Valorizing ‘Omics Visualization for Discovery

Abstract

Scientists from diverse backgrounds are joining the field of data science. This leads to advances in data science being actualized in the context of many different domains. Conclusions from datasets using innovative algorithms are obvious aspects but advances in data science can take on many different forms such as new methods for data interpretation, new data integration and processing technologies, or as will be the topic of this editorial, data visualization techniques. The parity and complementary relationship between techniques from all domains provide ways to improve discovery although quantifying the contributions to discovery process from each technique can be elusive. The experiences described here come from a visualizing life science multi-omics data, but most of the remarks can be associated with visualization methods in general. From the perspective that visualization serves as an important method for shaping data science interpretations, this paper sets out some of the difficulties encountered in creating and valorizing new visualization implementations for scientific discovery from multi-omics datasets.

Benefits of visualization

All fields and domains require the use and analysis of data; however, not all domain experts are statisticians or algorithm experts. The omics technologies (genomics, transcriptomics, proteomics, metabolomics, lipidomics, etc) have generated many multifactorial experiments that necessitate effective visual exploration by life science experts to successfully extract knowledge [CITATION Sch13 \l 4108] [CITATION Vis \l 4108] [CITATION Exp \l 4108]. The challenge in practical terms is how to present the data at the right level of detail, in a cohesive, insightful manner. In general, transforming spreadsheet data into visual representations can facilitate new knowledge discovery [CITATION Tho \l 4108]. The discovery often comes from seeing novel and unexpected patterns in datasets by visually interpreting data in a different way. As there is only limited utility in seeing the expected, one often seeks out outliers, oddities, unusual events and patterns, places where the data do not match expectations [CITATION Coo07 \l 4108].

Human working memory has limited capacity and transient storage properties for simultaneous interpretation of multiple hypotheses and huge amounts of evidence linked together by numerous relationships [CITATION Neu \l 4108]. Data for biological systems are organized as complex networks of molecular and functional interactions making the intuitive interpretation of multi-omics datasets difficult without help. Visual displays provide a method to extend the working memory capacity by establishing a placeholder for information patterns [CITATION War12 \l 4108]. More evidence can be viewed in concert. Research can advance more quickly if the barrier to the effective exploration by any scientists is minimized. Therefore, insights from the emerging field of visual analytics [CITATION Kei08 \l 4108], which specifically studies the role of visualization in the larger process of understanding and interpreting data, can bear significant rewards. Visual analytics methods have begun to be applied to studying the connection between visualization and analytical reasoning in systems biology [CITATION Kei08 \l 4108] [CITATION Geh10 \l 4108].

Characteristics of the data

In the field of multi-omics data assemblage and evaluation, common data characteristics surface; the complexity of the data is related to multidimensionality and multivariate nature, where variance in the measurements can be attributed to other numerous explanatory variables and possible confounders. The data complexity and the multitude of questions to be addressed means static visualization is often insufficient. The user needs to explore the data interactively in order to assess a wide range of questions. In addition to the high dimensionality of the data, information overload, data interconnectivity, and pattern extraction pose major hurdles to developing effective visualizations [CITATION Ogh16 \l 4108]. Here, one of the main difficulties lies in the design of graphical layouts that contain the complete information space[CITATION Sel16 \l 4108], although there are implementations, for example in variant genomics, that understand and elegantly address these issues [CITATION Fer13 \l 4108].

For intuitiveness and usefulness, it is likely that there is no single generic layout that will cover the requirements needed to answer the range of biological questions. Often the better-known representations, bar and pie charts, histograms, line and scatter plots are used to carry out simple statistical visualization and to report trends and summaries[CITATION App11 \l 4108]. Node-link tree and graph visualizations, in both 2D and 3D [CITATION Fre07 \l 4108], can display hierarchically structured data such as ontologies and networks. Other visualization methods that have been tested successfully, but typically incorporate just one omics type, include; heat maps and matrices[CITATION Cyd12 \l 4108]; parallel co-ordinates [CITATION Ins85 \l 4108]; timeline and topology plots[CITATION Rin11 \l 4108]; map and landscape views that build on the metaphor of cartography; space-filling visualizations such as tree maps, hive plots [CITATION Krz12 \l 4108], icicle, bubble and sunburst plots [CITATION Glu16 \l 4108]; iconography, including star and glyph plots [CITATION Vil15 \l 4108] [CITATION Car99 \l 4108][CITATION Hee10 \l 4108][CITATION Tuf01 \l 4108]. Specific use cases for high-dimensional data may require visualization such as parallel co-ordinates, while pie charts and scatter plots can be used for associated clinical variables to examine only a small number of dimensions simultaneously. Most novel visualization applications often employ or build on some of the simpler, well-known techniques that are organized together in innovative combinations. Overall, the choice of visualization for multi-omics data needs to reflect the complex organization of biological phenomena and, importantly, the user must have their own internal representation of the biological phenomena in order to reason about it while exploring the data. In general, experts will have built up from extensive experience a set of patterns for exploring the important elements found in their data and these must be taken into account when providing a visualization [CITATION Sac14 \l 4108].

Building visualizations

Modelling how a scientist thinks about biology plays a big role on how people interpret and interact with an interface. The application of human–computer interaction (HCI) methods enables a process approach to solve the difficult problems of omics visualization. Scientists want to answer questions with their datasets. While detecting trends is important, ultimately researchers want to see the causal relationships of how A has an effect on B. To address these knowledge discovery needs appropriately, it is useful to understand the current discussions pertaining to design study methodology. Sedlmair et al., [CITATION Sed12 \l 4108] propose a clear definition of design studies as well as practical guidance for conducting them effectively. They stress the need to understand the contributions design studies can make to visualizations, when design studies are the appropriate method to use, and how design studies are unique from other approaches. Following on from this, design studies should strive to understand

the life scientists' usage of multi-omics as applied to a specific real-world problem, validate the visualization design to confirm that it addresses the problem, and then reflect about process in order to refine visualization design guidelines. Based on the design study, it is possible to identify critical areas that are the most important with respect to user issues and plan a research agenda to pursue the most effective solutions. Frequently to be effective, visualizations benefit from a combination of problem-solving research and technique driven research. Although, when the validation criterion depends on the calculating the new knowledge derived due to the application of a visualization tool, measuring the impact can be elusive.

Quantifying visualization in the scientific discovery process

The power and value of visualization is often described by its ability to foster insight into and improve understanding of data, which then should lead to enabling intuitive, effective knowledge discovery and analytical activity. This can partly be achieved by removing the cognitive load encountered in managing the large amounts of complex, heterogeneous data, which are commonly delivered by multiple omics experiments [CITATION And12 \l 4108]. More challenging is that knowledge discovery is seldom an instantaneous event, but requires studying and manipulating the data repetitively from multiple perspectives and possibly using multiple tools. Streamlining repetitive tasks may be a benefit that is linked to discovery but the contribution of this may not be easily traceable back to the visualization. The introduction of data visualization tool may trigger changes in work practices, exacerbating the problem of identifying their contribution to discovery. One measure of success for a visualization could be that users can formulate and answer questions they didn't anticipate before looking at the visualization [CITATION Pla04 \l 4108]. If users need to look at the same data from different perspectives and over a long time, they must be motivated and actively intellectually engaged in experimenting with the visualization tool [CITATION Ise11 \l 4108]. Conducting longitudinal studies that record each and every finding by the users over a longer period of time to see how visualization tools influence knowledge acquisition can be very valuable[CITATION Sar04 \l 4108] [CITATION Per09 \l 4108]. These studies should be conducted with scientists analyzing their own experimental results for the first time. Several studies [CITATION Rie96 \l 4108][CITATION Sar051 \l 4108] [CITATION Kob01 \l 4108][CITATION Ger07 \l 4108] have conducted such longitudinal studies with evaluations that included frequent user interviews, diary studies, and 'Eureka' reports. Overall, measuring the impact of visualizations on discovery is a difficult task but a range of evaluation methods are being tested to measure success [CITATION Pla04 \l 4108].

Users adopt applications that have intuitive interfaces and deliver appropriate context and personalization via a rich end-user interaction. This usually means that the application has been perfectly simplified. The tasks being performed via the interface are streamlined. Irrelevant features or uncertainty does not distract user focus over where to click for the information for answering the next question. Real-time interactive features bring engaging, time-sensitive, or contextual biological information to the forefront [CITATION Gon03 \l 4108]. The mental model that users build up whilst interacting feels natural to the way they think without realizing it. Creating this type of visualization takes time, much trial and error, and an attention to psychological as well as the scientific detail. Measuring these attributes has been a current focus in evaluation practices [CITATION Lam12 \l 4108].

Finally, Dörk et al., [CITATION Dör13 \l 4108], have outlined an approach for HCI that promotes; disclosure of bias and decisions made about the visualization (disclosure), the enabling of multiple interpretations (plurality), a range of possible ways to interact with the visualization (contingency), and

allowing users to derive their own hypotheses (empowerment). The principles of disclosure and plurality largely address insight by promoting comprehensible representations, while contingency and empowerment are guiding principles driving impact through flexible interactions and empowering user experiences [CITATION Dör13 \l 4108].

Bias as a confounding issue

As with any domain of data science, visualizations are to some extent subjective and interpretive. No visualization captures all aspects of a particular dataset from all possible perspectives. Each visualization encompasses some assumptions of the developer and it is important to avoid potentially biasing users with a particular line of thought [CITATION Tor04 \l 4108]. With high dimensional data there may be many reasonable approaches to analyzing it. The scientist's perception is biased towards interpretation of information into existing (internal) models of biology and existing expectations. However, human reasoning is subject to a variety of well-documented heuristics and biases [CITATION Tve74 \l 4108] that cause people to deviate from how they should rationally make decisions. Therefore, a major challenge to any scientist is to be open to new and important insights while simultaneously avoiding being misled by the tendency to see structure in randomness and to find meaningful patterns in meaningless noise, such that confirmation bias leads to false conclusions [CITATION Mun17 \l 4108]. There appears to be little guidance and material that teaches people how to do actual exploratory analysis work[CITATION Whi06 \l 4108], let alone with an understanding of their biases. People are fixated with complex statistical models and blindly applying machine learning to data problems when in fact what we need to improve and perfect is our ability to reason with data and make rational decisions under conditions of uncertainty. Complementarily, visualizations are challenged to incorporate a notion of confidence or certainty because the factors that influence the certainty or uncertainty of data vary with the type of information and the type of decisions being made [CITATION Tho05 \l 4108]. Statisticians see the world in the light of confirmatory analysis and regard exploration as an inferior approach to analysis. Visualization researchers, too busy building innovative implementations to cope with the new data overload, have done little to teach users how to run actual data exploration methods. Part of the solution to this conundrum may depend on the visualization researchers adopting the philosophy that their implementations must teach as well as systematically guide exploratory data analysis in ways that make the process as effective, reliable, and rational as possible.

Visualization as a valuable asset to be rewarded

As discussed above, many aspects must be taken into consideration when developing an interface. A good multidimensional omics visualization tool must maximize simplicity, familiarity, intuitiveness, effectiveness, data correctness [CITATION Ber11 \l 4108] as well as minimize bias from both the developer and end user. Even when doing all this, visualization tools can be overlooked and not interpreted as a valuable publishable scientific effort in the context of data science. Clearly, visualizations are necessary for the adoption, use, and efficacy of uptake of computational methods in data science. Major efforts have been made in recent years to create visualization tools that can extract useful knowledge from the vast amount of data generated by high-throughput technologies [CITATION Sch13 \l 4108][CITATION Vis \l 4108][CITATION Exp \l 4108]. However, more progress is required to create new tools to meet the changing needs of the field. Incremental improvements of visualization software is highly important, but requires great effort from developers for low scientific reward when compared to the development of new methods. There must be acknowledgement that the investment to

the study and effort dedicated to the development and maintenance of new tools, as well as user training and support, will be adequately compensated to encourage advancement of the field. Long-term investment and funding are needed to guarantee the maintenance, improvement, and evolution of visualization tools beyond their first publication[CITATION Sch13 \l 4108].

Conclusion

As the size and complexity of omics datasets continues to increase, the development of user interfaces and interaction techniques that expedite the process of exploring that data must receive new attention. Novel approaches also need to take into consideration the technological challenges and opportunities given by new interaction contexts, ranging from mobile, touch [CITATION Kee10 \l 4108][CITATION Ise11 \l 4108], and gesture interaction to visualizations on large displays, and encompassing highly responsive web applications. Regardless of the speed of rendering and context, it is important to coherently organize the visual process of exploration to give insight about the data to a user and address psychological aspects of the user experience. Measures to access impact of visualizations remain a challenge [CITATION Nor06 \l 4108] and so it follows valorization may not be proportional to the effort put in for development. Overall, to quote Nils Gehlenborg [CITATION Geh10 \l 4108]: “The challenge is to create clear, meaningful and integrated visualizations that give biological insight, without being overwhelmed by the intrinsic complexity of the data”.

Acknowledgement

I would like to thank the reviewers Alexander Lex and Rafael Martins, and the editor Tobias Kuhn for their helpful comments, which have contributed to an improved and contemporaneous manuscript.

References

[1] M. P. Schroeder, A. Gonzalez-Perez and N. Lopez-Bigas, "Visualizing multidimensional cancer genomics data," *Genome Medicine*, vol. 5, 2013.

[2] G. A. Pavlopoulos, D. Malliarakis, N. Papanikolaou, T. Theodosiou, A. J. Enright and I. Iliopoulos, "Visualizing genome and systems biology: technologies, tools, implementation techniques and trends, past, present and future," *Gigascience*, vol. 4, no. 1, pp. 1-27, 2015.

[3] W. Dunn, A. Burgun, M.-O. Krebs and B. Rance, "Exploring and visualizing multidimensional data in translational research platforms," *Brief Bioinform*, 2016.

[4] J. J. Thomas and K. A. Cook, "A visual analytics agenda," *Computer Graphics and Applications, IEEE*, vol. 26, no. 1, pp. 10-13, 2006.

[5] K. Cook, R. Earnshaw and J. Stasko, "Guest editors’ introduction: Discovering the unexpected," *Computer Graohics and Applications, IEEE*, vol. 27, no. 5, pp. 15-19, 2007.

[6] E. K. Vogel and T. Möller, "Neural activity predicts individual differences in visual working memory capacity," *Nature*, vol. 428, pp. 748-751.

[7] C. Ware, "Information visualization: perception for design," 2012.

[8] D. Keim, G. Andrienko, J. D. Fekete, C. Görg, J. Kohlhammer and G. Melançon, "Visual Analytics: Definition, Process, and Challenges," in *Information Visualization: Human-Centered Issues and Perspectives*, A. Kerren, J. Stasko, J. D. Fekete and C. North, Eds., Berlin Heidleberg, Springer, 2008, pp. 154-175.

[9] N. Gehlenborg, S. I. O'Donoghue, N. S. Baliga and A. Goesmann, "Visualization of omics data for systems biology," *Nature Methods*, vol. 7, no. 3 Suppl, pp. S56-68, 2010.

[10] M. Oghbaie, M. J. Pennock and W. B. Rouse, "Understanding the efficacy of interactive visualization for decision making for complex systems," *Systems Conference (SysCon) Annual IEEE*, pp. 1-6, 2016.

[11] H. X. Self, J. Zeitz, L. House, S. Leman and C. North, "Designing usable interactive visual analytics tools for dimension reduction," *Human Centered Machine Learning at CHI*, 2016.

[12] J. A. Ferstay, C. B. Nielsen and T. Munzner, "Variant View: Visualizing Sequence Variants in their Gene Context," *IEEE Transactions on Visualization & Computer Graphics*, vol. 19, no. 12, pp. 2546-2555, 2013.

[13] A. S. Dadzie and M. Rowe, "Approaches to visualizing linked data: A survey," *Semantic Web*, vol. 2, no. 2, pp. 89-124, 2011.

[14] F. TC, G. L, v. D. S. Brosch M, M. P, G. RJ, F. S, T. J and E. AJ, "Construction, visualisation, and clustering of transcription networks from microarray expression data," *PLoS Comput Biol.*, vol. 3, no. 10, pp. 2032-2042, 2007.

[15] C. Nielsen and B. Wong, "Points of view: Managing deep data in genome browsers," *Nature Methods*, vol. 9, p. 512, 2012.

[16] A. Inselberg, "The plane with parallel coordinates," *The Visual Computer*, vol. 1, pp. 69-91, 1985.

[17] A. Rind, W. Aigner, S. Miksch, S. Wiltner, M. Pohl, T. Turic and F. Drexler, "Visual Exploration of Time-Oriented Patient Data for Chronic Diseases: Design Study and Evaluation.," in *Lecture Notes in Computer Science,*, Berlin, Heidelberg, Springer, 2011, pp. 301-320.

[18] K. M, B. I, J. SJ and M. MA, "Hive plots – rational approach to visualizing networks," *Brief Bioinform.*, vol. 13, no. 5, p. 627–644, 2012.

[19] M. Glueck, P. Hamilton, F. Chevalier, S. Breslav, A. Khan, D. Wigdor and M. Brudno, "PhenoBlocks: Phenotype Comparison Visualizations," *IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS*, vol. 22, no. 1, pp. 101-110, 2016.

[20] J. M. Villaveces, P. Koti and B. H. Habermann, "Tools for visualization and analysis of molecular networks, pathways, and -omics data," *Adv Appl Bioinform Chem*, vol. 8, p. 11–22, 2015.

[21] S. K. Card, J. D. Mackinlay and B. Shneiderman, Reading in information visualization, Morgan Kaufmann Publishers, Inc., 1999.

[22] J. Heer, M. Bostock and V. Ogievetsky, "A tour through the visualization zoo," *Communications of the ACM*, vol. 53, no. 6, pp. 59-67, 2010.

[23] E. Tufte, The visual display of quantitative information, Cheshire, CT: Graphics Press, 2001.

[24] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis and D. A. Keim, "Knowledge Generartion Model for Visual Analytics," in *IEEE Transactions on Visualization and Computer Graphics*, 2014, pp. 1604-1613.

[25] M. Sedlmair, M. Meyer and T. Munzner, "Design study methodology: Reflections from the trenches and the stacks.," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, p. 2431–2440, 2012.

[26] E. W. Anderson, "Evaluating Scientific Visualization Using Cognitive Measures," *BELIV Workshop: Beyond Time and Errors-Novel Evaluation Methods for Visualization BELIV*, vol. 12, no. 10.1145, pp. 2442576-2442581, 2012.

[27] C. Plaisant, "The challenge of information visualization evaluation," *Proceedings of the working conference on Adavanced visual analytics*, pp. 109-116, 2004.

[28] T. Isenberg, "Postion Paper: Touch interaction in Scientific Visualization," *Procedings of the Workshop on Interactive Surfaces*, pp. 24-27, 2011.

[29] P. Saraiya, C. North and K. Duca, "An Evaluation of Microarry Visualization Tools for Biological Insight," *INFOVIS 04: Proceedings of the IEEE Symposium on Information Visualization*, 2004.

[30] A. Perer and B. Shneiderman, "Integrating Statistics and Visualization for Exploratory Power: From Long-Term Case Studies to Design Guidelines," *IEEE Computer Graphics and Applications*, vol. 29, no. 3, pp. 39-51, 2009.

[31] J. Rieman, "A field study of exploatory learning strategies," *ACM Transactions on the Computer-Human Interaction*, vol. 3, pp. 189-218, 1996.

[32] P. Saraiya , C. North and K. Duca, "An insight-based methodology for evaluating bioinformatics visualizations," *IEEE Trans Vis Comput Graph*, vol. 11, no. 4, pp. 443-56, 2005.

[33] A. Kobsa, "An empirical comparison ot three commercial information visualization systems," *Proceedings of InfoVis*, pp. 123-130, 2001.

[34] J. Gerken, P. Bak and H. Reiterer, "Longitudinal evaluation methods in human-computer studies and visual analytics," *InfoVis*, 2007.

[35] V. Gonzales and A. Kobsa, "A workplace study of the adoption of information visualization systems," *Proceeding of IKNOW'03: 3rd International Conference of Knowledge Management*, pp. 92-102, 2003.

[36] H. Lam, E. Bertini, P. Isenberg, C. Plaisant and S. Carpendale, "Empirical studies in information visualization: Seven scenarios," *IEEE Transactions on Visualization and Computer Graphics*, vol. 9, no. 18, pp. 1520-1536, 2012.

[37] M. Dörk, P. Feng, C. Collins and S. Carpendale, "Critical InfoVis: Exploring the Politics of the Visualization," *CHI '13 Extended Abstracts of Human Factors on Computing Systems (CHI EA '13)*, pp. 2189-2198, 2013.

[38] M. Tory and T. Möller, "Human factors in visualization research," *IEEE Trans Vis Comput Graph*, vol. 10, no. 1, pp. 72-84, 2004.

[39] A. Tversky and D. Kahneman, "Judgement under uncertainty: Heuristics and bias," *Science*, vol. 185, pp. 1124-1131, 1874.

[40] M. R. Munafo, B. A. Nosek, D. V. M. Bishop, K. S. Button, C. D. Cbambers, N. Percie du Sert, U. Simonsohn, E. J. Wagenmakers, J. J. Ware and J. P. A. Ioannidis, "A manifesto for resproducible science," *Nature Human Behavior*, vol. 1, no. 21, 2017.

[41] R. W. White, B. Kules, S. M. Drucker and M. C. Schraefel, "Supporting Exploratory Search, Introduction," *Communications of the ACM*, vol. 49, no. 4, pp. 36-39, 2006.

[42] J. Thomson, E. Hetzler, A. MacEachren, M. Gahegan and M. Pavel, "A typology for visualizing uncertainty," *Proceedings SPIE 5669 Visualization and Data Analytics*, vol. 146, 2005.

[43] A. Bertini, D. Tatu and A. Keim, "Quality metrics in high-dimensional data visualization: An overview and systematization," *IEEE Trans Vis Computer Graphics*, vol. 17, no. 12, pp. 2203-2212, 2011.

[44] D. F. Keefe, "Integrating Visualization and Interaction Research to Improve Scientific Workflows," *IEEE*

Computer Graphics and Applications, vol. 30, pp. 8-13, 2010.

[45] C. North, "Toward measuring visualization insight," *Computer Graphics and Applications*, vol. 26, no. 3, pp. 6-9, 2006.

[46] P. Saraiya, C. North and K. Duca, "Visualizing biological pathways: requirements analysis, systems evaluations and research agenda," *Information Visualization*, vol. 4, no. 3, pp. 191-205, 2005.

[47] P. Pirolli and D. M. Russell, "Introduction to Special Issue on Sensemakeing Human-Computer Interaction," vol. 26, no. 1-2, 2011.