Genomic Data Is Going Google

Ask Bigger Biological Questions

Google Cloud Platform

You know your research could have
a significant scientific impact and answer
questions that may redefine how
a disease is diagnosed or treated.
DNA sequencing data is streaming
in and researchers around the world
are interested in working with you.
If only there were new technologies
that could help you ask bigger questions
and process your data faster.

Google Genomics can help.

Biology & Big Data

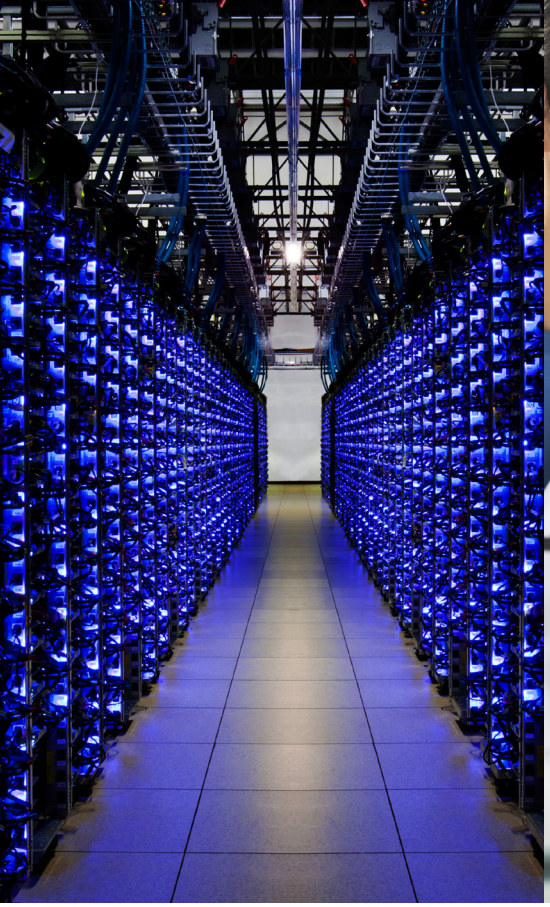
For most of its history, biology has been a qualitative, rather than quantitative, field. Scientists studied a single patient or a single gene or a single trait, because that's what technology allowed for. Today, DNA sequencing advances have made biology a data-rich field for the first time, and the volume of genomic data available is growing exponentially. With better, cheaper tools, you can now look across hundreds of thousands or even millions of patients, studying not just one gene but all of their genes.

Massive new studies are launching all the time. Just a few examples include the Million Veteran Program, the Autism Speaks MSSNG Project, and the Resilience Project. The bottleneck has now shifted from DNA sequence production to data analysis and management. Processing millions of genomes, and handling massive all-by-all comparisons of genomic information across them, takes more compute power than even the best university or private clusters offer.

This is the kind of data-quantity phase transition that Google has seen before with search, video, and email. As scientists scale up their studies and query thousands or millions of genomes, you'll need more scalable technologies than a local compute cluster. Cloud computing provides a useful, elastic resource for manipulating enormous datasets without the time or cost of moving that data from place to place.

That's why we started Google Genomics, to help the life science community organize the world's genomic information and make it accessible and useful. Through our extensions to Google Cloud Platform, you can apply the same technologies that power Google Search and Maps to securely store, process, explore, and share large, complex biological datasets. Whether you are working with one genome or one million, you can ask better questions and get quicker answers.

Processing millions of genomes, and handling massive all-by-all comparisons of genomic information across them, takes more compute power than even the best university or private clusters offer.



The Power of Google Genomics

Google Genomics is based on Google Cloud Platform, which enables users to scale experiments and data analysis at will by using the same high-powered infrastructure that Google runs on. Cloud Platform includes speedy Google Compute Engine VMs for data analysis, the ability to develop new apps easily, and vast storage resources. By using Google Genomics to access Cloud Platform, you can interrogate enormous data sets in as little time as tiny data sets.

Built for bioinformatics scientists, programmers and researchers, Google Genomics enables you to advance your work in the life sciences. If your research involves asking questions of one or a million genomes, you can benefit from our technologies.

Here's how Google Genomics can help:

Get Results Sooner:

Query the complete genomic information of large patient cohorts in seconds. Process as many genomes and experiments as you like in parallel.

Scale to Power Any Project:

No matter how many genomes you're studying, Google Genomics provides access to the power and flexibility you need to advance your work.

Work with Open and Interoperable Standards:

Google Genomics supports open industry standards, including those developed by the Global Alliance for Genomics and Health, so you can share your tools and data with your group, collaborators, or the broader community, if and when you choose.

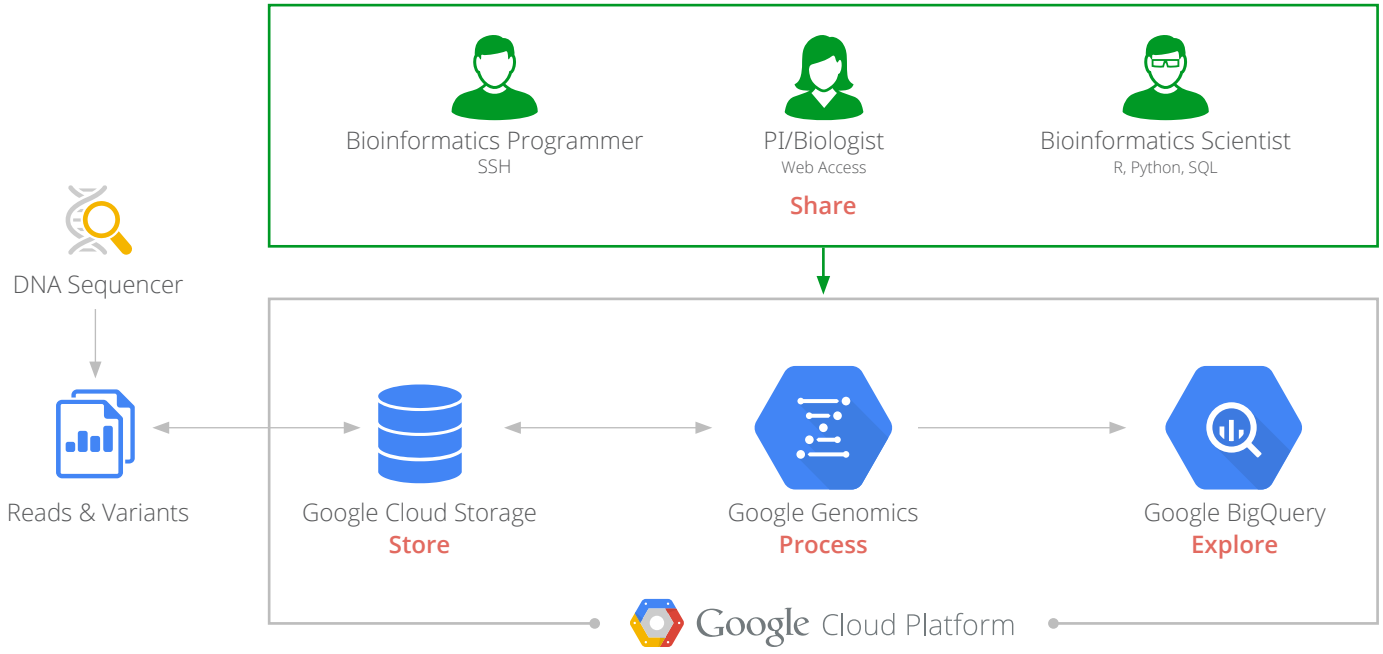
Protect Your Patients and Your Business:

Google's infrastructure provides reliable information security that can meet or exceed the requirements of frameworks such as HIPAA.

Inside Google Genomics

Google Genomics allows you to tap into the leading infrastructure available through Cloud Platform. Take advantage of Compute Engine, our Infrastructure as a Service (IaaS), to run large-scale workloads on virtual machines and pay only for what you use. Our cloud storage offers global edge-caching for fast access to data from any location and reliable versioning. You'll also get sophisticated load balancing along with the performance and consistency of Google's worldwide fiber network.

In addition, Google Genomics features a number of tools to make your life easier. Our open-source tools and cloud compute power can enable users to make a qualitative change in their analyses, switching from more limited batch processing to interactive processing that lets you make changes on the fly and see the effects of those edits immediately.



Here's a Quick Tour:

The Google Genomics API serves as a gateway for researchers to access and implement powerful tools from the Google portfolio. The web-based API is an implementation of the Global Alliance for Genomics and Health standards and allows users to process, store, explore, and share genomic data at massive scale. Designed to foster innovation, the API lets scientists build their own tools, craft completely new ways of exploring genomic data, and securely share data with collaborators next door or halfway around the world.

BigQuery, a tool that enables very fast SQL-like queries of massive data sets, lets you interactively explore large datasets of population variants to find patterns that shed light on disease correlation, epidemiology, and more. It has been tested on the 1,000 Genomes data and yielded accurate results in seconds, far less time than other computational infrastructure requires.

MapReduce, a programming model for processing large volumes of data in parallel, allows you to discover relationships in genomic data through machine learning and other approaches. It was also demonstrated on a 1,000 Genomes data set and was able to quickly and correctly separate individuals by ancestry using principal component analysis.

Security and the Google Cloud Platform

Genomic scientists need solutions that not only scale effortlessly but also meet the most rigorous security and compliance demands. To meet these high standards, Cloud Platform has made a tremendous investment in security. The same security that protects Gmail and Google Apps is used to protect data uploaded to Cloud Platform. Cloud Platform meets or exceeds ISO 27001, PCI, and HIPAA requirements. Based on our deep expertise in data privacy and security, Google can provide end-to-end protection for users' information.

Insight

Remove the Barriers, Unleash Creativity

By Jonathan Bingham, Product Manager
Google Genomics

At Google we've seen a few major revolutions based on data. Biology is reaching a similar transition where data challenges and technology needs are changing rapidly. When you're operating at a small scale, using a compute cluster with a few nodes works just fine. Scale that up by a factor of 1,000 or 1,000,000 and you really need different kinds of technologies to work with the data.

Tackling these large studies — looking at hundreds of thousands or even millions of people — is essential for understanding complex genetic variability. There's a need for algorithms that can handle these large sample sizes and find the signal in noisy data.

The biggest thing we're hoping to achieve with Google Genomics is to change the kinds of questions people can ask. Right now, there are lots of custom-built APIs for bioinformatics scientists and programmers, but they aren't interoperable. We developed an implementation of the API from Global Alliance for Genomics and Health to allow interoperability across multiple genome repositories. Without that kind of tool, scientists have to use FTP download to call data and query it on the local cluster, which can take weeks. Imagine if infrastructure were a solved problem. Imagine if massive, complex data analysis were a solved problem. I think we're going to unleash a tremendous amount of creative insight into science by making this genomic information available quickly and at scale.

The exciting thing is that we don't even know what questions people will ask. The science that's going to come out of this will be amazing. I can imagine a future where a high school science fair project can include analyses on cohorts of millions of patients and find effects that no PhD researcher at Harvard or Stanford ever thought to look for.

Harnessing Big Data for Autism Research

With the MSSNG Project launched in 2014, Autism Speaks aims to sequence 10,000 whole genomes from autism patients and families and to make the data available to the global research community. Autism Speaks, a leading science and advocacy organization, manages the world's largest private collection of autism-related DNA samples.

The MSSNG Project, expected to help connect DNA and clinical data, could easily surpass a petabyte of data. "To manage this scale, we had to reach beyond academia and the life sciences. We had to forge a new collaboration with experts in storing, analyzing, and providing access to big data," said Robert Ring, Chief Science Officer of Autism Speaks.

Working through Google Genomics, the MSSNG Project will use powerful technologies to securely store, process, explore, and share complex biological datasets. Autism Speaks has already uploaded nearly 100 terabytes of data from more than 1,300 genomes onto Google Cloud Storage and has an additional 2,000 samples in the sequencing queue. At completion, the MSSNG database will hold the whole genomes of 10,000 individuals, making it the world's largest single repository of autism-related DNA sequencing data.

Central to the MSSNG project is sharing these data with the global autism research community. Through Google Genomics, the autism community will have web-based access to the MSSNG database to instantly power research projects with new analysis tools and genomic data from thousands of individuals.

"I am immensely excited because for the first time, any scientist anywhere in the world will be able to collaborate and perform analyses with these data in a 'common cloud,'" says Stephen Scherer, MSSNG Program Director. "Thanks to the Google Cloud Platform and our work with the Google Genomics team, this vast sea of information will be made accessible for free to researchers everywhere. The greatest minds in science from around the world will be able to study trillions of data points in one single database."

> "To manage this scale, we had to reach beyond academia and the life sciences. We had to forge a new collaboration with experts in storing, analyzing, and providing access to big data."

— Robert Ring
Chief Science Officer,
Autism Speaks

DNAstack Tackles Massive, Complex Datasets

DNAstack was founded in 2014 to help scientists around the world interpret genomic data faster and more easily.

Without a platform like DNAstack, scientists must complete several complicated steps to link DNA sequence data to a person's disease risk. The process requires many people with various types of expertise, from bioinformatics to genomics to technical skill. "In our experience, the communication between all these different individuals with all these skill sets and different jargon is a significant bottleneck," says Marc Fiume, CEO and founder of DNAstack.

In collaboration with partners, DNAstack aims to build a "sequencer-to-scientist" workflow that would automatically handle routine steps, such as read alignment and variant calling, to get clear results to scientists faster — without needing all those other experts along the way.

The DNAstack platform runs on Google Cloud Platform, which frees up the team to design new apps for expanded functionality. "With Google involved, our expertise does not have to be in data storage anymore — it's going to be in making the data accessible, and that's what we're good at," says Fiume.

They began using Google because their custom-built, proof-of-concept system had a significant performance drop-off when data sets grew past 100 million genetic variants. "When we saw the Google system that was basically the industrial-strength version of what we had built, it was a no-brainer to transfer our platform to Google," Fiume says.

With Google, DNAstack is able to crank through much larger data sets. "We're getting 4-second turnaround times on very complex searches with BigQuery, compared to tens of seconds or even minutes previously," he adds. "That fulfills our need for a scalable system."

The DNAstack team is also using the Google Genomics API, which was developed in accordance with Global Alliance for Genomics and Health standards to promote interoperability and data sharing. "I was so excited when Google stepped in and provided their solution," Fiume says.

> "When we saw the Google system that was basically the industrial-strength version of what we had built, it was a no-brainer to transfer our platform to Google Genomics."

— Marc Fiume
CEO and Founder,
DNAstack

Put Google Genomics to Work for Your Research

Getting started with Google Genomics is simple. Just log into your Google account and create a genomics-enabled project to start using the Genomics API and other tools we've made available via GitHub.

For more information about how you can use Google Genomics to to securely store, process, explore, and share large, complex biological datasets, check out the additional resources at: cloud.google.com/genomics.