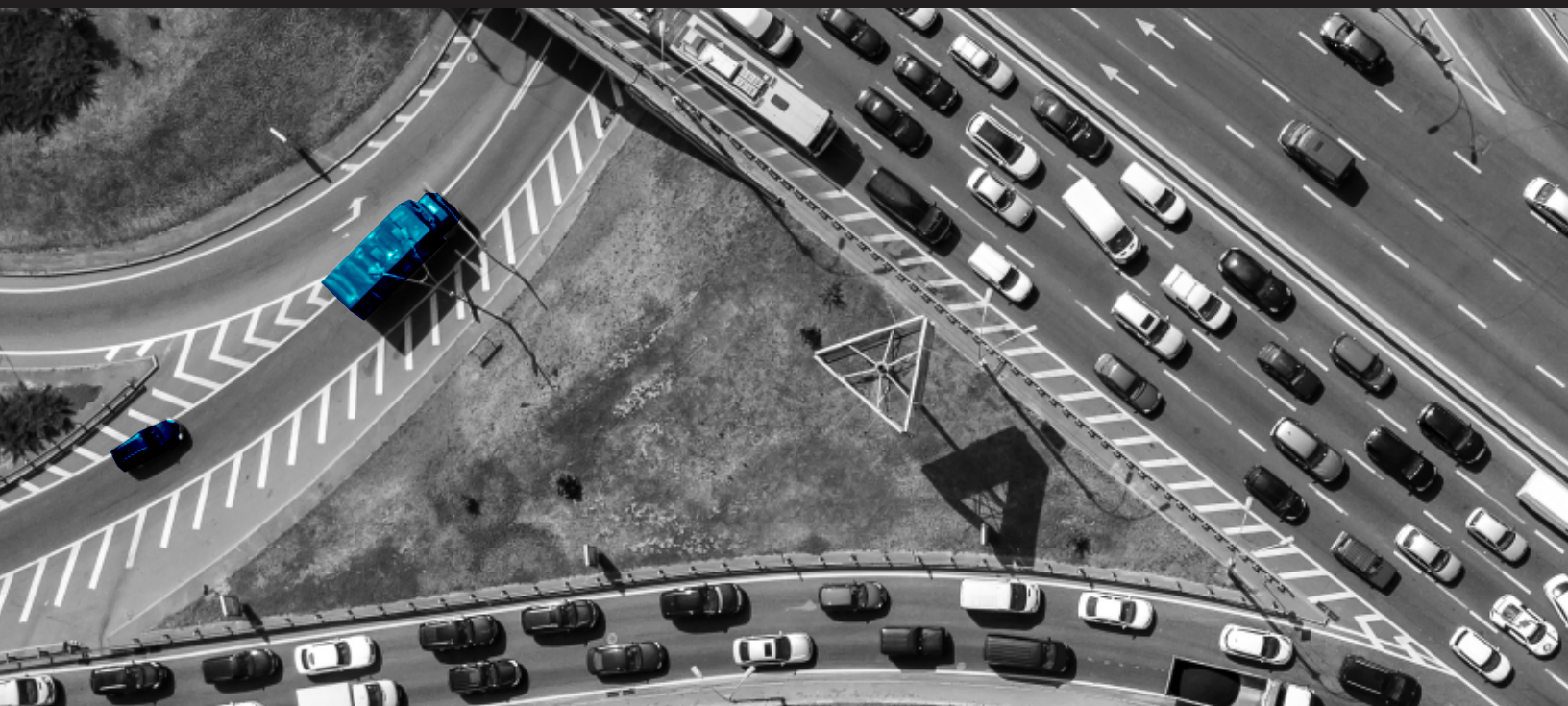
WHITE PAPER

How Regeneron Bypasses Bottlenecks to Iterate at the Scale and Speed of Science

DNAnexus®



Contents

Introduction

Regeneron Genetics Center

At the Regeneron Genetics Center (RGC), one of the world's largest genomic research efforts, success is not necessarily measured by how many jobs they run each day (10,000), how many samples they sequence (500,000 per year), or how many reads they generate each hour (500 million). It's about how many important genetic targets they discover, how much research they enable for more than 100 academic and industry partners and doing right by those who have consented to have their valuable genetic data used for science. The RGC is a uniquely integrated research initiative that seeks to improve patient care by using genomic approaches to speed drug discovery and development.

The infrastructure and automation requirements that enable the scale and speed of operations at the RGC is truly impressive. "About every minute we're seeing a [genetic] variant that we haven't seen in our database before. The scale that we do genetics really is unprecedented," said William Salerno, Senior Director of Genome and Sequencing Informatics at the RGC.

How did the RGC achieve such scale in just seven years? How does it operate now, and how is it preparing for the future?

Salerno shared many of his insights and advice with Endpoints publisher Arsalan Arif at a recent webinar, now available to watch on demand.

1 Data Infrastructure

Data generation is often a bottleneck for large-scale genetics operations. The RGC has overcome this through a combination of local infrastructure, cloud computing, and the workhorse of their production and analysis workflows, the DNAnexus Platform.

Understanding how a system might fail is just as important as understanding how it runs and preparing for failure is a top priority in Salerno's group.

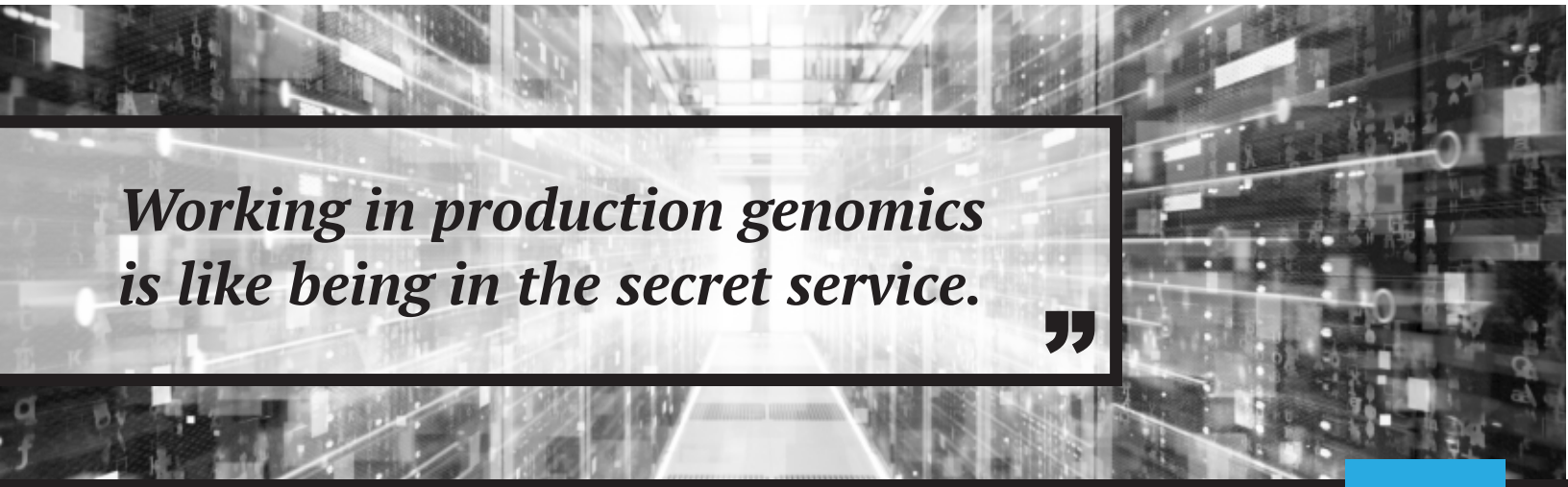
"Working in production genomics is like being in the secret service. You spend 99% of your time planning for crazy things that could go wrong and watching them not go. And then 1% of the time everything goes wrong at once and you have to be prepared for an emergency," Salerno said. "Mistakes will happen. Have you planned for it? Are you prepared?"

The platform also has metadata and version control built in that enables extensive logging and troubleshooting.

"There's a whole database on the backend where every job, every object, everything we've ever done, is queryable. It's organized. You can search it easily. You can restructure it. You can relaunch things. And that's been incredibly important."

As part of their infrastructure, the RGC and DNAnexus have also developed software called GLnexus that enables merging of data on a large scale.

"We have scale tested GLnexus on over a million samples so that we know that it is a solution that is going to serve our needs moving forward."



2 Policy Infrastructure

In addition to the hardware, there are important policy considerations to think about before building data platforms, Salerno said. This includes: cost, compliance, compute, security, and storage.

In terms of cost, Salerno urged viewers to think beyond unit costs of individual tasks and core resources, but to also consider audits, redundancy, troubleshooting, disaster recovery, managed services and other infrastructure costs.

Another cost to consider is compliance.

“Anyone who's had to deal with a FISMA certification or GDPR knows that security and compliance is not a one-time cost, it's perpetual. So, make sure you factor that into your solution.”

The RGC infrastructure addresses this part of the process by including features such as metadata tracking and version control. The platform is fully secure and allows users from around the world to access and work with the data.

For example, the RGC created an autonomous cloud environment for a partner that needed compute space to analyze genomic and phenotypic data related to COVID-19. The partner was able to easily import data into the environment and control who could access it.

“Genetics is a relatively new science. There are not standard answers to every question you would want to ask of the data, and we might not even know all of the questions yet,” Salerno added. “You have to be flexible, mobile, and you have to have a base infrastructure that can support a lot of different use cases and needs.”

3 Operating FAIRly

> *If we want our data to be as useful and actionable as possible, it has to be transparent.*

In its research, the RGC is committed to ensuring that its work is equitable, open access and transparent.

"If we want our data to be as useful and actionable as possible, it has to be transparent. It has to abide by the FAIR acronym: Findable, Accessible, Interoperable and Reusable," Salerno said.

It starts with open-source software and methods and continues with secure data sharing. The right infrastructure can enable more data, more hands on the data, and more value from that data. The RGC makes open-source versions of its different pipelines available to the public so that others can use these tools in their analysis.

"As more and more individuals opt to get their genome sequenced and share that for research purposes, we have an increased responsibility to protect those data and to make sure that it is being used responsibly, shared equitably, and that we are taking care of the interests of those participants."

Furthermore, the RGC can quickly address new research questions as they come up. In one project described during the presentation, scientists had to reanalyze data from 50,000 individuals using a new bioinformatics protocol. With their platform, they completed the analysis in less than two days and for a reasonable price.

"We are equipped to iterate at scale," Salerno said. "We have the infrastructure that lets us support doing, asking questions and iterating at the scale of the science. And I think that's one of the key components to industrial research genomics. It is enabling the scale of the research to be at the scale of the science.

About DNAnexus®

DNAnexus® has built the world's most secure cloud platform and global network for scientific collaboration and accelerated discovery. We embrace challenges and partnership to tackle the most exciting opportunities in human health.

We are scientists, engineers, cloud experts, compliance specialists, and thought leaders dedicated to the acceleration of scientific discovery. The work we do enables many important breakthroughs in health science, the development of life-saving cures, and access to critical data needed for new discovery.

Dedicated to Enabling Your Success.

Start the process with a brief scientific consultation to determine how we can help.

Contact us at: info@dnanexus.com

For more information about DNAnexus solutions, visit
www.dnanexus.com/industries/pharmaceutical-companies
www.dnanexus.com/product-overview/titan