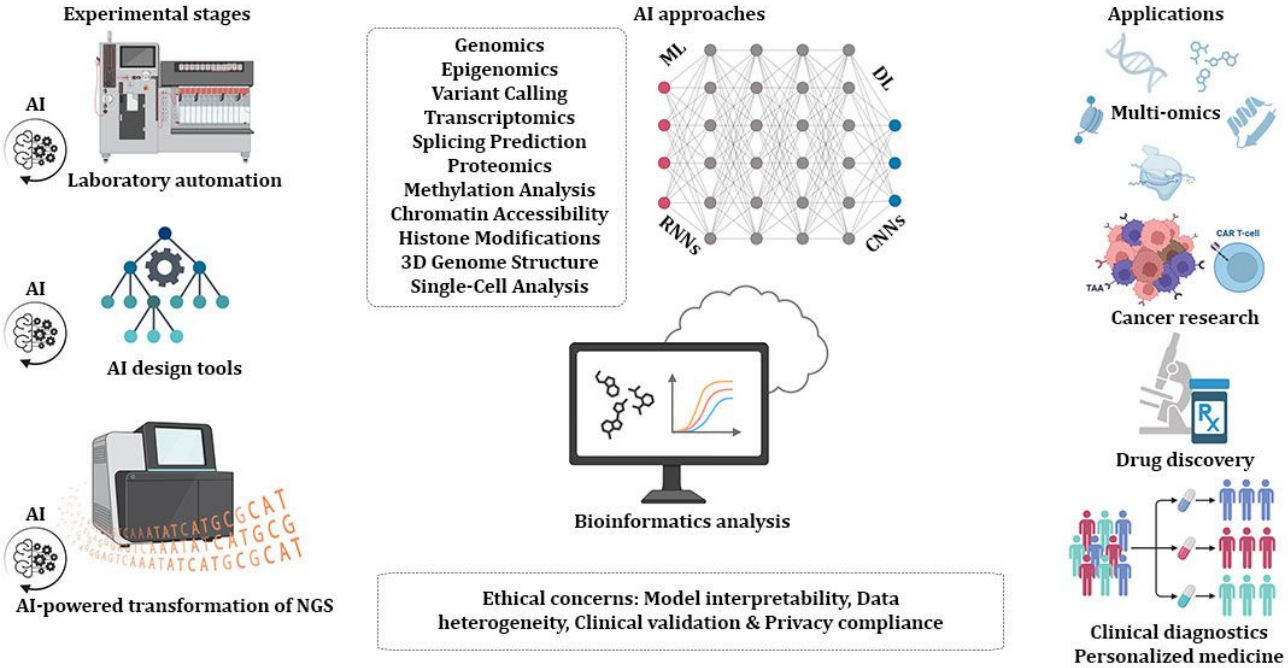
Introduction to AI in Genomics

A Comprehensive Overview

26.11.2025

The AI Revolution in Genomics: Unlocking the Code of Life

AI is quickly becoming an indispensable tool for turning vast genomic data into **actionable, life-changing medicine**



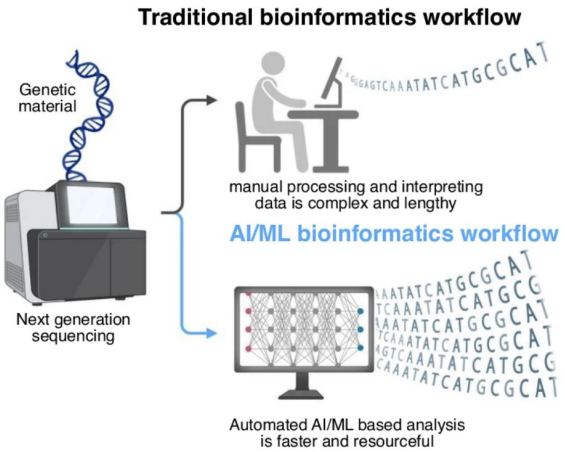
Athanasopoulou, K., Michalopoulou, V.-I., Scorilas, A., & Adamopoulos, P. G. (2025). Integrating Artificial Intelligence in Next-Generation Sequencing: Advances, Challenges, and Future Directions. *Current Issues in Molecular Biology*, *47*(6), 470. https://doi.org/10.3390/cimb47060470

Introduction to Genomics & The Data Deluge

Data Explosion
High-Throughput Sequencing (HTS) has created an unprecedented volume, velocity, and complexity of genomic data.

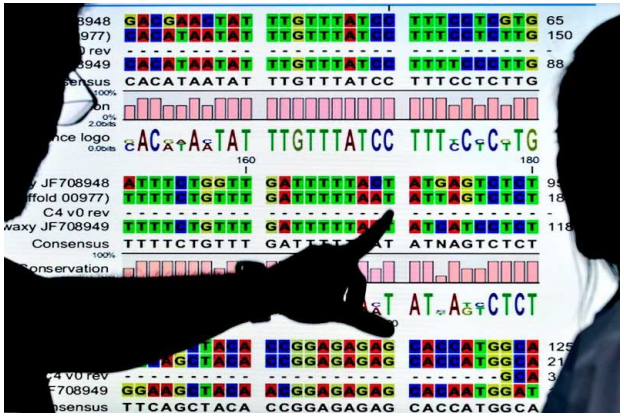
Bottleneck
Conventional computational methods are overwhelmed. They struggle to find the complex, non-linear patterns hidden within terabytes of raw data.

AI Solution
AI, particularly Deep Learning, is adapted from fields like computer vision to achieve "superhuman" pattern recognition in the genome.



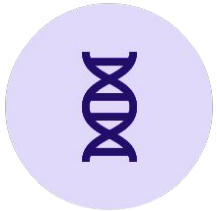
'how do we ***get*** the data?'

'how do we make ***sense*** of this data?'



O'Connor, O., McVeigh, T.P. Increasing use of artificial intelligence in genomic medicine for cancer care- the promise and potential pitfalls. *BJC Rep* 3, 20 (2025). https://doi.org/10.1038/s44276-025-00135-4

Why AI is Indispensable in Genomics



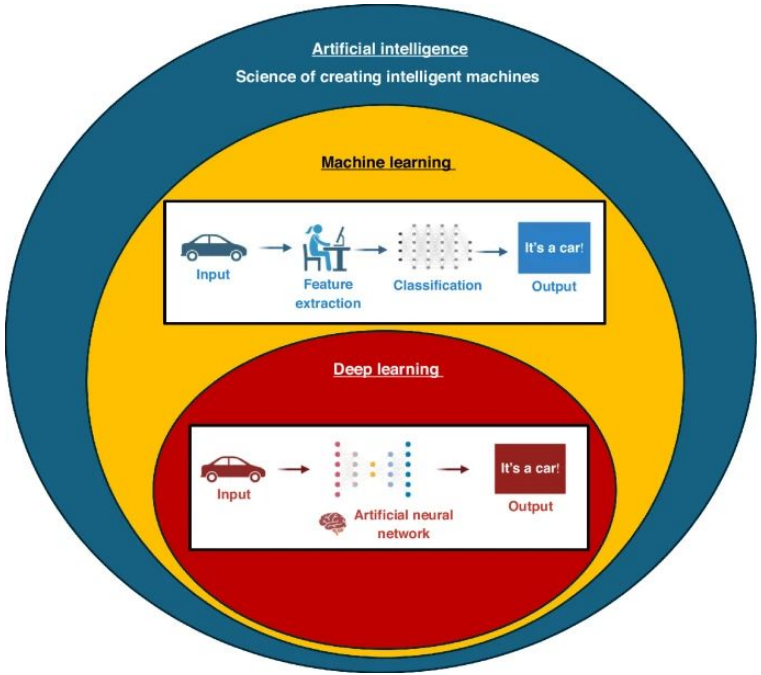
Massive Data Volumes:
Traditional methods cannot keep pace with the exponential growth of sequencing data.

Complex Pattern Recognition:
AI (especially deep learning) excels at detecting subtle, non-linear relationships in data that hint at disease mechanisms.

Acceleration of Discovery:
Automates time-consuming analysis tasks, speeding up research and diagnosis.

Integration of Multi-Omics Data:
AI is the only effective way to combine diverse data types (genomics, proteomics, clinical records) for a holistic view.

AI/ML/DL Hierarchy & Definitions in Genomic Context



O'Connor, O., McVeigh, T.P. Increasing use of artificial intelligence in genomic medicine for cancer care- the promise and potential pitfalls. *BJC Rep* 3, 20 (2025). https://doi.org/10.1038/s44276-025-00135-4

Term	Definition	Genomics Application Example
Artificial Intelligence (AI)	Systems that mimic human intelligence to perform complex tasks.	End-to-end automation of the entire sequencing pipeline (LIMS, QC, Analysis).
Machine Learning (ML)	Algorithms trained on data to identify patterns and make predictions.	Predicting protein degradability using intrinsic feature analysis.
Deep Learning (DL)	A subset of ML utilizing multi-layered neural networks to learn complex feature hierarchies.	Base Calling (Nanopore RNNs) and Variant Calling (DeepVariant CNNs).

AI is embedded across all stages of a genomic workflow

Stage I: Sample collection and preparation (wet lab)

Stage II: Sequencing and primary analysis (base calling, QC)

Stage III: Secondary analysis (alignment, variant calling)

Stage IV: Tertiary analysis (variant annotation, interpretation)

FASTQ

Raw sequencing output. Contains sequence (A, C, G, T) **plus** ASCII quality scores for each base.

BAM / CRAM

Compressed, aligned reads relative to a reference genome. The standard input for variant calling.

VCF

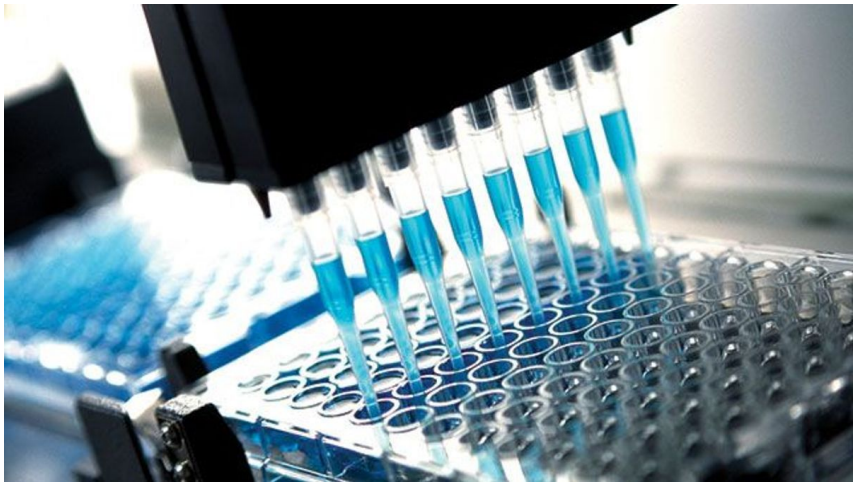
Variant Call Format. Final standardized output listing all variations (SNPs, InDels, SVs).

Stage I: Sample collection and preparation

Lab automation, AI-based QC, automated traceability

Tasks and Challenges

- Sample and metadata collection (e.g. patient details, tracking information)
- Sample disruption, nucleic acid isolation (DNA/RNA), QC, and NGS library prep.
- Anomaly detection and treceability



AI solutions

A. Lab automation: Robotic **liquid handling**

B. AI-based imaging for QC (**Image Recognition**)

A. Lab Automation

Challenges of Manual Sample Handling

- Increases variability and risk of contamination or mix-ups
- Slows processing and limits throughput
- Unsuitable for high-volume workflows such as neonatal screening

Example: automated DBS Processing for NGS by *Revvity Panthera-Puncher*™

➔ Automates Dried Blood Spot preparation up to 9 plates at once
➔ Boosts throughput and reduces prep time
➔ Enhances standardization and sample integrity
Optimizes workflows for downstream NGS in precision medicine

B. AI-based imaging for QC (Image Recognition)

Challenge:

- Accurate DNA quantification and integrity assessment is critical before sequencing. Manual and visual methods are slow and subjective.
- Labs produce large volumes of unstructured documents; manual metadata entry is slow and error-prone.

AI-Assisted Imaging for QC

➔ QC-flagged samples are automatically linked to genomic data in platforms.

Example: Metadata Extraction with ML (Amazon Textract)

➔ Ensures reliable digitization of key metadata (e.g., extraction lot, clinical notes, consent).

➔ Provides the crucial link between wet-lab efficiency (Revvity) and downstream computational analysis (DeepVariant, SpliceAI, NLP).

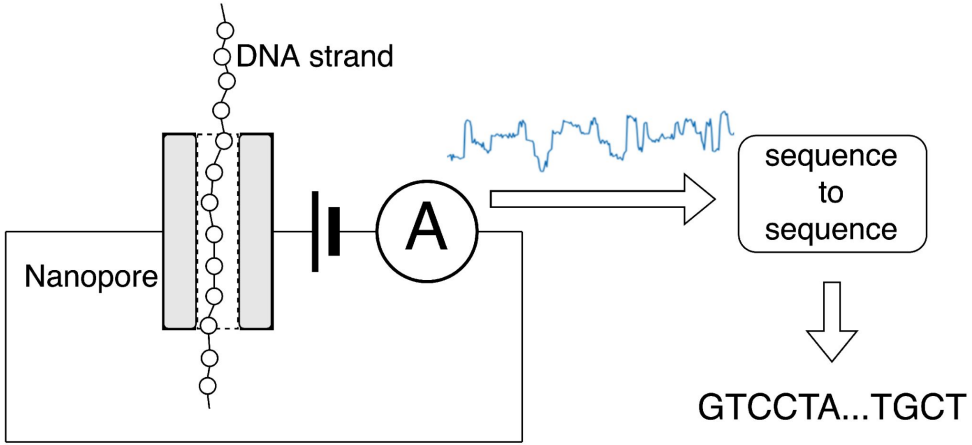
Amazon Textract

Stage II: Sequencing and primary analysis

DL basecalling and sequence correction

Tasks and Challenges

- First step that converts raw sequencing signals into DNA/RNA bases (A, T, G, C)
- Translates physical measurements (e.g., current shifts, fluorescence) into nucleotide reads
- **Key challenge:** accurately converting complex, continuous signals into precise, discrete bases



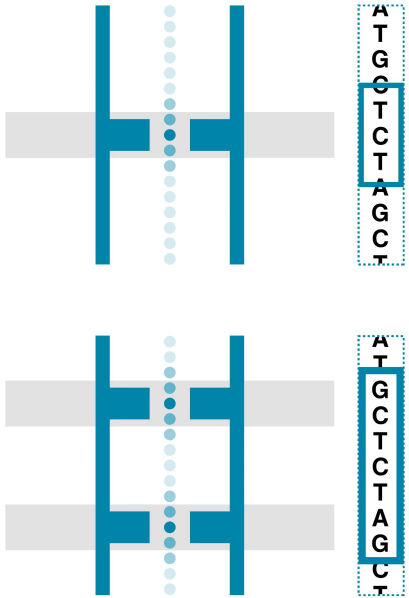
Napieralski, A., & Nowak, R. (2022). Basecalling Using Joint Raw and Event Nanopore Data Sequence-to-Sequence Processing. *Sensors*, *22*(6), 2275. https://doi.org/10.3390/s22062275

AI solutions

➔ Deep Learning models (e.g., **Dorado**) ⟶ treat this as a **blind deconvolution problem**, decoding complex, overlapping signals at high speed.

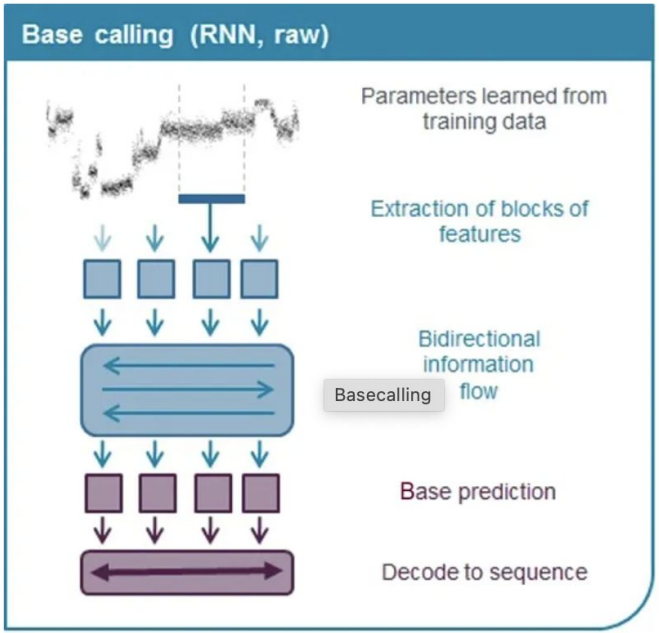
➔ Tools like **HERRO** ⟶ use AI again to "polish" the sequence and correct errors, ensuring high final accuracy.

Base Calling: The Deep Learning Deconvolution Task



Sequencing Stage	Raw Input Data (The Deconvolution Target)	AI Technique	Output (Deconvolution Goal)
Nanopore Sequencing	Electrical current signals (complex time series, "squiggle")	Recurrent Neural Networks (RNNs) and Transformers: RNNs / Transformers (Dorado) + Herro post-processing	Canonical DNA/RNA base sequence (ATCG) and chemical base modifications
General Primary Analysis	Complex analog signals (spectra, images, voltage potential)	Deep Learning Models (various architectures)	Quality-scored base calls (A, T, C, G)

Dorado



- **Speed vs. accuracy**: the fundamental trade-off
- Computational cost and **bottlenecks**

Fast → highest throughput, lower accuracy
High Accuracy → balanced performance
Super Accuracy → maximal accuracy, highest computational cost

*Base calling is demanding, creating a fundamental choice: **Speed** (fast model) or **Accuracy** (sup model). This single step can consume **up to 43%** of the total genome sequencing time.*

HERRO (Post-Basecalling Corrector)

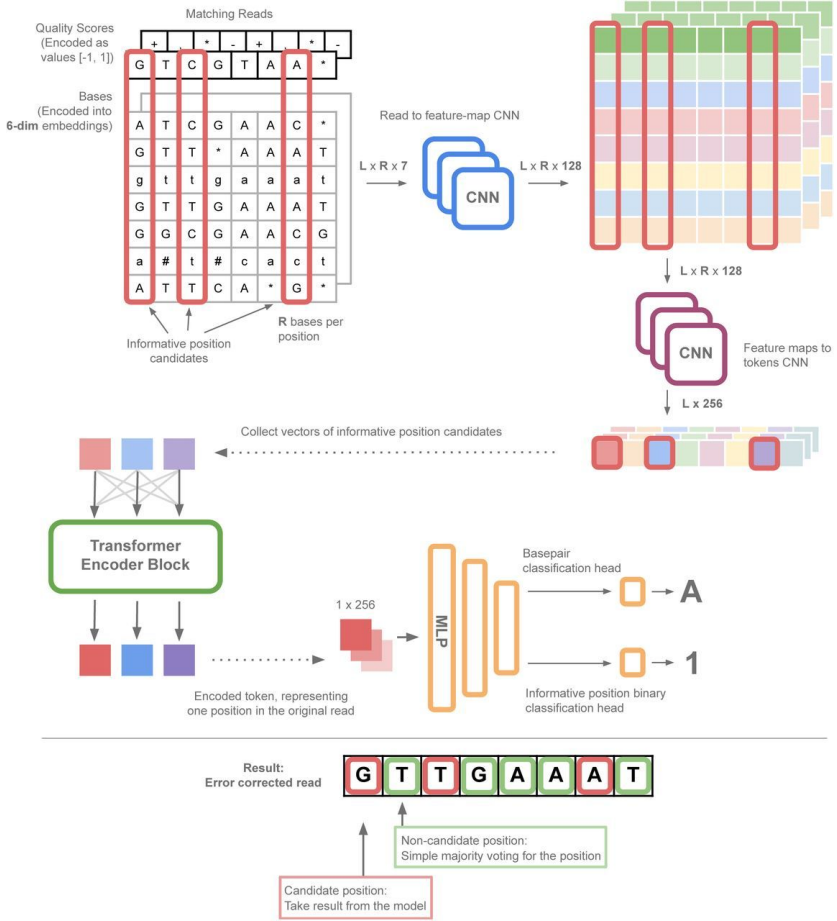
- Complementary tool focused on post-processing, including **signal refinement and chemical modification detection**.
- Enhances base call quality and downstream analysis reliability.

Why It's Needed

This layered design **decouples speed from final accuracy**:

- **Dorado** → fast, approximate basecalling
- **HERRO** → slow, high-precision correction

Together, they provide both the throughput required for real-time workflows and the rigor needed for downstream variant calling, assembly, and clinical interpretation.

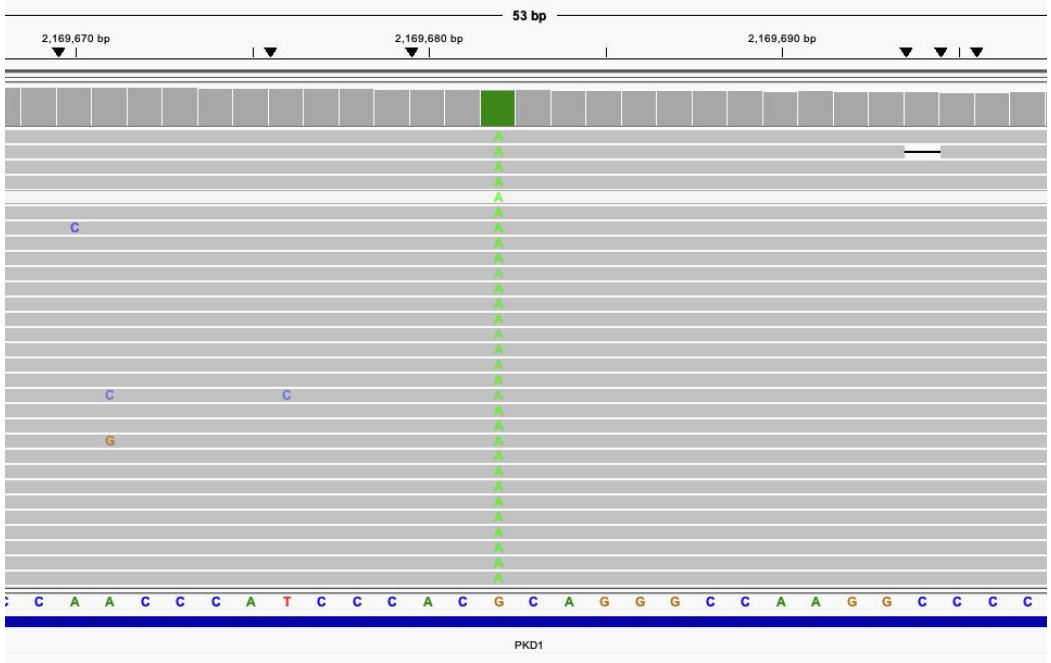


Stage III: Secondary analysis

Variant calling as an image classification problem

Tasks and Challenges

- Alignment of raw sequencing reads to a reference genome
- Identification of genomic variants (e.g., Small Nuclear Polymorphisms - SNPs, insertions or deletions - Indels)
- **Key challenge:** distinguishing true biological variation from sequencing or alignment errors



AI solutions

➔ DeepVariant (Google) reframes the variant calling task as an **image classification problem**

➔ DRAGEN (Illumina) uses ML in a **filtering step** to refine the final variant calls and improve sensitivity

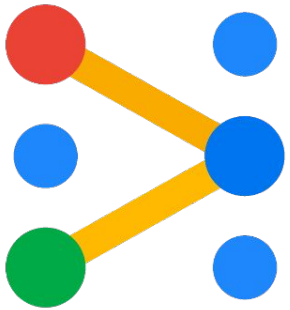
Deepvariant

Traditional statistical models struggle with sequencing errors, complex alignments, and low-coverage areas, leading to false positives and negatives.

DeepVariant treats variant-calling as an **image classification problem**.

It creates **multi-channel tensors** (called "pileup images") for each genomic locus.

A **Convolutional Neural Network (CNN)** analyzes the "image" and classifies the true underlying variation.



github.com/google/deepvariant

Input Reads

Aligned BAM/CRAM file.

Make Examples

Convert reads to pileup images (tensors).

Call Variants

Classify images with CNN to get probabilities.

Output VCF

Post-process probabilities into a final VCF file.

Phase I: Make examples

1. Read Base
(A, C, G, T)

2. Base Quality
(Sequencing confidence)

3. Mapping Quality
(Alignment confidence)

4. Strand of Alignment
(Forward / Reverse)

5. Read Supports Variant
(Does the read agree?)

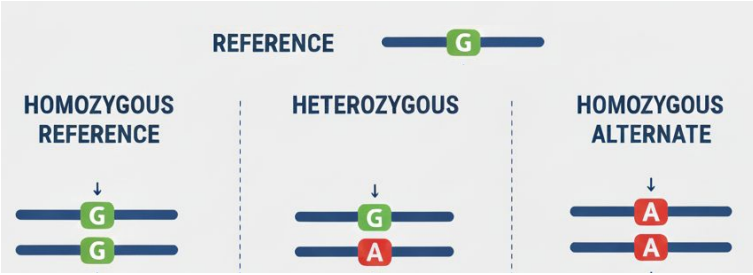
6. Base Differs from Ref
(Is it different?)

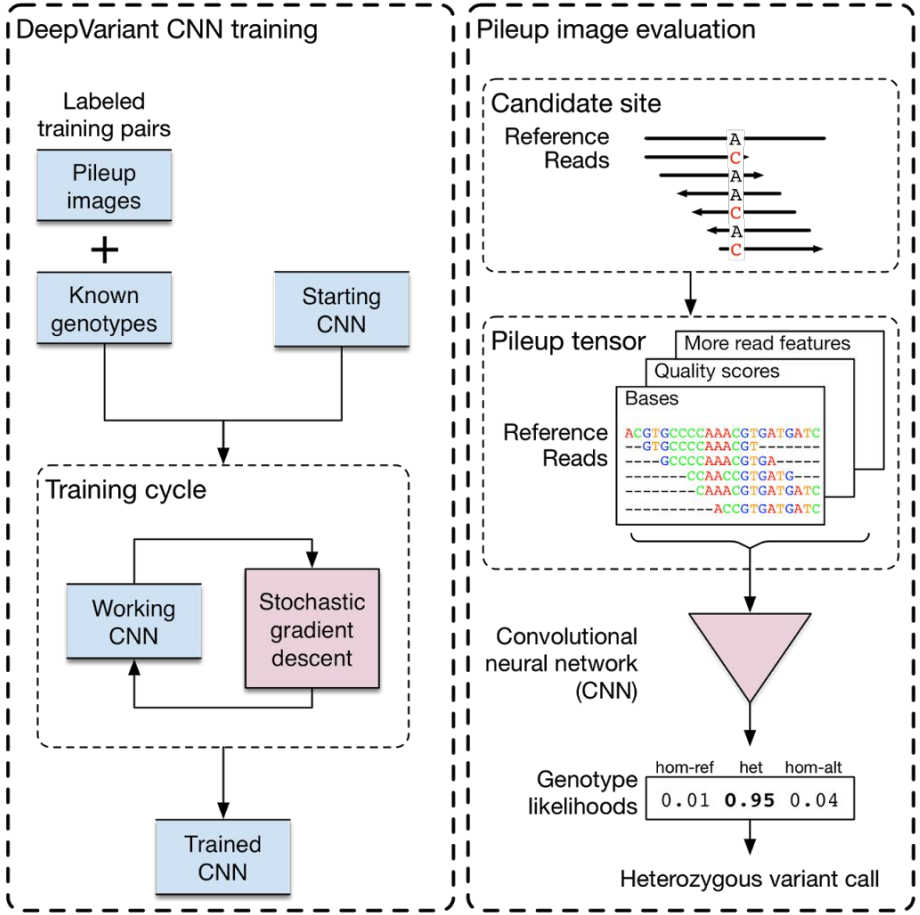
Looking through DeepVariant's eyes - https://tinyurl.com/3mca2bkj

Phase II: Call Variants

The pileup images are fed into a pre-trained **Convolutional Neural Network (CNN)**, trained on millions of high-confidence **"truth" variants** (e.g. NIST GIAB).

The CNN is tasked with classifying the locus into one of the true underlying **genomic variations** (e.g., homozygous alternate, heterozygous, or homozygous reference).





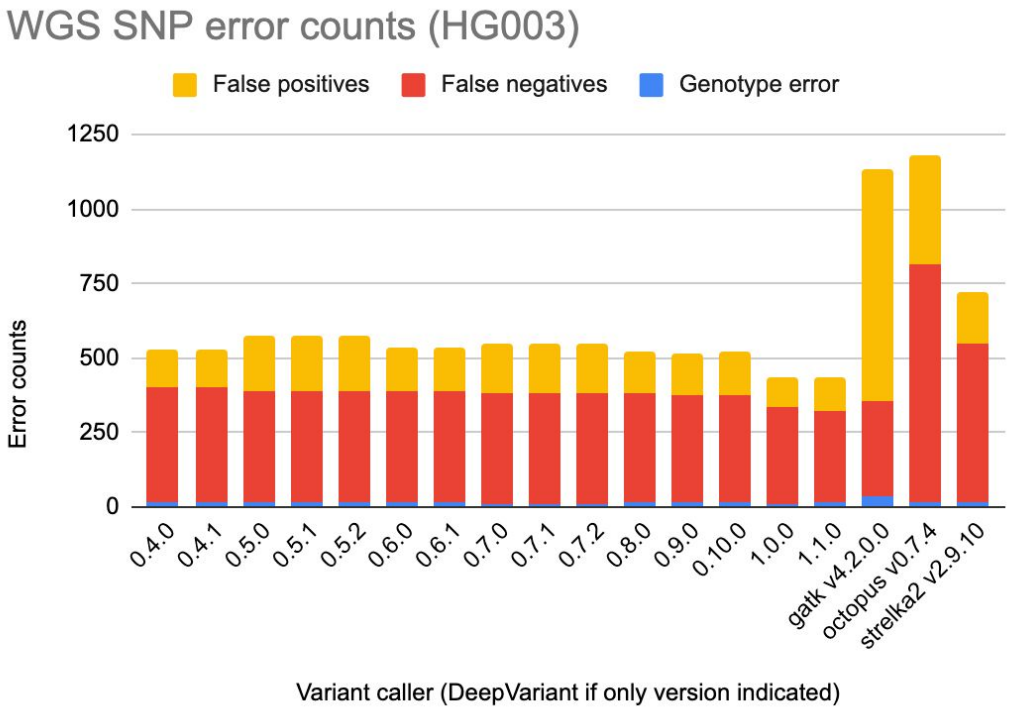
Poplin et al, 2018, doi.org/10.1038/nbt.4235

Phase III: Post-process Variants

The CNN outputs **probabilities** for each genotype. The genotype with the highest probability is "called." This information is consolidated and formatted into a standard **VCF file**.


```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="All filters passed">
#CHROM POS     ID      REF    ALT    QUAL   FILTER INFO    FORMAT        SAMPLE
chr1   3350173       .       T      C      34.5   PASS   AN=2;AC=1     GT:AD:DP:GQ:PL:VAF  0/1:87,62:151:34:34,0,51:0.410596
chr2   189910066     .       ATG    A      6.7    PASS   AN=2;AC=1     GT:AD:DP:GQ:PL:VAF  0/1:84,144:231:6:5,0,16:0.623377
chr14  23854341      .       A      AT     8.3    PASS   AN=2;AC=1     GT:AD:DP:GQ:PL:VAF  0/1:146,13:159:8:7,0,24:0.081761
```

Benefits over other variant callers



https://google.github.io/deepvariant/posts/2021-06-08-deepvariant-over-the-years/

- **Superior error modeling** ⟶ higher accuracy and reduction in **false positives**
- **Ease of re-training** ⟶ can be easily re-trained on platform-specific truth sets
- **Fast pipeline implementation** ⟶ no need for post-processing steps like variant recalibration

Stage IV: Tertiary analysis

AI predictions, data integration and conversational NLP

Tasks and Challenges

- Integrate genomic variants with diverse, **biologically-relevant data sources** (e.g., clinical records, species conservation scores, scientific literature)
- Filter, classify, and prioritize **variants of interest** based on established criteria and the needs of the end-user

AI solutions

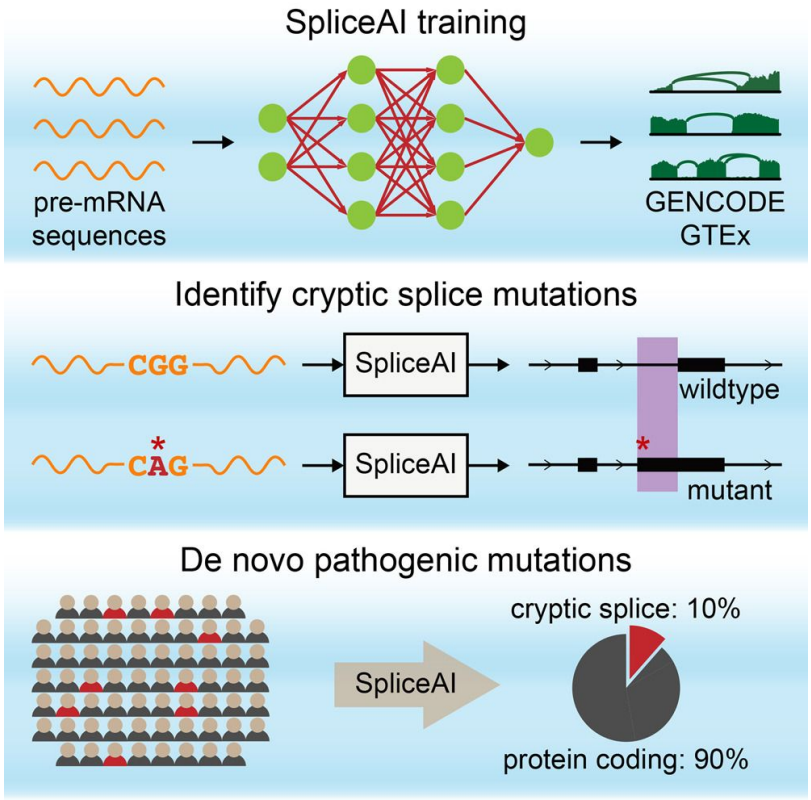
- ➔ **Functional prediction** tools to facilitate variant interpretation (e.g. SpliceAI, AlphaMissense, MetaRNN, etc)
- ➔ Omics platforms to integrate **multi-omics data** (genomic, phenomic, epigenomic, imaging data)
- ➔ **Conversational AI** to improve the accessibility of variant interpretation

Functional prediction

Functional predictors are tools or algorithms that use **biological features** (e.g. sequence conservation, molecular location, and biochemical properties) to predict the **effect or pathogenicity** of a genetic variant.

Example: functional prediction of cryptic splice sites with SpliceAI

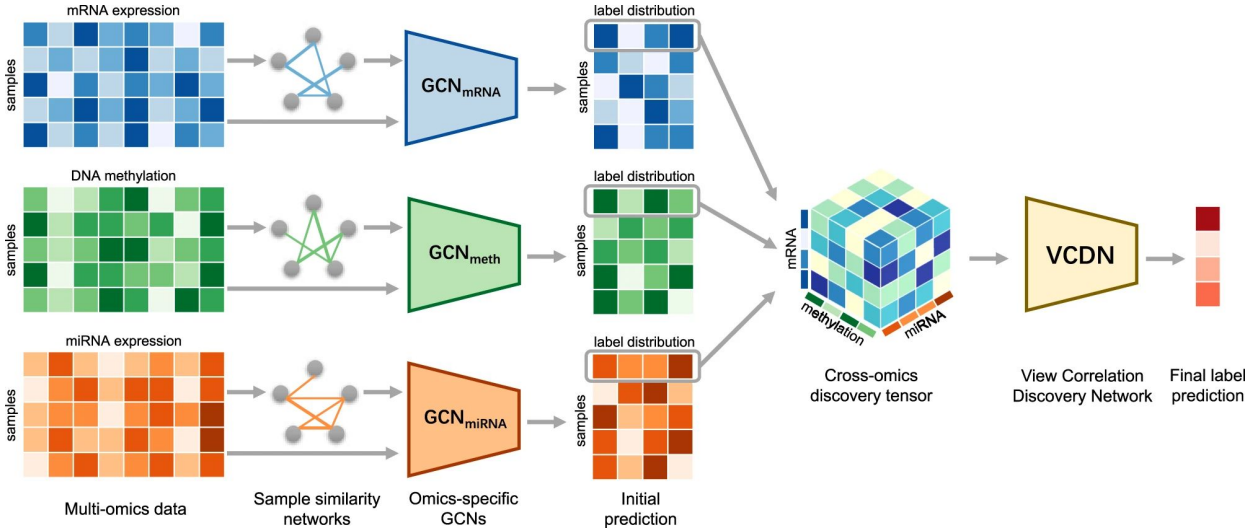
- ➔ Supervised sequence-to-sequence ML framework
- ➔ Capable of targeted or genome-wide splice site predictions
- ➔ Particularly effective for predictions of cryptic splicing, a key mechanism in genetic disorders



Jaganathan, Kishore et al., 2018, doi: 10.1016/j.cell.2018.12.015.

Multi Omics integration

Using ML and high-performance computing, researchers can efficiently **combine diverse datasets** (genomic, clinical, phenomic) to rapidly uncover complex biological insights and create highly accurate predictive models.



Wang et al., 2021, doi.org/10.1038/s41467-021-23774-w

Example: multi-omics integration with MOGONET

- ➔ Multi-Omics data integration using Graph CONvolutional NETworks (GCNs)
- ➔ Dedicated GCN for each omics type to capture features and correlations, integrated into a final prediction with a separate View Correlation Discovery Network (VCDN)
- ➔ Can be used to identify biomarkers for a wide range of biomedical problems

Conversational AI

Large Language Models (LLMs) act as a "**Great Equalizer**" in genomics by enabling Conversational Interfaces.

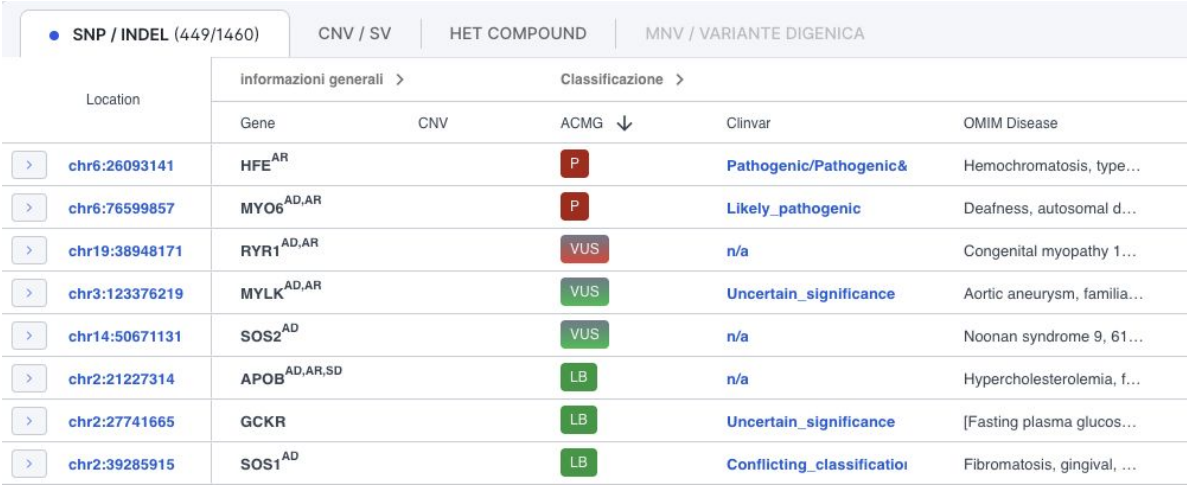
Using simple **natural language processing** (**NLP**), LLMs allow biologists and clinicians without coding expertise to directly query complex datasets with **high-level biological questions**.

Example: a semantic model is trained on genomic concepts (genes, pathways, functional annotations)

"Which variants are high confidence and likely pathogenic in the TP53 pathway?"
"How many variants in this sample are in genes associated with cardiovascular diseases?"
"Which variants are homozygous alternate and in genes connected with recessive mendelian disorders?"

Advantages of using AI in a genomic workflow

- Improved **accuracy** and **reproducibility**
- **Faster** workflows and diagnostic **turnaround times**
- New **insights** through multi-omic analysis
- Conversational AI democratizes **data exploration**



SNP / INDEL (449/1460) | CNV / SV | HET COMPOUND | MNV / VARIANTE DIGENICA

Location	informazioni generali >		Classificazione >		
	Gene	CNV	ACMG ↓	Clinvar	OMIM Disease
chr6:26093141	HFE^{AR}		P	Pathogenic/Pathogenic&	Hemochromatosis, type…
chr6:76599857	MYO6^{AD,AR}		P	Likely_pathogenic	Deafness, autosomal d…
chr19:38948171	RYR1^{AD,AR}		VUS	n/a	Congenital myopathy 1…
chr3:123376219	MYLK^{AD,AR}		VUS	Uncertain_significance	Aortic aneurysm, familia…
chr14:50671131	SOS2^{AD}		VUS	n/a	Noonan syndrome 9, 61…
chr2:21227314	APOB^{AD,AR,SD}		LB	n/a	Hypercholesterolemia, f…
chr2:27741665	GCKR		LB	Uncertain_significance	[Fasting plasma glucos…
chr2:39285915	SOS1^{AD}		LB	Conflicting_classificatio	Fibromatosis, gingival, …

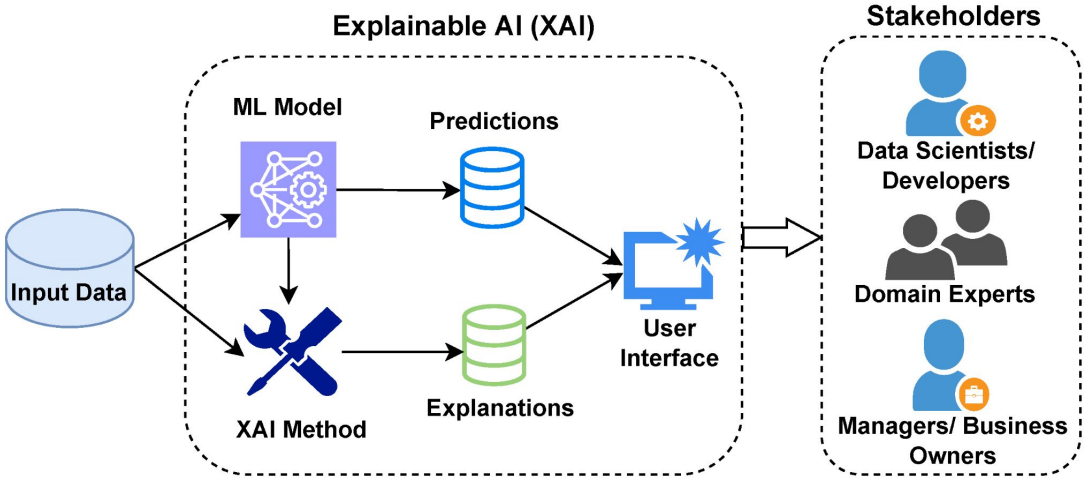
JuliaOmix integrates several AI-based tools into its genomic workflows, for variant calling, annotation and classification

Ethical concerns and future directions

Representation bias ⟶ models can learn implicit biases when training datasets lack diversity

Data privacy ⟶ sensitive health data must be safeguarded against unwanted disclosure

Black-box effect ⟶ DL models are not transparent by nature because of their increased complexity



Clement et al, 2023, doi.org/10.3390/make5010006

Explainable AI (XAI)

- *what data was used to train this model and why?*
- *what factors contributed to this prediction?*
- *how confident is the model in this prediction?*

Final summary

Stage I: Sample collection and preparation → optimized automation and execution efficiencies in the wet lab

Stage II: Sequencing and primary analysis → high-speed signal deconvolution for base calling and sequence correction

Stage III: Secondary analysis → converting bioinformatic data into an image classification problem for high-accuracy variant calling

Stage IV: Tertiary analysis → democratizing complex interpretation and cross-omic discovery through functional prediction and conversational interfaces

Increase in analytical **accuracy**, accelerated **throughput**, and enhanced **accessibility** of complex biological knowledge

Q&A

“Questions?”

Bibliography

1. Athanasopoulou, K., Michalopoulou, V.-I., Scorilas, A., & Adamopoulos, P. G. (2025). Integrating Artificial Intelligence in Next-Generation Sequencing: Advances, Challenges, and Future Directions. Current Issues in Molecular Biology, 47(6), 470. https://doi.org/10.3390/cimb47060470
2. Chen, Y., & Roberts, S. (2015). High-performance computing and big data in omics-based medicine. Frontiers in Bioinformatics / PMC – NIH. https://pmc.ncbi.nlm.nih.gov/articles/PMC4284979/
3. Clement, T., Kemmerzell, N., Abdelaal, M., & Amberg, M. (2023). XAIR: A Systematic Metareview of Explainable AI (XAI) Aligned to the Software Development Process. Machine Learning and Knowledge Extraction, 5(1), 78-108. https://doi.org/10.3390/make5010006
4. Fitzgerald, R. J., & Liu, T. (2023). Julia for Biologists. PubMed Central (PMC) – NIH. https://pmc.ncbi.nlm.nih.gov/articles/PMC10216852/
5. Harris, K., & Lenci, A. (2020). Infrastructure for semantic annotation in large-scale knowledge graphs. ACL Anthology. https://aclanthology.org/2020.lrec-1.855.pdf
6. Jaganathan K, et al. (2019) Predicting Splicing from Primary Sequence with Deep Learning. Cell. Jan 24;176(3):535-548.e24. doi: 10.1016/j.cell.2018.12.015.
7. JuliaOmix. (2025, November 12). About us. Retrieved November 12, 2025, from https://www.juliaomix.com/about-us/
8. Kato, S., Kim, K. H., Lim, H. J., Boichard, A., Goodman, A., & Kurzrock, R. (2022). Artificial intelligence–assisted serial analysis of clinical cancer genomics data identifies changing treatment recommendations and therapeutic targets. PLoS Computational Biology / PMC. https://pmc.ncbi.nlm.nih.gov/articles/PMC9167716/
9. Lifebit. (2025, November 12). AI in Genomics 2.0: What's next after the sequencing revolution. Retrieved November 12, 2025, from https://lifebit.ai/blog/ai-in-genomics-20-whats-next-after-the-sequencing-revolution/
10. Martínez, A., Zhao, L., & Singh, R. (2025). Optimizing DNA sequence classification via a deep learning hybrid of LSTM and CNN architecture. Applied Sciences, 15(15), 8225. MDPI. https://www.mdpi.com/2076-3417/15/15/8225
11. O'Connor, O., McVeigh, T.P. Increasing use of artificial intelligence in genomic medicine for cancer care- the promise and potential pitfalls. *BJC Rep* 3, 20 (2025). https://doi.org/10.1038/s44276-025-00135-4
12. Poplin, R., Chang, PC., Alexander, D. et al. (2018). A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol 36, 983–987 https://doi.org/10.1038/nbt.4235
13. Wang, H., Li, P., & Zhao, J. (2024). Integrating natural language processing and genome analysis enables accurate bacterial phenotype prediction. bioRxiv. https://www.biorxiv.org/content/10.1101/2024.12.07.627346v1.full-text
14. Wang, T., Shao, W., Huang, Z. et al. (2021) MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nat Commun 12, 3445 https://doi.org/10.1038/s41467-021-23774-w
15. Zhang, Y., Lin, D., & Chen, X. (2024). Genomic language models: Opportunities and challenges. PubMed Central (PMC). https://pmc.ncbi.nlm.nih.gov/articles/PMC11275703/