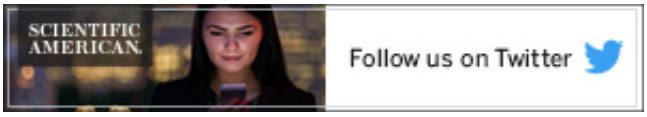

STAT

BIOLOGY

The Human Genome Was Never Completely Sequenced

The effort completed in 2003 used the best technology available but now scientists could do more

By Sharon Begley, STAT on June 20, 2017

Credit: Andrew Brookes Getty Images

The feat made headlines around the world: “Scientists Say Human Genome is Complete,” the New York Times announced in 2003. “The Human Genome,” the journals Science and Nature said in identical ta-dah cover lines unveiling the historic achievement.

There was one little problem.

“As a matter of truth in advertising, the ‘finished’ sequence isn’t finished,” said Eric Lander, who led the lab at the Whitehead Institute that deciphered more of the genome for the government-funded Human Genome Project than any other. “I always say ‘finished’ is a term of art.”

“It’s very fair to say the human genome was never fully sequenced,” Craig Venter, another genomics luminary, told STAT.

“The human genome has not been completely sequenced and neither has any other mammalian genome as far as I’m aware,” said Harvard Medical School bioengineer George Church, who made key early advances in sequencing technology.

Read more: Geneticist Craig Venter helped sequence the human genome. Now he wants yours

What insiders know, however, is not well-understood by the rest of us, who take for granted that each A, T, C, and G that makes up the DNA of all 23 pairs of human chromosomes has been completely worked out. When scientists finished the first draft of the human genome, in 2001, and again when they had the final version in 2003, no one lied, exactly. FAQs from the National Institutes of Health refer to the sequence’s “essential completion,” and to the question, “Is the human genome completely sequenced?” they answer, “Yes,” with the caveat — that it’s “as complete as it can be” given available

technology.

Perhaps nobody paid much attention because the missing sequences didn't seem to matter. But now it appears they may play a role in conditions such as cancer and autism.

"A lot of people in the 1980s and 1990s [when the Human Genome Project was getting started] thought of these regions as nonfunctional," said Karen Miga, a molecular biologist at the University of California, Santa Cruz. "But that's no longer the case." Some of them, called satellite regions, misbehave in some forms of cancer, she said, "so something is going on in these regions that's important."

Miga regards them as the explorer Livingstone did Africa — terra incognita whose inaccessibility seems like a personal affront. Sequencing the unsequenced, she said, "is the last frontier for human genetics and genomics."

Church, too, has been making that point, mentioning it at both the May meeting of an effort to synthesize genomes, and at last weekend's meeting of the International Society for Stem Cell Research. Most of the unsequenced regions, he said, "have some connection to aging and aneuploidy" (an abnormal number of chromosomes such as what occurs in Down syndrome). Church estimates 4 percent to 9 percent of the human genome hasn't been sequenced. Miga thinks it's 8 percent.

The reason for these gaps is that DNA sequencing machines don't read genomes like humans read books, from the first word to the last. Instead, they first randomly chop up copies of the 23 pairs of chromosomes, which total some 3 billion "letters," so the machines aren't overwhelmed. The resulting chunks contain from 1,000 letters (during the Human Genome Project) to a few hundred (in today's more advanced sequencing machines). The chunks overlap. Computers match up the overlaps, assembling the chunks into the correct sequence.

That's between difficult and impossible to do if the chunks contain lots of repetitive segments, such as TTAATATTAATATTAATA, or TTAATA three times. "The problem is, when you have the same exact words, it's hard to assemble," said Lander, just as if jigsaw puzzle pieces show the same exact blue sky.

In 2004, the genome project reported that there were 341 gaps in the sequence. Most of

the gaps — 250 — are in the main part of each chromosome, where genes make the proteins that life runs on. These gaps are tiny. Only a few gaps — 33 at last count — lie in or near each chromosome's centromere (where the two parts of a chromosome connect) and telomeres (the caps at the end of chromosomes), but these 33 are 10 times as long in total as the 250 gaps.

That makes the centromeres in particular the genome's uncharted Zambezi. Evan Eichler of the University of Washington said every chromosome has such sequence-defying repetitive elements — think of them as DNA stutters — including an infamous one that's 171 letters long and repeated end-to-end for thousands of letters.

At the beginning of the Human Genome Project, said Lander, now director of the Broad Institute of MIT and Harvard, "it became very clear these highly repetitive sequences would not be tractable with existing technology. It wasn't a cause of a great deal of agonizing at the time," since he and other project leaders expected the next generation of scientists to find a solution.

That hasn't really happened, partly because there hasn't been much motivation to map these regions. "I'm between agnostic and a little skeptical that these bits will be important for disease, but maybe I'm saying that because we can't read them," Lander said.

As new sequencing technology has begun allowing scientists to peek into unsequenced territory, however, they have seen that "these tough-to-sequence regions frequently have important genes," said Michael Hunkapiller, chairman and CEO of Pacific Biosciences, which makes DNA sequencers. (In 1998, Hunkapiller recruited Venter to his new company, Celera Genomics, to race the government-backed genome project.)

PacBio's "reason for being" is to increase the length of DNA segments that can be read and assemble them, Hunkapiller said. Longer reads have an effect like enlarging jigsaw puzzle pieces; even though the pieces still contain a lot of repeated blue sky, the greater size makes it more likely they'll also contain something sufficiently novel to make assembling them easier. PacBio's maximum DNA read is now about 60,000 letters, Hunkapiller said, and averages 15,000.

With such long reads, Lander said, "you could get through a lot of these nasty [unsequenced] regions."

Read more: 'Genome writers' gather in New York to pitch bomb-sniffing plants and more. Where's the funding?

That's looking more and more like a worthy undertaking, and not only because the unsequenced regions might contain actual protein-making genes. There is evidence that the non-gene parts — especially the DNA stutters — "clearly have disease implications," Hunkapiller said. "Three-quarters of the [genome] differences between one person and another are in [such] variants" rather than the single-letter spelling differences in A's, T's, C's, and G's which get all the attention. In a 2007 paper, Venter and his team showed that there are more person-to-person differences like this, called structural variants, than there are single-letter changes.

Yet about 90 percent of the structural variants, the vast majority of which weren't sequenced by either the genome project or a later effort called the 1000 Genomes Project, "have been missed," Eichler and his colleagues reported last year.

One reason the stutters are unusually influential is that this repetitive DNA can move around, make copies of itself, flip its orientation, and do other acrobatics that "can have quite dramatic functional effects," Hunkapiller said. For one thing, repetitive elements around the centromeres, called satellites, might cause a dividing cell to become cancerous, Miga said, because they can destabilize the entire genome.

When researchers at Stanford University tried to find the genetic cause of a young man's mysterious disease, which caused non-cancerous tumors to grow throughout his body, they found nothing using the standard whole-genome sequencing, Hunkapiller said. But the "long reads" made possible by the PacBio machines "looked for structural variants and found the problem right away," he said.

The stutters might even be what makes us human. Some of these complex duplications "appear to be important for the evolution of higher neuroadaptive function" — aka brain development, Eichler said. A gene called ARHGAP11B, which was created by one such duplication, causes the cortex to develop the myriad folds that support complex thought; SRGAP2C, also a duplication, triggers brain development.

"These are new genes that evolved specifically in our lineage over the last few million

years," said Eichler. The same duplications can also produce DNA rearrangements "associated with neurodevelopmental disorders such as autism and intellectual disability."

"Finish the sequence!" hasn't become a rallying cry, but maybe it should be, Venter said: "I'd be the last one to give you a quote saying that we don't need to bother with these [unsequenced] regions."

Republished with permission from STAT. This article originally appeared on June 20, 2017

ABOUT THE AUTHOR(S)

Sharon Begley

Recent Articles

"A Feature, Not a Bug": George Church Ascribes His Visionary Ideas to Narcolepsy

Does Living in a City Make You Psychotic?

Scientists Gather for Genome Writing Conference, but Funding Is Scarce