OPEN ACCESS Freely available online

ISSN: 2329-6607
Cardiovascular Pharmacology: Open Access

Commentary

Big Data Challenges in Human Genomic Research

Li Wei*

Department of Genomics Peking University, Beijing, China

DESCRIPTON

The advent of next-generation sequencing technologies has revolutionized the field of human genomics by enabling the generation of massive amounts of data at unprecedented speed and lower costs. Projects that once required years of effort and vast resources can now be completed within weeks. While this technological progress has transformed human genomics research, it has also introduced substantial challenges in the management, analysis and interpretation of data. The human genome contains more than three billion base pairs and large-scale sequencing projects generate terabytes or even petabytes of raw information. Efficient strategies for storage, processing and meaningful interpretation are therefore becoming just as critical as the sequencing itself. One of the most immediate concerns in genomic research is data quality. Sequencing technologies, despite their efficiency, are prone to various sources of error. Inaccuracies can arise during base calling, read alignment, or variant detection, which may result in false positives or missed variants. Incomplete coverage of certain genomic regions further complicates reliable analysis. To address this, quality control pipelines have been developed to filter out low-quality reads, remove contaminants and correct errors. These pipelines often require high levels of computational power and skilled expertise as even small inaccuracies can propagate into significant misinterpretations in downstream analysis. Continuous development of improved error-correction algorithms remains an area of active research.

Beyond data quality, the interpretation of genetic variants presents a significant challenge. Databases of known variants, such as dbSNP and ClinVar, provide valuable references, but many newly discovered variants lack functional annotation. This creates uncertainty about their biological significance or potential link to disease. Computational prediction methods can provide insights into whether a variant is likely to be benign or pathogenic, but these predictions are far from definitive. Experimental validation, while more reliable, is time-consuming and resource-intensive, making it difficult to scale for the millions of variants identified across populations. Population-scale projects have amplified the complexity of genomic data. As sequencing initiatives expand to include diverse global populations, researchers encounter significant variability in genetic architecture. Differences in allele frequency, haplotype structure and linkage disequilibrium patterns must be carefully accounted for when conducting variant calling or imputation. Failure to do so risks introducing biases that could lead to incomplete or inaccurate findings. Ensuring equitable representation of underrepresented populations in genomic studies is critical for avoiding disparities in both research outcomes and clinical applications.

Data sharing is another major issue in the genomic sciences. Large consortia and international collaborations often generate extensive datasets that hold immense scientific value. However, concerns surrounding privacy, consent and data security limit how widely such information can be shared. Since genomic data is inherently identifiable, protecting individual privacy is more complex than in other biomedical domains. Innovative approaches such as differential privacy, homomorphic encryption and secure multi-party computation are being developed to enable responsible data sharing without compromising personal information. The creation of federated learning frameworks, which allow models to be trained on distributed datasets without centralizing the raw data, represents another promising solution to this problem. The interpretation of genomic findings often requires integration with other types of biological and clinical data, such as transcriptomic, proteomic and epigenetic information. This integration adds another dimension of complexity, as data from different sources vary in scale, format and noise characteristics. Advanced data mining techniques, supported by high-performance computing, are essential for combining these heterogeneous data types into meaningful insights. Artificial intelligence and machine learning are increasingly applied to identify patterns across these multidimensional datasets, yet the speed at which data is generated still often outpaces our analytical capabilities.

In addition to computational challenges, sustainability and infrastructure considerations must also be addressed. The storage of genomic data requires vast amounts of digital space

Correspondence to: Li Wei, Department of Genomics Peking University, Beijing, China, E-mail: li.wei@pku.edu.cn

Received: 29-May-2025, Manuscript No. JDMGP-25-29760; **Editor assigned:** 31-May-2025, PreQC No. JDMGP-25-29760; **Reviewed:** 14-Jun-2025, QC No. JDMGP-25-29760; **Revised:** 20-Jun-2025, Manuscript No JDMGP-25-29760; **Published:** 28-Jun-2025, DOI: 10.35248/2153-0602.25.16.379

Citation: Wei L (2025). Big Data Challenges in Human Genomic Research. Journal of Data Mining in Genomics & Proteomics. 16:379.

and maintaining these resources over long periods raises both economic and environmental concerns. Cloud-based platforms have emerged as scalable alternatives to traditional local storage, allowing institutions to share resources and reduce costs. However, reliance on cloud infrastructure also raises questions of data ownership, jurisdiction and long-term accessibility. Establishing global standards and best practices for genomic data management will be vital for sustaining future research. As human genomic projects continue to expand globally, addressing the challenges posed by big data will become increasingly urgent. Improving data quality pipelines, developing reliable annotation methods, ensuring diverse population representation and implementing secure data-sharing practices are critical steps toward maximizing the utility of genomic information. At the same time, integrating genomic data with other biological and clinical information will remain a priority for translating discoveries into actionable medical benefits. While challenges remain significant, continued collaboration between biologists, data scientists, ethicists and policymakers will be central to ensuring that the potential of human genomics is fully realized.