Briefings in Bioinformatics, 2026, 27, bbag229
https://doi.org/10.1093/bib/bbag229
Published: 14 May 2026
Review

OXFORD

Artificial Intelligence in genomics: a comprehensive survey of methods, resources, challenges, and prospects

Md Ishtyaq Mahmud [1,*] and Tania Banerjee [1,2]

[1]Department of Electrical and Computer Engineering, University of Houston, 4226 Martin Luther King Boulevard, 77204-4005 TX, United States
[2]Department of Information Science Technology, University of Houston, 14000 University Boulevard, 77479-0800, TX, United States
*Corresponding author. Department of Electrical and Computer Engineering, University of Houston, 4226 Martin Luther King Boulevard, 77204-4005 TX, United States.
E-mail: mmahmud4@cougarnet.uh.edu

Abstract

Artificial intelligence (AI) is reshaping genomics by enabling unprecedented insights into disease mechanisms, therapeutic design, and precision medicine. This review provides a comprehensive survey of cutting-edge AI methodologies, including machine learning, deep learning (DL), natural language processing, large language models, generative frameworks, and explainable AI, and their applications across genomics. We systematically summarize how these technologies advance key domains, such as gene sequencing, variant detection, gene expression analysis, personalized medicine, and CRISPR-based genome editing. Core computational tools, benchmark datasets, and open-source frameworks supporting AI-driven genomic research are detailed. Despite remarkable progress, challenges persist in data quality, interpretability, ethical governance, and computational scalability. Integrating multi-omics data through advanced architectures, such as graph neural networks and multimodal DL promises deeper biological understanding. Emerging paradigms, e.g. synthetic genomics and digital twins, highlight AI's potential to deliver predictive and personalized healthcare.

Keywords genomics, artificial intelligence, tools and benchmark datasets

Introduction to AI in genomics

The intersection of artificial intelligence (AI) and genomics represents a convergence of rapid innovation in biological data generation and computational analysis. The completion of the Human Genome Project in 2003 marked a historic milestone, producing the first complete sequence of human DNA and establishing a foundation for all subsequent genomic research [1]. However, the massive scale and complexity of genomic data soon exceeded the capabilities of traditional statistical approaches.

The advent of next-generation sequencing technologies dramatically reduced sequencing costs, from millions of dollars per genome to less than a thousand, leading to a surge in data from large-scale initiatives such as the 1000 Genomes Project and the UK Biobank [2]. Alongside this biological revolution, AI underwent its own transformation. The success of AlexNet in 2012 ignited the deep learning (DL) revolution [3], followed by the rise of powerful open-source frameworks such as TensorFlow and PyTorch [4], which democratized access to advanced model development.

The application of AI in genomics has since evolved from early machine learning (ML) models for variant classification to sophisticated DL and Transformer-based architectures capable of learning long-range genomic dependencies. Notable breakthroughs include Google's DeepVariant, which employs DL to identify genomic variants with exceptional precision [5], and DeepMind's AlphaFold2, which achieved near-experimental accuracy in protein structure prediction [6]. Recent adaptations of Transformer models, such as Enformer, extend these capabilities by learning the "language of the genome" and predicting the regulatory impact of non-coding variants [7].

While the trajectory of AI in genomics is transformative, the shift from classical bioinformatics to DL entails distinct technological trade-offs that this review will critically examine. Foundation Models offer unprecedented capacity to capture long-range genomic dependencies and context (genomic grammar) via self-attention, yet they introduce quadratic computational complexity ($O(N^2)$) and require massive compute resources that challenge accessibility [8, 9]. Similarly, graph neural networks (GNNs) provide a superior inductive bias for modeling non-Euclidean data like 3D protein structures and interactomes, but they face unique scalability hurdles, such as over-smoothing and neighbor explosion, when applied to large-scale biological networks [6, 10]. Finally, while Multi-Modal Learning promises holistic "digital twins" by fusing diverse omics layers, it necessitates solving complex alignment problems between heterogeneous data spaces which are

Received: November 24, 2025. **Revised:** February 18, 2026. **Accepted:** April 15, 2026

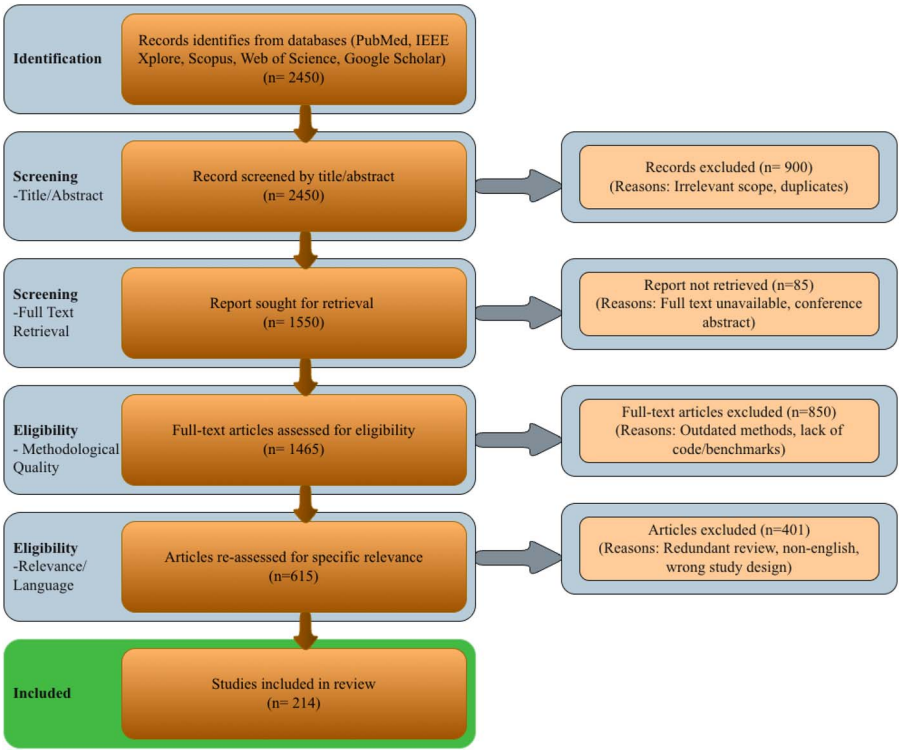

Figure 1 PRISMA flow diagram summarizing the systematic literature selection process.

challenges that simpler unimodal models avoid [11, 12]. By navigating these trade-offs, this survey aims to guide researchers in selecting the appropriate architecture for their specific genomic inquiries.

Genomic data are high-dimensional and inherently non-linear, often exceeding the modeling capacity of conventional statistical methods [13]. DL approaches such as variational autoencoders (VAEs) enable integration of heterogeneous multi-omics modalities into shared latent representations [14], while autoencoder-based architectures mitigate batch effects and technical noise [15]. Transfer learning further addresses data scarcity in rare disease settings by leveraging representations learned from larger cohorts [16].

Graph neural networks are well suited for modeling spatial transcriptomic architecture, and attention-based sequence models capture long-range regulatory dependencies beyond the reach of local motif-based approaches. End-to-end frameworks can directly infer phenotypes from raw sequencing inputs such as FASTQ files or expression matrices [17]. To improve interpretability, explainable AI methods including SHAP and Integrated Gradients are increasingly incorporated into genomic pipelines. Tools such as DeepSEA and Basenji exemplify the shift toward predictive modeling of regulatory DNA function [18, 19].

Literature search methodology

We conducted a systematic literature review in accordance with PRISMA guidelines. Searches were performed across PubMed, IEEE Xplore, Scopus, Web of Science, and Google Scholar, focusing on studies published between January 2015 and early 2026.

Keywords combined AI architectures and genomic applications: ("Artificial Intelligence" OR "Deep Learning" OR "Transformer" OR "Generative AI" OR "Foundation Models") AND ("Genomics" OR "Variant Calling" OR "Spatial Transcriptomics" OR "Precision Medicine"). Peer-reviewed articles and high-impact conference proceedings were prioritized, while foundational studies published prior to 2015 were included where relevant. As shown in Fig. 1, studies were excluded if they lacked algorithmic validation, focused on non-molecular imaging, or were unavailable in English. Of 2450 initial records, 214 met the inclusion criteria for qualitative synthesis.

This comprehensive review work provides numerous important contributions to the field of AI in genomics, including the following: (i) Provides a framework for AI in genomics, demonstrating a structured pipeline that begins with patient data and ends with customized treatment suggestions. (ii) Presents a comprehensive summary of the most recent AI technologies and the ways in which they may be applied in genomics, making difficult topics more understandable. (iii) Provides a summary of commonly used tools, datasets, and benchmarks to assist researchers in finding resources for their work. (iv) Indicates future research areas by highlighting the main obstacles and restrictions of using AI in genomics. (v) Highlights the promise for integrated and customized healthcare by providing a clear picture of forthcoming advances such as explainable AI and digital twins.

Figure 2 illustrates the structural roadmap of this study. Additionally, Table 1 provides the list of abbreviations with their full forms that are used in this study.

Artificial intelligence framework in genomics

The AI framework in genomics spans data acquisition, quality control, alignment, multi-modal integration, and model-driven inference. Standardized and privacy-preserved datasets are analyzed using ML and DL models to detect patterns, stratify patients, and infer clinically relevant biomarkers.

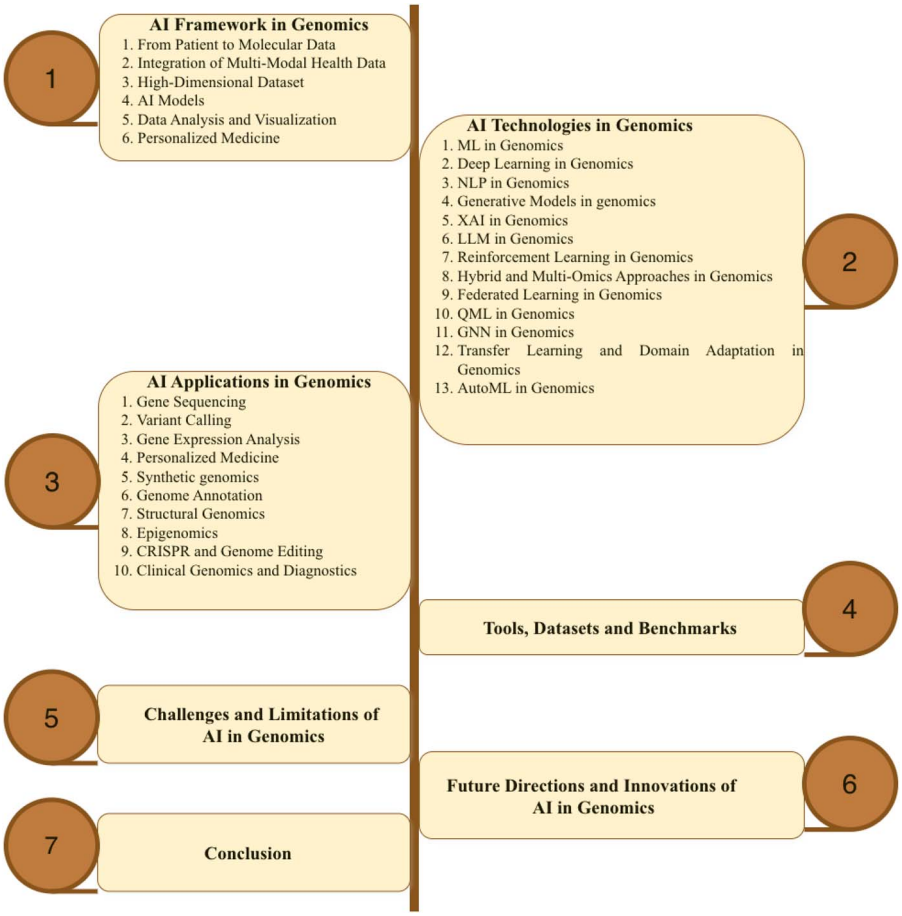

Figure 2 Structural roadmap of the review illustrating the logical transition from foundational AI frameworks to translational implementation in genomics.

Table 1 List of abbreviations used in this review.

Abbreviation	Full form
AI	artificial intelligence
ML	machine learning
LLMs	large language models
NLP	natural language processing
AutoML	automated machine learning
DL	deep learning
FL	federated learning
FDA	food and drug administration
GNN	graph neural network
GAE	graph autoencoder
GANs	generative adversarial networks
QML	quantum machine learning
RL	reinforcement learning
SHAP	Shapley additive explanations
TCGA	The cancer genome atlas
TMA	tissue microarray
VAE	variational autoencoder
XAI	explainable artificial intelligence

Once prepared, the multi-modal datasets are evaluated using ML and DL models to detect biological patterns, forecast disease risk, stratify patient subtypes, and infer therapeutically relevant biomarkers. Established metrics are used to evaluate model performance, while explainable AI approaches like SHAP, LIME, and Integrated Gradients give interpretability by highlighting influential genomic. The final stage incorporates these insights into personalized medicine, allowing for specific therapy options based on an individual's genetic and phenotypic profile. AI uses this integrated workflow to convert raw molecular data into meaningful, reliable clinical recommendations, improving diagnostic precision and treatment results.

Artificial intelligence technologies in genomics

In this section, we discuss each technology in terms of its fundamental concepts, impact, and sample models. This section discusses various methods to demonstrate the advantages and disadvantages of the available AI solutions, as well as how they are influencing the future development of precision medicine and computational genomics. Figure 3 displays various AI technologies utilized in genetics, along with the standard models utilized for each. Table 2 provides a comprehensive overview of these technologies, highlighting their main features and how they are applied.

Machine learning in genomics

ML enables modeling of high-dimensional, noisy genomic datasets beyond the limits of classical statistical frameworks [20]. In practice, supervised models are applied to variant classification, biomarker

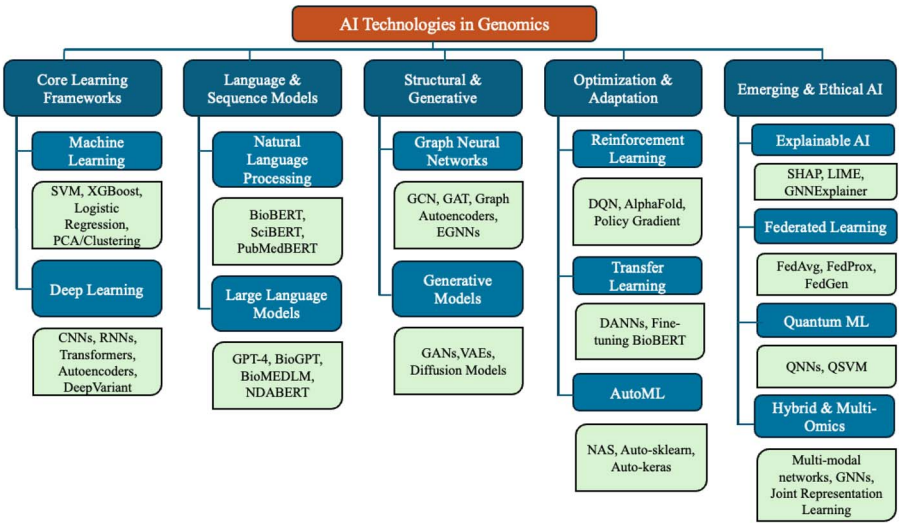

Figure 3 Taxonomy of AI technologies applied in genomics.

Table 2 Comprehensive overview of AI technologies in genomics.

Technology	Description	Typical applications	Techniques/Models	Citations
ML	Identifying patterns and relationships	Variant detection, biomarker discovery, disease prediction	SVM, RF, LR, XGBoost	[20, 22]
DL	NNs modeling complex relationships.	Gene expression analysis, variant calling, protein structure prediction	CNNs, RNNs, Transformers, AE, DeepVariant	[25–27]
NLP	Linguistic methods adapted for biological text and literature analysis.	Literature mining, gene annotation, functional extraction	BioBERT, SciBERT, PubMedBERT	[32, 33]
Generative Models	Generating synthetic genomic data.	Data augmentation, regulatory sequence modeling	GANs, VAEs, Diffusion Models	[36, 37, 57]
XAI	Making AI predictions interpretable and transparent.	Clinical diagnostics, gene interaction interpretation	SHAP, LIME, Integrated Gradients GNNExplainer	[45–48]
LLMs	Large-scale pretrained language models capturing biomedical context.	summarization, hypothesis generation	GPT-4, BioGPT, BioMedLM, NDABERT	[50, 51, 58, 59]
RL	Algorithms optimizing genomics processes through reward-based learning.	Drug–target discovery, sequencing optimization	DQN, Policy Gradient, AlphaFold	[55, 56, 60]
Hybrid and Multi-Omics Approaches	Integration of multi-omics datasets for biological insight.	Precision medicine, biomarker discovery	Multi-modal Networks, GNNs, Joint Representation Learning	[61–63]
FL	Collaborative, privacy-preserving decentralized learning.	Multi-institutional genomic studies, privacy-sensitive analyses	FedAvg, FedProx, FedGen	[64, 65]
QML	Quantum-enhanced methods for genomic data processing.	Sequence alignment, biological modeling, drug discovery	QNNs, QSVM	[66–68]
GNNs	Explicitly modeling relational biological networks.	Gene interactions, cell–cell/LR interaction modeling	GCN, GAT, GAE, EGNNs	[69–71]
Transfer Learning and Domain Adaptation	Leveraging prior knowledge from related genomic tasks for improved model performance.	Efficient training on limited data, functional prediction	Fine-tuning BioBERT, DANNs	[72, 73]
AutoML	Automating model selection, tuning, and deployment.	Rapid model building, biomarker discovery	Auto-Sklearn, Auto-Keras, NAS	[74, 75]

discovery, and disease risk prediction, while unsupervised approaches support clustering and dimensionality reduction in large-scale omics data [21, 22]. Cohorts such as TCGA and the 1000 Genomes Project have leveraged algorithms including SVMs, Random Forests, and gradient-boosting methods to identify genotype–phenotype associations [23]. Compared with traditional linear models, these approaches improve generalization and feature selection in high-dimensional settings [24].

Deep learning in genomics

Deep learning models directly capture non-linear biological structure from raw sequencing data, reducing reliance on manual feature engineering [25, 26]. Convolutional Neural Network (CNN)-based frameworks such as DeepVariant achieve near human-level SNP detection accuracy from sequencing reads [27]. Transformer architectures, including DNABERT, model long-range DNA dependencies critical for

enhancer–promoter interaction prediction [28]. VAEs further support unsupervised modeling of single-cell RNA-seq data for cell clustering and trajectory inference [26].

Beyond general-purpose architectures, task-specific attention models have emerged. The DFT_ANPD framework [29] employs a dual-feature two-sided attention mechanism to jointly model global and local molecular dependencies, demonstrating improved accuracy in anticancer natural product detection.

Natural language processing in genomics

Natural language processing (NLP) enables scalable extraction of structured information from unstructured biomedical literature and clinical records, which are impractical to curate manually [30, 31]. Transformer-based models such as BioBERT, pretrained on PubMed corpora, improve gene–disease relation extraction and clinical question answering [30, 32]. SciBERT extends this approach to broader scientific text domains [33]. Supporting resources such as PubTator Central and scispaCy provide curated annotations and named entity recognition pipelines for integrating literature-derived knowledge into genomic workflows [34, 35].

Generative models in genomics

Generative models learn underlying data distributions to synthesize biologically plausible genomic sequences, addressing the scarcity and cost of labeled experimental data [36, 37]. Architectures such as GANs, VAEs, and diffusion models have been applied to DNA sequence generation, dataset augmentation, and protein structure modeling [38–41]. In single-cell RNA-seq analysis, VAEs support denoising and latent structure discovery in high-dimensional expression data [40]. As these generative frameworks mature, rigorous evaluation becomes paramount. A recent systematic review on computational drug design [42] highlights that the success of generative architectures is heavily dependent on the choice of molecular representations (e.g. SMILES versus Graphs) and performance assessment metrics.

Explainable artificial intelligence in genomics

Explainable artificial intelligence (XAI) aims to make AI system decision-making transparent and interpretable. Clinicians must trust and understand sophisticated AI model outputs. XAI approaches reveal whether genes influence model predictions, boosting biological relevance, and clinical accountability [43, 44]. SHAP, LIME, and Integrated Gradients are popular XAI approaches. SHAP has been used to analyze cancer subtype, disease progression, and variation pathogenicity genomic models [45]. LIME uses a local linear surrogate model to explain predictions. It helps manual clinical pipeline curation by interpreting pathogenic variant classification [46]. While SHAP and LIME are effective for tabular or sequence data, graph-based architectures require specialized interpretability tools. GNNExplainer addresses this by providing a model-agnostic approach that identifies compact subgraphs and crucial node features influencing predictions. By maximizing mutual information between the GNN's prediction and the distribution of possible subgraphs, it highlights biologically relevant interaction modules (e.g. specific protein–protein interaction pathways) that drive model decisions [47]. Gradient-based integrated gradients assess feature attribution in DL models. It explains regulatory variation effects and chromatin accessibility predictions [48].

Large language models in genomics

Large language models (LLMs) were developed to represent natural language using vast text corpora. More recently, these models have been modified to comprehend biological sequences and biomedical literature. In genomics, LLMs are utilized for various tasks, including the extraction of gene–disease relations, the development of hypotheses from unstructured data, and the summarization of prior research [32, 33]. Millions of full-text articles and abstracts from PubMed that have been used to pretrained notable models like as BioBERT, PubMedBERT, and BioGPT [32, 49, 50]. Building upon foundational BERT models, NDABERT (Nucleotide-DNA BERT) introduces a specialized architecture designed to capture functional genomic elements with higher fidelity. Unlike standard DNABERT which relies on k-mer tokenization, NDABERT optimizes species-aware embeddings and curriculum contrastive learning to better identify regulatory motifs and differentiate species in metagenomic contexts. This architecture significantly enhances performance in tasks requiring fine-grained understanding of non-coding DNA syntax [51]. Large-scale ML systems like BioGPT are able to generate fluent summaries of research papers, extract structured insights from unstructured material, and provide answers to domain-specific queries [50]. Recent results revealed that OpenAI's GPT-4 performed exceptionally well on medical reasoning and USMLE-style questions, indicating that it has the potential to provide clinical genomic decision assistance [52]. In addition, LLMs are being modified so that they can have direct effects on biological sequences. Moreover, LLMs are being investigated for their potential use in the automated annotation of genetic variations and in the EHR-based patient stratification [53].

Reinforcement learning in genomics

Reinforcement learning (RL) formulates genomic tasks as sequential optimization problems in which agents learn policies through interaction with biological environments. Algorithms such as Deep Q-Networks (DQN) and Policy Gradient methods are applied to high-dimensional design and drug discovery challenges [54]. RL is particularly suited to settings with limited labeled data, where reward-driven learning guides model refinement [55]. Analogous iterative feedback mechanisms are incorporated in protein structure prediction; e.g. AlphaFold employs confidence scoring to iteratively improve structural accuracy [56].

Hybrid and multi-omics approaches in genomics

Hybrid and multi-omics frameworks integrate genomics, transcriptomics, proteomics, and metabolomics to model cross-layer molecular interactions [63]. Such integration supports patient stratification and identification of multi-modal biomarkers in complex diseases [76]. Advanced ML architectures perform joint representation learning, where multi-modal autoencoders project heterogeneous omics data into a shared latent space [61]. These representations enhance signal extraction while mitigating noise and batch effects, enabling more predictive, and mechanistic modeling of disease biology.

Federated learning in genomics

Federated learning (FL) allows model training across several decentralized datasets without exchanging raw data. This is particularly helpful in genomics, where sharing patient information and

protecting their privacy are major obstacles to pooling big datasets [64]. FL has been used for genome-wide association studies (GWAS) that protect privacy and multi-institutional clinical research [65]. Hospitals and labs can train models cooperatively using FedAvg, FedProx, and FedGen algorithms without disclosing individual data.

Despite its benefits, FL in genomics suffers data heterogeneity, communication overhead, and model drift. Robust aggregation and tailored federated models are being developed to overcome these limitations. FL is typically used with differential privacy to improve confidentiality. NVIDIA Clara and Flower are being tested for FL in healthcare and life sciences [77, 78].

Quantum machine learning in genomics

Quantum machine learning (QML) explores hybrid quantum–classical approaches for high-dimensional genomic analysis [66, 67]. Quantum SVMs and quantum neural networks (QNNs) have been investigated for sequence alignment, evolutionary modeling, and drug–target interaction tasks [79, 80]. Quantum kernel methods and circuit-based models have been applied to cancer expression classification and protein structure prediction [81]. Applications in single-cell clustering and genomic feature selection remain experimental due to hardware limitations [82].

Graph neural networks in genomics

GNNs are perfect for simulating biological networks since they are specialized for graph-structured data. In genomics, nodes can represent genes, proteins, or cells, while edges model their interactions. GNNS have been effectively applied for gene function prediction, cell–cell communication modeling, and disease gene prioritization [69, 70]. Moreover, GNNs in spatial transcriptomics help to model cell neighborhoods and the regulatory settings in which they are located. For instance, the STAGATE framework makes use of GATs to learn spatial domains and interactions from Visium or Slide-seq documents [83]. For tasks involving 3D molecular structures, standard GNNs often fail to capture geometric symmetries. Geometric equivariant graph neural networks (EGNNs) solve this by ensuring that the model's outputs remain invariant or equivariant to rotation and translation of the input coordinates. This capability is critical for structural genomics, enabling highly accurate protein quality assessment and molecular dynamics simulations without computationally expensive data augmentation [71]. In recent research, self-supervised training have been investigated as potential methods for enhancing scalability and generalizability [69, 70]. GNNs are becoming increasingly important for relational reasoning in the field of biology as the complexity and network structure of genomic information continue to further develop.

Transfer learning and domain adaptation in genomics

TL enables the reuse of models that have been trained in one domain in another, thereby significantly reducing the computational and data burden. BioBERT and DNABERT outperform generic models on gene annotation and variant classification due to domain-specific pretraining [28, 32]. TL enhances accuracy and generalizability by fine-tuning models using task-specific datasets. DANNs match source and target dataset distributions to improve cross-species gene prediction and cohort harmonization [84]. Moreover, Pretrained models are used to predict tumor subtypes in cancer genomics.

AutoML in genomics

AutoML automates model selection, feature engineering, and hyperparameter tuning in genomic workflows. Platforms such as Auto-Sklearn, Auto-Keras, and neural architecture search (NAS) enable rapid model development for gene expression classification, variant effect prediction, and biomarker discovery [85, 86]. Some systems incorporate explainability modules to support model interpretation [87]. However, exhaustive search strategies remain computationally demanding and prone to overfitting in small datasets. Integration of AutoML with federated learning (FL), transfer learning (TL), and XAI frameworks may enhance scalable precision genomics [88].

Comparative analysis: traditional bioinformatics versus AI-driven genomics

Traditional bioinformatics methods remain advantageous in low-sample regimes ($n \ll p$) and settings requiring statistical interpretability. Approaches such as PCA, HMMs, and DESeq2 provide robust performance for bulk RNA-seq and related analyses where over-parameterized DL models may overfit [89, 90]. In contrast, AI methodologies capture high-dimensional, non-linear biological interactions and automate feature learning from raw sequencing data, enabling tasks such as protein folding and multi-omics integration [91]. Table 3 summarizes these trade-offs across scalability, modeling capacity, and interpretability.

Applications of artificial intelligence in genomics

AI methods including ML, DL, NLP, and LLMs are applied across key genomic domains, such as gene sequencing, variant calling, genome editing, and clinical diagnostics. Table 4 summarizes representative applications, associated models, limitations, and clinical relevance, while Fig. 4 illustrates the major application categories.

Gene sequencing

AI-based models enhance base-calling accuracy and error correction in high-throughput sequencing workflows. CNNs, RNNs, and Transformer architectures are applied to model sequence context, with attention-based frameworks capturing long-range genomic dependencies relevant for functional prediction [28]. Such scalability supports clinical sequencing applications, including rare disease diagnosis and newborn screening [100].

Variant calling

The main goal of variant calling is to find genetic differences from a reference genome. These differences can be single-nucleotide polymorphisms (SNPs), insertions, deletions, and structural variants. Understanding the etiology of diseases, drug responses, and inherited disorders can be significantly influenced by these variants [101]. DL has revolutionized this field, enhancing accuracy far beyond traditional methods. In particular, CNNs and RNNs are effective for sequencing and image-encoded genetic alignments. DeepVariant, a tool made by Google, uses CNNs to treat variant calling as a picture classification task. It does a better job with Illumina and PacBio data [5]. Similarly, Clairvoyante uses multi-task CNNs to predict variant type, zygosity, and quality scores [102]. While accurate detection is crucial, interpreting the functional impact of variants remains a bottleneck. AlphaMissense addresses this by adapting structural

Table 3 Comparison between traditional bioinformatics methods and AI-driven genomic approaches.

Feature/criteria	Traditional bioinformatics	AI-driven genomics
Representative methods	PCA, HMM, DESeq2	LLMs, GNNs, Transformers.
Data handling	Structured, and static; optimized for tabular, lower-dimensional data such as bulk RNA-seq [91] counts. Limited capability for unstructured inputs (e.g. raw reads, histopathology images).	Unstructured, and multi-modal; integrates heterogeneous data including raw DNA sequences, 3D protein structures, imaging data, and clinical text into unified representations [92, 93].
Feature engineering	Manual and expert-driven; relies on predefined features such as k-mers or position weight matrices (PWMs). Highly interpretable but constrained by existing biological knowledge [94].	Automated representation learning; deep architectures (CNNs, transformers) discover non-linear and previously unknown features such as distal regulatory motifs [95].
Modeling non-linearity	Primarily linear or simplified probabilistic assumptions (e.g. PCA, Markovian dependencies), limiting detection of complex interactions [91].	High-capacity non-linear modeling; captures epistasis, protein folding, and 3D chromatin interactions with greater fidelity [56, 96].
Scalability	Computationally efficient for small to moderate cohorts ($n < 1000$) but performance plateaus with large-scale datasets [94].	Data-hungry yet scalable; performance improves with increasing data volume (scaling laws). Training is resource-intensive, while inference is fast and GPU-parallelizable [97, 98].
Interpretability	White-box models with statistically grounded outputs, including *P*-values, confidence intervals, and effect sizes [90].	Black-box models; interpretation relies on post-hoc XAI methods (e.g. SHAP, GNNExplainer), which may introduce instability [45].
Context awareness	Local or short-range focus; models often treat genomic elements independently or capture limited motif context [95].	Global and long-range context modeling; genomic language models capture regulatory grammar and dependencies spanning >100 kb [8, 99].

Table 4 Comprehensive overview of AI applications in genomics.

Application	Description	AI methods	Examples	Limitations	Clinical impact	Ref.
Gene Sequencing	Enhancing DNA sequencing, and accuracy.	CNN, RNN, Transformers	DeepVariant, Bonito	Performance drop with noisy data.	Enables early diagnosis of genetic diseases.	[27, 105]
Variant Calling	Detecting disease-associated genetic variants.	CNN, RNN, Ensemble models	Clairvoyante, DeepVariant, AlphaMis-sense	Hard to detect rare variants.	Critical for rare disease diagnostics.	[27, 103, 106]
Gene Expression Analysis	Analyzing gene activity across tissues.	PCA, Autoencoders, GNN	scVI, Scanpy, CellBERT	High dimensionality; interpretability challenges.	Enhances biomarker discovery.	[40, 107, 108]
Personalized Medicine	Tailoring treatments based on individual genomes.	Predictive modeling, XAI	Biomarker prediction	Privacy issues.	Precision therapies with fewer side effects.	[109, 110]
Synthetic Genomics	Designing synthetic genes and sequences.	GANs, RL	AlphaFold, RosettaFold	Ethical concerns.	Supports drug, and vaccine development.	[56, 111]
Genome Annotation	Assigning function to genomic sequences.	NLP, Transformers, Embeddings	Regulatory element prediction	Hard to annotate non-coding regions.	Identifies disease-relevant regions and targets.	[99, 112]
Structural Genomics	Predicting 3D protein structures.	DL	AlphaFold, OpenFold, ESM3	High resource usage; dependent on quality data.	Advances rational drug design.	[56, 113, 114]
Epigenomics	Modeling gene regulation via epigenetics.	GNN, Attention-based models	DeepChrome, GRNBoost	Complex and context-specific.	Enables early detection of epigenetic disorders.	[115, 116]
CRISPR and Genome Editing	Enhancing genome editing accuracy.	Predictive models, RL	DeepCRISPR, CRISPOR	Off-target effect prediction still limited.	Enables safe and precise gene therapies.	[117, 118]
Clinical Genomics and Diagnostics	Using genomic data for disease risk and diagnosis.	Classifiers, Ensemble models	Disease subtype prediction	Model biases.	Improves diagnostics and patient stratification.	[119, 120]

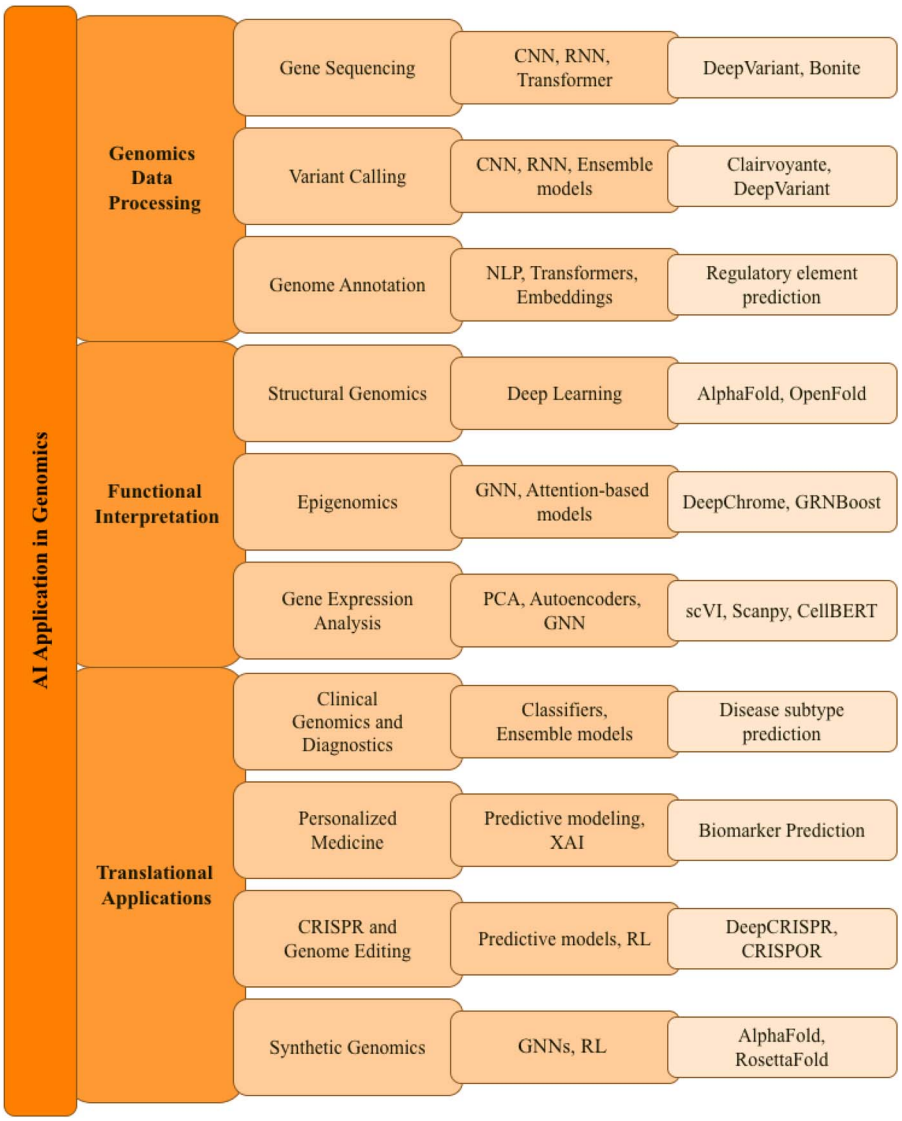

Figure 4 Overview of major AI application domains in genomics.

biology models to predict the pathogenicity of missense variants. By leveraging structural context and evolutionary constraints, it categorizes variants as likely benign or pathogenic with state-of-the-art accuracy, thereby illuminating molecular effects and aiding in the identification of novel disease-causing genes in clinical settings [103]. By mixing results from different model, and datasets, ensemble methods make predictions even more stable [104]. Overall, AI-powered variant calling is pushing the limits of genetics, from collecting data to finding insights that are useful for patients.

Gene expression analysis

Gene expression analysis quantifies RNA transcripts to characterize regulatory states and disease-associated patterns [121, 122]. In oncology, expression profiling supports tumor subtyping, prognosis estimation, and therapeutic target discovery [123]. Large-scale resources such as GTEx and TCGA link expression variation to genetic differences through eQTL mapping [124]. Single-cell RNA-seq enables high-resolution analysis of cellular heterogeneity in complex tissues [125]. Deep generative models (e.g. scVI) and GNN-based approaches improve denoising and dimensionality reduction in single-cell datasets [40]. Pretrained language models such as CellBERT further enhance annotation and contextual interpretation of expression profiles [126]. These advances strengthen biomarker discovery and support data-driven therapeutic stratification.

Personalized medicine

Personalized medicine leverages individual genomic information to guide diagnosis and therapy [127]. In oncology, tumor sequencing enables mutation-targeted treatment of alterations such as *EGFR* and *BRAF* [128]. Whole-exome and whole-genome sequencing further identify pathogenic germline variants in hereditary disorders, supporting early diagnosis and tailored interventions [129]. AI-based risk prediction and multi-omics integration improve patient stratification and clinical decision-making [130]. Explainable AI approaches are increasingly incorporated to enhance transparency and clinical trust in genomic prediction models [131].

Synthetic genomics

Synthetic genomics enables the design and construction of engineered DNA, RNA, and whole genomes for controlled investigation of gene function and regulatory mechanisms. Synthetic sequences are widely used to model mutations and optimize expression systems in functional genomics [132]. Structure-aware design is accelerated by protein prediction frameworks such as RosettaFold and AlphaFold, which infer tertiary structure from engineered sequences [56, 111]. Generative models further extend this paradigm by creating novel DNA patterns with desired structural or functional properties. Ethical and biosecurity considerations remain important in the deployment of synthetic genomic technologies [133].

Genome annotation

Genome annotation assigns functional interpretation to genes and regulatory elements within DNA sequences [134]. A major challenge lies in characterizing non-coding regions, which comprise most of the human genome and frequently harbor disease-relevant regulatory elements [95]. Transformer and NLP-based models have improved annotation accuracy by capturing complex sequence dependencies and long-range regulatory relationships [99]. Such advances support identification of pathogenic variants and regulatory hotspots with translational relevance [135].

Structural genomics

Structural genomics determines the 3D structures of proteins encoded by an organism's genome to understand gene activity at the molecular level. This field shows how genetic variants impact protein folding, stability, and interactions, connecting sequence data to biological processes [136]. AlphaFold and OpenFold DL methods predict protein structures with near-experimental accuracy from amino acid sequences [6, 113]. Beyond structure prediction, the field is evolving toward generative design. ESM3, a multimodal generative language model, represents a paradigm shift by reasoning simultaneously over sequence, structure, and function. Unlike earlier folding-only models, ESM3 extracts deep semantic features from amino acid sequences to simulate evolution, enabling the programmable generation of complex proteins and predicting tertiary structures for sequences with low evolutionary homology [114]. These techniques enable *in silico* mutational impact and protein–protein interaction network screening. Also, provide scalable structural coverage across the proteome despite their high computing costs and dependencies. Identifying allosteric regions and directing ligand docking simulations with structural knowledge aids rational drug design [137]. Structural and genetic data predict how patient-specific variations affect protein function and treatment response, enabling customized therapy.

Epigenomics

Epigenomic datasets provide critical insight into regulatory element activity and non-coding genome function [138]. Computational models have been developed to extract predictive signals from chromatin features: DeepChrome infers gene expression from histone modification patterns, and GRNBoost reconstructs gene regulatory networks from chromatin-derived signals [115, 116]. Recent advances incorporate GNNs and attention-based frameworks to model long-range chromatin interactions and high-order regulatory structure within epigenomic landscapes [139]. Such approaches have clarified the role of aberrant epigenetic regulation in tumor progression and therapy resistance [140, 141].

CRISPR and genome editing

CRISPR-based genome editing has transformed modern genomics, enabling precise, programmable DNA variations at specific genomic loci through the use of guide RNAs [142]. CRISPR is extensively utilized in genomics research for functional genomics screens that systematically identify gene–phenotype relationships [143]. Predictive models like DeepCRISPR facilitate the selection of guide RNAs that exhibit optimal on-target activity, thereby enhancing experimental reliability [117]. CRISPOR integrates genomic context and sequence features to improve guide RNA design, thereby increasing the accuracy and reproducibility of genome editing [144]. RL and DL methodologies are currently being advanced to enhance target selection and forecast editing outcomes *in silico*. CRISPR is leading in gene therapy, with clinical trials focused on monogenic disorders such as sickle cell disease, β-thalassemia, and Leber's congenital amaurosis [145].

Clinical genomics and diagnostics

Clinical genomics utilizes an individual's genomic data to diagnose disease, assess risk predisposition, and guide personalized patient care. WGS, WES are two types of sequencing technologies that help doctors find the harmful variations that cause rare genetic diseases, inherited cancers, and heart problems [146, 147]. Combining genetic data with clinical tests helps find diseases earlier, especially when phenotypic markers are not clear. SVM, RF, and ensemble classifiers are some of the ML models that are often used to divide patients into groups based on genetic traits and predict what kind of disease they will have [148, 149]. Clinical decision support tools, such as Pathoscore and AI-integrated systems within ClinVar, facilitate variant analysis, and prioritization for clinical reporting [150].

Clinical case studies and translational outcomes

While the diagnostic potential of AI is evident in research, its true value is realized through successful translation into clinical workflows. We highlight four landmark implementations that demonstrate this transition from algorithmic theory to patient care:

Rapid diagnostics in critical care (Project Baby Bear): The "Project Baby Bear" initiative at Rady Children's Hospital utilized AI-accelerated rapid whole-genome sequencing (rWGS) to reduce diagnostic turnaround times for critically ill infants to under 14 h. By deploying deep learning pipelines for variant prioritization, the project facilitated immediate targeted treatments and significantly lowered healthcare costs [151].

Multi-Cancer Early Detection (GRAIL's Galleri Test): Moving beyond symptomatic diagnosis, the GRAIL Galleri test employs ML classifiers to analyze methylation patterns in cell-free DNA (cfDNA) from liquid biopsies. Clinical validation demonstrates the model's ability to detect signals for over 50 cancer types with a false-positive rate of $<1\%$, showcasing AI's potential in shifting oncology toward non-invasive, early-stage screening [152].

Resolving diagnostic odysseys (100 000 Genomes Project): For rare diseases, the UK's 100 000 Genomes Project leveraged AI-driven variant prioritization tools to resolve long-standing "diagnostic odysseys". This approach successfully identified causal variants in a

significant cohort of previously undiagnosed patients, validating the scalability of automated analysis for population genomics [153].

AI-driven therapeutics (Insilico Medicine): In drug discovery, INS018_055 by Insilico Medicine marks a significant breakthrough as the first fully AI-generated drug candidate to reach Phase II clinical trials. Designed as a small-molecule inhibitor for Idiopathic Pulmonary Fibrosis (IPF), this milestone illustrates how generative AI is used for both target identification and molecular design can drastically compress discovery timelines and bridge the gap between genomic insight and therapeutic intervention [154].

Tools, datasets, and benchmarks of artificial intelligence in genomics

This section and Table 5 present a structured overview of the essential tools, datasets, and benchmarks that support AI-driven research in genomics and facilitate its clinical application.

Tools

AI-driven genomics leverages established deep learning frameworks such as TensorFlow and PyTorch for sequence modeling, variant calling, graph neural networks, and multi-omics integration [4, 155].

Single-cell and spatial transcriptomics analyses are supported by scvi-tools for probabilistic modeling and batch correction [156], Scanpy and Seurat for large-scale preprocessing and integration [107, 157], and Giotto for spatial domain and cell–cell interaction analysis [158]. For bulk RNA-seq, DESeq2 remains a standard for differential expression analysis [90]. Model interpretability is commonly addressed using SHAP and LIME [45, 46].

Datasets

AI applications in genomics are enabled and validated by diverse publicly available datasets spanning cancer genomics, population transcriptomics, single-cell atlases, structural biology, and spatial omics. The Cancer Genome Atlas (TCGA) remains one of the largest multi-omic resources, comprising over 11 000 tumor samples across 33 cancer types with matched clinical annotations, and is widely used for subtype classification, survival modeling, and integrative analyses [159, 160]. The Gene Expression Omnibus (GEO) serves as a large-scale repository for functional genomics data and supports meta-analysis, benchmarking, and TL studies [161, 162].

The GTEx project provides a comprehensive tissue-wide expression atlas across 54 human tissues, enabling eQTL mapping, and transcriptome-wide association studies [163–165]. Complementing bulk tissue resources, the human cell atlas (HCA) profiles millions of cells to support cell-type annotation, developmental trajectory inference, and cross-species comparisons [166–168]. For structural genomics, AlphaFold DB supplies high-confidence structural predictions for over 214 million proteins, facilitating variant interpretation and drug–target modeling [6, 169].

Spatial transcriptomics resources further extend these capabilities. DeepSpaceDB, HEST-1k, Xenium, and Visium collectively provide standardized spatially resolved transcriptomic and histopathology datasets that support benchmarking of spatial GNNs, multimodal pretraining, and tumor microenvironment analysis [170–179].

Finally, to overcome the limitations of linear reference genomes, the Human Pangenome Reference Consortium (HPRC) provides a diverse collection of diploid assemblies. This dataset is critical for benchmarking graph-based genomic analyses, enabling the detection of complex structural variations often missed by standard references [180].

Benchmarks

Benchmark datasets and community-driven challenges are essential for standardized evaluation of AI models in genomics. The DREAM Challenges provide community-wide evaluation frameworks for tasks such as network inference, mutation impact prediction, and drug response modeling [181–183]. The ENCODE project establishes regulatory genomics standards through curated annotations of promoters, enhancers, and chromatin states, serving as a benchmark for transcription factor binding and motif discovery [184–186]. GenBench further extends benchmarking to genomic foundation models through modular tasks in sequence modeling and gene expression prediction [99, 187]. CAMDA complements these efforts by organizing large-scale data analysis competitions using clinically relevant biological datasets [188].

Beyond these community challenges, specialized benchmarks are required to evaluate modern AI architectures. Genome in a Bottle (GIAB) serves as the gold standard for germline variant calling, providing highly curated reference genotypes [189]. For protein biology, ProteinGym and TAPE (Tasks Assessing Protein Embeddings) establish rigorous standards for evaluating LLMs on fitness prediction and structure modeling tasks [190, 191]. In single-cell genomics, the scIB (Single-cell Integration Benchmark) suite metrics are essential for assessing batch-correction efficacy [192]. Furthermore, to standardize genomic foundation models, GUANinE and UEE (Universal Expression Embeddings) provide de-noised tasks for sequence-to-function modeling and gene expression analysis, ensuring models are tested on generalizable biological signals [193, 194].

Challenges and limitations of artificial intelligence in genomics

Limited availability of high-quality, annotated genomic datasets remains a central constraint. Batch effects, noise, and platform-specific biases reduce model robustness and reproducibility [62, 198]. Obtaining labeled data often requires labor-intensive experimental validation, particularly for disease phenotypes and regulatory elements [199].

DL models are frequently criticized for limited interpretability [45]. Integrating heterogeneous multi-omics data (proteomics, DNA, RNA, epigenetics) remains challenging due to differences in scale, resolution, and noise characteristics [62]. Although XAI tools such as SHAP and LIME provide *post-hoc* explanations, they are not always biologically intuitive in genomic contexts [200]. Model uncertainty is often insufficiently quantified in high-stakes settings such as clinical diagnostics [201]. Reproducibility issues, inconsistent reporting standards, and limited tool accessibility further hinder translational adoption [202].

Computational efficiency and scalability analysis

Beyond data and interpretability, the widespread adoption of DL is constrained by significant computational bottlenecks. While foundational models offer superior predictive power, they introduce a

Table 5 Summary of tools, datasets, and benchmarks for AI in genomics.

Category	Name/Resource	Description	Key features	Typical application area	Citation
Tools	TensorFlow	Large-scale framework	Multi-device execution; rich ecosystem	Variant calling, sequence modeling, protein-fold prediction	[4]
	PyTorch	Imperative DL library	Dynamic graphs; TorchServe; TorchScript	Gene-expression modeling, graph neural nets on omics	[155]
	scvi-tools	Probabilistic models	PyTorch-Lightning and AnnData; variational inference	Batch correction, multi-omic integration	[156]
	Scanpy	Single-cell analysis toolkit	Highly scalable; AnnData format	Clustering, trajectory inference, QC pipelines	[107]
	Seurat	Single-cell analysis toolkit	Integration anchors; multimodal support	Cross-condition integration, cell-type annotation	[195]
	Giotto	Spatial transcriptomics toolbox	Interactive viewer, niche analysis, ligand–receptor tools	Cell–cell interaction in spatial omics	[158]
	DESeq2	RNA-seq differential expression	generalized linear models	Disease–control DE, pathway analysis	[90]
	SHAP/LIME	Model-agnostic explainability	local surrogate models	Interpreting CNNs on sequence	[45, 46]
Datasets	TCGA	Multi-omic cancer atlas	>11 k tumors; matched clinical data	Cancer subtype discovery, survival models	[159]
	GEO	Public expression repository	5 M+ samples; MIAME compliant	Meta-analysis, TL	[161]
	GTEx	Tissue-wide expression atlas	54 tissues; eQTLs	Tissue-specific regulation, TWAS	[163]
	HCA	Single-cell reference atlas	Millions of cells	Cell-type reference, cross-species mapping	[166]
	AlphaFold DB	Predicted 3-D protein structures	214 M structures; per-residue confidence	Variant effect prediction, drug target modeling	[169]
	DeepSpaceDB	Curated spatial-omics archive	Visium and Xenium cohorts	Benchmarking spatial GNNs, domain adaptation	[170]
	HEST-1k	1229 spatial transcriptomes	whole-slide images; annotations	multimodal pretraining	[172]
	Xenium	*In-situ* single-cell spatial data	protein + RNA	Subcellular localization studies	[174]
	Visium	Slide-based spatial RNA-seq	55 μm spots; whole-section capture	tumor micro-environment	[177]
	HPRC	Human Pangenome Reference	47+ diploid assemblies; graph-based	Graph genomics, Structural Variation detection	[180]
Benchmarks	DREAM Challenges	Community prediction contests	Blind test sets	Network inference, mutation impact	[181]
	ENCODE Benchmarks	Regulatory genomics standards	Motif discovery	Functional element prediction	[184]
	GenBench	Genomic foundation-model benchmarks	Modular tasks	Zero-shot and few-shot generalization	[196]
	CAMDA	Annual data-analysis contests	Very-large datasets; open methods	Meta-omics, clinical outcome prediction	[188, 197]
	GIAB	Reference standards for variants	High-confidence calls; diverse ancestries	Benchmarking germline variant calling accuracy	[189]
	ProteinGym	Protein design benchmark	2.7 M+ mutated sequences; zero-shot	Evaluating Protein LLMs and fitness prediction	[190]
	TAPE	Protein embedding tasks	5 downstream tasks (structure, fluorescence)	Evaluating self-supervised protein models	[191]
	scIB	Integration metrics suite	14 metrics for bio-conservation and batch removal	Benchmarking multi-omics integration methods	[192]
	GUANinE	Genomic AI benchmark	Sequence-to-function tasks; de-noised	Benchmarking DNA Foundation Models	[193]
	UEE	Universal Expression Embeddings	Multi-species, multi-tissue tasks	Benchmarking gene expression models	[194]

distinct trade-off between model complexity and resource efficiency. Transformer-based architectures (e.g. ESM3, DNABERT), which rely on self-attention mechanisms, suffer from quadratic computational complexity ($O(N^2)$) with respect to sequence length [203]. Training these models requires massive HPC clusters equipped with enterprise-grade GPUs (e.g. NVIDIA A100/H100), creating a steep barrier for clinical environments and resource-constrained laboratories [204].

In contrast, DeepVariant [5] exhibits linear scalability and is highly optimized for standard hardware, making it more feasible for routine diagnostic workflows. GNNs face unique scalability challenges; the "neighbor explosion" problem during message passing on large biological networks often exceeds the VRAM capacity of consumer-grade hardware [205].

To address these limitations, the field is pivoting toward efficiency-optimization techniques. Methods such as Low-Rank Adaptation (LoRA) [206], Knowledge Distillation, and 4-bit Quantization [207] are demonstrating the ability to reduce the memory footprint of genomic foundation models by ∼50% with negligible loss in accuracy. Adopting these "Green AI" strategies is essential for democratizing access to state-of-the-art genomic tools beyond elite research centers.

Future directions and innovations of artificial intelligence in genomics

AI in genomics is advancing toward more integrative and interpretable modeling frameworks. Continued progress in XAI aims to enhance transparency, enabling clearer attribution of genomic variants, and expression patterns to model predictions [208].

The field is shifting from task-specific architectures to genomic foundation models (GFMs) [209]. Leveraging self-supervised pretraining on large-scale unlabeled sequences, models such as HyenaDNA [8] and Nucleotide Transformer [9] learn generalizable representations of genomic syntax [210]. In single-cell biology, scGPT extends this paradigm to characterize cellular heterogeneity and predict gene perturbations [93].

Parallel progress in multi-modal integration addresses cross-domain data fusion. Pathomic Fusion combines histopathology and genomic features for improved prognostic modeling [211], while generative models such as ESM3 jointly model sequence, structure, and function to enable programmable protein design [114].

Digital twin frameworks represent a systems-level extension of these approaches, integrating multi-omics profiles with electronic health records and longitudinal phenotypic data into continuously updated patient-specific models [212, 213]. Such representations support simulation of disease trajectories and individualized therapy optimization within genomic contexts [214].

Conclusion

AI integration into genomics has advanced variant interpretation, protein structure prediction, and data-driven therapeutic design through deep learning, foundation models, and multi-omics frameworks. Contemporary AI-enabled genomics is defined by scalable architectures, curated datasets, and standardized benchmarks that support clinical translation.

However, challenges in interpretability, bias control, reproducibility, and data governance continue to constrain deployment in healthcare settings. Progress will require standardized validation protocols, transparent benchmarking, and robust regulatory oversight to ensure reliability and equity across populations.

Future advances will increasingly rely on multimodal integration, graph-based learning, and generative design to model complex biological systems. Emerging paradigms, including synthetic genomics and digital twin frameworks, extend AI from predictive analytics toward mechanistic simulation and individualized intervention. Sustained collaboration across computational, biological, and clinical disciplines will be essential to translate these capabilities into scalable and responsible genomic medicine.

Key Points

- This paper demonstrated a paradigm for AI in genomics, illustrating a structured pipeline that starts with patient data and concludes with personalized therapy recommendations.
- This paper offers an extensive overview of AI methods, datasets, and technologies utilized in essential genomic activities.
- We presented algorithms derived from technologies that expedite genomics research in healthcare, while simultaneously emphasizing their strengths and limits.
- Challenges like interpretability, generalizability, and dataset bias are reviewed, followed with suggestions for further advancement.

Funding

This work was supported by the National Institutes of Health through the Clinical and Translational Science Awards (CTSA) Program under award number UM1TR004539. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Data availability

Does not produce or analyze any new data for this review work.

1. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* 2004;**431**:931–45.
2. Akintunde O, Tucker T, Carabetta VJ. The evolution of next-generation sequencing technologies. In: Carabetta VJ, Akintunde O (eds.), *High Throughput Gene Screening: Methods and Protocols*, pp. 3–29. New York, NY: Springer, 2024.
3. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM* 2017;**60**:84–90.
4. Abadi M, Barham P, Chen J *et al.* TensorFlow: a system for large-scale machine learning. In: *Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation (OSDI'16)*, pp. 265–83. USENIX Association, 2016.
5. Poplin R, Chang PC, Alexander D *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* 2018;**36**:983–7.
6. Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with alphafold. *Nature* 2021;**596**:583–9.

7. Zeisel A, Muñoz-Manchado AB, Codeluppi S *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 2015;**347**:1138–42.
8. Nguyen E, Poli M, Faizi M *et al.* HyenaDNA: long-range genomic sequence modeling at single nucleotide resolution. *Adv Neural Inf Proces Syst* 2023;**36**:43177–43,201.
9. Dalla-Torre H, Gonzalez L, Mendoza-Revilla J *et al.* Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nat Methods* 2025;**22**:287–97.
10. Li MM, Huang K, Zitnik M. Graph representation learning in biomedicine and healthcare. *Nat Biomed Eng* 2022;**6**:1353–69.
11. Kline A, Wang H, Li Y *et al.* Multimodal machine learning in precision health: a scoping review. *NPJ Digit Med* 2022;**5**:171.
12. Armeni P, Polat I, De Rossi LM *et al.* Digital twins in healthcare: is it the beginning of a new era of evidence-based medicine? A critical review. *J Pers Med* 2022;**12**:1255.
13. Karthik A, Sahoo SK, Kumar A *et al.* Unified approach for accurate brain tumor multi-classification and segmentation through fusion of advanced methodologies. *Biomedical Signal Processing and Control* 2025;**100**:106872.
14. Shah M, Sureja N. A comprehensive review of bias in deep learning models: methods, impacts, and future directions. *Arch Comput Methods Eng* 2025;**32**:255–67.
15. Hanna MG, Pantanowitz L, Dash R *et al.* Future of artificial intelligence (AI)-machine learning (ML) trends in pathology and medicine. *Mod Pathol* 2025;100705.
16. Wang Y, Yao Q, Kwok JT *et al.* Generalizing from a few examples: a survey on few-shot learning. *ACM Comp Surveys* 2020;**53**:1–34.
17. Gabriel L, Becker F, Hoff KJ *et al.* Tiberius: end-to-end deep learning with an HMM for gene prediction. *Bioinformatics* 2024;**40**: btae685.
18. Barbadilla-Martínez L, Klaassen N, vanSteensel B *et al.* Predicting gene expression from DNA sequence using deep learning models. *Nat Rev Genet* 2025;1–15.
19. Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res* 2016;**26**:990–9.
20. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet* 2015;**16**:321–32.
21. Mahmud MI, Kochat V, Satpati S *et al.* Benchmarking dimensionality reduction techniques for spatial transcriptomics. In: *Proceedings of the 16th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB '25)*, pp. 1–10. New York, NY, USA: Association for Computing Machinery, 2025. https://doi.org/10.1145/3765612.3767237
22. Schrider DR, Kern AD. Supervised machine learning for population genetics: a new paradigm. *Trends Genet* 2018;**34**:301–12.
23. Bishop CM, Nasrabadi NM. *Pattern Recognition and Machine Learning*, Vol. **4**. New York, NY: Springer, 2006.
24. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–94. San Francisco, CA, USA: ACM, 2016.
25. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; **521**:436–44.
26. Zou J, Huss M, Abid A *et al.* A primer on deep learning in genomics. *Nat Genet* 2019;**51**:12–8.
27. Poplin R, Chang P-C, Alexander D *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* 2018;**36**:983–7.
28. Ji Y, Zhou Z, Liu H *et al.* DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* 2021;**37**:2112–20.
29. Norouzi R, Norouzi R, Abbasi K *et al.* DFT_ANPD: a dual-feature two-sided attention network for anticancer natural products detection. *Comput Biol Med* 2025;**194**:110442.
30. Wei L, Ke H, Cha C *et al.* Text mining for discovering gene-disease associations: a review. *Curr Bioinforma* 2020;**15**:140–8.
31. Demner-Fushman D, Ide NC, You B-H *et al.* Extracting genomic medicine information from full text articles. *BMC Bioinformatics* 2019;**20**:130.
32. Lee J, Yoon W, Kim S *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;**36**:1234–40.
33. Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3615–20. 2019, November.
34. Wei C-H, Kao H-Y, Zhiyong L. PubTator central: a web service for visualizing and retrieving curated biomedical information from the entire PubMed corpus. *Nucleic Acids Res* 2019; **47**:W588–93.
35. Neumann M, King D, Beltagy I *et al.* ScispaCy: fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pp. 319–27. 2019, August.
36. Killoran N, Lee LJ, Delong A *et al.* Generating and designing DNA with deep generative models. *arXiv preprint arXiv:1712.06148*, 2017.
37. Grønbech CH, Vording MF, Timshel PN *et al.* scVAE: variational auto-encoders for single-cell gene expression data. *Bioinformatics* 2020;**36**:4415–22.
38. Gupta A, Zou J. Feedback GAN for DNA optimizes protein functions. *Nat Mach Intell* 2019;**1**:105–11.
39. Davidsen K, Matsen FAIV. Deep generative models for T cell receptor protein sequences. *Front Immunol* 2019;**10**:2019.
40. Lopez R, Regier J, Cole MB *et al.* Deep generative modeling for single-cell transcriptomics. *Nat Methods* 2018;**15**:1053–8.
41. Schneuing A, Harris C, Du Y *et al.* Structure-based drug design with equivariant diffusion models. *Nature Computational Science* **2024**;**4**:899–909.
42. Abbasi K, Razzaghi P, Gharizadeh A *et al.* Computational drug design in the artificial intelligence era: a systematic review of molecular representations, generative architectures, and performance assessment. *Pharmacol Rev* 2026;**78**:100095.
43. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.
44. Samek W, Montavon G, Lapuschkin S *et al.* Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. *IEEE Signal Process Mag* 2021;**38**:77–88.
45. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Proces Syst* 2017;**30**:4768–77.
46. Ribeiro MT, Singh S, Guestrin C. ' Why should i trust you?' Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–44. San Francisco, CA, USA: Association for Computing Machinery, 2016.
47. Ying Z, Bourgeois D, You J *et al.* GNNExplainer: generating explanations for graph neural networks. *Adv Neural Inf Proces Syst* 2019;**32**:9240–51.

48. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: *International Conference on Machine Learning*, pp. 3319–28. Sydney, Australia: PMLR, 2017.
49. Yu G, Tinn R, Cheng H *et al.* Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans Comput Healthc* 2021;**3**:1–23.
50. Luo R, Sun L, Xia Y *et al.* BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief Bioinform* 2022;**23**:bbac409.
51. Zhou Z, Weimin W, Ho H *et al.* DNABERT-S: pioneering species differentiation with species-aware DNA embeddings. *Bioinformatics* 2025;**41**:i255–64.
52. Singhal K, Azizi S, Tao T *et al.* Large language models encode clinical knowledge. *Nature* 2023;**620**:172–80.
53. Smoller JW. The use of electronic health records for psychiatric phenotyping and genomics. *Am J Med Genet B Neuropsychiatr Genet* 2018;**177**:601–12.
54. Zhavoronkov A, Ivanenkov YA, Aliper A *et al.* Deep reinforcement learning for de novo drug design. *Mol Pharm* 2019;**16**:4282–91.
55. Angermueller C, Dohan D, Belanger D *et al.* Model-based reinforcement learning for biological sequence design. In: *International conference on learning representations*. New Orleans, LA, USA, 2019.
56. Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9.
57. Anand N, Huang P. Generative modeling for protein structures. *Adv Neural Inf Proces Syst* 2018;**31**.
58. Singhal K, Azizi S, Tao T *et al.* Large language models encode clinical knowledge. *Nature* 2023;**620**:172–80.
59. Nori H, King N, McKinney SM *et al.* Capabilities of GPT-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.
60. Blanchard AE, Shekar MC, Gao S *et al.* Automating genetic algorithm mutations for molecules using a masked language model. *IEEE Trans Evol Comput* 2022;**26**:793–9.
61. Zitnik M, Nguyen F, Wang B *et al.* Machine learning for integrating data in biology and medicine: principles, practice, and opportunities. *Inf Fusion* 2019;**50**:71–91.
62. Kang M, Ko E, Mersha TB. A roadmap for multi-omics data integration using deep learning. *Brief Bioinform* 2022;**23**:bbab454.
63. Subramanian I, Verma S, Kumar S *et al.* Multi-omics data integration, interpretation, and its application. *Bioinform Biol Insights* 2020;**14**:1177932219899051.
64. Rieke N, Hancox J, Li W *et al.* The future of digital health with federated learning. *NPJ Digit Med* 2020;**3**:119.
65. Sheller MJ, Brandon Edwards G, Reina A *et al.* Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci Rep* 2020;**10**:12598.
66. Emani PS, Warrell J, Anticevic A *et al.* Quantum computing at the frontiers of biological sciences. *Nat Methods* 2021;**18**:701–9.
67. Flöther FF. The state of quantum computing applications in health and medicine. *Research Directions: Quantum Technologies* 2023;**1**:e10. https://doi.org/10.1017/qut.2023.4
68. Outeiral C, Strahm M, Shi J *et al.* The prospects of quantum computing in computational molecular biology. *Wiley Interdiscip Rev Comput Mol Sci* 2021;**11**:e1481.
69. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* 2018;**34**:i457–66.
70. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.
71. Satorras VG, Hoogeboom E, Welling M. E (n) equivariant graph neural networks. In: *International Conference on Machine Learning*, pp. 9323–32. Brookline, MA, USA: PMLR, 2021.
72. Pan SJ. Transfer learning. *Learning* 2020;**21**:1–2.
73. Weiss K, Khoshgoftaar TM, Wang DD. A survey of transfer learning. *J Big Data* 2016;**3**:1–40.
74. Feurer M, Klein A, Eggensperger K *et al.* Efficient and robust automated machine learning. *Adv Neural Inf Proces Syst* 2015;**28**: 2962–70.
75. Jin H, Song Q, Xia H. Auto-Keras: an efficient neural architecture search system. In: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1946–56. New York, NY, USA: Association for Computing Machinery, 2019.
76. Wang Y, Zhang Z, Lin S *et al.* A review of deep learning for multi-omics data integration. *Brief Bioinform* 2021;**22**:bbab286.
77. Corporation NVIDIA. NVIDIA Clara: AI and HPC for medical imaging, genomics, and the development and deployment of smart sensors. https://developer.nvidia.com/clara. 2023. (June 2025, date last accessed).
78. Beutel DJ, Topal T, Mathur A *et al.* Flower: a friendly federated learning framework. https://flower.dev. 2020. (June 2025, date last accessed).
79. Li RY, Di Felice R, Rohs R. Quantum computing for genomics. *Trends Genet* 2021;**37**:883–94.
80. Emani PS, Warrell J, Anticevic A. Quantum computing at the frontiers of biological sciences. *Nat Methods* 2021;**18**:701–9.
81. Schuld M, Bocharov A, Svore KM *et al.* Circuit-centric quantum classifiers. *Phys Rev A* 2020;**101**:032308.
82. Mari A, Bromley TR, Izaac J *et al.* Transfer learning in hybrid classical-quantum neural networks. *Quantum* 2020;**4**:340.
83. Dong Z, Zhang H. STAGATE: spatial transcriptomics analysis via graph attention auto-encoder. *Bioinformatics* 2022;**38**:3494–502.
84. Ganin Y, Ustinova E, Ajakan H *et al.* Domain-adversarial training of neural networks. *J Mach Learn Res* 2016;**17**:2096–30.
85. Feurer M, Eggensperger K, Falkner S *et al.* Auto-sklearn: efficient and robust automated machine learning. In: *Automated Machine Learning, Hutter F, Kotthoff L, Vanschoren J*, pp. 113–34. Cham, Switzerland: Springer, 2019.
86. Olson RS, Bartley N, Urbanowicz RJ *et al.* Evaluation of a tree-based pipeline optimization tool for automating data science. In: *Proceedings of the Genetic and Evolutionary Computation Conference 2016*, pp. 485–92. New York, NY, USA: Association for Computing Machinery, 2016.
87. Liu D, Xu C, He W *et al.* Autogenome: an automl tool for genomic research. *Artificial Intelligence in the Life Sciences* 2021;**1**:100017.
88. Ivan T, Mosca and coworkers. Automated machine learning for genome-wide association studies. *Bioinformatics* 2023;**39**:7261–70.
89. Mahmud MI, Kochat V, Satpati S *et al.* hSNMF: hybrid spatially regularized NMF for image-derived spatial transcriptomics. *arXiv preprint arXiv:2602.02638*, 2026.
90. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:1–21.
91. Eraslan G, Avsec Ž, Gagneur J *et al.* Deep learning: new computational modelling techniques for genomics. *Nat Rev Genet* 2019;**20**:389–403.
92. Chen RJ, Lu MY, Wang J *et al.* Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Trans Med Imaging* 2022;**41**: 757–70.

93. Cui H, Wang C, Maan H *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods* 2024;**21**:1470–80.
94. Whalen S, Schreiber J, Noble WS *et al.* Navigating the pitfalls of applying machine learning in genomics. *Nat Rev Genet* 2022;**23**:95–108.
95. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* 2015;**12**:931–4.
96. Fudenberg G, Kelley DR, Pollard KS. Predicting 3D genome folding from DNA sequence with Akita. *Nat Methods* 2020;**17**:1111–7.
97. Dalla-Torre H, Gonzalez L, Mendoza-Revilla J *et al.* Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nat Methods* 2025;**22**:287–97. https://doi.org/10.1038/s41592-024-02523-z
98. Tay Y, Dehghani M, Bahri D *et al.* Efficient transformers: a survey. *ACM Comput Surv* 2022;**55**:1–28.
99. Avsec Ž, Agarwal V, Visentin D *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* 2021;**18**:1196–203.
100. Dias R, Torkamani A. Artificial intelligence in clinical and genomic diagnostics. *Genome Med* 2019;**11**:1–13.
101. McKenna A, Hanna M, Banks E *et al.* The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;**20**:1297–303.
102. Luo R, Wong T-H, Wong YL *et al.* Clairvoyante: a multi-task convolutional deep neural network for variant calling in single molecule sequencing. *Nat Commun* 2019;**10**:998.
103. Cheng J, Novati G, Pan J *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* 2023;**381**:eadg7492.
104. Wenger AM, Peluso P, Rowell WJ *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* 2019;**37**:1155–62.
105. Zhimeng X, Mai Y, Liu D *et al.* Fast-Bonito: a faster deep learning based basecaller for nanopore sequencing. *Artif Intell Life Sci* 2021;**1**:100011.
106. Zheng Z, Li S, Su J *et al.* Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. *Nat Comput Sci* 2022;**2**:797–803.
107. Wolf FA, Angerer P, Theis FJ. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biol* 2018;**19**:1–5.
108. Yang F, Wang W, Wang F *et al.* Scbert as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat Mach Intell* 2022;**4**:852–66.
109. Chen R, Mias GI, Li-Pook-Than J *et al.* Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* 2012;**148**:1293–307.
110. Bhalla S, Laganà A. Artificial intelligence for precision oncology. In: *Computational Methods for Precision Oncology*, pp. 249–68. Springer, 2022.
111. Baek M, DiMaio F, Anishchenko I *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 2021;**373**:871–6.
112. Kelley DR, Reshef YA, Bileschi M *et al.* Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res* 2018;**28**:739–50.
113. Ahdritz G, Bouatta N, Floristean C *et al.* Openfold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *Nat Methods* 2024;**21**:1514–24.
114. Hayes T, Rao R, Akin H *et al.* Simulating 500 million years of evolution with a language model. *Science* 2025;**387**:850–8.
115. Singh R, Lanchantin J, Robins G *et al.* Deepchrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics* 2016;**32**:i639–48.
116. Moerman T, Santos SA, González-Blas CB *et al.* GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics* 2019;**35**:2159–61.
117. Chuai G, Ma H, Yan J *et al.* DeepCRISPR: optimized CRISPR guide RNA design by deep learning. *Genome Biol* 2018;**19**:1–18.
118. Concordet J-P, Haeussler M. CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Res* 2018;**46**:W242–5.
119. Esteva A, Robicquet A, Ramsundar B *et al.* A guide to deep learning in healthcare. *Nat Med* 2019;**25**:24–9.
120. Vassy JL, Christensen KD, Schonman EF *et al.* The impact of whole-genome sequencing on the primary care and outcomes of healthy adult patients: a pilot randomized trial. *Ann Intern Med* 2017;**167**:159–69.
121. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;**10**:57–63.
122. Marioni JC, Mason CE, Mane SM *et al.* RNA-Seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* 2008;**18**:1509–17.
123. Venet D, Dumont JE, Detours V. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput Biol* 2011;**7**:e1002240.
124. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 2020;**369**:1318–30.
125. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* 2019;**15**:e8746.
126. Chen H, Altschuler SJ, Lei W. CellBERT: a pre-trained language model for single-cell RNA sequencing data. *Nat Methods* 2022;**19**:1059–67.
127. Collins FS, Manolio TA. The case for a US prospective cohort study of genes and environment. *Nature* 2004;**429**:475–7.
128. Dienstmann R, Rodon J, Barretina J *et al.* Precision oncology: a review of genomically guided cancer care. *Am J Cancer Res* 2018;**8**:1466–76.
129. Green RC, Goddard KA, Jarvik GP *et al.* Clinical sequencing exploratory research consortium: accelerating evidence-based practice of genomic medicine. *Am J Hum Genet* 2013;**93**:667–76.
130. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet* 2018;**19**:581–90.
131. Ginsburg GS, Phillips KA. Genomic medicine: a primer. *N Engl J Med* 2018;**379**:1442–50.
132. Gibson DG, Glass JI, Lartigue C *et al.* Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* 2010;**329**:52–6.
133. Baldwin C, Blackburn J, Fair J *et al.* Synthetic biology and biosecurity: learning from history, anticipating the future. *Front Bioeng Biotechnol* 2020;**8**:1104.
134. Yandell M, Ence D. Genome annotation—primer for the novice. *Nat Rev Genet* 2012;**13**:329–42.
135. Gill P, Datta A, Hurley JA *et al.* Identification of non-coding variants affecting gene expression and disease risk using deep learning. *Nat Commun* 2022;**13**:7001.
136. Baker D, Sali A. Protein structure prediction and structural genomics. *Science* 2001;**294**:93–6.
137. Mullard A. What does AlphaFold mean for drug discovery? *Nat Rev Drug Discov* 2021;**20**:725–8.

138. Roadmap EC, Kundaje A, Meuleman W *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* 2015;**518**:317–30.
139. Karbalayghareh A, Sahin M, Leslie CS. Chromatin interaction–aware gene regulatory modeling with graph attention networks. *Genome Res* 2022;**32**:930–44.
140. Sharma S, Kelly TK, Jones PA. Epigenetics in cancer. *Carcinogenesis* 2010;**31**:27–36.
141. Holliday R. Epigenetics: a historical overview. *Epigenetics* 2006;**1**:76–80.
142. Le Cong F, Ran A, Cox D *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* 2013;**339**:819–23.
143. Boettcher M, McManus MT. Choosing the right tool for the job: RNAi, TALEN, or CRISPR. *Mol Cell* 2015;**58**:575–85.
144. Haeussler M, Schönig K, Eckert H *et al.* Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool crispor. *Genome Biol* 2016;**17**:1–12.
145. Frangoul H, David Altshuler M, Cappellini D *et al.* CRISPR-Cas9 gene editing for sickle cell disease and β-thalassemia. *N Engl J Med* 2021;**384**:252–60.
146. Bamshad MJ, Ng SB, Bigham AW *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* 2011;**12**:745–55.
147. Rehm HL, Berg JS, Brooks LD *et al.* Clingen—The clinical genome resource. *N Engl J Med* 2015;**372**:2235–42.
148. Ashley EA. Towards precision medicine. *Nat Rev Genet* 2016; **17**:507–22.
149. Alipanahi B, Delong A, Weirauch MT *et al.* Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nat Biotechnol* 2015;**33**:831–8.
150. Landrum MJ, Lee JM, Benson M *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* 2018;**46**:D1062–7.
151. Dimmock D, Caylor S, Waldman B *et al.* Project baby bear: rapid precision care incorporating rWGS in 5 California children's hospitals demonstrates improved clinical outcomes and reduced costs of care. *Am J Hum Genet* 2021;**108**:1231–8.
152. Klein EA, Richards D, Cohn A *et al.* Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. *Ann Oncol* 2021;**32**:1167–77.
153. 100,000Genomes Project Pilot Investigators. 100,000 genomes pilot on rare-disease diagnosis in health care—Preliminary report. *N Engl J Med* 2021;**385**:1868–80.
154. Zuojun X, Ren F, Wang P *et al.* A generative AI-discovered TNIK inhibitor for idiopathic pulmonary fibrosis: a randomized phase 2a trial. *Nat Med* 2025;1–9.
155. Paszke A, Gross S, Massa F *et al.* Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems* 2019;32.
156. Gayoso A, Lopez R, Xing G *et al.* A python library for probabilistic analysis of single-cell omics data. *Nat Biotechnol* 2022;**40**:163–6.
157. Stuart T, Butler A, Hoffman P *et al.* Comprehensive integration of single-cell data. *Cell* 2019;**177**:1888–902.
158. Dries R, Zhu Q, Dong R *et al.* Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol* 2021;**22**:1–31.
159. Weinstein JN, Collisson EA, Mills GB *et al.* The cancer genome atlas pan-cancer analysis project. *Nat Genet* 2013;**45**:1113–20.
160. Zeng Z, Mao C, Vo A *et al.* Deep learning for cancer type classification. *BioRxiv* 2019;612762.
161. Barrett T, Wilhite SE, Ledoux P *et al.* NCBI HEO: archive for functional genomics data sets—Update. *Nucleic Acids Res* 2012;**41**:D991–5.
162. Costa-Silva J, Domingues D, Lopes FM. RNA-seq differential expression analysis: an extended review and a software tool. *PLoS One* 2017;**12**:e0190152.
163. GTEx Consortium et al. Genetic effects on gene expression across human tissues. *Nature* 2017;**550**:204–13.
164. Lappalainen T, Sammeth M, Friedländer MR *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 2013;**501**:506–11.
165. Gutierrez-Arcelus M, Ongen H, Lappalainen T *et al.* Tissue-specific effects of genetic and epigenetic variation on gene regulation and splicing. *PLoS Genet* 2015;**11**:e1004958.
166. Regev A, Teichmann SA, Lander ES *et al.* The human cell atlas. *Elife* 2017;**6**:e27041.
167. Karlsson M, Zhang C, Méar L *et al.* A single–cell type transcriptomics map of human tissues. *Sci Adv* 2021;**7**:eabh2169.
168. Cao J, Diana R *et al.* A human cell atlas of fetal gene expression. *Science* 2020;**370**: eaba7721.
169. Varadi M, Anyango S, Deshpande M *et al.* AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* 2022;**50**:D439–44.
170. Honcharuk V, Zainab A, Horimoto Y *et al.* DeepSpaceDB: a spatial transcriptomics atlas for interactive in-depth analysis of tissues and tissue microenvironments. *Nucleic Acids Research* 2026;**54**:D1017–30. https://doi.org/10.1093/nar/gkaf1117
171. Yuan Z, Zhao F, Senlin Lin Y *et al.* Benchmarking spatial clustering methods with spatially resolved transcriptomics data. *Nat Methods* 2024;**21**:712–22.
172. Jaume G, Doucet P, Song A *et al.* HEST-1k: a dataset for spatial transcriptomics and histology image analysis. *Adv Neural Inf Proces Syst* 2024;**37**:53798–53,833.
173. Zahedi R, Ghamsari R, Argha A *et al.* Deep learning in spatially resolved transcriptomics: a comprehensive technical view. *Brief Bioinform* 2024;**25**:bbae082.
174. 10x Genomics. Xenium In Situ: Spatially Resolved Transcriptomics. https://www.10xgenomics.com/products/xenium, 2024. (29 June 2025, date last accessed).
175. Cassella L, Ephrussi A. Subcellular spatial transcriptomics identifies three mechanistically different classes of localizing RNAs. *Nat Commun* 2022;**13**:6355.
176. Crosetto N, Bienko M, Van Oudenaarden A. Spatially resolved transcriptomics and beyond. *Nat Rev Genet* 2015;**16**:57–66.
177. 10x Genomics. Visium spatial gene expression. https://www.10xgenomics.com/products/spatial-gene-expression. 2024. Accessed: 2025-06-29
178. Cable DM, Murray E, Zou LS *et al.* Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat Biotechnol* 2022;**40**:517–26.
179. Wang Y, Liu B, Zhao G *et al.* Spatial transcriptomics: technologies, applications and experimental considerations. *Genomics* 2023;**115**:110671.
180. Liao W-W, Asri M, Ebler J *et al.* A draft human pangenome reference. *Nature* 2023;**617**:312–24.
181. Stolovitzky G, Monroe DON, Califano A. Dialogue on reverse-engineering assessment and methods: the dream of high-throughput pathway inference. *Ann N Y Acad Sci* 2007;**1115**:1–22.

182. Prill RJ, Marbach D, Saez-Rodriguez J *et al.* Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PLoS One* 2010;**5**:e9202.
183. Marbach D, Costello JC, Küffner R *et al.* Wisdom of crowds for robust gene network inference. *Nat Methods* 2012;**9**:796–804.
184. ENCODE Project Consortium*et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012; **489**:57.
185. Maurano MT, Humbert R, Rynes E *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* 2012;**337**:1190–5.
186. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* 2012;**9**:215–6.
187. Liu Z, Li J, Li S *et al.* GenBench: a benchmarking suite for systematic evaluation of genomic foundation models. *arXiv preprint arXiv:2406.01627*, 2024.
188. Walker AR, Datta S. Identification of city specific important bacterial signature for the MetaSUB CAMDA challenge microbiome data. *Biol Direct* 2019;**14**:1–16.
189. Zook JM, Catoe D, McDaniel J *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data* 2016;**3**:160025.
190. Notin P, Kollasch AW, Ritter D *et al.* ProteinGYM: large-scale benchmarks for protein design and fitness prediction. *Adv Neural Inf Proces Syst* 2023;**36**.
191. Rao R, Bhattacharya N, Thomas N *et al.* Evaluating protein transfer learning with tape. In: *Adv Neural Inf Proces Syst* 2019;**32**: 9681–91.
192. Luecken MD, Büttner M, Chaichoompu K *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* 2022;**19**:41–50.
193. Robson ES, Ioannidis NM, Guanine v1.0: benchmark datasets for genomic AI sequence-to-function models. *Proc Mach Learn Res* 2024;**240**:250–66.
194. Zhong J, Li L, Dannenfelser R *et al.* Benchmarking gene embeddings from sequence, expression, network, and text models for functional prediction tasks. *bioRxiv* 2025.
195. Butler A, Hoffman P, Smibert P *et al.* Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 2018;**36**:411–20.
196. Marin FI, Teufel F, Horlacher M *et al.* BEND: Benchmarking DNA language models on biologically meaningful tasks. In: *International Conference on Learning Representations (ICLR)*. 2024.
197. CAMDA Consortium. Critical assessment of massive data analysis. 2024. https://camda.info/ (29 June 2025, date last accessed).
198. Leek JT, Scharpf RB, Bravo HC *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* 2010;**11**:733–9.
199. Galla EP, Boddapati VN, Patra GK *et al.* AI-powered insights: leveraging machine learning and big data for advanced genomic research in healthcare. *Educational Administration: Theory and Practice 2023* 2023;**29**.
200. Tonekaboni S, Joshi S, McCradden MD *et al.* What clinicians want: contextualizing explainable machine learning for clinical end use. In: *Machine Learning for Healthcare Conference*, pp. 359–80. Brookline, MA, USA: PMLR, 2019.
201. Gündüz HA, Giri S, Binder M *et al.* Uncertainty quantification for deep learning models predicting the regulatory activity of DNA sequences. In:*2023 International Conference on Machine Learning and Applications (ICMLA)*, pp. 566–73. Piscataway, NJ, USA: IEEE, 2023.
202. Reddy S. Foundation for artificial intelligence-driven democratization of healthcare access. *SSRN* 2025;**5277581**.
203. Zhuang B, Liu J, Pan Z *et al.* A survey on efficient training of transformers. *arXiv preprint arXiv:2302.01107*, 2023.
204. Strubell E, Ganesh A, McCallum A. Energy and policy considerations for deep learning in NLP. In: *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 3645–50. Florence, Italy: Association for Computational Linguistics, 2019.
205. Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs. *Adv Neural Inf Proces Syst* 2017;**30**:1024–34.
206. Hu EJ, Shen Y, Wallis P *et al.* LoRA: low-rank adaptation of large language models. *ICLR* 2022;**1**:3.
207. Dettmers T, Pagnoni A, Holtzman A *et al.* QLoRA: efficient finetuning of quantized LLMs. *Adv Neural Inf Proces Syst* 2023;**36**: 10088–10,115.
208. Holzinger A, Langs G, Denk H *et al.* Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip Rev Data Min Knowl Discov* 2019;**9**:e1312.
209. Vishniakov K, Viswanathan K, Medvedev A *et al.* Genomic foundationless models: pretraining does not promise performance. *bioRxiv* 2024, 2024–12.
210. Kirk R, Zhang A, Grefenstette E *et al.* A survey of zero-shot generalisation in deep reinforcement learning. *J Artif Intell Res* 2023;**76**:201–64.
211. Chen RJ, Lu MY, Wang J *et al.* Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Trans Med Imaging* 2020; **41**:757–70.
212. Cellina M, Cè M, Alì M *et al.* Digital twins: the new frontier for personalized medicine? *Appl Sci* 2023;**13**:7940.
213. Shen S, Qi W, Liu X *et al.* From virtual to reality: innovative practices of digital twins in tumor therapy. *J Transl Med* 2025;**23**:348.
214. Corral-Acero J, Margara F, Marciniak M *et al.* The 'digital twin' to enable the vision of precision cardiology. *Eur Heart J* 2020;**41**:4556–64.

Briefings in Bioinformatics, 2026, **27**, bbag229
https://doi.org/10.1093/bib/bbag229
Review