Photo by Illumina

Illumina leads the industry in genomic AI

The company has proven its ability to decode the unknown regions of the genome

By Kyle Farh, distinguished scientist and vice president of Artificial Intelligence

FOR MOST OF THE HISTORY OF GENOMICS, the biggest hurdle was simply getting the data. It took 13 years and billions of dollars to map the first human reference genome. Illumina built its name and reputation on improving the technology, and now we can sequence a genome in hours, with exceptional accuracy, for an infinitesimal fraction of what it once cost.

The industry now generates upward of 40 billion gigabytes of genomic data every year.[1] Getting the data is no longer the problem—making sense of it is. Compared to reference genomes, each of us has about 4 million genetic variants that make us an individual, and the function of 99.9% of those variants is unknown. We know that devastating genetic diseases can be caused by differences as tiny as a single one of them. But how do we sort through the avalanche of data to separate the benign variants from the pathogenic ones?

The task is beyond any human. The key to the next era of genomic science is artificial intelligence (AI).

AI hype has reached a fever pitch in the last couple years, with the widespread adoption of large language models like ChatGPT. It's hard to deny that applications like these, which draw upon the corpus of all published works to generate new text, images, and even video, are truly revolutionary. The difference is, those human-made sources begin in a highly structured state. The biology of the genome is orders of magnitude more complex.

But we can crack the code. We're already doing it.

We recently announced BioInsight, a new business within Illumina that brings the company's software, informatics, AI, pharma data partnerships, and large national genomics initiatives teams together to meet the industry's demand for large, comprehensive data solutions. I'm thrilled to continue leading the AI branch of this business, which has been developing the world's leading genomic AI algorithms for seven years and counting.

We apply these algorithms across the genomic workflow. During sequencing, they correct errors for quality control and translate raw signals into data faster. In secondary analysis, they detect variants more accurately. But it's in tertiary analysis—interpreting the significance of those variants, narrowing down the handful that can cause disease—where our flagship algorithms really shine. And it's not just about the

1. genome.gov/about-genomics/fact-sheets/Genomic-Data-Science

technical capability. It's about the impact genomics AI can have for human health.

Splice AI, PrimateAI-3D, and PromoterAI

In 2019 we released SpliceAI. This deep neural network identifies cryptic splice mutations, which are variants that do not themselves code for proteins, but that specify the protein coding sequence. Over 98% of the genome is noncoding—but noncoding variants can still be pathogenic, so we ignore them at our peril. Despite lying beyond the splice nucleotides, these variants still disrupt the normal pattern of mRNA splicing and are known to play a role in childhood developmental disorders and in cancer. Comparable third-party tools at the time predicted splice junctions for pre-mRNA transcripts with 22% to 30% accuracy. SpliceAI is 95% accurate.[2]

We followed that success in 2023 after realizing that the best way to predict which protein-coding variants are pathogenic in humans…is to look beyond humans.

Homo sapiens shares more than 90% of our DNA with other primates, even though our most recent common ancestors lived about 60 million years ago. At Illumina, we reasoned that if we cross reference as many primate genomes as possible, we could find the variants they have in common—and we can infer that if these sequences have not been eliminated through natural selection, they must be benign in humans. In turn, ruling out these benign variants helps us zero in on the pathogenic ones we're looking for.

So, we undertook the largest primate sequencing effort to date: over 800 individuals from 233 species across all 16 families from every corner of the globe. With this data, our next flagship algorithm, PrimateAI-3D, was effectively trained on evolution itself, and we published our findings on PrimateAI-3D in *Science* that June.[3]

Most recently, just this summer we further expanded our investigation into the noncoding genome with PromoterAI. This deep learning model finds pathogenic variants in promoter regions, which are regulatory sequences that precede a gene, defining where gene transcription begins and enabling it to make RNA and proteins.

Even if the protein-coding sequence of a gene is free from variants, mutations in that gene's promoter region can prevent it from being properly expressed. In fact, the accompanying paper we published in *Science* found that promoter segments contribute up to 6% of the genetic causes of rare diseases.[4] But that same research study demonstrated that, when used together, PromoterAI, PrimateAI-3D, and SpliceAI can effectively double the diagnostic yield compared to using protein-truncating variants alone.

PHOTO BY ILLUMINA

Unmatched in the market

Innumerable companies are now investing in AI technology—even genomic AI. But for many of our peers, genomics is crowded among several other projects they're pursuing at once. At Illumina, genomics is our core business. It's been our expertise for 27 years, and we proudly put our algorithms up against the best in the world.

Here's one of the most astounding facts about PrimateAI-3D: Its study in *Science* compared it against 15 other machine-learning methods to detect pathogenic variants across six different clinical benchmarks and four cohorts (the UK Biobank, a neurodevelopmental disorders cohort, an autism spectrum disorders cohort, and a congenital heart disease cohort). Not only did PrimateAI-3D detect the most variants in every category—not a single competitor scored second place in more than one category.[3]

At the same time, we're not going it alone. The data challenges of genomics (and multiomics) will require an entire ecosystem of stakeholders to drive insights that make an impact on human health. We're building partnerships, such as the Alliance for Genomic Discovery, to share proprietary data. We're working with NVIDIA to advance platforms for multiomic data

2. illumina.com/science/genomics-research/articles/predict-splicing-primary-sequence-deep-learning.html
3. science.org/doi/10.1126/science.abn8197
4. science.org/doi/10.1126/science.ads7373

analysis and interpretation, and with Tempus to leverage their data platform to help train our algorithms. And we're combining our knowledge of AI-based analysis techniques with that of the brilliant minds at AstraZeneca.

Pharmaceutical companies are starting to realize the potential that AI tools hold for drug development. Bolstered by the collaborative strength of the talented minds across its constituent departments, Illumina's new BioInsight business will develop the large-scale data assets and AI solutions that our partners and customers can use to deliver the next wave of biological discoveries.

I'm extremely proud of all that our AI team has accomplished, and I can't wait to share what they're working on next. The evidence is undeniable: The next era of genomics is here, and we're leading the way. ➤

Use of forward-looking statements

This article may contain forward-looking statements that involve risks and uncertainties. Among the important factors to which our business is subject that could cause actual results to differ materially from those in any forward-looking statements are: (i) challenges inherent in developing and launching new products and services, including modifying and scaling manufacturing operations, and reliance on third-party suppliers for critical components; (ii) our ability to manufacture robust instrumentation and consumables; and (iii) the acceptance by customers of our newly launched products, which may or may not meet our and their expectations, together with other factors detailed in our filings with the Securities and Exchange Commission, including our most recent filings on Forms 10-K and 10-Q, or in information disclosed in public conference calls, the date and time of which are released beforehand. We undertake no obligation, and do not intend, to update these forward-looking statements, to review or confirm analysts' expectations, or to provide interim reports or updates on the progress of the current quarter.