Daily Memphian

SUBSCRIBERS ONLY BUSINESS

UTHSC researchers building software to scan thousands of genomes

By Jane Roberts, Daily Memphian
Updated: October 14, 2021 8:35 AM CT | Published: October 14, 2021 4:00 AM CT

The human genome has been mapped for close to two decades. A second, more complete roadmap of DNA in the 23 chromosomes was finished this year.

It's now possible for U.S. citizens to have their whole genome mapped for free (Allofus.gov), and in the not-so-distant future, it's likely individual treatment will be based on the patient's own genome.

The University of Tennessee Health Science Center is one of three institutions carrying the work to the next level by building a supercomputer that in minutes will be able to compare one person's genome against thousands of others instead of a single reference genome.

Federal government asks to intervene in case against Methodist

That will help health care providers quickly see how rare a particular gene might be or how often it is associated with disease or poor health.

Since the human genome was mapped, scientists have searched for abnormalities by comparing the genome of one person against the reference genome, a compilation of a handful of people who volunteered to be the representative sample of humanity.

"The current reference is a bit like the genome of James Holden from 'The Expanse.' He has eight parents who all contributed equally," says Erik Garrison, one

of two Memphis-based researchers spending the next five years building the supercomputer.

He and Pjotr Prins, both assistant professors at UTHSC, have received funding from National Science Foundation to build software that will make comparisons possible against a pangenome, or a collection of many whole human genomes.

The reference genome represents the best the researchers could do with the technology of the time. To avoid biasing it toward one individual, they used a collection of donors, taking perhaps 50% DNA from one person and 10% from another to simulate a single human genome.

“They appear to have attempted to choose a genetically diverse set, although by practical necessity, most of the donors came from one geographic area,” said Garrison, a Harvard University graduate who earned a Ph.D. in genomics from Cambridge.

“The flaw with any reference genome is simply that it is a single genome, which means that it represents some genomes better than others. With pangenomics, we are tackling this issue by greatly increasing the number of individual genomes in our reference system.”

TennCare enrollment went up 15 months in a row

Comparing a sample to a single genome is limiting, the two say, because it can miss differences contained in other genomes.

For instance, when a collection of genomes from African people is compared to the reference, “about 10% of data from Africa we cannot map against reference genome. It’s just missing,” says Prins, a bioinformatician.

Because most variants in human DNA are shared with others, adding more genomes to the reference makes it more representative.

“To bring all this information together, we put the genomes of the pangenome into a unified model that lets us understand the sequences and variation between them,” Garrison said.

“Building and working with that model is hard, and it could benefit from a computing system that’s specifically designed for the problem — that’s the motivation behind our NSF grant.”

For context, most laptops and smartphones have four cores or central processing units. This computer will have a million.

For perspective, the assemblies from the first year of the Human Pangenome Project, completed in 2020, are around 300 gigabytes.

“The total raw sequencing data from sequencing machines is much larger than these assemblies,” Garrison said.

The computer is being built in partnership with Cornell University and the University of Washington, where it will be based in the university’s Center for Clinical Genomics in Seattle. Prins and Garrison from UTHSC are creating the biomedical software.

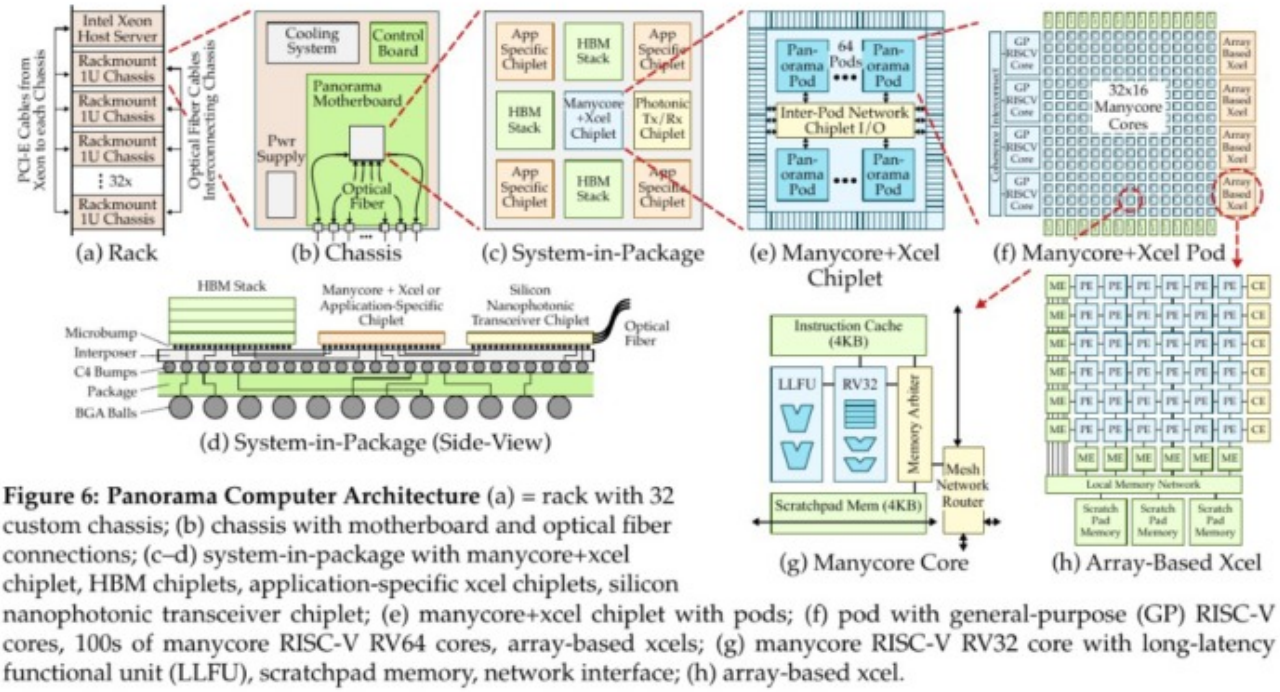

Figure 6: Panorama Computer Architecture (a) = rack with 32 custom chassis; (b) chassis with motherboard and optical fiber connections; (c–d) system-in-package with manycore+xcel chiplet, HBM chiplets, application-specific xcel chiplets, silicon nanophotonic transceiver chiplet; (e) manycore+xcel chiplet with pods; (f) pod with general-purpose (GP) RISC-V cores, 100s of manycore RISC-V RV64 cores, array-based xcels; (g) manycore RISC-V RV32 core with long-latency functional unit (LLFU), scratchpad memory, network interface; (h) array-based xcel.

This is a graphic of the design of the computer. It was submitted to the National Science Foundation in the grant proposal for funding. The portion of the work being done at Unvieristy of Tennessee Health Science Center was funded with more than $600,000. (Submitted)

For months, they have been working in a pandemic-induced exile. Both returned to their family homes in Europe during the lockdown and now are waiting for the U.S. to reopen its borders.

Prins is in the Netherlands. Garrison is in Italy.

The science of whole genome testing is quickly becoming mainstream. St. Jude Children's Research Hospital sequences the whole genome of each new cancer patient, which takes more than a week. Then it spends about a month checking it against the reference and looking up the variations to see how often they have been found and under what conditions.

"Doing it in minutes would be pretty amazing," says David Wheeler, director of the St. Jude Precision Genomics team.

"Clinicians need to have the information sooner not later. The fact that is it taking us, from the start of sequencing to the final result, roughly six weeks right now, that is impeding the ability of the oncologists to use the information in the diagnostic and upfront treatment."

Local developers feel sting of high construction costs

In less than 20 years, millions of patients will have their genome data on a piece of plastic, like a credit card, which they will present to their health care provider, says Dr. Robert Williams, who supervises Prins and Garrison as chair of the UTHSC Department of

Genetics, Genomics & Informatics.

"It doesn't do you a whole lot of good — or a health care provider a lot of good — without understanding the relationship between differences you happen to have inherited and outcome for diseases," he said.

Questions about who is likely to get Alzheimer's or be seriously affected by smoking are better answered by looking at large groups, he says.

"The only way to figure out what's good for you is to actually study millions of people and understand the trend lines for those millions of people.

"The supercomputer will allow you to weave together all the genomes in a giant fabric of humanity," Williams said.

Robert Williams

UTHSC does some of the work already on a computer large enough to do a simple analysis across 60 people. It takes about a day.

"If you double the number of individuals, the time will not double," Prins said. "It will get even slower, maybe four times as slow."

Analyzing 1,000 people would probably take a month, he says. "We want to be able to do it in minutes."

Their hope is that the pangenome computer will replace the single reference genome, which has been updated numerous times.

“There are multiple editions of the same book, with little difference between new editions,” Wheeler said.

“The problem with an update is, tons of research go into annotating the reference genome and the millions of base variants of which we know,” he said.

The updates generally add small bits of information, but they also throw off the numbering of approximately 3 billion bases, which means it takes times to shift the old information to the new release, Wheeler says.

For that reason, St. Jude, he says, is just now getting ready to shift to the 2013 update.

When scientists study genomes by comparing them to a reference, they end up looking a bit more like the reference genome than they really are, a wrinkle scientists call the reference bias.

“To avoid this, we want to relate genomic data to a much larger, more inclusive pangenome. We build this from many whole genomes, not just a single one,” Garrison said.

Garrison is a member of the Human Genome Reference Consortium, which is assembling the genomes of hundreds of people from diverse backgrounds.

They will be integrated into the computerized pangenome. So will genomes already available in public repositories. As more are added through high-fidelity measures, including processes that now can read the chromosome from end to end, a more complete picture of humanity from around the world will emerge.

“It is as if we are looking at the dark side of the moon,” Prins said. “We are seeing a huge chunk of DNA that we’ve never seen before.

“It turns out that even closely related people within families can have very diverse DNA. There’s much more variation than we thought.”

Thanks to popular genetic sites like 23andMe.com and large-scale research projects, several hundred thousand people in the U.S. have had their genomes sequenced. In 20 years, Wheeler says millions of people will this data on themselves.

For now, he says, it's hard to know how much benefit medical science will gain from the knowledge.

"We don't have a large population of people that have already been sequenced that we can follow longitudinally through their lives to see what benefit it is."

The only way to know the benefit, he says, is to sequence large numbers of people and measure the gains.

To that point, the National Institutes of Health launched Allofus.gov in 2018, hoping to collect the genomes of a million or more U.S. residents by 2025.

Participants give a blood or saliva sample, although blood is preferred, through an approved clinic. Participants may choose to receive their personal information. It is sent in batches over time.

In 2020, the first people to sign up, starting in 2018, received their data.

All of Us is one of several large genome-collecting projects happening now around the world.

"It's still research now. We are still asking the questions. But people who know about this field think the benefit is going to be enormous. I share that sentiment," said Wheeler, who has had his sequence done.

So has Williams, who says when he looks in the mirror in the morning, he wants to see the whole picture.

"If you take your shoes off, you want to be able to see your toes, right?"

Eventually, science will be able to add details about the levels of stress in a patient's

life, giving health care a better look at the genetic and environmental factors in people's lives, Williams says.

"Did your mom and dad love you and treat you right? Were you sexually abused? Did you fight in a war on behalf of America? All that matters a huge amount."

There is no timeline on when stress data will be available, but when it is, Williams says, health care will change from disease treatment to disease prevention.

"The whole point of being human is to know stuff. That is our job. If we don't know our own genome, what do we know?"

TOPICS

PAN-GENOME **DNA** **UTHSC** **SUBSCRIBER ONLY**

Thank you for supporting local journalism.

Subscribers to The Daily Memphian help fund our newsroom of over 35 full-time, local journalists plus more than 20 freelancers, all of whom work around the clock to cover the issues impacting our community. Subscriptions - and **donations** - also help fund our **community access programs** which provide free access to K-12 schools, senior-living facilities and more. Thank you for making our work possible.

Jane Roberts

Longtime journalist Jane Roberts is a Minnesotan by birth and a Memphian by choice. She's lived and reported in the city more than two decades. She covers healthcare and higher education for The Daily Memphian.