Medicine (NPM) Program (MOH-000588–01). Additional funding was provided by the European Union through the European Regional Development Fund under grant KK.01.1.1.01.0009 (DATACROSS), as well as by the Croatian Science Foundation under grant IP-2018–01-5886 (SIGMA).

Author information

Authors and Affiliations

Genome Institute of Singapore (GIS), Agency for Science, Technology and Research (A*STAR), 60 Biopolis Street, Genome, Singapore 138672, Singapore
Prasad Sarashetti , Qi Jia, Ling Wang, Zhe Li, Lovro Vrcek, Fei Yao, YY Sia, Dehui Lin, Xinyi Zhang, Daniel Muliaditan, Hwee Meng Low, See Ting Leong, Chiea Chuen Khor, Eleanor Wong, Patrick Tan, Mile Sikic, and Jianjun Liu

Laboratory for Bioinformatics and Computational Biology, Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia
Josipa Lipovac, Kresimir Krizanovic, and Mile Sikic

Precision Health Research, Singapore, 139234, Singapore
Weiling Zheng, John C Chambers, and Patrick Tan

Population and Global Health, Lee Kong Chian School of Medicine, Nanyang Technological University, 308232, Singapore
John C Chambers

Department of Hematology, Singapore General Hospital, Singapore 169856, Singapore
William Y K Hwang

Cancer & Stem Cell Biology, Duke-NUS Medical School, Singapore, Singapore
William Y K Hwang

Department of Hematology, National Cancer Centre Singapore, Singapore 168583, Singapore
William Y K Hwang

Regenerative Medicine Institute of Singapore, SingHealth Duke-NUS Academic Medical Centre, Singapore 169857, Singapore
William Y K Hwang

SingHealth Duke-NUS Institute of Precision Medicine, 169609, Singapore
Patrick Tan

Cancer & Stem Cell Biology Program, Duke-NUS Medical School, 169857, Singapore
Patrick Tan

Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117597, Singapore
Jianjun Liu

Contributions

JJ.L. and M.Š. conceived the project. W.YK.H., contributed to the sample recruitment and collection. YY.S. contributed to establishing and growing cell lines. F.Y., H.W.L., S.T.L., C.C.K., contributed to library preparation and sequencing. P.S. designed the pipeline. P.S., J.L. and D.H. performed genome assemblies. P.S. performed genome polishing, evaluation, centromere and repeat analysis. J.L., P.S. performed read binning. P.S., J.L.,

Q.J., Z.L., L.W., L.V., D.H., X.Z., K.K., M.Š., and JJ.L. performed manual curation. P.S., Q.J. performed in-silico rDNA CN estimation. Z.L. performed dPCR experiment. Z.L., D.M., performed FISH experiments and analysis. Q.J. performed rDNA assembly. Z.L. performed methylation analysis. X.Z. performed ancestry analysis. L.V. performed chromosomeY analysis. J.L. performed gene annotation. P.S., W.L., performed reference justification analysis. E.W., W.Z., contributed to project management. J.C.C., contributed to the recruitment of HELIOS samples. P.T., and JJ.L., secured the funding. P.S., J.L., Q.J., L.W., Z.L., L.V., M.Š., and JJ.L. drafted the manuscript with contribution from all authors. M.Š. and JJ.L. supervised the project and provided mentorship.

Ethics declarations

Ethics approval and consent to participate

All the participants were provided with informed consent for sample collection, and usage including making data publicly available via databases. Sample collection and usage were approved by SingHealth Centralised Institutional Review Board (IRB Reference: 2024–069) and A*STAR Institutional Review Board (IRB Reference: 2019-118). All experimental methods comply with the Helsinki Declaration.

Consent for publication

Not applicable.

Competing interests

M.Š. has been jointly funded by Oxford Nanopore Technologies and AI Singapore for the project AI-driven De Novo Diploid Assembler. The remaining authors declare no competing interests.

Supplementary Information

Additional File1 - Supplementary figures S1-S27

Additional File 2- Supplementary tables S1-S28

Additional methods - Detailed methodology