



Photo by Illumina

Independent testing proves DRAGEN provides a vastly more comprehensive genome

The new study published in Nature Biotechnology validates all of the software’s germline algorithms

ILLUMINA VICE PRESIDENT of bioinformatics James Han has been in the industry for decades—so he’s witnessed every stride the company has made in that time. “Illumina has worked tirelessly for the last 20 years to improve sequencing,” he says. “We brought the cost of a genome from hundreds of thousands—maybe even millions—down to \$200. Now, as labs move toward whole-genome sequencing, the bottleneck is no longer the lab workflow; it’s really in the informatics, in getting the insights from the data. This is why this work is so important.”

For over six years, DRAGEN Secondary Analysis software from Illumina¹ has stood out among its competitors for its speed and accuracy. But one comment its developers have heard from members of the scientific community has been that, because it’s a commercial product that uses proprietary code, the inner workings of its algorithms—the secrets to its success—are invisible to the user.

Senior Director of Bioinformatics Severine Catreux has even encountered the assumption that the DRAGEN team relies on open-source components to achieve such high

performance and rapid algorithmic evolution. “However, our approach is fully independent,” she says. “We develop all methods in house, from initial concept through to complete implementation.”

This perception needed to change. “Illumina believes in objective review of our methods,” Han says. “So, we’re being open with the scientific community about how the algorithms work.”

A new peer-reviewed study in the October 2024 issue of *Nature Biotechnology* publishes the results of an unprecedented third-party validation by the Human Genome Sequencing Center at Baylor College of Medicine of every single one of DRAGEN’s germline sequencing algorithms, rigorously comparing the software against other tools on the market and measuring it against industry-standard benchmarks.²

The most challenging genes require a custom solution

One of the study’s lead authors, and head of the validation effort at Baylor, is associate professor of molecular and human genetics Fritz Sedlazeck. His

1. illumina.com/products/by-type/informatics-products/dragen-secondary-analysis.html
 2. [nature.com/articles/s41587-024-02382-1](https://www.nature.com/articles/s41587-024-02382-1)

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

© 2024 Illumina, Inc. All rights reserved. All trademarks are the property of Illumina, Inc. or their respective owners. For specific trademark information, see www.illumina.com/company/legal.html.

research focuses on complex parts of the genome that are linked to rare diseases and cancer, and his group has never balked at the prospect of delving into even the largest and most challenging of these regions. He says, “I’m very curious and driven by the fact that we should study all types of variants that occur in our genomes, not just the small ones.”

Back in the mid-2010s, he found that the secondary analysis software available at the time often misreported these regions, and he concluded that better accuracy would require specialized algorithms and targeted variant callers. Eventually he got in touch with the DRAGEN team at Illumina to provide feedback, though he stresses that he’s not partial to any one company, saying, “I just want to get the science right.”

The DRAGEN team prioritized a shortlist of genes of greatest interest to the scientific community because of their potential relevance to medical research, and in the past few years, they’ve rapidly developed tailored solutions for characterizing these genes—for instance, *GBA*, which is linked to Gaucher disease and Parkinson disease; or *LPA*, the copy numbers of which are directly correlated with cardiovascular disease risk.

These high-priority genes are often present in multiple copies and/or located in highly homogenous regions of the genome. This repetitive genetic code can vary widely from person to person, and it constitutes a booby trap that confuses genomic sequencers during the demultiplexing step, which stitches the complete genome back together from the smaller fragments created during library preparation.

DRAGEN now includes over a dozen specialized variant callers, tailor-made to accurately read these most challenging sections of the genome. (For more detail on DRAGEN targeted callers, follow this link.³) Sedlazeck explains that researchers from various institutions have done well developing individual variant callers, but few to none of them are designed to work together. This paper shows that with DRAGEN, “you have a really comprehensive genomics workflow built together, where each component efficiently communicates with the others. DRAGEN is one of the few cases—or the only case—where these components are connected, and benefit from the innovations across them.”

Accurately reflecting the world’s genetic diversity—and showing our work

Another recently developed feature—and major

strength—of DRAGEN is its pangenome reference, sometimes called a multigenome reference or graph genome. This breakthrough aims to counter the Eurocentric bias present in most of the available reference data to date.

The reference genome most often used by researchers today, GRCh38, has its roots in the Human Genome Project of the 1990s. It has served researchers well for a long time, “but honestly, in my opinion it’s nothing more than a kind of ruler,” Sedlazeck says. Unique genetic sequences that appear in non-European ethnic groups are poorly represented or even absent in this data. “We compared every human genome to that ruler, and it doesn’t cope well with different variants across different ancestries.”

The pangenome reference in DRAGEN compares newly read genetic sequences against other known variations in that position, drawing from sample data that better captures the spectrum of people groups across the world. Sedlazeck says this feature, combined with the power of machine learning and Illumina’s proprietary algorithms, is a significant milestone that improves variant calling at any scale—from single-nucleotide variants all the way up to large copy number variations. “With this paper, we hope to show that DRAGEN has matured a lot in identifying more complex variants. This is a significant jump forward to get the whole picture of what a genome looks like in a certain individual; hopefully this catches on, and more and more studies will make use of it.”

In the spirit of scientific transparency, the paper in *Nature Biotechnology* references all the data and truth sets used, and all the command line parameters necessary for other computational biologists and geneticists to reproduce the same results. DRAGEN is built to handle everything from individual samples to population-size studies (such as the UK Biobank and the National Institutes of Health All of Us Research Program), encompassing hundreds of thousands of them.

The numbers don’t lie: DRAGEN’s speed and accuracy is in a class of its own

DRAGEN demonstrated higher accuracy and faster reads compared to eight other variant callers in this study—when tested on the same fully characterized benchmark data from the US National Institute of Standards, one third-party caller had 144% more false-positive and false-negative errors than DRAGEN; another showed 470% more errors.

3. [illumina.com/company/news-center/feature-articles/DRAGEN-targeted-callers.html](https://www.illumina.com/company/news-center/feature-articles/DRAGEN-targeted-callers.html)

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

© 2024 Illumina, Inc. All rights reserved. All trademarks are the property of Illumina, Inc. or their respective owners. For specific trademark information, see www.illumina.com/company/legal.html.

And it enables this laser-sharp analysis in just 30 minutes of computation time. “DRAGEN can go from raw reads to VCF files to variant reports in 30 minutes for the whole human genome,” Sedlazeck says. “For many people, this takes, like, half a week.” These comprehensive, scalable methods are necessary to detect medically relevant variants of all types, and to discover novel genetic markers and drug targets.

Catreux has been with Illumina ever since 2018, when the company acquired Edico Genome, the original developers of DRAGEN. She says that acquisition completely changed Illumina’s software landscape: “I’ve seen the whole evolution of DRAGEN, starting from

its infancy when nobody knew about it, to the point where we talk about it worldwide. We completely take ownership of how we help our customers with data processing, and we’re having a real impact in improving human health.”

She and Han are proud to share the latest fruits of their labor with the world—and they’re excited to tackle the challenges that lie ahead, all the while listening to customers and using feedback from leading labs like Baylor. Sedlazeck says, “I think the genomics you can now accomplish with this framework is astonishing. And I’m really looking forward to seeing how scientists will use it.” ♦