Scaling from Exome to Whole-Genome Sequencing with the DRAGEN™ Bio-IT Platform

The DRAGEN platform enables GeneDx to scale to whole-genome analysis and identify variants with precision.

Introduction

In 2010, Kyle Retterer joined GeneDx, a Maryland-based genomic analysis company. Starting as a bioinformatics engineer, he has supported the rise of the organization from a single-gene assay service company to one that now offers whole-genome sequencing (WGS) and analysis.

Today, he is Chief Innovation Officer at GeneDx where he oversees test development and genomic data science. Recently, the company faced a challenge in transitioning from exome sequencing to WGS. "The amount of WGS data generated by our NovaSeq™ 6000 System is roughly 25 times more data per sample than the amount of exome sequencing data and put a strain on our compute systems and processing times," Retterer said. "We saw the benefits for a more specialized, optimized approach to genomic data processing. The DRAGEN (Dynamic Read Analysis for GENomics) Bio-IT Platform has met our analysis turnaround needs and more."

iCommunity spoke with Mr. Retterer about the evolution of commercial genomic analysis at GeneDx and its adoption of the DRAGEN platform for secondary analysis of WGS data to support the company's product offering and growth

Q: What is the mission of GeneDx?
Kyle Retterer (KR): GeneDx was founded in 2000 by Sherri Bale and John Compton. They were NIH researchers who had developed genetic assays for ultra-rare disorders and wanted to offer them as a service. They gradually expanded from that, adding to the menu of rare disorders and broadening the focus over time as new technologies, such as chromosome microarray and next-generation sequencing (NGS), emerged. While relatively common disease areas, such as inherited cancer and cardiomyopathy, are now major areas of business, the largest growth in the last few years has been in our clinical genomics program focused on exome and genome sequencing.

Q: What is your role at the company?
KR: When I started, my role was to develop data processing and analysis pipelines and tools to support the transition to, and growth of, NGS. In the span of three years, we went from running primarily single-gene assays to small panels to exome sequencing. As exome sequencing and other complex activities have grown, I have become the Chief Innovation Officer responsible for the assay development group, which takes new assays from concept to completion, and for our data science group, which handles all the genomic data, including the recent WGS data.

Q: What bioinformatics platform is used at GeneDx?
KR: We have an on-premise high-performance compute grid that we've scaled up over the years. Typically, every time we buy another sequencer, we expand the compute and storage systems accordingly. We handle data from several NovaSeq 6000, HiSeq™ 2500, and MiSeq™ Systems, plus an iSeq™ 100 System along with Sanger sequencing and other non-sequencing platforms.

In 2011, we were trying to build up our exome sequencing analysis capabilities. It was largely an unsolved problem at the time. Given an exome's worth of data, how do we analyze it? There are several tools on the market now, but at the time the options were very limited. It was a buy vs. build decision, but there was nothing to buy, so we built our bioinformatics platform ourselves, working hand in hand with the geneticists here at GeneDx. We've made use of open-source tools such as GATK and BWA for secondary analysis where it makes sense. We've also built some custom programs, such as variant callers, to handle some clinically important scenarios. On top of that, we also developed a proprietary tertiary clinical analysis platform.

Kyle Retterer is Chief Innovation Officer at GeneDx in Gaithersburg, Maryland.

Q: What issues did you face when you started to produce WGS data?
KR: It takes less than a day to process exome data for analysis on commodity hardware. Putting the WGS data through our current architecture, just letting the genomes run through as though they were exomes, took two weeks.

One way to improve that would be to dedicate more compute resources from the high-performance computer (HPC) to processing genomes instead of exomes. Doing that can create bottlenecks, however, where genomes could end up hogging all our resources for too long.

In contrast, if we just let it run and waited for it to finish, that would impact sample turnaround time negatively. It would also be a problem if we had poor quality data off the sequencer. For instance, we might not detect a contaminated sample until those two weeks of processing were over.

> "...DRAGEN platform enables us to scale the analysis architecture and improve the speed to handle the growth of our WGS business...it also gives us flexibility because we are not purchasing depreciating capital equipment and are only paying for the level of compute that we need..."

Q: What were your bioinformatics options?
KR: In addition to the DRAGEN platform, we considered buying more traditional hardware, shifting to cloud, or adding GPU-based systems. To support one genome, we would need about 25 times more hardware than what was required for one exome. If we had purchased 25 times more hardware for WGS, but only got half of the genome sample volume we were expecting, then the hardware would sit idle most of the time.

Cloud was another option. We could scale infinitely with cloud, up to a point. For WGS data, the cloud doesn't make much sense because of the amount of data we would need to transfer. Having on-premise computing to do the heavy lifting would be more time and cost effective. It would also make our data security team happy.

We had already looked at some of the GPU-based systems. They were a little too specialized for us as our bioinformatics development team is focused more towards clinical applications rather than computational algorithm implementation.

Q: How did you decide on the DRAGEN platform?
KR: Our ultimate choice, the DRAGEN platform, has been around for a few years. We had already talked to the DRAGEN team several times before it became part of Illumina. Adding the DRAGEN platform enables us to scale the analysis architecture and improve the speed to handle the growth of our WGS business. The DRAGEN platform also gives us flexibility because we are not purchasing depreciating capital equipment and are only paying for the level of compute that we need, somewhat like you would with cloud, but with the advantage of being on-premise. It also implements much of the same GATK-like workflow that we have already been running and integrates well with our existing pipeline infrastructure both up and downstream.

All our WGS is currently performed on the NovaSeq 6000 Systems with the data going through the DRAGEN pipeline. Our other data are still processed through the standard HPC system. The DRAGEN platform is integrated directly into our SLURM HPC system, meaning that we can take advantage of DRAGEN processing speed on an as-needed basis. This also made it easy to integrate into our existing NGS pipeline.

Q: How has the DRAGEN platform performed?
KR: The speed of the DRAGEN platform is as fast as promised. We are able to process whole genomes in a few hours. The DRAGEN platform has exceeded our expectations in the quality of the variant calls, which is the ultimate measure for us.

Very rare variants are important in the analysis of Mendelian disorders. However, random stochastic noise in the data are hard to filter out. Consider a WGS trio test: if we have an extra 20 variants from noise, and they all look like *de novo* mutations, then we have to look at each of those 20 putative mutations and figure out whether it is relevant. Is it a real variant or is it just noise? This leads to extra analysis time and extra Sanger confirmations, increasing our costs and decreasing our turnaround time.

We were able to clean up most of that extra noise using the DRAGEN platform out of the box, with only minor parameter tuning. We have fewer variant calls that need to be evaluated, and we're not losing anything as a result. On top of that, we’re seeing slightly better sensitivity for WGS processed through the DRAGEN platform than our previous pipeline.

> "The DRAGEN platform is integrated directly into our SLURM HPC system, meaning that we can take advantage of DRAGEN processing speed on an as-needed basis."

Q: How did the DRAGEN platform compare to your existing analysis pipeline?
KR: We benchmarked the DRAGEN platform against our current pipeline using Genome in a Bottle samples. For indels, we got a small boost in recall and a significant increase in precision. The recall for indels on our current pipeline is around 98% and we saw it go up to 98.5% with the DRAGEN platform. The significant

gain was in precision for indels, which went from 85% to 99%, and that was uniform across all the samples.

Q: How does the DRAGEN platform fit in with your existing architecture?
KR: Like most people, our pipeline is broadly similar to GATK "Best Practices." The DRAGEN platform is GATK-like, enabling us to integrate it with our existing compute grid.

We use a layered approach. We have custom algorithms that we've developed. We can send off some jobs to the DRAGEN platform and others to traditional compute nodes as needed.

Our HPC architecture is the SLURM Workload Manager, and we didn't have any real issues getting that hooked together. We can pick up workflow description language (WDL) workflows and execute them through Cromwell with the DRAGEN platform. It just plugs right in.

> "With the DRAGEN platform onboard, we can now consider offering rapid whole-genome analysis, too."

Q: How does the cost compare?
KR: The real saving is that we're not buying new hardware to perform WGS analysis. It's additional capital that we don't have to dedicate to computing. When we bought another NovaSeq 6000 System, we didn't have to go out and buy more compute blades for it. Instead, we chose the DRAGEN platform. If we were to double our WGS volume, we would just increase our license for the DRAGEN server and there would be no need to bring in any more hardware.

Q: What is the future of genome analysis?
KR: We offer an "express exome" service that has a seven-day turnaround time. It has been a successful program for us and our clients. With the DRAGEN platform onboard, we can now consider offering rapid whole-genome analysis, too. It's something Rady Children's Institute for Genomic Medicine has been doing. They've been using the DRAGEN platform as well.[1]

We're seeing more analysis for Mendelian disorders shift towards exome or genome as a first-line method rather than going with a targeted approach. Instead of a tiered testing approach, people can order an exome or genome and start there then follow-up with more targeted testing if needed. This is often a more cost-effective approach and provides a quicker path to diagnosis.

Ultimately, I see more and more genetic testing moving towards WGS as a first-line test. Someone might order a targeted analysis, but the data generation will be whole genome. If the cost of genome sequencing drops very low, as some are predicting, there will eventually be no reason to run exomes. Over the next several years, genomes are where everything is going, and we want to be ready for that future.

Learn more about the software and systems mentioned in this article:

DRAGEN Bio-IT Platform, www.illumina.com/products/by-type/informatics-products/dragen-bio-it-platform.html

NovaSeq 6000 System, www.illumina.com/systems/sequencing-platforms/novaseq.html

MiSeq System, www.illumina.com/systems/sequencing-platforms/miseq.html

iSeq 100 System, www.illumina.com/systems/sequencing-platforms/iseq.html

References

1. Rady Team Automates Rapid Pediatric Sequence Interpretation for Rare Disease Dx. *GenomeWeb*. April 24, 2019. Accessed April 24, 2019.

For Research Use Only. Not for use in diagnostic procedures.