A Telomere-to-Telomere Diploid Reference Genome and Centromere Structure of the Chinese Quartet

Bo Wang[1,2], Peng Jia[3], Stephen J. Bush[1], Xia Wang[4], Yi Yang[4,5], Yu Zhang[4], Shijie Wan[6], Xiaofei Yang[6], Pengyu Zhang[1], Yuanting Zheng[7], Leming Shi[7], Lianhua Dong[4,5*], Kai Ye[1,2,3,8*]

[1] *School of Automation Science and Engineering, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China.*

[2] *MOE Key Lab for Intelligent Networks & Networks Security, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China.*

[3] *Department of Gynecology and Obstetrics, Center for Mathematical Medical, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710061, China.*

[4] *Center for Advanced Measurement of Science, National Institute of Metrology, Beijing 100029, China*

[5] *Shenzhen Institute for Technology Innovation, National Institute of Metrology, Shenzhen 518107, China*

[6] *School of Computer Science and Technology, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China*

[7] *State Key Laboratory of Genetic Engineering, Human Phenome Institute and School of Life Sciences, Fudan University, Shanghai, China*

[8] *Faculty of Science, Leiden University, Leiden 2311 EZ, The Netherlands.*

* Corresponding authors.
E-mail: donglh@nim.ac.cn (Dong L); kaiye@xjtu.edu.cn (Ye K)

Running title: *Wang B et al / T2T assemblies of Chinese Quartet*

Total word count: 5840

Number of figures: 2

Number of tables: 1

Number of references: 39

Number of letters in the article title: 84

Number of letters in the running title: 29

Number of keywords: 5

Words in abstract: 207

Abstract

Recent advances in sequencing technologies have enabled the complete assembly of human genomes from telomere to telomere (T2T), resolving previously inaccessible regions such as centromeres and segmental duplications. While T2T assemblies of the Chinese population (e.g., T2T-CN1 and T2T-YAO) have been reported, haplotype-resolved familial genomes—critical for studying inheritance patterns and structural variation—remain unexplored. Here, we present a high-quality, haplotype-phased T2T assembly of the Chinese Quartet (T2T-CQ), a family cohort comprising monozygotic twins and their parents, generated using high-coverage ultra-long Nanopore and high-fidelity sequencing. The T2T-CQ assembly achieves haplotype quality values exceeding 70, comparable to the T2T-CHM13 reference, enabling comprehensive annotation of centromeric regions. Our analysis reveals novel higher-order repeat (HOR) patterns, including expanded and contracted HOR arrays, shedding light on the dynamic nature of centromere evolution. Furthermore, through haplotype-resolved assembly and comparative analysis of the complete centromeric sequences between the twins, we identify potential copy number variations in HOR structures, with distinct patterns differentiating maternal and paternal inheritance. This work not only provides a critical genomic resource for the Han Chinese population but also establishes a foundation for forensic identification and complex case resolution. Additionally, our dataset contributes to the construction of a T2T pan-Chinese reference genome, facilitating future studies on population-specific structural variations and centromere biology.

KEYWORDS: monozygotic twins; genome assembly; complex genomic regions; centromere; higher-order repeats

Introduction

The completion of the first telomere-to-telomere (T2T) human genome assembly, T2T-CHM13 [1], was a landmark achievement in genomics, providing an updated reference for exploring the full complexity of the human genome, including previously unresolved regions such as centromeres, telomeres, rDNAs, and segmental duplications. This breakthrough was made possible by advances in long-read sequencing technologies, particularly Oxford Nanopore Technologies (ONT) and PacBio high-fidelity (HiFi), which enabled the resolution of highly repetitive and structurally complex regions [2, 3]. Following this, and to close gaps in ethnic diversity, the T2T-CN1 [4]. and T2T-Yao [5] assemblies were released, representing complete diploid genomes of Han Chinese individuals and highlighting the importance of high-quality, population-specific reference genomes to capture the full spectrum of human genetic variation. Building on these achievements, we present the genome T2T-CQ, also referred to as CQ v3.0, a high-quality, haplotype-resolved T2T assembly generated from a family of Han Chinese descent [6, 7]. The Chinese Quartet project (http://chinese-quartet.org/) is a suite of DNA, RNA, protein and metabolite reference materials from two monozygotic twin daughters and their biological parents, extensively detailed in previous publications [6-11]. Our initial assembly of the CQ family [9], hereafter CQ v2.0, laid the groundwork for the present study in which we refine and expand it. By incorporating high-depth ultralong (> 100 kb) ONT reads we significantly enhance both the contiguity and completeness of CQ v2.0 to T2T-level quality, with both maternal and paternal haplotypes attaining average quality values over 70.

Centromeres, which are essential for chromosome segregation during cell division, remain among the most challenging regions to assemble due to their highly repetitive nature and structural complexity [2, 12]. T2T assemblies therefore enable comprehensive annotations of centromeric regions, including the identification of higher-order repeats (HORs), which are essential for understanding centromere structure and function [3, 13, 14]. As such, we focus on the characterization of

centromeric regions in the T2T-CQ genome, identifying HORs apparently novel to the Han Chinese population. Notably, we investigated the centromeric HOR structures between monozygotic twins, revealing structural variations between them. The Chinese Quartet presents an intriguing model for studying centromere biology, genetics and evolution.

Results and discussion

A diploid T2T genome assembly for the Chinese Quartet

In previous work [9], we presented CQ v2.0, haplotype-resolved assemblies of the monozygotic twins. However, although these assemblies were highly contiguous, they fell short of T2T standard, with the maternal haplotype assembly (CQ v2.0_mat) containing 34 gaps, and the paternal (CQ v2.0_pat) 51 gaps (**Table 1**). Moreover, 17 and 20 telomere motifs were missing in the maternal and paternal CQ v2.0 assemblies, respectively (**Table 1**). In this study, we supplemented our previously used set of 10.3× ONT ultralong reads (> 100 kb) with an additional 57.3× (**Table S1**), aiming to assemble a T2T-level haplotype-resolved genome to serve as a new reference genome for the CQ We constructed the initial assemblies using state-of-the-art tools, verkko [15] and hifiasm [16], in trio mode. As described in the methods and illustrated in (**Figure S1A**), the contigs from the two assembly results were merged for the maternal haplotype, leaving 9 gaps distributed across chromosomes 1, 6, 10, 11, 13, 14, 15, 21, and 22. Except for the short arms of acrocentric chromosomes (SAACs) 13, 14, 15, 21, and 22, gaps in the other chromosomes were filled using TGS-GapCloser [17] in conjunction with binning ONT reads and the ONT-based hifiasm assembly (**Figure S1B**). For the paternal haplotype, 8 gaps remained, located on chromosomes 4, 5, 13, 14, 15, 21, 22, and X, which we also filled using TGS-GapCloser (**Figure S2**). The rDNA clusters in the SAACs of both maternal and paternal haplotypes were finalized by adopting the strategy employed for T2T-CN1 [4]. Briefly, this involved predicting the rDNA copy number (CN) using droplet digital PCR (ddPCR), followed by evaluating the rDNA copy number for each chromosome using the T2T-CHM13 major rDNA morphs [18]. The ddPCR results show that, as expected, the copy ratios

of the two twins were similar, with values of 276.92 and 283.73, respectively. Since we combined data from the twins (see Materials and methods), we used their average value of 280 as the rDNA CN. Based on the "rDNA analysis" method, the rDNA CN for CQ v3.0_mat was 123, and for CQ v3.0_pat it was 157 (**Table 1**). The five rDNA major morphs used for genome filling correspond to the five SAACs. The rDNA CN for CQ v3.0_mat and CQ v3.0_pat are detailed in (**Table S2, Table S3**), respectively. We then polished the assembled genome using NextPolish2 [19], ultimately obtaining T2T-CQ diploid human genome assemblies, which were designated CQ v3.0_mat and CQ v3.0_pat, respectively (**Figure 1A**). These two haplotyped assemblies were further validated using Hi-C contact maps and manual inspection, which confirmed the absence of significant assembly errors (**Figure S2**).

The polished T2T-CQ was subsequently evaluated using Merqury [20], a k-mer-based reference-free assessment pipeline, using both Illumina and HiFi reads to identify assembly errors (**Figure S3A**) and switching errors between the CQ v3.0_mat and CQ v3.0_pat haplotypes (**Figure S3B**). Compared to CQ v2.0, the contig N50 and QV of CQ v3.0 were significantly improved (**Table 1**). The contig N50 values in CQ v3.0_mat and CQ v3.0_pat are 155.11 Mb and 155.17 Mb, respectively, surpassing those of CQ v2.0_mat (132.84 Mb) and CQ v2.0_pat (132.84 Mb). The QV averages of chromosomes in CQ v3.0_mat and CQ v3.0_pat are 71.72 and 70.82, respectively, which are higher than those of CQ v2.0_mat (59.07) and CQ v2.0_pat (59.43), as well as being significantly higher than that of the first T2T diploid human genome, CN1 (**Figure 1B**, **Table 1**). The QV values of CQ v3.0 were nearly comparable to those of T2T-Yao [5] and approached the quality of T2T-CHM13 (QV = 73.94) in a parallel comparison [21]. Moreover, CQ v3.0 also exhibits longer assembled contigs compared to CQ v2.0, particularly in the centromeric regions (**Figure 1C**).

A comparison between the CQ v3.0 and CQ v2.0 genomes revealed two megabase-scale structural variants (inversions) on the paternal haplotypes of chromosomes 8 and 16 (**Figure 1D**, **Figure S4**, **Figure S5**), which we confirmed through read alignment and manual validation to be correct (**Figure S6**). We next used SyRI [22] to identify all inversion variants > 100 kb between CQ v3.0 and CQ v2.0. A total of 7 and 13

inversions were detected in the maternal and paternal haplotypes, respectively (**Table S4**), which we again confirmed with manual verification (**Figure S7, FigureS8**).

The heterozygosity rate between the two CQ v3.0 haplotype genomes, calculated using structural variant (SV) count method (https://github.com/T2T-CN1/CN1/tree/main/heterozygosity), revealed that the most diverse regions are located in the centromeres (**Figure 1E**). The heterozygosity rate in the centromeric regions is 5.3 times higher than that in the non-centromeric regions (on average 12.18 vs. 2.29 SVs per 500 kb) (**Figure S9**), which is close to the 5.4-fold difference observed in T2T-CN1 [4]. The high heterozygosity in centromeric regions suggests that these areas are genetically diverse and may play a significant role in maintaining genetic variation within a population.

Centromeric annotations and unique HOR patterns in the Chinese Quartet

With the generation of T2T-CQ, we were able to characterize the landscape of centromeric sequences in the Chinese Quartet. First, we annotated the centromeric α-satellite regions in CQ3.0_mat and CQ3.0_pat using the CenMAP pipeline, which ranged from 2.7 Mb to 10.2 Mb in CQ3.0_mat (**Table S5**) and from 2.5 Mb to 8.4 Mb in CQ3.0_pat (**Table S6**). Sequence identity maps generated using StainedGlass [23] are shown for each centromere of CQ3.0_mat and CQ3.0_pat in (**Figure S10, Figure S11, Figure S12, Figure S13)**. The length of the α-satellite HOR array is well known to exhibit variation across human populations [3, 24]. Accordingly, we annotated the HOR array lengths in both haplotypes and compared them with the HOR lengths of 102 additional human samples (**Figure 2A**). We found that the HOR length of CQ3.0_mat ranged from 787 kb to 6.3 Mb (**Table S7**), while that of CQ3.0_pat ranged from 183 kb to 5.1 Mb (**Table S8**). Interestingly, the HOR lengths of the two haplotypes of CQ differ significantly, with the exception of chromosome 2, where the HOR lengths are similar (1,999,710 bp in CQ3.0_mat and 1,978,117 bp in CQ3.0_pat), both close to the mean value (**Figure 2A**). We then visualized the α-satellite HOR array structures of both haplotypes and compared them with those of T2T-CHM13 (**Figure 2B**). For chromosome 1, the HOR structures of T2T-CHM13

exhibit nearly 2 Mb of unique sequence compared to CQ3.0_mat and CQ3.0_pat (**Figure 2B**). These regions are primarily composed of 2-mer HOR S1C1/5/19H1L.5_6/4 and S1C1/5/19H1L.5-6, as well as 6-mer HOR S1C1/5/19H1L.1-5_6/4. Notably, the S1C1/5/19H1L.5_6/4 motif is entirely absent in both haplotypes of CQ v3.0 (**Figure S14**). For chromosome 17, a unique region of less than 2 Mb is observed in CQ v3.0_mat, composed of the 13-mer HOR S3C17H1L.1-10_15_15-16 (**Figure 2B**). The canonical HOR pattern for chromosome 17 is S3C17H1L.1-16, whereas S3C17H1L.1-10_15_15-16 lacks monomers 11, 12, 13, and 14, and has monomer 15 duplicated once (**Figure 2C**). The S3C17H1L.1-10_15_15-16 is exclusively present in CQ v3.0_mat, while it is entirely absent in both T2T-CHM13 and CQ v3.0_pat. This suggests that S3C17H1L.1-10_15_15-16 is a unique HOR specific to CQ v3.0_mat (**Figure 2BC**). Interestingly, both alleles of chromosome 17 in T2T-CN1 were found to contain S3C17H1L.1-10_15_15-16 (**Figure S15**), suggesting that this HOR might be specific to the Han Chinese. The HOR S3C17H1L.1-13_15_15-16, which replaces monomer 14 with monomer 15 in the canonical pattern S3C17H1L.1-16, has a repeat length of 234 on chromosome 17 of CQ v3.0_mat (**Figure 2C**), a significant expansion compared to its length in both T2T-CHM13 (n = 3) and CQ v3.0_pat (n = 10). Furthermore, we identified a unique 4-mer HOR, CQ-specific pattern, S2C2H1L.1-2_1/3_4, on chromosome 2 (**Figure 2C**). This HOR is formed by the fusion of monomer 1 with monomer 3 from the canonical HOR S2C2H1L.1-4, resulting in a hybrid monomer 1/3 composed of monomers 1 and 3 (**Figure 2C**). The quantities of the 2-mer HOR S2C2H1L.1/3_4 and S2C2H1L.1_2/4 are contracted in both haplotypes of CQ v3.0 compared to T2T-CHM13 (**Figure 2C**). Interestingly, these two 2-mer HOR were not detected in T2T-CN1 [4], further emphasizes the genetic diversity within the Han Chinese population. We also identified a CQ-specific novel 10-mer HOR on chromosome 20, S2C20H1L.7-16. This HOR is derived by deleting monomers 1 to 6 from the canonical S2C20H1L.1-16 pattern (**Figure 2C**). Whether these CQ-specific HORs belong to Chinese population-specific HORs, or whether the presence/absence or expansion/contraction variation in HOR patterns is population-related, remains to be

further investigated. To do so, additional T2T-complete assembled human genome data from diverse populations are required for a systematic evaluation of whether these structural variations in HOR patterns exhibit population specificity.

In a previous study, we detected 68 *de novo* mutations shared between the monozygotic twins in non-centromeric regions [9]. In this study, we aimed to investigate structural variations in the centromeric regions between the twins using T2T genome assembly data. We independently generated haplotype-resolved assemblies for LCL5 and LCL6. A total of 15 T2T-level chromosomes were assembled for the maternal haplotypes of both LCL5 and LCL6 (**Figure S16**), while 14 were assembled for the paternal haplotypes (**Figure S17**). Using the T2T assembled chromosomes, we annotated higher-order repeats in the centromeric regions. Among the 15 maternal haplotypes, six chromosomes exhibited differences in HOR copy numbers between LCL5 and LCL6 (**Figure 1D, Figure S16**). Among the 14 maternal haplotypes, half of the chromosomes exhibited differences in HOR copy numbers between LCL5 and LCL6 (**Figure 1E, Figure S17**). Interestingly, both the maternal and paternal haplotypes of chromosome 15 in LCL5 and LCL6 exhibited variations in HORs (**Figure 1DE)**. Local synteny analysis revealed that the maternal haplotypes of LCL5 and LCL6 contained seven deletion variants, occurring in the 4-mer HOR S2C15H1L.1-4 and the 11-mer HOR S2C15H1L.1-11 (**Figure 1F**). In contrast, the paternal haplotypes of LCL5 and LCL6 also showed seven structural variants, but these consisted of two deletions and five insertions (**Figure 1G**). These findings suggest that, based on the assembly results, centromeric regions between the twins might exhibit differences in HOR copy numbers, and these variations differ between the two haplotypes. However, further experimental validation is required to confirm these observations.

Conclusions

In summary, this study advances genomic research by delivering a high-quality telomere-to-telomere assembly of the Chinese Quartet. By leveraging advanced sequencing technologies, we uncovered novel structural variations in centromeric

regions, as well as HOR expansions and contractions. Comparative analysis between the twins revealed potential HOR copy number variations, highlighting distinct patterns between maternal and paternal lineages. These findings enhance our understanding of centromere diversity, inheritance, and twin genomics. Furthermore, this work provides a valuable resource for forensic applications and supports the development of a comprehensive T2T-pan-Chinese reference genome, contributing to population-scale genomics and precision medicine [25, 26].

Materials and methods

Sample collection

The Chinese Quartet family, comprising a 60-year-old father (LCL7), 60-year-old mother (LCL8), and their two 30-year-old monozygotic twin daughters (LCL5 and LCL6), all from the Fudan Taizhou cohort, are Certified Reference Materials according to the State Administration for Market Regulation in China (reference numbers GBW09900-GBW09903). The process of establishing the cell lines was described in a prior study [27].

ONT ultralong library preparation and sequencing

ONT ultralong DNA was extracted using the GrandOmics Genomic DNA Kit following the manufacturer's guidelines. The quality and quantity of the total DNA were assessed to ensure suitability for ultra-long Nanopore library construction, which was performed according to the established protocols [28]. DNA libraries (approximately 400 ng) were constructed and sequenced on the PromethION platform (Oxford Nanopore Technologies) at the Genome Center of GrandOmics (Wuhan, China).

Hi-C sequencing

The Hi-C library was prepared from cross-linked chromatins of the cells following a standard Hi-C protocol [29]. The library was subsequently sequenced on the MGI-2000 platform. The Hi-C sequencing data were used to map the assembled genome

using Juicer v1.6 [30]. The assemblies were then manually reviewed and visualized using Juicebox v1.11.08 [31].

Genome assembly

Since monozygotic twins are generally considered genetically identical, with only limited somatic substitutions [32], we merged the sequencing reads from the two children to generate a high-quality haplotype-resolved genome [9]. The initial assemblies were constructed using state-of-the-art tools, Verkko v2.2.1 [15] and hifiasm v0.19.9 [16], in trio mode. The scaffolding of the Verkko assemblies was enhanced using the hifiasm assembly, as the contig N50 of the Verkko-assembled genome (131.8 Mb) was greater than that of hifiasm (107.5 Mb). For each haplotype, hifiasm and Verkko assemblies were assigned to their respective chromosomes and aligned to the CHM13 reference genome (**Figure S1A**). Within each chromosome, contigs were merged into longer scaffolds based on overlaps between hifiasm and Verkko assemblies. Additionally, supplementary assemblies were generated using hifiasm v0.24.0 with the '--ONT' parameter, incorporating canu-based [33] binned ONT reads. The gaps were initially filled using TGS-GapCloser v 1.2.1 [17], leveraging the hifiasm assemblies with ONT data and binned ONT reads. The highly repetitive nature of rDNA, combined with heterozygosity, presents significant challenges for resolution [1]. Therefore, only the rDNA copy number was estimated to populate each chromosome with corresponding identical copies [4]. For the assembly of the twins LCL5 and LCL6, we employed hifiasm v0.24.0 in trio mode separately. Subsequently, RagTag v2.0.1 [34] was utilized to scaffold the assemblies using CQ v3.0. The assemblies were then filtered for T2T completeness using 'seqtk telo' and 'seqtk gap' (https://github.com/lh3/seqtk). Following this, the same analytical procedures as those applied to CQ v3.0 were conducted for further analysis.

Genome polishing

Since our genome assembly was performed using the trio-binning approach, the two haplotype assembly results were polished separately. First, the yak v0.1-r56

(https://github.com/lh3/yak) was used to bin HiFi reads based on parental information. Subsequently, the 'getseq' function from fxTools v0.1.0 (https://github.com/moold/fxTools) was employed to extract the binned HiFi reads. The "Racon + Merfin" pipeline was then executed to generate the HiFi mapping file [21], "racon.meryl.iter_1.winnowmap.bam". Finally, the NextPolish2 workflow was run to produce the polished genome files [19].

Genome assessment

Meryl v1.4.1 was utilized to collect 21-mers from the Illumina and HiFi reads of the twins, constructing a hybrid kmer database. Subsequently, Merqury was employed to evaluate the assemblies [20], following the methodology outlined in previous studies [4, 5].

Genomic variation

To comprehensively identify the full spectrum of heterozygosity between the CHM13 and the diploid T2T-CQ genomes, we performed a direct comparison of these three assemblies using minimap2 v2.20-r1061 [35] with the following parameters: -x asm5 -t 64 --cs. The paf2delta.pl script (https://github.com/gorliver/paf2delta) was used to convert the PAF format into the delta format. Subsequently, the 'show-snps' function from MUMmer v4.0.0rc1 [36] and the 'Transform --SNP' function from GenomeSyn [37] were employed to generate SNP results. Additionally, the 'show-diff' function from MUMmer and the 'Transform --PAV' function from GenomeSyn were utilized to produce PAV (Presence-Absence Variation) results. Following this, GenomeSyn was utilized to conduct a whole-genome comparison between the CHM13 genome and the assembled genome in this study. Structural variants > 100 kb, specifically inversions, between the CQ v3.0 and CQ v2.0 genome assemblies were identified using SyRI v1.7.0 [22].

Centromere analysis

CenMAP v0.2.4 (https://github.com/logsdon-lab/CenMAP), a centromere mapping

and annotation pipeline designed for T2T human genome assemblies, was utilized to predict potential centromeric α-satellite regions in the genome. The α-satellite sequences were annotated using the pipeline hmmer-run.sh (https://github.com/fedorrik/HumAS-HMMER_for_AnVIL) with the hmm file 'AS-HORs-hmmer3.3.2-120124.hmm'. Subsequently, HOR structural variants were identified using the tool StV (https://github.com/fedorrik/stv). StainedGlass v0.6 was employed to visualize the centromeric regions [23], and custom R scripts (https://github.com/BoWangXJTU/T2T-CQ/tree/main/03.centromere_analysis) were utilized to plot the HOR length and StV annotations. The HOR length was calculated using 'censtats' from CenMAP v0.2.4.

Estimation of rDNA copy number using ddPCR

The ddPCR experimental protocol followed a previously published method [4], with the modification that the single-copy gene *RPPH1* was used instead of the *TBP1* gene. The rDNA forward primer was 5'-AACGTGAGCTGGGTTTAG-3', the rDNA reverse primer was 5'-CTCGTACTGAGCAGGATTAC-3', and the rDNA probe was 5'-CACATCATCAGTAGGGT-3'. The forward primer, reverse primer, and probe for *RPPH1* were 5'-GAGGGAAGCTCATCAGTGG-3', 5'-CCCTAGTCTCAGACCTTCC-3', and 5'-CCACGAGCTGAGTGC-3', respectively. The rDNA copy number was calculated as the ratio of rDNA copies/μL to *RPPH1* copies/μL.

rDNA analysis

Ribotin was used to extract binned ONT reads containing rDNA sequences [18]. Subsequently, the extracted ONT reads were mapped to the major morphs of CHM13 (https://github.com/maickrau/ribotin_paper_experiments) using minimap2 v2.20-r1061 [35]. The ONT read counts and base counts for each major morph of CHM13 were calculated, and the rDNA copy number for each morph was determined (https://github.com/BoWangXJTU/CQ_T2T/tree/main/04.rDNA_analysis). Finally, the script 'rDNA_CN_fill_genome_assembly_v2.py' was used to fill the rDNA regions in the genome assembly.

Acknowledgements

This study was supported by the National Key R&D Program of China (grant numbers 2023YFF0613300 and 2022YFC3400300) and the National Natural Science Foundation of China (grant numbers 32200510, 32400509, and 32125009).

CRediT author statement

Bo Wang: Investigation, Writing – original draft, Methodology, Visualization, Funding acquisition. **Peng Jia:** Investigation, Methodology. **Stephen J. Bush:** Writing – review & editing. **Xia Wang:** Validation, Investigation. **Yi Yang:** Validation, Investigation. **Yu Zhang:** Validation, Investigation. **Shijie Wan:** Methodology, Visualization. **Xiaofei Yang:** Methodology, Visualization. **Pengyu Zhang:** Methodology. **Yuanting Zheng:** Investigation, Data curation. **Leming Shi:** Investigation, Data curation. **Lianhua Dong:** Conceptualization, Supervision, Investigation, Data curation, Writing – review & editing, Resources, Funding acquisition. **Kai Ye:** Conceptualization, Supervision, Methodology, Writing – review & editing, Resources, Funding acquisition. All authors have read and approved the final manuscript.

Code availability

The pipelines and code for genome assembly, polishing, centromere analysis, and rDNA analysis are available on GitHub: https://github.com/BoWangXJTU/T2T-CQ. 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: BT007908), which is publicly accessible at https://ngdc.cncb.ac.cn/biocode/tools/BT007908.

Data availability

The raw sequencing data of the Chinese Quartet have been deposited in the Genome Sequence Archive for Human [38] (GSA-Human) at the NGDC, BIG, CAS / CNCB,

under accessions HRA010594. The CQ v3.0 genome sequences have been deposited in the Genome Warehouse [39] (GWH) at the NGDC, BIG, CAS / CNCB, under accessions GWHFQEY00000000.1 and GWHFQEX00000000.1.

Ethical statement

The Quartet Project received approval from the Institutional Review Board of the School of Life Sciences at Fudan University (BE2050). The study was conducted in accordance with the principles outlined in the Declaration of Helsinki. Four healthy volunteers from a family quartet, participating in the Taizhou Longitudinal Study in Taizhou, Jiangsu, China, were enrolled. Peripheral blood samples were collected from these individuals to establish human immortalized B-lymphoblastoid cell lines. All four donors provided signed informed consent forms prior to participation.

Competing interests

The authors declare no competing interests.

ORCID

0000-0002-9041-878X (**Bo Wang**)
0000-0002-3429-919X (**Peng Jia**)
0000-0001-9341-2562 (**Stephen J. Bush**)
0009-0003-1935-1249 (**Xia Wang**)
0000-0003-0863-0878 (**Yi Yang)**
0000-0002-8807-6389 (**Yu Zhang**)
0009-0005-0834-892X **(Shijie Wan)**
0000-0002-5118-7755 **(Xiaofei Yang)**
0000-0001-8696-4983 **(Pengyu Zhang)**
0000-0003-4480-8303 **(Yuanting Zheng)**
0000-0002-2981-4150 **(Leming Shi)**
0000-0002-8609-9552 **(Lianhua Dong)**
0000-0002-2851-6741 (**Kai Ye**)

References

[1] Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science 2022;376(6588):44-53.
[2] Miga KH. Centromere studies in the era of 'telomere-to-telomere' genomics. Exp Cell Res 2020;394(2):112127.
[3] Logsdon GA, Rozanski AN, Ryabov F, Potapova T, Shepelev VA, Catacchio CR, et al. The variation and evolution of complete human centromeres. Nature 2024;629(8010):136-45.
[4] 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(10):745-61.
[5] He Y, Chu Y, Guo S, Hu J, Li R, Zheng Y, et al. T2T-YAO: A Telomere-to-telomere Assembled Diploid Reference Genome for Han Chinese. Genomics Proteomics Bioinformatics 2023;21(6):1085-100.
[6] Zheng Y, Liu Y, Yang J, Dong L, Zhang R, Tian S, et al. Multi-omics data integration using ratio-based quantitative profiling with Quartet reference materials. Nat Biotechnol 2024;42(7):1133-49.
[7] Ren L, Duan X, Dong L, Zhang R, Yang J, Gao Y, et al. Quartet DNA reference materials and datasets for comprehensively evaluating germline variant calling performance. Genome Biol 2023;24(1):270.
[8] Yang J, Liu Y, Shang J, Chen Q, Chen Q, Ren L, et al. The Quartet Data Portal: integration of community-wide resources for multiomics quality control. Genome Biol 2023;24(1):245.
[9] Jia P, Dong L, Yang X, Wang B, Bush SJ, Wang T, et al. Haplotype-resolved assemblies and variant benchmark of a Chinese Quartet. Genome Biol 2023;24(1):277.
[10] Ren L, Shi L, Zheng Y. Reference Materials for Improving Reliability of Multiomics Profiling. Phenomics 2024;4(5):487-521.
[11] Zhang N, Chen Q, Zhang P, Zhou K, Liu Y, Wang H, et al. Quartet metabolite reference materials for inter-laboratory proficiency test and data integration of metabolomics profiling. Genome Biol 2024;25(1):34.
[12] Talbert PB, Henikoff S. What makes a centromere? Exp Cell Res 2020;389(2):111895.
[13] 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(7857):101-7.
[14] McKinley KL, Cheeseman IM. The molecular basis for centromere identity and function. Nat Rev Mol Cell Biol 2016;17(1):16-29.
[15] 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(10):1474-82.
[16] Cheng H, Asri M, Lucas J, Koren S, Li H. Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. Nat Methods 2024;21(6):967-70.

[17] Xu M, Guo L, Gu S, Wang O, Zhang R, Peters BA, et al. TGS-GapCloser: a fast and accurate gap closer for large genomes with low coverage of error-prone long reads. GigaScience 2020;9(9):giaa094.
[18] Rautiainen M. Ribotin: automated assembly and phasing of rDNA morphs. Bioinformatics 2024;40(3):btae124.
[19] Hu J, Wang Z, Liang F, Liu SL, Ye K, Wang DP. NextPolish2: a repeat-aware polishing tool for genomes assembled using HiFi long reads. Genomics Proteomics Bioinformatics 2024;22(1):qzad009.
[20] Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol 2020;21(1):245.
[21] 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(6):687-95.
[22] Goel M, Sun H, Jiao W-B, Schneeberger K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol 2019;20:1-13.
[23] Vollger MR, Kerpedjiev P, Phillippy AM, Eichler EE. StainedGlass: interactive visualization of massive tandem repeat structures with identity heatmaps. Bioinformatics 2022;38(7):2049-51.
[24] Gao S, Zhang Y, Bush SJ, Wang B, Yang X, Ye K. Centromere landscapes resolved from hundreds of human genomes. Genomics Proteomics Bioinformatics 2024;qzae071.
[25] Taylor DJ, Eizenga JM, Li Q, Das A, Jenike KM, Kenny EE, et al. Beyond the Human Genome Project: The age of complete human genome sequences and pangenome references. Annu Rev Genoma Hum G 2024;25.
[26] Wang B, Dang N, Yang X, Xu S, Ye K. The human pangenome reference: the beginning of a new era for genomics. Sci Bull (Beijing) 2023;68(14):1484-7.
[27] Pan B, Ren L, Onuchic V, Guan M, Kusko R, Bruinsma S, et al. Assessing reproducibility of inherited variants detected with short-read whole genome sequencing. Genome Biol 2022;23(1):2.
[28] Wang B, Yang X, Jia Y, Xu Y, Jia P, Dang N, et al. High-quality Arabidopsis thaliana Genome Assembly with Nanopore and HiFi Long Reads. Genomics Proteomics Bioinformatics 2022;20(1):4-13.
[29] Gong G, Dan C, Xiao S, Guo W, Huang P, Xiong Y, et al. Chromosomal-level assembly of yellow catfish genome using third-generation DNA sequencing and Hi-C analysis. GigaScience 2018;7(11):giy120.
[30] Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst 2016;3(1):95-8.
[31] Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst 2016;3(1):99-101.

[32] van Dongen J, Slagboom PE, Draisma HH, Martin NG, Boomsma DI. The continuing value of twin studies in the omics era. Nat Rev Genet 2012;13(9):640-53.
[33] 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(9):1291-305.
[34] Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S, et al. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. Genome Biol 2022;23(1):258.
[35] Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 2018;34(18):3094-100.
[36] Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast and versatile genome alignment system. PLoS Comput Biol 2018;14(1):e1005944.
[37] Zhou Z, Yu Z, Huang X, Liu J, Guo Y, Chen L, et al. GenomeSyn: a bioinformatics tool for visualizing genome synteny and structural variations. J Genet Genomics 2022;49(12):1174-6.
[38] 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(4):578-83.
[39] Chen M, Ma Y, Wu S, Zheng X, Kang H, Sang J, et al. Genome Warehouse: A Public Repository Housing Genome-scale Data. Genomics Proteomics Bioinformatics 2021;19(4):584-9.

Figure legends

Figure 1. Overview of CQ v3.0 and its comparison with CQ v2.0. (**A**) Comparison of the CQ v3.0 and CHM13 reference genomes. Each set of chromosomes includes the maternal haplotype of CQ v3.0 (CQ v3.0_mat, top), the paternal haplotype of CQ v3.0 (CQ v3.0_pat, bottom), and CHM13 (middle). (**B**) Comparison of the QV for each chromosome of the maternal and paternal haplotypes in CQ v3.0 and CQ v2.0. (**C**) Illustration of the advantage of CQ v3.0 over CQ v2.0 in centromeric region assembly. (**D**) Illustration of large inversion variation between CQ v3.0 and CQ v2.0. (**E**) Visualization of heterozygous regions between two haplotypes using bubbles. Regions with different heterozygous variants are displayed in distinct colors, while centromeric regions are represented by black lines.

Figure 2. Variation in centromeric higher-order repeat (HOR) length, structure, and novel HORs in CQ v3.0 and between the twins. (**A**) The length of the HOR arrays in CQ v3.0_mat (red), CQ v3.0_pat (purple), and 102 human samples. (**B**) The HOR array structure is shown on the axes, along with the organization of each centromeric region in CHM13, CQ v3.0_mat, and CQ v3.0_pat. (**C**) The heatmap showing the HOR number in CHM13, CQ v3.0_mat, and CQ v3.0_pat, along with the HOR patterns. Red solid circles indicate novel HOR patterns. (**D**) The heatmap showing chromosomes with centromeric differences and HOR counts in the maternal haplotypes of LCL5 and LCL6. (**E**) The heatmap showing chromosomes with centromeric differences and HOR counts in the paternal haplotypes of LCL5 and LCL6. Green shapes represent variations in the centromeres of maternal and paternal chromosome 15 between LCL5 and LCL6. (**F**) Local synteny of centromeres on maternal chromosome 15 in LCL5 and LCL6, highlighting HOR deletion variants. (**G**) Local synteny of centromeres on paternal chromosome 15 in LCL5 and LCL6, highlighting HOR deletion and insertion variants.

Tables

Table 1. Comparison of assembly quality among human telomere-to-telomere genome assemblies.

Statistic	CQ v2.0_mat	CQ v2.0_pat	**CQ v3.0_mat**	**CQ v3.0_pat**	CN1 v0.8_mat	CN1 v0.8_pat
Assembly size (Gb)	3	3	3.03	3.03	3.04	2.94
No. of gaps	34	51	0	0	0	0
No. of telomeres	29	26	46	46	46	46
Contig N50 (Mb)	132.84	132.84	155.11	155.17	157.37	145.8
QV (range among chromosomes)	51.93-66.05	50.85-65.11	58.25-84.87	57.57-81.97	47.33-70.83	51.91-73.02
QV (average of chromosomes)	59.07	59.43	71.72	70.82	64.98	63.17
Chromosome composition	22 + X	22 + X	22 + X	22 + X	22 + X	22 + Y
Haplotype switch error (%)	0.15872	0.137578	0.192231	0.172954	*NA*	*NA*
rDNA copy number	*NA*	*NA*	123	157	132	117

Note: Merqury results of CQ v2.0 and CQ v3.0 from the same pipeline used in this study.

Supplementary material

Figure S1. Illustration of the process of merging hifiasm and Verkko assemblies, followed by subsequent gap-filling steps. (**A**) The hifiasm and Verkko assemblies for maternal chromosome 11 were aligned to CHM13 (middle) to organize and orient the contigs. The red block in CHM13 indicates the centromeric region. (**B**) The gap in the maternal chromosome 1 was closed using hifiasm assemblies in ONT mode with binned ONT ultralong reads.

Figure S2. The Hi-C contact maps of the CQ v3.0 maternal (**A**) and paternal (**B**) haplotypes.

Figure S3. The quality of the diploid T2T-CQ assembly. (**A**) The spectrum plots of k-mers in the assembly and reads for evaluating assembly errors. (**B**) Hap-mer blob plot of the T2T-CQ assembly.

Figure S4. Karyotype alignments of CQ v3.0 maternal (top) and CQ v2.0 maternal

(bottom) haplotypes. Blue regions represent centromeric regions.

Figure S5. Karyotype alignments of CQ v3.0 paternal (top) and CQ v2.0 paternal (bottom) haplotypes. Blue regions represent centromeric regions.

Figure S6. IGV map of binned ONT reads mapped to CQ v3.0 and CQ v2.0. We manually confirmed that the large inversions between chr8 and chr16 of the paternal haplotypes in CQ v3.0 and CQ v2.0 were due to assembly errors in CQ v2.0, with CQ v3.0 being correct.

Figure S7. Four IGV map illustrations of binned ONT reads aligned to the CQ v3.0 and CQ v2.0 maternal haplotypes demonstrate the correction of assembly errors in CQ v2.0, specifically for inversions greater than 100 kb.

Figure S8. Four IGV map illustrations of binned ONT reads aligned to the CQ v3.0 and CQ v2.0 paternal haplotypes demonstrate the correction of assembly errors in CQ v2.0, specifically for inversions greater than 100 kb.

Figure S9. Heterozygosity comparison between centromeric and non-centromeric regions. The heterozygosity rate was calculated as the SV count in each 500 kb window. Red squares represent the mean values of the data. This comparison highlights the significantly higher heterozygosity in centromeric regions compared to non-centromeric regions.

Figure S10. StainedGlass plots of the centromeric (CEN) regions of chromosomes 1 to 6 in T2T-CQ

Figure S11. StainedGlass plots of the centromeric (CEN) regions of chromosomes 7 to 12 in T2T-CQ

Figure S12. StainedGlass plots of the centromeric (CEN) regions of chromosomes 13 to 18 in T2T-CQ

Figure S13. StainedGlass plots of the centromeric (CEN) regions of chromosomes 19 to 22 and chromosome X in T2T-CQ

Figure S14. Heatmap showing the number of HORs and their patterns in chromosome 1 of CHM13, CQ v3.0_mat, and CQ v3.0_pat.

Figure S15. The higher-order repeat array structures of the chromosome 17 centromeric region were compared among CHM13, CQ v3.0, and CN1.

Figure S16. The heatmap illustrates the number of higher-order repeats (HORs) in 15 telomere-to-telomere maternal chromosomes of LCL5 and LCL6. Labels in red indicate the presence of HOR variations.

Figure S17. The heatmap illustrates the number of higher-order repeats (HORs) in 14 telomere-to-telomere paternal chromosomes of LCL5 and LCL6. Labels in red indicate the presence of HOR variations.

Table S1. Sequencing depth statistics used in this study.

Table S2. The predicted rDNA copy numbers for the five short arms of acrocentric chromosomes in the CQ v3.0 maternal haplotype.

Table S3. The predicted rDNA copy numbers for the five short arms of acrocentric chromosomes in the CQ v3.0 paternal haplotype.

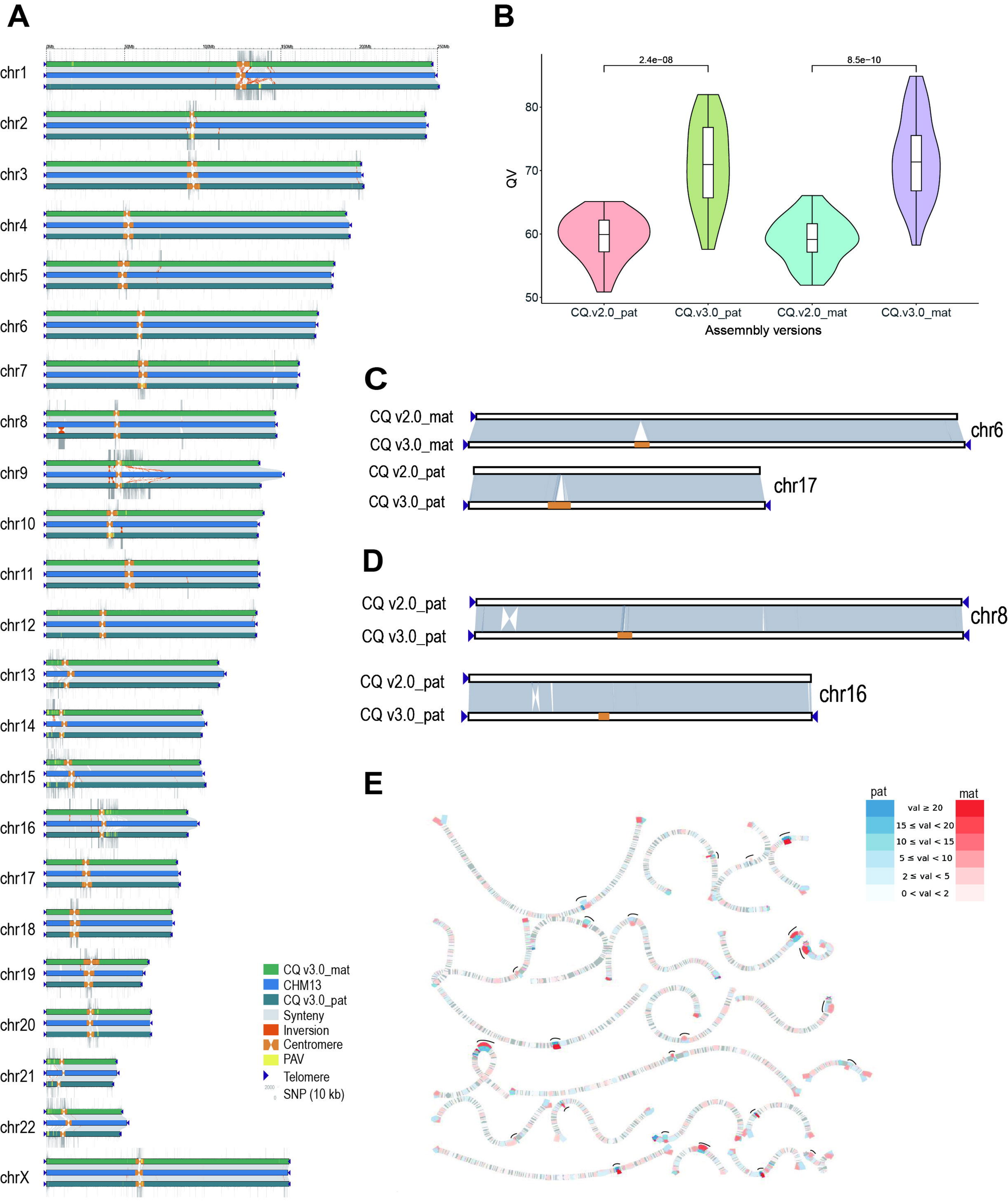
Table S4. The coordinates of inversion variants (> 100 kb) identified between the CQ v2.0 and CQ v3.0 genomes.

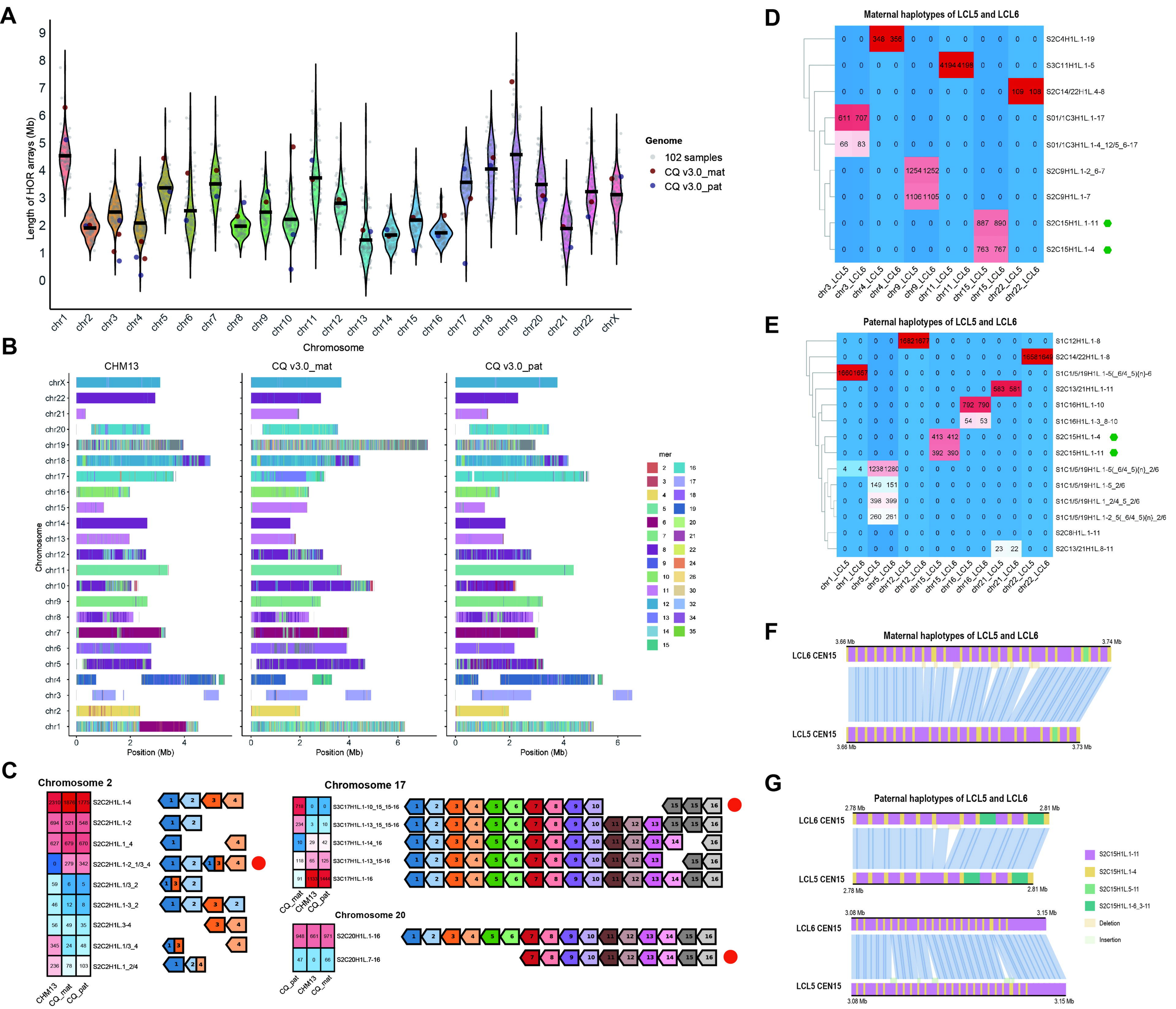
Table S5. CQ v3.0 maternal haplotype centromeric regions and their lengths.

Table S6. CQ v3.0 paternal haplotype centromeric regions and their lengths.

Table S7. CQ v3.0 maternal haplotype centromeric higher-order repeat coordinates.

Table S8. CQ v3.0 paternal haplotype centromeric higher-order repeat coordinates.


A
0Mb
50Mb
100Mb
150Mb
200Mb
250Mb
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
CQ v3.0_mat
CHM13
CQ v3.0_pat
Synteny
Inversion
Centromere
PAV
Telomere
SNP (10 kb)
B
2.4e−08
8.5e−10
QV
80
70
60
50
CQ.v2.0_pat
CQ.v3.0_pat
CQ.v2.0_mat
CQ.v3.0_mat
Assemnbly versions
C
CQ v2.0_mat
CQ v3.0_mat
chr6
CQ v2.0_pat
CQ v3.0_pat
chr17
D
CQ v2.0_pat
CQ v3.0_pat
chr8
CQ v2.0_pat
CQ v3.0_pat
chr16
E
pat
mat
val ≥ 20
15 ≤ val < 20
10 ≤ val < 15
5 ≤ val < 10
2 ≤ val < 5
0 < val < 2


A
Length of HOR arrays (Mb)
Chromosome
Genome
102 samples
CQ v3.0_mat
CQ v3.0_pat
B
CHM13
CQ v3.0_mat
CQ v3.0_pat
Chromosome
Position (Mb)
mer
C
Chromosome 2
Chromosome 17
Chromosome 20
CHM13
CQ_mat
CQ_pat
D
Maternal haplotypes of LCL5 and LCL6
E
Paternal haplotypes of LCL5 and LCL6
F
Maternal haplotypes of LCL5 and LCL6
LCL6 CEN15
LCL5 CEN15
G
Paternal haplotypes of LCL5 and LCL6
LCL6 CEN15
LCL5 CEN15
S2C15H1L.1-11
S2C15H1L.1-4
S2C15H1L.5-11
S2C15H1L.1-6_3-11
Deletion
Insertion