Effective storytelling of genomic datasets through visualization

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

[1]PGD Data Science, Health and Climate Change for Social Impact, Indraprastha Institute of Information Technology, New Delhi – 110020, India.

[2]School of Biological Sciences, University of Edinburgh, United Kingdom.

[#]Corresponding author: contact@rrmalinda.com

Abstract

Genomic data are inherently multidimensional and complex, therefore, presenting researcher 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.

Keywords: Data visualization, Genetics, Genomics, DataViz, AI Tools.

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; K. C. 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 (Auton et al., 2015; W. 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 & Munzner, 2013; Durant et al., 2022a; Munzner T., 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; B. Wong, 2012; K. C. 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 and better representation of complexity contained within the genomic research.

Use of data visualization tools

Massive data from studies, such as Next-Generation Sequencing (NGS) (Abbasi & 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., 2022b; 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 & Marti-Renom, 2017; Krzywinski et al., 2009; Pearce et al., 2019; Y. Wang et al., 2018; Yokoyama & 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 & 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) (F. Li et al., 2023), are some of the examples.

As we understand, the massive datasets demand a more convincing organizational approach, making further analysis more straightforward to navigate in the ocean of unorganized datasets, and these tools facilitate such conditions. 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 & 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 (X. 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, such kind of similar tools may 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 (X. 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.

Author Contributions

RRM: Conceptualization, and Writing – original and edited drafts. DM: Writing – Original and edited drafts.

Funding

None.

Competing Interests

The authors declare no competing interests.

References

Abbasi, S., & Masoumi, S. (2020). Next-generation sequencing (NGS). *International Journal of Advanced Science and Technology*, *29*(3). https://doi.org/10.1007/978-981-16-1037-0_23

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., … Schloss, J. A. (2015). A global reference for human genetic variation. In *Nature* (Vol. 526, Issue 7571). https://doi.org/10.1038/nature15393

Brehmer, M., & Munzner, T. (2013). A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, *19*(12). https://doi.org/10.1109/TVCG.2013.124

Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulyte, A., Applebaum, T., Pritzel, A., Wong, L. H., Zielinski, M., Sargeant, T., Schneider, R. G., Senior, A. W., Jumper, J., Hassabis, D., Kohli, P., & Avsec, Ž. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, *381*(6664). https://doi.org/10.1126/science.adg7492

Cumsille, A., Durán, R. E., Rodríguez-Delherbe, A., Saona-Urmeneta, V., Cámara, B., Seeger, M., Araya, M., Jara, N., & Buil-Aranda, C. (2023). GenoVi, an open-source

automated circular genome visualizer for bacteria and archaea. *PLoS Computational Biology*, *19*(4). https://doi.org/10.1371/JOURNAL.PCBI.1010998

Durant, E., Rouard, M., Ganko, E. W., Muller, C., Cleary, A. M., Farmer, A. D., Conte, M., & Sabot, F. (2022a). Ten simple rules for developing visualization tools in genomics. *PLoS Computational Biology*, *18*(11). https://doi.org/10.1371/journal.pcbi.1010622

Durant, E., Rouard, M., Ganko, E. W., Muller, C., Cleary, A. M., Farmer, A. D., Conte, M., & Sabot, F. (2022b). Ten simple rules for developing visualization tools in genomics. *PLOS Computational Biology*, *18*(11), e1010622. https://doi.org/10.1371/journal.pcbi.1010622

Genereux, D. P., Serres, A., Armstrong, J., Johnson, J., Marinescu, V. D., Murén, E., Juan, D., Bejerano, G., Casewell, N. R., Chemnick, L. G., Damas, J., Di Palma, F., Diekhans, M., Fiddes, I. T., Garber, M., Gladyshev, V. N., Goodman, L., Haerty, W., Houck, M. L., … Karlsson, E. K. (2020). A comparative genomics multitool for scientific discovery and conservation. *Nature*, *587*(7833). https://doi.org/10.1038/s41586-020-2876-6

Goel, M., & Schneeberger, K. (2022). plotsr: visualizing structural similarities and rearrangements between multiple genomes. *Bioinformatics*, *38*(10), 2922–2926. https://doi.org/10.1093/bioinformatics/btac196

Goodstadt, M., & Marti-Renom, M. A. (2017). Challenges for visualizing three-dimensional data in genomic browsers. In *FEBS Letters* (Vol. 591, Issue 17). https://doi.org/10.1002/1873-3468.12778

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 Systems*, *279*, 110937. https://doi.org/10.1016/j.knosys.2023.110937

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873). https://doi.org/10.1038/s41586-021-03819-2

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., & Marra, M. A. (2009). Circos: An information aesthetic for comparative genomics. *Genome Research*, *19*(9). https://doi.org/10.1101/gr.092759.109

Langer, C. C. H., Mitter, M., Stocsits, R. R., & Gerlich, D. W. (2023). HiCognition: a visual exploration and hypothesis testing tool for 3D genomics. *Genome Biology*, *24*(1). https://doi.org/10.1186/s13059-023-02996-9

Li, F., Hu, H., Xiao, Z., Wang, J., Liu, J., Zhao, D., Fu, Y., Wang, Y., Yuan, X., Bu, S., Zhou, X., Zhao, J., & Wang, S. (2023). Visualization and review of reads alignment on the graphical pan-genome with VAG. *BioRxiv*. https://doi.org/10.1101/2023.01.20.524849

Li, W., Zhang, S., Liu, C. C., & Zhou, X. J. (2012). Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics*, *28*(19). https://doi.org/10.1093/bioinformatics/bts476

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. *Briefings in Bioinformatics*, *22*(3). https://doi.org/10.1093/bib/bbaa124

L'Yi, S., Wang, Q., Lekschas, F., & Gehlenborg, N. (2022). Gosling: A Grammar-based Toolkit for Scalable and Interactive Genomics Data Visualization. *IEEE Transactions on Visualization and Computer Graphics*, *28*(1). https://doi.org/10.1109/TVCG.2021.3114876

Malinda, R. R. (2023). Biological data studies, scale-up the potential with machine learning. In *European Journal of Human Genetics* (Vol. 31, Issue 6). https://doi.org/10.1038/s41431-023-01361-5

Meyer, M., Sedlmair, M., & Munzner, T. (2012). The four-level nested model revisited: Blocks and guidelines. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/2442576.2442587

Munzner, T. (2014). Visualization Analysis & Design. In *Visualization Analysis and Design*. https://doi.org/10.1201/b17511

Munzner T. (2014). *Visualization analysis and design.* CRC Press.

Nielsen, C. B., Cantor, M., Dubchak, I., Gordon, D., & Wang, T. (2010). Visualizing genomes: Techniques and challenges. In *Nature Methods* (Vol. 7, Issue 3). https://doi.org/10.1038/nmeth.1422

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks,

S. Y., … Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, *376*(6588). https://doi.org/10.1126/science.abj6987

Nusrat, S., Harbig, T., & Gehlenborg, N. (2019). Tasks, techniques, and tools for genomic data visualization. *Computer Graphics Forum*, *38*(3). https://doi.org/10.1111/cgf.13727

O'Donoghue, S. I. (2021). Grand Challenges in Bioinformatics Data Visualization. *Frontiers in Bioinformatics*, *1*. https://doi.org/10.3389/fbinf.2021.669186

Parsons, P. (2022). Understanding Data Visualization Design Practice. *IEEE Transactions on Visualization and Computer Graphics*, *28*(1). https://doi.org/10.1109/TVCG.2021.3114959

Pearce, T. M., Nikiforova, M. N., & Roy, S. (2019). Interactive Browser-Based Genomics Data Visualization Tools for Translational and Clinical Laboratory Applications. *Journal of Molecular Diagnostics*, *21*(6). https://doi.org/10.1016/j.jmoldx.2019.06.005

Qu, Z., Lau, C. W., Nguyen, Q. V., Zhou, Y., & Catchpoole, D. R. (2019). Visual Analytics of Genomic and Cancer Data: A Systematic Review. In *Cancer Informatics* (Vol. 18). https://doi.org/10.1177/1176935119835546

Rhie, A., Nurk, S., Cechova, M., Hoyt, S. J., Taylor, D. J., Altemose, N., Hook, P. W., Koren, S., Rautiainen, M., Alexandrov, I. A., Allen, J., Asri, M., Bzikadze, A. V., Chen, N. C., Chin, C. S., Diekhans, M., Flicek, P., Formenti, G., Fungtammasan, A., … Phillippy, A. M. (2023). The complete sequence of a human Y chromosome. *Nature*, *621*(7978). https://doi.org/10.1038/s41586-023-06457-y

Singh, R., Lanchantin, J., Robins, G., & Qi, Y. (2016). DeepChrome: Deep-learning for predicting gene expression from histone modifications. *Bioinformatics*, *32*(17). https://doi.org/10.1093/bioinformatics/btw427

Tanjo, T., Kawai, Y., Tokunaga, K., Ogasawara, O., & Nagasaki, M. (2021). Practical guide for managing large-scale human genome data in research. *Journal of Human Genetics*, *66*(1), 39–52. https://doi.org/10.1038/s10038-020-00862-1

Tumescheit, C., Firth, A. E., & Brown, K. (2022). CIAlign: A highly customisable command line tool to clean, interpret and visualise multiple sequence alignments. *PeerJ*. https://doi.org/10.7717/peerj.12983

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. In *Nature Reviews Methods Primers* (Vol. 1, Issue 1). Springer Nature. https://doi.org/10.1038/s43586-021-00056-9

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. *Frontiers of Computer Science*, *17*(6). https://doi.org/10.1007/s11704-023-2691-y

Wang, Y., Song, F., Zhang, B., Zhang, L., Xu, J., Kuang, D., Li, D., Choudhary, M. N. K., Li, Y., Hu, M., Hardison, R., Wang, T., & Yue, F. (2018). The 3D Genome Browser: A web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biology*, *19*(1). https://doi.org/10.1186/s13059-018-1519-9

Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., Highland, H. M., Patel, Y. M., Sorokin, E. P., Avery, C. L., Belbin, G. M., Bien, S. A., Cheng, I., Cullina, S., Hodonsky, C. J., Hu, Y., Huckins, L. M., Jeff, J., Justice, A. E., … Carlson, C. S. (2019). Genetic analyses of diverse populations improves discovery for complex traits. *Nature*, *570*(7762). https://doi.org/10.1038/s41586-019-1310-4

Wong, B. (2012). Visualizing biological data. *Nature Methods*, *9*(12). https://doi.org/10.1038/nmeth.2258

Wong, K. C. (2019). Big data challenges in genome informatics. In *Biophysical Reviews* (Vol. 11, Issue 1). https://doi.org/10.1007/s12551-018-0493-5

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 Bioinformatics*, *22*(1). https://doi.org/10.1186/s12859-021-04408-w

Yokoyama, T. T., & Kasahara, M. (2020). Visualization tools for human structural variations identified by whole-genome sequencing. In *Journal of Human Genetics* (Vol. 65, Issue 1). https://doi.org/10.1038/s10038-019-0687-0

Yuan, Y., Shi, Y., Li, C., Kim, J., Cai, W., Han, Z., & Feng, D. D. (2016). Deepgene: An advanced cancer type classifier based on deep learning and somatic point mutations. *BMC Bioinformatics*, *17*. https://doi.org/10.1186/s12859-016-1334-9

Zhou, L., Feng, T., Xu, S., Gao, F., Lam, T. T., Wang, Q., Wu, T., Huang, H., Zhan, L., Li, L., Guan, Y., Dai, Z., & Yu, G. (2022). Ggmsa: A visual exploration tool for multiple sequence alignment and associated data. *Briefings in Bioinformatics*, *23*(4). https://doi.org/10.1093/bib/bbac222