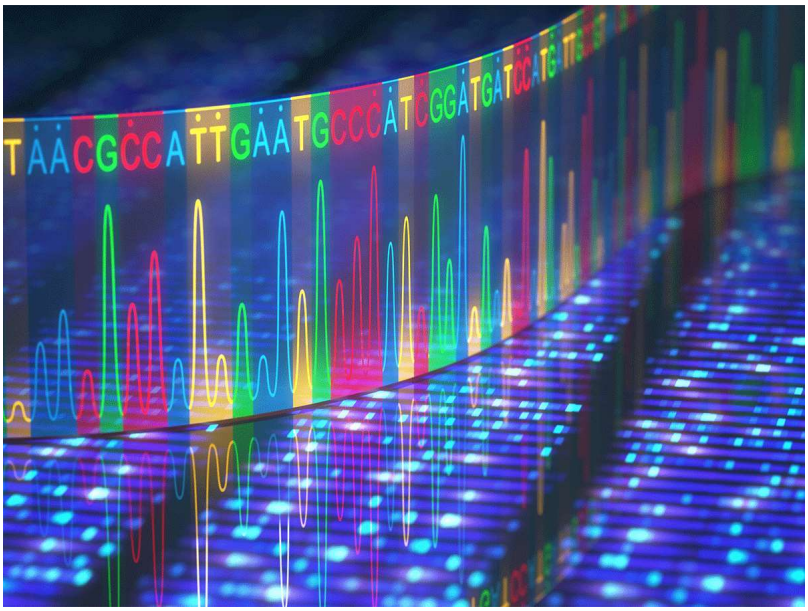
www.bioworld.com/articles/517527-telomere-to-telomere-the-human-genome-is-done



Telomere to telomere, the human genome is done

By Anette Breindl (/authors/2-anette-breindl)

April 1, 2022

There is a project management joke that the first 90% of a project takes 90% of the time, whereas the last 10% of the project takes the other 90% of the time.

So it went with the comprehensive sequence of the human genome.

A first human genome sequence was published in 2003, 19 years after the Alta Summit, a 1984 meeting that the National Institute of Human Genome Research lists as the first key moment in the history of the Human Genome Project.

But that "complete" sequence, although it was a big step up from the draft sequence first published in 2001, was actually only about 92% complete.

Another 19 years later, in the March 31, 2022, issue of *Science*, researchers published a series of papers on a new "telomere to telomere" (T2T) reference genome that includes the last 8%.

The complete genome was reported by the T2T consortium, which is comprised of researchers at the National Human Genome Research Institute (NHGRI), the University of California, Santa Cruz, and the University of Washington at Seattle.

In 2003, "we called it an essentially complete (or near-complete) human genome sequence," NHGRI director Eric Green told reporters at a press conference announcing the full completion of the human genome. "In some ways, this press briefing and these publications might be considered the long-awaited closing cere-

mony (or perhaps encore) to that incredibly audacious project, which determined as much of the human genome sequence as was possible with the tools in hand at the time."

In their lead paper, the authors wrote that the new genome, which goes by T2T-CHM13, "includes gapless assemblies for all chromosomes except Y, corrects errors in the prior references, and introduces nearly 200 million base pairs of sequence containing 1,956 gene predictions, 99 of which are predicted to be protein-coding."

The chromosomes of the human genome are tens to hundreds of millions of base pairs in length, and reading DNA sequences, until fairly recently, was done about a thousand base pairs at a time.

"Shotgun" sequencing, the most common sequencing form, in practice means cutting up many, many copies of a chromosome into bits of different lengths, sequencing the bits, and then reassembling them by overlapping the ends of different fragments.

That works well for many parts of the genome. But in some parts -- including centromeres, transposable elements and duplications -- there are repeat stretches of DNA. Once those repeat sequences are cut, they cannot easily be reassembled, because it is not clear whether how many repeats there are.

Technological advances that have enabled sequencing of up to 20,000 base pairs with near-perfect accuracy, and up to 1 million base pairs with lower but still useful accuracy, enabled the sequence of those challenging regions.

In their papers, the consortium also reported genomic, epigenetic and transcriptional analyses of the newly sequenced regions.

Next up

In a commentary that accompanied the papers, Deanna Church of Inscripta wrote that the T2T reference genome is "poised to have its own substantial impact on genome analysis and represents an important step to assembly models that represent all humans, which will better support personalized medicine, population genome analysis and genome editing."

An important part of representing all humans will be to T2T genomes from diverse populations. Several of the leaders of the T2T consortium, including the University of Washington's Evan Eichler and the University of California at Santa Cruz' Karen Miga, are on the steering committee of the Human Pangenome Reference Consortium, whose goal it is to generate several hundred T2T genomes that include persons of different ancestries.

The same long-read advances that have enabled sequencing the last bit of the human genome have enabled researchers to discover an unexpected level of genomic diversity in populations with different ancestries.

The group of Evan Eichler, professor at the University of Washington, co-chair of the T2T consortium and corresponding author of several of the papers now published in *Science*, has shown that long-read sequencing can detect three times the number of variants that older methods can, and have used it to identify novel genes in specific populations.

Understanding genomic variation in populations of different ancestries is a matter of scientific accuracy, as well as indispensable for precision medicine.

But it is also in the self-interest of the biopharmaceutical industry, because it provides a rich trove of variants that can be the basis of blockbuster drug discovery efforts.

The initial identification of PCSK9, target of the blockbuster cholesterol drug Inclisiran (Novartis), for example, was made in individuals with African ancestry. Early case studies linking PCSK9 mutations to very low LDL cholesterol, and the publication demonstrating that they lowered the risk of heart disease, studied Black Americans. About 2% of Black Americans have mutations in PCSK9, a much higher rate than European ancestry populations (Nurk, S. et al. Science 2022, 376(6588): 44; Aganezov, S. et al. Science 2022, 376(6588): eabl3533; Vollger, M.R. et al. Science 2022, 376(6588): eabj6965; Altemose, N. et al. Science 2022, 376(6588): eabl4178; Hoyt, S.J. et al. Science 2022, 376(6588): eabk3112; Gershman, A. et al. Science 2022, 376(6588): eabj5089).