ARTICLE

Genomic data, effective communication through visualization

Raj Rajeshwar Malinda[1],*, Dipika Mishra[2]

[1]Indraprastha Institute of Information Technology Delhi, New Delhi 110020, India.
[2]Independent researcher, Odisha 766001, India.

ABSTRACT　　　　Genomic data are inherently multidimensional and complex, therefore, presenting researchers with significant challenges in analysis and interpretation. Data visualization of genomic datasets can unravel the complexity and provide meaningful insights for effective communication. Here, we discuss that, in data-driven genomic studies, effective storytelling of formulated hypotheses can be significantly enhanced by using suitable data visualization tools. Further, with the ongoing advancement of technology, we argue that, the integration of these tools with artificial intelligence or machine learning concepts could potentially revolutionize the visualization trends within the field of genomic research.
　　　　　　　　　　　　　　　　　　　　　　　Acta Biol Szeged 68(2):88-91 (2024)

Unfolding the genomic data complexity

Genomic data, considered as blueprints of life, are more complex and mysterious in the way they hold information. The vast and intricate datasets generated within the domain of genomics, through various high-throughput technologies, are overwhelming for researchers (Tanjo et al. 2021; Wong 2019). The complexity of genomic data comprises multidimensional layers such as sequencing, variations, epigenetic modifications, spatial and temporal data, and therefore, the correct method of analysis and interpretation is critical and remains a challenge for researchers (The 1000 Genomes Project Consortium 2015; Li et al. 2012; Wojcik et al. 2019). Data visualization (DataViz) emerges as one of the essential tools to unravel the complexity of genetic and genomic data and convey meaningful insights through various graphical representations (Brehmer and Munzner 2013; Durant et al. 2022; Munzner 2014; Nielsen et al. 2010; Nusrat et al. 2019; O'Donoghue 2021). In data-driven studies, including in the field of genomics, the primary goal remains the effective communication of research hypothesis, and therefore, a clear interpretation is necessary when extracting meaningful insights from complex datasets (Nielsen et al. 2010; Wong 2012; Wong 2019). Transiting the intricate nature of genomic datasets through relevant visualization enhances the visibility of research (Meyer et al. 2012; Munzner 2014; Nielsen et al. 2010). Furthermore, we emphasize that, genomic studies are interdisciplinary in nature and involve active participation from other fields, including molecular biology, biochemistry, statistics, and computer sciences, and therefore, associating with DataViz professionals would bring a community effort towards better learning and understanding of genomic research. In this article, we strive to highlight the significance of data visualization tools in elucidating the complexity of genomic datasets. We further emphasize the use of AI-powered integration in these DataViz tools for enhanced representation of complexity contained within the genomic research.

Use of data visualization tools

Massive data from studies, such as Next-Generation Sequencing (NGS) (Abbasi and Masoumi 2020), Genome-wide Association Studies (GWAS) (Uffelmann et al. 2021), and Comparative and Single-Cell Genomics (Genereux et al. 2020), significantly generate millions of genetic variations from interpretation. Genomic data and visualization interests overlap and actively engage in elucidating and simplifying complex data structures (Durant et al. 2022; Nusrat et al. 2019). Some significant examples include the Human Genome Project (Nurk et al. 2022), and sequencing of the human Y chromosome (Rhie et al. 2023). High-dimensional genomic studies are generally challenging to understand through textual representations. Therefore, selecting an appropriate visualization method becomes a significant characteristic when communicating findings to the research community and other engaged audiences,

such as science journalists and communicators (Nusrat et al. 2019; Parsons 2022). Visualization tools are critical to stay on course, and play a substantial role in observing correlations, patterns, trends, or any hidden messages in massive datasets, and serve as an effective means of communication. For example, identifying genes of interest through visualizing tools is also effective and can further assist in associating their functions in physiological and pathological conditions (Qu et al. 2019).

Some of the currently available major and popular visualization tools for genetics and genomic studies, which have been mostly used by the genomic researchers for decades (Goodstadt and Marti-Renom 2017; Krzywinski et al. 2009; Pearce et al. 2019; Wang et al. 2018; Yokoyama and Kasahara 2020). However, we highlight some of the recently developed data visualization tools or software applications and their appropriate use for genomic datasets. In this list, CoolBox for multi-omics data (Xu et al. 2021), Gosling for interactive visualization (L'Yi et al. 2022), plotsr helps to visualize various structural similarities and rearrangements across multiple genome databases (Goel and Schneeberger 2022), ggmsa explores multiple sequence alignment data visualization (Zhou et al. 2022), GenoVi can be used in bacterial and archaeal genome (Cumsille et al. 2023), and VAG (Visualizing read alignments in graph genomes) (Li et al. 2023), are some of the examples.

DataViz platforms are now essential for minimizing efforts and facilitating meaningful interpretation. As mentioned, these tools are helpful to understand and visually represent data in various contexts and create visualizations (Goodstadt and Marti-Renom 2017; Pearce et al. 2019). However, it is certainly debatable whether a certain level of software skills, in terms of programming language knowledge, coding efficiency, and basic statistical competency are required to visualize biological data, including genomic data. Although several newly developed tools offer some relief in coding, or understanding file formats and are accessible through a pre-build user interface to minimize effort and maximize productivity (Langer et al. 2023; Tumescheit et al. 2022), however, they may limit the options for researchers in terms of potential outcomes. It is notable that, as the technology progresses, these tools may offer integrated more user-friendly interfaces and layouts for analyzing genomic data (Langer et al. 2023). Indeed, genomic data, with its interpretation, are quite complex and dynamic to comprehend, and therefore, automated visuals can facilitate a better understanding. Communicating research findings in a cognizable manner is also a significant component of DataViz to enhance the knowledge of complex research by the common audience. Furthermore, we stress that, integrating AI-based concepts into such tools could revolutionize the visualizing

efforts and trends, as would demand minimum efforts, and yield maximum outcomes in data-driven genomic research (Wang et al. 2023).

Integration of AI-powered concepts in data visualization tools

One recent advancement is artificial intelligence (AI), which makes most performative tasks easier, manageable, comprehensible, and accessible in research data analysis. AI-based applications aid in conducting influential and significant analyses of gene-based databases. In structural biology, AlphaFold, which is based on AI, is considered a breakthrough for predicting protein's structure in 3-dimensional space and has gained significant attention from researchers (Jumper et al. 2021). In genomic studies for clinical researchers, similar tools might be still in progress, however, an adopted version called AlphaMissense can predict the pathological conditions linked to genetic alterations (Cheng et al. 2023). Here, we mentioned some of visualization tools or software for genomic data integrated with the power of AI technology to enhance performance and capabilities. Among the others, DeepGene for identifying disease-causing genetic variants (Yuan et al. 2016), DeepChrome for predicting genetic modifications (Singh et al. 2016), and DeepTorrent predict DNA N4-methylcytosine sites (Liu et al. 2021), are based on the deep learning concept. These AI-featured tools or software may not entirely be dedicated to visualization standalone but combined to extract the complex phenomenon of genomic datasets. We signify the use of AI-driven DataViz tools, as it reduces the major burden among researchers those are not familiar with traditional visualizing tools (Wang et al. 2023). These AI-adopted tools are not only convenient for skilled researchers, but also for beginner-level and enthusiastic researchers who aim to convey their ideas convincingly through visualization. Nonetheless, researchers are benefiting from AI-based applications for identifying, characterizing, and analyzing genetic and genomic data. In this regard, AI-driven DataViz tools tailored to genomic data may find acceptance among researchers more in the future. Furthermore, the prediction of new targets for drug delivery and other therapeutics developments would represent a new interplay between AI technology and genomic data analysis (Guo et al. 2023).

Conclusion

In conclusion, the visualization of gene-associated data has emerged as a more effective, communicative, and

better method of storytelling for the audience. DataViz in genetics and genomics enhances clarity by transforming complexity into an understandable knowledge of the field. The integration and handling of massive and intricate genetic data necessitates the development of more compelling visualization tools that are yet to reach their full potential. A clear and concise approach to addressing the research question would facilitate a better comprehension of DataViz. The incorporation of AI-based and machine-learning techniques can effectively handle the complex and enormous amount of datasets that are being generated on a daily basis, and with convenience of researcher's learnings (Malinda, 2023). In conjunction with other practical tools for communicating research findings in genetics and genomics, DataViz would constitute a significant addition to the toolkit for effectively conveying information to the research community.

References

Abbasi S, Masoumi S (2020) Next-generation sequencing (NGS). Inter J Adv Sci Tech 29(3):6364-6377.

The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. Nature 526:68-74.

Brehmer M, Munzner T (2013) A multi-level typology of abstract visualization tasks. IEEE Trans Visual Comp Graph 19(12):2376-2385.

Cheng J, Novati G, Pan J, Bycroft C, Akvilė Žemgulytė A, et al. (2023) Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science 381(6664):eadg7492.

Cumsille A, Durán RE, Rodríguez-Delherbe A, Saona-Urmeneta V, Cámara B, et al. (2023) GenoVi, an open-source automated circular genome visualizer for bacteria and archaea. PLoS Comput Biol 19(4):e1010998.

Durant E, Rouard M, Ganko EW, Muller C, Cleary AM, et al. (2022b) Ten simple rules for developing visualization tools in genomics. PLOS Comput Biol 18(11):e1010622.

Genereux DP, Serres A, Armstrong J, et al. (2020) A comparative genomics multitool for scientific discovery and conservation. Nature 587(7833):240-245.

Goel M, Schneeberger K (2022) plotsr: visualizing structural similarities and rearrangements between multiple genomes. Bioinformatics 38(10):2922-2926.

Goodstadt M, Marti-Renom MA (2017) Challenges for visualizing three-dimensional data in genomic browsers. FEBS Letters 591(17):2505-2519.

Guo K, Wu M, Soo Z, Yang Y, Zhang Y, Zhang Q, Lin H, Grosser M, Venter D, Zhang G, Lu J (2023) Artificial intelligence-driven biomedical genomics. Knowledge-Based Sys 279:110937.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, et al. (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596(7873):583-589.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19(9):1639-1645.

Langer CCH, Mitter M, Stocsits RR, Gerlich DW (2023) HiCognition: a visual exploration and hypothesis testing tool for 3D genomics. Genome Biol 24(1):158.

Li F, Hu H, Xiao Z, Wang J, Liu J et al. (2023) Visualization and review of reads alignment on the graphical pan-genome with VAG. Preprint BioRxiv:2023.01.20.524849.

Li W, Zhang S, Liu CC, Zhou XJ (2012) Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. Bioinformatics 28(19):2458-2466.

Liu Q, Chen J, Wang Y, Li S, Jia C, Song J, Li F (2021) DeepTorrent: a deep learning-based approach for predicting DNA N4-methylcytosine sites. Brief Bioinform 22(3):bbaa124.

L'Yi S, Wang Q, Lekschas F, Gehlenborg N (2022) Gosling: a grammar-based toolkit for scalable and interactive genomics data visualization. IEEE Trans Vis Comput Graph 28(1):140-150.

Malinda RR (2023) Biological data studies, scale-up the potential with machine learning. Euro J Hum Genet 31(6):619-620.

Meyer M, Sedlmair M, Munzner T (2012) The four-level nested model revisited: blocks and guidelines. ACM Inter Conf Proc Series (11):1-6.

Munzner T (2014) Visualization Analysis and Design. CRC Press, New York.

Nielsen CB, Cantor M, Dubchak I, Gordon D, Wang T (2010) Visualizing genomes: techniques and challenges. Nat Methods 7(3):S5-S15.

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, et al. (2022) The complete sequence of a human genome. Science 376(6588):44-53.

Nusrat S, Harbig T, Gehlenborg N (2019) Tasks, techniques, and tools for genomic data visualization. Comput Graph Forum 38(3):781-805.

O'Donoghue SI (2021) Grand challenges in bioinformatics data visualization. Front Bioinform 1:669186.

Parsons P (2022) Understanding data visualization design practice. IEEE Trans Vis Comp Graph 28(1):665-675.

Pearce TM, Nikiforova MN, Roy S (2019) Interactive browser-based genomics data visualization tools for translational and clinical laboratory applications. J Mol Diagn 21(6):985-993.

Qu Z, Lau CW, Nguyen QV, Zhou Y, Catchpoole DR (2019) Visual analytics of genomic and cancer data: A Systematic Review. Cancer Inform (18):1176935119835546.

Rhie A, Nurk S, Cechova M et al. (2023) The complete sequence of a human Y chromosome. Nature 621(7978):344-354.

Singh R, Lanchantin J, Robins G, Qi Y (2016) DeepChrome: deep-learning for predicting gene expression from histone modifications. Bioinformatics 32(17):i639-i648.

Tanjo T, Kawai Y, Tokunaga K, Ogasawara O, Nagasaki M (2021) Practical guide for managing large-scale human genome data in research. J Hum Genet 66(1):39-52.

Tumescheit C, Firth AE, Brown K (2022) CIAlign: a highly customisable command line tool to clean, interpret and visualise multiple sequence alignments. PeerJ 10:e12983.

Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, Martin HC, Lappalainen T, Posthuma D (2021) Genome-wide association studies. Nat Rev Methods Primers 1:59.

Wang X, Wu Z, Huang W, Wei Y, Huang Z, Xu M, Chen W (2023) VIS+AI: integrating visualization with artificial intelligence for efficient data analysis. Front Comput Sci 17:176709.

Wang Y, Song F, Zhang B, Zhang L, Xu J, et al. (2018) The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. Genome Biol 19(1):151.

Wojcik GL, Graff M, Nishimura KK, Tao R, Haessler J, et al. (2019) Genetic analyses of diverse populations improves discovery for complex traits. Nature 570(7762):514:518.

Wong B (2012) Visualizing biological data. Nat Methods 9:1131.

Wong KC (2019) Big data challenges in genome informatics. Biophys Rev 11(1):51-54.

Xu W, Zhong Q, Lin D, Zuo Y, Dai J, Li G, Cao G (2021) CoolBox: a flexible toolkit for visual analysis of genomics data. BMC Bioinform 22(1):489.

Yokoyama TT, Kasahara M (2020) Visualization tools for human structural variations identified by whole-genome sequencing. J Hum Genet 65(1):49-60.

Yuan Y, Shi Y, Li C, Kim J, Cai W, Han Z, Feng DD (2016) Deepgene: An advanced cancer type classifier based on deep learning and somatic point mutations. BMC Bioinform 17(Suppl 1):476.

Zhou L, Feng T, Xu S, Gao F, Lam TT et al. (2022) ggmsa: A visual exploration tool for multiple sequence alignment and associated data. Brief Bioinform 23(4)bbac222.