



Review

Computational frameworks for modern genomics: From sparse DNA sequences to epigenomic landscapes

Heikham Russiachand Singh ^{1,*}¹ Department of Plant Science, McGill University, Raymond Building, 21111, Lakeshore Road, Ste. Anne de Bellevue, Quebec, Canada.* **Correspondence:** heikham.singh@mail.mcgill.ca (H.R.S.)

Citation: Singh, H.R. Computational frameworks for modern genomics: From sparse DNA sequences to epigenomic landscapes. *Glob. Jour. Bas. Sci.* 2025, 1(11). 1-6.

Received: July 03, 2025

Revised: September 21, 2025

Accepted: September 23, 2025

Published: September 28, 2025

doi: 10.63454/ijbs20000059

ISSN: 3049-3315

Abstract: The rapid evolution of high-throughput sequencing technologies has transformed genomics into a data-intensive discipline, generating massive volumes of heterogeneous biological data. Computational genomics has emerged as an indispensable field that integrates algorithms, statistics, and machine learning to extract biological meaning from sparse DNA sequences and complex epigenomic landscapes. This review provides a comprehensive overview of the computational frameworks that underpin modern genomics, spanning sequence analysis, genome assembly, variant detection, transcriptomics, epigenomics, and integrative multi-omics approaches. We discuss classical bioinformatics pipelines alongside emerging artificial intelligence-driven models, highlighting their applications in functional genomics, disease biology, and precision medicine. Furthermore, we examine key challenges related to data sparsity, scalability, interpretability, and reproducibility, and outline future directions that will shape the next generation of genomic research.

Keywords: Computational genomics; DNA sequencing; epigenomics; machine learning; multi-omics; bioinformatics

1. Introduction

The completion of the Human Genome Project in 2003 marked a foundational milestone in biological science, ushering in an era where genomic data became central to understanding life processes and disease mechanisms [1]. Since then, advances in next-generation sequencing (NGS) and third-generation sequencing technologies, such as single-molecule real-time (SMRT) and nanopore sequencing, have dramatically reduced sequencing costs while exponentially increasing throughput [2]. This technological democratization has enabled large-scale, population-level projects such as ENCODE (Encyclopedia of DNA Elements), GTEx (Genotype-Tissue Expression), The Cancer Genome Atlas (TCGA), and the 1000 Genomes Project, generating petabyte-scale datasets that map genomic variation, gene regulation, and disease associations across human populations [3,4]. However, the sheer scale, complexity, and multidimensional nature of these datasets have necessitated the development of robust, scalable, and sophisticated computational frameworks capable of managing, analyzing, and interpreting genomic information [5].

Computational genomics lies at the dynamic intersection of biology, computer science, mathematics, and statistics. It encompasses a broad spectrum of methods designed to analyze DNA sequences, RNA transcripts, chromatin states, and epigenetic modifications [6]. Early computational approaches, developed in the late 20th and early 21st centuries, focused primarily on fundamental tasks like sequence alignment, genome assembly, and gene prediction. In contrast, modern computational frameworks address higher-order biological questions, including the deciphering of complex regulatory networks, dynamic epigenomic landscapes, three-dimensional genome architecture, and integrative multi-omics analyses [7]. This evolution reflects the transition from a static, linear view of the genome to a dynamic, systems-level understanding of its function. This review aims to synthesize current computational strategies used across the field of genomics, from processing sparse sequence-level data to modeling multidimensional epigenomic profiles. We emphasize key methodological advances, persistent challenges, and their transformative biological applications, particularly in human health and disease.

2. DNA sequencing technologies and data characteristics

Modern high-throughput sequencing platforms, primarily Illumina (short-read), PacBio (HiFi long-read), and Oxford Nanopore Technologies (ultra-long-read), generate vast quantities of sequence data with distinct technical profiles and associated computational challenges [8,9]. Illumina sequencing offers high accuracy (>99.9%) and massive parallel throughput but produces short reads (50-300 bp), which limits the ability to resolve complex genomic regions such as repetitive elements, segmental duplications, and structural variants [10]. In contrast, PacBio and Oxford Nanopore generate reads spanning thousands to millions of bases, providing superior contiguity for genome assembly and direct detection of structural variants and base modifications, albeit with higher raw error rates (5-15%) that require specialized computational error correction and consensus algorithms [11,12].

From a computational perspective, genomic data are inherently sparse, noisy, and high-dimensional. Key statistical properties include coverage variability (uneven read distribution across the genome), sequencing bias (e.g., GC-content bias, fragmentation bias), and platform-specific technical artifacts (e.g., homopolymer errors in nanopore reads) [13]. Furthermore, data from multi-omic

assays—such as paired DNA and RNA sequencing from the same sample—introduce layers of correlation and noise that must be jointly modeled. Effective computational frameworks must therefore incorporate rigorous quality control (e.g., FastQC), sophisticated normalization (e.g., GC correction, depth normalization), and statistical error correction steps (e.g., using k-mer spectra or machine learning models) to ensure reliable biological inference [14]. Understanding these fundamental data characteristics is critical for designing robust downstream algorithms for core tasks like genome assembly, sequence alignment, and variant detection.

3. Genome assembly and sequence alignment algorithms

De novo genome assembly, the reconstruction of complete genomic sequences from millions of fragmented sequencing reads, represents one of the most computationally intensive challenges in bioinformatics (Figure 1). The two dominant algorithmic paradigms are the overlap-layout-consensus (OLC) approach, historically used for Sanger and long-read data, and the de Bruijn graph (DBG) approach, which is highly efficient for short, high-coverage Illumina reads [15]. Widely used short-read assemblers such as Velvet, SOAPdenovo, and SPAdes construct and simplify de Bruijn graphs by breaking reads into k-mers (substrings of length k) to manage the immense data volume and resolve repeats [16]. For long-read data, assemblers like Canu, Flye, and wtdbg2 employ OLC or string graph methods, often incorporating iterative error correction and consensus polishing steps (using tools like Racon or Medaka) to achieve high-quality, contiguous assemblies [17].

Sequence alignment—the mapping of sequencing reads to a reference genome—remains a cornerstone of virtually all reference-based genomic analyses, enabling applications from variant calling to transcript quantification. The Burrows-Wheeler Transform (BWT), which enables highly memory-efficient indexing of large genomes, underpins fast and accurate aligners like BWA and Bowtie2 for DNA sequencing [18]. For RNA sequencing data, spliced aligners such as STAR and HISAT2 are essential, as they must detect exon-exon junctions by splitting reads across introns, using reference transcriptome annotations or *de novo* junction discovery [19]. Recent algorithmic developments focus on improving scalability and accuracy through techniques like adaptive seed-and-extend, minimum exact matching, and hardware acceleration (e.g., GPU-based alignment), which are critical for processing large-scale population cohorts and single-cell datasets [20].

4. Variant detection and functional annotation

Identifying genetic variation—including single-nucleotide variants (SNVs), small insertions/deletions (indels), copy number variants (CNVs), and complex structural variants (SVs)—is critical for understanding population genetics, evolutionary biology, and disease susceptibility. Computational variant calling requires sophisticated statistical models that distinguish true biological variants from sequencing errors and alignment artifacts. Widely adopted pipelines, such as the GATK Best Practices workflow, employ a multi-step process including duplicate marking, base quality score recalibration, and haplotype-aware variant calling using Bayesian or hidden Markov models [21]. Alternative tools like FreeBayes (a Bayesian haplotype-based caller) and SAMtools mpileup provide complementary approaches, while specialized callers like Delly, Manta, and Sniffles are designed for detecting structural variants from paired-end, split-read, or long-read data [22,23].

The functional interpretation of millions of discovered variants represents a subsequent, major computational challenge. Annotation frameworks such as ANNOVAR, the Ensembl Variant Effect Predictor (VEP), and SnpEff integrate genomic coordinates with a wealth of biological context [24]. They annotate variants based on their location relative to genes (e.g., exonic, intronic, intergenic), predicted impact on protein sequence (e.g., missense, nonsense), evolutionary conservation scores (e.g., PhyloP, GERP++), population allele frequencies (e.g., from gnomAD), and overlap with regulatory elements from ENCODE (e.g., promoters, enhancers) [25]. To prioritize variants with potential pathogenic relevance, machine learning-based predictors like CADD (Combined Annotation Dependent Depletion), REVEL, and AlphaMissense have been developed. These tools aggregate diverse annotation signals to model complex genotype-phenotype relationships, significantly aiding in the identification of disease-causing mutations in both monogenic and complex disorders [26].

5. Computational transcriptomics

Transcriptomics, powered by RNA sequencing (RNA-seq), provides a dynamic snapshot of gene expression and RNA processing. Core computational tasks include transcriptome reconstruction, quantification of transcript abundance, and detection of differential expression (DE) or alternative splicing across experimental conditions. Quantification methods are broadly divided into alignment-based and alignment-free (pseudoalignment) strategies. Tools like HTSeq and featureCounts count reads aligned to genomic features, whereas Kallisto and Salmon use lightweight pseudoalignment to rapidly estimate transcript abundances by matching reads to a pre-built k-mer index of the transcriptome, offering substantial gains in speed without sacrificing accuracy [27,28].

The advent of single-cell RNA sequencing (scRNA-seq) has introduced a new layer of computational complexity. scRNA-seq data are characterized by extreme sparsity due to "dropout" events (where transcripts are not detected), significant technical batch

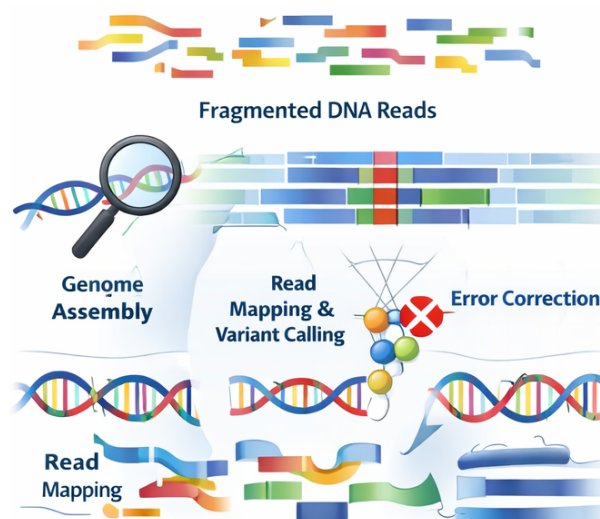


Figure 1. DNA sequence analysis.

effects, and the need to dissect cellular heterogeneity [29]. A standard analytical pipeline involves quality control (filtering low-quality cells), normalization (e.g., using SCTransform or scran), dimensionality reduction (Principal Component Analysis, t-distributed Stochastic Neighbor Embedding - t-SNE, Uniform Manifold Approximation and Projection - UMAP), and clustering (e.g., with Leiden or Louvain algorithms) to identify cell types or states [30]. Beyond static classification, trajectory inference algorithms like Monocle3, PAGA, and Slingshot model dynamic processes such as differentiation, providing pseudotemporal ordering of cells along developmental or transitional pathways [31]. These frameworks have revolutionized our understanding of tissue organization, tumor microenvironments, and cellular responses.

6. Epigenomics and chromatin landscape modeling

Epigenomic profiling techniques capture regulatory information beyond the primary DNA sequence, providing insights into the functional state of the genome. Chromatin immunoprecipitation sequencing (ChIP-seq) maps protein-DNA interactions (e.g., transcription factors, histone modifications); Assay for Transposase-Accessible Chromatin sequencing (ATAC-seq) identifies open chromatin regions; bisulfite sequencing (BS-seq) quantifies DNA methylation; and Chromosome Conformation Capture (Hi-C) reveals three-dimensional genome architecture [32].

The computational analysis (Figure 2) of each assay involves specialized steps. For ChIP-seq and ATAC-seq, peak calling algorithms like MACS2 and HOMER use statistical models to identify regions of significant enrichment over background noise [33]. For DNA methylation, pipelines such as Bismark or MethylDackel align bisulfite-treated reads and calculate methylation proportions at individual cytosine sites. Hi-C data analysis requires correcting for technical biases (e.g., using HiC-Pro or Juicer tools) before constructing contact matrices and identifying topologically associating domains (TADs) and chromatin loops [34]. Integrating these disparate datasets to model the complete epigenomic landscape is a major focus. Computational frameworks employ multivariate approaches, including hidden Markov models (e.g., ChromHMM, Segway) to segment the genome into discrete chromatin states (e.g., active promoters, repressed heterochromatin) based on combinatorial epigenetic marks [35]. More recently, deep learning architectures like DeepSEA and Basenji have been trained to predict transcription factor binding and chromatin accessibility directly from DNA sequence, revealing the complex cis-regulatory code [36].

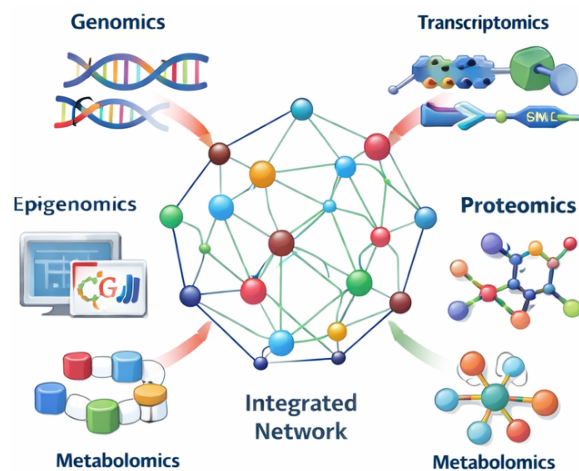


Figure 2. Integrative genomics analysis.

7. Machine learning and artificial intelligence in genomics

Machine learning (ML) and artificial intelligence (AI) have become indispensable for extracting meaningful patterns from the high-dimensional, non-linear data that define modern genomics. Supervised learning models, including support vector machines (SVMs) and random forests, are routinely used for classification tasks such as distinguishing functional from non-functional genomic elements or predicting pathogenic variants [37]. Unsupervised methods like clustering and matrix factorization are key for discovering novel subtypes of cancer from transcriptomic data or identifying latent factors in population genetics.

Deep learning, a subset of AI based on artificial neural networks, has achieved state-of-the-art performance in numerous genomic prediction tasks. Convolutional neural networks (CNNs) excel at learning spatial patterns from sequence data for tasks like predicting transcription factor binding sites and splice sites [38]. Recurrent neural networks (RNNs) and their advanced variants (e.g., Long Short-Term Memory networks - LSTMs) model sequential dependencies, making them suitable for analyzing regulatory grammar. Large language models (LLMs), pre-trained on vast corpora of biological sequences, are emerging as powerful tools for generative and predictive tasks in genomics [39]. However, these powerful "black-box" models face significant challenges regarding interpretability (understanding why a prediction was made), data bias (models trained on under-represented populations may not generalize), and clinical applicability. The burgeoning field of explainable AI (XAI) aims to address these issues by developing methods to attribute predictions to specific input features, thereby enhancing trust and facilitating the translation of AI discoveries into biological insights and clinical tools [40].

8. Integrative multi-omics frameworks

Biological systems are defined by the complex interplay of molecular layers—genomics, transcriptomics, epigenomics, proteomics, and metabolomics. A siloed analysis (Figure 2) of any single layer provides an incomplete picture. Integrative multi-omics computational frameworks are therefore essential for a holistic understanding of cellular states and disease mechanisms [41]. These frameworks must address the "curse of dimensionality," heterogeneous data types, and missing values across modalities.

Common computational strategies for integration include:

- **Matrix Factorization:** Methods like Joint Non-negative Matrix Factorization (jNMF) and Multi-Omics Factor Analysis (MOFA) decompose multi-omics datasets into a set of shared latent factors that capture coordinated biological variation across assays [42].

- **Network-Based Approaches:** Biological knowledge graphs or co-expression networks integrate different data types as nodes and edges, allowing for the identification of key regulatory hubs and pathways. Tools like Cytoscape with its plugins facilitate such network visualization and analysis [43].
- **Graph Neural Networks (GNNs):** These advanced ML models operate directly on graph structures, making them ideal for integrating multi-omics data where relationships between molecules (e.g., protein-protein interactions) can be explicitly encoded [44].
- **Multi-view Learning:** This class of algorithms learns a unified representation from multiple "views" (omics datasets) of the same biological sample, improving clustering, classification, and outcome prediction [45].

Integrative genomics has proven particularly transformative in oncology. By combining genomic alterations, transcriptomic subtypes, epigenomic silencing, and proteomic signaling, researchers can identify master regulators of tumorigenesis, stratify patients into molecular subtypes with prognostic and therapeutic implications, and uncover mechanisms of drug resistance [46]. As multi-omic profiling becomes more routine in biomedical research, the development of scalable, user-friendly, and statistically rigorous integrative frameworks will be paramount for unlocking the full potential of systems biology.

9. Challenges and limitations

Despite the transformative progress in computational genomics, the field grapples with persistent and emerging challenges that complicate analysis, interpretation, and translation. A primary obstacle is data heterogeneity and integration. Multi-omic datasets derived from different platforms (e.g., Illumina vs. Nanopore), protocols, and laboratories exhibit profound technical variability in the form of batch effects, differing noise profiles, and non-biological correlations. Integrating sparse single-cell data with bulk sequencing data, or combining discrete epigenetic marks into a coherent regulatory model, requires sophisticated normalization and batch-correction algorithms (e.g., ComBat, Harmony) that can inadvertently remove subtle biological signals if applied incorrectly [47-48]. Furthermore, missing data is a ubiquitous issue, particularly in multi-omics studies where not all assays are performed on every sample, creating incomplete matrices that challenge conventional statistical and machine learning models.

Scalability and computational resource demands represent another critical bottleneck. The exponential growth of sequencing data, exemplified by initiatives like the UK Biobank and All of Us Research Program, generates datasets on the exabyte scale. Processing, storing, and analyzing this data demands immense memory, processing power, and efficient I/O operations. While cloud computing (e.g., AWS, Google Cloud Platform) and high-performance computing clusters offer solutions, they introduce challenges of data transfer costs, software portability, and the need for parallelized, memory-efficient algorithms [49]. Many widely used tools were not designed for this scale, necessitating continual re-engineering.

A foundational crisis in the field is the lack of standardization and reproducibility. Subtle variations in bioinformatics pipelines—such as choice of aligner (BWA vs. Bowtie), quality filtering thresholds, or reference genome build—can lead to significantly different variant calls or differential expression results for the same raw data [50]. This "analysis archaeology" undermines the comparability of studies and hinders meta-analyses. The movement towards containerization (Docker, Singularity) and workflow management systems (Nextflow, Snakemake, WDL) aims to encapsulate complete computational environments, promoting reproducibility and portability across systems [51].

Finally, the rapid pace of genomic discovery has outstripped the development of robust ethical and governance frameworks. Key concerns include data privacy and re-identification risk, even from anonymized genomic data; informed consent for future, unspecified research uses; and the equitable access to the benefits of genomic medicine, which risks exacerbating health disparities if diverse populations are underrepresented in reference databases [52,53]. Addressing these challenges requires not just technical solutions, but also ongoing dialogue among bioinformaticians, clinicians, ethicists, and policy makers to establish responsible and equitable computational practices.

10. Future Perspectives

The future trajectory of computational genomics will be defined by the convergence of advanced algorithms, scalable infrastructure, and collaborative, open science. A major focus will be the development of interpretable and explainable AI models. Moving beyond "black-box" deep learning, future frameworks will integrate biological priors—such as known pathway relationships or 3D chromatin constraints—directly into model architectures, leading to more generalizable predictions and actionable biological hypotheses [54]. The rise of foundation models pre-trained on massive, diverse genomic datasets promises a shift from task-specific training to flexible, context-aware inference for a wide array of downstream applications, from variant effect prediction to *de novo* protein design [55].

Technologically, the field will increasingly leverage cloud-native and federated analytics. Cloud platforms enable elastic scaling and democratize access to state-of-the-art pipelines, while federated learning approaches allow models to be trained on distributed datasets across institutions without sharing raw, sensitive genomic data, directly addressing privacy concerns [55-56]. Furthermore, as spatial omics (e.g., 10x Genomics Visium, MERFISH) and live-cell imaging technologies mature, computational frameworks must evolve to capture the temporal and spatial dimensions of genome regulation. This will require novel algorithms from computer vision and spatial statistics to model gene expression and chromatin dynamics within the native tissue architecture, providing unprecedented views of cellular ecosystems in development and disease [57].

Ultimately, the integration of these computational innovations with deep biological insight will be essential. The next generation of computational genomics will not merely process data but will actively generate testable mechanistic models of biological systems.

This synergy will be crucial for realizing the promise of precision medicine, enabling the move from population-level associations to patient-specific dynamic models that can predict disease trajectories and optimize therapeutic interventions, thereby fully unlocking genomics for scientific discovery and human health [58].

11. Conclusion

Computational frameworks are the indispensable engine of modern genomics, enabling the transformation of raw, massive sequencing data into coherent biological knowledge and clinical insight. From the foundational tasks of assembling sparse DNA sequences and mapping genetic variation to the integrative modeling of complex epigenomic landscapes and single-cell atlases, these methods have radically expanded our understanding of genome structure, function, and regulation. They have illuminated the molecular underpinnings of development, evolution, and disease, transitioning genomics from a descriptive to a predictive and mechanistic science. The journey ahead requires sustained interdisciplinary collaboration among biologists, computer scientists, statisticians, and clinicians. Continued methodological innovation to tackle the challenges of scale, integration, and interpretation, coupled with a steadfast commitment to ethical, reproducible, and equitable practices, will be paramount. By harnessing these principles, the field of computational genomics is poised to continue its pivotal role in driving scientific discovery and translating the promise of the genome into tangible advances in biology and medicine.

Author Contributions: Conceptualisation, H.R.S.; software, H.R.S.; investigation, H.R.S.; writing—original draft preparation, H.R.S.; writing—review and editing, H.R.S.; visualisation, H.R.S.; supervision, H.R.S.; project administration, H.R.S. The author has read and agreed to the published version of the manuscript.

Funding: Not applicable.

Acknowledgments: We are grateful to the Department of Plant Science, McGill University, Raymond Building, 21111, Lakeshore Road, Ste. Anne de Bellevue, Quebec, Canada for providing us all the facilities to carry out the entire work.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All the related data are supplied in this work or have been referenced properly.

References

1. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001;409(6822):860-921.
2. Shendure J, Balasubramanian S, Church GM, et al. DNA sequencing at 40: past, present and future. *Nature*. 2017;550(7676):345-353.
3. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57-74.
4. The Cancer Genome Atlas Research Network. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*. 2013;45(10):1113-20.
5. Stephens ZD, Lee SY, Faghri F, et al. Big Data: Astronomical or Genomical? *PLoS Biol*. 2015;13(7):e1002195.
6. Pevsner J. *Bioinformatics and functional genomics*. 3rd ed. Hoboken, NJ: Wiley-Blackwell; 2015.
7. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet*. 2015;16(6):321-32.
8. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*. 2016;17(6):333-51.
9. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. *Nat Rev Genet*. 2020;21(10):597-614.
10. Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet*. 2010;11(1):31-46.
11. Eid J, Fehr A, Gray J, et al. Real-time DNA sequencing from single polymerase molecules. *Science*. 2009;323(5910):133-8.
12. Jain M, Koren S, Miga KH, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol*. 2018;36(4):338-345.
13. Sims D, Sudbery I, Iltott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet*. 2014;15(2):121-32.
14. Conesa A, Madrigal P, Tarazona S, et al. A survey of best practices for RNA-seq data analysis. *Genome Biol*. 2016;17:13.
15. Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. *Genomics*. 2010;95(6):315-27.
16. Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. 2012;19(5):455-77.
17. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol*. 2019;37(5):540-546.
18. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-60.
19. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21.
20. Alser M, Rotman J, Deshpande D, et al. Technology dictates algorithms: recent developments in read alignment. *Genome Biol*. 2021;22(1):249.
21. Van der Auwera GA, O'Connor BD. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. Sebastopol, CA: O'Reilly Media; 2020.
22. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. [arXiv:1207.3907 \[q-bio.GN\]](https://arxiv.org/abs/1207.3907). 2012.
23. Rausch T, Zichner T, Schlattl A, et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*. 2012;28(18):i333-i339.
24. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164.
25. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17(1):122.
26. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res*. 2019;47(D1):D886-D894.
27. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*. 2016;34(5):525-7.
28. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods*. 2017;14(4):417-419.
29. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med*. 2017;9(1):75.
30. Stuart T, Butler A, Hoffman P, et al. Comprehensive Integration of Single-Cell Data. *Cell*. 2019;177(7):1888-1902.e21.

31. Trapnell C, Cacchiarelli D, Grimsby J, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol.* 2014;32(4):381-386.
32. Kundaje A, Meuleman W, Ernst J, et al. Integrative analysis of 111 reference human epigenomes. *Nature.* 2015;518(7539):317-30.
33. Zhang Y, Liu T, Meyer CA, et al. Model-based Hi-C analysis of ChIP-Seq (MACS). *Genome Biol.* 2008;9(9):R137.
34. Servant N, Varoquaux N, Lajoie BR, et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* 2015;16:259.
35. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods.* 2012;9(3):215-6.
36. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods.* 2015;12(10):931-4.
37. Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. *Mol Syst Biol.* 2016;12(7):878.
38. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol.* 2015;33(8):831-8.
39. Rives A, Meier J, Sercu T, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A.* 2021;118(15):e2016239118.
40. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip Rev Data Min Knowl Discov.* 2019;9(4):e1312.
41. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome Biol.* 2017;18(1):83.
42. Argelaguet R, Velten B, Arnol D, et al. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol.* 2018;14(6):e8124.
43. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003;13(11):2498-504.
44. 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.
45. Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res.* 2018;46(20):10546-10562.
46. Hoadley KA, Yau C, Hinoue T, et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell.* 2018;173(2):291-304.e6.
47. 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(10):733-9.
48. Korsunsky I, Millard N, Fan J, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods.* 2019;16(12):1289-1296.
49. Schatz MC, Philipp AM. The rise of a digital immune system. *Gigascience.* 2021;10(5):giab049.
50. Patro R, Mount SM, Kingsford C, Salzberg SL. The devil in the details of RNA-seq. *Nat Biotechnol.* 2014;32(9):882-4.
51. Di Tommaso P, Chatzou M, Floden EW, et al. Nextflow enables reproducible computational workflows. *Nat Biotechnol.* 2017;35(4):316-319.
52. Gymrek M, McGuire AL, Golan D, et al. Identifying personal genomes by surname inference. *Science.* 2013;339(6117):321-4.
53. Popejoy AB, Fullerton SM. Genomics is failing on diversity. *Nature.* 2016;538(7624):161-164.
54. Avsec Ž, Weilert M, Shrikumar A, et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat Genet.* 2021;53(3):354-366.
55. Rives A, Meier J, Sercu T, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A.* 2021;118(15):e2016239118.
56. Warnat-Herresthal S, Schultze H, Shastry KL, et al. Swarm Learning for decentralized and confidential clinical machine learning. *Nature.* 2021;594(7862):265-270.
57. Moses L, Pachter L. Museum of spatial transcriptomics. *Nat Methods.* 2022;19(5):534-546.
58. Schork NJ. Artificial Intelligence and Personalized Medicine. *Cancer Treat Res.* 2019;178:265-283.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of Global Journal of Basic Science and/or the editor(s). Global Journal of Basic Science and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Copyright: © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).