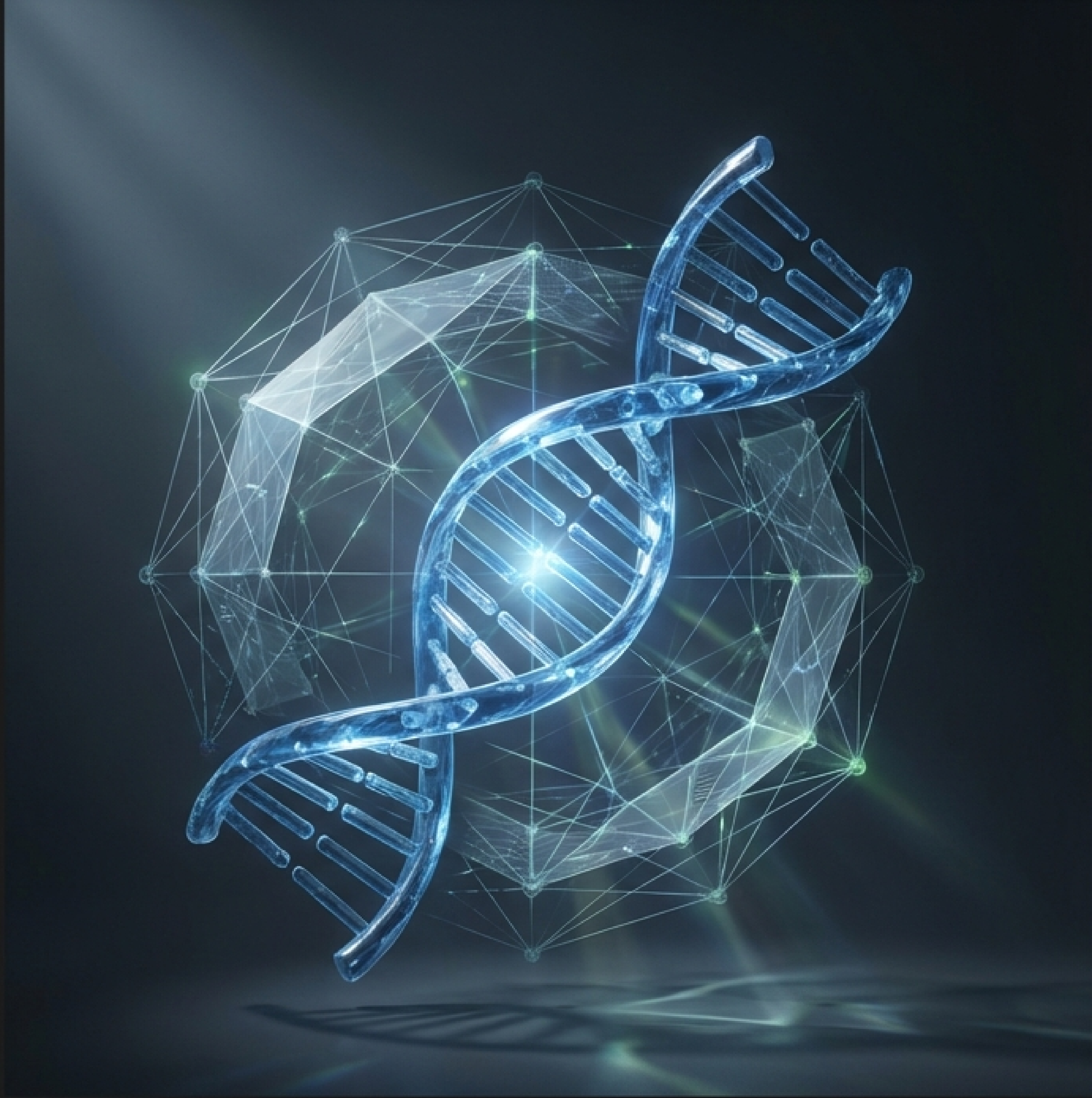
Velocity & Efficiency: The New Economics of Sequencing

How Deep Learning is Redefining the Cost-Benefit of Genomics



From Brute Force to Intelligent Analysis

Deep learning-based variant calling delivers a step-change in performance by fundamentally altering the trade-offs between cost, speed, and accuracy. This new paradigm enables:

1. Unprecedented Efficiency

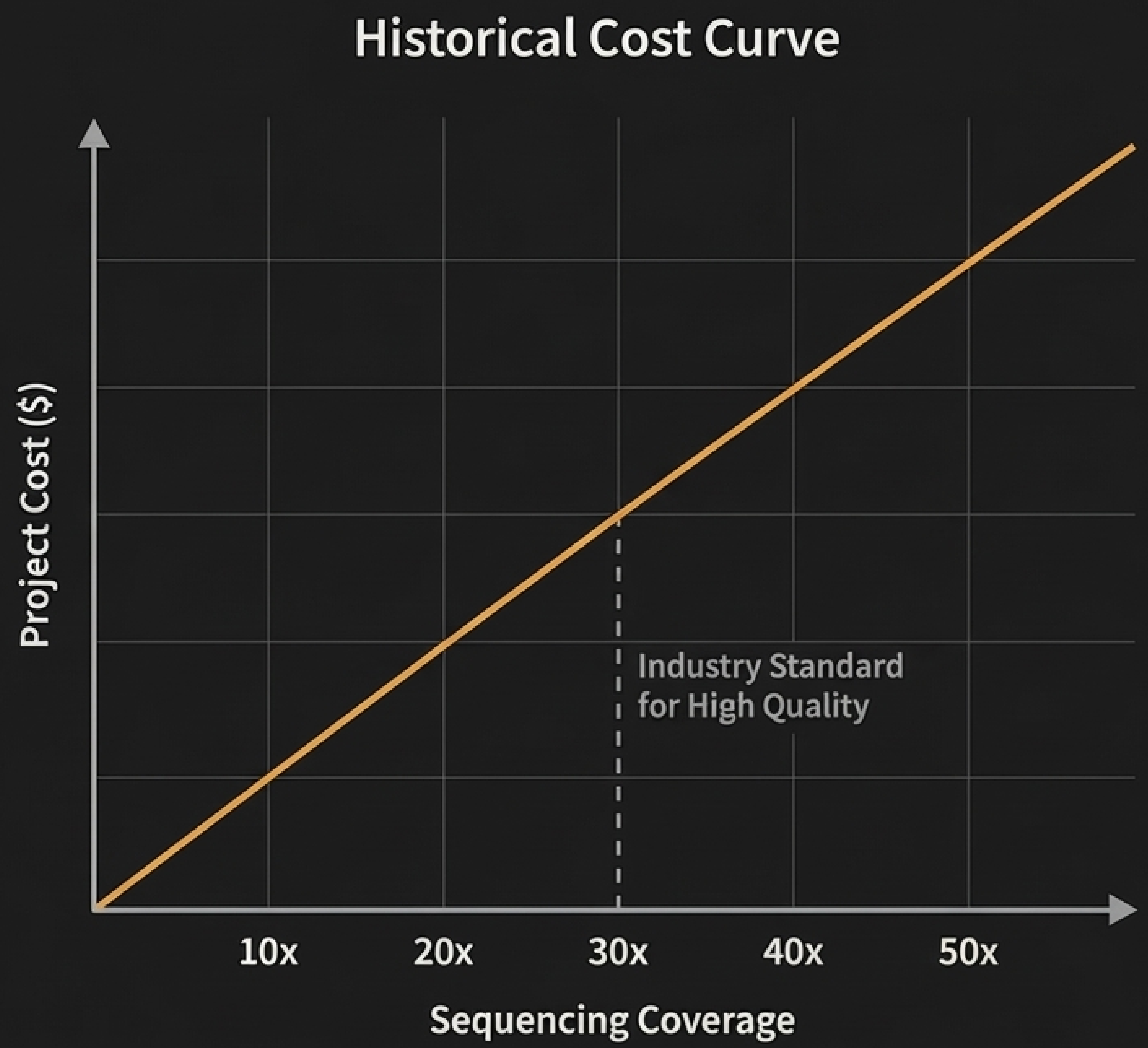
Achieve higher accuracy with significantly lower sequencing coverage, directly reducing instrument and reagent costs.

2. Clinical Velocity

Shrink diagnostic timelines from days to hours with ultra-rapid, massively parallelized analysis pipelines, making genomics viable in acute care.

3. Future-Proof Adaptability

A single, universal, and trainable framework that achieves state-of-the-art accuracy across diverse and emerging sequencing technologies.



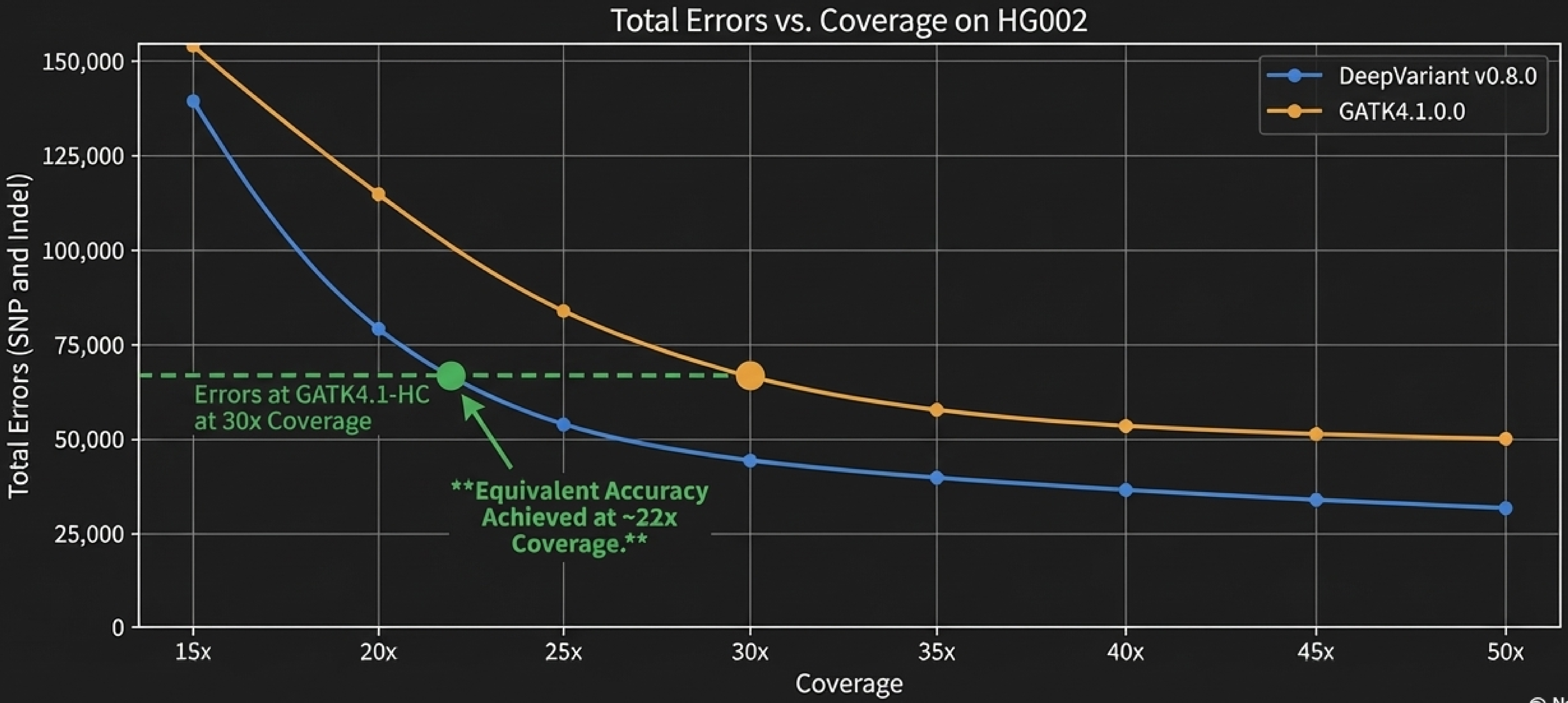
The Old Dogma: The Coverage-Cost Trade-Off

Historically, high-quality variant calling demanded high sequencing coverage, with ~30x coverage becoming the accepted standard for a ‘high quality’ genome. This created a direct, linear, and expensive trade-off between project cost and data quality.

This standard was dictated by the analytical limitations of handcrafted statistical models, which required deep data to overcome systematic errors.

Redefining the Standard: 22x is the New 30x

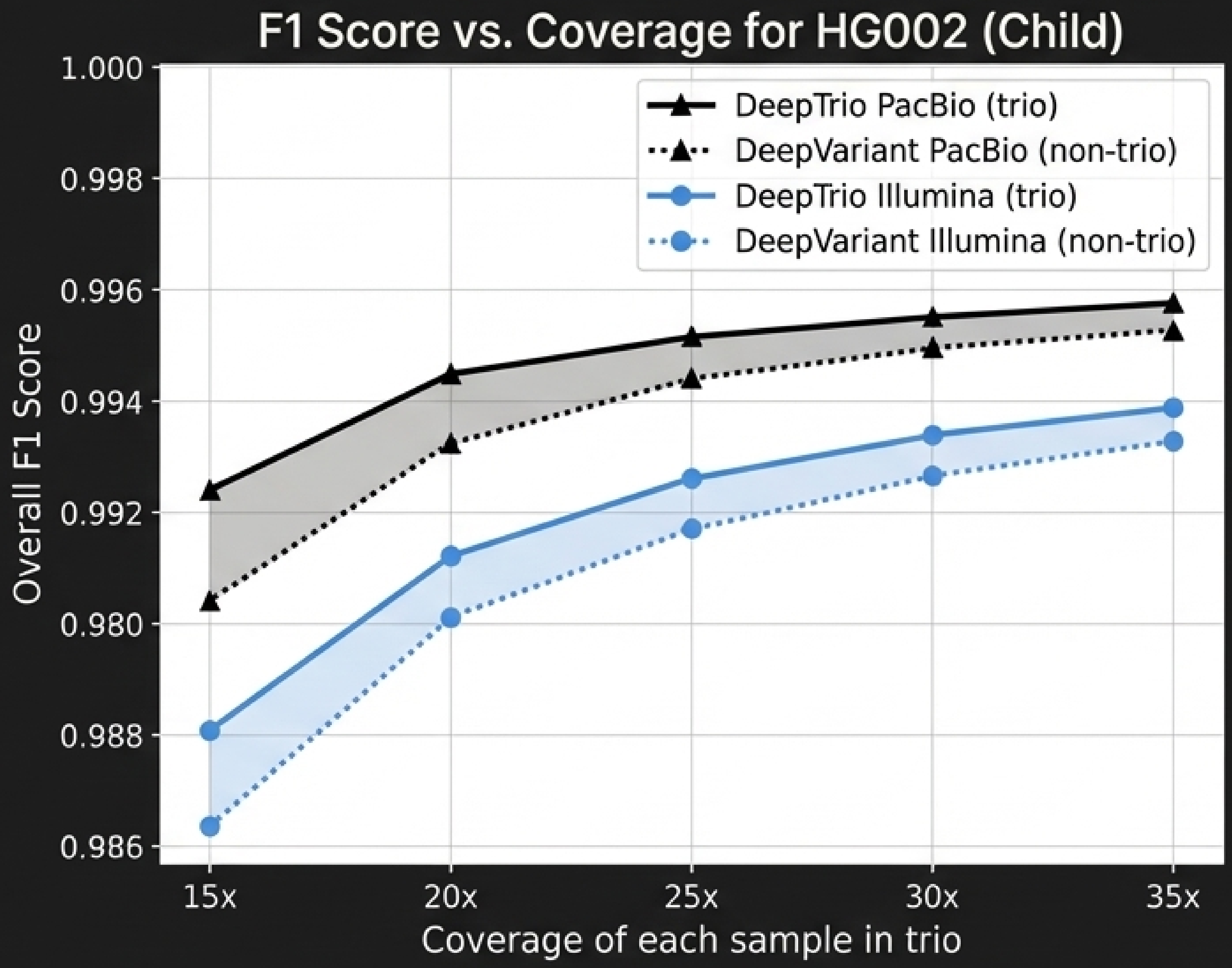
DeepVariant's deep learning model achieves the same or better accuracy at ~22x coverage as traditional methods (GATK4 HaplotypeCaller) at 30x coverage. This translates to a >25% reduction in sequencing costs without sacrificing quality.



Trio Analysis: Superior Accuracy at Even Lower Coverage

By jointly analyzing child-mother-father genomes, DeepTrio learns Mendelian inheritance patterns to further boost accuracy and resolve ambiguity. This added context allows the model to achieve superior performance with even less data.

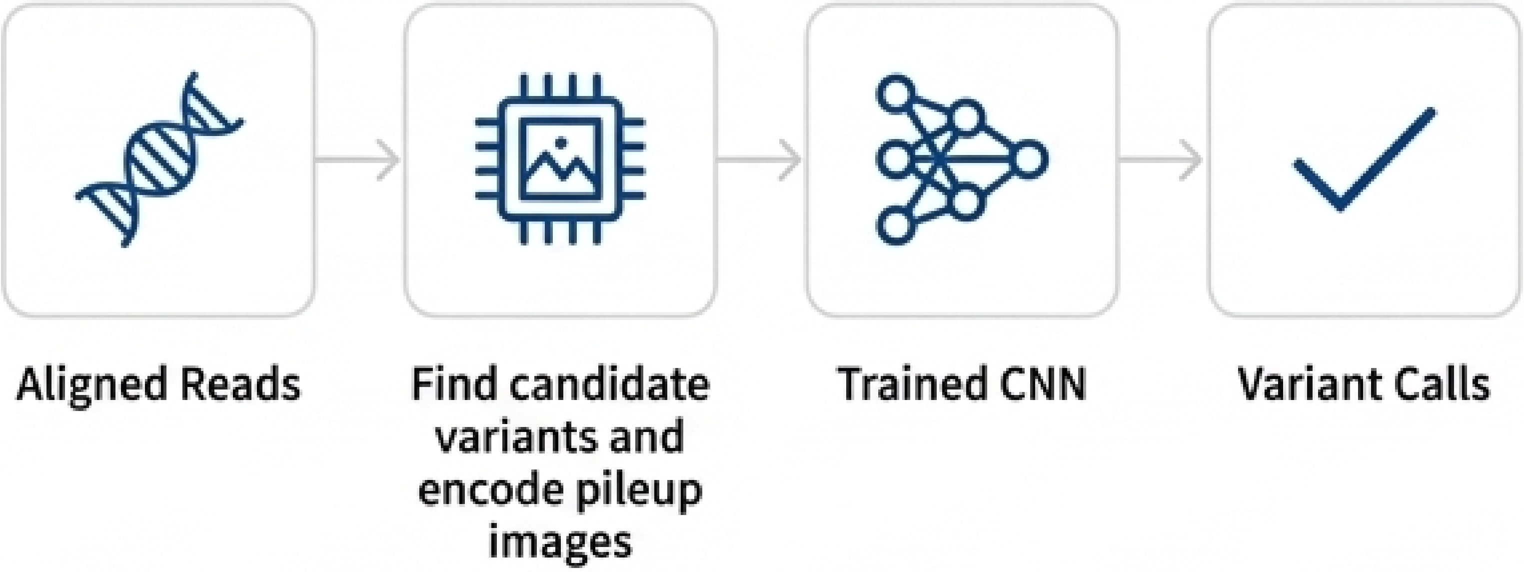
The result is that ~20x coverage with DeepTrio is roughly equivalent to a 30x single-sample analysis with DeepVariant, compounding the cost savings for family-based studies.



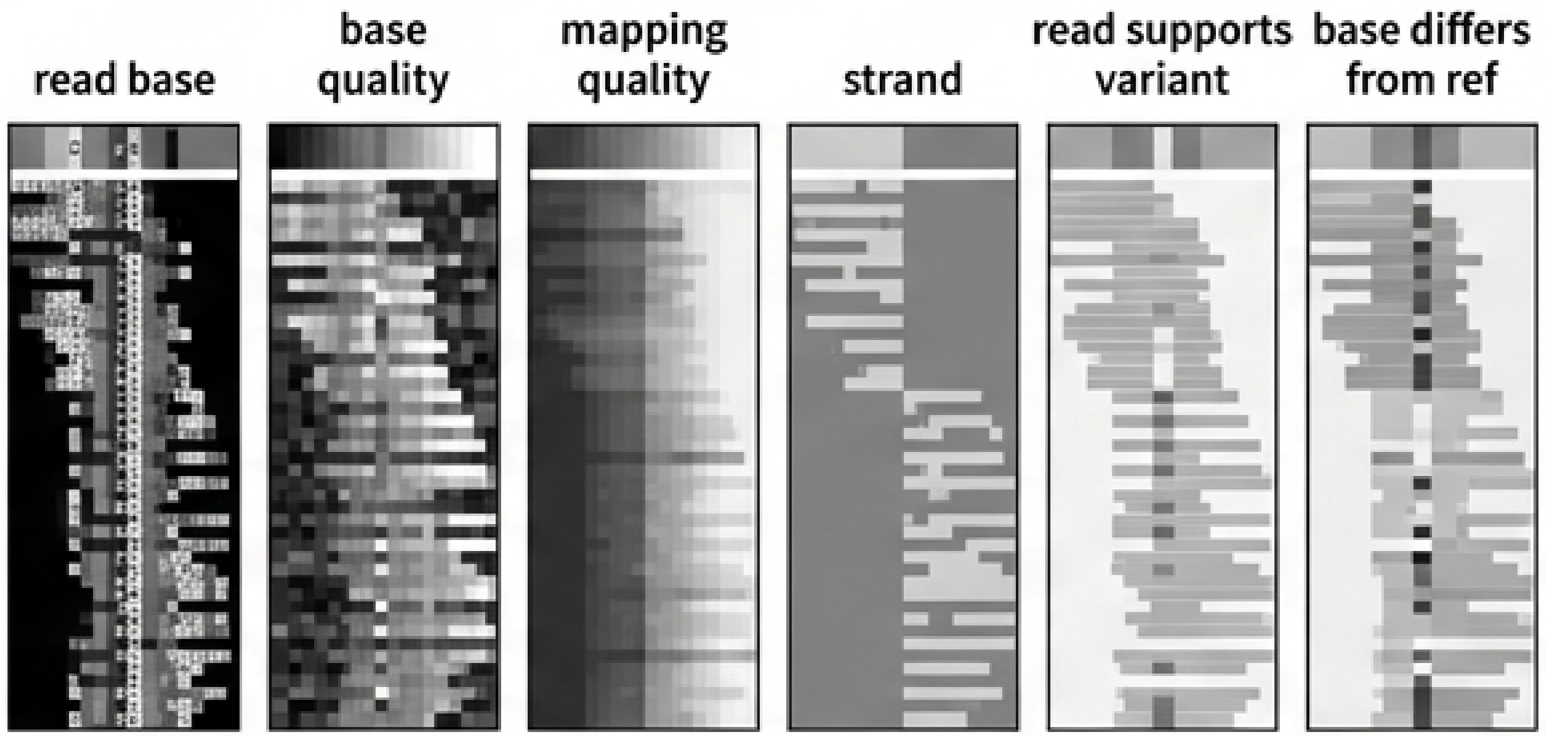
The Engine: Variant Calling as Image Classification

DeepVariant transforms aligned read data around a candidate variant into a multi-channel tensor, or "pileup image". Each channel encodes a different feature of the sequencing data (e.g., read base, base quality, mapping quality). A Convolutional Neural Network (CNN), similar to those used in image recognition, is then trained to distinguish the visual signatures of true variants from sequencing errors.

Simplified Workflow



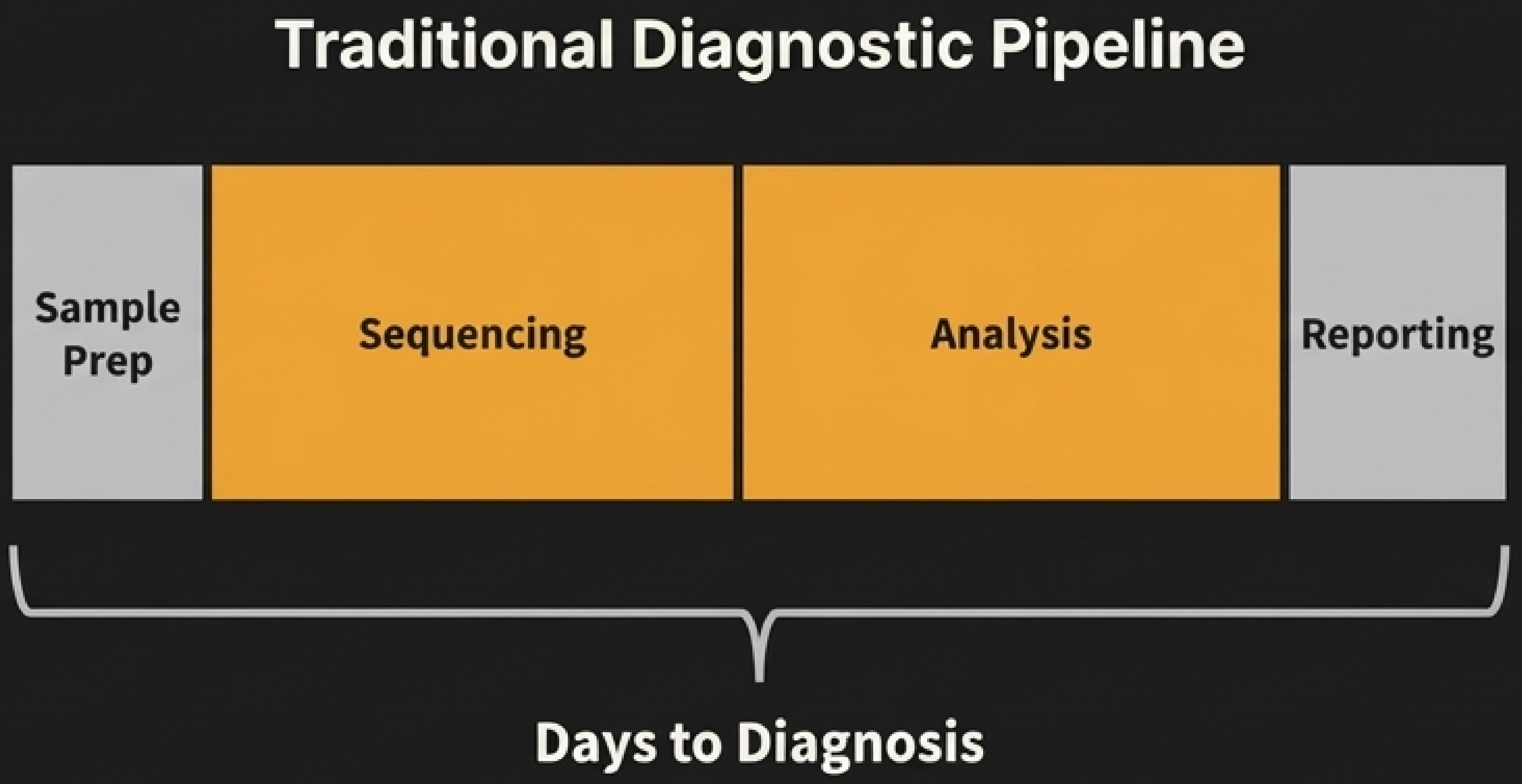
The Pileup Image: What the Model Sees



The Clinical Imperative: When Hours Matter

For critically ill patients, especially newborns in the NICU, rapid genetic diagnosis can be life-saving. Traditional sequencing and analysis pipelines, with turnaround times of days or even weeks, are often too slow to guide acute clinical intervention.

The analysis phase has historically been a significant bottleneck in this process.



World Record Speed: Sample to Diagnosis in 7 Hours, 18 Minutes

In a study led by Stanford University School of Medicine, a massively parallelized pipeline using PEPPER-Margin-DeepVariant on Oxford Nanopore data achieved a new record for sample-to-diagnosis. The rapid identification of a likely disease-causing variant in the TNNT2 gene directly influenced patient care, reducing the need for further invasive procedures like cardiac biopsy. This represents a ~2x speed improvement over the previous record.

0:00 — **Sample preparation started with DNA extraction.** 2 ml of blood sample collected.

0:50 — **DNA shearing started.** 28 µg of DNA extracted.

1:02 — **Library preparation started.** 28 µg of input DNA for library.

2:32 — **Sequencing, base calling, alignment started.** 50% of library recovered and 333 ng of mass loaded per flow cell.

4:48 — **Sequencing completed.** 201.5 GB of data generated with a read N50 of 24 kb.

5:05 — **Base calling, alignment completed.** 89% of reads passed (>Q7) with a 60x coverage.

6:55 — **Variant calling and annotation completed.** 4,316,464 small variants and 35,780 SVs identified for filtration and prioritization.

7:05 — **Manual review started.** 22 small variants and 16 SVs prioritized for manual review.

7:05 — **Variant calling and annotation completed.** 4,316,464 small variants and 35,780 SVs identified for filtration and prioritization.

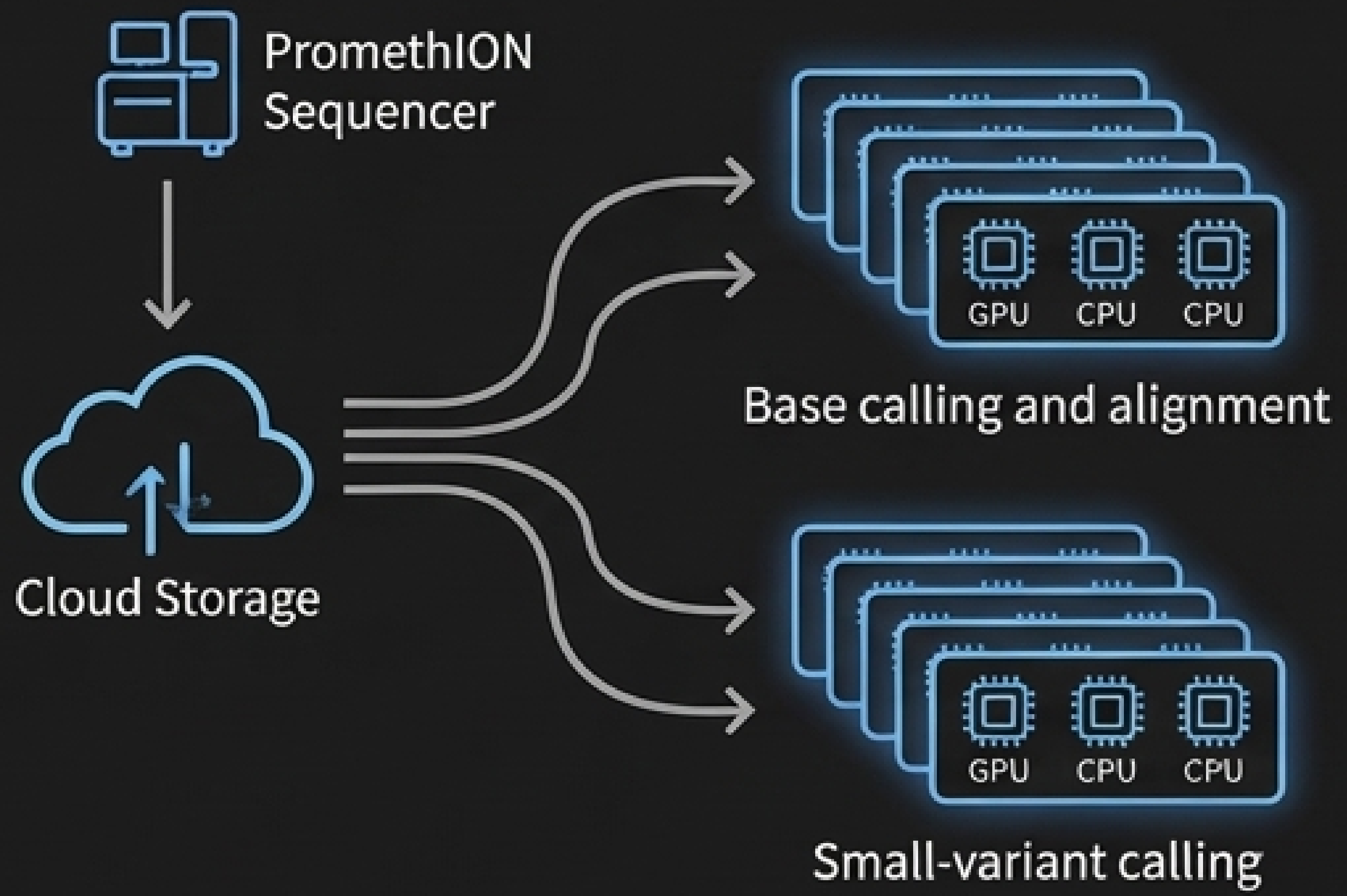
7:18 — **Candidate variant identified.** Variant identified in gene TNNT2.

Engineered for Velocity: Parallelism and Hardware Acceleration

The record-breaking speed is enabled by a software pipeline designed for massive cloud parallelism and hardware acceleration.

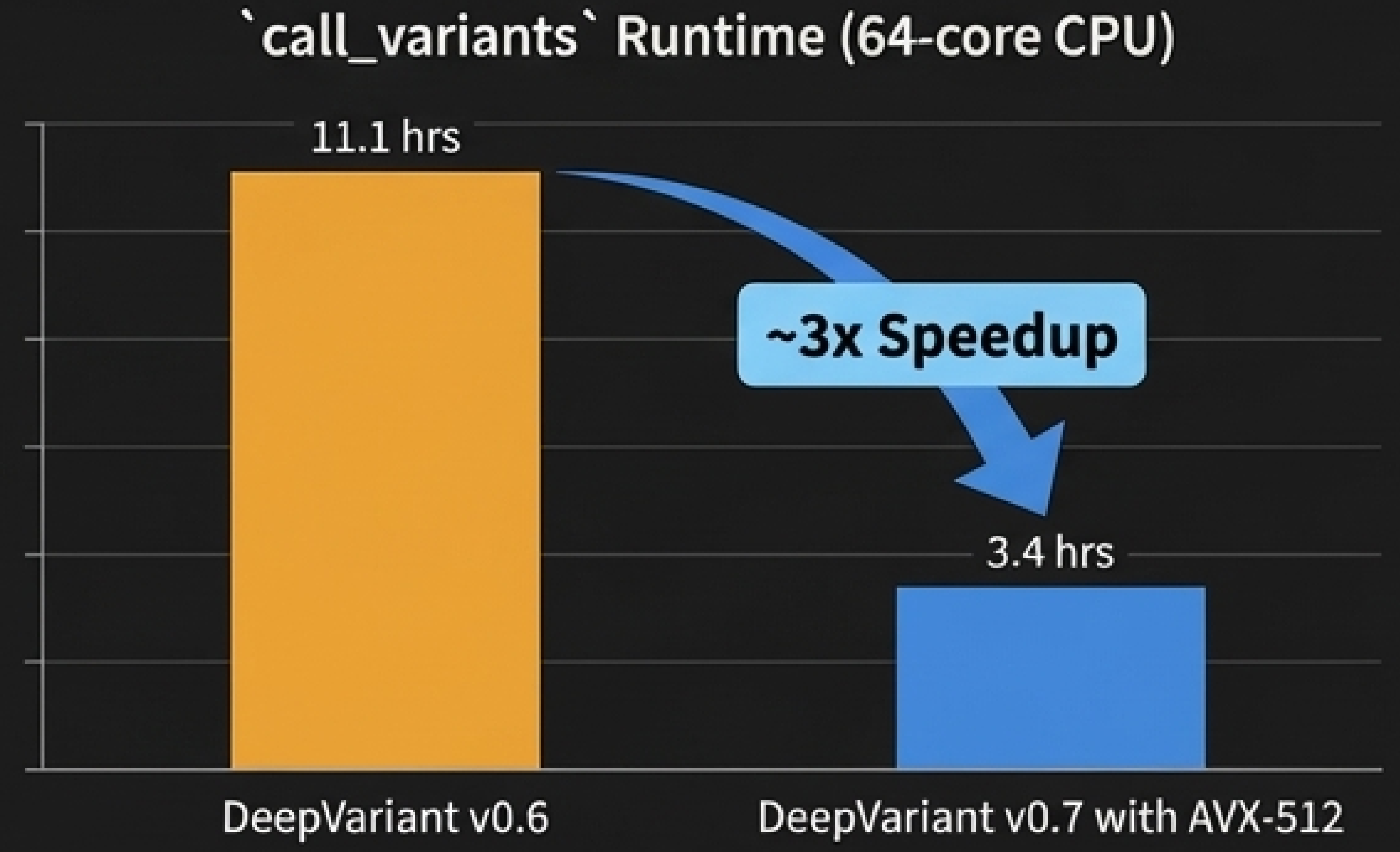
Cloud Parallelism

The workflow distributes compute-intensive base calling, alignment, and variant calling across dozens of GPU and CPU instances, running concurrently with sequencing.



Hardware Acceleration

DeepVariant leverages modern CPU instruction sets like Intel AVX-512, which alone provides a ~3x speedup in the core `call_variants` step compared to previous versions.



A Universal Caller for a Multi-Platform World

Unlike traditional statistical models that are finely tuned for a specific technology's error profile, DeepVariant's learning-based framework is technology-agnostic. By re-training on new, representative data, it achieves state-of-the-art accuracy across all major sequencing platforms.

illumina

PacBio

OXFORD NANOPORE

Beyond Variant Calling: Improving Raw Read Accuracy

The same transformer-based architectures used in variant calling can be applied upstream to "polish" raw data from the sequencer. DeepConsensus, applied to PacBio HiFi data, learns the systematic error profiles of the instrument to produce more accurate reads.

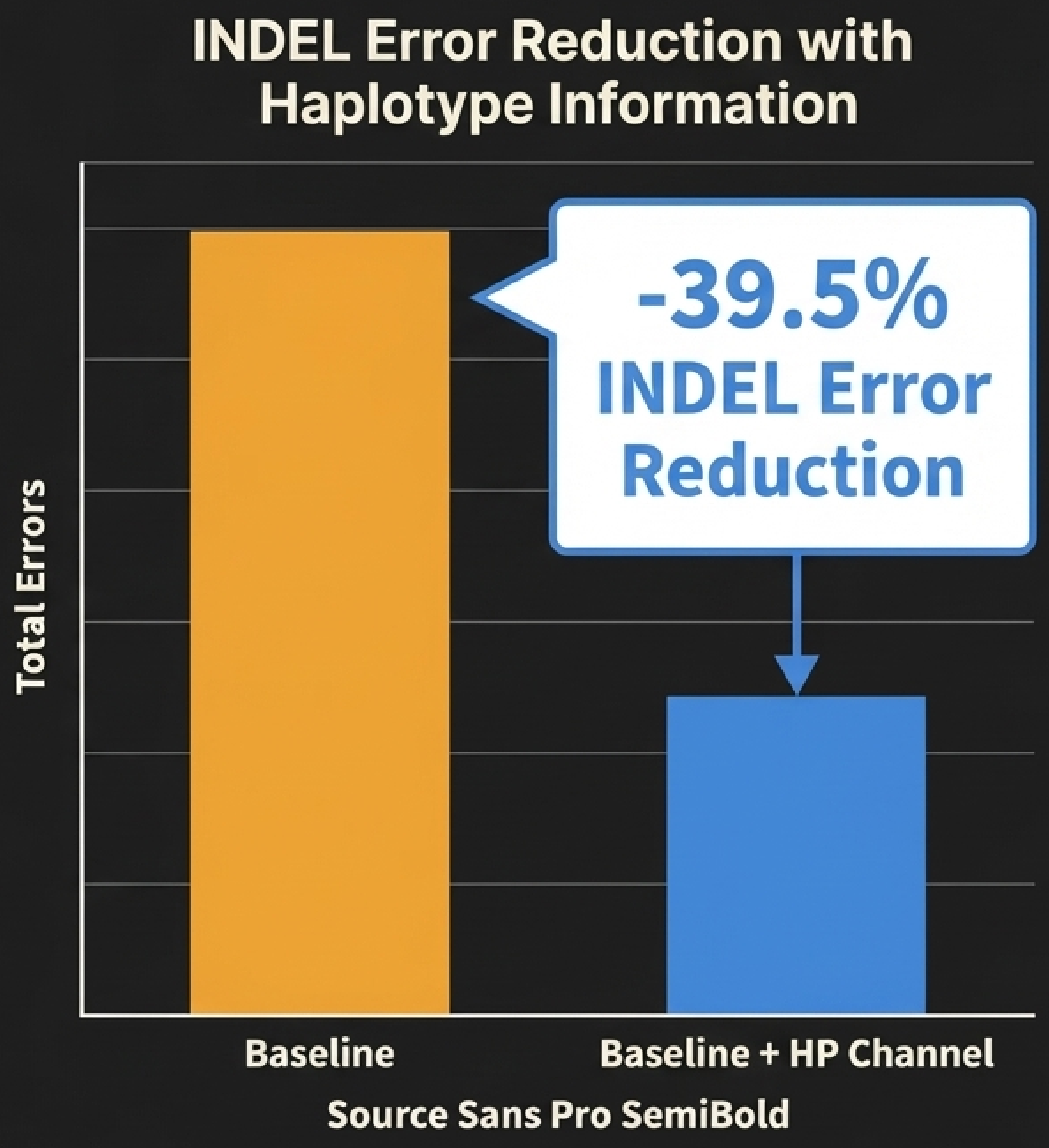
DeepConsensus reduces overall read errors by 41.9%, with the most significant reduction in non-homopolymer insertions (70.9%). More accurate reads lead to better assemblies and more accurate downstream variant calls.

Error Rate per Read (bp/read)			
Metric	**pbccs**	**DeepConsensus**	**% Decrease**
Mismatch	1.82	0.99	45.6%
Homopolymer Deletion	9.41	6.74	28.4%
Homopolymer Insertion	10.28	5.01	51.3%
Non-Homopolymer Deletion	1.05	0.97	7.6%
Non-Homopolymer Insertion	2.06	0.60	70.9%
All Errors	**24.61**	**14.31**	**41.9%**

A System That Continuously Improves

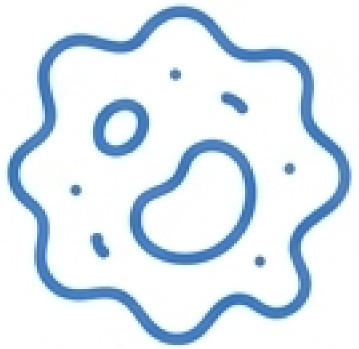
DeepVariant's performance is not fixed. Its accuracy continually evolves by incorporating new training data or adding new information channels. This allows the model to learn from new biological contexts, sequencing preparations, or population data.

"Adding haplotype information for long-read data provides crucial context for resolving ambiguous sites, particularly complex indels. This single change reduced INDEL errors by nearly 40%."



From Germline to New Frontiers

The core deep learning framework is being extended from inherited germline variants to new, challenging domains. This demonstrates the versatility of the approach to tackle a wider range of biological questions and clinical needs. Key areas of expansion include:

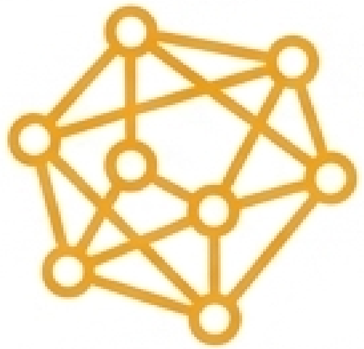


Somatic Variants

Adapting the model to detect low-frequency mutations in cancer tissue (DeepSomatic).

RNA-Seq Variants

Calling germline variants directly from RNA-sequencing data, adding value to transcriptomic experiments.



Pangenomes

Improving mapping and variant calling accuracy by using graph-based pangenome references that better represent human diversity.

The New Economics of Sequencing

- **Efficiency:** Sequence more samples for the same budget, or save on costs by achieving target accuracy at lower coverage (~22x vs. 30x).
- **Velocity:** Enable new clinical applications where speed is paramount, moving genomics from a research tool into acute care with sub-8-hour diagnoses.
- **Versatility:** De-risk technology choices and future-proof analysis pipelines with a single, adaptable framework that excels across Illumina, PacBio, and Nanopore platforms.

NotebookLM

Explore the Technology

DeepVariant and its ecosystem are open-source and actively developed. Find code, documentation, pre-trained models, and key publications at the following resources.

DeepVariant on GitHub
github.com/google/deepvariant

Google Health AI Blog
ai.googleblog.com/search/label/health

Key Publications:

- Poplin, R. et al. *A universal SNP and small-indel variant caller using deep neural networks*. Nat Biotechnol (2018).
- Goenka, S.D. et al. *Accelerated identification of disease-causing variants with ultra-rapid nanopore genome sequencing*. Nat Biotechnol (2022).
- Kolesnikov, A. et al. *DeepTrio: Variant Calling in Families Using Deep Learning*. bioRxiv (2021).