# Completing human genomes

*Nature Methods* is pleased to publish several papers presenting methods developed by members of the Telomere-to-Telomere (T2T) Consortium, which facilitated the generation and analysis of the first complete human genome.

The genome sequence of a species not only provides the fundamental basis for the genomics field, but also bears fruit for fields beyond genomics. Proteomics, genome editing and medical genetics studies all benefit from the resource of a high-quality genome sequence. Although the existing human reference genome has been widely used and richly annotated, it is not complete, with gaps and errors in many highly repetitive and complex genomic regions such as centromeres and segmental duplications. The recalcitrance of those regions to traditional sequencing strategies poses one of the major challenges in genomic research, hindering a full understanding of the human genome.

Method development has long been one of the driving forces for reading genome sequences with even higher accuracy. While short-read sequencing is still a workhorse of many genome sequencing projects, long-read sequencing and de novo genome assembly approaches are increasingly gaining traction. Despite their power to improve the quality of genome sequences and uncover sequences missing in previous genome assemblies, each currently available sequencing technology still has its own weakness. The endeavor of devising and implementing new experimental and computational strategies calls for collaboration by the community via initiatives such as the T2T Consortium.

Since its inception, the T2T Consortium has been releasing and sharing the data they've generated. Recently, after several years' work, the consortium published a series of papers, with the main one describing the first complete human genome. As a journal devoted to publishing methods and tools for the life sciences, we are pleased to publish a part of this package in this issue, highlighting the efforts of method development by the consortium towards the generation of the first complete human genome.

In their Article, Ann Mc Cartney and colleagues provide a detailed description of the bioinformatic strategy applied to evaluate and polish the T2T genome assembly. This strategy uses a semiautomatic workflow to identify and remove errors while minimizing overcorrection. Leveraging their experience from building this workflow, the authors also suggest best practices for other genome sequencing projects. An accompanying News & Views from Li Fang and Kai Wang provides an overview of this work and highlights its impact. Another Article, by Giulio Formenti and colleagues, presents Merfin, a *k*-mer-based variant-filtering method that is used in the above workflow. Winnowmap2, a computational tool for mapping long-read data, is described in an Article by Chirag Jain and colleagues. It achieves enhanced performances in repetitive regions and improves downstream variant calling, which is also used in the pipeline. These and other existing and newly developed computational methods taken together boosted the quality of the final T2T genome assembly.

The consortium also developed experimental methods to take full use of the rich data they generated for functional genomics analysis. One example, DiMeLo-seq, is a method developed by Nicolas Altemose and colleagues for profiling chromatin-binding proteins and histone modifications on the basis of long-read data. In a News & Views, Kami Ahmad explains why this method will be useful in studying genome function. We hope the methods developed by the consortium will further aid research in genome assemblies and functional genomics. With the potential of this work to reach other research areas, a wave of method development inspired by the new T2T data may well be on the horizon.

The series of T2T Consortium papers also includes a comprehensive analysis of the complete human genome from multiple aspects, such as centromeres, repeat elements and epigenetic patterns. Furthermore, the complete genome facilitates accurate identification of small and large genetic variants in samples of different ancestries, including in medically relevant genes. These and many other applications showcase how broadly enabling the T2T resource is, enticing exploration by various communities.

Yafei Mao and Guojie Zhang discuss the potential that a complete genome will bring about in the area of evolutional genomics in a Comment. New complete genomes fill important gaps in the previous reference genomes, which can shine more light on the evolution of chromosome structure, genomic duplication and sex-specific chromosomes. Further, the areas of population and evolutionary genomics welcome similar high-quality genome sequences in different populations and species to uncover the full spectrum of intra- and interspecific genomic diversity.

Echoing this need, a few large community-initiated projects or consortia such as the Vertebrate Genomes Project and the Human Pangenome Reference Consortium are also setting out to generate high-quality genomes. Their goals, species of interest and scales vary, and are focused on various research topics. Owing to other practical considerations, the quality of genomes they aim to achieve is also different. Even so, the experience attained by the T2T consortium should help maximize other projects' potential. The best practices recommended by the T2T authors regarding implementation with limited resources should also apply to the vast number of individual sequencing projects.

The currently released T2T assembly (T2T-CHM13) is from a homozygous cell line. While this greatly simplifies many bioinformatics tasks, it also leaves the future goal of generating a complete phased genome from a heterozygous individual in the human population, which warrants further method development. New sequencing technologies will keep pushing the limit of longer read length, lower error, and higher throughput. Other approaches that are capable of using long-range and/or phasing information will also benefit. Advances in genomics informatics can accelerate progress in phased genome assembly with higher accuracy and speed. Large-scale population and phylogenomics studies will be energized by improved scalability and integrative analysis of genetic variation. In areas like medical genomics, the ability to better detect genetic variation in complex medically relevant regions holds potential for clinical translation. Researches targeting other -omes, such as the transcriptome and proteome, will thrive on the availability of complete genomes.

We are excited for the age of more complete genomes, and for the opportunities for more method development to accelerate this transformation. ❐

Published online: 10 June 2022
https://doi.org/10.1038/s41592-022-01537-9