ILLUSTRATION BY THE PROJECT TWINS

TERRA MAKES CLOUD-BASED 'OMICS' ACCESSIBLE

The web-based tool allows scalable, user-friendly computation across multiple data sets. **By Jeffrey M. Perkel**

Biomedical research produces massive amounts of data. But actually sharing them can be a challenge.

"It doesn't matter how enthusiastic you are about sharing data, actually sharing data is logistically a pain," says Elinor Karlsson, a geneticist at the University of Massachusetts Chan Medical School in Worcester.

Karlsson is a co-lead of the genetics core for the Dog Aging Project, a longitudinal study funded by the US National Institutes of Health (NIH), that aims to understand the genetics of healthy ageing by analysing several thousand genome sequences and health data from 100,000 pets.

The project involves researchers from multiple institutions, all of whom need to be able to access and analyse the data. Conventional data-sharing strategies – such as shared servers, data downloads and even shipping physical disk drives – simply weren't up to the task, Karlsson says.

Advances in DNA sequencing mean it's much easier to collect genetic data than it is to analyse them. As data sets have ballooned, Karlsson says, "We went from it being inconvenient to copy data to my server, to being cost-prohibitive. Institutions didn't have the resources to keep up."

So the project turned to a system that could: Terra.

Terra began life as a service called FireCloud. It was developed by the Data Sciences Platform (DSP) team at the Broad Institute of MIT and Harvard in Cambridge, Massachusetts, in collaboration with Microsoft and Verily Life Sciences – a subsidiary of Google's parent company, Alphabet. In 2019 it was renamed and folded into the National Human Genome Research Institute's Analysis, Visualization and Informatics Lab-Space (AnVIL). Terra's web-based interface provides user-friendly access to scalable workflows, collaboration and analysis tools. It allows researchers to quickly integrate and computationally analyse vast 'omics' data sets on the Google Cloud Platform. Rather than forcing researchers to find and download data to analyse them locally, Terra lets them work on the data *in situ*, using as much or as little computing power as they need.

"It's really about increasing access and breaking down [data kept in separate] silos so that you can do more interesting science with the data that's being generated," says Geraldine Van der Auwera, communications director for the DSP at the Broad Institute. Terra provides access to data sets such as the Cancer Genome Atlas, the Genome Aggregation Database and the All of Us Research Program, the last of which comprises some 3.7 petabytes alone. Users can also upload their own data; access to 'sensitive' (personally identifiable) information is limited to people with proper authorization.

Terra can be used in 'batch' mode, running scripts written in the Workflow Description Language (a tool for specifying data-processing workflows) to drive anything from one to tens of thousands of virtual computers. Alternatively researchers can explore data interactively using tools such as Jupyter Notebook, RStudio and the graphical Galaxy workflow engine.

Michael Schatz, a computational biologist at Johns Hopkins University in Baltimore, Maryland, was part of a team that used Terra and other AnVIL components to process data for the Telomere-to-Telomere consortium, which in May 2021 documented the first gap-free assembly of the human genome. That assembly adds some 200 million base pairs of DNA to the human genome, and the Telomere-to-Telomere consortium used it to search for variants in the 3,202 genomes sequenced as part of the 1000 Genomes Project. A computing cluster within the Maryland Advanced Research Computing Center (MARCC) in Baltimore, a multimillion-dollar high-performance facility, would have needed up to a year to perform those analyses, Schatz estimates. But using Terra it took just weeks, including optimization time. "If we had to do it again, we could probably get it all done in about a week," he says.

For Karlsson, Terra eases data access and collaboration problems. Her PhD student, Kathleen Morrill, has created scripts to automate the uploading and analysis of new data sets, for instance to determine the sex of the sequenced animals. Researchers who wish to explore those data further can simply log in, copy her shared workspace and get to work. "In the physical laboratory, you'd make your data reproducible through your lab notebook," Morrill says. "By having the analysis up on the cloud in these Terra workspaces it becomes something that's easier to track and share and reproduce."

C. Titus Brown, a bioinformatician at the University of California, Davis, and his team are creating tutorials for Terra as part of their work for the NIH Common Fund Data Ecosystem, a data-reuse initiative. Although he finds the platform promising, Brown cautions that cloud computing is still new, and says that it will be some time before it's the best option for everybody. "There's a technological adoption curve," he explains, "and we're still [in the] very early stages." Few scientists are comfortable working in the cloud, even with Terra's relatively user-friendly interface. And researchers and institutions are more familiar with buying computers than renting time on them.

And then there's the fact that the cost of cloud computing can be opaque. Alexander Bick, a human geneticist at Vanderbilt University Medical Center in Nashville, Tennessee, recalls one student who inadvertently over-requested computing resources on Terra, racking up a US$500 tab for a job that should have cost just $5. "It was among the most expensive runs that we had," he says. But it was also comparable to wasting a vial of antibodies, he notes. "It's all relative."

"If you can spend $3,000 on Terra to do your computing, that's clearly a win."

Schatz advises starting small, for instance by analysing a few samples to assess the per-sample cost, then scaling up. Establish firewalls so intensive workflows don't burn through your budget. And optimize your workflow, adds Ryan Collins, a computational geneticist and PhD student at Harvard Medical School in Boston. "Every additional gigabyte of RAM that you request is going to translate directly to a cost that will get billed back to you," he says.

Terra's website lists workspaces that bundle different workflows together with model data, including pricing details to help users estimate the cost for their own data. Some 56 such featured workspaces have been created, Van der Auwera says, including two from the Human Cell Atlas project.

For Brown, Terra shines when working with data that are either sensitive or extremely large. Privacy rules preclude working with sensitive health data on insecure systems, he notes. And downloading data from the cloud can be expensive, with some data sets costing tens of thousands of dollars to download because of their size. "If instead you can spend $3,000 on Terra to do your computing, that's clearly a win."

Besides, in many cases there is no other practical option, Karlsson notes, even for institutions with high-performance local computing resources. Powerful on-premises clusters, such as MARCC, can still struggle with modern genomics data sets. "Whether Terra or something else, we need something to get to this huge data-scale environment," she says.

Jeffrey M. Perkel is Technology Editor for *Nature*.

Correction

This Technology feature neglected to mention that Terra was developed in collaboration with Microsoft and Verily Life Sciences. It also misnamed the group involved. The team is part of the Data Sciences Platform, not the Data Services Platform.

Corrected 5 January 2022