A Complete Telomere-to-Telomere Diploid Reference Genome for South Asian Population

Prasad Sarashetti[1#], Josipa Lipovac[2#], Qi Jia[1,11], Ling Wang[1], Zhe Li[1], Lovro Vrcek[1], Ying Chen[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,], Jonathan Göke[1,11], Mile Sikic[1,2*], Jianjun Liu[1,12*]

[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]Department of Statistics and Applied Probability, National University of Singapore, Singapore, 117546, Singapore

[12]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.

Keywords

Diploid genome assembly, Reference genome, Telomere-to-telomere, Indian Ancestry, South Asian population, Haplotype-resolved methylome, Imprinting, Reference bias

Introduction

A human reference genome serves as the foundation for genomic analyses, such as gene and regulatory sequence annotations, transcriptome quantification, and variation analysis, making it an essential resource for biomedical and clinical research to understand human genetic diversity, 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 unresolved regions, including ribosomal DNA (rDNA) arrays, as well as pericentromeric and subtelomeric regions[2], which are critical for understanding fundamental cellular processes. Trying to overcome these limitations, there have been several efforts on generating new human reference genomes with better completeness and quality, such as CHM13[2], HG002[3], CN1[4] and YAO[5] 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)[6,7].

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[8,9]. 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[10–18], and currently no complete T2T reference genome is 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[4,7,19]. These limitations of reference bias hamper genetic studies and translation to clinical applications in underrepresented populations and call for the generation of ancestry-specific T2T reference genomes.

87

88 Here we present the first high-quality, gapless, telomere-to-telomere, haplotype-resolved
89 reference genome (I002C) with 44 autosomes and XY chromosomes for South Asia,
90 using a biologically normal Indian sample, from a father-mother-child trio samples set
91 recruited in Singapore. We employed state-of-the-art sequencing technologies, including
92 PacBio HiFi circular consensus long-read sequencing and Oxford Nanopore ultralong
93 sequencing, and advanced algorithms for de novo assembling, including Verkko[20] and
94 Hifiasm[21,22] to construct chromosome-level contigs, including the long and highly
95 repetitive centromeric regions and the short arms of five acrocentric chromosomes. We
96 utilized both parental sequencing data information for phasing and generating the
97 paternal and maternal haploid genomes, and we used long read RNA-Seq data combined
98 with reference genome annotations to provide a comprehensive set of transcripts for the
99 I002C genome. We further demonstrated that with this reference genome, South Asian
100 (SAS)-specific novel variants can be identified with greater accuracy, and sequencing
101 results from SAS samples can be analyzed with higher mapping rates and lower clipping
102 rates. This effort has helped to overcome a significant gap in the presentation of human
103 genome diversity by providing the first reference genome for the SAS population, the
104 largest one in the world by accounting for 25% of the world population.

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, complemented with 70.53 GB of Nanopore cDNA data. 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—SAS (Figure 1). Local ancestry analysis based on data from the 1000 Genomes Project[23] 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[24] (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). We also performed the PCA analysis of I002C and the 5 SAS subpopulation samples from the 1000 Genomes Project. The analysis showed that I002C clustered with the SAS group overall, with a closer affinity to Gujarati Indians from Houston (GIH), Punjabis from Lahore (PJL), and Indian Telugu from the UK (ITU) (Additional file 1: Figure S3). This clustering pattern further supports the South Asian origin of I002C.

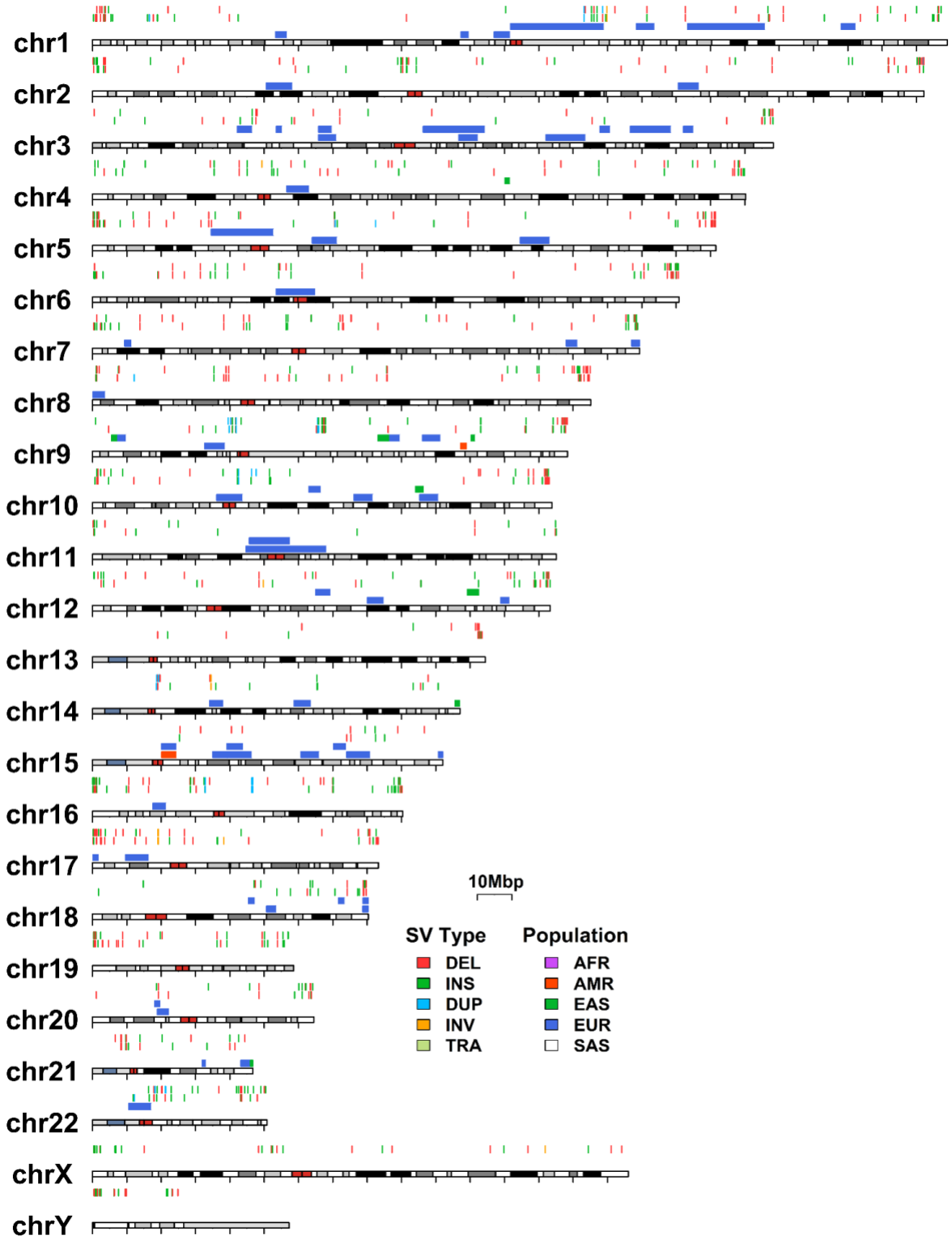


Figure 1: Ancestral composition and unique structural variants of I002C. Overview of ancestry markers and structural variants across I002C haplotypes. Chromosomal

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

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

144 Initial assemblies were generated using Verkko[20] and Hifiasm[21,22] in trio mode. For each
145 haplotype, we selected chromosome-scale sequences from either assembler based on
146 quality, contiguity and presence of telomere motif on both ends. This resulted in 20
147 maternal chromosomes (excluding the p-arms of chromosomes 13, 15 and 22) with 39
148 gaps, and 19 paternal chromosomes (excluding the p-arms of chromosomes 4, 13, 14
149 and 15) with 14 gaps (Additional methods: **Contig construction)**. The remaining
150 chromosomes were generated using template-based scaffolding (Additional Methods:
151 **Scaffolding**). The gaps were closed using combinatorial methods, including TGS-
152 Gapcloser[25], contigs from local reassemblies of binned reads and ultra-long read
153 alignments (Additional methods: **Gap-filling**). The remaining telomeres were patched
154 using subtelomere ultra-long read mapping and local assemblies (Additional methods:
155 **Telomere patching and extension**) (Additional file 2: Table S2). The rDNA arrays were
156 verified and corrected by local re-assembly and manual curation (Additional methods:
157 **rDNA**). The final assembly includes full T2T sequences for all 22 autosome pairs, both
158 sex chromosomes (X and Y), and the mitochondrial genome.
159
160 To further improve assembly quality, we performed multiple rounds of polishing using
161 binned HiFi, ONT reads (duplex, corrected and uncorrected ultra-long), and MGI short
162 reads (Additional methods: **Read binning**, Additional file 2: Table S3-S4). The polishing
163 workflow included two rounds of automated polishing[26], one round of manual curation
164 (Additional file 1: Figure S4), 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 (average of two haplotypes)(Figure 2a), as assessed by Merqury[27] using hybrid k-mers (k=21) 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 S5). A range of evaluation tools confirmed the high quality of the final I002C assembly (Additional file 1: Figure S6-S7). 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[7] (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[26] (Additional file 1: Figure S8). 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). Merqury based quality assessment (k=21) resulted in genome wide QV scores of 82.05 for the maternal and 83.08 for the paternal I002C assemblies, exceeding other high-quality human genomes that are either published or publicly released (Figure 2c). To increase the sensitivity in homopolymer-rich regions, we also performed QV estimation using k-mer size of 31-mers. As expected, QV scores based on 31-mers were generally lower than those from 21-mers across all the genomes. Despite the overall decrease, I002C still exhibited higher QV values than the other genomes (Additional file 1: Figure S9). 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 (k=21) and 70.16 (k=31). These results establish I002C as the highest quality human genome currently available.

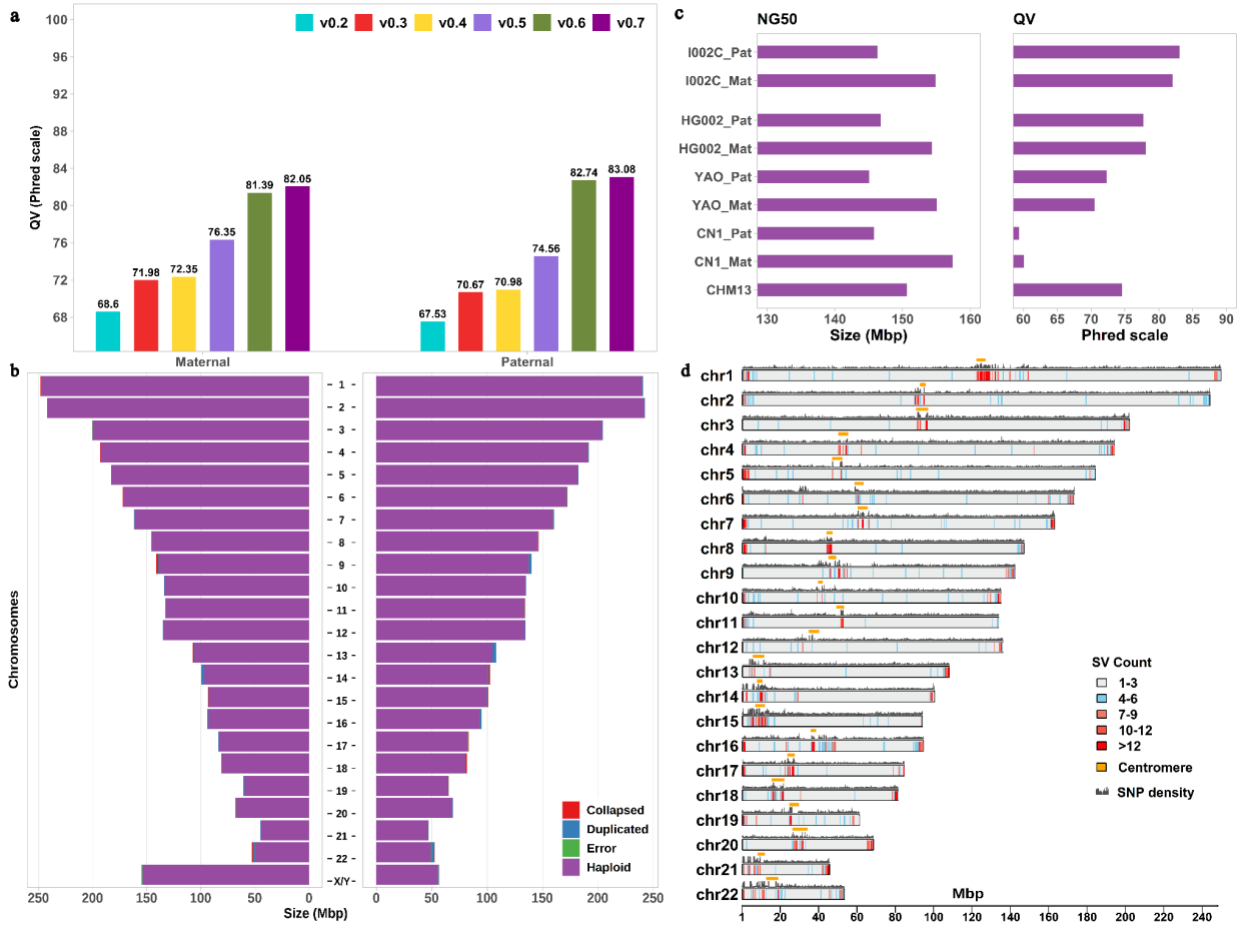


Figure 2: Overview of the haplotype-resolved I002C genome. a) Progression of I002C genome versions with polishing, showing improvements in QV scores assessed using hybrid k-mers. b) Assessment of haplotype reliability based on read mapping. For each chromosome, the left bar represents the maternal haplotype and the right bar the paternal haplotype. Regions marked as haploid (purple) are considered reliable, comprising over 99% of each haplotype on average. c) NG50 and QV (k=21) values for I002C haplotypes compared with other high-quality human genomes (HG002, YAO, CN1 and CHM13), ordered by year of publication or public release with most recent shown at the top. 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 and subtelomere 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[28,29], 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 file 1: Figure S10-S11). 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

235 spans 1.8 Mbp compared to 143.8 Kbp in CHM13. In contrast, the maternal chromosome
236 14 monomeric αSat measured 6.1 Kbp, substantially shorter than the 91.6 Kbp in CHM13.
237 Despite these length differences, most peri/centromeric satellites are conserved between
238 the I002C and CHM13 genomes, although some are uniquely present or absent. For
239 example, maternal chromosome 14 and paternal chromosome 21 carry around ~2 Kbp
240 of HSat2 not found in CHM13, and both I002C haplotypes also include about ~2 Kbp of
241 HSat2 on chromosome 15 and 55 bp of γSat on chromosome 16 not present in CHM13
242 (Figure 3b, Additional file 1: Figure S12-S15).
243
244 To further explore the αSat HOR array structure between genomes, we compared
245 centromeres of chromosome pairs between I002C and CHM13. This analysis showed
246 differences in HOR organization between CHM13 and I002C genomes, including HOR
247 SVs that are deviated from canonical patterns [30,31] (Additional file 2: Table S12). Notably,
248 some HOR monomers found in I002C were completely absent from CHM13. For instance,
249 CHM13 chromosome 1 was primarily composed of 6-mer HORs such as
250 S1C1/5/19H1L.1-5(_6/4_5){n}-6, S1C1/5/19H1L.1-6 while I002C haplotypes contained
251 distinct 2-mer variants like S1C1/5/19H1L.5_6/4 and S1C1/5/19H1L.5-6, which were
252 missing from CHM13 (Figure 3c). The active HOR array of chromosome 17 in CHM13
253 mainly consists of the canonical 16-mer, whereas I002C haplotypes contain a mix of
254 variant 16-mers along with 15-, 14-, and 13-mers. Notably, a novel 13-mer HOR variant
255 (S3C17H1L.1-10_15_15-16) was found in both I002C haplotypes but not in CHM13
256 (Figure 3d). In contrast, chromosome 21 in both I002C haplotypes was primarily made up
257 of 5-mer variants, compared to the canonical 11-mer in CHM13. A 5-mer variant
258 (S2C13/21H1L.1-5) observed in I002C was completely absent in CHM13, and additional
259 HORs unique to the paternal haplotype were not found in either the maternal genome or
260 CHM13 (Figure 3e). Despite these differences, monomers within these HOR variants
261 remain sequence-identical, suggesting structural polymorphism without extensive
262 sequence divergence.
263
264 Similar differences were observed among the I002C maternal and paternal haplotypes
265 (Additional file 1: Figure S16). Overall, 52% of the unique HOR SVs to I002C were specific

266 to either the maternal or paternal haplotype. The dominant αSat HORs also differed
267 between haplotypes—for instance, chromosome 5 had an 8-mer and a 4-mer;
268 chromosome 10 showed a 6-mer and an 8-mer; and chromosome 17 carried a 16-mer
269 and 13-mers in maternal and paternal haplotypes, respectively (Figure 3b). Together,
270 these findings highlight the polymorphic nature of centromeric components among
271 different human genomes and across haplotypes within the same genome.
272
273 Taking advantage of single-molecule long-read sequencing, which preserves native DNA
274 methylation signals, we investigated methylation patterns across centromeric regions of
275 both haplotypes. To identify putative kinetochore regions within the centromeres, we
276 analyzed methylation patterns for signatures associated with centromere function. A
277 characteristic centromeric methylation dip—marked by a sharp decrease in methylation
278 levels within the active HOR subregions relative to their flanking sequences-was
279 observed consistently across all chromosomes (Figure 3f, Additional file 1: Figure S17).
280 The size of the methylation dip ranged from 0.1 to 0.4 Mb, consistent with previous reports
281 in CHM1 and CHM13[32] (Additional file 1: Figure S18).
282
283 Ribosomal DNA (rDNA) arrays located on the short arms of acrocentric chromosomes
284 are among the most challenging regions to assemble due to their high repeat content,
285 extensive segmental duplications, and complex satellite structures. The total rDNA copy
286 number (CN) in I002C was estimated to be 416 for the diploid genome using digital PCR,
287 which was consistent with estimations from the k-mer-based approach[2]. Based on the
288 coverage of binned rDNA reads, we estimated 157 and 259 CNs for maternal and paternal
289 haplotypes, respectively. Chromosome-level rDNA CNs were inferred by integrating read
290 clustering with coverage-based and fluorescence in situ hybridization (FISH)-based
291 estimates, and further cross-validated using independent FISH results (Additional
292 methods: rDNA). Comparison of the estimated CN with the CN observed in the
293 assembled genome revealed that the CNs of rDNAs were underrepresented for most of
294 the chromosomes, except for the maternal chromosomes 13 and 21 (Additional file 2:
295 Table S13). The maternal chromosome 21 rDNA sequence was found to be completely
296 assembled in the initial assembly, supported by the coverage based estimates, FISH

297 estimates, and full coverage of overlapping ultra-long reads. As for maternal chromosome
298 13, the reassembled rDNA sequence was nearly identical to the assembled genome in
299 terms of both sequences and CNs, despite the absence of the full coverage of the region
300 by overlapping ultra-long reads. The rDNA sequences from the assembled genome were
301 retained for the maternal chromosomes 13 and 21. As for the remaining acrocentric
302 chromosomes, due to the difficulty of resolving base-level variation among highly similar
303 rDNA units, rDNA morphs, defined as representative sequences of one or more nearly
304 identical units, were identified through further read clustering (Additional methods: **rDNA**).
305 Arrays were reconstructed using consensus morph sequences scaled by their estimated
306 CNs. These reconstructed arrays were then incorporated into the final assembly by
307 replacing the originally assembled rDNA segments (Figure 3g, Additional file 2: Table
308 S14).
309
310 In general, the typical rDNA array structure consisted of consecutive ~45 kb rDNA units,
311 flanked by a ~4 kb rDNA fragment upstream and a ~6 kb rDNA fragment downstream,
312 and further bordered by the distal and proximal junction sequences (DJ, PJ). In total, we
313 identified 19 major rDNA morphs that are different with >200 edit distance, including 7
314 maternal and 12 paternal morphs in I002C. Comparison with the 8 major morphs in
315 CHM13 revealed sequence diversity, with all major morphs in I002C showing >200 edit
316 distance[33] from their CHM13 counterparts (Additional file 1: Figure S19). Despite this
317 divergence, several large SVs were shared between I002C and CHM13 major morphs
318 when compared to the canonical rDNA reference (Additional file 1: Figure S20). For
319 example, a 1,530 bp deletion in the tandem repeat block 1 (TR1) is shared among I002C
320 Mat_chr22_A, chm13-model15c and chm13-model22a morphs. In contrast, a 988 bp
321 deletion in the long repeat (LR) region present in chm13-model15a and chm13-model15b
322 was not observed in I002C major morphs. Additionally, we identified a ~900 kbp deletion
323 in the TR2 region of I002C Pat_chr22_B1 and Pat_chr22_B2 that is not present in CHM13
324 major morphs.

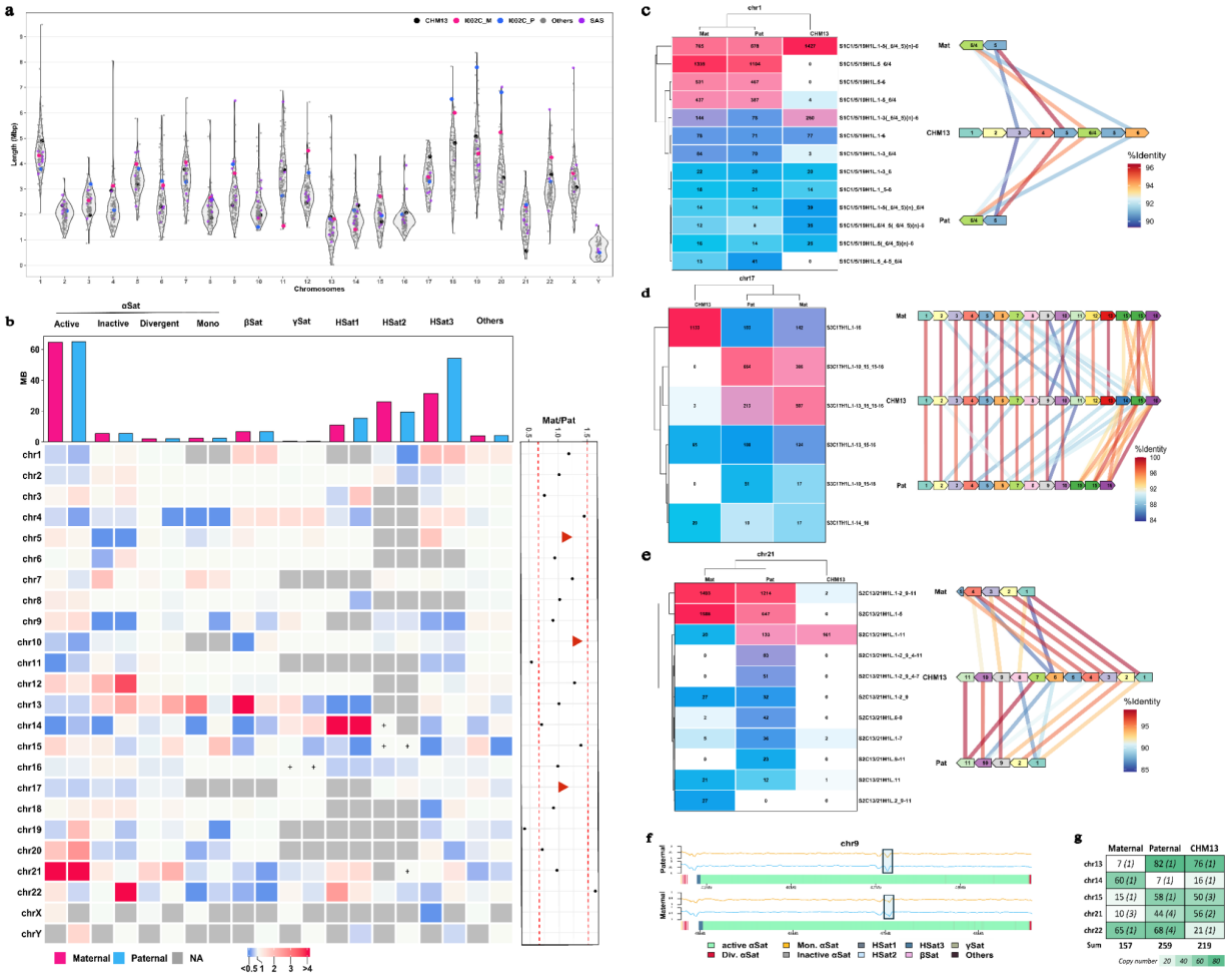


Figure 3: Peri/Centromeric satellite diversity, methylation, and rDNA variation among the I002C maternal, paternal and CHM13. a) Comparison of HOR lengths across I002C haplotypes and 165 human genomes. **b)** Heatmap showing the ratio of peri/centromeric satellite lengths in I002C haplotypes relative 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)** Left: Heatmaps showing the abundances of top 10 most frequent HORs in each of I002C haplotypes and CHM13. Right: Structural variation and sequence similarity of the most abundant HORs. **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

339 indicated by shaded boxes. **g)** Heatmap of ribosomal DNA (rDNA) copy numbers, as well
340 as major morph count [CN (morph count)] across I002C haplotypes and CHM13.

341 Chromosome Y analysis

342 Assembling the human Y chromosome to a T2T chromosome remains challenging due
343 to its highly repetitive architecture, especially within large heterochromatic regions and
344 was only recently completed[19]. The I002C assembly provides a T2T, gapless Y
345 chromosome from a South Asian individual, representing a valuable addition to the
346 current catalogue of Y chromosome diversity[34]. The total length of I002C-Y is 56.6 Mb,
347 consistent with other recent Y chromosome assemblies ranging from 45 to 85 Mb[34]
348 (Additional file 1: Figure S21). Notably, it is comparable in size to other South Asian
349 individuals, including HG03009 (58.8 Mb), HG02492 (51.66 Mb), HG03732 (51.5 Mb),
350 and HG03492 (51.2 Mb).
351

352 Comparative analysis of I002C-Y and T2T-Y from CHM13 (v2.0) showed concordance in
353 overall structure (Additional file 1: Figure S22), with the largest size differences observed
354 in ampliconic and heterochromatic subregions (Additional file 2: Table S15). Specifically,
355 the AMPL2 subregion in I002C-Y is 11.7% (375 kb) shorter, the centromeric
356 HET1_centro2 is 18.5% (327 kb) longer, HET2_DYZ19 is 52.7% (140 kb) longer, and
357 HET3_Yq12 is 18.1% (6.25 Mb) shorter than their counterparts in T2T-Y. The multi-
358 megabase variation observed in Yq12 is expected, as this region is composed of
359 alternating repeat arrays known to be the largest heterochromatic block in the human
360 genome and exhibits extensive inter-individual variability[34]. Sequence identity between
361 subregions of I002C-Y and T2T-Y exceeds 97% for most regions, with the average
362 identity being 96.65%. The most divergent region is HET2_DYZ19, with 89.4% identity,
363 followed by two centromeric regions HET1_centroDYZ3 (93.9%) and HET1_centro2
364 (94.5%). In addition, we identify two large inversions when compared to T2T-Y, one in
365 AMPL6 and the other one in AMPL7, and the one in AMPL6 was also observed when
366 comparing to GRCh38 (Additional file 1: Figure S22). These inversions belong to

367 palindromes P5 and P1, respectively, and are noticed in other recent Y chromosome
368 assemblies, including South Asian ones[34].

369

370 Finally, haplogroup assignment based on I002C-Y sequence places this individual in the
371 R-Y6 lineage, a Y-chromosome haplogroup prevalent in South Asian populations and
372 estimated to have emerged ~4,000–4,500 years ago, underscoring the deep-rooted
373 paternal ancestry in the region.

374 Structural Variation in I002C Relative to CHM13 and Between the Paternal 375 and Maternal Haplotypes

376

377 In addition to variations in peri/centromeric repeats, whole-genome comparison between
378 I002C haplotypes and CHM13 identified 14,943 structural variants (SVs; ≥50 bp)
379 spanning ~19.50 Mbp in the maternal and 14,243 SVs spanning ~18.74 Mbp in the
380 paternal genome (excluding large repeats >10kb), with ~95% of these being insertions
381 and deletions (Additional file 2: Table S16–S17). Comparing I002C SVs against the
382 combined SVs from public databases, including gnomAD[35], HGSVC[28,36], and HPRC[7],
383 showed that approximately 13.63% (3,236) of SVs were found to be unique to I002C
384 (Additional file 2: Table S18). These I002C-specific SVs were primarily duplications and
385 deletions (~34% each), followed by insertions (30%) (Figure 1; Additional file 2: Table
386 S18). Further repeat annotation showed that 72% of these unique SV regions overlapped
387 with various repeat types (Additional file 2: Table S19).

388

389 Comparative analysis of the two complete haplotypes revealed approximately 2.90 million
390 SNVs, 329,328 small indels (<50 bp), and 16,308 SVs. The heterozygosity rate,
391 calculated as the number of SVs per 500 kb, showed that highly divergent regions are
392 enriched in centromeric and subtelomeric regions (Figure 2d). In these regions, the SV
393 density was more than four fold higher, and the SNV density more than two fold higher,
394 compared to the genome-wide average (Additional file 1: Figure S23).

395

396 **Population-specific Reference Reduces Variant Calling Bias**

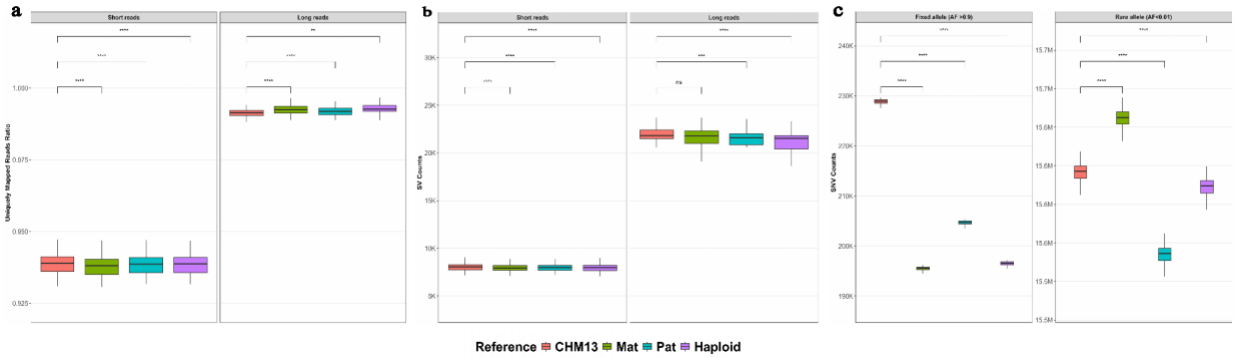
397 To evaluate the impact of reference bias on read mapping and variant calling (SNV and
398 SV), we compared four complete haploid reference genomes: three I002C haplotypes
399 (maternal, paternal, and haploid version including mitochondrial DNA) and CHM13, using
400 both short-read (NGS) and long-read (ONT) sequencing data. We randomly selected a
401 subset of publicly available NGS data from the 1000 Genomes Project (1KGP)[23],
402 comprising 50 individuals from each of five South Asian (SAS) subpopulations (BEB, GIH,
403 ITU, PJL, STU) (Additional file 2: Table S20). Additionally, we used ONT data from the
404 Singapore HELIOS study[37], including 26 Chinese, 31 Indian, and 37 Malay individuals,
405 as well as NGS data from the same study including 39 Chinese, 39 Indian and 42 Malay
406 individuals. Raw sequencing reads from these individuals were mapped to each reference
407 genome, and variants were identified from each alignment.
408

409 For NGS data, we observed that the rate of uniquely mapped reads was marginally higher
410 on CHM13 in both SAS populations and Singaporeans compared to either the maternal
411 or paternal I002C haplotype. In SAS, the median of the rate reduced 0.09%, 0.02% and
412 0.06% for maternal, paternal and haploid, respectively, while in Singaporean samples, it
413 reduced 0.02%, 0.05% and 0.02%, respectively. However, for long read (ONT) data of
414 the Asian samples from Singapore the mapping rate is significantly lower on CHM13 than
415 I002C (Additional file 1: Figure S24). The median of the rate increased 0.12% for
416 maternal, 0.06% for paternal and 0.14% for haploid (Figure 4a). Interestingly, in the same
417 SAS samples with short-read data, the number of detected SVs was generally lower when
418 using the I002C haplotypes compared to CHM13 (Figure 4b, Additional file 1: Figure S25).
419 An exception was observed in the PJL subgroup, where a slightly higher number of SVs
420 was identified using the paternal or haploid genome. A similar trend was observed in long-
421 read data and short-read data across the three Singaporean ethnic populations (Chinese,
422 Malays and Indians), although the number of SVs was slightly higher in Chinese samples
423 mapped to the maternal genome in long-read data or to the paternal or haploid genome
424 in short-read data (Additional file 1: Figure S25).

425 To evaluate reference bias in variant calling at both individual and population cohort
426 levels, variants were identified using deepvariant from mapping files of SAS, Singapore
427 Indian, Chinese and Malay samples. For the SAS population including the Singapore
428 Indian samples, an average of 25.76 million bi-allelic SNVs were identified across four
429 reference genomes (Additional File 2: Table S21). However, allele frequency profiles
430 differed significantly, especially for fixed or nearly fixed variants (AF > 0.9) between
431 CHM13 and I002C genomes (Figure 4c). Using CHM13 as reference, 230,031 high-
432 frequency variants were detected, compared with an average of 199,918 when using the
433 I002C assemblies (maternal, paternal, and haploid) (Additional File 2: Table S21). In
434 contrast, I002C references yielded a greater number of rare alleles (MAF < 0.01),
435 consistent with improved representation of population-matched haplotypes. This allele
436 frequency spectrum, indicating reduction in fixed variants and an enrichment of rare
437 alleles, suggests that the I002C assemblies better capture the genetic landscape of the
438 SAS population compared to CHM13 (Figure 4c). Similar results were observed for
439 Singapore Chinese and Malay population (Additional file 2: Table S22-S23, Additional file
440 1: Figure S26).

441 We then assessed SVs that were specifically called by using CHM13 as reference
442 (CHM13-specific SVs) after projecting their coordinates to GRCh38. SVs were merged
443 across all SAS samples (250 samples with 183,811 SVs) and all Singaporean samples
444 (94 samples with 138,909 SVs for long reads data, 120 samples with 91039 SVs for short
445 reads data), respectively. Of these SVs, 21.2% (39,051) from SAS short reads data, 21%
446 (19,082) and 15.4% (21,353) from Singaporeans short reads or long reads data were
447 unique to CHM13. Of these CHM13-specific SVs, 5.5% in SAS short-read data, 9.4% and
448 18.8% in Singaporeans short- and long-read data were shared across at least 50% of the
449 population. These additional SVs identified using CHM13 as the reference are likely
450 variants that are either fixed or highly prevalent in the SAS or Singaporean populations
451 and thus also present in the I002C haplotype genomes, while others may also be truly
452 private to CHM13. Notably, 71 and 152 clinically relevant genes[38] were affected by fixed
453 or common SVs observed in over 90% of SAS or Singaporean individuals, respectively,
454 indicating potential population-specific relevance in clinical interpretation. Notably, the
455 gene TSC2—included in the ACMG SF v3.2[39] list of actionable genes recommended for

456 opportunistic screening was affected by an SV present in 99% of the SAS population or
457 94% of Singaporean individuals. Additionally, two other ACMG genes, TRDN and STK11,
458 were affected by SVs present in 100% of Singaporean individuals. Further, OPRM1 and
459 DPYD, both included in the Clinical Pharmacogenetics Implementation Consortium
460 (CPIC) gene panel, were associated with SVs commonly found in 87% of SAS samples
461 and 91.5% of Singaporeans. These SVs, particularly the clinically relevant ones, would
462 not be called by using I002C as a reference, eliminating the need to do subsequent clinical
463 annotation.



464

465 **Figure 4: Comparative performance of I002C and CHM13 references in South Asian**
466 **genomes. a)** Mapping statistics of short and long-read sequencing data from SAS
467 samples against I002C and CHM13 references. **b)** Allele frequency spectrum showing
468 differences in the number of fixed or nearly fixed variants (AF > 0.9) and rare alleles (MAF
469 < 0.01) across references. c) Comparison of structural variant counts detected using
470 I002C and CHM13 references. The paired Wilcoxon test was used to assess differences
471 in medians between the two groups. Significance levels: ns (P > 0.05), * (P ≤ 0.05), ** (P
472 ≤ 0.01), *** (P ≤ 0.001), **** (P ≤ 0.0001).

473 Genome-wide Analysis of Allelic DNA Methylation

474 Besides the analysis of methylation patterns within the centromere regions, we also
475 analyzed the genome-wide methylation pattern in the I002C genomes. Single-base
476 resolution DNA methylation patterns were extracted from both Nanopore R10.4.1 and
477 PacBio Revio data. The results showed high concordance between the two data, with a

478 Pearson correlation coefficient of 0.953 when analyzed using 1000 bp sliding windows
479 with 100 bp steps (Additional file 2: Table S24).
480

481 To enable parental-origin–aware haplotype-resolved methylation profiles, we assigned
482 sequencing reads to maternal and paternal haplotypes using parental-specific k-mers.
483 This analysis identified 70,400 single-CpG differentially methylated loci (DMLs) between
484 paternal and maternal haplotypes. These DMLs were not randomly distributed, but
485 instead clustered into regions with consistent haplotype-specific methylation. Such
486 differentially methylated regions (DMRs) may have functional significance, and their
487 relevance to disease development is exemplified by well-characterized imprinting
488 disorders, including Prader-Willi syndrome (OMIM 176270)[40–42], Angelman syndrome
489 (OMIM 105830)[41,43], Beckwith-Wiedemann syndrome (OMIM 130650)[44,45], Silver-Russell
490 syndrome (OMIM 180860)[45,46], Temple syndrome (OMIM 616222)[47], Kagami–Ogata
491 syndrome (OMIM 608149)[48], transient neonatal diabetes mellitus (OMIM 601410)[49], and
492 pseudohypoparathyroidism Ib (OMIM 603233)[50]. Using these known imprinting regions
493 as benchmarks, we calibrated our thresholds and identified 968 autosomal DMRs
494 (Additional file 1: Figure S27), of which 86.4% overlapped regulatory elements such as
495 promoters and enhancers (Additional file 2: Table S25). To evaluate the robustness of
496 our approach, we intersected these DMRs with 93 previously reported imprinting regions
497 supported by multiple studies[51–56]. Reassuringly, all 46 overlapping regions (Additional file
498 1: Figure S28, Additional file 2: Table S26) matched the reported methylated allele,
499 demonstrating strong concordance. The remaining reported regions were not detected,
500 likely due to the stringent criteria applied (≥97.5% methylation difference at each DML
501 and ≥5 DMLs per DMR).

502

503 Intersecting the 968 DMRs with experimentally validated promoter and first exon regions
504 in GRCh38 revealed 2,229 transcripts (596 genes) potentially regulated by these DMRs

505 (Additional file 1: Figure S29). In the I002C cell line, 203 transcripts (67.2% of those
506 expressed) exhibited monoallelic expression, and 142 (70.0% of 203) showed concordant
507 allele-specific methylation and transcription, corresponding to 94 genes (Additional file 2:
508 Table S27). In addition, among the 596 candidate genes, 53 have previously been
509 reported as imprinted[57], including 22 genes directly linked to the eight imprinting disorders
510 (Additional file 2: Table S28), highlighting the sensitivity of our approach and its potential
511 in suggesting novel candidates of imprinting genes.

512

513 Gene Annotation

514 To provide a comprehensive set of gene and transcript annotations, we first mapped
515 Gencode annotations (v48) to the I002C maternal and paternal genome assemblies using
516 liftOver[58] and liftoff[59]. This process successfully transferred 98.92% of annotated
517 transcripts to the I002C maternal and 96.41% to paternal assembly, with an overall
518 99.39% of annotated transcripts to the I002C (Additional file 2: Table S29).

519 To obtain a more complete set of genome annotation, we generated long read RNA-Seq
520 data from I002C and used Bambu (v3.10.1)[60] to build additional transcript models (Figure
521 5a). We observed that the population-matched reference genome (I002C) resulted in
522 higher alignment rates and lower alignment errors than GRCh38 and CHM13 (Additional
523 file 2: Table S30). Using a more conservative threshold (NDR=0.3), we identify 6190 novel
524 transcripts in the maternal genome, and 5832 in the paternal genome, with more than
525 21,000 novel transcripts when a sensitive threshold is used (NDR=1, Additional file 2:
526 Table S31). We then combined reference annotations from the maternal and paternal
527 genomes to obtain a single haploid reference genome annotation for I002C), with 24.5%
528 (NDR = 0.3) of transcripts being expressed in the I002C cell line (Figure 5b, c). In addition,
529 209 annotated transcripts in v48 are re-discovered by RNA-seq analysis, which failed
530 liftOver. This integrated Lifting and long read RNA-seq approach improved annotation
531 completeness, leading to the final numbers of 386477 transcript counts and 79221 genes

532 for maternal haplotype, and 376337 transcripts and 76937 genes for paternal haplotype,
533 when the more conservative threshold (NDR=0.3) is used (Additional file 2: Table S31).

534 Next we studied the genome divergence between I002C, GRCh38 and CHM13 for I002C-
535 specific transcript annotations. We aligned the transcript cDNA sequences both
536 annotated and newly discovered by Bambu with different NDR thresholds from I002C
537 against GRCh38 and CHM13, and calculated the number of alignment mismatch
538 positions to estimate the sequence variation, and the number of deletion and insertion
539 positions to estimate structural variation. While the majority of transcripts are shared
540 between I002C and GRCh38 (96%), we observe a number of transcripts with substantial
541 sequence variation (Figure 5d-g). Among these are for example novel transcripts for the
542 CASP8AP2 and CES1P1 gene which contain a previously unannotated exon that is
543 missing from the GRCh38 and CHM13 reference genomes respectively (Figure 5f, g).
544 These novel transcripts will not be quantified if GRCh38 or CHM13 is used as reference,
545 highlighting the value of a population-matched reference genome for transcriptomic
546 analysis.

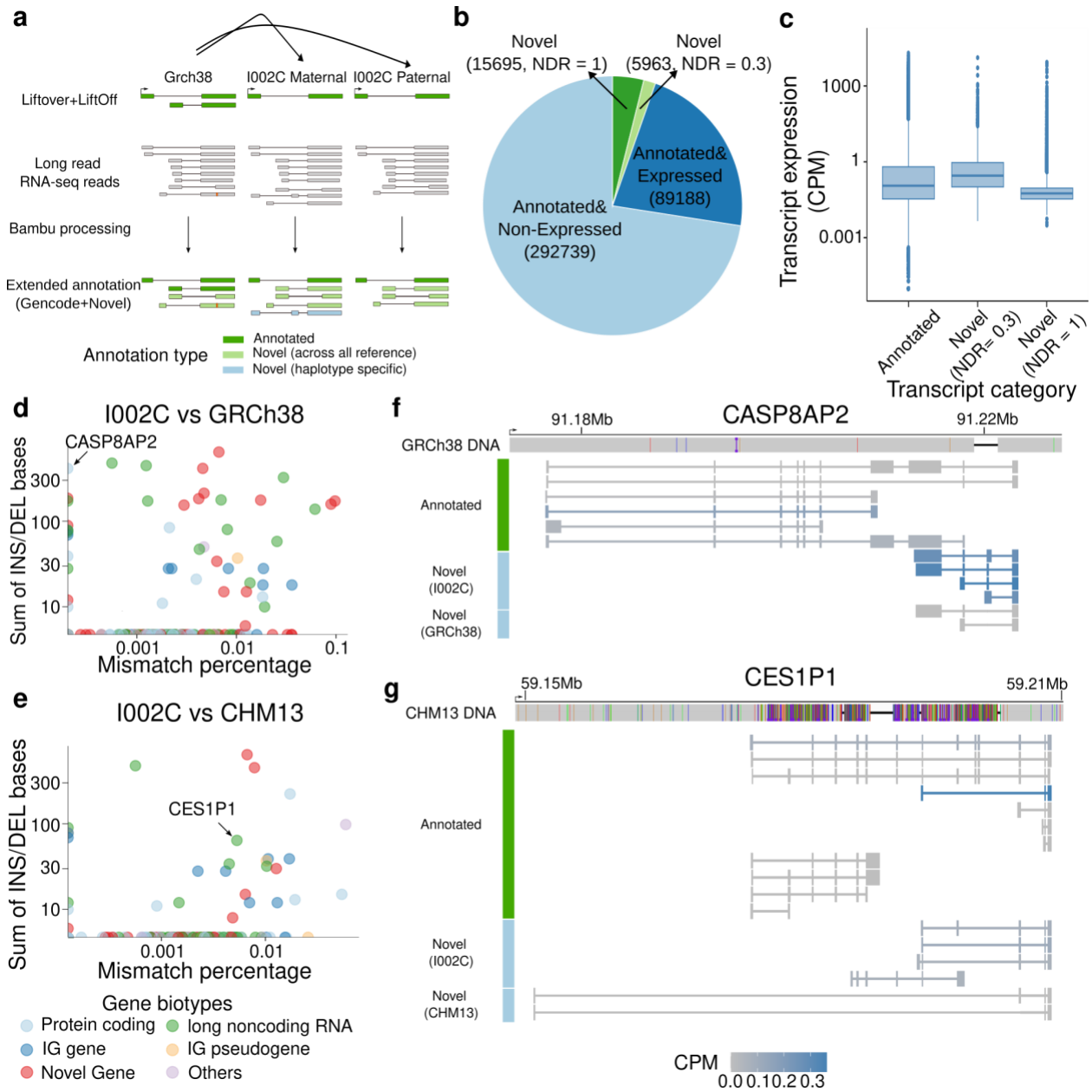


Figure 5. Gene annotation using long-read RNA-seq reads a) Schematic of RNA-seq processing pipeline: 1) Firstly, we performed liftOver and liftoff of annotations from GRCh38 to I002C Maternal and Paternal haplotype separately; 2) We then processed long-read RNA-seq reads with Bambu using GRCh38, I002C Maternal and Paternal as reference respectively; 3) Extended annotation obtained from Bambu can be categorised by annotated, or novel (shared across all references), novel transcripts specific to individual haplotype **b)** Pie chart of I002C haplotype transcript number categorised by annotation and expression level in RNA-seq analysis. Light blue represents annotated

556 and non-expressed (CPM = 0) transcript, dark blue represents annotated and expressed
557 (CPM > 0) transcript, light green represents novel transcript identified with NDR = 0.3,
558 and dark green represents additional novel transcript identified with NDR = 1 **c)** Boxplot
559 of transcript expression in count per million (CPM) for annotated transcript, novel
560 transcript identified when NDR = 0.3, and novel transcript identified when NDR = 1 **d-e)**
561 Scatter plot of percentage of mismatch vs sum of insertion and deletion bases (small
562 indels < 5bp filtered before sum) when I002C haplotype-specific novel transcripts aligned
563 against d) GRCh38 and e) CHM13 reference, focusing on transcripts containing at least
564 3 exons and CPM greater than 0.04. Note that for each gene, the transcript showing
565 largest deviation (rank by sum of insertion and deletion bases and then followed by
566 mismatch percentage) is kept in the plot for visualization. Light blue represents protein
567 coding genes, dark blue represents IG genes, green represents long noncoding RNAs,
568 orange represents IG pseudogenes, pink represents novel genes, and purple represents
569 other pseudogene types. **f-g)** Top: f) GRCh38 and g) CHM13 DNA sequence aligned
570 against I002C maternal reference. Bottom: From top to bottom, transcript annotations for
571 annotated transcripts, novel transcript specific to I002C maternal haplotype, novel
572 transcripts specific to f) GRCh38, or g) CHM13. Transcript annotations are colored by
573 CPM, with darker values indicating higher CPM.

574

575 ## Discussion

576 In this study, we present the first telomere-to-telomere (T2T), fully phased diploid genome
577 assembly of an Indian individual using a trio sample set of Indian ethnicity from Singapore.
578 The majority of Singaporean Indians descend from migrants during the British colonial
579 period, primarily of Telugu and Tamil origin, with some contributions from Northern groups
580 such as Sikhs and Pathans[12]. It is therefore assuring to see that I002C is well embedded
581 within the 5 SAS population samples from the 1000 Genomes Project. More interestingly,
582 I002C is genetically located in the middle area between the Western and Northern SAS
583 samples (GIH and PJL) and the Eastern, Southeastern and Southern SAS samples (BEB,
584 ITU and STU). As such, the I002C genome reflects broader Indian subcontinental

585 ancestry and offers value to individuals of similar backgrounds in South Asia and
586 worldwide. Additionally, it contributes to improving representation for these South Asian
587 populations in human genomic datasets.

588 Using high-quality long read data, trio-aware assembly, ultra long-read gap filling, and
589 multi-platform polishing, we achieved complete, gapless assemblies of all the autosomes,
590 sex chromosomes, and a fully resolved circular mitochondrial genome representative of
591 the individual. The final assemblies reached NG50 values of over 154 Mb for maternal
592 146 Mb for paternal and Merqury QV scores exceeding 82 for both haplotypes,
593 establishing I002C as the highest-quality diploid and haploid human genome assembled
594 to date. To the best of our knowledge, the maternal chromosome 21 rDNA sequence of
595 the I002C was the first fully assembled rDNA cluster sequence that has been published.
596 Furthermore, I002C also contributes one of the few complete Y chromosome assemblies
597 from a South Asian individual, expanding current knowledge of Y-linked diversity and
598 structural variation.

599 As the first reference genome for the South Asian population, I002C highlights important
600 features that remain incompletely characterized in many reference genomes. For
601 example, in addition to providing the first fully assembled rDNA cluster on Chr21
602 (maternal), I002C also revealed 19 major rDNA morphs, compared to the 8 major rDNA
603 morphs for CHM13. Comparison of I002C genome to CHM13 revealed that over 13% of
604 SVs in I002C are novel, with many falling in repeat-rich regions, which are often sources
605 of structural complexity in the genome. When comparing two I002C haplotypes, the
606 analysis revealed 2.9 million SNVs, more than 329 thousands small indels and more than
607 16 thousands SVs. Many of these variants were concentrated in centromeric and
608 subtelomeric regions, identifying them as hotspots of structural divergence. In addition,
609 the fully resolved diploid genome enabled the generation of haplotype-resolved
610 methylome and long read RNA expression data. This data enables the identification of
611 parent-of-origin–aware differentially methylated regions (DMRs) that show high
612 concordance with DMRs reported in previous studies, where parent-of-origin information
613 was obtained using strand-seq, germline variant–based allelic determination, or special
614 uniparental samples such as uniparental disomy, complete hydatidiform moles, and

615 ovarian teratomas, which contain chromosomes from a single parent. We have further
616 demonstrated that the DMRs within core transcription regulatory sequences showed high
617 concordance with haplotype-resolved long read RNA expression data. These DMRs
618 have captured many of known imprinting genes, particularly the imprinting disease-
619 associated ones, and suggest additional candidates of novel imprinting genes. Together,
620 these findings highlight the value of fully phased diploid genomes for exploring complex
621 and polymorphic regions.

622 Our analysis has demonstrated that the I002C is a better reference than CHM13 for the
623 genomic analysis of South Asian samples as well as other Asian samples by reducing
624 population bias. For example, population-specific analysis, based on 250 randomly
625 selected SAS individuals from the 1000 Genomes Project, showed comparable rates of
626 uniquely mapped short reads across references. However, for Singapore samples, the
627 mapping rate for long read sequencing data is higher for I002C than CHM13, which is
628 consistent with the fact that I002C carries a good number of SVs that are unique or much
629 more common in South Asian and Southeast Asian samples than other populations.
630 These SVs are expected to have a negative impact on mapping long reads than short
631 reads by using CHM13 as reference. In addition, analysis of SNVs showed that allele
632 frequency profiles differed between references, with CHM13 yielding more high-
633 frequency variants and I002C assemblies capturing more rare alleles, consistent with
634 improved representation of population-matched haplotypes. Similar patterns were also
635 observed for Singapore Chinese and Malay populations. The number of called SVs was
636 reduced when mapping SAS samples to I002C instead of CHM13, reflecting the presence
637 of population-specific sequences in I002C. Reduction of variant calls by using I002C as
638 a reference can help to reduce the effort on identifying potential pathogenic variant(s) for
639 genomic diagnostic analysis of SAS patients. The additional SVs identified with CHM13
640 as a reference included variants in medically relevant genes, such as TSC2, which was
641 affected in 99% of SAS samples. Such additional SVs, particularly the medically relevant
642 ones, may lead to misdiagnosis for SAS patients. Together, these findings suggest that
643 the I002C haplotypes more accurately represent the genomic architecture of the SAS
644 population and also serve as a suitable reference for Southeast Asian population
645 particularly the Singapore Malay and Chinese samples. This underscores the value of

646 population-specific assemblies in reducing reference bias and improving accuracy in
647 downstream analyses such as variant discovery and association studies.

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

656 # Methods

657 ## Sample selection and sequencing

658 The details of sample selection, Pacbio HiFi, Oxford Nanopore duplex, ultra-long and
659 Omni-C data generation are described in the previous study[61].

660 ## Karyotyping

661 Cells were cultured in RPMI 1640 medium (#11875-093, Gibco, USA) supplemented with
662 10% fetal bovine serum (#A5670801, Gibco) and 1% penicillin–streptomycin (#15140-
663 122, Gibco), and maintained in suspension at 37 °C in a humidified incubator with 5%
664 CO_2. Approximately 5 million cultured cells were treated with KaryoMAX™ Colcemid™
665 Solution in HBSS (#15210040, Gibco) for 2 hours to induce mitotic arrest. Following
666 standard protocols from *The AGT Cytogenetics Laboratory Manual* (4th Edition)[62], routine
667 G-banded metaphase preparations were performed[62].

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

671 **RNA Sequencing**

672 RNA was extracted using the QIAGEN PAXgene Blood RNA Kit according to the
673 manufacturer's instructions. Total RNA was used as input material for the RNA sample
674 preparations. rRNA was removed from total RNA by using TIANSeq rRNA Depletion Kit
675 (Animal). Fragmentation was carried out using divalent cations under elevated
676 temperature in the First Strand Synthesis Reaction Buffer. First strand cDNA was
677 synthesized using random hexamer primer and reverse transcriptase, followed by the
678 synthesis of the second strand by adding buffer and dNTPs (with dTTP replaced by
679 dUTP). The synthesized double-stranded cDNA was treated by terminal repair, A-tailing,
680 adaptor ligation, and PCR enrichment after fragment selection. The strand-specific library
681 was obtained by selecting cDNA fragments of 370-420bp by purification with AMPure XP
682 beads. The qualified library was pooled and sequenced on the Illumina NovaSeq X Plus
683 platform with a 150 × 2 paired-end configuration.
684

685 **Nanopore cDNA sequencing**

686 Total RNA was extracted from the I002C lymphoblastoid cell line using the Bioline Isolate
687 II RNA Mini Kit (Meridian Biosciences) according to manufacturer's instructions. Total
688 RNA quantification, purity and integrity assessment was performed using the Qubit 3.0
689 Fluorometer and Qubit BR RNA Assay Kit (ThermoFisher Scientific), NanoDrop 2000
690 spectrophotometer (ThermoFisher Scientific) and Agilent 4200 TapeStation system with
691 the RNA screentape assay kit (Agilent). 39μg of total RNA was brought forward to
692 poly(A)+ RNA enrichment using the Dynabeads mRNA Purification Kit (ThermoFisher
693 Scientific) according to manufacturer’s instructions. Nanopore cDNA-PCR library was
694 generated from 50ng of poly(A)+ RNA using the SQK-PCB114.24 kit, (Oxford Nanopore
695 Technologies) according to manufacturer’s instructions. Briefly, complementary strand
696 synthesis and strand switching were performed on input full-length poly(A)+ RNA using
697 kit-supplied oligonucleotides, 16-cycle PCR amplification using primers containing 5’ tags

698 to generate a double-stranded cDNA library, followed by ligase-free attachment of rapid
699 sequencing adaptors. Final cDNA library was loaded onto a FLO-PRO114M R10.4.1
700 flowcell (Oxford Nanopore Technologies) and sequenced on a PromethION P24 A100
701 instrument, PromethION release version 25.03.7, Dorado server 7.8.3, equivalent to
702 Dorado release 0.9.2, with live-basecalling in super-accurate (SUP) mode and barcoding
703 trimming.

704 **WGS Analysis by Next Generation Sequencing (NGS)**

705

706 WGS PCR-free libraries for the gDNA samples were prepared using MGIEasy FS PCR-
707 free DNA library prep set (MGI; Cat. no. 1000013455) as per manufacturer's protocol.
708 The WGS libraries were quality checked by the HSD 1000 Screentape and reagent
709 (Agilent Technologies; cat. no. 5067-5585 and 5067-5584) on the Tapestation 4200
710 (Agilent Technologies). The single strand circularized DNA product of the library was
711 measured with Qubit™ ssDNA HS Assay Kit (Thermo Fisher Scientific; Cat. no. Q10212)
712 on the Qubit fluorometer 4 (Thermo Fisher Scientific). DNA nanoball (DNB) was made for
713 each library from the single strand circularized DNA using the DNBSEQ DNB Rapid Make
714 Reagent Kit (FCL PE150) (MGI; 1000028453) and concentration of DNB were determined
715 by the Qubit™ ssDNA HS Assay Kit (Thermo Fisher Scientific; Cat. no. Q10212) on the
716 Qubit fluorometer 4 (Thermo Fisher Scientific). DNBs were pooled based on the
717 sequencing data requirement, and the pooled DNBs were sequenced 150bp paired-end
718 (PE150) on the MGI DNBSEQ-T7 sequencer using the DNBSEQ-T7RS High-throughput
719 Sequencing Kit (FCL PE150) V1.0 sequencing reagent (MGI; Cat. no. 1000016106) and
720 DNBSEQ-T7RS Sequencing Flow Cell (T7 FCL) (MGI; Cat. no. 1000016269).

721 **Genome Assembly and Polishing**

722

723 The initial assemblies were constructed using PacBio HiFi and uncorrected ONT ultra-
724 long reads in trio mode with Verkko (v1.4.1)[20] and Hifiasm (v0.19.6-r595)[21,22] algorithms.
725 Contigs were further assembled into full-length chromosomes through manual curation
726 with 3D-DNA[63] and a template-based approach using the CHM13 genome to determine
727 the correct contig order for joining.

728 To support gap filling and manual curation, ONT ultra-long reads were also corrected
729 using HERRO[64] and assembled in trio mode with Verkko. In addition, all long reads were
730 haplotype-binned and reassembled using both Hifiasm and LJA (v0.2)[65].
731 Some gaps were falsely introduced during the scaffolding process, as confirmed by long-
732 read and Hifiasm and LJA reassembled contig alignments to the assembly gap regions.
733 These false gaps appeared as large deletions in the alignments and were subsequently
734 removed. We considered all the remaining gaps to be real gaps. Real gaps were first
735 closed using TGS-GapCloser (v1.2.1)[25]. The remaining gaps were manually filled by
736 aligning corrected ultra-long reads, Verkko HERRO reads assembly, as well as Hifiasm
737 and LJA reassemblies to the 10 kb flanking regions around each gap.
738 Alignments were manually examined in IGV (v2.15.2)[66], and only reciprocal best hits were
739 selected for gap filling. Telomeres were patched and extended using ultra-long, duplex
740 and HiFi reads containing the (TTAGGG)n motif at either end, identified using
741 Tel_Sequences[67]. These reads were aligned to the first or last 250 kb of each
742 chromosome, and consensus sequences were generated using SPOA (v4.1.4)[68] for
743 patching, resulting in 23 gapless telomere-to-telomere chromosome level assemblies for
744 each haplotype.
745
746 The assembly was polished through two rounds of automated processes using the T2T
747 polishing pipeline[26], followed by one round of manual curation and an additional round of
748 ONT-based assembly patching. In brief, haplotype-specific binned long reads (HiFi,
749 duplex and ultra-long) reads were mapped to their respective haplotypes using
750 winnowmap2 (v2.03)[69] with default platform-specific parameters. PCR-free MGI short
751 reads were aligned using BWA MEM (v0.7.17-r1188)[70], and the resulting sorted HiFi and
752 short alignment files were merged using SAMtools (v1.18)[71]. SNV + indel (<50 bps) were
753 called using DeepVariant (v1.6.1)[72] in "HYBRID_PACBIO_ILLUMINA" for merged HiFi,
754 short reads and in “ONT_R104” mode for duplex reads, while ultra-long reads were
755 processed with PEPPER-Margin-DeepVariant (v0.8)[73]. SVs were identified using
756 Sniffles2 (v2.2)[74] and subsequently refined and merged using the Iris (v1.0.5) - Jasmin
757 (v1.1.5)[75]. These variants were filtered using merfin (v1.0)[76] and used for genome
758 polishing with bcftools (v1.10.2)[77]. During the manual curation round, variants were

759 identified as previously described and carefully examined in IGV for genome polishing.
760 Subsequently, remaining erroneous regions within 1 kb window were merged using
761 BEDTools (v2.30)[78], and 1 kb flanking sequences, including the erroneous regions, were
762 extracted. These sequences were aligned to the corresponding haplotype assemblies
763 from corrected ultra-long assemblies using Minimap2 (v2.28)[79] to identify reciprocal best
764 matches. Only error-free regions from the ultra-long assemblies were used for patching
765 (Additional methods: **Genome assembly**).

766 **Evaluation of Assembly**

767 Assembly quality (QV) was assessed using hybrid k-mers derived from HiFi and MGI
768 reads, with completeness and phasing statistics calculated using Merqury (v1.3)[27].
769 Assembly errors and overall quality were also evaluated using CRAQ (v1.0.9-alpha)[80],
770 which utilizes read clipping information, and Inspector (v1.3)[81], which detects both large
771 (SV) and small-scale (SNV) errors based on reads-to-assembly alignments. Additionally,
772 coverage-based tools such as Flagger (v0.3.3)[7] and GCI (v1.0)[82] were employed to
773 identify errors based on coverage inconsistencies. Haplotype assemblies were further
774 validated using Omni-C contact maps, which were visualized in Juicebox (v1.11.08)[83].

775 **Repeat Annotation and Centromere Analysis**

776

777 Repeat annotation for both haplotypes was performed using RepeatMasker (v4.1.5)[84] with
778 a curated repeat library from Dfam (v3.7)[85]. Centromere alpha satellite (αSat) monomers
779 were identified using Hum-AS-HMMER[86], and high-order repeats (HORs) were classified
780 as active/live, inactive, divergent and monomeric based on previous studies[30,31]. HSat2
781 and HSat3 were annotated following established methods[31] while other centromeric
782 satellites were identified by parsing RepeatMasker output. Structural variants within active
783 centromeric HORs were predicted using Stv[87].

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)[88] and svim-asm (v1.0.3)[89]. 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)[35], HGSVC (v1–3)[28,36], and HPRC[7]. Since SV coordinates for I002C and HPRC were based on CHM13, whereas gnomAD and HGSVC SVs were based on GRCh38, the latter were lifted over to CHM13 using Picard LiftOverVCF (v2.26.9)[90]. 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. Moreover, the NGS data from the Singapore HELIOS cohort were used to facilitate a comprehensive evaluation . The

812 corresponding CRAM files were converted to FASTQ format using samtools (v1.18).
813 Paired-end reads from each sample were aligned separately to the four reference
814 genomes using BWA-MEM (v0.7.17). Duplicate reads were marked using GATK
815 MarkDuplicates (v4.1.8.1)[91].

816 For long-read analysis, we used ONT data from the Singapore HELIOS cohort. The reads
817 were aligned to each reference using minimap2 (v2.28). Mapping statistics were collected
818 from each BAM file using samtools, and the percentage of uniquely mapped reads was
819 calculated using the formula ($N_{Mapped\ reads}$ -N_{MQ0}) / $N_{Mapped\ reads}$.
820 ***Variant calling*** Short-read variants were identified with DeepVariant from duplicate-
821 marked read alignments to four reference genomes. Long-read variants were called using
822 Clair3[92] with the "r941_prom_sup_g5014" model. Bi-allelic SNPs were extracted with the
823 bcftools view "-m2 -M2 -v snps -f PASS" options. Variants with read depth (DP) ≥ 10 and
824 genotype quality (GQ) > 10 were retained. Allele frequency (AF) were calculated
825 separately for SAS and Singapore HELIOS cohorts (Chinese and Malay). Variants with
826 AF > 0.9 were classified as fixed or nearly fixed, while those with AF < 0.01 were
827 considered rare.

828 ***Structural variant calling*** To further assess reference genome performance, structural
829 variants (SVs) were called from the NGS alignments using Manta (v1.6.0)[93], and from
830 ONT alignments using Sniffles (v2.6.2). To evaluate the reference specific SVs, we first
831 projected SVs on GRCh38 coordinates using CrossMap (v0.7.0)[94]. The chain files from
832 I002C to GRCh38 were generated following the nf-LO pipeline[95].

833 A combined SV set for each haplotype was generated by merging SV across all samples
834 using Jasmine (v1.1.5) with 'max_dist=1000'. The same merging strategy was used to
835 merge three reference-based SV files to identify the reference-specific SVs.
836

Gene Annotation

To annotate genes on the I002C maternal and paternal genome assemblies, as part of the gene annotation pipeline (Additional file 1: Figure S30; Additional methods: **Gene annotation**), we first lifted gene models from GRCH38 (Gencode annotations (v48)) to our assembly using liftOver (conda version 377). Chain files for both haplotypes were generated from GRCH38 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 gencode annotations (v48) onto the I002C assemblies. In parallel, we also applied liftoff (v1.6.3), which directly maps gene models by aligning genomic features rather than relying on precomputed chain files, allowing a more flexible transfer in structurally divergent regions. Finally, annotations from both approaches were merged using a set of predefined rules (Additional methods: **Gene annotation**), with the procedure giving greater weight to liftoff annotations.

Reference genome annotations were then extended by using long read RNA-seq data. First, reads were aligned to each haplotype-specific assembly using Minimap2. Transcript models were then constructed and quantified using Bambu (v3.10.1)[60]. Here we run Bambu with two thresholds for high precision (NDR=0.3) and high sensitivity (NDR=1). We then extracted cDNA sequences for Bambu extended annotations using each haplotype-specific assembly and aligned against GRCh38 and CHM13 separately to identify potential sequence variation for annotated transcripts and also to check whether novel transcripts identified using each haplotype-specific assembly is haplotype-specific or shared in GRCh38 or CHM13. For each transcript, we calculated the mismatch percentage and the sum of insertion and deletion bases to measure the sequence variation. Note that for this analysis, small insertion or deletion less than 5 bp were not counted, to minimize the impact of small alignment errors. Alignment of these transcript models were then processed with Bambu with trackReads set to TRUE, and without discovery and quantification. We then defined novel transcripts as haplotype-specific if there does not exist any equalMatches (i.e., not complete splicing junction chain match)

868 to extended annotations obtained when using GRCh38 or CHM13 (including both
869 annotated and novel transcripts), CPM > 0.04, and containing at least 3 exons.
870

871 **Ancestry Analysis**

872

873 Local ancestry inference for the I002C genome was performed by calling variants relative
874 to the GRCh38 reference using dipcall (v0.3)[96]. The resulting variant set was merged with
875 biallelic SNPs from 1000 Genomes Project (1KGP) samples representing African (AFR),
876 American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS) super-
877 populations. Ancestry was inferred at the super-population level using RFMix (v2.03-r0)[97].
878

879 For the PCA analysis, we performed variant calling for the individual I002C and the 1KGP
880 South Asian (250 SAS) samples using DeepVariant, followed by joint genotyping with
881 GLnexus (v1.4.1)[98]. To minimize potential bias from close relatives, we excluded first- and
882 second-degree related samples, resulting in a final dataset of 248 individuals. Variants
883 were retained if they met the thresholds of allelic quality (AQ) > 20, read depth (DP) > 10,
884 and genotype quality (GQ) > 20. To obtain an independent set of SNPs for PCA, we
885 excluded long-range linkage disequilibrium (LD) regions (projected from GRCh38) and
886 applied additional filtering with the parameters:
887 --maf 0.1 --geno 0.03 --hwe 1e-3 --indep-pairwise 1500 150 0.2 --snps-only --autosome.
888 PCA was then conducted in PLINK (v1.9)[99] using the final dataset of 247 SAS samples
889 and I002C.
890

891 **Methylation**

892 Genome-wide and centromeric DNA methylation profiling for I002C genomes was
893 performed using PacBio HiFi and ONT simplex reads with platform specific pipelines. To
894 ensure reliable quantification, methylation ratios were calculated only for CpG sites
895 supported by at least four methylation-informative reads (containing MM/ML tags). To
896 evaluate platform concordance, Pearson correlation coefficients were computed for

897 methylation levels across shared CpG sites and predefined genomic windows. CDR for
898 each chromosome was manually examined and annotated as described in the previous
899 study[31]. Haplotype-specific differentially methylated loci (DMLs) were identified by
900 aligning parental reads (excluding homozygous reads) to the GRCh38 reference,
901 requiring a minimum methylation difference of 97.5% between the haplotypes and at least
902 five informative reads per site. Differentially methylated regions (DMRs) were defined as
903 regions containing at least five DMLs and majority of DMLs within this region to have
904 consistent methylated allele. To identify candidate genes potentially regulated by allelic
905 methylation, DMRs were intersected with core promoter and first exon regions of
906 experimentally validated transcripts (Additional methods: **Methylation**).

907 **ChrY analysis**

908

909 The subregions of I002C-Y were annotated using subregion definitions from T2T-Y. The
910 T2T-Y subregions were mapped onto I002C-Y using minimap2 (v2.28) with the
911 parameters "-x asm20 -Y -N1 --cs -c --eqx --paf-no-hit." To assess sequence
912 conservation, we computed average sequence identity for each subregion based on
913 pairwise alignments between I002C-Y and T2T-Y using nucmer (v3.1)[100] with "--
914 maxmatch -l 100" parameters. For each subregion, identity was calculated as a length-
915 weighted average across all alignments. Additionally, I002C-Y subregions were mapped
916 onto GRCh38 chromosome Y (NCBI GCF_000001405.40) using the same method as
917 described above, and the alignment identities were also computed accordingly.

918

919 To further characterize the paternal lineage, we determined the Y chromosome
920 haplogroup using short read data aligned to the I002C-Y assembly. Haplogroup
921 assignment was performed using Yleaf (v3.2.1)[101] by comparing SNPs in the non-
922 recombining region of the Y chromosome to those in the YFull phylogenetic reference
923 tree (v10.01)[102].

924

925 ## Data availability

926 The sequencing reads have been submitted to NCBI under Project ID PRJNA1150503.
927 The I002C assemblies are available at https://github.com/LHG-GG/I002C. Additionally,
928 the publicly available data resources used in this study are listed in Additional file 2: Table
929 S32, along with their corresponding access URLs.

930 ## References

931 1. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies
932 demonstrates the enduring quality of the reference assembly. *Genome Res* **27**, 849–864
933 (2017).
934 2. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
935 3. HG002. MarBL (2025).
936 4. Yang, C. *et al.* The complete and fully-phased diploid genome of a male Han Chinese. *Cell*
937 *Res* **33**, 745–761 (2023).
938 5. He, Y. *et al.* T2T-YAO: A Telomere-to-Telomere Assembled Diploid Reference Genome for
939 Han Chinese. *Genomics, Proteomics & Bioinformatics* **21**, 1085–1100 (2023).
940 6. Wang, T. *et al.* The Human Pangenome Project: a global resource to map genomic
941 diversity. *Nature* **604**, 437–446 (2022).
942 7. Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
943 8. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164
944 (2016).
945 9. Fatumo, S. *et al.* Diversity in Genomic Studies: A Roadmap to Address the Imbalance. *Nat*
946 *Med* **28**, 243–250 (2022).
947 10. Yew, C.-W. *et al.* Genomic structure of the native inhabitants of Peninsular Malaysia and
948 North Borneo suggests complex human population history in Southeast Asia. *Human*
949 *genetics* **137**, 161–173 (2018).

950 11. Wall, J. D. *et al.* The GenomeAsia 100K Project enables genetic discoveries across Asia.
951 *Nature* **576**, 106–111 (2019).
952 12. Wu, D. *et al.* Large-Scale Whole-Genome Sequencing of Three Diverse Asian Populations
953 in Singapore. *Cell* **179**, 736-749.e15 (2019).
954 13. Le, V. S. *et al.* A Vietnamese human genetic variation database. *Human mutation* **40**, 1664–
955 1675 (2019).
956 14. Deng, L. *et al.* Analysis of five deep-sequenced trio-genomes of the Peninsular Malaysia
957 Orang Asli and North Borneo populations. *BMC Genomics* **20**, 842 (2019).
958 15. Piriyapongsa, J. *et al.* Pharmacogenomic landscape of the Thai population from genome
959 sequencing of 949 individuals. *Sci Rep* **14**, 30683 (2024).
960 16. Azmi, M., Chen, L., Idris, A. & Lu, Z. H. Genomic Insights of Bruneian Malays.
961 2022.06.01.492266 Preprint at https://doi.org/10.1101/2022.06.01.492266 (2022).
962 17. Bhattacharyya, C. *et al.* Mapping genetic diversity with the GenomeIndia project. *Nat Genet*
963 **57**, 767–773 (2025).
964 18. Ardiansyah, E. *et al.* Sequencing whole genomes of the West Javanese population in
965 Indonesia reveals novel variants and improves imputation accuracy. *Front. Genet.* **15**,
966 (2025).
967 19. Rhie, A. *et al.* The complete sequence of a human Y chromosome. *Nature* **621**, 344–354
968 (2023).
969 20. Rautiainen, M. *et al.* Telomere-to-telomere assembly of diploid chromosomes with Verkko.
970 *Nature Biotechnology* **41**, 1474–1482 (2023).
971 21. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo
972 assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170–175 (2021).
973 22. Cheng, H., Asri, M., Lucas, J., Koren, S. & Li, H. Scalable telomere-to-telomere assembly
974 for diploid and polyploid genomes with double graph. *Nature Methods* **21**, 967–970 (2024).
975 23. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

24. 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).
25. 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).
26. Mc Cartney, A. M. *et al.* Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. *Nature Methods* **19**, 687–695 (2022).
27. 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).
28. 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).
29. Gao, S. *et al.* Centromere Landscapes Resolved from Hundreds of Human Genomes. *Genomics Proteomics Bioinformatics* **22**, qzae071 (2024).
30. 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).
31. Altemose, N. *et al.* Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178.
32. Logsdon, G. A. *et al.* The variation and evolution of complete human centromeres. *Nature* **629**, 136–145 (2024).
33. Šošić, M. & Šikić, M. Edlib: a C/C ++ library for fast, exact sequence alignment using edit distance. *Bioinformatics* **33**, 1394–1395 (2017).
34. Hallast, P. *et al.* Assembly of 43 human Y chromosomes reveals extensive complexity and variation. *Nature* **621**, 355–364 (2023).
35. Collins, R. L. *et al.* A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).

1002 36. Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of
1003 structural variation. *Science* **372**, eabf7117 (2021).
1004 37. Wang, X. *et al.* The Health for Life in Singapore (HELIOS) Study: delivering Precision
1005 Medicine research for Asian populations. *medRxiv* 2024.05. 14.24307259 (2024).
1006 38. Wagner, J. *et al.* Curated variation benchmarks for challenging medically-relevant
1007 autosomal genes. *Nat Biotechnol* **40**, 672–680 (2022).
1008 39. Miller, D. T. *et al.* ACMG SF v3.2 list for reporting of secondary findings in clinical exome
1009 and genome sequencing: A policy statement of the American College of Medical Genetics
1010 and Genomics (ACMG). *Genet Med* **25**, 100866 (2023).
1011 40. Butler, M. G. Prader-Willi Syndrome: Obesity due to Genomic Imprinting. *Curr Genomics* **12**,
1012 204–215 (2011).
1013 41. Buiting, K. *et al.* Epimutations in Prader-Willi and Angelman Syndromes: A Molecular Study
1014 of 136 Patients with an Imprinting Defect. *Am J Hum Genet* **72**, 571–577 (2003).
1015 42. Bieth, E. *et al.* Highly restricted deletion of the SNORD116 region is implicated in Prader–
1016 Willi Syndrome. *European Journal of Human Genetics* **23**, 252–255 (2015).
1017 43. Meng, L. *et al.* Truncation of Ube3a-ATS unsilences paternal Ube3a and ameliorates
1018 behavioral defects in the Angelman syndrome mouse model. *PLoS genetics* **9**, e1004039
1019 (2013).
1020 44. Viljoen, D. & Ramesar, R. Evidence for paternal imprinting in familial Beckwith-Wiedemann
1021 syndrome. *Journal of medical genetics* **29**, 221–225 (1992).
1022 45. Chiesa, N. *et al.* The KCNQ1OT1 imprinting control region and non-coding RNA: new
1023 properties derived from the study of Beckwith–Wiedemann syndrome and Silver–Russell
1024 syndrome cases. *Human molecular genetics* **21**, 10–25 (2012).
1025 46. Ishida, M. New developments in Silver–Russell syndrome and implications for clinical
1026 practice. *Epigenomics* **8**, 563–580 (2016).
1027 47. Ioannides, Y., Lokulo-Sodipe, K., Mackay, D. J., Davies, J. H. & Temple, I. K. Temple

1028 syndrome: improving the recognition of an underdiagnosed chromosome 14 imprinting
1029 disorder: an analysis of 51 published cases. *Journal of medical genetics* **51**, 495–501
1030 (2014).
1031 48. Kilich, G. *et al.* Kagami Ogata syndrome: a small deletion refines critical region for
1032 imprinting. *NPJ Genomic Medicine* **9**, 5 (2024).
1033 49. Yorifuji, T., Higuchi, S., Hosokawa, Y. & Kawakita, R. Chromosome 6q24-related diabetes
1034 mellitus. *Clinical Pediatric Endocrinology* **27**, 59–65 (2018).
1035 50. Hanna, P., Francou, B., Delemer, B., Jüppner, H. & Linglart, A. A novel familial PHP1B
1036 variant with incomplete loss of methylation at GNAS-A/B and enhanced methylation at
1037 GNAS-AS2. *The Journal of Clinical Endocrinology & Metabolism* **106**, 2779–2787 (2021).
1038 51. Court, F. *et al.* Genome-wide parent-of-origin DNA methylation analysis reveals the
1039 intricacies of human imprinting and suggests a germline methylation-independent
1040 mechanism of establishment. *Genome Res* **24**, 554–569 (2014).
1041 52. Joshi, R. S. *et al.* DNA Methylation Profiling of Uniparental Disomy Subjects Provides a Map
1042 of Parental Epigenetic Bias in the Human Genome. *The American Journal of Human*
1043 *Genetics* **99**, 555–566 (2016).
1044 53. Hernandez Mora, J. R. *et al.* Characterization of parent-of-origin methylation using the
1045 Illumina Infinium MethylationEPIC array platform. *Epigenomics* **10**, 941–954 (2018).
1046 54. Zink, F. *et al.* Insights into imprinting from parent-of-origin phased methylomes and
1047 transcriptomes. *Nature genetics* **50**, 1542–1552 (2018).
1048 55. Akbari, V. *et al.* Genome-wide detection of imprinted differentially methylated regions using
1049 nanopore sequencing. *eLife* **11**, e77898.
1050 56. Akbari, V. *et al.* Parent-of-origin detection and chromosome-scale haplotyping using long-
1051 read DNA methylation sequencing and Strand-seq. *Cell Genomics* **3**, 100233 (2023).
1052 57. Rosenski, J. *et al.* Atlas of imprinted and allele-specific DNA methylation in the human body.
1053 *Nature Communications* **16**, 2141 (2025).

1054 58. Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools.
1055 *Briefings in Bioinformatics* **14**, 144–161 (2013).
1056 59. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics*
1057 **37**, 1639–1643 (2021).
1058 60. Chen, Y. *et al.* Context-aware transcript quantification from long-read RNA-seq data with
1059 Bambu. *Nature methods* **20**, 1187–1195 (2023).
1060 61. Sarashetti, P., Lipovac, J., Tomas, F., Šikić, M. & Liu, J. Evaluating data requirements for
1061 high-quality haplotype-resolved genomes for creating robust pangenome references.
1062 *Genome Biology* **25**, 312 (2024).
1063 62. Lawce, H. J. & Brown, M. G. Cytogenetics: an overview. *The AGT cytogenetics laboratory*
1064 *manual* 25–85 (2017).
1065 63. Dudchenko, O. *et al.* De novo assembly of the Aedes aegypti genome using Hi-C yields
1066 chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
1067 64. Stanojević, D., Lin, D., Nurk, S., Florez de Sessions, P. & Šikić, M. Telomere-to-Telomere
1068 Phased Genome Assembly Using HERRO-Corrected Simplex Nanopore Reads. *bioRxiv*
1069 2024.05.18.594796 (2024) doi:10.1101/2024.05.18.594796.
1070 65. Bankevich, A., Bzikadze, A. V., Kolmogorov, M., Antipov, D. & Pevzner, P. A. Multiplex de
1071 Bruijn graphs enable genome assembly from long, high-fidelity reads. *Nat Biotechnol* **40**,
1072 1075–1081 (2022).
1073 66. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV):
1074 high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*
1075 **14**, 178–192 (2013).
1076 67. Sarashetti, P. T2T Sequences.
1077 68. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly
1078 from long uncorrected reads. *Genome research* **27**, 737–746 (2017).
1079 69. Jain, C., Rhie, A., Hansen, N. F., Koren, S. & Phillippy, A. M. Long-read mapping to

1080 repetitive reference sequences using Winnowmap2. *Nature Methods* **19**, 705–710 (2022).
1081 70. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
1082 Preprint at https://doi.org/10.48550/arXiv.1303.3997 (2013).
1083 71. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–
1084 2079 (2009).
1085 72. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks.
1086 *Nature Biotechnology* **36**, 983–987 (2018).
1087 73. Shafin, K. *et al.* Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables
1088 high accuracy in nanopore long-reads. *Nature Methods* **18**, 1322–1332 (2021).
1089 74. Smolka, M. *et al.* Detection of mosaic and population-level structural variants with Sniffles2.
1090 *Nature Biotechnology* **42**, 1571–1580 (2024).
1091 75. Kirsche, M. *et al.* Jasmine and Iris: population-scale structural variant comparison and
1092 analysis. *Nature Methods* **20**, 408–417 (2023).
1093 76. Formenti, G. *et al.* Merfin: improved variant filtering, assembly evaluation and polishing via
1094 k-mer validation. *Nature Methods* **19**, 696–704 (2022).
1095 77. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008
1096 (2021).
1097 78. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic
1098 features. *Bioinformatics* **26**, 841–842 (2010).
1099 79. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–
1100 3100 (2018).
1101 80. Li, K., Xu, P., Wang, J., Yi, X. & Jiao, Y. Identification of errors in draft genome assemblies
1102 at single-nucleotide resolution for quality assessment and improvement. *Nature*
1103 *Communications* **14**, 6556 (2023).
1104 81. Chen, Y., Zhang, Y., Wang, A. Y., Gao, M. & Chong, Z. Accurate long-read de novo
1105 assembly evaluation with Inspector. *Genome Biology* **22**, 312 (2021).

1106 82. Chen, Q., Yang, C., Zhang, G. & Wu, D. GCI: a continuity inspector for complete genome
1107 assembly. *Bioinformatics* **40**, btae633 (2024).
1108 83. Durand, N. C. *et al.* Juicebox Provides a Visualization System for Hi-C Contact Maps with
1109 Unlimited Zoom. *Cell Systems* **3**, 99–101 (2016).
1110 84. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0. 2013–2015. (2015).
1111 85. Storer, J., Hubley, R., Rosen, J., Wheeler, T. J. & Smit, A. F. The Dfam community resource
1112 of transposable element families, sequence models, and genome annotations. *Mobile DNA*
1113 **12**, 2 (2021).
1114 86. Ryabov, F. HumAS-HMMER_for_AnVIL. (2024).
1115 87. Ryabov, F. stv. (2024).
1116 88. Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. SyRI: finding genomic rearrangements
1117 and local sequence differences from whole-genome assemblies. *Genome Biol* **20**, 277
1118 (2019).
1119 89. Heller, D. & Vingron, M. SVIM-asm: structural variant detection from haploid and diploid
1120 genome assemblies. *Bioinformatics* **36**, 5519–5521 (2020).
1121 90. Broad Institute. Picard Tools.
1122 91. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing
1123 next-generation DNA sequencing data. *Genome Res* **20**, 1297–1303 (2010).
1124 92. Zheng, Z. *et al.* Symphonizing pileup and full-alignment for deep learning-based long-read
1125 variant calling. *Nature Computational Science* **2**, 797–803 (2022).
1126 93. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and
1127 cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
1128 94. Zhao, H. *et al.* CrossMap: a versatile tool for coordinate conversion between genome
1129 assemblies. *Bioinformatics* **30**, 1006–1007 (2014).
1130 95. Chen, N.-C. & Vollger, M. nfLO.
1131 96. Li, H. *et al.* A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat*

1132 *Methods* **15**, 595–597 (2018).
1133 97. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: A Discriminative
1134 Modeling Approach for Rapid and Robust Local-Ancestry Inference. *Am J Hum Genet* **93**,
1135 278–288 (2013).
1136 98. Yun, T. *et al.* Accurate, scalable cohort variant calls using DeepVariant and GLnexus.
1137 *Bioinformatics* **36**, 5582–5589 (2020).
1138 99. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based
1139 linkage analyses. *The American journal of human genetics* **81**, 559–575 (2007).
1140 100. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol*
1141 **5**, R12 (2004).
1142 101. Ralf, A., Montiel González, D., Zhong, K. & Kayser, M. Yleaf: Software for Human Y-
1143 Chromosomal Haplogroup Inference from Next-Generation Sequencing Data. *Molecular*
1144 *Biology and Evolution* **35**, 1291–1294 (2018).
1145 102. Yfull. Haplogroup YTree v10.01.00. https://www.yfull.com/arch-10.01/tree/ (2022).
1146 103. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer
1147 weighting and repeat separation. *Genome research* **27**, 722–736 (2017).
1148 104. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nature*
1149 *methods* **9**, 676–682 (2012).
1150 105. Akalin, A. *et al.* methylKit: a comprehensive R package for the analysis of genome-wide
1151 DNA methylation profiles. *Genome Biol* **13**, R87 (2012).

1152 Acknowledgement

1153 We want to thank the anonymous volunteers who contributed their blood samples to this
1154 study. We would also like to thank the A*STAR Microscopy Platform (AMP) and the AMP
1155 staff, especially Dr Goh Wah Ing, for the consultation, advice, and technical support. AMP
1156 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, Ying Chen, Fei Yao, YY Sia, Dehui Lin, Xinyi Zhang, Daniel Muliaditan, Hwee Meng Low, See Ting Leong, Chiea Chuen Khor, Eleanor Wong, Patrick Tan, Jonathan Göke, 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

1183 Population and Global Health, Lee Kong Chian School of Medicine, Nanyang
1184 Technological University, 308232, Singapore
1185 John C Chambers

1186 Department of Hematology, Singapore General Hospital, Singapore 169856, Singapore
1187 William Y K Hwang

1188 Cancer & Stem Cell Biology, Duke-NUS Medical School, Singapore, Singapore
1189 William Y K Hwang

1190 Department of Hematology, National Cancer Centre Singapore, Singapore 168583,
1191 Singapore
1192 William Y K Hwang

1193 Regenerative Medicine Institute of Singapore, SingHealth Duke-NUS Academic Medical
1194 Centre, Singapore 169857, Singapore
1195 William Y K Hwang

1196 SingHealth Duke-NUS Institute of Precision Medicine, 169609, Singapore
1197 Patrick Tan

1198 Cancer & Stem Cell Biology Program, Duke-NUS Medical School, 169857, Singapore
1199 Patrick Tan

1200 Department of Statistics and Applied Probability, National University of Singapore,
1201 117546, Singapore

1202 Jonathan Göke

1203 Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117597,
1204 Singapore
1205 Jianjun Liu, Qi Jia

1206 ## 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. and Y.C. 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., Y.C., M.Š., J.G., 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-S30

Additional File2 - Supplementary tables S1-S33

Additional methods - Detailed methodology