Computational Molecular Biology 2024, Vol.14, No.4, 155-162
http://bioscipublisher.com/index.php/cmb

Review Article Open Access

Big Data in Genomics: Overcoming Challenges Through High-Performance Computing

Liting Wang ✉, Haimei Wang
Hainan Institute of Biotechnology, Haikou, 570206, Hainan, China
✉ Corresponding author: liting.wang@hitar.org
Computational Molecular Biology, 2024, Vol.14, No.4 doi: 10.5376/cmb.2024.14.0018
Received: 01 Jun., 2024
Accepted: 15 Jul., 2024
Published: 31 Jul., 2024
Copyright © 2024 Wang and Wang, This is an open access article published under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Preferred citation for this article:
Wang L.T., and Wang H.M., 2024, Big data in genomics: overcoming challenges through high-performance computing, Computational Molecular Biology, 14(4): 155-162 (doi: 10.5376/cmb.2024.14.0018)

Abstract The rapid development of genomics has brought enormous challenges in storage, management, and processing of massive genomic data. High performance computing (HPC) technology aims to address key issues in genomics big data analysis. By introducing the applications of HPC in data storage, parallel computing, sequence alignment, and assembly, this study explores how to overcome the bottleneck of data analysis using HPC technology. The focus is on the application prospects of HPC in personalized medicine, evolutionary genomics, and population genetics, and looks forward to the potential of combining quantum computing and artificial intelligence with HPC in the future. Suggestions for further optimizing the application of HPC in the field of genomics are also proposed.
Keywords Genomic big data; High-performance computing; Personalized medicine; Evolutionary genomics; Quantum computing

1 Introduction

The advent of next-generation sequencing (NGS) technologies has revolutionized the field of genomics, leading to an unprecedented explosion of genomic data. These high-throughput technologies can generate billions of short DNA or RNA fragments, resulting in datasets that can exceed several terabytes in a single run. The decreasing cost of sequencing, now around $1 000 per genome, has made large-scale genomic projects feasible, further contributing to the data deluge (Schmidt and Hildebrandt, 2017). This massive influx of data presents significant challenges in terms of storage, management, and analysis (Wong, 2018; Xu, 2020). The complexity and volume of genomic data necessitate the development of sophisticated computational tools and algorithms to extract meaningful insights (Ward et al., 2013).

High-Performance Computing (HPC) has emerged as a critical enabler in addressing the challenges posed by big data in genomics. HPC systems provide the computational power required to process and analyze large-scale genomic datasets efficiently. The integration of HPC with big data technologies, such as Apache Hadoop and cloud computing, allows for distributed and parallelized data processing, making it possible to handle petabyte-scale datasets (O'Driscoll et al., 2013). Moreover, HPC facilitates the development of advanced predictive analytics and deep learning models, which are essential for tasks such as gene prediction and the identification of genomic variants (Koumakis, 2020; Leung et al., 2020). The use of HPC in genomics not only accelerates data analysis but also enhances the accuracy and reliability of the results (Leung et al., 2020).

Regarding the current status of big data in genomics and the crucial role of high-performance computing in overcoming related challenges, we will explore various computational methods and tools developed for managing and analyzing large genomic datasets, with a focus on their success and ongoing challenges. This study will discuss the future direction and potential progress of HPC and genomics integration, emphasizing the importance of collaborative methods and improving computing infrastructure. Identify the transformative impact of HPC on genomics research and its potential to drive future discoveries in personalized medicine and other related fields.

2 The Challenges of Big Data in Genomics

2.1 Data storage and management

2.1.1 Current storage technologies

The rapid advancement in next-generation sequencing (NGS) technologies has led to an unprecedented increase in the volume of genomic data. This explosion of data presents significant challenges in terms of storage and

management. Traditional storage systems are often inadequate to handle the petabytes (PB) of data generated by high-throughput sequencing instruments. For instance, the storage of genomic data in plain text files can quickly become unmanageable due to the sheer size of the datasets, which can reach gigabytes (GB) per genome (Wong, 2018). Additionally, the need for secure storage solutions is paramount, as genomic data often contains sensitive information. Secure storage mechanisms, such as encrypted databases, have been proposed to address these concerns, offering both scalability and data protection (Wong et al., 2018).

2.1.2 Cloud computing and genomic databases

Cloud computing has emerged as a viable solution to the storage and management challenges posed by big data in genomics. Cloud platforms offer scalable storage solutions that can flexibly expand to accommodate growing datasets. For example, cloud-based genomic databases can store and manage petabytes of data, freeing researchers from the burden of maintaining physical storage infrastructure (Yang, 2019). Moreover, cloud computing facilitates the use of distributed and parallelized data processing frameworks, such as Apache Hadoop, which can efficiently handle large-scale genomic data analysis (Yeo and Crawford, 2015). These cloud-based solutions not only provide the necessary computational power but also offer cost-effective and secure storage options, making them an attractive choice for the genomics community (Tariq et al., 2020).

2.2 Data integration across platforms

One of the significant challenges in genomics is the integration of data across various platforms and technologies. Genomic data is often generated from multiple sources, including NGS, third-generation sequencing (TGS), and proteomics (Ellegren, 2014). The diversity of data types and formats complicates the integration process, making it difficult to perform comprehensive analyses. Effective data integration requires robust bioinformatics tools and platforms that can harmonize disparate datasets. For instance, proteogenomics, which combines proteomics and genomics data, faces scalability issues due to the large size of the integrated datasets. High-performance computing (HPC) solutions have been proposed to address these bottlenecks, ensuring that integrated analyses can be performed efficiently (Godhandaraman et al., 2017). Additionally, big data analytics platforms are being developed to facilitate the seamless integration and analysis of diverse genomic datasets, enabling more comprehensive and accurate genomic research (He et al., 2017).

2.3 Scalability issues in genomic research

Scalability is a critical issue in genomic research, particularly as the volume of data continues to grow exponentially. Traditional bioinformatics tools and computing infrastructures often struggle to keep pace with the increasing data demands. To address these challenges, various computational strategies have been explored, including the use of parallel distributed computing and specialized hardware (Shi and Wang, 2019). For example, the MapReduce framework, implemented on platforms like Apache Hadoop, has shown promise in scaling genomic analysis workflows, such as short read sequence alignment and assembly (Yeo and Crawford, 2015). These frameworks enable the efficient processing of large datasets by distributing the computational load across multiple nodes, thereby enhancing scalability and performance. However, the development and optimization of these scalable solutions require ongoing research and innovation to keep up with the ever-growing data landscape in genomics (Godhandaraman et al., 2017; Xu, 2020).

3 High-Performance Computing in Genomics

High-performance computing (HPC) has become indispensable in genomics due to the massive data generated by next-generation sequencing (NGS) technologies. The integration of HPC with genomics enables the efficient processing, analysis, and interpretation of large-scale genomic data, which is crucial for advancements in personalized medicine, evolutionary biology, and other fields.

3.1 Parallel computing for genomic data processing

3.1.1 Distributed algorithms for big data analysis

Distributed algorithms are essential for managing and analyzing the vast amounts of data generated in genomics. These algorithms leverage multiple computing nodes to perform tasks concurrently, enhancing both speed and efficiency. For instance, the use of Message Passing Interface (MPI) in tools like QUARTIC allows for the

alignment and sorting of high-throughput sequencing data with significant speed-ups, ensuring reproducibility and scalability (Jarlier et al., 2020). Additionally, the integration of distributed computing with GPU-based devices has shown promising results in drug discovery applications, providing cost-effective and scalable solutions (Merelli et al., 2014).

3.1.2 Cluster and supercomputing applications

Cluster and supercomputing applications are pivotal in genomics for tasks that require immense computational power. These systems utilize a network of interconnected computers to perform complex calculations at high speeds. For example, the use of multicore clusters and supercomputers has been demonstrated to improve the efficiency of distance matrix computations and sequence alignment tasks, which are fundamental in multiple sequence alignment and systems biology (Yelick et al., 2020). Moreover, the application of high-level parallel programming patterns, such as the master-worker FastFlow pattern, has been shown to enhance the performance of widely used alignment tools like Bowtie2 and BWA (Merelli et al., 2014).

3.2 Accelerating sequence alignment and assembly

Sequence alignment and assembly are critical steps in genomic analysis that benefit greatly from HPC. The development of specialized algorithms and hardware accelerators has led to significant improvements in these processes. For instance, the use of algorithm-architecture co-design has been proposed to accelerate genome analysis, integrating multiple steps of the analysis pipeline to reduce data movement and energy consumption (Mutlu and Firtina, 2023). Additionally, the implementation of parallel computing models for genome sequence alignment and preprocessing has been shown to enhance computing efficiency and scalability (Zou et al., 2021).

3.3 Role of GPUs and FPGAs in genomics

Graphics Processing Units (GPUs) and Field-Programmable Gate Arrays (FPGAs) play a crucial role in accelerating genomic computations. These hardware accelerators are designed to handle parallel tasks efficiently, making them ideal for the computationally intensive tasks in genomics.

GPUs have been successfully used to accelerate various genomic applications, such as the exploration of perturbed conditions in biological systems and the simulation of reaction-diffusion systems (Xu, 2020). The use of GPUs in deterministic systems biology simulators has achieved remarkable speed-ups, demonstrating their potential in large-scale genomic simulations. Similarly, FPGAs have been employed to enhance the performance of genome analysis pipelines, providing fast and accurate results with lower power consumption (Ward et al., 2013).

4 Overcoming Data Analysis Bottlenecks

The rapid advancement in genomics has led to an unprecedented increase in the volume and complexity of data generated. This surge has created significant bottlenecks in data analysis, necessitating the development of advanced tools and methodologies to manage and interpret large-scale genomic datasets efficiently. High-performance computing (HPC) has emerged as a critical solution to these challenges, enabling the processing of vast amounts of data in a timely and scalable manner.

4.1 Tools and pipelines for large-scale genomic analysis

The development of specialized tools and pipelines is essential for the efficient analysis of large-scale genomic data. Tools like DISSECT have been designed to leverage distributed-memory parallel computational architectures, significantly reducing the time required for complex genomic analyses. For instance, DISSECT can analyze simulated traits from 470 000 individuals in approximately four hours using 8 400 processor cores, achieving high prediction accuracies (Canela-Xandri et al., 2015). Similarly, platforms like DolphinNext offer a modular approach to building and deploying complex workflow (Figure 1), ensuring flexibility, portability, and reproducibility in high-throughput data processing (Yukselen et al., 2019; Yukselen et al., 2020). These tools address the need for scalable and efficient data processing frameworks, which are crucial for handling the growing volume of genomic data.

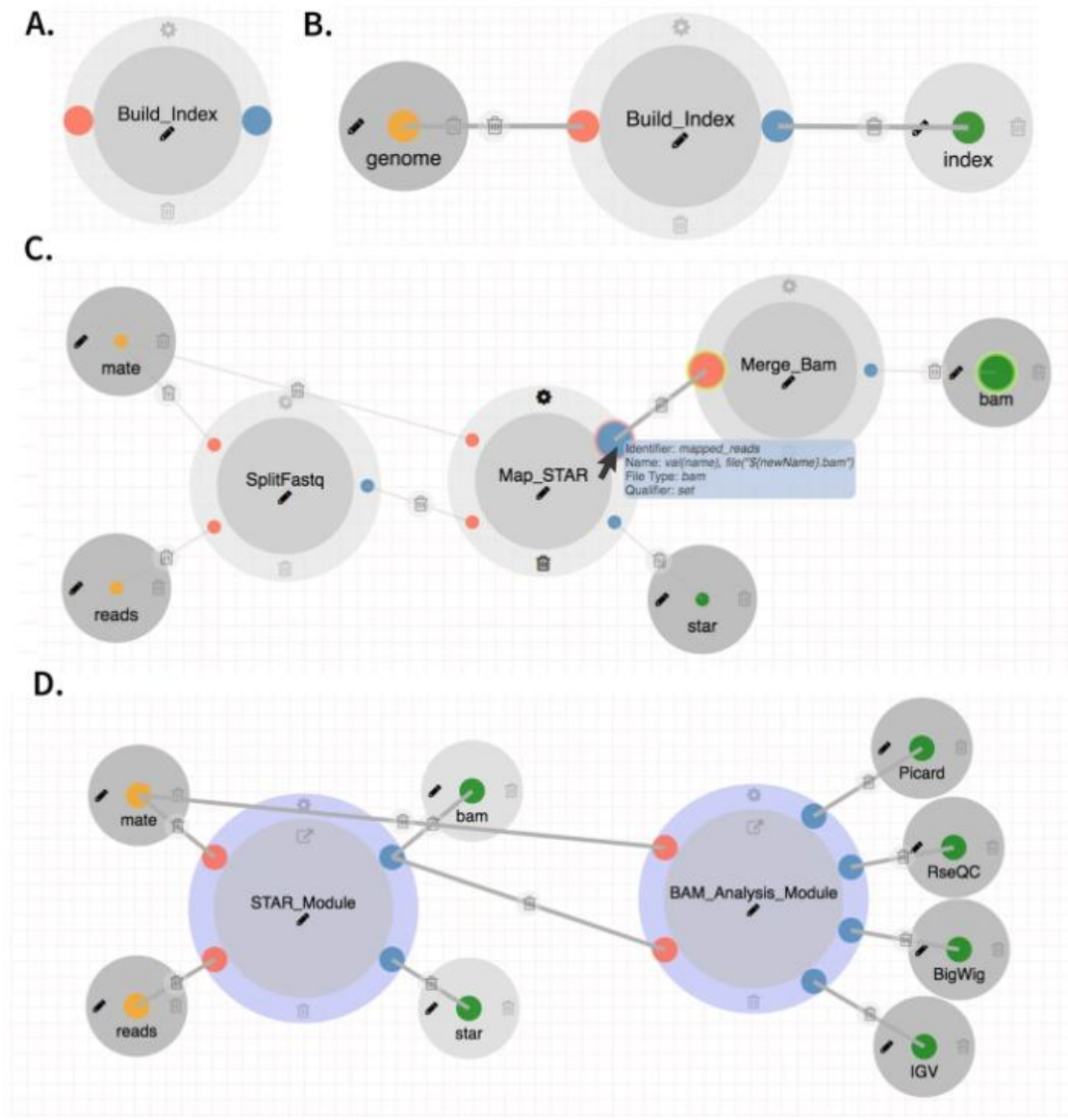

Figure 1 A process for building index files b Input and output parameters attached to a process c The STAR alignment module connected through input/output with matching parameter types. d The RNA-Seq pipeline can be designed using two nested pipelines: the STAR pipeline and the BAM analysis pipeline (Adopted from Yukselen et al., 2019)

4.2 Integration of machine learning with HPC

Machine learning (ML) has become an invaluable asset in genomics, offering powerful techniques for analyzing large and complex datasets. The integration of ML with HPC can further enhance the efficiency and accuracy of genomic analyses. ML algorithms can assist in various tasks, such as annotating sequence elements and analyzing epigenetic, proteomic, or metabolomic data (Libbrecht and Noble, 2015). By leveraging HPC resources, these algorithms can process large datasets more quickly and accurately, facilitating the discovery of novel insights and patterns in genomic data (Fu et al., 2024). This integration is particularly beneficial for tasks that require extensive computational power, such as predictive modeling and data mining.

4.3 Data security and privacy concerns in genomics

As genomic data becomes increasingly integral to personalized medicine and clinical applications, ensuring data security and privacy is paramount. The integration of diverse genomic data with electronic health records (EHRs) poses significant challenges in terms of data manipulation, management, and analysis (He et al., 2017). It is crucial to implement robust security measures to protect sensitive information from unauthorized access and breaches. Additionally, maintaining data privacy while enabling data sharing and collaboration among researchers

is essential for advancing genomic research. Addressing these concerns requires a combination of technological solutions, such as secure data storage and transmission protocols, and regulatory frameworks that govern data usage and access (Huttenhower and Hofmann, 2010). Overcoming data analysis bottlenecks in genomics requires a multifaceted approach that includes the development of advanced tools and pipelines, the integration of machine learning with high-performance computing, and stringent data security and privacy measures. By addressing these challenges, researchers can unlock the full potential of genomic data, paving the way for significant advancements in biological research and personalized medicine.

5 Applications of High-Performance Computing in Genomics

5.1 Personalized medicine and genomic diagnostics

High-performance computing (HPC) plays a crucial role in personalized medicine and genomic diagnostics by enabling the analysis of large-scale genomic data to identify clinically actionable genetic variants. The integration of next-generation sequencing (NGS) data with electronic health records (EHRs) allows for the development of individualized diagnostic and therapeutic strategies. For instance, big data analytics can uncover hidden patterns and correlations within vast datasets (Figure 2), facilitating the identification of genetic markers relevant to specific diseases and tailoring treatments accordingly (Tariq et al., 2020; Xu, 2020). The decreasing cost of sequencing, now around $1000 per genome, has made large population-scale projects feasible, further enhancing the precision of personalized medicine (Davis-Turak et al., 2017).

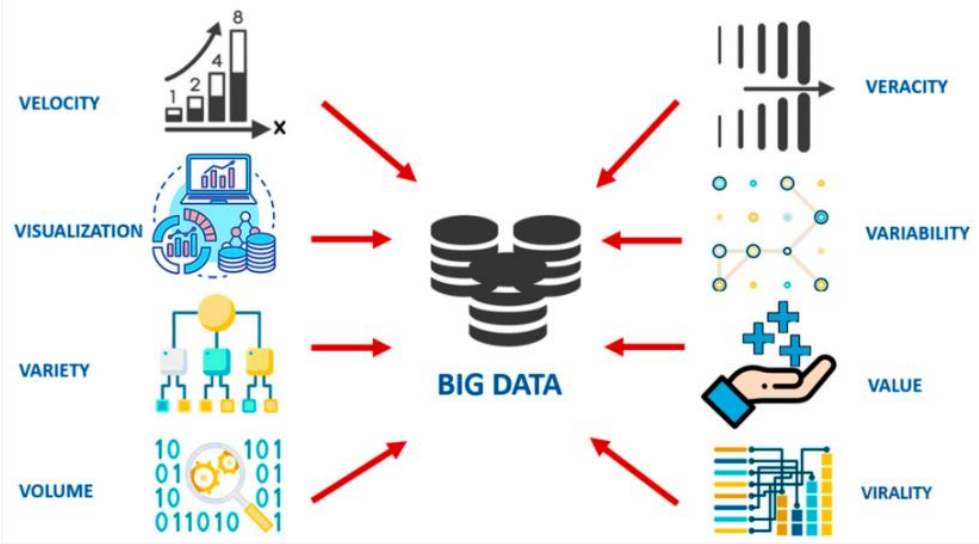

Figure 2 Representation of distinct dimensions of big data (Adopted from Hassan et al., 2022)

5.2 Evolutionary genomics and comparative studies

HPC is indispensable in evolutionary genomics and comparative studies, where the analysis of massive genomic datasets is required to understand evolutionary relationships and genetic diversity. The ability to process and analyze large-scale genomic data sets, such as those generated by the Human Genome Project, has provided insights into the evolutionary history of species and the genetic basis of adaptation (Miller et al., 2017). The use of distributed and parallelized data processing technologies, such as Apache Hadoop, allows researchers to handle petabyte-scale data efficiently, facilitating comprehensive comparative genomic analyses (Hassan et al., 2022).

5.3 Genomic data in population genetics and epidemiology

In population genetics and epidemiology, HPC enables the analysis of extensive genomic data to study genetic variation within and between populations, as well as the spread of genetic traits and diseases. The integration of genomic data with epidemiological data helps in understanding the genetic factors contributing to disease susceptibility and resistance, as well as tracking the spread of infectious diseases (Ward et al., 2013; Xu, 2020). The development of robust computational infrastructure and collaborative approaches is essential for making sense of the data and delivering actionable insights in these fields.

6 Future Trends and Technologies

6.1 Quantum computing in genomic research

Quantum computing holds the potential to revolutionize genomic research by providing unprecedented computational power to handle the massive datasets generated by next-generation sequencing technologies. Traditional computing methods struggle with the complexity and volume of genomic data, but quantum computing can offer solutions through its ability to perform complex calculations at significantly faster rates. This can lead to more efficient data analysis, enabling researchers to uncover new genetic insights and accelerate the development of personalized medicine (Stephens et al., 2015; Schmidt and Hildebrandt, 2017).

6.2 Integration of artificial intelligence with genomic big data

The integration of artificial intelligence (AI) with genomic big data is another promising trend. AI algorithms, particularly machine learning and deep learning, can analyze vast amounts of genomic data to identify patterns and correlations that might be missed by human researchers. This can enhance the accuracy of genetic predictions and the identification of disease biomarkers. AI can also streamline the data processing pipeline, making it easier to manage and interpret large datasets, which is crucial for advancing genomic medicine and personalized healthcare (Shi and Wang, 2019; Xu, 2020).

6.3 Real-time genomic data analysis and streaming

Real-time genomic data analysis and streaming represent a significant advancement in the field of genomics. The ability to analyze and stream data in real-time can facilitate immediate decision-making in clinical settings, such as during surgeries or in the diagnosis of genetic disorders. This requires robust high-performance computing infrastructure and sophisticated algorithms capable of handling continuous data flow without compromising accuracy. The development of such technologies will be essential for the future of real-time genomic applications (Godhandaraman et al., 2017; Maia et al., 2017).

7 Concluding Remarks

High-performance computing (HPC) has become indispensable in the field of genomics, addressing the challenges posed by the exponential growth of biological data. Current HPC solutions leverage the power of supercomputers, computer clusters, and parallel processing techniques to manage and analyze massive datasets efficiently. For instance, platforms like the one described in provide scalable and reconfigurable HPC infrastructures that significantly accelerate genomic sequencing and protein structure analysis. Similarly, the use of Hadoop clusters for parallel processing, as demonstrated in, showcases the benefits of HPC in speeding up data analysis tasks that would otherwise be infeasible on traditional computing systems. Moreover, the integration of HPC with big data technologies, such as Apache Spark and MPI, has shown promising results in metagenomics, offering faster and more memory-efficient solutions compared to traditional methods.

Despite the advancements, several challenges remain in the realm of big data genomics. One of the primary issues is the scalability of current HPC systems to handle the ever-increasing volume of data generated by next-generation sequencing technologies. As highlighted in, the transition to exascale computing systems presents both opportunities and challenges, requiring new design and implementation strategies to exploit their full potential. Additionally, the complexity of biological data necessitates the development of more sophisticated algorithms and tools that can efficiently process and analyze these datasets. The integration of proteogenomics data, for example, still faces significant scalability bottlenecks that need to be addressed. Furthermore, ensuring the robustness, stability, and maintainability of HPC systems, especially in shared environments, remains a critical concern.

To effectively implement HPC in genomic research, several recommendations can be made. First, it is essential to develop scalable and reconfigurable HPC platforms that can adapt to the growing data demands and provide efficient data analysis capabilities. Second, leveraging parallel processing frameworks like Hadoop and Spark can significantly enhance the performance of bioinformatics algorithms, but it is crucial to address their scalability limitations and optimize memory usage. Third, the transition to exascale computing should be accompanied by the development of new software solutions that can fully utilize the computational power of these systems while

maintaining ease of use. Finally, fostering collaboration between biologists and computer scientists is vital to ensure that the developed HPC solutions are tailored to the specific needs of genomic research and can provide profound insights into biological functions.

Acknowledgments

The guidance and feedback from Professor Wendy Yang were instrumental to this work. I also thank the anonymous reviewers for their insightful comments and suggestions.

Conflict of Interest Disclosure

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

References

Canela-Xandri O., Law A., Gray A., Woolliams J., and Tenesa A., 2015, A new tool called DISSECT for analysing large genomic data sets using a big data approach, Nature Communications, 6(1): 10162.
https://doi.org/10.1038/ncomms10162.

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.

Ellegren H., 2014, Genome sequencing and population genomics in non-model organisms, Trends in Ecology and Evolution, 29(1): 51-63.
https://doi.org/10.1016/j.tree.2013.09.008.

Fu J., Hong Z.M., and Huang W.Z., 2024, Harnessing genomic tools for Cassava improvement: advances and prospects, Bioscience Evidence, 14(1): 32-38.
https://doi.org/10.5376/be.2024.14.0005

Godhandaraman T., Pruthviraj N., Praveenkumar V., Banuprasad A., and Karthick K., 2017, Application of cloud computing in biomedicine big data analysis cloud computing in big data, 2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET), IEEE, 2017: 1-3.
https://doi.org/10.1109/ICAMMAET.2017.8186739.

Hassan M., Awan F.M., Naz A., deAndrés-Galiana E.J., Álvarez Ó., Cernea A., Fernández-Brillet L., Fernández-Martínez J., and Kloczkowski A., 2022, Innovations in genomics and big data analytics for personalized medicine and health care: a review, International Journal of Molecular Sciences, 23(9): 4645.
https://doi.org/10.3390/ijms23094645.

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

Huttenhower C., and Hofmann O., 2010, A quick guide to large-scale genomic data mining, PLoS Computational Biology, 6(5): e1000779.
https://doi.org/10.1371/journal.pcbi.1000779.

Jarlier F., Joly N., Fedy N., Magalhaes T., Sirotti L., Paganiban P., Martin F., McManus M., and Hupé P., 2020, Quartic: QUick pArallel algoRithms for high-throughput sequencing data processing, F1000 Research, 9: 240.
https://doi.org/10.12688/f1000research.22954.2.

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.

Leung C.K., Sarumi O.A., and Zhang C.Y., 2020, Predictive analytics on genomic data with high-performance computing, 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2020: 2187-2194.
https://doi.org/10.1109/BIBM49941.2020.9312982.

Libbrecht M., and Noble W., 2015, Machine learning applications in genetics and genomics, Nature Reviews Genetics, 16: 321-332.
https://doi.org/10.1038/nrg3920.

Maia A.T., Sammut S.J., Jacinta-Fernandes A., and Chin S., 2017, Big data in cancer genomics, Current Opinion in Systems Biology, 4: 78-84.
https://doi.org/10.1016/J.COISB.2017.07.007.

Merelli I., Pérez-Sánchez H., Gesing S., and D'Agostino D., 2014, High-performance computing and big data in omics-based medicine, BioMed Research International, 2014: 2014.
https://doi.org/10.1155/2014/825649.

Miller M., Zhu C., and Bromberg Y., 2017, Clubber: removing the bioinformatics bottleneck in big data analyses, Journal of Integrative Bioinformatics, 14(2): 20170020.
https://doi.org/10.1515/jib-2017-0020.

Mutlu O., and Firtina C., 2023, Invited: accelerating genome analysis via algorithm-architecture co-design, 2023 60th ACM/IEEE Design Automation Conference (DAC), 2023: 1-4.
https://doi.org/10.1109/DAC56929.2023.10247887.

O'Driscoll A., Daugelaite J., and Sleator R., 2013, 'Big data' hadoop and cloud computing in genomics, Journal of Biomedical Informatics, 46(5): 774-81.
https://doi.org/10.1016/j.jbi.2013.07.001.

Schmidt B., and Hildebrandt A., 2017, Next-generation sequencing: big data meets high performance computing, Drug Discovery Today, 22(4): 712-717.
https://doi.org/10.1016/j.drudis.2017.01.014.

Shi L.Z., and Wang Z., 2019, Computational strategies for scalable genomics analysis, Genes, 10(12): 1017.
https://doi.org/10.3390/genes10121017.

Stephens Z.D., Lee S.Y., Faghri F., Campbell R.H., Zhai C., Efron M., Iyer R., Schatz M., Sinha S., and Robinson G., 2015, Big data: astronomical or genomical, PLoS Biology, 13(7): e1002195.
https://doi.org/10.1371/journal.pbio.1002195.

Tariq M., Haseeb M., Aledhari M., Razzak R., Parizi R., and Saeed F., 2020, Methods for proteogenomics data analysis challenges and scalability bottlenecks: a survey, IEEE Access : Practical Innovations Open Solutions, 9: 5497-5516.
https://doi.org/10.1109/ACCESS.2020.3047588.

Ward R., Schmieder R., Highnam G., and Mittelman D., 2013, Big data challenges and opportunities in high-throughput sequencing, Systems Biomedicine, 1: 29-34.
https://doi.org/10.4161/sysb.24470.

Wong K.C., 2018, Big data challenges in genome informatics, Biophysical Reviews, 11(1): 51-54.
https://doi.org/10.1007/s12551-018-0493-5.

Xu H.Y., 2020, Big data challenges in genomics, Handbook of Statistics, Elsevier, 43: 337-348.
https://doi.org/10.1016/BS.HOST.2019.08.002.

Yang J.T., 2019, Cloud computing for storing and analyzing petabytes of genomic data, Journal of Industrial Information Integration, 15: 50-57.
https://doi.org/10.1016/J.JII.2019.04.005.

Yelick K., Buluç A., Awan M., Azad A., Brock B., Egan R., Ekanayake S., Ellis M., Georganas E., Guidi G., Hofmeyr S., Selvitopi O., Teodoropol C., and Oliker L., 2020, The parallelism motifs of genomic data analysis, Philosophical Transactions, Series A Mathematical Physical and Engineering Sciences, 378(2166): 20190394.
https://doi.org/10.1098/rsta.2019.0394.

Yeo H., and Crawford C.H., 2015, Big data: cloud computing in genomics applications, 2015, IEEE International Conference on Big Data (Big Data), 2015: 2904-2906.
https://doi.org/10.1109/BigData.2015.7364117.

Yukselen O., Turkyilmaz O., Ozturk A., Garber M., and Kucukural A., 2019, Dolphinnext: a distributed data processing platform for high throughput genomics, BMC Genomicsm, 21: 1-16.
https://doi.org/10.1186/s12864-020-6714-x.

Zou Y., Zhu Y.J., Li Y.H., Wu F.X., and Wang J.X., 2021, Parallel computing for genome sequence processing, Briefings in Bioinformatics, 22(5): bbab070.
https://doi.org/10.1093/bib/bbab070.

Disclaimer/Publisher's Note

The statements, opinions, and data contained in all publications are solely those of the individual authors and contributors and do not represent the views of the publishing house and/or its editors. The publisher and/or its editors disclaim all responsibility for any harm or damage to persons or property that may result from the application of ideas, methods, instructions, or products discussed in the content. Publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.