BGI-Tech P&A T2T genome Research Solution

Research Protocol

Contents

Product Background 2

1. Introduction 2
2. Case study 2

Case 1. The complete sequence of a human genome 2

Case 2. The genetic and epigenetic landscape of the Arabidopsis centromeres 4

Research Methods 6

Reference Strategy 7

Bioinformatics Content 9

Product Background

1. Introduction

Telomere-to-telomere (T2T) genome refers to a high-quality complete genome with high genomic accuracy, high continuity, and high integrity. This is conducive to the in-depth study of highly repetitive sequence regions in the genome and helps to resolve centromeres variation characteristics and evolutionary patterns of complex structures such as telomeric and telomeres. 20 years after the draft human genome sequence had been published, the Telomere-to-Telomere (T2T) Consortium published the latest complete human genome sequence CHM13V1.1, which not only contains all the unresolved sequences, but also corrects the original assembly errors, making it the most complete human genome sequence to date.[1]

Today, long-read sequencing technologies such as Pacific Biosciences (PacBio) HiFi and Oxford Nanopore Technologies (ONT) have overcome the low throughput of Sanger sequencing and the short read size of short-read sequencing. The continuous optimization of combined assembly algorithms has greatly facilitated the *De novo* assembly of genomes. In more and more species, assembly gaps caused by complex regions of the genome, complex sex chromosomes, and haploid genomes in polyploid assembly have been analyzed [5-8]. These findings are related to various levels such as disease occurrence and species evolution. Therefore, the construction of a "perfect reference genome" containing T2T, haplotype information and sex chromosome information has become a research trend in genomics.

2. Case study

Case 1. The complete sequence of a human genome

Since its initial release in 2000, the human reference genome has covered only the euchromatic fraction of the genome, leaving important heterochromatic regions unfinished.

Addressing the remaining 8% of the genome, the Telomere-to-Telomere (T2T) Consortium presents a complete 3.055 billion–base pair sequence of a human genome, T2T-CHM13, that includes gapless assemblies for all chromosomes except Y. It corrects errors in the prior references, and introduces nearly 200 million base pairs of sequence containing 1956 gene predictions, 99 of which are predicted to be protein coding. The completed regions include all centromeric satellite arrays, recent segmental duplications, and the short arms of all five acrocentric chromosomes, unlocking these complex regions of the genome to variational and functional studies.

In the complete picture of the X chromosome, the initial assembly of the genome was performed using the 50X ONT Ultra-long data combined with the 70X PacBio HiFi data. The assembled genomic errors were subsequently corrected using 10X Genomics and Bionano Genomics data [2]. Finally, a complete X chromosome map with NG50 = 75Mb was obtained, a ~3.1 Mb centrosomal satellite DNA array (DXZ1) was reconstructed, and 29 gaps in the GRCh38 reference genome were filled.

The assembly method of human chromosome 8 is different from that of the X chromosome. Researchers utilized the respective advantages of ONT Ultra-long and PacBio HiFi data to fill five blank regions on chromosome 8 [3]. In addition, we have also confirmed the methylation and arrangement patterns of the centromeric region, completed the high-quality draft assembly of the homologous centromeres of chimpanzee, orangutan, and macaque chromosome 8. This resolved the evolutionary pattern of the centromeric region.

Finally, the T2T consortium first completed the string graph of the whole human genome using PacBio HiFi data with high precision and long reading length. The string graph was then analyzed using ONT data and other technologies to construct 22 T2T genomes. The efficiency and accuracy of this assembly method were evaluated.
In the future, the Human Pangenome Reference Consortium will sequence the genomes of more than 300 individuals from different races to understand the genetic diversity of different races, individuals, and provide greater support for future precision medicine goals.

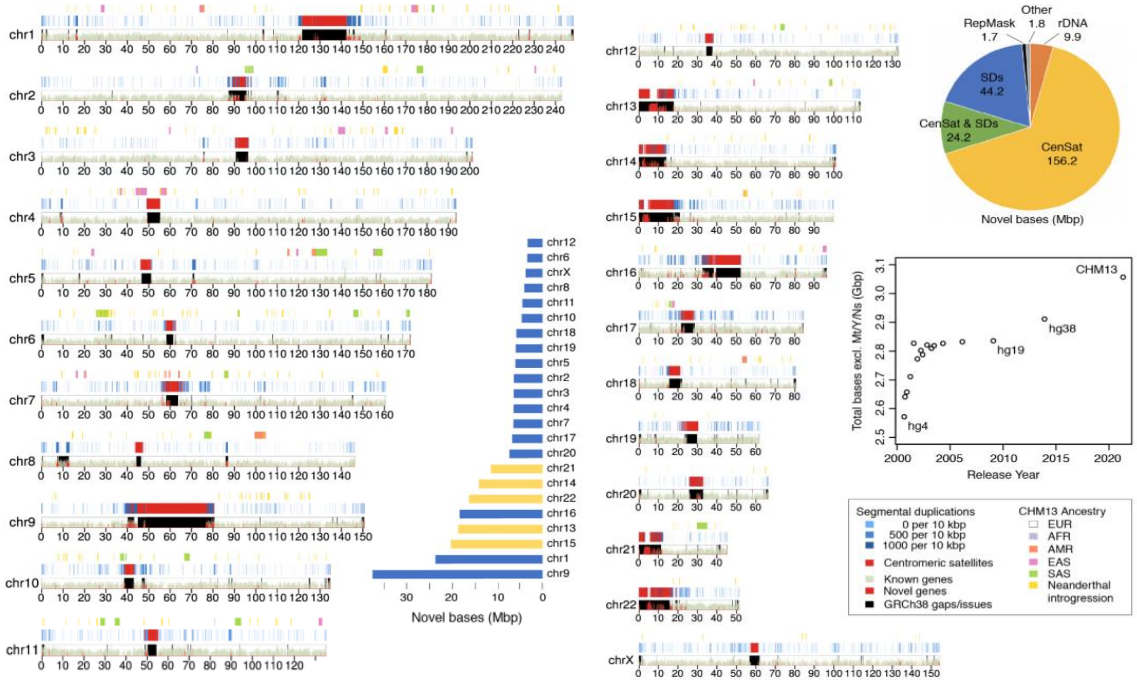

Figure 1. Human CHM13 genome.

Case 2. The genetic and epigenetic landscape of the Arabidopsis centromeres[4]

The initial assembly of Arabidopsis chromosomes was performed using ONT Ultra-long, followed by scaffold assembly, and correction using PacBio HiFi data. The final genome version was named Col-GEN v1.2. Arabidopsis near-complete figure containing five centromeres: chromosomes 1,3 and 5 contain complete telomere-to-telomere sequences, while chromosomes 2 and 4 remain unassembled in the 45s rDNA region of the short arm and the adjacent telomere region.

Satellite arrays consisting of millions of bases are present in the centrism of Arabidopsis, in which single satellite repeats are about 180bp in length, so these sequences are called CEN180. By identifying the CEN180 sequence on the centromere, the researchers found that the CEN180 sequence on different chromosomes was significantly different, while the CEN180 sequence within the same chromosome showed a trend of homogenization. The CEN180 content of centromor on chromosome 5 in Arabidopsis is very low (12% to 22% of other centromor levels). Further studies showed that the invasion of the retrotransposon ATHILA in this region promotes the evolution of CEN180 sequence and the organization

pattern of epigenetic modifications. It also affected the homogenization of CEN180 sequence in the centromere of chromosome 5 in Arabidopsis.

Finally, researchers proposed a recombination-based homogenization model of CEN180: the centromere repeat region of Arabidopsis was a homogenization process, but the invasion of the retrotransposon ATHILA led to the diversification of CEN180 sequence, which jointly promoted the evolution of centromere structure and function

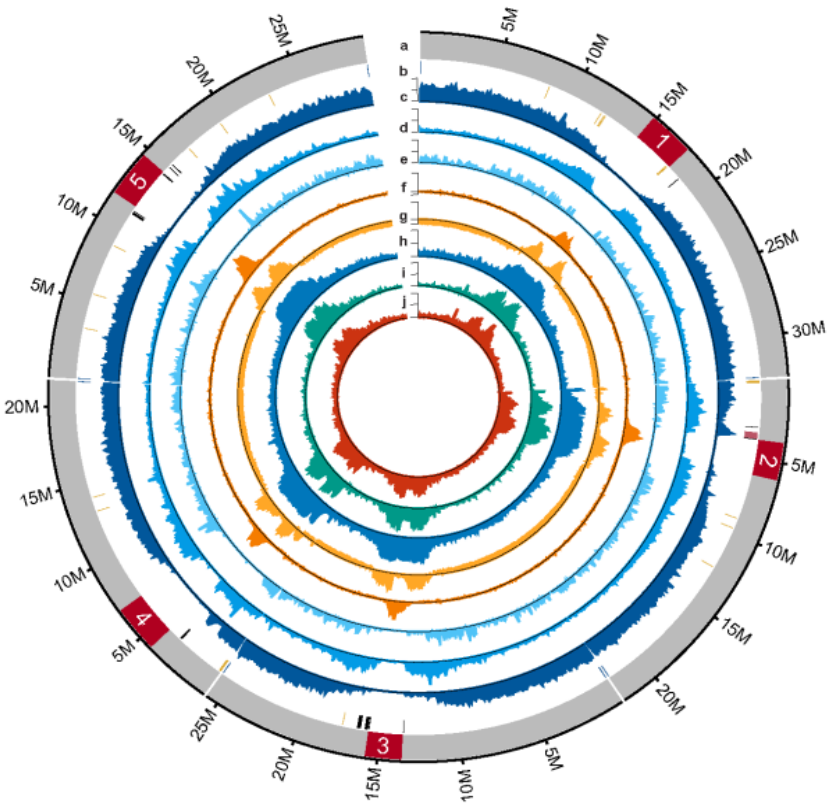

Figure 2. Complete map of the Arabidopsis genome.

In addition, near-complete genome maps of many species have also been published, such as rice, watermelon, maize, barley, etc. Some chromosomes have reached the T2T level, but some chromosomal telomeric regions or rDNA repeat regions still have assembly defects.[1] [5-8]

Research Methods

T2T genome assembly combines the advantages of a variety of sequencing technologies, which can sequence complex regions such as repetitive sequences, telomeres and centromeric, significantly increase the integrity of genome sequencing sequence, reduce the Gap region in the assembled genome, and provide reference sequence level genomes for further study of the origin, evolution, traits, and characteristics of species. At present, the published T2T genome is mostly based on PacBio HiFi + ONT Ultra-long + Hi-C + NGS multiple sequencing strategies, which are characterized by the following:

PacBio HiFi: Because of the long read length and high quality of HiFi data, there is no error correction in the genome assembly, which can achieve rapid genome assembly. It can be applied to different genome types, and higher quality and continuity of genomes can be obtained in the assembly of high-repeat genomes, large complex genomes, allopolyploid genomes, and even autopolyploid genomes. For heterozygous genomes, two sets of haplotype genomes can be assembled. In terms of genome continuity, the Contig N50 of HIFI may be slightly lower than that of CLR mode because the read length is shorter than that of CLR, but its completeness is better than that of CLR.

ONT Ultra-long: ONT Ultra-long sequencing can generate very long sequencing fragments that easily span the large repetitive regions in the genome. It can significantly improve the assembly effect of species genome and fill the gaps in the genome, making T2T (Telomere to Telomere) gap-free genome assembly possible.

Hi-C: Using high-throughput sequencing technology, combined with bioinformatics analysis methods, to study the spatial position of the whole chromatin DNA on a genome-wide scale and obtain high-resolution three-dimensional chromatin structure information.

Reference strategy

Genome assembly usually includes DNBSEQ™ based genome survey, PacBio HiFi genome assembly, Hi-C auxiliary assembly. For the T2T genome, assembling also needs combining with ONT Ultra-long sequencing, which contains the advantage of Ultra-long data.

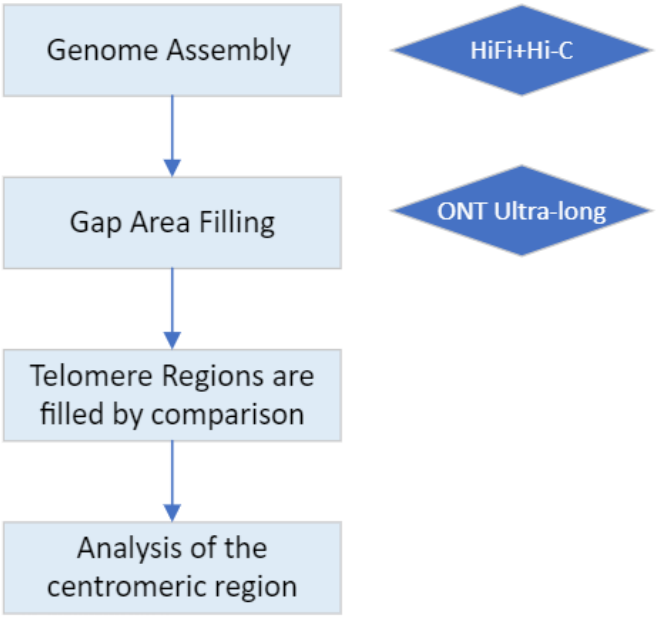

Figure 3. Main steps of T2T assembly.

Therefore, for different species and different levels of complexity of the genome, we recommend a variety of comprehensive solutions:

Research Methods	Sequencing technology	Platform	Depth
Genome Assembly	PacBio HiFi	PacBio	40-60X
	ONT Ultra-long	ONT	40 -100X
Auxiliary Assembly	Hi-C	DNBSEQ	100x

T2T strategy for common genomes:

1. HiFi data are assembled by Hifiasm software ----Contig level

2. Juicer processed Hi-C data and 3D-DNA software is used for scaffolding----Super-scaffold level

3. The LR_Gapcloser software is used to “fill holes” combined with ONT Ultra-long data----nearly complete T2T

4. tidk software is used for telomere identification

(The above steps are for reference only.)

T2T genomes are currently targeted at mammals and common genomes below 1.5 Gb (such as rice, soybean, etc.). Other species need to be evaluated to adopt different assembly strategies based on different species’ characteristics.

After the completion of T2T genome assembly (genome completion map), if it is necessary to verify the integrity of telomere and centromere assembly, it is usually necessary to combine a variety of methods to verify the centromeric and telomere regions, including FISH, ChIP-Seq, etc.

Bioinformatics Analysis Content

Bioinformatics service content	
Genome Assembly (PacBio HiFi data + ONT Ultra-long data+ Hi-C data)	1. Assembly 2. Assessment by short reads alignment 3. BUSCO assessment 4. Auxiliary Assembly Hi-C data auxiliary assembly 5. Identification of telomere region
Gene Annotation	1. Repeat annotation 2. Gene structure annotation 3. Gene function Annotation 4. Transcription factors (plant)
Evolution	1. Gene family identification (Animal TreeFam; Plant OrthoMCL; ≤10 species); 2. Phylogenetic tree construction 3. Estimation of divergence time 4. Genome synteny analysis 5. Whole genome duplication analysis 6. Gene family expansion and contraction analysis

References

[1] Miga, K. H. et al. Telomere-to-telomere assembly of a complete human X chromosome. Nature 585, 79–84 (2020).
[2] Logsdon, G. A. et al. The structure, function and evolution of a complete human chromosome 8. Nature 593, 101–107 (2021).
[3] Nurk, S. et al. The complete sequence of a human genome. Science 376, 44–53 (2022).
[4] Naish, M. et al. The genetic and epigenetic landscape of the Arabidopsis centromeres. Science 374, eabi7489 (2021).
[5] Deng, Y. et al. A telomere-to-telomere gap-free reference genome of watermelon and its mutation library provide important resources for gene discovery and breeding. Mol. Plant S1674205222001927 (2022).
[6] Li, K. et al. Gapless indica rice genome reveals synergistic contributions of active transposable elements and segmental duplications to rice genome evolution. Mol. Plant 14, 1745–1756 (2021).
[7] Song, J.-M. et al. Two gap-free reference genomes and a global view of the centromere architecture in rice. Mol. Plant 14, 1757–1767 (2021).
[8] Zhang, Y. et al. The telomere - to - telomere gap - free genome of four rice parents reveals SV and PAV patterns in hybrid rice breeding. Plant Biotechnol. J. pbi.13880 (2022).

www.bgi.com
info@bgi.com

BGI Americas

One Broadway, 14th Floor
Cambridge, MA 02142
U.S.A.

BGI Japan

Kobe KIMEC Center BLDG. 8F 1-5-2
Minatojima-minamimachi,
Chuo-ku, Kobe 650-0047 Japan

BGI Europe

Jutrzenki 12 A,
02-230 Warszawa,
Poland

BGI Australia

L6,CBCRC, 300 Heston Road,
erston, Brisbane,
Queensland 4006, Australia

BGI Asia

Building NO.7, BGI Park,
Yantian District Shenzhen,
Guangdong Province, China