Genomics Proteomics Bioinformatics 21 (2023) 1085–1100

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

# T2T-YAO: A Telomere-to-telomere Assembled Diploid Reference Genome for Han Chinese

**Yukun He [1,#], Yanan Chu [2,#], Shuming Guo [3,4,#], Jiang Hu [5,#], Ran Li [1,6,#], Yali Zheng [7,#], Xinqian Ma [1,6], Zhenglin Du [8], Lili Zhao [1,6], Wenyi Yu [1,6], Jianbo Xue [1,6], Wenjie Bian [1,6], Feifei Yang [1,6], Xi Chen [1,6], Pingan Zhang [1,6], Rihan Wu [1,6], Yifan Ma [1,6], Changjun Shao [2], Jing Chen [2], Jian Wang [2], Jiwei Li [7], Jing Wu [7], Xiaoyi Hu [7], Qiuyue Long [7], Mingzheng Jiang [7], Hongli Ye [7], Shixu Song [7], Guangyao Li [3], Yue Wei [3], Yu Xu [9], Yanliang Ma [1,6], Yanwen Chen [1,6], Keqiang Wang [1,6], Jing Bao [1,6], Wen Xi [1,6], Fang Wang [1,6], Wentao Ni [1,6], Moqin Zhang [1,6], Yan Yu [1,6], Shengnan Li [1,6], Yu Kang [2,10,*], Zhancheng Gao [1,4,6,*]**

[1] *Department of Respiratory and Critical Care Medicine, Peking University People's Hospital, Beijing 100044, China*
[2] *CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation, Beijing 100101, China*
[3] *Linfen Clinical Medicine Research Center, Linfen 041000, China*
[4] *Institute of Chest and Lung Diseases, Shanxi Medical University, Taiyuan 030001, China*
[5] *GrandOmics Biosciences Co., Ltd, Wuhan 430076, China*
[6] *Beijing Key Laboratory of Genome and Precision Medicine Technologies, Beijing 100101, China*
[7] *Department of Respiratory, Critical Care and Sleep Medicine, Xiang'an Hospital of Xiamen University, School of Medicine, Xiamen University, Xiamen 361101, China*
[8] *Institute of PSI Genomics, Wenzhou 325024, China*
[9] *Beijing Jishuitan Hospital, Capital Medical University, Beijing 100035, China*
[10] *University of Chinese Academy of Sciences, Beijing 100490, China*

Received 18 July 2023; revised 1 August 2023; accepted 8 August 2023
Available online 16 August 2023

Handled by Yu Jiang

**KEYWORDS**

Reference genome;

**Abstract** Since its initial release in 2001, the human reference genome has undergone continuous improvement in quality, and the recently released telomere-to-telomere (T2T) version —

* Corresponding authors.
E-mail: zcgao@bjmu.edu.cn (Gao Z), kangy@big.ac.cn (Kang Y).
# Equal contribution.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2023.08.001

Telomere-to-telomere assembly;
Han Chinese;
Haplotype-resolved assembly;
Diploid

T2T-CHM13 — reaches its highest level of continuity and accuracy after 20 years of effort by working on a simplified, nearly homozygous genome of a hydatidiform mole cell line. Here, to provide an authentic complete **diploid** human genome reference for the **Han Chinese**, the largest population in the world, we assembled the genome of a male Han Chinese individual, T2T-YAO, which includes T2T assemblies of all the 22 + X + M and 22 + Y chromosomes in both haploids. The quality of T2T-YAO is much better than those of all currently available diploid assemblies, and its haploid version, T2T-YAO-hp, generated by selecting the better assembly for each autosome, reaches the top quality of fewer than one error per 29.5 Mb, even higher than that of T2T-CHM13. Derived from an individual living in the aboriginal region of the Han population, T2T-YAO shows clear ancestry and potential genetic continuity from the ancient ancestors. Each haplotype of T2T-YAO possesses ~ 330-Mb exclusive sequences, ~ 3100 unique genes, and tens of thousands of nucleotide and structural variations as compared with CHM13, highlighting the necessity of a population-stratified reference genome. The construction of T2T-YAO, an accurate and authentic representative of the Chinese population, would enable precise delineation of genomic variations and advance our understandings in the hereditability of diseases and phenotypes, especially within the context of the unique variations of the Chinese population.

## Introduction

A complete and accurate reference genome has been a long-standing goal in the biomedical research community since the initiation of the Human Genome Project (HGP) three decades ago. However, the limitations of sequencing technology have made it challenging to achieve this level of completeness and accuracy [1–3]. Recently, a groundbreaking publication from the Telomere-to-Telomere (T2T) Consortium described the first-ever complete haploid human genome of exceptional quality, known as T2T-CHM13v1.1. This success, attributed to significant improvements in sequencing technology and supporting bioinformatics tools, has fulfilled 8% of the previously unknown highly repetitive region in the human genome [4] and achieved a high quality of Q73.94, which means one error per 24.8 Mb [5]. The resulting T2T-CHM13, the first-ever complete reference for genetic research and various omics studies, allows for more accurate localization of variation signals, especially in regions of repeats and duplications [6–11].

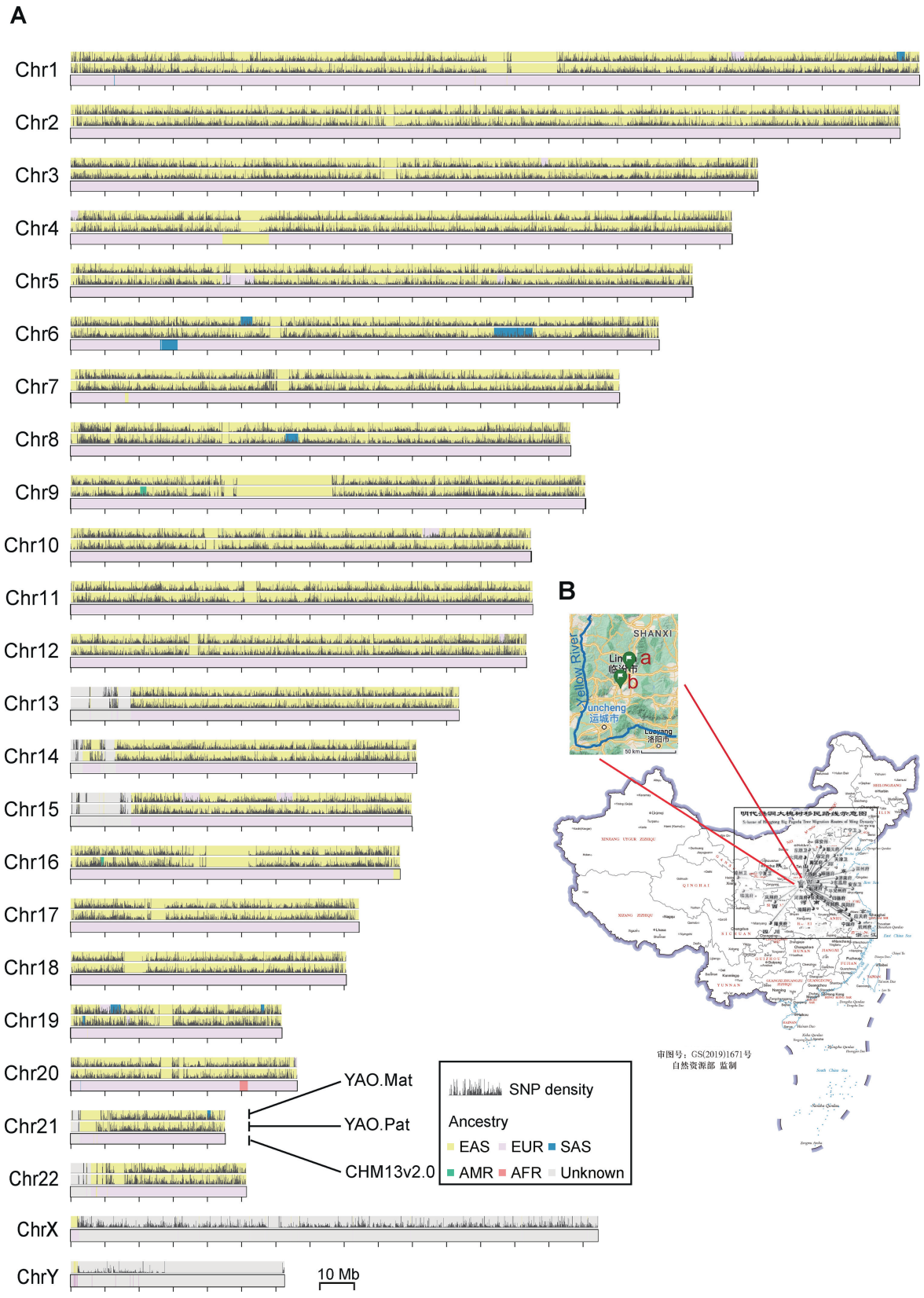
T2T-CHM13, while a remarkable scientific achievement, is not a representative genome of a real human individual, but a haploid genome of the complete hydatidiform mole (CHM) lacking the Y chromosome [12]. The CHM13 cell line is of Northern European origin, and even by adding the complementary Y chromosome of HG002, which is of Eastern European Ashkenazi Jewish ancestry [13], the T2T-CHM13v2.0 is not enough for representing all individuals worldwide, as emphasized by the initial release of the Human Pangenome Reference Consortium (HPRC), which includes draft genomes from 47 individuals worldwide [13]. The draft genomes in the pangenome study, although not completed, reveal millions of single-nucleotide polymorphisms (SNPs), small indels, and tens of thousands of structural variations (SVs) per haplotype, with at least 100 Mb of sequences (~ 3%) not represented in CHM13 [13]. Similar findings have also been reported in a recent pangenome study of Chinese populations [14], highlighting the importance of creating distinct reference genomes for each major population. Population-stratified references would allow for a more comprehensive understanding of genome variations across different populations for in-depth biological research and medical applications. However, assembling a high-quality diploid T2T reference genome for an individual has yet to be achieved, even for the extensively sequenced and assembled diploid genome HG002 despite the significant advancements in sequencing technology [15].

The Han Chinese is the largest population in the world with billions of descendants worldwide, and yet remains underrepresented in current human reference genomes, including GRCh38 and the HPRC, especially lack of samples from their aboriginal regions. Despite extensive efforts being made to create a reference genome for the Han population, including NH1.0 [16], HJ [17], Han1 [18], and CN1 [19], technical limitations make this task difficult for an accurate and complete one. Here, we constructed a diploid T2T reference genome with a Y chromosome for the Han population (**Figure 1**A). Our sample comes from a healthy young man living in an ancient village in Shanxi Province that has been inhabited by the Han population for tens of generations, as recorded in their family genealogy. We named this reference genome "T2T-YAO" after the sampling point, which is located near the ruin capital of the Emperor YAO from thousands of years ago. This area is also significant as it marks the starting point of the Hongtong migration during the Ming Dynasty. This migration, which lasted for nearly half a century (AD 1370–1417), saw a large amount of emigrants throughout China and into Southeast Asia. Thus, the T2T-YAO genome is expected to provide a comprehensive representation of the Han population, and the top quality of Q74.69 that it achieves, which means one error per 29.5 Mb, makes it a truly accurate and authentic reference genome of the Chinese population and enables its applications in future medical research and clinical practice for this vast population group.

## Results

### Sample and sequencing

The goal of this work is to construct a complete and accurate reference genome for the Han Chinese population. Unlike hydatidiform mole cell lines, which have unstable genomes and a high chance of accumulating mutations and translocations during construction and passages, as revealed by sequencing data in CHM13 [20], our study specifically collected a peripheral blood mononuclear cell (PBMC) sample from a real person. This approach ensures greater accuracy


A
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
Chr13
Chr14
Chr15
Chr16
Chr17
Chr18
Chr19
Chr20
Chr21
Chr22
ChrX
ChrY
YAO.Mat
YAO.Pat
CHM13v2.0
10 Mb
SNP density
Ancestry
EAS EUR SAS
AMR AFR Unknown
B
SHANXI
Yellow River
Linyi
临汾市
a
b
Yuncheng
运城市
Luoyang
洛阳市
50 km
审图号：GS(2019)1671号
自然资源部 监制

and reliability in the generated reference genome. Furthermore, to achieve a more representative reference genome, this study specifically collected samples from Hongtong County, which is the starting point of the latest country-wide mass migration in China (Figure 1B). This migration, whose destinations had involved 28.5% of counties in Ming territory, has had a significant impact on the population distribution of their descendants throughout China and beyond, making Hongtong an ideal location for sampling. Additionally, the proximity of Hongtong to the ruins of the capital city where the legendary Emperor YAO lived over 4200 years ago (Figure 1B) also adds to its significance as a location for sampling. After this, we named the new reference genome "T2T-YAO".

To construct the T2T diploid genome, we conducted a trio-based assembly by collecting multiple PBMC samples from the trio (son and both parents). Before sequencing, we conducted a karyotyping test with 400 bands on all three samples and ruled out any chromosomal disease in the family (Figure S1). This was followed by sequencing using multiple technologies, including 92× high-fidelity (HiFi) sequencing from PacBio [21,22], 336× Oxford Nanopore Technologies (ONT) ultra-long read sequencing [23] with a subset of super-long reads above 250 kb (42×) included, 584× Illumina Arima Genomics high-through chromosome conformation capture (Hi-C) sequencing [24], BioNano optical maps [25], and DNBSEQ-T7 150 bp for the son and parents (with 278× and ~ 116× coverage, respectively). All the aforementioned sequencing coverage was calibrated in terms of an average haploid human genome, which is 3 Gb. In pursuing high continuity, we specifically selected long DNA fragments greater than 100 kb for the library construction and applied the seemingly unrelated Poisson (SUP) model for the base-calling in the ONT sequencing [26]. The final ONT sequencing data achieved an N50 of 128 kb with a median base-calling accuracy of 98% for single reads.

### Diploid assembly, polishing, and assessment

To construct the T2T-YAO diploid genome, we followed a similar strategy to the assembly of CHM13 but phased the filial ONT reads according to the haplotype-specific markers identified by the HiSeq sequencing data of each parent. Then we constructed a haploid-resolved de Bruijn graph using HiFi reads [27] and simplified the assembly graph by iteratively integrating the phased ONT reads into it [28]. The automated pipeline, Verkko (https://github.com/marbl/verkko), has integrated these steps and assembled the sequencing reads of YAO into a diploid draft genome [maternal (Mat), 22 + X + M; paternal (Pat), 22 + Y], leaving only 90 gaps within the 46 chromosomes. Most of the gaps were due to large scale repeats, particularly in the centromeres, the Yq12 region, and the short arms of acrocentric chromosomes (SAACs) 13, 14, 15, 21, and 22. Next, using only super-long ONT reads (> 250 kb) and HiFi reads with unique or low-frequency *k*-mers, we successfully closed all the gaps except the ribosomal DNA (rDNA) clusters by extending the overlapping HiFi reads guided by phased ONT reads from the boundary of the gaps. The rDNA clusters in SAACs were finalized by following the strategy used in T2T-CHM13, except in a haplotype-phased manner. Finally, we validated the telomere sequences at the ends of all 46 chromosomes and confirmed no intervening model sequences or gaps, achieving the T2T assembly of YAO, which includes the complete sequences of 44 autosomes, chromosomes X and Y, as well as the mitochondria.

To improve the quality of the T2T-YAO genome, we conducted several major steps (Figure S2) modified from the polishing process for CHM13 [5]. First, Winnowmap2 [29] was used to map HiFi and ONT reads to the T2T-YAO genome in a repeat-sensitive manner. Next, pipelines of NextPolish2 [30], DeepVariant [31], and PEPPER-Margin-DeepVariant [32] were used to call single-nucleotide variant (SNV)-like errors in a haplotype-sensitive manner, and Sniffles2 [33] was used to call SV-like errors. Finally, SNV-like and SV-like errors were respectively filtered by Merfin [34] and Jasmine [35], and manually checked. This process identified a total of 32,861 and 29,087 SNV-like errors, 8 and 7 SV-like errors in the Mat and Pat haploids, respectively, which were corrected by BCFtools [36,37]. BioNano and Hi-C data were also used to identify potential SV-like errors, and only 28 discordant variations supported by ONT and HiFi reads were corrected manually. Ultimately, we obtained a fully-resolved diploid genome, T2T-YAO, with its total length of Mat and Pat haploids of 3.02 Gb and 2.92 Gb, respectively. The size of both T2T-YAO haploids is comparable to that of T2T-CHM13, and the difference between the two haploids is mainly due to the different length of sexual chromosomes (Table S1).

The highly polished T2T-YAO was then evaluated by Merqury [38], a *k*-mer-based reference-free assessment pipeline, using both HiSeq and HiFi reads to determine its completeness, assembly errors (**Figure 2**A), and switching errors between haplotypes (Figure 2B) following the same process and parameters used in evaluating T2T-CHM13 [5]. T2T-YAO reaches the reference-quality accuracy with its quality value (QV; a log-scaled probability of error for the consensus base calls) as Q70.49 for Mat haplotype and Q72.28 for Pat

**Figure 1 Overview of T2T-YAO and its ancestry**

**A.** Comparison of the ancestry markers of the T2T-YAO and CHM13v2.0 reference genomes. Each set of chromosomes includes a maternal haplotype of YAO (YAO.Mat, top), a paternal haplotype of YAO (YAO.Pat, middle), and CHM13 (bottom). All feature regions are lifted over to CHM13, and mitochondrial DNA is not shown due to its short length relative to the large scale. Chromosomal regions are colored according to their ancestry prediction (EAS, EUR, SAS, AMR, and African) and overlaid with the density of SNPs against CHM13 indicated by black vertical bars. **B.** Map of the sampling site and the routes of Hongtong migration in the Ming Dynasty. Insert shows the sampling site of T2T-YAO (a) and the adjacent ruins of Emperor YAO's capital (b). T2T, telomere-to-telomere; CHM, complete hydatidiform mole; Chr, chromosome; SNP, single-nucleotide polymorphism; Mat, maternal; Pat, paternal; EAS, East Asian; EUR, European; SAS, South Asian; AMR, admixed American; AFR, African.

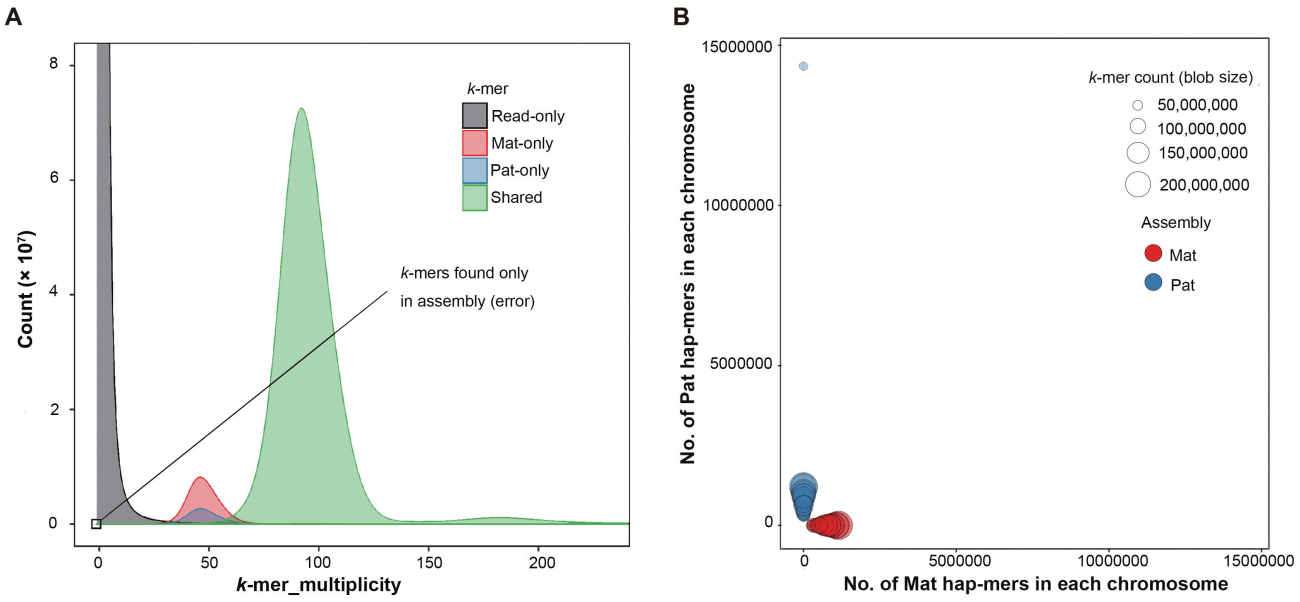



**Figure 2 The quality of the diploid T2T-YAO assembly**

**A.** The spectrum plots of *k*-mers in assembly and reads for the evaluation of assembly error and completeness. The distributions of *k*-mers present in the reads and Mat/Pat assemblies are separately colored. The X-axis represents the *k*-mer_multiplicity, which means the number of times a *k*-mer occurs in the read set. The *k*-mers at the left end of the X-axis with zero multiplicity in reads are erroneous ones found only in assemblies. Read-only means *k*-mers found only in reads; Shared means *k*-mers shared by Mat/Pat assemblies. **B.** Hap-mer blob plot of the T2T-YAO assembly. Mat and Pat assemblies are represented in red and blue, respectively. The numbers of Mat and Pat hap-mers in each chromosome are plotted as pink and blue blobs, respectively, and the blob size is proportional to the number of total *k*-mers of each chromosome.

**Table 1 Comparison of the assembling quality among high-quality human genome assemblies**

| Statistic | YAO.Mat | YAO.Pat | HG002.Mat | HG002.Pat | CN1.Mat | CN1.Pat | YAO-hp | CN1 | T2T-CHM13v1.1 |
|---|---|---|---|---|---|---|---|---|---|
| Assembly size (Gb) | 3.02 | 2.92 | 3.06 | 2.96 | 3.04 | 2.94 | 3.06 | 3.09 | 3.05 |
| No. of gaps | 0 | 0 | 109 | 117 | 0 | 0 | 0 | 0 | 0 |
| Length of gaps (Mb) | 0 | 0 | 16.4 | 25.3 | 0 | 0 | 0 | 0 | 0 |
| Contig N50 (Mb) | 155.06 | 145.07 | 62.88 | 81.56 | INA | INA | 155.06 | INA | 154.26 |
| QV | 70.49 | 72.28 | 59.3 | 58.6 | 60.11 | 59.37 | 74.69 | 63.35 | 73.94 |
| Chromosome composition | 22 + X | 22 + Y | 22 + X | 22 + Y | 22 + X | 22 + Y | 22 + XY | 22 + XY | 22 + X |
| Completeness (%; haploid) | 99.65 | 99.59 | 99.7 | 99 | INA | INA | – | – | – |
| False duplication (%) | 0.3 | 0.2 | 0.49 | 0.37 | INA | INA | – | – | 0.15 |
| Haplotype switch error (%) | 0.0192 | 0.0113 | 0.028 | 0.018 | INA | INA | – | – | – |
| Length of collapsed regions with common repeats | | | | | | | | | |
| Collapsed regions (Mb) | 4.71 | 4.96 | 18.53 | 17.64 | INA | INA | – | – | – |
| Expanded regions (Mb) | 19.13 | 21.57 | 51.01 | 46.35 | INA | INA | – | – | – |
| Length of collapsed regions without common repeats | | | | | | | | | |
| Collapsed regions (Mb) | 0.56 | 0.35 | 7.61 | 7.81 | INA | INA | – | – | – |
| Expanded regions (Mb) | 1.91 | 2.92 | 15.52 | 15.88 | INA | INA | – | – | – |
| rDNA copy number | 79 | 149 | INA | INA | 132 | 117 | 98 | 182 | 219 |

*Note*: The data for HG002.Mat and HG002.Pat were obtained from [15]; the data for T2T-CHM13v1.1 were obtained from [28]. INA, information not available; –, not applicable; QV, quality value; Mat, maternal; Pat, paternal; Chr, chromosome; hp, haploid; T2T, telomere-to-telomere; CHM, complete hydatidiform mole; rDNA, ribosomal DNA.

haplotype (**Table 1**). The quality is much higher than those of HG002 [15] and the recently published CN1 (the up-to-date highest-quality diploid human genome) [19] (Table 1). T2T-YAO achieved high quality, even after filling all the gaps (a process often introduces many assembly errors). The QV scores of more than one-third of the chromosomes in T2T-YAO are greater than 75, and the QV scores of Chr18 in Mat and Chr3, Chr6, and Chr8 in Pat even reach infinity,

which means no assembly error (Table S1). More importantly, by merging the diploid T2T-YAO into a haploid reference through selecting the better assembly for each autosome, we generated T2T-YAO-hp, whose QV reaches Q74.69. The quality is even higher than that of T2T-CHM13v1.1 (Q73.94) in parallel comparison [5], placing T2T-YAO-hp the top-quality haplotype reference of the human genome in the world (Table 1, Table S1).

In addition to QV, we also evaluated other key quality metrics for diploid assemblies, including: (1) switch errors, which describe incorrectly switched *k*-mers between Mat and Pat haplotypes; (2) false duplications, which are estimated by the *k*-mers with incorrect higher copy numbers in assembly; and (3) collapsed repeats, which have additional copies in the actual genome that have been ''collapsed'' into a single copy. The T2T-YAO exhibited ~ 40% less false duplications, one-third less switch errors, and about 20-fold shorter collapsed regions than HG002 (Table 1), although information regarding these metrics is not available for CN1 for a direct comparison. Unlike CN1, in which the rDNA sequences were filled by rDNA units with estimated copies [19], T2T-YAO boasts a more comprehensive rDNA sequence, which illustrates the heterogeneity of rDNA copy numbers among haplotypes. The copy numbers of several important multicopy genes in T2T-YAO, including the rDNA cluster, and the lengths of satellite regions in the X chromosome were estimated by performing droplet digital polymerase chain reaction (ddPCR) in comparison to *in silico* PCR (Table S2). All the assessments of T2T-YAO ensure that it is highly qualified as a reference genome.

## The ancestry markers

The local ancestry inference based on random forest models [39] trained by the SNPs called from the 1000 Genomes Project (1KGP) dataset [40] demonstrates that the majority of genome YAO is of East Asia origin and admixed with sporadic predicted markers of South Asia, Europe, and America (Figure 1A). The positions of the ancestry marker admixture are not consistent between the two haplotypes due to the inter-individual diversity in SNP profiles in the Han population. The markers of South Asia are a little more than those of Europe and America in both haplotypes, suggesting more genetic exchange between the two Asia ethnic groups than with groups in other continents, which is consistent with a previously reported study [41]. When inferred with the same random forest model, the ancestry of CHM13 is mainly of European origin, admixed with a few predicted markers of East Asia, South Asia, and Africa.

We further identified the Y haplogroup of YAO using yHaplo [42], which is a tool to build a tree of phylogenetically significant SNPs accumulated in the non-recombining portion of the male-specific region (MSY) [43], and compared it to the Y-DNA haplogroup tree in the International Society of Genetic Genealogy (ISOGG) database [44]. The Y haplogroup of YAO is identified as O-F2137 (O2a2b1a1a1a2a), which is one of the main descendant groups of O-M122, the predominant Y haplogroup in China and Asia [45,46], and is consistent with the established ancestry of YAO. Interestingly, the O-F2137 haplogroup is also identified in one of the ancient DNA samples in the Shimao Shengedaliang, a Neolithic site in Shenmu County, Shaanxi Province [47]. The site is dated to ~ 2000 BC, a contemporaneous site close to the ruin capital of Emperor YAO and near the village where we sampled. This finding suggests a potential genetic continuity in the region dating back to the earliest days of human habitation in this part of China.

## Exclusive sequences and genes

We performed pairwise alignment among the three haplotypes (Mat and Pat in YAO and CHM13) using MUMmer [48] to get a full view of the sequence similarity among them. The dot plots of the alignments exhibit a two-phase pattern for all pairings in terms of the length and identity of alignment: one is of perfect alignments that are longer than 50 kb with high identity (mostly > 99.5%); the other is of poor alignments that are less than 50 kb with low identity (90%–99.5%) (Figure S3). The perfect alignments indicate the orthologous regions that constitute most of the chromosomes, and the rest 280–350-Mb sequences (~ 10%) in each haplotype are not or poorly aligned to others, which mainly locate in the regions of heterochromosomes, such as centromeres and the Yq12 region (**Figure 3**A). Heterochromosome regions, comprised of highly repetitive sequences, are more instable during chromosomal duplication and diverse among individuals than the other parts of the genome, which has been found to contribute to aging, neurodegeneration, and other disorders [9,49,50].

It is notable that the total lengths in Mat and Pat haploids of poorly aligned sequences to CHM13 in the 22 autosomes are 326.9 Mb and 332.6 Mb, respectively, but reduce to 299.4 Mb and 306.3 Mb of them when aligned to each other, indicating better alignment between Mat and Pat of YAO than that to CHM13 (Table S3). Furthermore, in the perfect alignments longer than 50 kb, the weighted average identity between the two haplotypes of YAO is 99.94%, higher than that of 99.83% between YAO and CHM13, suggesting more nucleotide-level variations between YAO and CHM13 (Figure 3A). We then investigated the sequence similarity in the centromere regions with poor alignments among the haplotypes. The result showed longer alignments and higher similarity in the Mat *vs.* Pat pairing than in the YAO *vs.* CHM13 pairings (Figure S4). All these alignment results indicate that ~ 10% of sequences in each haplotype are of unique and represents most of the inter-individual genome diversity. The haploids of YAO share more sequences and possess a higher identity than comparisons to CHM13, implying a greater genomic distance between the ethnic groups compared with that within the Han population.

We next annotated the T2T-YAO genome for both repeat and non-repeat portions. RepeatMasker and BISER were first utilized to infer repeat sequences and segmental duplications (SDs), respectively. The results showed a similar proportion in all categories of repeat sequences and a comparable amount of SDs to CHM13 (Table S4). For gene annotations in the non-repeat regions, Comparative Annotation Toolkit (CAT) [51] and Liftoff [52] were applied to project the GENCODE v43 annotation of reference GRCh38 [53] onto the T2T-YAO assembly. By merging the annotations obtained from both tools, a total of 64,131 and 62,284 genes in Mat and Pat haploids, respectively, are identified in the T2T-YAO gen-

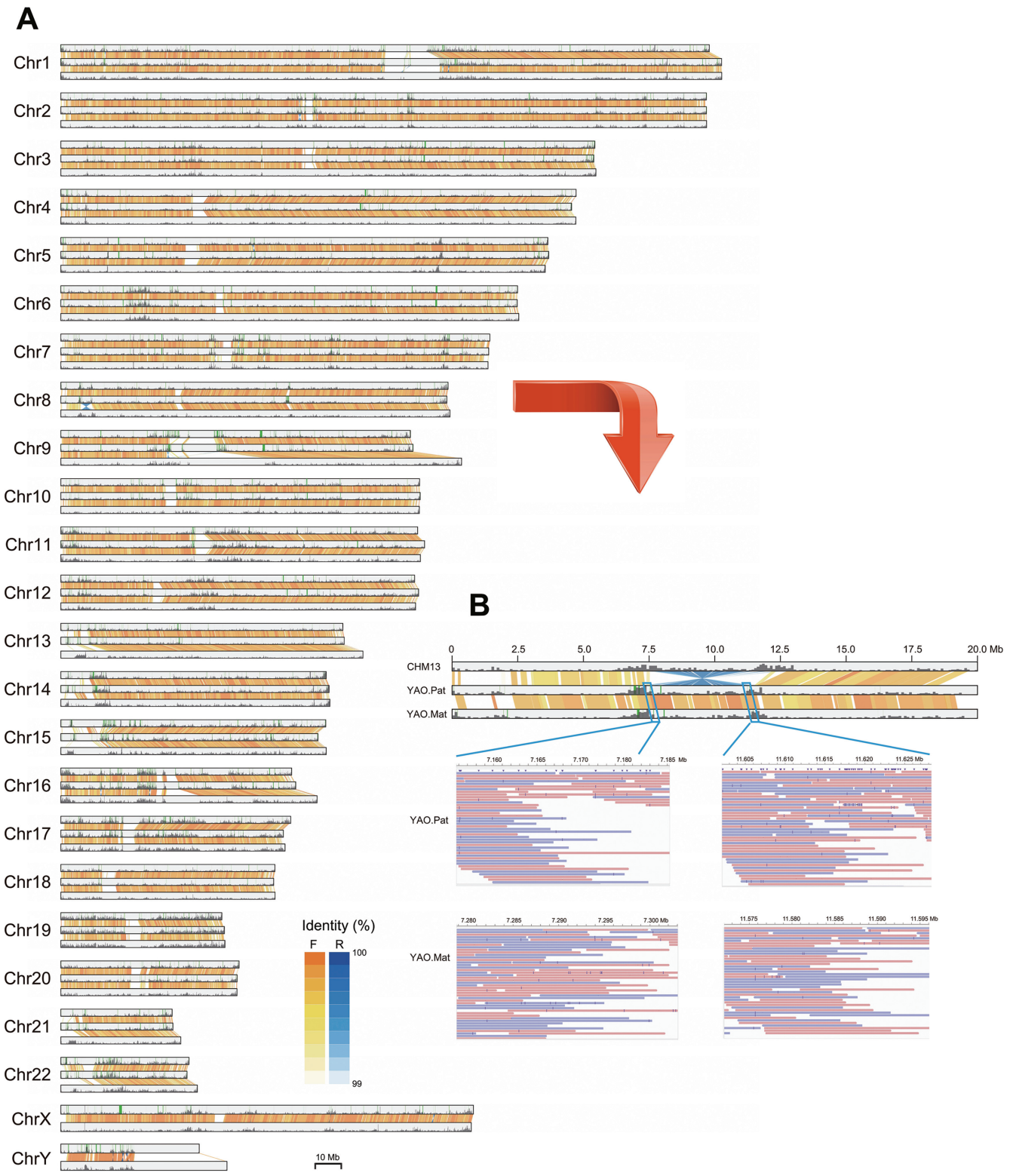



**Figure 3 The overview of genomic variations between YAO and CHM13**

**A.** The alignments between the chromosomes of the YAO.Mat, YAO.Pat, and CHM13 v2.0 haplotypes. Each set of chromosomes includes a maternal haplotype of YAO (YAO.Mat, top), a paternal haplotype of YAO (YAO.Pat, middle), and CHM13 (bottom). All chromosomes are plotted in scale to their original length, and alignments with length > 100 kb and identity > 99% are linked by lines between the chromosomes. The orange and blue colors and their depth of the lines indicate the direction and identity of the alignment as displayed in the legend. The black vertical bars that overlaid on chromosomes indicate the density of genes with exclusive genes labeled by green lines. **B.** The huge inversion in the distal region (0–25 Mb) of Chr8 short arm of YAO. The upper panel is the alignment map showing the large inversion in both Mat and Pat haploids of YAO compared with CHM13. The corresponding region in CHM13 is aligned at the top of the map for comparison. The bottom panel shows the HiFi reads mapped to the break points in the Mat and Pat assemblies of YAO. F, forward; R, reverse; HiFi, high-fidelity.

**Table 2 Comparison of the gene annotations between T2T-YAO and CHM13v2.0**

| | YAO.Mat | YAO.Pat | T2T-CHM13v2.0 |
|---|---|---|---|
| No. of genes | 64,131 | 62,284 | 64,213 |
| No. of protein-coding genes | 20,045 | 19,309 | 20,067 |
| No. of exclusive genes (YAO *vs.* CHM13) | 3193 | 3083 | 2529 |
| No. of exclusive protein-coding genes (YAO *vs.* CHM13) | 234 | 227 | 175 |
| No. of transcripts | 253,892 | 247,035 | 233,615 |
| No. of protein-coding transcripts | 89,223 | 86,167 | 86,245 |
| No. of exclusive transcripts (YAO *vs.* CHM13) | 28,283 | 27,770 | 8627 |
| No. of exclusive protein-coding transcripts (YAO *vs.* CHM13) | 8542 | 8209 | 5612 |

ome, similar to the number of genes found in CHM13 (**Table 2**). Either Mat or Pat haplotype has ~ 3100 exclusive genes other than CHM13, among which, 2646 are shared between Mat and Pat haploids (Figure S5). Furthermore, Mat and Pat haplotypes share a similar pattern for which their exclusive genes dispersed along chromosomes, with some hotspots in the peri-centromere regions of low identity to CHM13 (Figure 3A), suggesting that haplotypes of Han origin are much closer to each other than to CHM13 in terms of the genes that they harbor.

Among the 2646 exclusive genes that are shared by both haploids of YAO, 157 are predicted to be protein-coding (Table S5), whereas the rest are non-coding genes. Mat and Pat haploids also have a few hundred exclusive genes to each other, including < 10% protein-coding genes (Tables S6 and S7). According to the annotations of these exclusive protein-coding genes, most of them are pseudogenes, gained gene copies, or members of a large gene family, such as the six genes in the olfactory receptor family. Notably, 78 exclusive protein-coding genes are annotated as "novel protein" with unknown functions, and 76 out of them are single-copy ones. These novel proteins exclusive to YAO potentially highlight the unique features in the genomes of the Han population in terms of function and deserve further studies in the future. The aforementioned GRCh38-based annotation only paints a partial picture of the exclusive genes in T2T-YAO, but it illustrates the genome dissimilarity between the genomes of different ethnic groups in terms of the genes that they contain. Thorough exploration of the genome dissimilarity and distinct genes of the Han population requires further in-depth investigations in population genomics and biological function studies.

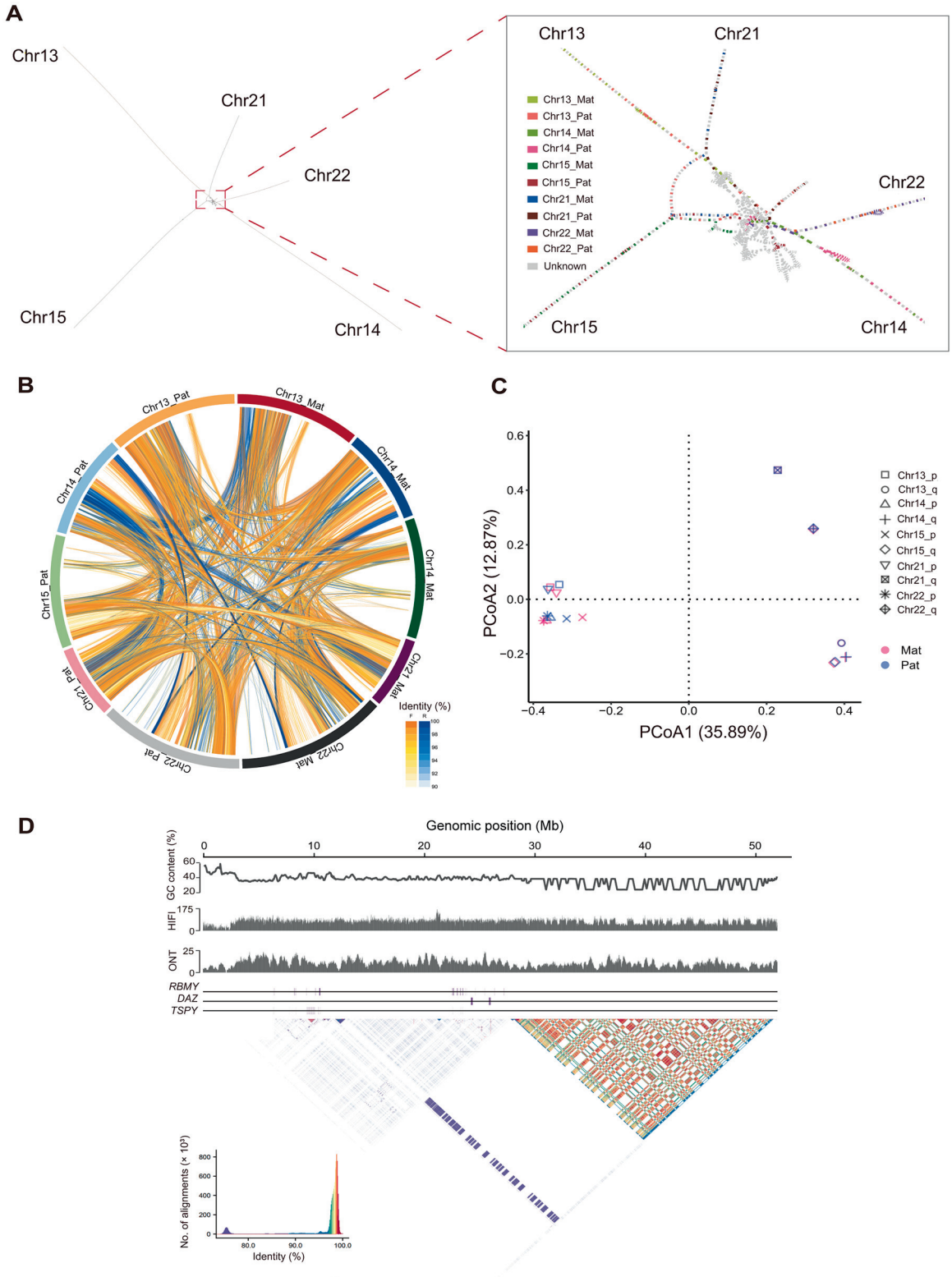
### Genome variations in homologous regions

We first compared the small variations, including SNVs and small indels (< 50 bp), in the orthologous sequences among the three haploids, for which we used GRCh38 as the reference to annotate the variations. A total of 3.03 Mb, 2.94 Mb, and 2.92 Mb of small variations were identified in YAO.Mat, YAO.Pat, and CHM13, respectively, largely comparable to those averagely identified in each haplotype in the pangenome studies [13,14]. When only considering the autosomes, the total length reduces to 2.92 Mb, 2.93 Mb, and 2.82 Mb, respectively, among which YAO.Mat and YAO.Pat share another 0.646 Mb of small variations in-between besides the 0.95 Mb shared by the three, but only share 0.381 Mb and 0.383 Mb with CHM13, respectively, suggesting a greater genome dissimilarity between Han and European populations (Figure S6). Most of the small variations locate in the intergenic regions and introns, and only ~ 14,000 small variations per haplotype, mostly SNPs, are in the coding sequence (CDS) regions of protein-coding genes. Among the SNVs in CDS regions, more than a hundred are detrimental nonsense or frameshift mutations, whose numbers are similar among the three haplotypes and within the range of such mutations identified in the recent pangenome study [13] (Table S8).

To investigate the overall genome dissimilarity between the three haplotypes, we further compared their *k*-mer composition. As the *k*-mer composition varies between repetitive and non-repeat sequences, we first separated the repeat and non-repeat regions for each haploid genome. Then we calculated a weighted dissimilarity matrix for the non-repeat section and an unweighted matrix for the repeat section. In the principal co-ordinates analysis (PCoA) plot, most homologous chromosomes are very close, indicating similar *k*-mer composition in each chromosome, except for Chr1, Chr2, and the acrocentric chromosomes. These chromosomes showed a greater distance between CHM13 and the YAO haplotypes in both repeat and non-repeat sections (Figure S7).

Next, the T2T assembly of YAO enables a more thorough analysis of the SVs (50–100,000 bp) among the three haplotypes in a pairwise manner. Using SVIM-asm [54], we inferred 22,555 and 24,355 SVs in Mat and Pat haploids of YAO, respectively, when compared with CHM13. Most of the SVs are indels in length of 50–300 bp, whereas the longer ones are sporadically distributed (Figure S8). SVs between Mat and Pat are 21,453, fewer than their comparison to CHM13, indicating closer genome similarity among the YAO's haplotypes. We further investigated the larger SVs (> 100 kb) and identified 91 (30.7 Mb) and 83 (35.1 Mb) SVs in Mat and Pat haploids, respectively, when compared with CHM13 using SyRI [55] (Table S9). A fraction of 87 (29.6 Mb) and 76 (31.0 Mb) SVs locate in the autosomes, much longer than the 70 SVs (18.2 Mb) between Mat *vs.* Pat comparison. The difference is mainly due to the larger inversions and translocations in YAO *vs.* CHM13 pairings than those between Mat *vs.* Pat, making their total length millions of bases longer.

The largest inversion that we identified, which spans 4 Mb in length, is located on the short arm of Chr8 (8p23) of both YAO haploids when compared with CHM13. Careful verification of the assemblies in both Mat and Pat haploids of YAO shows strong support from multiple coverages of HiFi and ONT reads in comparison to the break points in CHM13 where are entirely devoid of reads (Figure 3B, Figure S9). Interestingly, the 8p23 inversion has also been reported as a structural polymorphism in previous genetic research [56,57]


A
Chr13
Chr21
Chr22
Chr15
Chr14
Chr13_Mat
Chr13_Pat
Chr14_Mat
Chr14_Pat
Chr15_Mat
Chr15_Pat
Chr21_Mat
Chr21_Pat
Chr22_Mat
Chr22_Pat
Unknown
B
Identity (%)
C
PCoA1 (35.89%)
PCoA2 (12.87%)
Chr13_p
Chr13_q
Chr14_p
Chr14_q
Chr15_p
Chr15_q
Chr21_p
Chr21_q
Chr22_p
Chr22_q
Mat
Pat
D
Genomic position (Mb)
GC content (%)
HIFI
ONT
RBMY
DAZ
TSPY
No. of alignments (× 10³)
Identity (%)

and in the recently released high-quality genomes of Han Chinese individuals [18,19]. Near both sites of the break, we identified many exclusive member genes of the olfactory receptor family unique to CHM13; however, the inversion does not disrupt or fuse any genes, although it does switch the intervening genes to the opposite strand of the chromosome.

### Homology mosaics in the SAACs

The assembly of the SAACs is a challenging task due to their highly repetitive sequences and extensive recombination among homologous and heterologous acrocentric chromosomes, such as the Robertsonian translocation between Chr14 and Chr21 observed in cytogenetic studies [58]. The highly entangled contig components of SAACs in the assembly graph (**Figure 4**A) demonstrates the complexity of unraveling their complete sequences, which has not been fully resolved by all the previously published human diploid genomes. Relying on the deep coverage of our highly accurate ultra-long ONT reads and the clear phasing between Mat and Pat haplotypes, we are able to successfully complete all ten SAAC regions in T2T-YAO. Pairwise alignment of the complete SAACs revealed the presence of almost identical sequences across heterologous chromosomes, forming homology mosaics with a large amount of inversions, duplications, and translocations, especially among chromosomes 13, 14, 21, and 22 (Figure 4B), making them hotspots of chromosomal abnormalities usually observed in genetic disorders and tumors. The completion of the diploid assembly of T2T-YAO has provided direct sequence-based evidence for the homology mosaics nature of SAACs, explaining the mechanisms underlying the active recombination events in these regions [59]. We further investigated the *k*-mer composition of the ten SAAC regions, and their pairwise distance in the PCoA plot indicated the homogeny among them, which cluster together. In contrast, the long arms of homologous chromosomes show almost identical positions but are far from the heterogenous ones (Figure 4C), confirming the unique homology mosaics nature of SAACs.

Each SAAC in T2T-YAO contains a locus of clustered rDNA genes, arranged in large stretches of tandem repeats without any rDNA cluster outside these regions, like CHM13. These rDNA clusters are highly homogenized, except for a few variations in the intergenic non-coding regions, rendering them susceptible to copy loss during chromosomal duplication [60]. Being of paramount importance for cellular biology, the copy number of rDNA genes is tightly regulated to maintain homeostasis and is associated with the state of cell proliferation and DNA methylation [61,62]. The complete sequence of T2T-YAO shows that the copy number of rDNA clusters in each chromosome is highly variable, ranging from 7 to 52. The total copy numbers in Mat and Pat haploids are 79 and 149, respectively, significantly fewer than the 219 copies in CHM13, but consistent with our ddPCR results (Table S2) and previous reports [60]. The amplification of the rDNA gene in CHM13 is possibly an adaption to the rapid proliferation of the hydatidiform mole cells.

### The architecture of chromosome Y

The assembly of chromosome Y appears to be particularly challenging due to the high degree of repetitive sequence content, especially in the Yq12 region. The T2T assembly of our YAO-Y has a length of 51 Mb, which is smaller than the T2T-Y in CHM13v2.0 by about 10 Mb [63], but still within the recently reported range of length polymorphism of chromosome Y (45.2–84.9 Mb) [64]. The difference is primarily due to a contraction in the Yq12 heterochromatic region, which spans 24 Mb in YAO. Similar to previous observations [63,64], the Yq12 region in YAO is also comprised of alternating human satellite 1 and 3 blocks, which are repetitive sequences with high identity. The correctness of our YAO-Y assembly is supported by the evenly covered HiFi and ONT reads mapping to it (Figure 4D), and some segments in the Yq12 region showing decreased HiFi coverage are composed of DYZ2 repetitive satellite sequences of high A/T content, which are known to be biased in HiFi sequencing [5] (Figure 4D). The architecture of YAO-Y is consistent with the previous report, containing pseudoautosomal regions (PARs) at both ends, X-transposed regions, ampliconic sequences, heterochromatic satellites, and X-degenerate regions. Most of the orthologous blocks were aligned perfectly with those in CHM13-Y, except for some translocations and inversions in the ampliconic region and the low-identity region of Yq12 (Figure 3A). In YAO-Y, the copy numbers of the ampliconic genes of *TSPY*, *DAZ*, and *RBMY* are 53, 4, and 32 copies, respectively, consistent with those in CHM13-Y and confirmed by our *in-silico* PCR and ddPCR results (Table S2). The ampliconic structures of these genes are also fully preserved in the ampliconic regions (Figure 4D), like the pattern observed in CHM13-Y [63].

**Figure 4 Assembly of the SAACs and chromosome Y**

**A.** Bandage visualization of acrocentric chromosomes. The left panel shows that the short arms of the ten acrocentric chromosomes are entangled together, while the long arms of each pair of homologous chromosomes are separately twisted. The right panel shows the enlargement of the entangled short arms. The constitute block of chromosomes is colored according to the haplotype and chromosome it is assigned. **B.** Sequence similarity among the SAACs. The ten SAACs are permutated along the circumference. Pairwise alignments between heterogenous chromosomes with length > 10 kb and identity > 90% are linked by curves. The orange and blue colors and their depth of the curves indicate the direction and identity of the alignment as displayed in the legend. **C.** The PCoA plot of the short arms and long arms of the acrocentric chromosomes based on their distance in *k*-mer composition. **D.** The structure of the complete chromosome YAO-Y. Peak charts show the GC content, coverage of mapped HiFi and ONT reads (> 100 kb) along YAO-Y. The middle panel shows the positions of the ampliconic genes *RBMY*, *DAZ*, and *TSPY* on YAO-Y. The bottom panel is the alignment dot plot of the tandem repeats on YAO-Y visualized by StainedGlass, with the dot color indicating sequence identity. SAAC, short arms of acrocentric chromosome; PCoA, principal co-ordinates analysis; ONT, Oxford Nanopore Technologies.

### A truly complete reference-quality diploid genome for the Han population

T2T-YAO includes *de novo* T2T assembly for the diploid human genome of a single Han individual, with 44 + XY chromosomes that contain 22 autosomes and X/Y in each Mat/Pat haplotype, inherited separately from the mother and father. This diploid assembly poses significant challenges due to the regions of high repetitiveness and extensive recombination, particularly in the SAAC regions. Despite these challenges, T2T-YAO has been completed thanks to the advent of high-accuracy long-reads sequencing technology and pipelines for read phasing and diploid assembly. T2T-YAO achieved an even higher quality than T2T-CHM13 and had no gaps filled with model nucleotides, representing a truly complete and accurate genome sequence of a real individual, including the Y chromosome. The difference of allele genes in their effects on phenotype and function in various biological processes is still an open question, as well as the epigenetic modification and temporal-spatial expression of allele genes and all their transcripts in various tissues, emphasizing the importance of building a diploid T2T reference for human genome. T2T-YAO and its induced pluripotent stem (iPS) cell lines thus provide an applicable resource for studying the genotype–phenotype relationship and cellular biological processes in a haplotype-phased manner, with implications for understanding all kinds of biological processes and disease pathogenesis.

The differences of both YAO haplotypes to CHM13 are much greater than their in-between differences in terms of ancestry markers, exclusive sequences and genes, *k*-mer composition, sequence-level variations, and structure variations. The inter-ethnic difference, which has long been underestimated due to the preoccupation with the incomplete genomes assembled from short-read sequencing data, may contain significant genetic features associated with ethnic populations, diseases, and various phenotypes. The previously unknown distinct sequence and structure variations in YAO would undoubtedly provide valuable insights and further our understanding of human genomics and its applications in clinical medicine. The availability of T2T-YAO would of course improve mapping and analysis of short-read and long-read data from samples of Han individuals, highlighting the necessity of building a special T2T reference for the huge population. Furthermore, the comprehensive characterization of the genetic features of T2T-YAO is ongoing, including complementing the annotations of its exclusive genes by transcriptome sequencing of the iPS cells and investigating the genomic diversity of the Han population by variant calling against T2T-YAO. All these studies will provide a valuable resource for researchers to study the genomics of Han people, improving the practice in health research and precision medicine in China.

## Materials and methods

### Sample collection

For the YAO genome, a fresh blood sample was collected from a healthy male of northern Han and both of his parents. Routine blood test and liver and kidney function were conducted to exclude potential diseases.

### Karyotyping and iPS cell line

For karyotype analysis, peripheral blood lymphocytes (PBLs) were cultured. The fresh and anticoagulated blood samples were inoculated in the culture bottles, and then the culture bottles were placed in an incubator set at a constant temperature of 37 °C for 68 h. Prior to harvesting the PBLs, colchicine was added into culture medium for an additional 60 min of cultivation. Then the PBLs were harvested and treated with KCl solution, followed by fixation with Carnoy's fixative (methanol: acetic acid = 3:1). Chromosomes were stained using Giemsa, and images of chromosome spread were captured using an automated cytogenetic imaging system (Catalog No. GSL-10, Leica, Germany).

PBMCs were obtained from peripheral blood with Ficoll–Hypaque density gradient centrifugation and expanded in Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 10% fetal bovine serum, 100 U/ml penicillin, and 100 μg/ml streptomycin in a humidified incubator with 5% $CO_2$ at 37 °C. The PBMCs were reprogrammed by using the Sendai virus (CytoTune-iPS 2.0 Sendai Reprogramming Kit, Thermo Fisher Scientific, MA) according to the manufacturer's protocol, which includes four Yamanaka factors, OCT4, SOX2, KLF4, and c-MYC, which are sufficient for efficient reprogramming. Subsequently, the transfected iPS cells were suspended in Essential 8 medium (Catalog No. A1517001, Thermo Fisher Scientific) on 0.1 mg/ml Matrigel coated plates (Catalog No. 354277, Corning, NY) and incubated at 37 °C in 5% $CO_2$ atmosphere. From the next day, the erythroid medium was half-changed with ReproTeSR medium (Catalog No. 05926, StemCell Technologies, Vancouver, Canada) every other day. Fourteen days after transfection, iPS cell colonies were picked up and cultured in mTeSR Plus medium (Catalog No. 05825, StemCell Technologies) in vitronectin (1:100 dilution; Catalog No. A14700, Thermo Fisher Scientific) coated 6-well plates. Cells were passaged using Versene (Catalog No. 15040066, Thermo Fisher Scientific) and mTeSR Plus according to the manufacturer's instructions at a ratio of 1:10.

### Genomic DNA extraction

High-molecular-weight genomic DNA (gDNA) was prepared by the cetyltrimethylammonium bromide (CTAB) method and followed by purification with the QIAGEN Genomic Kit (Catalog No. 13343, QIAGEN, Hilden, Germany). Ultra-long DNA was extracted by the sodium dodecyl sulfate (SDS) method without purification step to sustain the length of DNA. DNA purity was detected using the NanoDrop One UV-Vis spectrophotometer (Thermo Fisher Scientific). DNA degradation and contamination of the extracted DNA were monitored on 1% agarose gels. At last, DNA concentration was further measured by a Qubit 4.0 fluorometer (Thermo Fisher Scientific).

### Ultra-long ONT sequencing

For each ultra-long Nanopore library, approximately 8–10 μg of gDNA was size-selected (> 50 kb) with the SageHLS HMW Library System (Sage Science, MA) and processed using the Ligation Sequencing Kit 1D (Catalog No. SQK-

LSK109, ONT, UK) according to the manufacturer's instructions. About 800 ng DNA libraries were constructed and sequenced on the PromethION (ONT) at the Genome Center of GrandOmics (Wuhan, China).

### PacBio HiFi sequencing

For PacBio HiFi whole-genome sequencing, SMRTbell target-size libraries were prepared according to PacBio's standard protocol (Pacific Biosciences, CA) using 15-kb preparation solutions. The main steps include (1) gDNA shearing by g-TUBEs (Covaris, MA); (2) DNA damage repair, end repair, and A-tailing; (3) ligation with hairpin adapters from the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences); (4) nuclease treatment of the SMRTbell library with the SMRTbell Enzyme Clean Up Kit; and (5) size selection by the BluePippin (Sage Science), and binding to polymerase. The SMRTbell library was then purified by AMPure PB beads, and Agilent 2100 Bioanalyzer (Agilent Technologies, CA) was used to detect the size of library fragments. Sequencing was performed on a PacBio Sequel II instrument with Sequencing Primer V2 and Sequel II Binding Kit 2.0 in GrandOmics.

### Short-read sequencing

For short-read sequencing, the DNA samples of the offspring and his parents were sequenced on DNBSEQ-T7. One microgram gDNA was randomly fragmented by Covaris, and then the fragments were selected by Magnetic beads to an average size of 200–400 bp. Eligible fragments were end repaired and then 3′ adenylated. Adaptors were ligated to the ends of these 3′ adenylated fragments, which were used to amplify the fragments. The PCR products were heat denatured and circularized by the splint oligo sequence. Ultimately, the single strand circle DNA (ssCir DNA) was formatted as the final library. The library was amplified with phi29 to make DNA nanoball (DNB), and then DNB was loaded into the patterned nanoarray, and paired-end 100/150 base reads were generated by combinatorial Probe-Anchor Synthesis (cPAS).

### Hi-C sequencing

For Hi-C sequencing, purified DNA was digested with 100 U DpnII and incubated with Biotin-14-dATP. The ligated DNA was sheared into fragments of 300–600 bp, and then blunt-end repaired and A-tailed, followed by purification through biotin–streptavidin-mediated pulldown. Finally, the Hi-C libraries were quantified and sequenced using the Illumina NovaSeq/MGI-2000 platform.

### BioNano

For BioNano analysis, extracted DNA was subject to manufacturer-recommended protocols for library preparation using BioNano Prep Animal Tissue DNA Isolation Kit (Catalog No. 80002, BioNano Genomics, CA) and optical scanning provided by BioNano Genomics (https://bionanogenomics.com), with the labeling enzyme Direct Label Enzyme (DLE) in BioNano PrepDLS Labeling DNA Kit (Catalog No. 80005, BioNano Genomics). Labeled DNA samples were loaded and run on the Saphyr system (BioNano Genomics) in GrandOmics.

### Genome assembly

Before assembling, HiSeq reads of the offspring and both parents were trimmed, and only bases with Phred QS > 30 were retained for the following *k*-mer calling. Hap-mers called according to the *k*-mers of each parent were input along with the HiFi and ONT reads, and the initial assemblies were constructed with Verkko (https://github.com/marbl/verkko) in the trio mode [28]. The output fully-phased contigs of the Mat and Pat genomes of YAO were further scaffolded, guided by CHM13. Then the gaps in either haplotype were closed in the following processes in a phased manner. To fill the gaps, ultra-long ONT reads > 250 kb and HiFi reads with low-frequency *k*-mers were first mapped to the extracted 5 kb of flanking regions, and then the gap was closed with the consensus sequences of HiFi reads scaffolded by the aligned ONT reads. Remaining gaps or deleted regions due to poor quality were fulfilled by contigs assembled with altered *k*-mer aligned with the gaps. All these alignments were visualized and manually examined. The rDNA array is highly repetitive and is output as short contigs of rDNA units, whose copy number can be roughly estimated by their coverage of HiFi reads. These rDNA contigs were then trimmed and aligned with the original ONT reads containing multiple rDNA units, constructing a graph in which the walks between the nodes were supported by at least two ONT reads. Then the contigs were clustered into arrays in each chromosome by the edges connecting them, and their order in each chromosome was finalized by manually selecting a route through the graph in which the walks visited each node the appropriate number of times consistent with its estimated copy number. All the final chromosomes were checked for telomere sequences at both ends to ensure their completeness and were free of any model nucleotides.

### Polishing and validation

To correct potential SNV and SV, we followed the polishing pipeline of McCartney and colleagues [5]. Long reads from the offspring are binned into paternal and maternal groups based on the presence of the haplotype-specific *k*-mers by TrioCanu [65]. The alignment was conducted between corresponding haplotypes and reads using Winnowmap2 [29]. DeepVariant [31] and PEPPER-Margin-DeepVariant [32] were used to call small errors from self-alignments of HiFi and ONT reads, respectively. Similarly, calls with low allele fraction support or a low genotype quality (GQ) score were removed [GQ < 30 for the HiFi calls and GQ < 25 for ONT SNP calls, and variant allele frequency (VAF) < 0.5]. The remaining SNVs were merged using the custom script [5], and then Merfin [34] was used to ensure that no new false *k*-mers were introduced due to the correction of the SNVs. The SVs were detected by Sniffles [33] from the alignments of ONT and HiFi, and finally, SVs with supportive reads less than 60% were removed. Jasmine [35] was used to merge the SVs. The SVs identified by both ONT and HiFi reads were manually

inspected and validated in Integrative Genomics Viewer (v2.6), and the true SVs were corrected.

#### Assessment

The 21-mers of reads were collected from next-generation sequencing (NGS) and HiFi reads of offspring and the parental NGS reads using Meryl, and Merqury [38] was used to calculate QV, completeness, and phasing statistics. Collapsed and expandable sequences were calculated using Segmental Duplication Assembler (SDA) [66]. SDs were detected by BISER [67].

#### ddPCR

The gDNA was isolated from the iPS cell line using the Magetic Univeral Genomic DNA Kit (Catalog No. DP705-01, TIANGEN, Beijing, China). Copy numbers of ampliconic genes were validated through ddPCR. Primers, gDNA concentrations, and restriction enzymes are referred to previous studies [4,68]. The ddPCR reactions were performed using the dye method, ddPCR Detection Kit (TargetingOne, Beijing, China), and TargetingOne ddPCR system. Briefly, each reaction consists of 15 μl 2× ddPCR Supermix, 0.2 μl of restriction enzyme for fragmentation, 3 μl of 10 μM primer mix, 1 μl of 2–50 ng DNA template, and 10.8 μl nuclease-free water. Mastermixes were then emulsified with 180-μl droplet generator oil using a droplet generator according to the manufacturer's instructions. After droplet generation, thermocycling was performed with the following parameters: 5 min at 37 °C, 10 min at 95 °C, 45 cycles consisting of a 30-s denaturation at 94 °C and a 60-s extension at 55 °C, followed by a hold at 12 °C for 5 min. Control reactions without the DNA were performed to rule out non-specific amplification.

#### Gene annotation

The final annotation combined the CAT result and the Liftoff annotation. First of all, alignment to GRCh38 was generated with Cactus [69] and used as input for CAT [51] along with the GENCODEv43 annotation. Liftoff [52] was run to map genes from the GENCODEv43 to the T2T-YAO. Predictions of Liftoff that did not overlap any CAT annotations were added using bedtools intersect. Common repeat elements were identified by RepeatMasker (v4.1.0).

#### Ancestry analysis

RFMix 2 [39] (https://github.com/slowkoni/rfmix) was used to infer the local ancestry of the T2T-YAO genome. A total of 2531 individuals with 30× sequenced in the Phase 3 release of the 1KGP were used as a set of reference samples for ancestry (https://www.internationalgenome.org/data-portal/sample) [40]. Biallelic SNVs against the GRCh38 reference were obtained from the 1KGP Consortium (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/20181203_biallelic_SNV/). A genetic map for GRCh38 was obtained from Beagle (https://bochet.gcc.biostat.washington.edu/beagle/genetic_maps/). T2T-YAO (Pat and Mat assemblies) and CHM13v2.0 variants were called on GRCh38 with dipcall (https://github.com/lh3/dipcall) [70]. RFMix2 was performed with "-c 50 -s 5", grouping the 1KGP reference panel into superpopulations (African, admixed American, East Asian, European, and South Asian). For chromosome X, we only used SNVs in PARs and female individuals in the reference panel for ancestry analysis. The computed ancestry regions were lifted over from GRCh38 to CHM13v2.0 using LiftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver).

#### Y haplogroup identification

T2T-YAO chromosome Y was aligned to the hg19 chromosome Y sequence with dipcall (https://github.com/lh3/dipcall) [70] to identify SNPs. We used yHaplo (v1.1.2) to build a tree and determine the haplogroup (https://github.com/23andMe/yhaplo).

#### Genomic variation

To call the full spectrum of heterozygosity among CHM13 and the T2T-YAO diploid genomes, we directly compared these three assemblies using MUMmer (v4.0.0rc1) [71]. Sequences of unalignments and SNPs were generated by "delta-filter -m -i 99 -l 100000" and followed by "dnadiff". The SnpEff v5.0 program [72] was adopted to infer functional annotation of any SNPs or small indels ($< 50$ bp) and any potential deleterious effect on protein structure.

SyRI (v1.5) [55] was used to detect SVs longer than 100 kb, including inversions, translocations, and duplications from MUMmer alignments. SVIM-asm was used to detect the SVs longer than 50 bp.

## Ethical statement

The application for the study was submitted to and approved by the Ethical Review Committee of Linfen Central Hospital, China (Approval No. 2022-20-1). The collection and storage of human samples were registered with and approved by the Human Genetic Resources Administration of China (HGRAC). Written informed consents were obtained from the participants.

## Code availability

The code to reproduce the pangenome from this work can be found at GitHub (https://github.com/ZCGAOlab/ChTY001). The code has also been submitted to BioCode at the National Genomics Data Center (NGDC), Beijing Institute of Genomics (BIG), Chinese Academy of Sciences (CAS) / China National Center for Bioinformation (CNCB) (BioCode: BT007453), which is publicly accessible at https://ngdc.cncb.ac.cn/biocode/tools/BT007453.

## Data availability

The raw sequencing data of T2T-YAO have been deposited in the Genome Sequence Archive for human [73] at the NGDC, BIG, CAS / CNCB (GSA-Human: HRA004987), and are publicly accessible at https://ngdc.cncb.ac.cn/gsa-human. The

T2T-YAO genome sequences have been deposited in the Genome Warehouse [74] at the NGDC, BIG, CAS / CNCB (GWH: GWHDQZJ00000000, GWHDOOG00000000, and GWHDQZI00000000), and are publicly accessible at https://ngdc.cncb.ac.cn/gwh.

## Competing interests

Jiang Hu is an employee of GrandOmics Biosciences, whose company played no role in the design of the experiments or the analysis, interpretation, and presentation of the data. All the other authors have declared no competing interests.

## CRediT authorship contribution statement

**Yukun He:** Investigation, Methodology, Data curation, Formal analysis, Software, Writing – original draft, Writing – review & editing. **Yanan Chu:** Investigation, Methodology, Data curation, Formal analysis, Software, Writing – original draft. **Shuming Guo:** Conceptualization, Formal analysis, Writing – original draft. **Jiang Hu:** Methodology, Data curation, Software, Writing – review & editing. **Ran Li:** Investigation, Methodology, Data curation, Software, Writing – original draft. **Yali Zheng:** Methodology, Formal analysis, Software. **Xinqian Ma:** Methodology, Data curation, Formal analysis. **Zhenglin Du:** Data curation, Formal analysis. **Lili Zhao:** Methodology, Formal analysis. **Wenyi Yu:** Formal analysis, Data curation. **Jianbo Xue:** Formal analysis, Data curation. **Wenjie Bian:** Formal analysis, Data curation. **Feifei Yang:** Formal analysis, Data curation. **Xi Chen:** Formal analysis, Data curation. **Pingan Zhang:** Formal analysis, Data curation. **Rihan Wu:** Data curation, Validation. **Yifan Ma:** Data curation, Validation. **Changjun Shao:** Data curation, Validation. **Jing Chen:** Data curation, Validation. **Jian Wang:** Data curation, Validation. **Jiwei Li:** Data curation, Formal analysis. **Jing Wu:** Data curation, Formal analysis. **Xiaoyi Hu:** Data curation, Formal analysis. **Qiuyue Long:** Data curation, Validation. **Mingzheng Jiang:** Data curation, Validation. **Hongli Ye:** Formal analysis, Data curation. **Shixu Song:** Formal analysis, Data curation. **Guangyao Li:** Formal analysis, Writing – review & editing. **Yue Wei:** Formal analysis, Writing – review & editing. **Yu Xu:** Formal analysis, Writing – review & editing. **Yanliang Ma:** Formal analysis, Visualization. **Yanwen Chen:** Formal analysis, Visualization. **Keqiang Wang:** Formal analysis, Visualization. **Jing Bao:** Formal analysis, Writing – review & editing. **Wen Xi:** Formal analysis, Writing – review & editing. **Fang Wang:** Data curation, Validation. **Wentao Ni:** Data curation, Validation. **Moqin Zhang:** Formal analysis, Visualization. **Yan Yu:** Formal analysis, Visualization. **Shengnan Li:** Formal analysis, Visualization. **Yu Kang:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Formal analysis, Writing – original draft, Writing – review & editing. **Zhancheng Gao:** Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Supervision, Writing – review & editing. All authors have read and approved the final manuscript.

## Acknowledgments

This study was supported by the grants from the Linfen Soft Science Research Project (Grant No. 2126), the National Natural Science Foundation of China (Grant No. 31970568), the National Key R&D Program of China (Grant No. 2021YFC2301000), the National and Provincial Key Clinical Specialty Capacity Building Project 2020 (Department of the Respiratory Medicine), and the Peking University People's Hospital Scientific Research Development Funds (Grant No. RDGS2022-11).

## Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2023.08.001.

## ORCID

ORCID 0000-0002-4164-2478 (Yukun He)
ORCID 0000-0002-9349-4307 (Yanan Chu)
ORCID 0009-0001-7931-3725 (Shuming Guo)
ORCID 0000-0002-8521-9161 (Jiang Hu)
ORCID 0000-0002-5085-0463 (Ran Li)
ORCID 0000-0002-0359-2799 (Yali Zheng)
ORCID 0000-0002-6356-6584 (Xinqian Ma)
ORCID 0000-0003-2147-3475 (Zhenglin Du)
ORCID 0009-0002-9016-0106 (Lili Zhao)
ORCID 0000-0003-3455-3087 (Wenyi Yu)
ORCID 0000-0003-2531-7808 (Jianbo Xue)
ORCID 0000-0001-6716-4489 (Wenjie Bian)
ORCID 0009-0000-1830-5978 (Feifei Yang)
ORCID 0009-0005-5421-8704 (Xi Chen)
ORCID 0000-0003-0524-4322 (Pingan Zhang)
ORCID 0009-0008-4902-9767 (Rihan Wu)
ORCID 0009-0009-7009-4341 (Yifan Ma)
ORCID 0000-0002-5836-9447 (Changjun Shao)
ORCID 0000-0003-2252-8001 (Jing Chen)
ORCID 0000-0002-0434-0139 (Jian Wang)
ORCID 0000-0002-4652-8842 (Jiwei Li)
ORCID 0000-0002-7075-3687 (Jing Wu)
ORCID 0009-0002-3904-0547 (Xiaoyi Hu)
ORCID 0000-0002-5361-3260 (Qiuyue Long)
ORCID 0009-0008-3273-0012 (Mingzheng Jiang)
ORCID 0009-0009-7538-2668 (Hongli Ye)
ORCID 0009-0005-0689-6869 (Shixu Song)
ORCID 0009-0006-1468-2163 (Guangyao Li)
ORCID 0009-0001-4213-6796 (Yue Wei)
ORCID 0009-0000-1297-1312 (Yu Xu)
ORCID 0000-0002-8491-391X (Yanliang Ma)
ORCID 0009-0000-1173-2096 (Yanwen Chen)
ORCID 0009-0001-5157-2565 (Keqiang Wang)
ORCID 0000-0003-2273-4577 (Jing Bao)
ORCID 0009-0002-1058-3345 (Wen Xi)
ORCID 0009-0003-8219-4637 (Fang Wang)
ORCID 0000-0002-0687-5196 (Wentao Ni)
ORCID 0000-0002-5811-9966 (Moqin Zhang)
ORCID 0009-0004-0898-7951 (Yan Yu)
ORCID 0009-0005-4034-9930 (Shengnan Li)
ORCID 0000-0001-5196-0376 (Yu Kang)
ORCID 0000-0001-7415-1416 (Zhancheng Gao)

## References


[1] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921.

[2] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science 2001;291:1304–51.

[3] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature 2004;431:931–45.

[4] Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science 2022;376:44–53.

[5] Mc Cartney AM, Shafin K, Alonge M, Bzikadze AV, Formenti G, Fungtammasan A, et al. Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. Nat Methods 2022;19:687–95.

[6] Gershman A, Sauria MEG, Guitart X, Vollger MR, Hook PW, Hoyt SJ, et al. Epigenetic patterns in a complete human genome. Science 2022;376:eabj5089.

[7] Vollger MR, Guitart X, Dishuck PC, Mercuri L, Harvey WT, Gershman A, et al. Segmental duplications and their variation in a complete human genome. Science 2022;376:eabj6965.

[8] Aganezov S, Yan SM, Soto DC, Kirsche M, Zarate S, Avdeyev P, et al. A complete reference genome improves analysis of human genetic variation. Science 2022;376:eabl3533.

[9] Altemose N, Logsdon GA, Bzikadze AV, Sidhwani P, Langley SA, Caldas GV, et al. Complete genomic and epigenetic maps of human centromeres. Science 2022;376:eabl4178.

[10] Hoyt SJ, Storer JM, Hartley GA, Grady PGS, Gershman A, de Lima LG, et al. From telomere to telomere: the transcriptional and epigenetic state of human repeat elements. Science 2022;376:eabk3112.

[11] Church DM. A next-generation human genome sequence. Science 2022;376:34–5.

[12] Fan JB, Surti U, Taillon-Miller P, Hsie L, Kennedy GC, Hoffner L, et al. Paternal origins of complete hydatidiform moles proven by whole genome single-nucleotide polymorphism haplotyping. Genomics 2002;79:58–62.

[13] Liao WW, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, et al. A draft human pangenome reference. Nature 2023;617:312–24.

[14] Gao Y, Yang X, Chen H, Tan X, Yang Z, Deng L, et al. A pangenome reference of 36 Chinese populations. Nature 2023;619:112–21.

[15] Jarvis ED, Formenti G, Rhie A, Guarracino A, Yang C, Wood J, et al. Semi-automated assembly of high-quality diploid human reference genomes. Nature 2022;611:519–31.

[16] Du Z, Ma L, Qu H, Chen W, Zhang B, Lu X, et al. Whole genome analyses of Chinese population and *de novo* assembly of a Northern Han genome. Genomics Proteomics Bioinformatics 2019;17:229–47.

[17] Yang X, Zhao X, Qu S, Jia P, Wang B, Gao S, et al. Haplotype-resolved Chinese male genome assembly based on high-fidelity sequencing. Fundam Res 2022;2:946–53.

[18] Chao KH, Zimin AV, Pertea M, Salzberg SL. The first gapless, reference-quality, fully annotated genome from a Southern Han Chinese individual. G3 (Bethesda) 2023;13:jkac321.

[19] Yang C, Zhou Y, Song Y, Wu D, Zeng Y, Nie L, et al. The complete and fully-phased diploid genome of a male Han Chinese. Cell Res 2023;33:745–61.

[20] Steinberg KM, Schneider VA, Graves-Lindsay TA, Fulton RS, Agarwala R, Huddleston J, et al. Single haplotype assembly of the human genome from a hydatidiform mole. Genome Res 2014;24:2066–76.

[21] Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol 2019;37:1155–62.

[22] Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Res 2020;30:1291–305.

[23] Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol 2018;36:338–45.

[24] Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, et al. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. PLoS Comput Biol 2019;15:e1007273.

[25] Lam ET, Hastie A, Lin C, Ehrlich D, Das SK, Austin MD, et al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nat Biotechnol 2012;30:771–6.

[26] Foster-Nyarko E, Cottingham H, Wick RR, Judd LM, Lam MMC, Wyres KL, et al. Nanopore-only assemblies for genomic surveillance of the global priority drug-resistant pathogen, *Klebsiella pneumoniae*. Microb Genom 2023;9:mgen000936.

[27] Bankevich A, Bzikadze AV, Kolmogorov M, Antipov D, Pevzner PA. Multiplex de Bruijn graphs enable genome assembly from long, high-fidelity reads. Nat Biotechnol 2022;40:1075–81.

[28] Rautiainen M, Nurk S, Walenz BP, Logsdon GA, Porubsky D, Rhie A, et al. Telomere-to-telomere assembly of diploid chromosomes with Verkko. Nat Biotechnol 2023;41:1474–82.

[29] Jain C, Rhie A, Hansen NF, Koren S, Phillippy AM. Long-read mapping to repetitive reference sequences using Winnowmap2. Nat Methods 2022;19:705–10.

[30] Hu J, Wang Z, Liang F, Liu S, Ye K, Wang DP. NextPolish2: a repeat-aware polishing tool for genomes assembled using HiFi long reads. Genomics Proteomics Bioinformatics 2024;22. https://doi.org/10.1093/gpbjnl/qzad009.

[31] Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol 2018;36:983–7.

[32] Shafin K, Pesout T, Chang PC, Nattestad M, Kolesnikov A, Goel S, et al. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. Nat Methods 2021;18:1322–32.

[33] Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods 2018;15:461–8.

[34] Formenti G, Rhie A, Walenz BP, Thibaud-Nissen F, Shafin K, Koren S, et al. Merfin: improved variant filtering, assembly evaluation and polishing via *k*-mer validation. Nat Methods 2022;19:696–704.

[35] Kirsche M, Prabhu G, Sherman R, Ni B, Battle A, Aganezov S, et al. Jasmine and Iris: population-scale structural variant comparison and analysis. Nat Methods 2023;20:408–17.

[36] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25:2078–9.

[37] Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience 2021;10:giab008.

[38] Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol 2020;21:245.

[39] Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am J Hum Genet 2013;93:278–88.

[40] 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature 2015;526:68–74.

[41] Feng Q, Lu Y, Ni X, Yuan K, Yang Y, Yang X, et al. Genetic history of Xinjiang's Uyghurs suggests Bronze Age multiple-way contacts in Eurasia. Mol Biol Evol 2017;34:2572–82.
[42] Poznik GD. Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or genotyped men. bioRxiv 2016; 088716.
[43] Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, et al. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature 2003;423:825–37.
[44] Tiirikka T, Moilanen JS. Human chromosome Y and haplogroups; introducing YDHS database. Clin Transl Med 2015;4:60.
[45] Shi H, Dong YL, Wen B, Xiao CJ, Underhill PA, Shen PD, et al. Y-chromosome evidence of southern origin of the East Asian-specific haplogroup O3–M122. Am J Hum Genet 2005;77:408–19.
[46] Yan S, Wang CC, Zheng HX, Wang W, Qin ZD, Wei LH, et al. Y chromosomes of 40% Chinese descend from three Neolithic super-grandfathers. PLoS One 2014;9:e105691.
[47] Ning C, Li T, Wang K, Zhang F, Li T, Wu X, et al. Ancient genomes from northern China suggest links between subsistence changes and human migration. Nat Commun 2020;11:2700.
[48] Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol 2004;5:R12.
[49] Vollger MR, Dishuck PC, Harvey WT, DeWitt WS, Guitart X, Goldberg ME, et al. Increased mutation and gene conversion within human segmental duplications. Nature 2023;617:325–34.
[50] Copley KE, Shorter J. Repetitive elements in aging and neurodegeneration. Trends Genet 2023;39:381–400.
[51] Fiddes IT, Armstrong J, Diekhans M, Nachtweide S, Kronenberg ZN, Underwood JG, et al. Comparative annotation toolkit (CAT) — simultaneous clade and personal genome annotation. Genome Res 2018;28:1029–38.
[52] Shumate A, Salzberg SL. Liftoff: accurate mapping of gene annotations. Bioinformatics 2021;37:1639–43.
[53] Frankish A, Carbonell-Sala S, Diekhans M, Jungreis I, Loveland JE, Mudge JM, et al. GENCODE: reference annotation for the human and mouse genomes in 2023. Nucleic Acids Res 2023;51: D942–9.
[54] Heller D, Vingron M. SVIM-asm: structural variant detection from haploid and diploid genome assemblies. Bioinformatics 2020;36:5519–21.
[55] Goel M, Sun H, Jiao WB, Schneeberger K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol 2019;20:277.
[56] Salm MP, Horswell SD, Hutchison CE, Speedy HE, Yang X, Liang L, et al. The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. Genome Res 2012;22:1144–53.
[57] Logsdon GA, Vollger MR, Hsieh P, Mao Y, Liskovykh MA, Koren S, et al. The structure, function and evolution of a complete human chromosome 8. Nature 2021;593:101–7.
[58] Jarmuz-Szymczak M, Janiszewska J, Szyfter K, Shaffer LG. Narrowing the localization of the region breakpoint in most frequent Robertsonian translocations. Chromosome Res 2014;22:517–32.
[59] Guarracino A, Buonaiuto S, de Lima LG, Potapova T, Rhie A, Koren S, et al. Recombination between heterologous human acrocentric chromosomes. Nature 2023;617:335–43.
[60] Nelson JO, Watase GJ, Warsinger-Pepe N, Yamashita YM. Mechanisms of rDNA copy number maintenance. Trends Genet 2019;35:734–42.
[61] Hori Y, Shimamoto A, Kobayashi T. The human ribosomal DNA array is composed of highly homogenized tandem clusters. Genome Res 2021;31:1971–82.
[62] Hori Y, Engel C, Kobayashi T. Regulation of ribosomal RNA gene copy number, transcription and nucleolus organization in eukaryotes. Nat Rev Mol Cell Biol 2023;24:414–29.
[63] Rhie A, Nurk S, Cechova M, Hoyt SJ, Taylor DJ, Altemose N, et al. The complete sequence of a human Y chromosome. Nature 2023;621:344–54.
[64] Hallast P, Ebert P, Loftus M, Yilmaz F, Audano PA, Logsdon GA, et al. Assembly of 43 human Y chromosomes reveals extensive complexity and variation. Nature 2023;621:355–64.
[65] Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. Genome Res 2017;27:722–36.
[66] Vollger MR, Dishuck PC, Sorensen M, Welch AE, Dang V, Dougherty ML, et al. Long-read sequence and assembly of segmental duplications. Nat Methods 2019;16:88–94.
[67] Iseric H, Alkan C, Hach F, Numanagic I. Fast characterization of segmental duplication structure in multiple genome assemblies. Algorithms Mol Biol 2022;17:4.
[68] Tomaszkiewicz M, Rangavittal S, Cechova M, Campos Sanchez R, Fescemyer HW, Harris R, et al. A time- and cost-effective strategy to sequence mammalian Y chromosomes: an application to the *de novo* assembly of gorilla Y. Genome Res 2016;26:530–40.
[69] Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, et al. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature 2020;587:246–51.
[70] Li H, Bloom JM, Farjoun Y, Fleharty M, Gauthier L, Neale B, et al. A synthetic-diploid benchmark for accurate variant-calling evaluation. Nat Methods 2018;15:595–7.
[71] Delcher AL, Salzberg SL, Phillippy AM. Using MUMmer to identify similar regions in large sequence sets. Curr Protoc Bioinformatics 2003;Chapter 10:Unit 10.3.
[72] Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain $w^{1118}$*; iso-2; iso-3*. Fly (Austin) 2012;6:80–92.
[73] Chen T, Chen X, Zhang S, Zhu J, Tang B, Wang A, et al. The Genome Sequence Archive Family: toward explosive data growth and diverse data types. Genomics Proteomics Bioinformatics 2021;19:578–83.
[74] Chen ML, Ma YK, Wu S, Zheng XC, Kang HG, Sang J, et al. Genome Warehouse: a public repository housing genome-scale data. Genomics Proteomics Bioinformatics 2021;19:584–9.