



doi:10.5281/zenodo.19500033

The Crucial Role Of Bioinformatics Tools In Deconstructing The Human Genome

¹Usman Y.M, ¹Nyango P.B, ²Yilgwan G, ²Tsoho F, ³Henry R, ⁴Galam N.Z, ⁵Shugaba A.I

¹Department of Human Anatomy, Faculty of Basic Medical Sciences, University of Jos

²Department of Human Physiology, Faculty of Basic Medical Sciences, University of Jos

³Department of Human Anatomy, Faculty of Basic Medical Sciences, Federal University Wukari

⁴Department of Human Physiology, Faculty of Basic Medical Sciences, Federal University Wukari

⁵Department of Human Anatomy, Faculty of Basic Medical Sciences, Federal University of Lafia

Corresponding Author: Dr Yohanna Musa Usman, Email: usmany@unijos.edu.ng, Telephone: +2348064817431

ABSTRACT

Bioinformatics is an interdisciplinary field that combines biology, computer science, and statistics to develop computational frameworks and analytical tools for managing and interpreting genomic data. This review surveys literature published between January 2000 and April 2026, covering the period after the completion of the Human Genome Project. The selected literatures were synthesised narratively, with key concepts, methodologies, and findings organised according to the main themes of the review. Bioinformatics tools enable the analysis of genetic variation within human populations, providing a comprehensive understanding of human history and diversity. These tools are fundamental in personalised medicine, supporting the interpretation of genomic profiles and the translation of this information into clinically actionable insights. They facilitate the tailoring of treatments to individual genetic profiles, which improves efficacy and reduces adverse effects. Furthermore, bioinformatics tools predict drug responses based on genetic variants, informing drug selection and dosing. They are also essential for diagnosing genetic disorders, particularly rare diseases caused by single-gene mutations. As genomic data continues to expand, emerging areas such as single-cell genomics, epigenomics, and multi-omics integration introduce new challenges and opportunities. Advances in artificial intelligence, particularly deep learning, are anticipated to significantly transform genomics and bioinformatics.

Keywords: bioinformatics, human genomics, genomic data analysis, computational biology, DNA sequencing

INTRODUCTION

The human genome serves as a complex blueprint encoded within approximately 3 billion base pairs of DNA (Collins and Fink, 1995). This nucleotide sequence determines biological characteristics, hereditary traits, and disease susceptibility. The completion of the Human Genome Project in 2003 marked a pivotal milestone in genomic research and underscored the considerable challenge of interpreting extensive genetic information (Lu et al., 2014). The advent of high-throughput sequencing technologies has facilitated unprecedented access to genomic data (Satam et al., 2023). Next-generation sequencing

methods have significantly reduced both the cost and time required to sequence entire genomes, resulting in a rapid expansion of available genomic information (Akintunde et al., 2023). This vast quantity of data would remain unmanageable and largely uninterpretable without the advanced analytical tools and methodologies provided by bioinformatics (McCue and McCoy, 2017).

Bioinformatics, an interdisciplinary field integrating biology, computer science, and statistics, is essential for understanding the complexities of human genomics (Bayat, 2002). It provides computational frameworks and analytical tools necessary to store, organise, analyse, and interpret the extensive datasets generated by genomic research (Zhang et al., 2025). Bioinformatics plays a critical role at every stage of genomic investigation, from managing large-scale sequencing data to identifying pathogenic mutations (Pereira et al., 2020). The integration of genomics and bioinformatics continues to drive significant scientific advancements (Baev, 2025). This review examines the diverse functions of bioinformatics in human genomics, with particular emphasis on its roles in data management, sequence analysis, comparative genomics, and clinical applications. By addressing these areas, the review demonstrates how bioinformatics has transformed the study of the human genome and its broader implications for medicine, evolutionary biology, and human health.

METHODOLOGY FOR REVIEW

A narrative approach was employed to facilitate a comprehensive and critical analysis of the literature, unconstrained by the limitations of systematic review methodologies. This method enabled the inclusion of diverse study types and supported the synthesis of multidisciplinary perspectives, thereby capturing the breadth and complexity of bioinformatics applications in human genomics. The literature search incorporated PubMed/MEDLINE, Web of Science, Scopus, IEEE Xplore Digital Library, and ACM Digital Library to ensure comprehensive coverage of biomedical, computer science, and interdisciplinary research relevant to bioinformatics and genomics.

The search strategy integrated MeSH (Medical Subject Headings) terms and keywords associated with bioinformatics and human genomics. Articles were selected based on relevance, contribution to the field, and their capacity to illustrate key concepts or recent advances. The review concentrated on literature published from January 2000 to April 2026, covering the period from the completion of the Human Genome Project to the present. Both primary research articles and authoritative review papers were included to ensure a comprehensive overview of the field.

Information from the selected literatures were synthesised narratively by organising key concepts, methodologies, and findings into the main themes of the review. This approach supported the integration of diverse perspectives and facilitated the identification of overarching trends in bioinformatics as applied to human genomics.

DATA MANAGEMENT AND STORAGE

Handling Big Data

One of the primary roles of bioinformatics in human genomics is the management and storage of large-scale datasets. Genomic research generates substantial volumes of data; for instance, sequencing a single human genome may require several gigabytes of storage space (Clark et al., 2024). In population-scale genomics projects that involve thousands or millions of individuals, data management complexity increases considerably (Wertenbroek et al., 2024). Bioinformatics addresses these challenges by offering advanced database systems and specialised file formats for genomic data (Davis-Turak et al., 2017). The Variant Call Format (VCF) and Binary Alignment Map (BAM) formats are widely used to efficiently store and compress genomic sequence and variation data. These formats reduce storage requirements and facilitate rapid data retrieval and analysis (Danecek et al., 2011; Beier et al., 2022).

Cloud computing platforms, such as Amazon Web Services' Genomics in the Cloud and Google Genomics, are integral to managing large-scale genomic data (Langmead and Nellore, 2018). These platforms offer scalable storage solutions and high-performance computing resources that are accessible remotely, enabling researchers to store and analyse extensive genomic datasets without the need for

substantial local infrastructure (Koppad et al., 2021). Data compression is also a critical aspect of genomic data management. Algorithms like CRAM (Compressed Reference-orientated Alignment Map) achieve high compression ratios by leveraging the similarity between an individual's genome and a reference genome. This approach reduces storage requirements and expedites data transfer and analysis (Brandon et al., 2009; Pinho et al., 2012).

Ensuring Data Integrity

Bioinformatics plays a critical role in maintaining data integrity throughout the genomic research process. It offers tools for quality control, error detection, and data cleaning, thereby ensuring that genomic information used in studies remains accurate and reliable. This reliability is fundamental for drawing valid conclusions from genomic analyses (Clark and Lillard, 2024; Zhang et al., 2025). Quality control commences at the sequencing stage, where bioinformatics tools evaluate the quality of raw sequencing reads (Guo et al., 2014; Zhang and Kang, 2021).

Programs such as FastQC identify issues including low-quality base calls, PCR artefacts, or contamination (Hong et al., 2015; Zhou et al., 2013). Tools like Trimmomatic and Cutadapt are subsequently employed to trim or filter out low-quality sequences (Bolger et al., 2014). During alignment and variant calling, bioinformatics pipelines integrate various quality metrics and filtering steps (Pinto et al., 2026). For instance, the Genome Analysis Toolkit (GATK) applies advanced statistical models to differentiate true genetic variants from sequencing errors or artefacts (McKenna et al., 2010).

Version control systems and data provenance tracking are essential for maintaining data integrity in genomics. These systems enable researchers to monitor different versions of datasets, analysis pipelines, and results, thereby ensuring reproducibility and facilitating the detection and correction of errors (Caliskan et al., 2023; Davis-Turak et al., 2017). Through robust data management and quality control solutions, bioinformatics establishes a solid and trustworthy foundation for genomic research, supporting reliable and meaningful scientific discoveries (Alkhatib and Gaede, 2024; Mulder et al., 2017).

SEQUENCE ANALYSIS AND ANNOTATION

Identifying Genetic Elements

Bioinformatics algorithms are essential for identifying and annotating genetic elements in the human genome, including protein-coding genes, regulatory regions, repetitive sequences, and non-coding RNAs. Genome annotation combines computational prediction with experimental evidence (Ejigu and Jung, 2020; Taher et al., 2015). BLAST (Basic Local Alignment Search Tool) is a key sequence analysis tool that compares newly sequenced DNA to existing databases to identify known genes and genetic elements. Advanced tools like GENSCAN use statistical models of gene structure to predict genes, considering features such as promoter sequences, exon-intron boundaries, and polyadenylation signals (Burge and Karlin, 1997; Guigo et al., 2000).

Tools such as MEME (Multiple EM for Motif Elicitation) identify regulatory regions by detecting recurring motifs in DNA sequences that may indicate transcription factor binding sites (Bailey et al., 2006). ChIP-seq data analysis pipelines determine the binding sites of transcription factors and other regulatory proteins across the genome (Bardet et al., 2013). Specialised tools like RepeatMasker detect repetitive elements, including transposons and tandem repeats, which play a significant role in genome structure and evolution (Liao et al., 2023; Hall et al., 2025). Non-coding RNAs, such as microRNAs and long non-coding RNAs, are identified using sequence homology searches and secondary structure prediction algorithms. Tools like Infernal (INFERENCE of RNA ALIGNMENT) use both sequence and structural information to identify RNA genes (Bao et al., 2012; Barquist et al., 2016).

Predicting Gene Function

Bioinformatics uses advanced machine learning algorithms and comparative genomics to predict the functions of newly discovered genes. This process, known as functional annotation, is essential for clarifying gene roles in biological processes and their potential links to diseases (Pilalis et al., 2025; Adugna et al., 2025). One common prediction method relies on sequence similarity. If a new gene shows significant sequence homology with a gene of known function in another organism, it likely performs a

similar role. Tools such as InterProScan detect known protein domains within gene sequences, providing insights into potential molecular functions (Pearson, 2013).

Machine learning algorithms, such as Support Vector Machines (SVMs) and Random Forests, combine features including sequence data, expression patterns, and protein-protein interaction data to predict gene functions. These models are trained on genes with known functions and then applied to uncharacterised genes (Mahood et al., 2020; Merumba et al., 2025). Comparative genomics examines evolutionary conservation across species. Genes that are highly conserved among diverse species are likely to have essential biological functions. Tools such as OrthoMCL identify orthologous genes across multiple species, supporting evolutionary analyses (Kanwal et al., 2017).⁴⁴

Network-based approaches use protein-protein interaction data, co-expression networks, and other biological networks to predict gene functions based on the principle of "guilt by association." Genes with similar interaction patterns or expression profiles are likely to share related functions (Gillis and Pavlidis, 2012; Ko and Brandizzi, 2020). By integrating these methods, bioinformatics generates valuable hypotheses about gene functions, guiding experimental research and accelerating genomic discovery (Hamid et al., 2009).

COMPARATIVE AND EVOLUTIONARY GENOMICS

Tracing Human Evolution

Bioinformatics enables comparison of human genomic sequences with those of other species, providing insight into evolutionary history. Comparative genomics, a branch of bioinformatics, has advanced understanding of human evolution and the genetic basis of human-specific traits (Muse, 2005; Mendoza et al., 2022). Whole-genome alignment tools such as LASTZ and MULTIZ allow researchers to compare entire genomes across species. These alignments reveal conserved regions that are likely functionally important, as well as areas of rapid evolution in the human lineage (Armstrong et al., 2019; Herrero et al., 2016). Phylogenetic analysis tools, including PHYLIP and MrBayes, use genomic data to reconstruct evolutionary relationships among species. These analyses have improved knowledge of human origins and relationships to other primates. For example, genomic comparisons have clarified the timing of the human-chimpanzee divergence and interbreeding events between modern humans and archaic hominins such as Neanderthals and Denisovans (Zou et al., 2024; Aris-Brosou and Xia, 2008).

Bioinformatics tools also help identify genomic regions showing signs of positive selection in humans. These regions often contain genes important to human evolution. For example, comparative genomic analyses have found human-specific changes in genes linked to brain development, language, and diet, offering insight into the genetic basis of uniquely human traits (Nielsen et al., 2007; Kelley et al., 2006). Advanced statistical methods, such as the McDonald-Kreitman test and dN/dS ratio analysis, detect subtle selection signals by comparing rates of synonymous and non-synonymous mutations within and between species. These methods have revealed selection pressures on genes related to immunity, metabolism, and cognitive functions (Egea et al., 2008; Fay, 2011).

Understanding Genetic Diversity

Bioinformatics tools support the analysis of genetic variation in human populations, including single-nucleotide polymorphisms (SNPs), insertions and deletions (indels), and larger structural variants. These analyses are essential for understanding human genetic diversity, migration patterns, and the genetic basis of traits and diseases (Li and Wei, 2015; Clifford et al., 2004). Population genetics software such as PLINK and ADMIXTURE enables the analysis of large-scale genomic data from diverse groups. These tools help identify population structure, estimate ancestry proportions, and detect signatures of natural selection (Cheng and Steinrucken, 2024; He et al., 2025). Haplotype analysis tools like SHAPEIT and BEAGLE reconstruct combinations of alleles inherited together. This information is vital for tracing human population history and mapping disease-associated genes (Feng et al., 2022).

Bioinformatics has been central to large-scale projects cataloguing human genetic variation, such as the 1000 Genomes Project and the Human Genome Diversity Project. These initiatives have generated comprehensive maps of genetic variation, clarifying the complex history of human migrations and

admixture events (Devuyst, 2015). Bioinformatics tools also integrate genomic data with archaeological and linguistic evidence, providing a more complete understanding of human history and diversity. For example, combining genomic and linguistic analyses has offered insights into the spread of Indo-European languages (Morozova *et al.*, 2016). By supporting the analysis and interpretation of genetic variation, bioinformatics continues to advance knowledge of human diversity and evolution, with important implications for medicine and anthropology.

CLINICAL APPLICATIONS

Personalised Medicine

Bioinformatics plays a vital role in personalised medicine by interpreting individual genomic profiles and translating them into clinically actionable insights. This approach tailors treatments to each patient's genetic makeup, improving efficacy and reducing adverse effects (Khan *et al.*, 2025; Jamalnia and Weiskirchen, 2025). In pharmacogenomics, bioinformatics tools predict drug responses based on genetic variants. Databases such as PharmGKB compile gene-drug interaction data, while algorithms integrate these data with individual profiles to guide drug selection and dosing. For example, CYP2C19 gene variants affect the metabolism of certain antidepressants and antiplatelet drugs, and bioinformatics tools can quickly identify patients who may need alternative therapies (Thorn *et al.*, 2010; Tong *et al.*, 2021). Bioinformatics is critical in cancer treatment for analysing tumour genomics. Tools such as MutSig and GISTIC identify driver mutations, distinguishing them from passenger mutations and guiding targeted therapy selection. For example, detecting the BRAF V600E mutation in melanoma indicates the potential effectiveness of BRAF inhibitors (Jimenez-Santos *et al.*, 2022; Castellani *et al.*, 2023). Bioinformatics also enables interpretation of polygenic risk scores (PRS), which estimate an individual's risk for complex diseases by aggregating multiple genetic variants. Tools like PRSice calculate these scores, and advanced machine learning methods are being developed to improve their accuracy and clinical relevance (Ndong-Sima *et al.*, 2024; Schuran *et al.*, 2025). Additionally, bioinformatics integrates genomic data with other biological data types, such as transcriptomics, proteomics, and metabolomics. This multi-omics approach provides a comprehensive view of an individual's biological state, supporting more precise and personalised interventions (Elrashedy *et al.*, 2025; Subramanian *et al.*, 2020).

Diagnosing Genetic Disorders

Bioinformatics pipelines are essential for diagnosing genetic disorders, particularly rare diseases caused by single-gene mutations. Clinicians compare a patient's genomic sequence with databases of known pathogenic mutations to identify potential genetic causes, which supports accurate diagnoses and targeted treatments (Kanzi *et al.*, 2020; Hong *et al.*, 2024). The diagnostic process typically begins with whole-genome or whole-exome sequencing of the patient's DNA. Bioinformatics tools then align sequencing reads to a reference genome and identify genetic variants. Annotation tools, such as ANNOVAR and VEP (Variant Effect Predictor), assess the potential functional impact of each variant (Clark and Lillard, 2024; Pereira *et al.*, 2020).

Variant filtering and prioritisation algorithms are then used to refine the list of potentially causative mutations. These algorithms consider factors such as variant rarity in the general population, predicted functional impact, and alignment with the patient's phenotype and the suspected inheritance pattern (Pedersen *et al.*, 2021). Databases such as OMIM (Online Mendelian Inheritance in Man), ClinVar, and HGMD (Human Gene Mutation Database) offer essential information on known pathogenic mutations and their associated phenotypes. Bioinformatics tools can automatically query these databases and flag variants previously linked to genetic disorders (Hamosh *et al.*, 2005).

For novel variants, tools such as SIFT, PolyPhen, and CADD use machine learning to predict pathogenicity based on genomic location, evolutionary conservation, and the predicted impact on protein structure and function (Murali *et al.*, 2024; Grimm *et al.*, 2015). If a single causative mutation is not found, bioinformatics tools can help detect compound heterozygous mutations or propose candidate genes through pathway analysis and known gene functions (van Driel and Brunner, 2006; Kobren *et al.*, 2024). Bioinformatics is also essential for analysing structural variants, including copy number variations

(CNVs), which are important in diagnosing certain genetic disorders. Tools such as PennCNV and XHMM can detect these larger genomic alterations from sequencing or microarray data (Pounraja et al., 2019; Tattini et al., 2015). These advanced analytical capabilities have significantly increased the diagnostic yield for genetic disorders, enabling clinicians to provide more precise diagnoses and targeted treatments for patients with rare and complex genetic conditions.

FUTURE PERSPECTIVES

As genomic data continues to grow and algorithms become more advanced, bioinformatics will play an increasingly important role in human genomics. Emerging areas such as single-cell genomics, epigenomics, and multi-omics integration present both challenges and opportunities for the field (Satam et al., 2023; Yetgin, 2025). Single-cell genomics allows researchers to study genetic variation and gene expression at the level of individual cells, offering new insights into cellular diversity, development, and complex diseases like cancer. Bioinformatics is crucial for managing large single-cell datasets, developing clustering algorithms, and identifying cell-type-specific gene expression or genetic variation patterns (Ortega-Batista et al., 2025; Wang et al., 2026).

Epigenomics, which studies heritable changes in gene function without altering the DNA sequence, is also expanding rapidly. Bioinformatics tools are essential for analysing data from ChIP-seq, ATAC-seq, and DNA methylation sequencing, enabling the mapping of epigenetic landscapes and providing insights into gene regulation and cellular identity (Wang and Chang, 2018; Lin and Liu, 2025). Integrating multiple omics data types, such as genomics, transcriptomics, proteomics, and metabolomics, is a major focus in bioinformatics. This multi-omics approach offers a more complete understanding of biological systems but also presents significant computational challenges. Machine learning and network analysis methods are being developed to integrate these data types and build predictive models.

Artificial intelligence, especially deep learning, is expected to transform genomics and bioinformatics by improving the prediction of genetic variant effects, modelling complex gene-environment interactions, and supporting personalised therapeutic design (Jiang et al., 2025; Quazi, 2022). As genomic sequencing becomes more common in clinical practice, bioinformatics will play a critical role in healthcare. Developing robust, scalable, and interpretable algorithms for clinical genomic data analysis is essential. There is also an increasing need for standardised pipelines and reporting formats to ensure reliable and reproducible genomic analyses in clinical settings (Davis-Turak et al., 2017; Bianconi et al., 2023). Privacy and security of genomic data are additional priorities. Developing secure storage, sharing methods, and privacy-preserving analysis techniques will be vital as genomic information becomes a routine part of personal health records (Bonomi et al., 2020; Kuo et al., 2022).

CONCLUSION

In conclusion, bioinformatics serves as a vital bridge between raw genomic data and meaningful biological insights. Its ability to manage, analyse, and interpret genomic information is essential for advancing human genomics and applying this knowledge in medicine and related fields. The ongoing collaboration between genomics and bioinformatics will continue to drive scientific discovery, deepen our understanding of human biology, and support transformative progress in healthcare.

REFERENCES

- Aduagna, A., Amare, G. A., & Jemal, M. (2025). Machine Learning Approach and Bioinformatics Analysis Discovered Key Genomic Signatures for Hepatitis B Virus-Associated Hepatocyte Remodelling and Hepatocellular Carcinoma. *Cancer Informatics*, 24, 11769351251333847. <https://doi.org/10.1177/11769351251333847>
- Akintunde, O., Tucker, T., & Carabetta, V. J. (2023). The evolution of next-generation sequencing technologies. *ArXiv*, arXiv:2305.08724v1.
- Alkhatib, R., & Gaede, K. I. (2024). Data Management in Biobanking: Strategies, Challenges, and Future Directions. *Biotech (Basel (Switzerland))*, 13(3), 34. <https://doi.org/10.3390/biotech13030034>

- Aris-Brosou, S., & Xia, X. (2008). Phylogenetic analyses: A toolbox expanding towards Bayesian methods. *International Journal of Plant Genomics*, 2008, 683509. <https://doi.org/10.1155/2008/683509>
- Armstrong, J., Fiddes, I. T., Diekhans, M., & Paten, B. (2019). Whole-Genome Alignment and Comparative Annotation. *Annual Review of Animal Biosciences*, 7, 41–64.
- Baev V. (2025). Bioinformatics Research in Bacterial Genomics and Metagenomics. *Current Issues in Molecular Biology*, 47(4), 258. <https://doi.org/10.3390/cimb47040258>
- Bailey, T. L., Williams, N., Misleh, C., & Li, W. W. (2006). MEME: discovering and analysing DNA and protein sequence motifs. *Nucleic Acids Research*, 34(Web Server issue), W369–W373. <https://doi.org/10.1093/nar/gkl198>
- Bao, M., Cervantes Cervantes, M., Zhong, L., & Wang, J. T. (2012). Searching for non-coding RNAs in genomic sequences using ncRNAscout. *Genomics, Proteomics & Bioinformatics*, 10(2), 114–121.
- Bardet, A. F., Steinmann, J., Bafna, S., Knoblich, J. A., Zeitlinger, J., & Stark, A. (2013). Identification of transcription factor binding sites from ChIP-seq data at high resolution. *Bioinformatics (Oxford, England)*, 29(21), 2705–2713.
- Barquist, L., Burge, S. W., & Gardner, P. P. (2016). Studying RNA Homology and Conservation with Infernal: From Single Sequences to RNA Families. *Current Protocols in Bioinformatics*, 54, 12.13.1–12.13.25. <https://doi.org/10.1002/cpbi.4>
- Bayat A. (2002). Science, medicine, and the future: Bioinformatics. *BMJ (Clinical research ed.)*, 324(7344), 1018–1022.
- Beier, S., Fiebig, A., Pommier, C., Liyanage, I., Lange, M., Kersey, P. J., Weise, S. ... & Scholz, U. (2022). Recommendations for the formatting of Variant Call Format (VCF) files to make plant genotyping data FAIR. *F1000Research*, 11, ELIXIR-231. <https://doi.org/10.12688/f1000research.109080.2>
- Bianconi, I., Aschbacher, R., & Pagani, E. (2023). Current Uses and Future Perspectives of Genomic Technologies in Clinical Microbiology. *Antibiotics (Basel, Switzerland)*, 12(11), 1580. <https://doi.org/10.3390/antibiotics12111580>
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, 30(15), 2114–2120.
- Bonomi, L., Huang, Y., & Ohno-Machado, L. (2020). Privacy challenges and research opportunities for genomic data sharing. *Nature Genetics*, 52(7), 646–654.
- Brandon, M. C., Wallace, D. C., & Baldi, P. (2009). Data structures and compression algorithms for genomic sequence data. *Bioinformatics (Oxford, England)*, 25(14), 1731–1738.
- Burge, C., & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, 268(1), 78–94.
- Caliskan, A., Dangwal, S., & Dandekar, T. (2023). Metadata integrity in bioinformatics: Bridging the gap between data and knowledge. *Computational and Structural Biotechnology Journal*, 21, 4895–4913.
- Castellani, G., Buccarelli, M., Arasi, M. B., Rossi, S., Pisanu, M. E., Bellenghi, M., Lintas, C., & Tabolacci, C. (2023). BRAF Mutations in Melanoma: Biological Aspects, Therapeutic Implications, and Circulating Biomarkers. *Cancers*, 15(16), 4026. <https://doi.org/10.3390/cancers15164026>
- Cheng, X., & Steinrücken, M. (2024). Population Genomic Scans for Natural Selection and Demography. *Annual Review of Genetics*, 58(1), 319–339.
- Clark, A. J., & Lillard, J. W., Jr (2024). A Comprehensive Review of Bioinformatics Tools for Genomic Biomarker Discovery Driving Precision Oncology. *Genes*, 15(8), 1036. <https://doi.org/10.3390/genes15081036>

- Clark, A. J., & Lillard, J. W., Jr (2024). A Comprehensive Review of Bioinformatics Tools for Genomic Biomarker Discovery Driving Precision Oncology. *Genes*, 15(8), 1036. <https://doi.org/10.3390/genes15081036>
- Clifford, R. J., Edmonson, M. N., Nguyen, C., Scherpbier, T., Hu, Y., & Buetow, K. H. (2004). Bioinformatics tools for single-nucleotide polymorphism discovery and analysis. *Annals of the New York Academy of Sciences*, 1020, 101–109.
- Collins, F. S., & Fink, L. (1995). The Human Genome Project. *Alcohol Health and Research World*, 19(3), 190–195.
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E. ... & 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, 27(15), 2156–2158.
- Davis-Turak, J., Courtney, S. M., Hazard, E. S., Glen, W. B., Jr, da Silveira, W. A., Wesselman, T., Harbin, L. P., Wolf, B. J., Chung, D., & Hardiman, G. (2017). Genomics pipelines and data integration: challenges and opportunities in the research setting. *Expert Review of Molecular Diagnostics*, 17(3), 225–237.
- Devuyst O. (2015). The 1000 Genomes Project: Welcome to a New World. *Peritoneal Dialysis International: Journal of the International Society for Peritoneal Dialysis*, 35(7), 676–677.
- Egea, R., Casillas, S., & Barbadilla, A. (2008). Standard and generalised McDonald-Kreitman test: a website to detect selection by comparing different classes of DNA sites. *Nucleic Acids Research*, 36(Web Server issue), W157–W162. <https://doi.org/10.1093/nar/gkn337>
- Ejigu, G. F., & Jung, J. (2020). Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing. *Biology*, 9(9), 295. <https://doi.org/10.3390/biology9090295>
- Elrashedy, A., Mousa, W., Nayel, M., Salama, A., Zaghawa, A., Elsify, A., & Hasan, M. E. (2025). Advances in bioinformatics and multi-omics integration: transforming viral infectious disease research in veterinary medicine. *Virology Journal*, 22(1), 22. <https://doi.org/10.1186/s12985-025-02640-x>
- Fay J. C. (2011). Weighing the evidence for adaptation at the molecular level. *Trends in Genetics: TIG*, 27(9), 343–349.
- Feng, C., Wang, X., Wu, S., Ning, W., Song, B., Yan, J., & Cheng, S. (2022). HAPPE: A Tool for Population Haplotype Analysis and Visualisation in Editable Excel Tables. *Frontiers in Plant Science*, 13, 927407. <https://doi.org/10.3389/fpls.2022.927407>
- Gillis, J., & Pavlidis, P. (2012). "Guilt by association" is the exception rather than the rule in gene networks. *PLoS Computational Biology*, 8(3), e1002444. <https://doi.org/10.1371/journal.pcbi.1002444>
- Grimm, D. G., Azencott, C. A., Aicheler, F., Gieraths, U., MacArthur, D. G., Samocha, K. E., Cooper, D. N. ... & Borgwardt, K. M. (2015). The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Human Mutation*, 36(5), 513–523.
- Guigó, R., Agarwal, P., Abril, J. F., Burset, M., & Fickett, J. W. (2000). An assessment of gene prediction accuracy in large DNA sequences. *Genome Research*, 10(10), 1631–1642.
- Guo, Y., Ye, F., Sheng, Q., Clark, T., & Samuels, D. C. (2014). Three-stage quality control strategies for DNA re-sequencing data. *Briefings in Bioinformatics*, 15(6), 879–889.
- Hall, L. L., Smith, K. P., & Lawrence, J. B. (2025). Emerging Functions of the Repeat Genome in Nuclear Structure: A View from the Human Karyotype. *Annual Review of Genomics and Human Genetics*, 26(1), 45–75.
- Hamid, J. S., Hu, P., Roslin, N. M., Ling, V., Greenwood, C. M., & Beyene, J. (2009). Data integration in genetics and genomics: methods and challenges. *Human Genomics and Proteomics: HGP*, 2009, 869093. <https://doi.org/10.4061/2009/869093>
- Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., & McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(Database issue), D514–D517. <https://doi.org/10.1093/nar/gki033>

- He, G., Chen, J., Duan, S., Yang, Q., Li, B., Luo, L., Zhong, J. ... & Wang, M. (2025). Largest-Scale Genomic Resource Reconstructing the Genetic Origin, Population Structure, and Biological Adaptations of the Hui People. *Molecular Biology and Evolution*, 42(10), msaf225. <https://doi.org/10.1093/molbev/msaf225>
- Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A. J. ... & Flicek, P. (2016). Ensembl comparative genomics resources. *Database: The Journal of Biological Databases and Curation*, 2016, bav096. <https://doi.org/10.1093/database/bav096>
- Hong, C., Manimaran, S., & Johnson, W. E. (2015). PathoQC: Computationally Efficient Read Preprocessing and Quality Control for High-Throughput Sequencing Data Sets. *Cancer Informatics*, 13(Suppl 1), 167–176.
- Hong, J., Lee, D., Hwang, A., Kim, T., Ryu, H. Y., & Choi, J. (2024). Rare disease genomics and precision medicine. *Genomics & Informatics*, 22(1), 28. <https://doi.org/10.1186/s44342-024-00032-1>
- Jamalnia, M., & Weiskirchen, R. (2025). Advances in personalised medicine: translating genomic insights into targeted therapies for cancer treatment. *Annals of Translational Medicine*, 13(2), 18. <https://doi.org/10.21037/atm-25-34>
- Jiang, J., Li, Y., Cao, S., Shan, Y., Liu, Y., Fei, T., Yu, Y. ... & Yuan, J. (2025). Artificial intelligence in bioinformatics: a survey. *Briefings in Bioinformatics*, 26(6), bbaf576. <https://doi.org/10.1093/bib/bbaf576>
- Jiménez-Santos, M. J., García-Martín, S., Fustero-Torre, C., Di Domenico, T., Gómez-López, G., & Al-Shahrour, F. (2022). Bioinformatics roadmap for therapy selection in cancer genomics. *Molecular Oncology*, 16(21), 3881–3908.
- Kanwal, S., Khan, F. Z., Lonie, A., Sinnott, R. O. (2017). Investigating reproducibility and tracking provenance - A genomic workflow case study. *BMC Bioinformatics*, 18(1), 337. <https://doi.org/10.1186/s12859-017-1747-0>
- Kanzi, A. M., San, J. E., Chimukangara, B., Wilkinson, E., Fish, M., Ramsuran, V., & de Oliveira, T. (2020). Next Generation Sequencing and Bioinformatics Analysis of Family Genetic Inheritance. *Frontiers in Genetics*, 11, 544162. <https://doi.org/10.3389/fgene.2020.544162>
- Kelley, J. L., Madeoy, J., Calhoun, J. C., Swanson, W., & Akey, J. M. (2006). Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Research*, 16(8), 980–989.
- Khan, A., Barapatre, A. R., Babar, N., Doshi, J., Ghaly, M., Patel, K. G., Nawaz, S. ... & Jamil, U. (2025). Genomic medicine and personalised treatment: a narrative review. *Annals of Medicine and Surgery (2012)*, 87(3), 1406–1414.
- Ko, D. K., & Brandizzi, F. (2020). Network-based approaches for understanding gene regulation and function in plants. *The Plant Journal: For Cell and Molecular Biology*, 104(2), 302–317.
- Kobren, S. N., Moldovan, M. A., Reimers, R., Traviglia, D., Li, X., Barnum, D., Veit, A. ... & Sunyaev, S. R. (2024). Joint, multifaceted genomic analysis enables diagnosis of diverse, ultra-rare monogenic presentations. *bioRxiv: The Preprint Server for Biology*, 2024.02.13.580158. <https://doi.org/10.1101/2024.02.13.580158>
- Koppad, S. B. A., Gkoutos, G. V., & Acharjee, A. (2021). Cloud Computing Enabled Big Multi-Omics Data Analytics. *Bioinformatics and Biology Insights*, 15, 11779322211035921. <https://doi.org/10.1177/11779322211035921>
- Kuo, T. T., Jiang, X., Tang, H., Wang, X., Harmanci, A., Kim, M., Post, K. ... & Ohno-Machado, L. (2022). The evolving privacy and security concerns for genomic data analysis and sharing, as observed from the iDASH competition. *Journal of the American Medical Informatics Association: JAMIA*, 29(12), 2182–2190.
- Langmead, B., & Nellore, A. (2018). Cloud computing for genomic data analysis and collaboration. *Nature Reviews. Genetics*, 19(4), 208–219.

- Li, L., & Wei, D. (2015). Bioinformatics tools for discovery and functional analysis of single-nucleotide polymorphisms. *Advances in Experimental Medicine and Biology*, 827, 287–310.
- Liao, X., Zhu, W., Zhou, J., Li, H., Xu, X., Zhang, B., & Gao, X. (2023). Repetitive DNA sequence detection and its role in the human genome. *Communications Biology*, 6(1), 954. <https://doi.org/10.1038/s42003-023-05322-y>
- Lin, L., & Liu, Y. (2025). Advances in Epigenomic Sequencing and Their Applications in Cancer Diagnostics. *Precision Chemistry*, 3(10), 581–603.
- Lu, Y. F., Goldstein, D. B., Angrist, M., & Cavalleri, G. (2014). Personalised medicine and human genetic diversity. *Cold Spring Harbour Perspectives in Medicine*, 4(9), a008581. <https://doi.org/10.1101/cshperspect.a008581>
- Mahood, E. H., Kruse, L. H., & Moghe, G. D. (2020). Machine learning: A powerful tool for gene function prediction in plants. *Applications in Plant Sciences*, 8(7), e11376. <https://doi.org/10.1002/aps3.11376>
- McCue, M. E., & McCoy, A. M. (2017). The Scope of Big Data in One Medicine: Unprecedented Opportunities and Challenges. *Frontiers in Veterinary Science*, 4, 194. <https://doi.org/10.3389/fvets.2017.00194>
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K. ... & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analysing next-generation DNA sequencing data. *Genome Research*, 20(9), 1297–1303.
- Mendoza, R. M., Kim, S. H., Vasquez, R., Hwang, I. C., Park, Y. S., Paik, H. D., Moon, G. S., & Kang, D. K. (2022). Bioinformatics and its role in the study of the evolution and probiotic potential of lactic acid bacteria. *Food Science and Biotechnology*, 32(4), 389–412.
- Merumba, S. B., Ahmed, H. O., Fu, D., & Yang, P. (2025). Recent Advances and Application of Machine Learning for Protein-Protein Interaction Prediction in Rice: Challenges and Future Perspectives. *Proteomes*, 13(4), 54. <https://doi.org/10.3390/proteomes13040054>
- Morozova, I., Flegontov, P., Mikheyev, A. S., Bruskin, S., Asgharian, H., Ponomarenko, P., Klyuchnikov, V. ... & Tatarinova, T. V. (2016). Toward high-resolution population genomics using archaeological samples. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, 23(4), 295–310.
- Mulder, N. J., Adebiyi, E., Adebiyi, M., Adeyemi, S., Ahmed, A., Ahmed, R., Akanle, B. ... & H3ABioNet Consortium, as members of the H3Africa Consortium (2017). Development of Bioinformatics Infrastructure for Genomics Research. *Global heart*, 12(2), 91–98.
- Murali, H., Wang, P., Liao, E. C., & Wang, K. (2024). Genetic variant classification by predicted protein structure: A case study on IRF6. *Computational and Structural Biotechnology Journal*, 23, 892–904.
- Muse S. (2005). GENOMICS AND BIOINFORMATICS. *Introduction to Biomedical Engineering*, 799–831.
- Ndong Sima, C. A. A., Step, K., Swart, Y., Schurz, H., Uren, C., & Möller, M. (2024). Methodologies underpinning polygenic risk scores estimation: a comprehensive overview. *Human Genetics*, 143(11), 1265–1280.
- Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C., & Clark, A. G. (2007). Recent and ongoing selection in the human genome. *Nature Reviews. Genetics*, 8(11), 857–868.
- Ortega-Batista, A., Jaén-Alvarado, Y., Moreno-Labrador, D., Gómez, N., García, G., & Guerrero, E. N. (2025). Single-Cell Sequencing: Genomic and Transcriptomic Approaches in Cancer Cell Biology. *International Journal of Molecular Sciences*, 26(5), 2074. <https://doi.org/10.3390/ijms26052074>
- Pearson W. R. (2013). An introduction to sequence similarity ("homology") searching. *Current Protocols in Bioinformatics*, Chapter 3, 3.1.1–3.1.8. <https://doi.org/10.1002/0471250953.bi0301s42>
- Pedersen, B. S., Brown, J. M., Dashnow, H., Wallace, A. D., Velinder, M., Tristani-Firouzi, M., Schiffman, J. D. ... & Quinlan, A. R. (2021). Effective variant filtering and expected candidate

- variant yield in studies of rare human disease. *NPJ Genomic Medicine*, 6(1), 60. <https://doi.org/10.1038/s41525-021-00227-3>
- Pereira, R., Oliveira, J., & Sousa, M. (2020). Bioinformatics and Computational Tools for Next-Generation Sequencing Analysis in Clinical Genetics. *Journal of Clinical Medicine*, 9(1), 132. <https://doi.org/10.3390/jcm9010132>
- Pilalis, E., Zisis, D., Andrinopoulou, C., Karamanidou, T., Antonara, M., Stavropoulos, T. G., & Chatziioannou, A. (2025). Genome-wide functional annotation of variants: a systematic review of state-of-the-art tools, techniques and resources. *Frontiers in Pharmacology*, 16, 1474026. <https://doi.org/10.3389/fphar.2025.1474026>
- Pinho, A. J., Pratas, D., & Garcia, S. P. (2012). GReEn: a tool for efficient compression of genome resequencing data. *Nucleic Acids Research*, 40(4), e27. <https://doi.org/10.1093/nar/gkr1124>
- Pinto, V., Sousa, L., & Silva, C. (2026). Variant calling in genomics: A comparative performance analysis and decision guide. *PloS one*, 21(2), e0339891. <https://doi.org/10.1371/journal.pone.0339891>
- Pounraja, V. K., Jayakar, G., Jensen, M., Kelkar, N., & Girirajan, S. (2019). A machine-learning approach for accurate detection of copy number variants from exome sequencing. *Genome Research*, 29(7), 1134–1143.
- Quazi S. (2022). Artificial intelligence and machine learning in precision and genomic medicine. *Medical oncology (Northwood, London, England)*, 39(8), 120. <https://doi.org/10.1007/s12032-022-01711-1> (Retraction published *Med Oncol.* 2025 Apr 26;42(6):180. doi: 10.1007/s12032-025-02732-2.)
- Satam, H., Joshi, K., Mangrolia, U., Waghoo, S., Zaidi, G., Rawool, S., Thakare, R. P. ... & Malonia, S. K. (2023). Next-Generation Sequencing Technology: Current Trends and Advancements. *Biology*, 12(7), 997. <https://doi.org/10.3390/biology12070997>
- Schuran, M., Goudey, B., Dite, G. S., & Makalic, E. (2025). A survey on deep learning for polygenic risk scores. *Briefings in Bioinformatics*, 26(4), bbaf373. <https://doi.org/10.1093/bib/bbaf373>
- Subramanian, I., Verma, S., Kumar, S., Jere, A., & Anamika, K. (2020). Multi-omics Data Integration, Interpretation, and Its Application. *Bioinformatics and Biology Insights*, 14, 1177932219899051. <https://doi.org/10.1177/1177932219899051>
- Taher, L., Narlikar, L., & Ovcharenko, I. (2015). Identification and computational analysis of gene regulatory elements. *Cold Spring Harbour Protocols*, 2015(1), pdb.top083642. <https://doi.org/10.1101/pdb.top083642>
- Tattini, L., D'Aurizio, R., & Magi, A. (2015). Detection of Genomic Structural Variants from Next-Generation Sequencing Data. *Frontiers in Bioengineering and Biotechnology*, 3, 92. <https://doi.org/10.3389/fbioe.2015.00092>
- Thorn, C. F., Klein, T. E., & Altman, R. B. (2010). Pharmacogenomics and bioinformatics: PharmGKB. *Pharmacogenomics*, 11(4), 501–505.
- Tong, H., Phan, N. V. T., Nguyen, T. T., Nguyen, D. V., Vo, N. S., & Le, L. (2021). Review on Databases and Bioinformatic Approaches on Pharmacogenomics of Adverse Drug Reactions. *Pharmacogenomics and Personalised Medicine*, 14, 61–75.
- van Driel, M. A., & Brunner, H. G. (2006). Bioinformatics methods for identifying candidate disease genes. *Human Genomics*, 2(6), 429–432.
- Wang, K. C., & Chang, H. Y. (2018). Epigenomics: Technologies and Applications. *Circulation Research*, 122(9), 1191–1199.
- Wang, X., Li, N., Wang, W., & Liu, B. (2026). Single-cell sequencing: accurate disease detection. *Clinical & Translational Oncology: official publication of the Federation of Spanish Oncology Societies and of the National Cancer Institute of Mexico*, 28(2), 404–423.
- Wertenbroek, R., Hofmeister, R. J., Xenarios, I., Thoma, Y., & Delaneau, O. (2024). Improving population-scale statistical phasing with whole-genome sequencing data. *PLoS Genetics*, 20(7), e1011092. <https://doi.org/10.1371/journal.pgen.1011092>

- Yetgin A. (2025). Revolutionising multi-omics analysis with artificial intelligence and data processing. *Quantitative Biology (Beijing, China)*, 13(3), e70002. <https://doi.org/10.1002/qub2.70002>
- Zhang, F., & Kang, H. M. (2021). FASTQuick: rapid and comprehensive quality assessment of raw sequence reads. *Giga Science*, 10(2), giab004. <https://doi.org/10.1093/gigascience/giab004>
- Zhang, S., Liu, K., Liu, Y., Hu, X., & Gu, X. (2025). The role and application of bioinformatics techniques and tools in drug discovery. *Frontiers in Pharmacology*, 16, 1547131. <https://doi.org/10.3389/fphar.2025.1547131>
- Zhou, Q., Su, X., Wang, A., Xu, J., & Ning, K. (2013). QC-Chain: fast and holistic quality control method for next-generation sequencing data. *PloS one*, 8(4), e60234. <https://doi.org/10.1371/journal.pone.0060234>
- Zou, Y., Zhang, Z., Zeng, Y., Hu, H., Hao, Y., Huang, S., & Li, B. (2024). Common Methods for Phylogenetic Tree Construction and Their Implementation in R. *Bioengineering (Basel, Switzerland)*, 11(5), 480. <https://doi.org/10.3390/bioengineering11050480>