THE STRIDES INITIATIVE SERIES:

Expanding Access to TOPMed Genomics Data Through Cloud Services

Journey to the Cloud to Support Biomedical Research

The University of Michigan was no stranger to using cloud for computing and storing data at scale for biomedical research, even prior to participating in the STRIDES Initiative to access cloud resources. As a part of the Trans-Omics for Precision Medicine (TOPMed) program, the University of Michigan works with study investigators to harmonize genotype data and to address their informatics challenges for heart, lung, blood, and sleep disorder research. This work requires expansive data storage, computing, and sharing.

Before transitioning to the cloud, the University of Michigan experienced problems scaling and sending data back to researchers. The TOPMed data—roughly 1.8 petabytes today—quickly grew and became difficult to manage, compute, and share because of its magnitude. Additionally, the University of Michigan's local compute resources were finite—approximately 5,000 central processing cores—which could not run multiple data analysis jobs simultaneously. At the time, the University of Michigan shuffled data from cluster to cluster between universities and researchers in order to share data. To reduce data processing time and to streamline the data sharing process, the team investigated cloud services to improve access to data and to decrease the need for local computation resources.

> "In the general sense, we have had a very good experience with the cloud, and for anyone that is approaching genomics data, rather than building out their local cluster, which will eventually become strained, we encourage them to go to cloud for the overall efficiencies for computing at scale. The STRIDES Initiative would be a good way to get started."
>
> —**Albert Vernon Smith**, research senior supervisor, University of Michigan

"In order to produce the results we needed in a reasonable and practical amount of time, we had to migrate what we were doing locally to the cloud. By moving to the cloud, we were able to compress a year's worth of data processing into a couple months," said Jonathan LeFaive, senior app programmer/analyst in the School of Public Health's Department of Biostatistics at the University of Michigan.

Making the Switch to the STRIDES Initiative

The University of Michigan ramped up its use of cloud for its computing needs in 2014. In 2019, the team began leveraging the STRIDES Initiative to access cost-effective cloud services, computation and analysis tools, and technical support offered by commercial providers.

"We've been working on scaling. Just to get the computers we need is really expensive and this is one of the advantages of us moving to the cloud. The STRIDES Initiative offers favorable discounts for cloud services which eventually benefits us as everyone's dollar goes further, and the overhead is better," said Albert Vernon Smith, research senior supervisor in the School of Public Health's Department of Biostatistics at the University of Michigan.

Challenges

- Scaling resources
- Getting data back to researchers easily
- Achieving results quickly

Key Results

- Central data repository that can increase reusability
- A high-level technical interface for research studies
- Workflow interoperability across the ecosystem

The Impact of Cloud on TOPMed Projects

All data generated by the TOPMed project are available through NHLBI's BioData Catalyst, a developing cloud-based platform providing researchers with tools, applications, and workflows in secure workspaces. At this time, only a small group of early users can access the system, but it will be more widely available in the future. The primary goal of BioData Catalyst is to build a data science ecosystem that creates efficiencies for research and ultimately leads to novel diagnostic tools, therapeutic options, and prevention strategies for heart, lung, blood, and sleep disorders."[1] In support of this goal, the University of Michigan uses both Amazon Web Services and Google Cloud for its TOPMed projects.

LeFaive noted that a goal of BioData Catalyst is to work across the cloud environments of multiple cloud providers, making data accessible through both Amazon Web Services (AWS) and Google Cloud. "We have been storing data in both cloud environments because we wanted the ecosystem we are creating to work on both clouds," he said.

The University of Michigan is also migrating its Imputation Server to the cloud under the STRIDES Initiative and integrating it with BioData Catalyst. Imputation is the process of filling in missing genetic information, and computing in the cloud is improving this process and allowing reuse of the data. While the Imputation Server is not yet publicly available, the University of Michigan is working with AWS to continue the integration with the BioData Catalyst ecosystem.

The cloud services and tools available positively impact the work the University of Michigan conducts. A few key outcomes include:

- Offsite backup of TOPMed data for security purposes and data retention.
- A mechanism to share data easily from a central location.
- A process for sending data submitted through jobs back to the researchers.
- Availability of compute resources and access to data for researchers at institutions that did not have access before.
- Reusability of the data generated and stored for multiple TOPMed projects and other initiatives, eliminating the need to duplicate the data and reducing costs associated with data transfers or storing locally.

Future Plans For Using Cloud

Many of the TOPMed research studies focus on genetic variants, disorders, and diseases, such as cystic fibrosis or sickle cell disease, that may often be overlooked in research. "These tremendously huge datasets we work on can offer insight into human variation, particularly as it applies to disease. This includes rare variations that are unique to a small number of individuals," Smith said. "We believe that those who work directly with these rare disorders can use this information to better understand what causes are related to different diseases. That is our hope."

The University of Michigan is expanding its work by using different data types and integrating them into their existing work. When the team needs to work at scale or make data available to other researchers, they "definitely turn to the cloud" to take advantage of the cloud storage and tools, and other cloud-based programs like BioData Catalyst.

For both LeFaive and Smith, the future possibilities for TOPMed and the use of cloud for biomedical research are exciting. With projects focused on expanding the sample size of the genomics dataset, such as TOPMed and *All of Us* Research Program, LeFaive looks forward to the challenge of figuring out how the research community can put this vast amount of data to great use and is thrilled to be a part of this field during its exponential growth.

About TOPMed

The National Heart, Lung, and Blood Institute's (NHLBI's) Trans-Omics for Precision Medicine (TOPMed) program aims to generate scientific resources to improve our understanding of heart, lung, blood, and sleep disorders through advancing precision medicine by collecting whole-genome sequencing and other -omics (e.g., RNA transcripts, proteins, metabolites, etc.) data. The University of Michigan manages the TOPMed Informatics Research Center and the Imputation Server, a part of BioData Catalyst.

1For more information, see https://biodatacatalyst.nhlbi.nih.gov/about.