Genomics and Applied Biology 2024, Vol.15, No.4, 172-181
http://bioscipublisher.com/index.php/gab

Research Article Open Access

Exploring the Future of Biostatistics in Genomic Research: Opportunities and Challenges

Manman Li
Hainan Institute of Biotechnology, Haikou, 570206, Hainan, China
Corresponding author: manman.li@hibio.org

Genomics and Applied Biology, 2024, Vol.15, No.4 doi: 10.5376/gab.2024.15.0019
Received: 03 May, 2024
Accepted: 18 Jun., 2024
Published: 07 Jul., 2024

Preferred citation for this article:
Li M.M., 2024, Exploring the future of biostatistics in genomic research: opportunities and challenges, Genomics and Applied Biology, 15(4): 172-181 (doi: 10.5376/gab.2024.15.0019)

Abstract The study discoveries highlight the integration of biostatistics with genomic data to enhance precision public health initiatives, the development of novel computational methods to handle large genomic datasets, and the critical role of biostatistics in genome-wide association studies (GWAS). The study also underscores the importance of addressing data integration, methodological rigor, and health equity to fully leverage genomic data in public health. The findings suggest that the future of biostatistics in genomic research is promising, with significant potential to advance our understanding of complex genetic diseases and improve public health outcomes. However, realizing this potential will require overcoming substantial challenges related to data management, methodological development, and interdisciplinary collaboration.

Keywords Biostatistics; Genomic research; Genome-wide association studies; Computational methods; Data integration

1 Introduction

Biostatistics plays a pivotal role in the field of genomics, providing the necessary tools and methodologies to analyze complex biological data. The integration of biostatistics with genomics has enabled researchers to make significant strides in understanding genetic variations and their implications for health and disease. The advent of high-throughput technologies, such as next-generation sequencing, has exponentially increased the volume of genomic data, necessitating advanced biostatistical methods to manage, analyze, and interpret these data effectively (Ziegler et al., 2008; Manzoni et al., 2016; Davis-Turak et al., 2017).

The field of biostatistics has evolved considerably over the past few decades, particularly with the rise of genomic research. Initially, biostatistical methods were primarily used for analyzing small-scale genetic studies. However, the completion of the Human Genome Project and subsequent technological advancements have transformed the landscape, leading to the development of genome-wide association studies (GWAS) and other large-scale genomic analyses. These advancements have posed new biostatistical challenges, such as managing multiple testing issues and detecting gene-environment interactions, which have been addressed through innovative statistical techniques (Baráth and Rosner, 1992; Ziegler et al., 2008; Duggal et al., 2019).

In modern genomic research, biostatistics is indispensable for ensuring the scientific rigor and validity of study findings. It aids in the design of experiments, data collection, and the interpretation of results, thereby enhancing the reliability of genomic studies. Biostatistics also plays a crucial role in the integration of various types of omics data, such as genomics, transcriptomics, and proteomics, facilitating a comprehensive understanding of biological systems. This integrative approach is essential for the advancement of precision medicine, where individualized treatment plans are developed based on a patient's genetic profile (Mandrekar and Mandrekar, 2009; McCarthy et al., 2013; Manzoni et al., 2016; Roberts et al., 2021).

This study aims to explore the future of biostatistics in genomic research, focusing on the opportunities and challenges that lie ahead. The objectives are to review the current state of biostatistical methods in genomics and their applications, identify the key challenges faced by biostatisticians in the era of big data and high-throughput technologies, discuss the potential opportunities for advancing biostatistical methodologies to better support genomic research, and highlight the importance of interdisciplinary collaboration in overcoming these challenges and maximizing the potential of genomic data.

2 Current Applications of Biostatistics in Genomic Research

2.1 Biostatistical approaches in genome-wide association studies (GWAS)

Genome-Wide Association Studies (GWAS) have become a cornerstone in genomic research, leveraging biostatistical methods to identify genetic variants associated with complex traits and diseases. The application of GWAS has been instrumental in uncovering genotype-phenotype associations across various species, including plants and animals. For instance, GWAS has been extensively used in maize to link genotypic variations to phenotypic differences, utilizing advanced statistical models to optimize study design and analysis (Xiao et al., 2017). Similarly, GWAS has facilitated significant discoveries in human genetics, identifying risk loci for diseases such as autism spectrum disorder and schizophrenia through large-scale meta-analyses (Anney et al., 2017). The integration of biostatistics in GWAS has also led to the development of new models and population designs, enhancing the detection of marker-trait associations and improving our understanding of genetic architecture (Cortes et al., 2021; Gupta, 2021).

2.2 Role in next-generation sequencing (NGS) data analysis

Next-Generation Sequencing (NGS) technologies have revolutionized genomic research by enabling the rapid and cost-effective sequencing of entire genomes. Biostatistics plays a crucial role in the analysis of NGS data, addressing challenges such as data quality, variant calling, and interpretation of results. NGS has been applied in various domains, including clinical genomics, cancer research, and infectious disease studies, providing detailed insights into genetic variations and gene expression profiles (Satam et al., 2023). In animal breeding, NGS has enhanced our ability to understand the genetic basis of traits, facilitating the identification of genetic loci associated with economically important traits and improving breeding programs (Khanzadeh et al., 2020). The continuous advancements in NGS technology, coupled with robust biostatistical methods, are expected to further drive innovations in genomics research (Müller et al., 2018).

2.3 Statistical methods for gene expression analysis

Gene expression analysis is another critical area where biostatistics is extensively applied. Statistical methods are used to analyze data from RNA sequencing (RNA-seq) and microarray experiments, identifying differentially expressed genes and understanding gene regulatory networks. These analyses provide insights into the functional roles of genes and their involvement in various biological processes and diseases. For example, biostatistical approaches have been employed to analyze gene expression data in crops, revealing the genetic architecture of complex traits and guiding future research in functional genomics (Liu and Yan, 2018). The integration of biostatistics in gene expression analysis ensures the accurate interpretation of high-dimensional data, facilitating the discovery of novel biomarkers and therapeutic targets.

2.4 Integrating biostatistics in epigenomic studies

Epigenomic studies investigate modifications to the genome that do not involve changes in the DNA sequence but can influence gene expression and phenotype. Biostatistics is essential in analyzing epigenomic data, such as DNA methylation and histone modification patterns, to understand their impact on gene regulation and disease. The application of biostatistical methods in epigenomic studies has provided valuable insights into the mechanisms underlying complex traits and diseases, contributing to the development of precision medicine approaches. For instance, advancements in NGS have enabled the comprehensive analysis of epigenetic modifications, with biostatistics playing a pivotal role in data interpretation and the identification of epigenetic markers (Satam et al., 2023) (Figure 1). The integration of biostatistics in epigenomic research continues to enhance our understanding of gene-environment interactions and their implications for health and disease.

3 Emerging Trends and Techniques

3.1 Machine learning and artificial intelligence in genomic biostatistics

Machine learning (ML) and artificial intelligence (AI) have become indispensable tools in genomic biostatistics, driven by the need to handle and interpret vast amounts of high-throughput sequencing data. These technologies facilitate the integration and analysis of multi-omics data, enabling the discovery of new biomarkers and the development of predictive models. For instance, ML methods such as autoencoders, random forests, and support

vector machines are employed to manage the high dimensionality and complexity of omics datasets, particularly in cancer research (Reel et al., 2021; Feldner-Busztin et al., 2023). Additionally, deep learning techniques are increasingly used to predict experimental outcomes and improve the reliability of analytical workflows in proteomics (Mann et al., 2021). The integration of ML in genomic research not only enhances the understanding of biological systems but also supports precision medicine by enabling accurate disease prediction and patient stratification (Mirza et al., 2019; Reel et al., 2021).

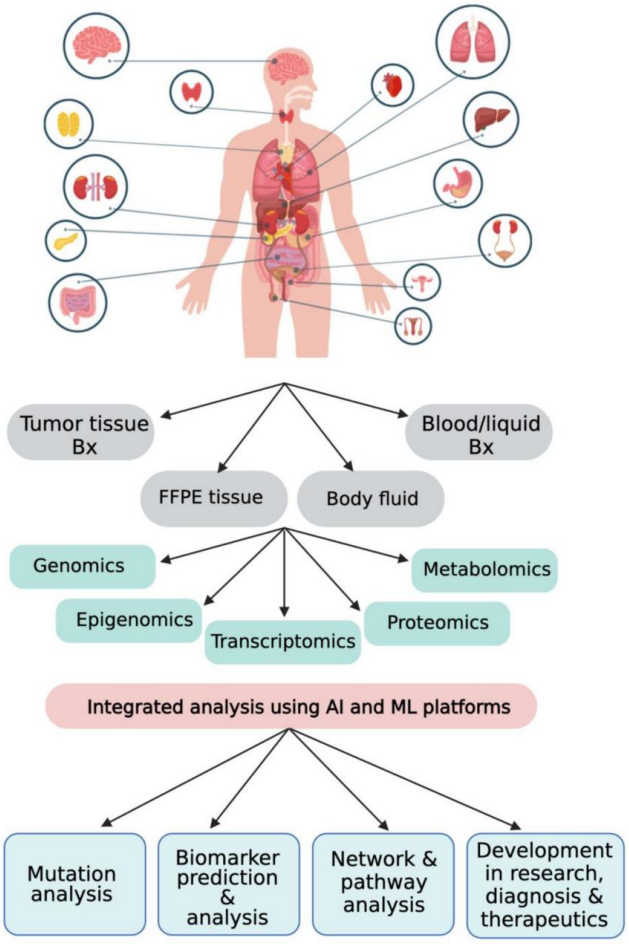

Figure 1 Role of NGS technology in cancer diagnosis, prognosis, and therapeutics using an integrative omics approach (Adopted from Satam et al., 2023)

Image caption: FFPE, formalin-fixed paraffin-embedded; Bx, biopsy; AI, artificial intelligence; Ml, machine learning (Adopted from Satam et al., 2023)

3.2 High-dimensional data analysis

The analysis of high-dimensional data is a significant challenge in genomic research due to the large number of features and relatively small sample sizes. Techniques such as dimensionality reduction are crucial for managing this complexity. Methods like principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and autoencoders are commonly used to reduce the feature space while preserving essential information (Feldner-Busztin et al., 2023). These techniques help in uncovering genotype-phenotype interactions and identifying true associations while minimizing false positives (Ritchie et al., 2015). The application of these methods is particularly evident in studies utilizing The Cancer Genome Atlas dataset, which provides a rich source of diverse experimental data (Feldner-Busztin et al., 2023).

3.3 Advances in multi-omics data integration

Multi-omics data integration is a rapidly evolving field that aims to combine data from various omics layers, such as genomics, transcriptomics, proteomics, and metabolomics, to gain a comprehensive understanding of biological

systems. Recent advances in this area include the development of integrative machine learning models that can handle heterogeneous data sources and complex biological interactions (Li et al., 2016; Nicora et al., 2020). Techniques such as network-based methods, matrix factorization, and deep neural networks are employed to fuse data from different omics layers, enabling the identification of biomarkers and the elucidation of disease mechanisms (Li et al., 2016; Mirza et al., 2019; Nicora et al., 2020). These integrative approaches are particularly valuable in oncology, where they support precision medicine by providing actionable insights for patient treatment and drug repurposing (Nicora et al., 2020).

3.4 Bayesian methods in genomic research

Bayesian methods offer a powerful framework for genomic research by incorporating prior information and modeling measurements with various distributions. These methods are particularly useful for integrating multi-view data and addressing the challenges of data heterogeneity and missing values. Bayesian models can infer direct and indirect associations in heterogeneous networks, making them suitable for complex biological data integration (Li et al., 2016). Additionally, Bayesian approaches are employed in the analysis of single-cell genomics data, where they help in trajectory inference, cell type classification, and gene regulatory network inference (Raimundo et al., 2021). The flexibility and robustness of Bayesian methods make them a valuable tool in the ongoing efforts to understand the genetic underpinnings of complex traits and diseases (Ritchie et al., 2015).

4 Challenges in Biostatistical Applications in Genomics

4.1 Handling big data and computational complexity

The rapid advancements in genomic technologies, particularly next-generation sequencing, have led to an explosion of genomic data. This data is not only vast in volume but also highly diverse, posing significant challenges in terms of computational complexity and data management. For instance, the large feature space of genome-wide data increases computational demands, making scalability a major issue. Novel approaches such as convolutional Wasserstein GANs (WGANs) and conditional RBMs (CRBMs) have been developed to address these challenges by generating high-quality artificial genomes while managing computational loads effectively (Yelmen et al., 2023). Additionally, the integration and manipulation of diverse genomic data and electronic health records (EHRs) require sophisticated Big Data analytics to uncover hidden patterns and clinically actionable insights (He et al., 2017).

4.2 Issues with data quality and standardization

The quality and standardization of genomic data are critical for reliable analysis and interpretation. High-throughput technologies generate large pools of sensitive information that are often difficult to interpret due to inconsistencies and lack of standardization. This issue is compounded by the need for sustainable infrastructure and state-of-the-art tools for efficient data management (Umbach et al., 2019). Moreover, the success of genomic research hinges on the reproducibility and interpretability of results, which are often hampered by the lack of standardized bioinformatics pipelines (Davis-Turak et al., 2017). Ensuring data quality and standardization is essential for bridging the gap between genotype and phenotype and for the effective clinical application of genomic data.

4.3 Addressing population structure and genetic diversity

Genomic data is inherently complex due to the diverse genetic backgrounds of different populations. This diversity poses challenges in accurately characterizing population structure and linkage disequilibrium, which are crucial for understanding genetic variations and their implications. Generative models like GANs and RBMs have shown promise in preserving complex characteristics of real genomes, such as population structure and selection signals, but there is still room for improvement in terms of genome quality and privacy preservation (Yelmen et al., 2023). Additionally, the state of population genetics theory needs substantial improvement to effectively handle the forensic use of genome-wide data, highlighting the need for better biostatistical modeling (Amorim and Pinto, 2018).

4.4 Ethical and privacy concerns in genomic data analysis

The sharing and analysis of genomic data raise significant ethical and privacy concerns. The potential for privacy infringement is high, given the sensitive nature of genetic information and its implications for individuals and their relatives. Effective privacy protection measures are essential to mitigate these risks. Current research highlights the major privacy threats and suggests various privacy-protection techniques for genomic data sharing, particularly in direct-to-consumer genetic testing and forensic analyses (Bonomi et al., 2020). Furthermore, the increasing commercialization of DNA technologies necessitates a security-by-design approach to protect the confidentiality, integrity, and availability of genomic data (Arshad et al., 2021). Addressing these ethical and privacy concerns is crucial for advancing genomic research within a safe and ethical framework.

5 Opportunities for Future Research and Development

5.1 Development of new statistical methods for genomic data

The rapid advancement of high-throughput technologies, such as next-generation sequencing, has resulted in the generation of vast and complex genomic datasets. Traditional bioinformatics pipelines are often insufficient to fully leverage these datasets, necessitating the development of novel statistical methods and computational paradigms. For instance, the transition from string-based to graph-based representations of reference genomes is a promising direction that could enhance the analysis of large-scale genomic data (Consortium, 2016). Additionally, the integration of machine learning techniques, particularly interpretable machine learning (iML), can help in making complex models more intelligible and actionable for genomic research (Watson, 2021) (Figure 2).

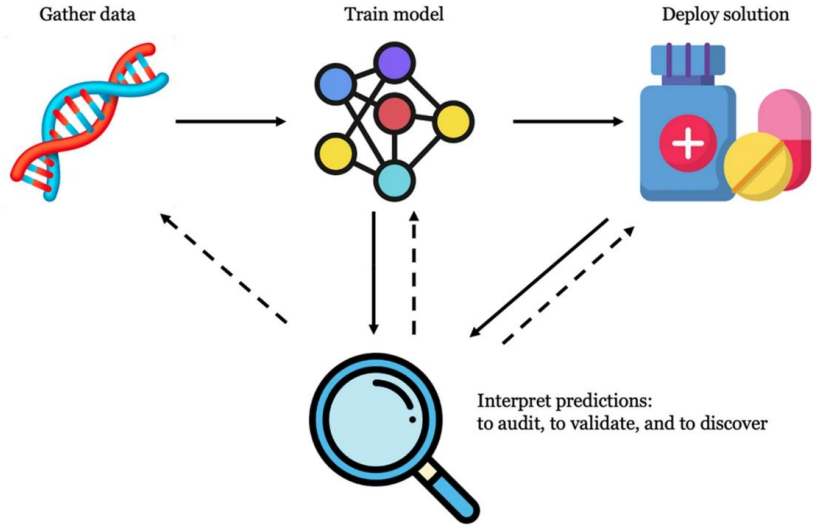

Figure 2 The classic bioinformatics workflow spans data collection, model training, and deployment. iML augments this pipeline with an extra interpretation step, which can be used during training and throughout deployment (incoming solid edges). Algorithmic explanations (outgoing dashed edges) can be used to guide new data collection, refine training, and monitor models during deployment (Adopted from Watson, 2021)

5.2 Improving the accuracy and interpretability of genomic predictions

The complexity and volume of genomic data require advanced statistical methods to improve the accuracy and interpretability of predictions. Interpretable machine learning (iML) is a burgeoning field that aims to make the predictions of machine learning models more understandable to end-users, which is crucial for the realization of precision medicine (Watson, 2021). Moreover, enhancing the interpretability of genomic predictions can aid in the timely identification and interpretation of genetic variants, which remains a significant challenge in diagnostic laboratories (Ahmed et al., 2021).

5.3 Expanding biostatistical approaches to understudied populations

Current genomic research often focuses on well-studied populations, leaving a gap in our understanding of genetic diversity across different human groups. Expanding biostatistical approaches to include understudied populations can provide a more comprehensive understanding of human genetic diversity and its implications for disease

susceptibility and treatment. This expansion is essential for the equitable application of precision medicine and public health initiatives. Additionally, addressing the unique challenges of managing and analyzing large-scale genomic data in diverse populations can further enhance the field of social science research (Liu and Guo, 2016).

5.4 Enhancing collaboration between statisticians and genomic scientists

The complexity of genomic data analysis necessitates close collaboration between statisticians and genomic scientists. Such interdisciplinary efforts can lead to the development of more robust and innovative methodologies for genomic research. For example, the integration of biostatistics with other public health domains, such as epidemiology and environmental health, can advance precision public health initiatives (Roberts et al., 2021). Furthermore, collaborative efforts can help address the challenges of data integration and management, which are critical for the effective implementation of genomic medicine (He et al., 2017).

5.5 Addressing privacy concerns in genomic data sharing

The sharing of genomic data holds great promise for advancing precision medicine, but it also raises significant privacy concerns. Developing effective privacy-protecting solutions is imperative to prevent data misuse and protect individuals' privacy. Research opportunities exist in creating robust privacy-protection techniques that can be applied to direct-to-consumer genetic testing and forensic analyses (Bonomi et al., 2020). Addressing these privacy challenges is crucial for fostering public trust and facilitating the widespread adoption of genomic data sharing practices.

6 Case Studies and Real-World Applications

6.1 Successful biostatistical applications in precision medicine

Biostatistics has played a pivotal role in the advancement of precision medicine, particularly through the integration of omics data. The rise of genomics, transcriptomics, and proteomics has enabled a more comprehensive understanding of biological systems, which is crucial for the development of precision medicine. For instance, the integration of these omics data has facilitated the identification of novel drug targets and the customization of drug treatments, thereby improving therapeutic outcomes (Manzoni et al., 2016). Additionally, the application of metabolomics in precision medicine has led to significant advancements in diagnosing diseases and understanding their mechanisms, further highlighting the importance of biostatistics in this field (Wishart, 2016).

6.2 Role of biostatistics in drug discovery and development

Biostatistics is integral to drug discovery and development, particularly through the application of pharmacogenomics. Despite the established role of pharmacogenomic variation in drug efficacy and safety, its integration into routine clinical care remains limited. Addressing this gap requires the creation of a global network of experts to drive basic pharmacogenomics research and clinical implementation. Such a network would enhance the visibility and relevance of pharmacogenomics, improve data quality, and facilitate the adoption of genomics-guided precision medicine (Chenoweth et al., 2019). Moreover, recent advances in metabolomics technologies have also contributed to drug discovery by identifying novel drug targets and monitoring therapeutic outcomes, underscoring the critical role of biostatistics in this domain (Wishart, 2016).

6.3 Biostatistics in public health genomics

The integration of biostatistics in public health genomics has the potential to revolutionize public health by improving population health outcomes. Precision public health, which combines genomics with public health concepts, aims to provide the right intervention to the right population at the right time. This approach has been particularly effective in the fight against communicable and noncommunicable diseases, as demonstrated by the application of genomic tools in tracking the origin, transmission, and evolution of the SARS-CoV-2 virus during the COVID-19 pandemic (Khoury and Holt, 2021). Furthermore, the successful implementation of newborn screening for treatable inherited conditions exemplifies the potential of biostatistics in public health genomics. However, to fully realize this potential, it is essential to address challenges such as developing policies for precision public health initiatives, improving data integration, and incorporating health equity considerations (Khoury and Holt, 2021; Roberts et al., 2021).

7 Future Directions and Vision

7.1 Predictive modeling and personalized genomics

Predictive modeling in genomics is rapidly advancing, driven by the integration of high-throughput sequencing technologies and sophisticated machine learning techniques. For instance, the development of predictive models using RNA sequencing (RNAseq) data has shown promising results in predicting disease outcomes, such as in Chronic Lymphocytic Leukemia (CLL) where models achieved up to 95% cross-validation accuracy (Kosvyra et al., 2020). Additionally, the use of genome and exome sequencing in predictive and precision medicine is gaining traction, with initiatives like Geisinger's MyCode and NHGRI's ClinSeq exploring the potential of genomic data to inform healthcare for healthy individuals (Baudhuin et al., 2019). The integration of these predictive models into clinical practice could revolutionize personalized medicine by providing tailored healthcare based on an individual's genetic profile.

7.2 Biostatistics in the era of digital and precision health

The era of digital and precision health is characterized by the convergence of genomics, big data, and advanced computational methods. Deep learning models, for example, have shown superior performance in genomics tasks, offering higher accuracies in disease prediction and treatment modeling (Koumakis, 2020). The integration of heterogeneous data sources, such as whole exome sequencing (WES) and RNAseq, into comprehensive patient profiles is another significant advancement. This approach not only summarizes large-scale datasets but also links genomic data with clinical information to build efficient predictive models (Kosvyra et al., 2019). Furthermore, the use of patient similarity networks, which cluster patients based on genomic and clinical features, represents a novel paradigm in precision medicine, enhancing both predictive performance and interpretability (Pai and Bader, 2018).

7.3 Potential impact of quantum computing on genomic biostatistics

Quantum computing holds the potential to revolutionize genomic biostatistics by providing unprecedented computational power to handle the complexity and scale of genomic data. The current challenges in genomics, such as the detailed understanding of genomic variations and their effects on disease, could be significantly mitigated by quantum computing. This technology could enhance the accuracy and speed of genomic data analysis, facilitating the discovery of new therapeutic targets and biomarkers (Chakravorty and Hegde, 2018). As quantum computing technology matures, it is expected to play a crucial role in advancing personalized medicine by enabling more precise and comprehensive genomic analyses.

8 Concluding Remarks

The field of biostatistics in genomic research is rapidly evolving, driven by advancements in high-throughput technologies and the increasing availability of large-scale genomic datasets. Key insights from the literature highlight the transformative potential of these technologies in bridging the gap between genotype and phenotype, thereby enhancing our understanding of cell biology, evolutionary history, and personalized medicine. However, the success of these technologies also brings significant challenges, particularly in data management, analysis provenance, and the reproducibility of results. Privacy concerns associated with genomic data sharing further complicate the landscape, necessitating robust privacy-protection techniques and policies.

Biostatistics plays a crucial role in addressing the challenges posed by the massive influx of genomic data. The development of novel computational methods and paradigms, such as computational pan-genomics, is essential for leveraging the full potential of these datasets. Biostatisticians are also pivotal in integrating human genomics with precision public health initiatives, which aim to improve population health through more personalized approaches. The application of single-cell genomics in cancer research exemplifies the need for advanced biostatistical methods to unravel complex clonal structures and tissue hierarchies, thereby driving progress in understanding disease mechanisms and treatment responses.

The future of biostatistics in genomic research is filled with both opportunities and challenges. On the one hand, the integration of genomics into healthcare promises to revolutionize patient care across all stages of life, from

preconception to adult medicine. On the other hand, the sheer volume and diversity of genomic data present unprecedented challenges in data analysis, requiring continuous innovation in biostatistical methodologies. Addressing these challenges will necessitate a collaborative, transdisciplinary approach that leverages the strengths of both precision medicine and public health. As we move forward, it is imperative to develop policies and frameworks that ensure the ethical use of genomic data while maximizing its potential to improve human health.

Acknowledgments

The BioSci Publisher would like to express my gratitude to the anonymous reviewers for their constructive comments and suggestions on this manuscript.

Conflict of Interest Disclosure

The author affirms that this research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Ahmed Z., Renart E., and Zeeshan S., 2021, Genomics pipelines to investigate susceptibility in whole genome and exome sequenced data for variant discovery, annotation, prediction and genotyping, PeerJ, 9: e11724.
https://doi.org/10.7717/peerj.11724

Amorim A., and Pinto N., 2018, Big data in forensic genetics, Forensic science international. Genetics, 37: 102-105.
https://doi.org/10.1016/j.fsigen.2018.08.001

Anney R.J.L., Ripke S., Anttila V., Grove J., Holmans P., Huang H., Klei L., Lee P., Medland S., Neale B., Robinson E., Weiss L., Zwaigenbaum L., Yu T., Wittemeyer K., Willsey A., Wijsman E., Werge T., Wassink T., Waltes R., Walsh C., Wallace S., Vorstman J., Vieland V., Vicente A., vanEngeland H., Tsang K., Thompson A., Szatmari P., Svantesson O., Steinberg S., Stefánsson K., Stefánsson H., State M., Soorya L., Silagadze T., Scherer S., Schellenberg G., Sandin S., Sanders S., Saemundsen E., Rouleau G., Rogé B., Roeder K., Roberts W., Reichert J., Reichenberg A., Rehnström K., Regan R., Poustka F., Poultney C., Piven J., Pinto D., Pericak-Vance M., Pejović-Milovančević M., Pedersen M., Pedersen C., Paterson A., Parr J., Pagnamenta A., Oliveira G., Nurnberger J., Nordentoft M., Murtha M., Mouga S., Mortensen P., Mors O., Morrow E., Moreno-De-Luca D., Monaco A., Minshew N., Merikangas A., McMahon W., McGrew S., Mattheisen M., Martsenkovsky I., Martin D., Mane S., Magnússon P., Magalhães T., Maestrini E., Lowe J., Lord C., Levitt P., Martin C., Ledbetter D., Leboyer M., Lecouteur A., Ladd-Acosta C., Kolevzon A., Klauck S., Jacob S., Iliadou B., Hultman C., Hougaard D., Hertz-Picciotto I., Hendren R., Hansen C., Haines J., Guter S., Grice D., Green J., Green A., Goldberg A., Gillberg C., Gilbert J., Gallagher L., Freitag C., Fombonne E., Folstein S., Fernandez B., Fallin M., Ercan-Sencicek A., Ennis S., Duque F., Duketis E., Delorme R., DeRubeis S., DeJonge M., Dawson G., Cuccaro M., Correia C., Conroy J., Conceição I., Chiocchetti A., Celestino-Soper P., Casey J., Cantor R., Café C., Bybjerg-Grauholm J., Brennan S., Bourgeron T., Bolton P., Bölte S., Bolshakova N., Betancur C., Bernier R., Beaudet A., Battaglia A., Bal V., Baird G., Bailey A., Bækvad-Hansen M., Bader J., Bacchelli E., Anagnostou E., Amaral D., Almeida J., Børglum A., Buxbaum J., Chakravarti A., Cook E., Coon H., Geschwind D., Gill M., Hallmayer J., Palotie A., Santangelo S., Sutcliffe J., Arking D., Devlin B., and Daly M., 2017, Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia, Molecular Autism, 8: 1-17.

Arshad S., Arshad J., Khan M., and Parkinson S., 2021, Analysis of security and privacy challenges for DNA-genomics applications and databases, Journal of Biomedical Informatics, 119: 103815.
https://doi.org/10.1016/j.jbi.2021.103815

Auton A., Abecasis G., Altshuler D., Durbin R., Bentley D., Chakravarti A., Clark A., Donnelly P., Eichler E., Flicek P., Gabriel S., Gibbs R., Green E., Hurles M., Knoppers B., Korbel J., Lander E., Lee C., Lehrach H., Mardis E., Marth G., McVean G., Nickerson D., Schmidt J., Sherry S., Wang J., Wilson R., Boerwinkle E., Doddapaneni H., Han Y., Korchina V., Kovar C., Lee S., and Muzny D., 2015, A global reference for human genetic variation, Nature, 526(7571): 68-74.

Baráth E., and Rosner B., 1992, Fundamentals of biostatistics, Biometrics, 48: 976.
https://doi.org/10.2307/2532371

Baudhuin L., Biesecker L., Burke W., Green E., and Green R., 2019, Predictive and precision medicine with genomic data, Clinical chemistry, 66(1): 33-41.
https://doi.org/10.1373/clinchem.2019.304345

Bonomi L., Huang Y., and Ohno-Machado L., 2020, Privacy challenges and research opportunities for genomic data sharing, Nature Genetics, 52: 646-654.
https://doi.org/10.1038/s41588-020-0651-0

Brown T.P., Rumsby P.C., Capleton A.C., Rushton L., and Levy L.S., 2006, Pesticides and Parkinson's disease--is there a link? Environ Health Perspect, 114(2): 156-164.
https://doi.org/10.1289/ehp.8095

Chakravorty S., and Hegde M., 2018, Inferring the effect of genomic variation in the new era of genomics, Human Mutation, 39: 756-773.
https://doi.org/10.1002/humu.23427

Chenoweth M., Giacomini K., Pirmohamed M., Hill S., Schaik R., Schwab M., Shuldiner A., Relling M., and Tyndale R., 2019, Global pharmacogenomics within precision medicine: challenges and opportunities, Clinical Pharmacology and Therapeutics, 107(1): 57-61.
https://doi.org/10.1002/cpt.1664

Consortium T., 2016, Computational pan-genomics: status, promises and challenges, Briefings in Bioinformatics, 19: 118-135.
https://doi.org/10.1101/043430

Cortes L., Zhang Z., and Yu J., 2021, Status and prospects of genome-wide association studies in plants, The Plant Genome, 14(1): e20077.
https://doi.org/10.1002/tpg2.20077

Davis-Turak J., Courtney S., Hazard E., Glen W., Silveira W., Wesselman T., Harbin L., Wolf B., Chung D., and Hardiman G., 2017, Genomics pipelines and data integration: challenges and opportunities in the research setting, Expert Review of Molecular Diagnostics, 17: 225-237.
https://doi.org/10.1080/14737159.2017.1282822

Duggal P., Ladd-Acosta C., Ray D., and Beaty T., 2019, The evolving field of genetic epidemiology: from familial aggregation to genomic sequencing, American Journal of Epidemiology, 188: 2069-2077.
https://doi.org/10.1093/aje/kwz193

Feldner-Busztin D., Nisantzis P., Edmunds S., Boza G., Racimo F., Gopalakrishnan S., Limborg M., Lahti L., and Polavieja G., 2023, Dealing with dimensionality: the application of machine learning to multi-omics data, Bioinformatics, 39(2): btad021.
https://doi.org/10.1093/bioinformatics/btad021

Godard B., Marshall J., and Laberge C., 2007, Community engagement in genetic research: results of the first public consultation for the Quebec CARTaGENE project, Commun Genetics, 10(3): 147-158.
https://doi.org/10.1159/000101756

Gupta P., 2021, GWAS for genetics of complex quantitative traits: Genome to pangenome and SNPs to SVs and k-mers, BioEssays, 43(11): 2100109.
https://doi.org/10.1002/bies.202100109

He K., Ge D., and He M., 2017, Big data analytics for genomic medicine, International Journal of Molecular Sciences, 18(2): 412.
https://doi.org/10.3390/ijms18020412

Khanzadeh H., Hosseinzadeh N., and Ghovvati S., 2020, Genome wide association studies, next generation sequencing and their application in animal breeding and genetics: a review, Iranian Journal of Applied Animal Science, 10: 395-404.

Khoury M., and Holt K., 2021, The impact of genomics on precision public health: beyond the pandemic, Genome Medicine, 13(1): 67.
https://doi.org/10.1186/s13073-021-00886-y

Kosvyra A., Maramis C., and Chouvarda I., 2019, Developing an integrated genomic profile for cancer patients with the use of NGS data, Emerging Science Journal, 3(3): 157-167.
https://doi.org/10.28991/esj-2019-01178

Kosvyra A., Maramis C., and Chouvarda I., 2020, A data-driven approach to build a predictive model of cancer patients' disease outcome by utilizing co-expression networks, Computers in biology and medicine, 125: 103971.
https://doi.org/10.1016/j.compbiomed.2020.103971

Koumakis L., 2020, Deep learning models in genomics; are we there yet? Computational and Structural Biotechnology Journal, 18: 1466-1473.
https://doi.org/10.1016/j.csbj.2020.06.017

Li Y., Wu F., and Ngom A., 2016, A review on machine learning principles for multi-view biological data integration, Briefings in Bioinformatics, 19: 325-340.
https://doi.org/10.1093/bib/bbw113

Liu H., and Guo G., 2016, Opportunities and challenges of big data for the social sciences: The case of genomic data, Social Science Research, 59: 13-22.
https://doi.org/10.1016/j.ssresearch.2016.04.016

Liu H., and Yan J., 2018, Crop genome-wide association study: a harvest of biological relevance, The Plant Journal, 97: 8-18.
https://doi.org/10.1111/tpj.14139

Mandrekar J., and Mandrekar S., 2009, Biostatistics: a toolkit for exploration, validation, and interpretation of clinical data, Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer, 4(12): 1447-1449.
https://doi.org/10.1097/JTO.0b013e3181c0a329

Mann M., Kumar C., Zeng W., and Strauss M., 2021, Artificial intelligence for proteomics and biomarker discovery, Cell Systems, 12(8): 759-770.
https://doi.org/10.1016/j.cels.2021.06.006

Manzoni C., Kia D., Vandrovcova J., Hardy J., Wood N., Lewis P., and Ferrari R., 2016, Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences, Briefings in Bioinformatics, 19: 286-302.
https://doi.org/10.1093/bib/bbw114

McCarthy J., McLeod H., and Ginsburg G., 2013, Genomic medicine: a decade of successes, challenges, and opportunities, Science Translational Medicine, 5: 189sr4-189sr4.
https://doi.org/10.1126/scitranslmed.3005785

Mirza B., Wang W., Wang J., Choi H., Chung N., and Ping P., 2019, Machine learning and integrative analysis of biomedical big data, Genes, 10(2): 87.
https://doi.org/10.3390/genes10020087

Müller B., Filho J., Lima B., Garcia C., Missiaggia A., Aguiar A., Takahashi E., Kirst M., Gezan S., Silva-Junior O., Neves L., and Grattapaglia D., 2018, Independent and Joint-GWAS for growth traits in Eucalyptus by assembling genome-wide data for 3373 individuals across four breeding populations, The New phytologist, 221(2): 818-833.
https://doi.org/10.1111/nph.15449

Narayan S., Liew Z., Bronstein J.M., and Ritz B., 2017, Occupational pesticide use and Parkinson's disease in the Parkinson environment gene (PEG) study, Environ Int., 107: 266-273.
https://doi.org/10.1016/j.envint.2017.04.010

Nicora G., Vitali F., Dagliati A., Geifman N., and Bellazzi R., 2020, Integrated multi-omics analyses in oncology: a review of machine learning methods and tools, Frontiers in Oncology, 10: 1030.
https://doi.org/10.3389/fonc.2020.01030

Pai S., and Bader G., 2018, Patient similarity networks for precision medicine, Journal of Molecular Biology, 430(18 Pt A): 2924-2938.
https://doi.org/10.1016/j.jmb.2018.05.037

Raimundo F., Meng-Papaxanthos L., Vallot C., and Vert J., 2021, Machine learning for single-cell genomics data analysis, Current Opinion in Systems Biology, 26: 64-71.
https://doi.org/10.1016/j.coisb.2021.04.006

Reel P., Reel S., Pearson E., Trucco E., and Jefferson E., 2021, Using machine learning approaches for multi-omics data analysis: A review, Biotechnology advances, 49: 107739.
https://doi.org/10.1016/j.biotechadv.2021.107739

Ritchie M., Holzinger E., Li R., Pendergrass S., and Kim D., 2015, Methods of integrating data to uncover genotype-phenotype interactions, Nature Reviews Genetics, 16: 85-97.
https://doi.org/10.1038/nrg3868

Roberts M., Fohner A., Landry L., Olstad D., Smit A., Turbitt E., and Allen C., 2021, Advancing precision public health using human genomics: examples from the field and future research opportunities, Genome Medicine, 13(1): 97.
https://doi.org/10.1186/s13073-021-00911-0

Satam H., Joshi K., Mangrolia U., Waghoo S., Zaidi G., Rawool S., Thakare R., Banday S., Mishra A., Das G., and Malonia S., 2023, Next-generation sequencing technology: current trends and advancements, Biology, 12(7): 997.
https://doi.org/10.3390/biology12070997

Tremblay M., and Rouleau G., 2017, Deep genealogical analysis of a large cohort of participants in the CARTaGENE project (Quebec, Canada), Ann Hum Biol., 44(4):357-365.
https://doi.org/10.1080/03014460.2017.1300326

Watson D., 2021, Interpretable machine learning for genomics, Human Genetics, 141: 1499-1513.
https://doi.org/10.1007/s00439-021-02387-9

Wishart D., 2016, Emerging applications of metabolomics in drug discovery and precision medicine, Nature Reviews Drug Discovery, 15: 473-484.
https://doi.org/10.1038/nrd.2016.32

Woodahl E.L., Lesko L.J., Hopkins S., Robinson R.F., Thummel K.E., and Burke W., 2014, Pharmacogenetic research in partnership with American Indian and Alaska Native communities, Pharmacogenomics, 15(9): 1235-1241.
https://doi.org/10.2217/pgs.14.91

Xiao Y., Liu H., Wu L., Warburton M., and Yan J., 2017, Genome-wide association studies in maize: praise and stargaze, Molecular Plant, 10(3): 359-374.
https://doi.org/10.1016/j.molp.2016.12.008

Xu H., 2020, Big data challenges in genomics, Elsevier, 43: 337-348.
https://doi.org/10.1016/bs.host.2019.08.002

Yelmen B., Decelle A., Boulos L., Szatkownik A., Furtlehner C., Charpiat G., and Jay F., 2023, Deep convolutional and conditional neural networks for large-scale genomic data generation, PLOS Computational Biology, 19(10): e1011584.
https://doi.org/10.1371/journal.pcbi.1011584

Ziegler A., König I., and Thompson J., 2008, Biostatistical aspects of genome-wide association studies, Biometrical Journal, 50(1): 8-28.
https://doi.org/10.1002/bimj.200710398