UK Biobank Advances Genomics Research with NVIDIA Clara Parabricks

World's largest repository of deidentified data for genetics research is providing scientists access to NVIDIA Clara Parabricks to fuel accelerated genomics and drug discovery.

Author: Craig Rhodes

UK Biobank is broadening scientists' access to high-quality genomic data and analysis by making its massive dataset available in the cloud alongside NVIDIA GPU-accelerated analysis tools.

Used by more than 25,000 registered researchers around the world, UK Biobank is a large-scale biomedical database and research resource with deidentified genetic datasets, along with medical imaging and health record data, from more than 500,000 participants across the U.K.

Regeneron Genetics Center, the high-throughput sequencing center of biotech leader Regeneron, recently teamed up with UK Biobank to sequence and analyze the exomes — all protein-coding portions of the genome — of all the biobank participants.

The Regeneron team used NVIDIA Clara Parabricks , a software suite for secondary genomic analysis of next-generation sequencing data, during the exome sequencing process.

UK Biobank has released 450,000 of these exomes for access by approved researchers, and is now providing scientists six months of free access to Clara Parabricks through its cloud-based Research Analysis Platform. It was developed by bioinformatics platform DNAnexus, which lets scientists use Clara Parabricks running on NVIDIA GPUs in the AWS cloud.

"As demonstrated by Regeneron, GPU acceleration with Clara Parabricks achieves the throughputs, speed and reproducibility needed when processing genomic datasets at scale," said Dr. Mark Effingham, deputy CEO of UK Biobank. "There are a number of research groups in the U.K. who were pushing for these accelerated tools to be available in our platform for use with our extensive dataset."

Regeneron's researchers used the DeepVariant Germline Pipeline from NVIDIA Clara Parabricks to run their analysis with a model specific to the genetic center's workflow.

Its researchers identified 12 million coding variants and hundreds of genes associated with health-related traits — certain genes were associated with increased risk for liver disease and eye disease, and others were linked to lower risk of diabetes and asthma.

The unique set of tools the researchers used for high-quality variant detection is available to UK Biobank registered users through the Research Analysis Platform . This capability will allow scientists to harmonize their own exome data with sequenced exome data from UK Biobank by running the same bioinformatics pipeline used to generate the initial reference dataset.

Researchers deciphering the genetic codes of humans — and of the viruses and bacteria that infect humans — can often be limited by the computational resources available to them.

UK Biobank is democratizing access by making its dataset open to scientists around the world, with a focus on further extending use by early-career researchers and those in low- and middle-income countries. Instead of researchers needing to download this huge dataset to use on their own compute resources, they can instead tap into UK Biobank's cloud platform through a web browser.

"We were being contacted by researchers and clinicians who wanted to access UK Biobank data, but were struggling with access to the basic compute needed to work with even relatively small-scale data," said Effingham. "The cloud-based platform provides access to the world-class technology needed for large-scale exome sequencing and whole genome sequencing analysis."

Researchers using the platform pay only for the computational cost of their analyses and for storage of new data they generate from the biobank's petabyte-scale dataset, Effingham said.

Using Clara Parabricks on DNAnexus helps reduce both the time and cost of this genomic analysis, delivering a whole exome analysis that would take nearly an hour of computation on a 32-vCPU machine in less than five minutes — while also reducing cost by approximately 40 percent.

For researchers studying links between genetics and disease, exome sequencing is a critical tool — and the UK Biobank dataset includes nearly half a million participant exomes to work with.

The exome is approximately 1.5 percent of the human genome, and consists of all the known genes and their regulatory elements. By studying genetic variation in exomes across a large, diverse population, scientists can better understand the population's structure, helping researchers address evolutionary questions and describe how the genome works.

With a dataset as large as UK Biobank's, it is also possible to identify the specific genetic variants associated with inherited diseases, including cardiovascular disease, neurodegenerative conditions and some kinds of cancer.

Exome sequencing can even shed light on potential genetic drivers that might increase or decrease an individual's risk of severe disease from COVID-19 infection, Effingham said. As the pandemic continues, UK Biobank is adding COVID case data, vaccination status, imaging data and patient outcomes for thousands of participants to its database.

Get started with NVIDIA Clara Parabricks on the DNAnexus-developed UK Biobank Research Analysis Platform . Learn more about the exome sequencing project by registering for this webinar , which takes place Feb. 17 at 8am Pacific.

Subscribe to NVIDIA healthcare news here .

Main image shows the freezer facility at UK Biobank where participant samples are stored. Image courtesy of UK Biobank.

Original URL:
https://blogs.nvidia.com/blog/2022/01/26/uk-biobank-advances-genomics-research-clara-parabricks/