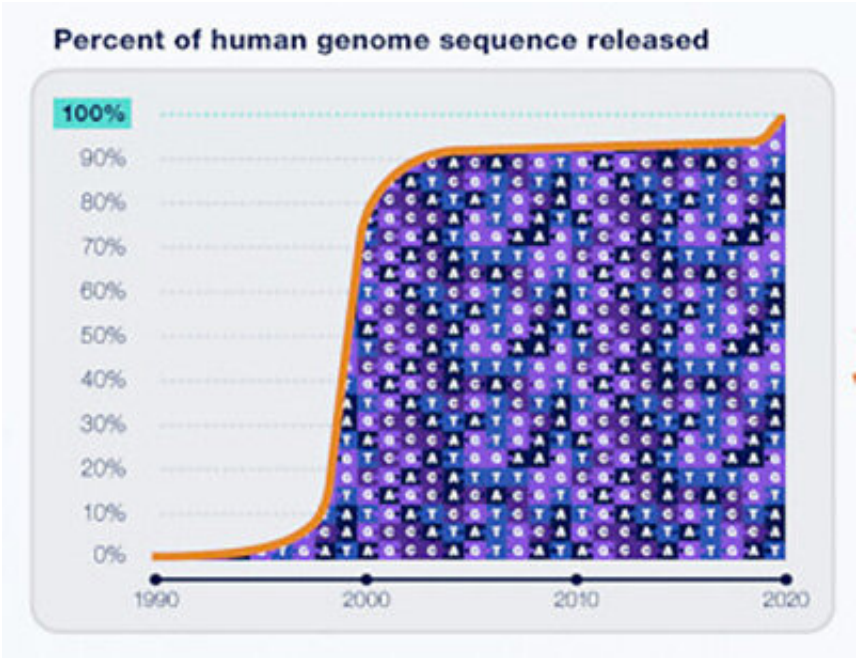
# The human genome is, at long last, complete

March 31 2022




It took almost twice as long to finish the last 8% of the human genome as it did to sequence the first 92%. New laboratory and computational technologies finally enabled genomic researchers to overcome obstacles such as highly repetitive DNA sequences and fill in the remaining gaps. Credit: NHGRI

When scientists declared the Human Genome Project complete two decades ago, their announcement was a tad premature. A milestone achievement had certainly been reached, with researchers around the world gaining access to the DNA sequence of most protein-coding genes in the human genome. But even after 20 years of upgrades, eight percent of our genome still remained unsequenced and unstudied. Derided by

some as "junk DNA" with no clear function, roughly 151 million base pairs of sequence data scattered throughout the genome were still a black box.

Now, a large international team led by Adam Phillippy at National Institutes of Health has revealed the final eight percent of the human genome in a paper published in *Science*. These long–missing pieces of our genome contain more than mere junk. Within the new data are mysterious pockets of noncoding DNA that do not make protein, but still play crucial roles in many cellular functions and may lie at the heart of conditions in which cell division runs amok, such as cancer.

"You would think that, with 92 percent of the genome completed long ago, another eight percent wouldn't contribute much," says Rockefeller's Erich D. Jarvis, a coauthor on the study who helped develop a number of techniques central to unlocking the final pieces of the human genome. "But from that missing eight percent, we're now gaining an entirely new understanding of how cells divide, allowing us to study a number of diseases we had not been able to get at before."

## On the shoulders of the HGP

The Human Genome Project essentially handed us the keys to euchromatin, the majority of the human genome, which is rich in genes, loosely packaged, and busy making RNA that will later be translated into protein. Left untouched, however, was a labyrinth of tightly wound, repetitive heterochromatin—a smaller portion of the genome, which does not produce protein.

Scientists had good reasons for initially deprioritizing heterochromatin. The euchromatic regions contained more genes and were simpler to sequence. Just as a puzzle with distinct pieces is easier to put together than a puzzle composed of similar ones, the genomics tools of the day

found euchromatic DNA easier to parse than its repetitive, heterochromatic cousin.

As a result, geneticists were left with a sizable hole in their knowledge of what drives some basic cellular functions. The heterochromatic sequences behind centromeres, which lie at the cruxes of chromosomes and conduct cell division, were all marked with longs runs of N for "unknown base" in the human reference genome. The sequences of the short arms of chromosomes 13, 14, 15, 21, and 22 were likewise omitted. "Not even all of the euchromatic genome was sequenced properly," Jarvis adds. "Errors, such as false duplications, needed to be fixed."

Then, about ten years ago, scientists began developing new techniques for producing longer sequence reads that filled in gaps in the genomes of humans and other species. One such initiative is the Vertebrate Genomes Project, helmed by Jarvis, which recently produced the first near error-free and near complete reference genomes for 25 animals. "That study was part of an international effort to develop new tools that produce the highest-quality gene assemblies," he says. "Compared to the methods that were used twenty years ago, modern genomics has high-fidelity long reads that are 99.9 percent accurate, better genome assembly tools, and more powerful algorithms that are better at distinguishing similar-looking puzzle pieces from one another."

With updated tools and renewed resolve, Jarvis and other scientists were able to help finish what the Human Genome Project started and describe, at long last, a truly complete human genome—its euchromatic regions revised, and its heterochromatic regions on full display.

"It's a big deal," Jarvis says. "Every single base pair of a human genome is now complete."

## Meeting Merfin

The flagship *Science* study was led by the Telomere-to-Telomere (T2T) Consortium, a group of researchers at various academic institutions and NIH. The Jarvis lab's contribution, published in Nature Methods, involved providing tools to help T2T refine messy genome sequences to produce error-free sequences.

One of these tools is Merfin, which they used to clean up some of the most difficult sequences in the human genome. "Genomes that we generate in the lab can have many errors in them," says Giulio Formenti, a postdoc in Jarvis' lab who developed Merfin. "If even just one or a few base pairs are wrong, that can have big consequences for the overall accuracy of the genomic sequence." Merfin makes it possible to test the accuracy of a sequence, sensing code that may be out of place and automatically correcting mistakes. Because the technologies that generate modern sequences are more accurate, Merfin is reserved for only the trickiest cases.

"Stretches of identical base pairs, such as AAA, are hard for existing technology to assess," Formenti says. "There are often errors in those sequences, even now. Merfin corrects them."

Jarvis and Formenti hope that their contribution will not only help tie a bow on the Human Genome Project, but also inform research into diseases linked to the heterochromatic genome—chief among them cancer, which is associated with centromere abnormalities. Cancer cells divide wildly when certain heterochromatic centromere genes are overexpressed, and a complete understanding of the centromere genome may open the door to novel therapies.

"We are finally digging into what we once called junk DNA, because we could not understand it or look at it accurately," Formenti says. "We now

know that many diseases are linked to structural repeats in the centromere and, now that these sequences are no longer missing from the human reference genome, we can begin to map the origins of these diseases."

Other co-authors in the Merfin study are: Arang Rhie, Brian P. Walenz, Françoise Thibaud-Nissen, Kishwar Shafin, Sergey Koren, Eugene W. Myers, and Adam M. Phillippy.

**More information:** Sergey Nurk et al, The complete sequence of a human genome, *Science* (2022). DOI: 10.1126/science.abj6987. www.science.org/doi/10.1126/science.abj6987

Provided by Rockefeller University