Implications of the first complete human genome assembly

The Telomere-to-Telomere (T2T) Consortium has recently announced the assembly and analysis of the first complete human genome assembly. The use of the functionally haploid CHM13hTERT cell line (CHM13), originally isolated from a hydatidiform mole, as well as ultra-long Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio) high-fidelity (HiFi) data, has resulted in the gapless assemblies of all 22 autosomes plus Chromosome X, including the centromeric regions, the short arms of five acrocentric human chromosomes, and almost 200 million base pairs of novel sequence, potentially harboring protein-coding genes.

In anticipation of the initial publications, the Editors of *Genome Research* invited researchers in diverse fields to share with us their viewpoints on the implications of the T2T-CHM13 assembly. We asked them to elucidate the biological questions that perhaps can now be answered with the new complete assembly, especially in areas of their expertise, and why it is important from their academic perspectives.

Can Alkan[1]: The complete assembly generated by the T2T Consortium has several implications in my research, namely algorithm development for genomics. First of all, the gaps in the Genome Reference Consortium Human Build 38 (GRCh38) assembly are problematic because any reads that originated from those regions remain unmapped, or due to the repeats in those regions, the reads map to incorrect paralogs. This in turn causes both missing real variants due to lack of information and generation of false positive calls within the wrong paralog. The problem of lower accuracy in structural variation discovery is even more pronounced than that of single nucleotide variation and small indels. A reference genome with no gaps would therefore limit such mapping errors and improve both precision and recall of any genome variation discovery. Eliminating such problems is important in our pursuit for precision medicine. It was recently demonstrated that using the new T2T-CHM13 assembly significantly reduced false positives in hundreds of medically relevant genes. Thanks to the T2T-CHM13 assembly, we also now have access to previously uncharacterized sequences that remained as gaps in GRCh38. This newly accessible sequence contains both functionally relevant genes, regulatory sequences, and structurally important segments, such as the centromeric repeats. Finally, for the methods developer, T2T-CHM13 assembly and the sequencing and optical mapping data used to generate it provide a gold standard for genome variation discovery algorithms as well as genome assembly tools.

Lucia Carbone[2]: For someone like me, who has been spending most of their time digging into the dark matter of the genome (i.e., centromeric repeats and transposable elements), the possibility of finally analyzing the actual, decompressed sequences for these elements will be life changing. More importantly, the hope is that, similar to what happened after the human genome project, the new ultralong sequencing technologies, as well as the novel assembly methods developed for the human T2T genome, will soon also be applied to other species. The availability of T2T genomes from multiple species will enable a new level of comparative genomics analyses, finally involving the troublesome regions of the genome.

Megan Dennis[3]: My research group focuses on the functional significance of duplicated genes in human disease and evolution. Segmental duplications in assemblies are historically error-ridden, concealing potentially important genes and making the study of variation at these loci extremely challenging. When we first heard of the T2T Consortium, established to complete the first telomere-to-telomere assembly from a single individual, we were excited to contribute. Throughout my career, in collaboration with the Genome Reference Consortium, I have been fortunate to witness and contribute to the painstaking efforts to finish the final frontier of the reference genome, including nearly identical sequences within segmental duplications. Through the collective hard work of many individuals in diverse institutes and countries, using the technologies available at the time, the current human reference assembly (GRCh38) has enabled important genetic discoveries and ushered in a genomics era. The new T2T-CHM13 genome represents a significant advancement in this effort, leveraging improvements in long-read sequencing technologies to resolve some of the most recalcitrant loci (centromeres!), identify new genes, reduce artifacts, and eliminate false variant calls across previously collapsed segmental duplications. For myself and others, this means that we can now begin to assay variation formerly hidden across newly resolved complex regions and connect it to mechanisms of gene regulation, diversity of traits, and etiologies of complex diseases, such as autism spectrum disorder.

The T2T Consortium follows a long history of open science, comprising over 100 investigators from across the globe who share ideas, results, and data, demonstrating the efficacy of this model to advance human genomics. Moreover, this effort also signifies an improvement in inclusivity, with near equal representation of women and men on manuscripts as both lead and senior authors, as well as co-chair leadership. For me, as a "first gen" academic trying to find a place in the larger genomics field, inclusion in the T2T has imbued in me a sense of community, belonging, and hope for the future of my own trainees. This milestone, however, is just the beginning. Moving forward, continued momentum by the community toward increasing diversity of not only the genomes we sequence but also the scientists participating in these important endeavors is necessary. In doing so, there is no doubt that we will see even more remarkable achievements in the years to come.

Article published online before print. Article and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.276723.122. Freely available online through the *Genome Research* Open Access option.

© 2022 Cold Spring Harbor Laboratory Press This article, published in *Genome Research*, is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Jason Ernst[4]: A gapless telomere-to-telomere assembly of the human genome by the T2T Consortium represents a remarkable technical achievement with broad and important implications to the field of genomics. From the perspective of someone who has developed and applied a variety of approaches for systematic annotation of the human genome, including chromatin and conservation states, for a range of applications including within epigenomics, regulatory genomics, and noncoding variant interpretation, the gaps in the genome assembly had previously represented an impenetrable barrier to comprehensive genome annotation and a limitation for many applications. Thanks to the efforts of the T2T Consortium such portions of the genome are no longer inaccessible, enabling more comprehensive annotations and study of the genome and epigenome. The T2T Consortium has demonstrated that some short-read data from large-scale epigenome mapping efforts can be used to begin to annotate portions of these newly assembled regions of the genome, in addition to long-read data that are better optimized for studying repetitive sequences that dominate these regions. The T2T Consortium has shown how its assembly leads to a more comprehensive view of the epigenome, including of traditionally hard-to-study repetitive and centromeric regions, while in parallel showing that it leads to a more comprehensive identification of genetic variation, including some disease-relevant genetic variants. While the T2T efforts are already providing some immediate insights, the full impact of its assembly will likely be seen once it gains greater community adoption, which will take time given practical considerations associated with the extensive legacy annotations available on prior assemblies. Nonetheless, the completion of the T2T assembly represents a pivotal moment in genomics and additional layers of annotations on the newly assembled regions will further enhance our ability to study the genome and epigenome and their relationship with disease.

Gilad Evrony[5]: The T2T-CHM13 assembly is a starting point for a new era in genomics and a remarkable achievement that will advance both basic research and clinical genomics. The new assembly has finally resolved regions of the genome, such as large segmental duplications, that have been a thorn in the side of anyone who has analyzed human genome data, leading to false positive and missed variants. The new assembly is not only enabling new basic research discoveries about human genome structure and function—it will also improve our ability to identify disease-causing variants in these challenging regions of the genome.

This milestone also reflects a coming of age for long-read sequencing technologies that made the assembly possible. While these technologies will continue to advance, the full benefits of the T2T-CHM13 assembly will be realized only when long-read sequencing becomes routine. Broader adoption of the new assembly will also critically depend on porting to T2T-CHM13 the key existing genome annotations (such as population variant databases, variant pathogenicity predictions, and many more) that are required by most analytic pipelines, and creating entirely new annotations. Annotating the new T2T-CHM13 assembly with these resources is a major undertaking that will be important for the genomics community to prioritize. Without these annotations, it will take a long time for researchers and clinical laboratories to transition to the new assembly.

Beyond the new assembly itself, the methodologies developed by the T2T Consortium will undoubtedly now be applied and further developed to create many T2T assemblies, T2T pangenome references, and individualized fully diploid T2T assemblies. A T2T diploid assembly at birth (or even before birth) will be the foundation for the future of genomic medicine. In the hopefully not too far future, T2T assemblies of single cells may even become possible for studying mosaic variation in the genome's most complex regions.

Santhosh Girirajan[6]: My research is focused on identifying individual genes and their genetic interactions in functionally relevant networks that contribute to complex neurodevelopmental disorders, such as autism and intellectual disability. Although genetic studies on large cohorts of individuals with neurodevelopmental disorders have identified associations with rare variants in coding and noncoding regions, as well as individual and polygenic contributions of common variants, these studies have relied on genetic architecture and haplotype structure as understood from the perspective of a single human genome reference assembly. The efforts from the T2T Consortium will enable us to identify new categories of variants for more routine genetic association studies and clinical diagnosis beyond currently assessed variants, such as single-nucleotide variants, copy-number variants, and large chromosomal aberrations. For example, we will be able to identify complex structural variants, altered counts of short tandem repeats, inversions, and copy-number changes of duplicated genes more accurately in patient genomes. The T2T work is especially relevant for evaluating classes of genes that confer functional novelty to humans, such as cortical brain development, particularly those embedded in complex regions that have historically been hard to genotype for association studies. From a mechanistic point of view, the structure and orientation of segmental duplications and other repeat elements such as LINEs and SINEs would enable better prediction of risk for disease-associated genomic rearrangements. A complete T2T genome structure will also provide more accurate and subpopulation-specific haplotype structures for fine-mapping associations of variants to traits. We currently have an imperfect understanding of the genetic etiology of human disease, because we have been looking at an incomplete picture of the human genome.

A gapless telomere-to-telomere assembly of the human genome will identify genes and genetic segments that are currently missing, misassigned, or misassembled, accurately map repeat expansions that have been implicated in several neurological disorders, identify duplicated regions that are evolutionarily active and potentially confer phenotypic novelty, and, for the first time, provide an exquisite map of the sequence and structure of centromeres and telomeres. Overall, the reference assembly and the methodological tools generated by the T2T Consortium will be valuable to researchers and clinicians alike, who generate or rely on complete genome assemblies for their genetic studies.

Danny Chi Yeu Leung[7] and Clooney C.Y. Cheng[7]: Recently, researchers of the T2T Consortium generated a seamless reference of the human reference genome. By leveraging long-read sequencing technologies, the Consortium has patched the gaps in Chromosome X, Chromosome 8, and the entire genome of the essentially haploid CHM13hTERT cells. Notably, the newly assembled T2T-CHM13 reference human genome contains approximately 200 million new bases, 75%–90% of which are repetitive elements. Collectively, the communal efforts have elucidated the locations and identities of many repetitive elements, opening new avenues for understanding the genetic variation and epigenetic signatures of repeats in the human genome.

The T2T-CHM13 assembly has introduced a compendium of 62 novel repeat classes, greatly expanding the atlas of repetitive sequences in the human genome. The term "composite repeats" was coined to describe tandem arrays of three or more repetitive elements. Future studies to integrate functional analyses will shed light on how these composite repeats contribute to biological processes. Notably, the T2T Consortium also catalyzed the advancement of technologies and computational tools to study the structures, functions, and evolution of repeats. A method called Directed Methylation with Long-read sequencing (DiMeLo-seq) was developed and applied to map CENPA binding and DNA methylation of centromeric alpha-satellite higher-order repeats (HORs). In addition to confirming the enrichment of CENPA at the hypomethylated centromere dip region (CDR), the T2T Consortium further noted the single-molecule heterogeneity of CENPA localization and the DNA methylation level of alpha-satellite HORs at a resolution unattainable with short-read sequencing. Moreover, other new HORs were discovered at the pericentromeric regions, raising questions about the distinctions between these HORs and the active alpha-satellite HORs demarcating the centromere. Interestingly, the active HORs exhibit DNA hypomethylation, which suggests a possible interplay between transcription of active HORs, kinetochore assembly, and proper mitotic distribution of genetic materials. Further mechanistic and functional studies are still needed. Taken together, the advancements of the T2T Consortium will provide tools and references that in turn allow researchers to comprehensively analyze the regulation and function of repetitive elements.

David MacAlpine[8]: Every cell division cycle more than 6 billion base pairs, spanning the 23 pairs of human chromosomes, must be copied accurately and in their entirety. The complete copying of the genome within the confines of S-phase is accomplished by the selection and activation of 50,000–100,000 DNA replication origins throughout the genome. The temporal activation of these DNA replication origins establishes the DNA replication timing program which contributes to the maintenance of the epigenetic landscape. Modern genomic technologies have enabled investigators to map replication origins and to determine the time of replication for a given sequence and the relative direction of replication forks as they progress across the genome. However, these studies have been limited to unique sequences largely representing euchromatic regions of the genome. The T2T assembly of the human genome expands the sequence coverage of the human genome by more than 200 megabases and includes gapless assemblies of the repetitive pericentromeric sequences. While the satellite repeats making up the pericentromeric heterochromatin are generally thought to replicate late in S-phase, we do not know the relative density of replication origins, fork speed, or directionality of the replication forks. The expanded genome coverage provided by the T2T assembly coupled with advances in long-read technology and optical DNA mapping will enable researchers to address these and other longstanding questions which will ultimately provide invaluable insights into the mechanisms of genetic and epigenetic inheritance.

Ting Ni[9]: The T2T Consortium's complete human genome assembly has the potential to significantly impact aging research. Aging is associated with heterochromatin loss, which in turn causes cellular malfunction. Among the missing ~200 million bases in the GRCh38p13 version compared to the new CHM13v1.1 version, about 131 million bases are repeats. While LINEs and SINEs show a small percentage of increase from GRCh38p13 to CHM13v1.1, satellite repeats, simple repeats, and rDNA regions show 96.6%, 112.9%, and 730.4% increases, respectively. Whether the newly identified repeat sequences display similar epigenetic changes during aging or whether the newly identified rDNA units show different alteration patterns compared with other repeat regions are interesting questions that deserve further investigation. The additional ~200 million bases also contain 2226 paralogous genes, of which 115 are potentially novel protein-coding genes. It would be interesting to explore whether any aging-associated protein-coding or noncoding genes exist in the newly added sequence and whether any genetic variants located in these genes are associated with aging or age-related diseases. RNA-seq and functional screening identify increasing numbers of genes that show differential expression during aging or cellular senescence, some of which are key regulators of aging or age-related diseases and could serve as potential targets for delaying or even reversing aging-associated functional decline. By using long-read RNA sequencing strategies and the CHM13 assembly, novel protein-coding or noncoding genes that are differentially expressed between aging and control samples can be identified. By applying further functional validation, the list of genes (both coding and noncoding) that are relevant to aging may expand and lead to novel discoveries. Despite the potentially significant impact the new assembly may have on the aging research field, one should also understand that technical challenges at both experimental and computational levels still exist. To efficiently transfer such sequence information to aging-related knowledge necessitates the development of new methods and tools. Together with the new assembly, researchers will have new opportunities to explore and understand ourselves during aging, an unavoidable part of life.

Michèle Ramsay[10]: The first and long awaited gapless T2T human genome has been completed. This was a monumental task that took several years and a large budget despite the significant advances in DNA sequencing technologies, including long-read single-molecule sequencing, and many improvements in bioinformatics assembly tools. It is a genome from an essentially haploid cell line (CHM13), a feature that made it easier to link DNA sequences together in a linear set of chromosomes. The T2T genome includes highly repetitive DNA sequences at telomeres and centromeres of the 22 autosomes and the X Chromosome. Many previously unknown genes were identified and unresolved repetitive regions completed.

Working on African genomes reminds me daily of the enormous genetic diversity across the human species and how no two genomes are entirely alike. Many of the differences are not phenotypically relevant or disease-associated and the key is to understand when it matters. This has been a challenge since the first near-complete human genomes were revealed in the early 2000s and will continue to occupy researchers for decades to come. Having the first T2T assembly is an exciting milestone, demonstrating that it is possible. Through the process much was learned that will make the next T2T genomes faster and cheaper. However, it is important to reflect on whether the time, effort, and cost of producing T2T genomes at scale will provide the insights that we are hoping for. Is it feasible, or desirable, to do this for hundreds and thousands of people who each have two unique genomes (one maternal and one paternal in origin)? Likely there will be diminishing scientific returns for more genomes sequenced, but if chosen carefully the next examples could contribute greatly. For example, it is vital that genomes from people with diverse ancestries are included early on in the process, as they are likely to reveal the most variation and potentially novel scientific insights.

Not all regions of the genome are equally likely to be clinically relevant. To sequence T2T genomes of a few people for the same cost of sequencing the regions we already recognize as contributing to disease or disease susceptibility would not make sense, especially in low resource settings. I am supportive of the development of a Human Pangenome Reference, an initiative already in progress, to develop a complex system and database for capturing genomic variation from all populations in all regions of the genome but do not see large-scale T2T genome projects as a priority until algorithms for accurate assembly of T2T genomes from diploid organisms improve and sequencing technologies become more affordable.

Helen Rowe,[11] Poppy Gould,[11] and Rocio Enriquez-Gasca[11]: Here, we summarize how the T2T assembly will promote breakthrough discoveries in understanding how human disease susceptibility is linked to structural variation, within noncoding regions of the human genome. Most DNA in the human genome still has unknown functions and is referred to as "genomic dark matter." Recent work has centered on uncovering key roles for this DNA, which is mainly derived from transposable elements (TEs) that were once mobile and invaded our genome. TEs are now being unveiled to serve not only as gene regulatory elements, but also as self-derived nucleic acids that can be sensed by the human innate immune system to induce anti-viral type I interferons. Assigning function to the dark matter of our genome is a cutting-edge area of research, yet a major caveat in this field is that this enigmatic DNA often lies within unresolved parts of the genome. Genome gaps encompass satellite repeat arrays, ribosomal gene clusters, and regions of segmental duplications that represent an unexplored dimension of human population variation, impacting health and disease. With their highly repetitive nature, unresolved regions are a black box in terms of their sequence and activity. Now, with ultra-long Oxford Nanopore sequencing, the T2T Consortium has been able to resolve gaps even across centromeres that are hotbeds of tandem satellite arrays. In the new T2T complete reference from CHM13hTERT cells with a normal karyotype, repetitive elements are found to comprise more of the genome than previously thought, with satellite and simple repeats particularly underestimated. But do we have evidence that dark matter can contribute to disease variation? Yes, indeed, several recent studies have shed light on how structural variation in TEs impacts on gene expression differences in disease settings. Our own unpublished recent work pinpoints satellite repeat arrays as platforms for the regulation of normal developmental fate transitions. The initial T2T data and follow-up complete assemblies of more genomes will allow us and others to investigate how previously unresolved dark matter DNA can vary in the human population. Studies assessing how repetitive RNA shapes 3D genome organization will also benefit from the resolved T2T reference sequences. Future breakthroughs building on the T2T initiative will ultimately lead to innovative therapies for diverse diseases but will also allow us to understand more broadly how genomes evolve, and how tissue-specific gene expression programs are controlled.

Beth Sullivan[12]: The recent accomplishment of a truly complete human genome assembly by the T2T Consortium closed the numerous gaps in the contiguous genome that existed for several decades. These gaps contained noncoding sequences including repeats, copy number/structural variants, and transposable elements that were computationally intractable and thus had been ignored and/or largely assumed to have little functional importance. However, the minimization of their significance conflicted with their locations at essential genetic loci, such as telomeres and centromeres that safeguard genome stability. The centromeric regions, defined primarily by alpha satellite DNA stretching for many megabases, were the largest unclosed gaps. The absence of complete contiguous centromere assemblies had made sequence-based mechanisms of centromere identity difficult to investigate.

A gapless T2T assembly will allow new and deeper exploration of genome structure and function, and an increasingly detailed description of the elements that define a chromosome. Human disorders that remain undiagnosed molecularly or unusual chromosomal variants whose molecular origins and functional impact have been elusive might now be explained by comprehensive, new information that T2T sequence, variation, and chromatin maps will provide.

I have been fortunate to be part of the T2T Consortium from the standpoint of centromere assembly. Having spent several decades studying the genomic basis, albeit in a limited fashion, of centromere identity, I believe the achievement of a gapless T2T assembly sets the stage for accomplishing additional T2T assemblies (pangenomes) from different human populations. In my own lab, we have focused on how a particular type of alpha satellite DNA variation on a specific human chromosome affects centromere quality and chromosome stability. With T2T (epi)genomic assemblies, we will soon be able to understand the *full extent* of structural and functional diversity at *all* centromeres more comprehensively—between individuals, in normal and diseased states, throughout development, and during stress and aging. For decades centromere biologists have approached centromere specification and function through the optics of cell biology, biochemistry, or epigenetics, and while we have gained major functional insight, the reality of T2T genomics feels like the missing lens that will bring a previously blurry picture of the centromere into sharper focus.

[1]Department of Computer Engineering, Bilkent University, Ankara 06800, Turkey; [2]Oregon National Primate Research Center, Oregon Health & Science University, Portland, Oregon 97239-3098, USA; [3]Department of Biochemistry and Molecular Medicine, University of California Davis Health, School of Medicine, Davis, California 95616, USA; [4]Department of Biological Chemistry, University of California, Los Angeles, Los Angeles, California 90095, USA; [5]Center for Human Genetics & Genomics, New York University Grossman School of Medicine, New York, New York 10016, USA; [6]Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, Pennsylvania 16802, USA; [7]Division of Life Science, The Hong Kong University of Science and Technology, Kowloon, Hong Kong, China; [8]Department of Pharmacology and Cancer Biology, Duke University Medical Center, Durham, North Carolina 27710, USA; [9]State Key Laboratory of Genetic Engineering, School of Life Sciences, Fudan University, Shanghai, China 200437; [10]Division of Human Genetics at the National Health Laboratory Service, University of the Witwatersrand, Johannesburg, 2000, South Africa; [11]Blizard Institute, Centre for Immunobiology, Queen Mary University of London, London E1 2AT, United Kingdom; [12]Department of Molecular Genetics and Microbiology, Duke University School of Medicine, Durham, North Carolina 27710, USA