A complete, telomere-to-telomere human genome sequence presents new opportunities for evolutionary genomics

The release of the first telomere-to-telomere (T2T) human genome sequence marks a milestone for human genomics research and holds promise of complete genomes for evolutionary genomic studies. Here we describe the advances that this new human genome assembly represents and explore the potential insights that the complete genome sequence could bring to evolutionary genomics. We also discuss the potential challenges to be faced in applying this new sequencing strategy to a broad spectrum of extant species.

Yafei Mao and Guojie Zhang

A complete genome assembly should in theory cover every base pair and allow the ordering of all sequences into individual chromosomes. Producing a gapless, complete human genome sequence has long been a goal for biologists, but has remained a pipe dream until now[1]. The first human genome sequence, released in 2001 by the Human Genome Project, was incomplete and included millions of unknown bases[2,3]. The greatest challenge in filling these gaps is the highly repetitive sequences spanning up to several megabases that are dispersed throughout the genome and cannot be accurately resolved with short-read sequencing technology[1,4]. These unresolved regions, which include centromeres, telomeres, tandem repeat arrays, segmental duplications and the p-arms of the acrocentric chromosomes, account for over 8% of the latest human reference genome (GRCh38). Owing to their complex genomic structures, many of the repetitive regions in the GRCh38 are either missing or incorrectly assembled, greatly limiting our understanding of their structural composition, epigenetic regulation, biological functions, population variation and evolution[1,5–9].

With advances in long-read sequencing technologies and computational algorithms, the Telomere-to-Telomere (T2T) Consortium has finally succeeded in filling all of the remaining gaps to produce the first complete human genome assembly (T2T-CHM13)[1] (Fig. 1a). This complete human genome sequence allows us to investigate the biological functions of regions that were previously inaccessible. With further development, the T2T sequencing strategy and pipeline can also be applied to large human populations, as well as to other species, and hence should open a new era of comparative genomic study in the coming years. Here we explore the potential areas to which this new technology, and complete genomes of every individual and species, could be applied.

Turning unknowns to knowns

Highly repetitive and/or gapped regions are usually excluded from population genetic and comparative genomic studies. It follows that a complete genome will provide unique opportunities to investigate both the variation and evolution of these previously unresolved regions. The release of the T2T-CHM13 genome has revealed the composition of these hitherto mysterious genomic regions. Surprisingly, despite extensive efforts to annotate the human reference gene catalog, nearly 100 protein-coding genes were newly detected in these regions, including several genes associated with neurodevelopmental disorders, such as *GPRIN2B* and *TBC1D3* (refs. [1,9]). The complex nature of these regions often results from complicated evolutionary scenarios, which can only be unraveled with comparisons of complete genome sequences. Therefore, we anticipate that replacing the current human reference genome with the T2T-CHM13 genome assembly in primate genome comparison studies could provide insights into the genomic innovations associated with the origins of human-specific traits[10,11] (Fig. 1b).

A complete human reference genome not only increases the number of detected variants but also greatly reduces the rate of inaccurate variant detection[5] (Fig. 1c). Artifacts in variant discovery are the main source of systematic bias for genome-wide association studies, selection estimation and disease-related variant discovery[12,13]. The complete assembly of pericentromeres and centromeres also allows us to characterize their variation and structure between and among entire chromosomes. These efforts have expanded our knowledge of kinetochore location and centromere evolution in primates[6,10].

Filling the gaps in genomic evolution

Although a standard T2T genome pipeline is not yet available for all non-human species, we expect that a complete genome will become the gold standard for future biodiversity genomics projects[14,15]. For example, the Vertebrate Genomes Project aims to generate near error-free genome assemblies for ~70,000 extant vertebrate species and has already produced over 150 high-quality reference genomes[14]. Although most of these reference genomes still contain some gaps and are not as complete as the T2T genome, the effort represents an important step toward achieving the goal of producing a complete genome for every vertebrate species. Once more T2T genomes have been generated, the cross-population and cross-species comparison of these complete genomes may be expected to address many evolutionary questions that remain challenging.

Comparison of complete genomes will broaden our knowledge of the function and evolution of repetitive elements. Comparative genomic studies have so far mainly focused on the evolution of genes and regulatory elements. To what extent the repetitive elements also contribute to phenotypic innovation remains less explored. It is well established that repetitive elements represent a significant source of genetic diversity and regulatory variation[16–18]. As selfish genetic entities, repetitive elements also function as mutators in evolution, harboring a broad

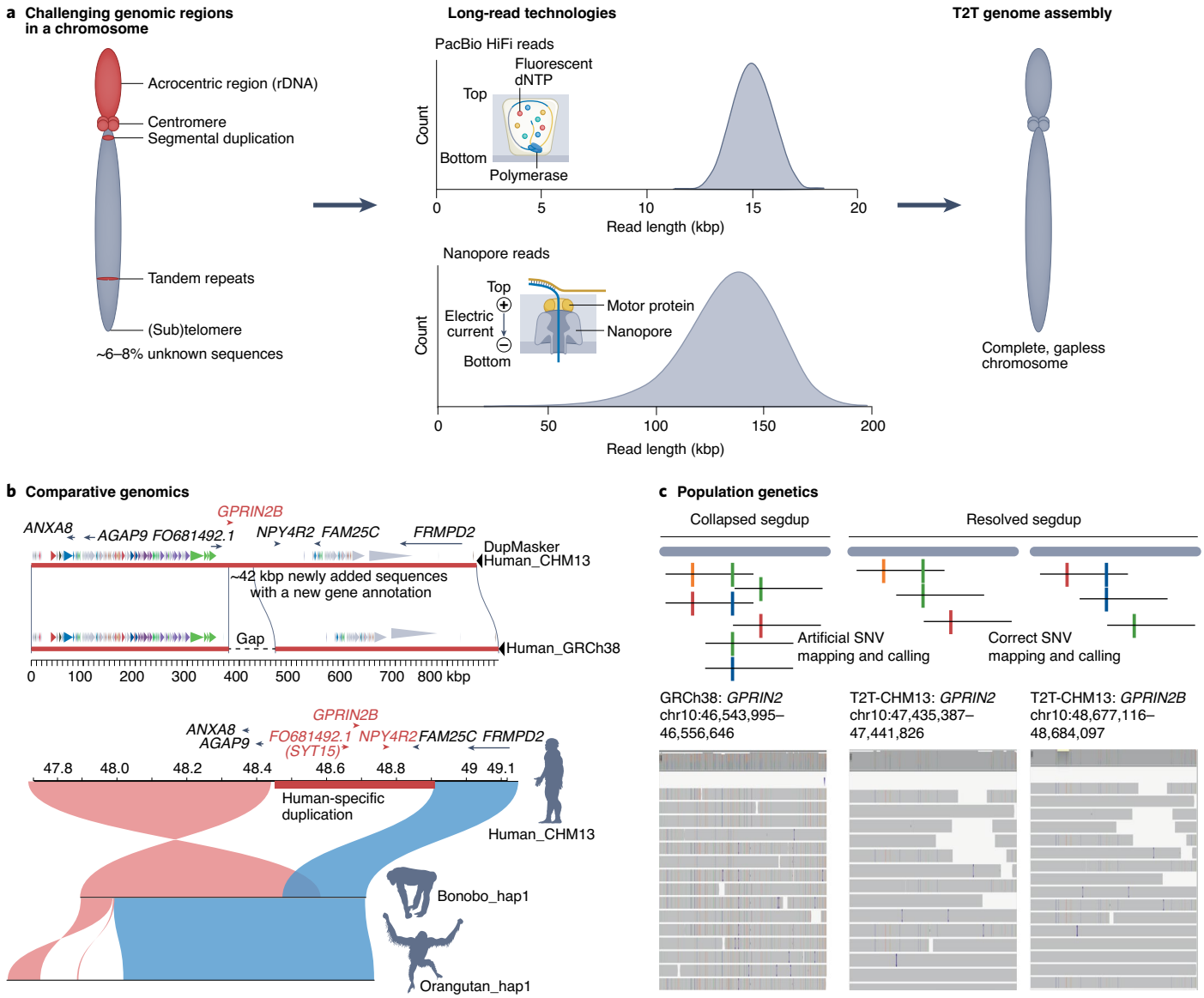

Fig. 1 | A complete telomere-to-telomere (T2T) human genome sequence potentiates new investigations in comparative genomics and population genetics. **a**, Left, previously unresolved regions in the human reference genome (red) total an estimated ~6–8% of the sequence. Middle, raw-read length distributions from PacBio and ONT technologies. Right, a complete T2T chromosome assembled with long-read sequencing technologies. **b**, A T2T genome reveals a new human-specific gene, *GPRIN2B*, residing within a gapped region in GRCh38 in the top panel. Synteny comparison among human, bonobo and orangutan genomes shows a human-specific region that includes *SYT15*, *GPRIN2B* and *NPY4R2* genes (bottom). Without the T2T genome effort, this region would have been missed in comparative genomic studies. **c**, A T2T genome reduces incorrect paralogous variant mapping and calling. Top, SNVs are represented by red, orange, blue and green vertical lines. Bottom, an example of read mapping over the *GRPIN2B* locus reveals how the complete T2T genome sequence enhances the sensitivity of paralogous read mapping and SNV calling. Segdup, segmental duplication.

spectrum of mutations large and small[19,20]. Therefore, these elements are important in both gene and genome evolution[20–22]. The complete assembly of these repetitive sequences should not only reveal their complex evolutionary processes but also shed light on their potential regulatory roles. Further, the processes by which repetitive elements (for example, higher order repeats in centromeres) evolve are different than those of other genomic regions[6,9,10,17,23]. Thus, complete genomes will facilitate studies of mutation patterns, as well as mechanisms of repetitive element evolution, between and among populations or species.

The availability of T2T genomes will open up new avenues in the investigation of molecular mechanisms of chromosome evolution. Chromosome fusion and fission events often occur during the speciation process[24–26]. Indeed, it has been proposed that the chromosome rearrangements mediated by repetitive elements have been involved in both speciation and the evolutionary divergence of different organisms[26–28]. Comparison of complete, T2T genomes should enable the determination of rearrangement breakpoints at the single-base-pair level, which would be invaluable for the investigation of the underlying mutational mechanisms, potentially including recombination, non-allelic homologous recombination,

mobile element association and segmental duplication association, among others[17,25,26]. T2T genomes will provide opportunities to study the processes involved in telomeric repeat gains and losses, as well as the repositioning of centromeres during evolution[6,10,17]. The T2T genome assembly should also potentiate the determination of chromatin structure and methylation patterns at the whole-genome level. This should enable us to evaluate how changes in chromatin structure influence large-scale chromosome rearrangements and how these changes contribute to the modification of gene regulatory networks.

Genomic duplication is a major source of genetic innovation. Owing to their recent duplication or/and gene conversion, a large proportion of tandemly duplicated genes are highly similar in sequence and are collapsed in the current genome assembly strategies[4,9,29]. The complete genome assembly increases the detection power for duplication events, particularly for lineage-specific duplications. For example, 182 new duplicated genes are found in the T2T-CHM13 genome, including *GPRIN2B* and *TBC1D3*, which are both human-specific genes[9] (Fig. 1b). Newly duplicated genes often retain subfunctions of their ancestral genes or have evolved neofunctionality by accumulating either mutations or modifications in regulatory regions leading to their functional divergence. Complete genomes will assist in the determination of the different methylation patterns on the duplicated genes and the exploration of the epigenetic fates of duplicated genes[6,7,9].

The T2T genome assembly strategy should also allow us to fully decode the sex-specific chromosomes (Y and W). Sex chromosomes represent some of the most challenging portions of vertebrate genomes to sequence[30–32]. As a result, they have often been omitted in most genome sequencing projects. The Y and W chromosomes are particularly difficult to sequence and assemble because they tend to accumulate a high abundance of repetitive sequences as a result of the suppression of recombination between sex chromosome homologs. The much poorer sequencing coverage of these chromosomes in genome projects belies their key evolutionary significances. The Y and W chromosomes often carry the sex determination locus and are subject to unique evolutionary pressures because of their sex-limited inheritance[31,32]. Decoding the T2T sequence of these chromosomes should provide important insights into the evolutionary history of sex chromosomes and their unique contributions to speciation and lineage-specific adaptation[33,34]. The current T2T-CHM13 genome sequence does not include the Y chromosome, but some promising progress has recently been made towards the production of a near-complete Y chromosome sequence with long-read technologies[31,32]. The trio-binning strategy has been demonstrated to be an efficient way to produce a near-complete Y chromosome, including the pseudoautosomal region, which has often been collapsed in X chromosome assemblies[31].

Finally, population-level phased T2T genomes promise to provide the ultimate solution for reference-free haplotype phase determination[35] and can substantially improve the evolutionary study of haplotypes, particularly for those highly variable genomic regions such as the major histocompatibility complex, which carries the most, and some of the longest, recognized disease-associated loci. These regions often harbor essential genes that contribute to the adaptive evolution of organisms[11,36].

Challenges ahead

While the release of the T2T- CHM13 genome sequence represents a tremendous success in the field, the routine assembly pipeline of T2T genomes that can be applied to any species has not yet been achieved. This is for a myriad of reasons. First, the current T2T- CHM13 genome was produced from a complete hydatidiform mole cell line that carries a diploid genome replicated from one set of chromosomes. Producing a complete genome that in reality only represents a 'haploid' genome is easier than the assembly of a natural diploid genome that includes two sets of non-identical chromosomes. It remains challenging to produce a haplotype-resolved diploid genome at the T2T level[30]. Second, different species have unique genomic features, particularly in relation to repeat families and centromeric repeats, which remain poorly annotated and require time-consuming manual assignment during the assembly process in non-human species[6,8,10]. Owing to differential repeat expansion and other duplication events, some species have much larger genome sizes than humans. For instance, the marbled lungfish has a genome size of ~130 Gbp, over 40 times larger than that of the human[37,38]. According to the genome size database, over 25% of vertebrate species have genome sizes that are larger than that of the human[14,39]. It remains to be seen whether the current T2T strategy will also work on species with extremely large or polyploid genomes. This assembly strategy combines several different sequencing technologies and requires a huge amount of high-quality biomaterial, which may pose a challenge[14,15]. Further improvement of sequencing technologies that allow low-input biomass at lower cost will make the sequencing strategy more feasible in practical terms for most species.

Despite an expectation to replace the GRCh38 assembly with T2T-CHM13 assembly as the human reference genome[1,5–10], the full utility of this new reference genome relies on the completeness of functional annotations, which include gene annotations (GENCODE_V39 and RefSeq), gene expression patterns and regulation (ENCODE, Hi-C, and single-cell RNA-seq), disease-related and common variants (gnomAD variants, dbSNP, CADD, ClinGen, ClinVar, HGMD, evolutionary conservation and others)[40]. Comprehensive and accurate annotations are essential for comparative genomics and biomedical genetics studies. Liftover of these annotations from one genome version to another is an easy way to address this deficiency; however, this is not possible for newly sequenced regions and the highly variable regions[1,8]. Therefore, de novo annotation of the T2T-CHM13 genome assembly and other complete human genomes is an essential step in realizing their full utility. An efficient annotation tool is required to overcome the considerable computational resources and labor costs required for current annotation pipelines.

Since additional complete human genomes with diverse population genetic backgrounds will soon be released[41], pan-genomic analyses will require an efficient reference-free alignment. The progressive Cactus aligner has shown itself capable of producing high-accuracy reference-free multiple alignments for thousands of vertebrate genomes[35]. This aligner can add new genomes to an existing multiple alignment. However, running this tool requires considerable computational resources, and improvements to the algorithm would be needed for large-scale genomic data. Another alignment challenge is the comparison of repetitive sequences, which are often masked before multiple alignments. Current repeat aligners still do not perform well on the highly repetitive sequences[42,43]. The development of more sophisticated sequence aligners specific for repetitive sequences will be required to study the evolutionary processes and selection forces acting on these regions. We also lack the ability to measure the variation among repeats (for example, higher order repeats in centromeres), in which all variants, including single nucleotide variants, insertion/deletions (indels), copy number variants, orientation switches and turnover rate, should be considered.

Various large-scale biodiversity genome projects[14,15,44–48], such as the Vertebrate Genomes Project[14], Bird 10,000 Genomes[44], Global Ant Genome Project[45], Bat1K Project[47], Primate Genome Project[48] and Earth BioGenome Project[15,46], are underway. Given the considerable cost of T2T genome assembly compared to the funding situation experienced by some other fields, it is as yet impossible to produce a complete genome for every species. It is, therefore, important to prioritize the assembly of rare or endangered species and some taxon-representative species at the T2T level. Before the sequencing cost is reduced to an affordable level for individual researchers and the required biomass is reduced to a feasible level, a high-quality draft reference genome will still represent a rational solution for most biodiversity genomic research[14,49,50]. But with the rapid development of sequencing technologies and the continued improvements in the computational tools, we can anticipate a bright future for T2T genomes in biodiversity research. ❐

Yafei Mao [1,2] ✉ and Guojie Zhang [3,4,5,6] ✉

[1]Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders, Ministry of Education, Shanghai Jiao Tong University, Shanghai, China. [2]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA. [3]Evolutionary & Organismal Biology Research Center, Zhejiang University School of Medicine, Hangzhou, China. [4]Liangzhu Laboratory, Zhejiang University Medical Center, Hangzhou, China. [5]Villum Center for Biodiversity Genomics, Section for Ecology and Evolution, Department of Biology, University of Copenhagen, Copenhagen, Denmark. [6]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China.
✉*e-mail: yafmao@sjtu.edu.cn; guojiezhang@zju.edu.cn*

Published online: 10 June 2022
https://doi.org/10.1038/s41592-022-01512-4

References

1. Nurk, S. et al. *Science* **376**, 44–53 (2022).
2. Lander, E. S. et al. *Nature* **409**, 860–921 (2001).
3. Venter, J. C. et al. *Science* **291**, 1304–1351 (2001).
4. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. *Nat. Rev. Genet.* **21**, 597–614 (2020).
5. Aganezov, S. et al. *Science* **376**, eabl3533 (2022).
6. Altemose, N. et al. *Science* **376**, eabl4178 (2022).
7. Gershman, A. et al. *Science* **376**, eabj5089 (2022).
8. Hoyt, S. J. et al. *Science* **376**, eabk3112 (2022).
9. Vollger, M. R. et al. *Science* **376**, eabj6965 (2022).
10. Logsdon, G. A. et al. *Nature* **593**, 101–107 (2021).
11. Mao, Y. et al. *Nature* **594**, 77–81 (2021).
12. Li, H. *Bioinformatics* **30**, 2843–2851 (2014).
13. Qi, J., Chen, Y., Copenhaver, G. P. & Ma, H. *Proc. Natl Acad. Sci. USA* **111**, 10007–10012 (2014).
14. Rhie, A. et al. *Nature* **592**, 737–746 (2021).
15. Lawniczak, M. K. et al. *Proc. Natl Acad. Sci. USA* **119**, e2115639118 (2022).
16. Wolffe, A. P. & Matzke, M. A. *Science* **286**, 481–486 (1999).
17. O'Neill, R. J., Eldridge, M. D. & Metcalfe, C. J. *J. Hered.* **95**, 375–381 (2004).
18. Bodega, B. & Orlando, V. *Curr. Opin. Cell Biol.* **31**, 67–73 (2014).
19. Kidwell, M. G. & Lisch, D. R. *Evolution* **55**, 1–24 (2001).
20. Kashi, Y. & King, D. G. *Trends Genet.* **22**, 253–259 (2006).
21. Soltis, P. S., Marchant, D. B., Van de Peer, Y. & Soltis, D. E. *Curr. Opin. Genet. Dev.* **35**, 119–125 (2015).
22. Xia, B. et al. Preprint at *bioRxiv* https://doi.org/10.1101/2021.09.14.460388 (2021).
23. Smith, G. P. *Science* **191**, 528–535 (1976).
24. Rieseberg, L. H. *Trends Ecol. Evol.* **16**, 351–358 (2001).
25. Raskina, O., Barber, J. C., Nevo, E. & Belyayev, A. *Cytogenet. Genome Res.* **120**, 351–357 (2008).
26. Fuller, Z. L., Koury, S. A., Phadnis, N. & Schaeffer, S. W. *Mol. Ecol.* **28**, 1283–1301 (2019).
27. Ventura, M., Archidiacono, N. & Rocchi, M. *Genome Res.* **11**, 595–599 (2001).
28. Carbone, L. et al. *Nature* **513**, 195–201 (2014).
29. Vollger, M. R. et al. *Nat. Methods* **16**, 88–94 (2019).
30. Jarvis, E. D. et al. Preprint at *bioRxiv* https://doi.org/10.1101/2022.03.06.483034 (2022).
31. Yang, C. et al. *Nature* **594**, 227–233 (2021).
32. Zhou, Y. et al. *Nature* **592**, 756–762 (2021).
33. Chen, S. et al. *Nat. Genet.* **46**, 253–260 (2014).
34. Wang, Z. et al. *J. Genet. Genomics* **49**, 109–119 (2022).
35. Armstrong, J. et al. *Nature* **587**, 246–251 (2020).
36. Zhou, F. et al. *Nat. Genet.* **48**, 740–746 (2016).
37. Meyer, A. et al. *Nature* **590**, 284–289 (2021).
38. Wang, K. et al. *Cell* **184**, 1362–1376.e1318 (2021).
39. Pellicer, J., Hidalgo, O., Dodsworth, S. & Leitch, I. J. *Genes (Basel)* **9**, 88 (2018).
40. Navarro Gonzalez, J. et al. *Nucleic Acids Res.* **49**(D1), D1046–D1057 (2021).
41. Miga, K. H. & Wang, T. *Annu. Rev. Genomics Hum. Genet.* **22**, 81–102 (2021).
42. Li, H. *Bioinformatics* **34**, 3094–3100 (2018).
43. Ren, J. & Chaisson, M. J. P. *PLOS Comput. Biol.* **17**, e1009078 (2021).
44. Feng, S. et al. *Nature* **587**, 252–257 (2020).
45. Boomsma, J. J. et al. *Myrmecol. News* **25**, 61–66 (2017).
46. Lewin, H. A. et al. *Proc. Natl Acad. Sci. USA* **115**, 4325–4333 (2018).
47. Jebb, D. et al. *Nature* **583**, 578–584 (2020).
48. Wu, D.-D. et al. *Zool. Res.* **43**, 147–149 (2022).
49. Stiller, J. & Zhang, G. *Diversity (Basel)* **11**, 115 (2019).
50. Formenti, G. et al. *Trends Ecol. Evol.* **37**, 197 (2022).

Acknowledgements

We acknowledge valuable comments from Glennis A. Logsdon (University of Washington School of Medicine). This work was supported by International Partnership Program of Chinese Academy of Sciences (no. 152453KYSB20170002) and a Villum Investigator Grant (no. 25900) from the Villum Foundation to G.Z.

Author contributions

Y.M. and G.Z. conceived the project. Y.M. and G.Z. contributed to the writing.

Competing interests

The authors declare no competing interests.