AI in genomics: Progress through innovation

Patients benefit from faster technical breakthroughs in genomics

NetApp

NVIDIA

High-performance GPU computing on genomic workloads can provide 30 to 50 times faster secondary analysis compared with other approaches.

With a growth rate in double digits and a global market forecast of more than US$62 billion by 2026,[1] genomics is one of the fastest-growing industries. But the real story is about more than just market share. Science and health experts are calling genomics a revolution that's only just begun. The ability to sequence DNA quickly and easily has opened up an array of applications in personalized medicine, cancer research, drug discovery, and more. COVID-19 is also highlighting the importance of sequencing as scientists work to understand the virus.[2]

It took more than 10 years to sequence the first human genome, at a cost of billions of dollars. Now, genome sequencing is widely available. It generally costs under $600 and the results take less than a week.[3] And the price keeps dropping. In late 2018, Veritas Genetics offered whole genome sequencing and interpretation for just $199, with a 2-day turnaround.[4] Not to be outdone, Veritas competitor BGI in China recently said it will offer results for $100.[5]

This consumerization of genomics is largely the result of faster, cheaper technology. But sequencing is only the first step, and over-the-counter tests yield limited information. Clinical whole genome sequencing (WGS) is where researchers, data scientists, and other experts are using artificial intelligence (AI) to improve the accuracy of disease diagnosis and precision medicine—the identification of the most effective patient treatment based on genetic, environmental, and other factors. The benefits include earlier disease detection, tailored treatment plans, reduced radiation doses, and much more.

Whole genome sequencing generates 300GB to 1TB of data per patient.

Challenge

Overall AI efforts in healthcare and genomics have barely scratched the surface of what will eventually be possible. The fundamental challenge of genomics is to take mountains of human sequence data and figure out which differences are important. Which gene variants, or combinations of genes, contribute to various medical conditions, and how can genomic information be used to individualize patient treatment?

Furthermore, AI-based algorithms are extraordinary in their ability to interpret complex data. However, their power and complexity can also result in spurious or biased conclusions when applied to human health data.6 Without careful consideration of the integrity of a trained AI system, the practical benefit of these systems in clinical diagnostics can be limited.

To get the most value from WGS in a clinical setting, operators must do it quickly, accurately, and inexpensively, largely because of technical factors. Most often, only limited genetic data is available to clinicians because of the time and cost to process and store WGS data. Since the amount of data generated per patient can be 300GB to 1TB, processing alone can take several days.

The data generated by WGS requires massive amounts of compute power, storage, and data management that can easily become a bottleneck. Although 300GB might seem manageable by itself, with thousands of subjects or patients, it scales quickly, and it could take years to get through every record. To keep up with the increased demand, organizations must be able to handle more sequencing jobs in less time without sacrificing accuracy or security.

In genomics, data management is a much more significant challenge compared to medical imaging or digital pathology. With larger datasets, WGS creates data management challenges in every part of the lifecycle. Although the file formats used in genomics are standardized, there is no equivalent to a picture archiving and communication system (PACS) or vendor-neutral archive (VNA) for managing sequence data. In other words, genomics is unique. Methods to store genomic data in electronic health record (EHR) systems are being investigated, but existing EHR databases can't effectively store these very large data files. It's likely that EHR systems will need to store shortcuts to the data files in external systems that are better suited to the task.

Once researchers start to systematically rethink their data needs, processes, and off-site systems, it will be natural to consider more modern solutions that involve cloud and integrated, high-performance storage. Currently, genomics data management operates from silos of data in isolated research centers. In some cases, there are data services models with shared data access.

In many big data applications, data loses its value over time. But in genomics, data never loses its value. Intermediate data generated during analysis is often used for reanalysis, enabling new insights. For example, data scientists in pharmaceutical companies frequently reanalyze genomic files to try to discover new mutations or biomarkers. By definition, this reanalysis poses a new set of scalability requirements.

> "What all of our customers have in common is the belief that a genomicsdriven approach will improve decision making in terms of truly understanding biological drivers of disease, and a hope that this approach could be a step change in R&D productivity, bringing more efficacious drugs, sooner, to patients who desperately need them."
>
> Rob Brainin
> CEO, WuXi NextCODE

Solution

After a person's genomic information is collected, AI and machine learning help analyze the data to determine personalized treatment options. The results can impact pharmacology, oncology, infectious diseases, and many other areas of healthcare. In practice, AI applications in genomics tend to target tasks that are difficult or impractical to complete using human intelligence alone. AI is also useful for tasks that are prone to error with standard statistical approaches, including variant calling, genome annotation, variant classification, and phenotype-to-genotype correspondence.

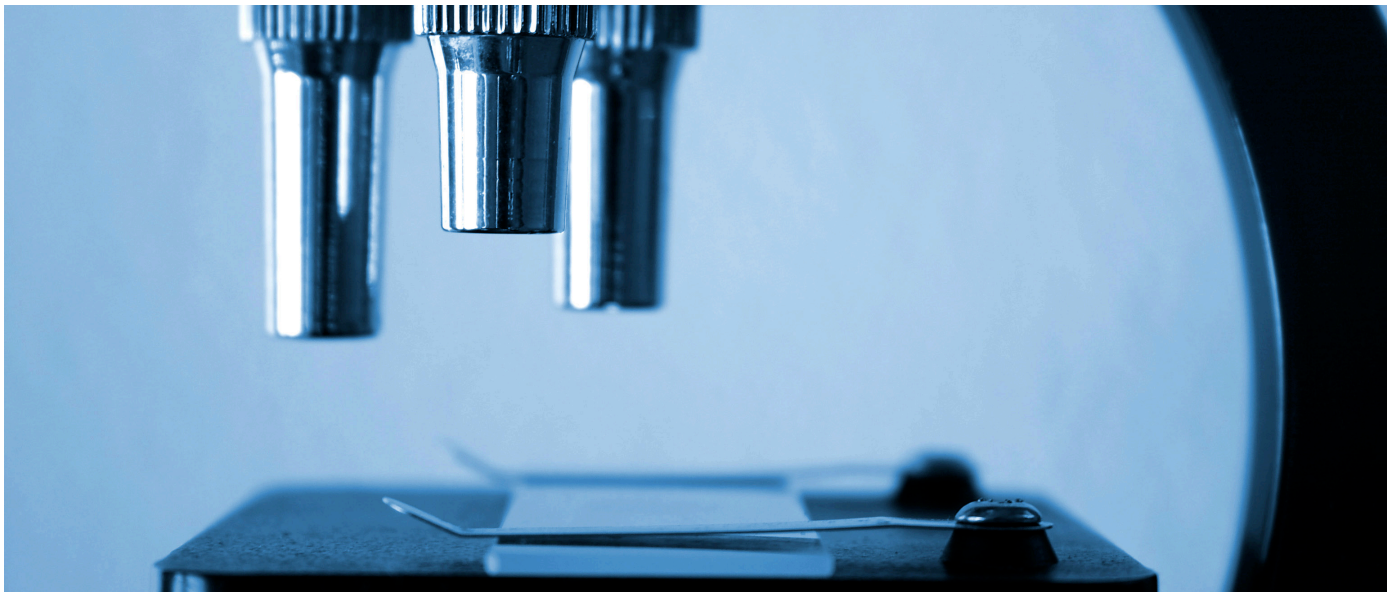
The result has been some innovative and very powerful AI projects in genomics. For example, NetApp customer WuXi NextCODE has created a unique platform specifically to organize, mine, share, and apply genomic data to improve human health. Over the past 20 years, it has amassed the world's largest database of human genome sequences. WuXi NextCODE uses NetApp® Cloud Volumes Service to help researchers quickly generate insights from processing unprecedented amounts of genomic data, discovering new ways to address disease.

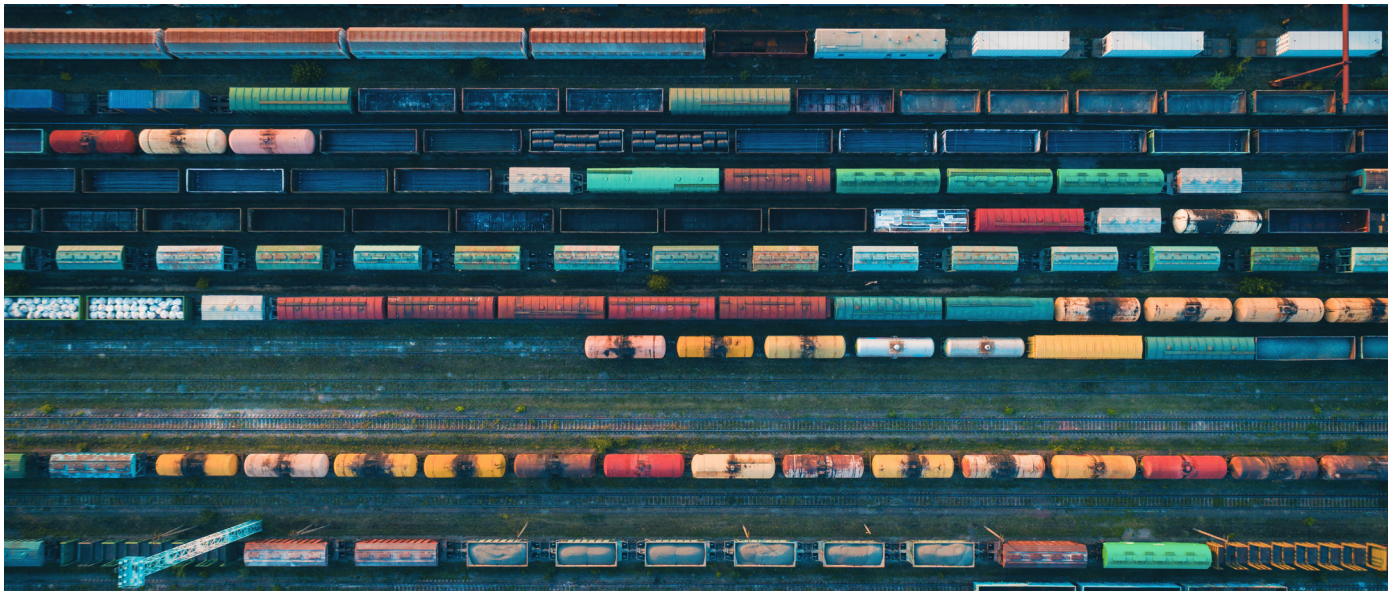
Across four continents, the company's partners can correlate genomic and phenotypic data at unprecedented scale and speed. WuXi NextCODE makes it possible to rapidly access this information to enable biopharma researchers to discover new drug targets and biomarkers that have the potential to improve healthcare worldwide.

In the case of ICON plc, headquartered in Dublin, Ireland, it's all about data, mountains of it, that drives clinical trials, accelerates proof of efficacy, and hastens the delivery of lifesaving new medicines and therapeutic devices to market. As a global leader in contract clinical research, ICON's data-modeling team runs algorithms on a software-as-a-service grid platform deployed on NetApp E-Series systems. This advanced data modeling requires extreme performance.

As another example, AstraZeneca accelerates pharmaceutical science with the cloud. By partnering with NetApp, AstraZeneca has been able to design and implement a data strategy for its hybrid multicloud environment. This dynamic movement of data to any cloud from any cloud allows faster analysis and discovery.

Another solution focuses on improving the speed of genomic analysis. NVIDIA Parabricks provides 30 to 50 times faster secondary analyses of sequencer-generated FASTQ files to variant call files. Additionally, Parabricks achieves results that are equivalent to the results of common secondary analysis tools like GATK4 and DeepVariant, while significantly increasing throughput. By using GPU-accelerated computing, Parabricks can provide throughput comparable to about 40 to 50 CPU servers with one GPU server, reducing IT management overhead and operating costs, including power and cooling. Parabricks is rapidly evolving as a preferred solution for analysis and reanalysis of next-generation sequencing data, vastly improving current pipelines in terms of efficiency, while also enabling user-driven customization.





Core Scientific gave researchers free access to AI infrastructure as a service, powered by NetApp ONTAP AI and NVIDIA Parabricks, for GPU-accelerated coronavirus-related research.

In combination with NetApp ONTAP® AI, Parabricks makes it possible to deploy a fully integrated, end-to-end AI solution tuned for compute and data-intensive genomics workloads. For example, in response to the impact of COVID-19, Core Scientific gave researchers free access to AI infrastructure as a service, powered by NetApp ONTAP AI and NVIDIA Parabricks, for GPU-accelerated coronavirus-related research.

These types of solutions are not prototypes. There are many real-life examples of AI improving genomics workloads—and helping patients. For example, San Diego researchers recently set a record for sequencing to diagnosis in a neonatal/pediatric ICU. The team created a pipelined workflow that included Illumina sequencers and Diploid Moon to automatically filter and rank the possible causative gene variants, compressing the entire workflow down to about 19 hours.

Many researchers are turning to the cloud to gain access to necessary storage and compute resources. Because of the large datasets and the amount of compute required, genetics researchers use a hybrid cloud approach to gain access to the necessary storage and other resources.

In clinical settings, secondary and tertiary analysis of patient genomic data has to be done close to the sequencer, at least in urgent cases. Hospitals and laboratories that want to perform sequencing for clinical use will need to have local compute and high-performance storage, and the results of genetic analysis must be integrated with EHRs so that they are quickly available to clinicians.
Regardless of the data demands, companies need very high performance and scalability, with the ability to integrate to appropriate cloud services.

Parabricks can provide throughput comparable to about 40 to 50 CPU servers with one GPU server.

Benefits

High-performing computers and AI systems make the analysis of genes—and the mutations that cause problems—much easier and faster to deal with. This genetic analysis helps the medical community make better-informed decisions about a myriad of health problems, including untold cancers, organ transplant rejection, cystic fibrosis, and much more.

In addition to the specific clinical benefits—medical insights that span healthcare and research—the more technical benefits of AI solutions and integrated data and cloud systems are helping transform genomics breakthroughs.
For example:

- **Accelerate genome sequencing.** With innovations like the NetApp and Parabricks solutions, companies can speed the performance of GPU-accelerated genomic sequencing an average of 50 times compared to CPU-only solutions.
- **Maximize throughput and minimize turnaround time.** Perform more genome analysis operations in less time with industry-leading, cloud-connected all-flash storage from NetApp and NVIDIA DGX servers.
- **Improve accuracy and security.** Protect sensitive genomic data while improving test accuracy for AI-powered precision calculations.
- **Lower total cost of ownership.** Data-efficiency technologies and a 25 times capacity advantage versus competitive systems mean lower TCO.
- **Simplify design and accelerate return on investment.** Accelerate ROI with simplified integration, automation, and orchestration of data in clouds and on premises.

Conclusion

The fact is that AI systems are outperforming traditional analytic methods. The FDA is giving clearance for a variety of AI-related clinical diagnostics. The building blocks are finally readily available: large functional genomics datasets, in conjunction with advances in AI algorithms and the GPU systems used to train them. Increased productivity and improved results are clearly visible. Currently, the most promising applications of AI in clinical genomics appear to be the extraction of deep phenotypic information from images, EHRs, and other medical devices that help inform downstream analysis.7 Deep-learning algorithms have also shown great promise in a variety of clinical genomics tasks. It is possible that more generalized AI tools will become the standard, especially where inference from complex data, such as variant calling, is a frequently recurring task.
Finally, the combination of breakthrough technology, AI best practices, and human persistence is starting to pay off in the field of genomics.

About NetApp

In a world full of generalists, NetApp is a specialist. We're focused on one thing, helping your business get the most out of your data. NetApp brings the enterprise-grade data services you rely on into the cloud, and the simple flexibility of cloud into the data center. Our industry-leading solutions work across diverse customer environments and the world's biggest public clouds.

As a cloud-led, data-centric software company, only NetApp can help build your unique data fabric, simplify and connect your cloud, and securely deliver the right data, services and applications to the right people—anytime, anywhere. www.netapp.com

1 https://www.globenewswire.com/news-release/2019/12/09/1957615/0/en/Genomics-Market-to-Exhibit-a-Stellar-CAGR-of-18-7-Demand-for-Genomic-Products-to-Augment-Owing-to-Increasing-Number-of-Product-Launches-Fortune-Business-Insights.html

2 https://hub.jhu.edu/2020/03/30/covid-19-gene-sequencing/

3 https://www.weforum.org/agenda/2019/06/today-you-can-have-your-genome-sequenced-at-the-supermarket/

4 https://www.wired.com/story/whole-genome-sequencing-cost-200-dollars/

5 https://www.technologyreview.com/s/615289/china-bgi-100-dollar-genome/

6 https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-019-0689-8

7 https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-019-0689-8

NetApp