EuroVis Workshop on Visual Analytics (2021)
J. Bernard and K. Vrotsou (Editors)

Immersive Analytics of Heterogeneous Biological Data Informed through Need-finding Interviews

Christine Ripken[1], Sebastian Tusk[1], Christian Tominski[2]

[1]Breakpoint One GmbH, Berlin, Germany
[2]Institute for Visual & Analytic Computing, University of Rostock, Germany

Abstract

The goal of this work is to improve existing biological analysis processes by means of immersive analytics. In a first step, we conducted need-finding interviews with 12 expert biologists to understand the limits of current practices and identify the requirements for an enhanced immersive analysis. Based on the gained insights, a novel immersive analytics solution is being developed that enables biologists to explore highly interrelated biological data, including genomes, transcriptomes, and phenomes. We use an abstract tabular representation of heterogeneous data projected onto a curved virtual wall. Several visual and interactive mechanisms are offered to allow biologists to get an overview of large data, to access details and additional information on the fly, to compare selected parts of the data, and to navigate up to about 5 million data values in real-time. Although a formal user evaluation is still pending, initial feedback indicates that our solution can be useful to expert biologists.

CCS Concepts
• ***Human-centered computing*** → ***Visual analytics;*** • ***Computing methodologies*** → *Virtual reality;*

1. Introduction

Understanding relations across different types of biological data, including cellular structures, numeric parameters, genome data as well as spatio-temporal transcriptome, proteome, and metabolome data for thousands of individuals, is challenging. In the AVATARS project, we investigate how virtual reality (VR), in particular immersive analytics (IA) can help tackle this challenge.

IA is a young field of research [MSD*18, LBDM19, FP21] and there is much potential in exploring the utility of immersive methods for specific application domains [SBKW18, EBC*21]. However, domain experts often cannot afford exploring new technologies on their own. Instead, the new technologies must be brought to the prospective users. Therefore, this work aims to investigate what immersive analytics technology can do for biologists working with large heterogeneous datasets.

In the first place, it is necessary to identify issues that researchers face in today's biological data analysis practice. To this end, we met with 12 biological experts and conducted need-finding interviews, an established design-thinking method. In a second step, we started developing an IA solution addressing the most pressing issues brought forward during the interviews. Inspired by existing software currently applied in the biologists' daily work, our prototype visualizes genome data along with transcriptome and phenome data in tabular form projected onto a virtual wall. About 5 million data values can be accessed in real-time, and appropriate interaction mechanisms allow biologists to link different information on the fly and study selected parts of the data in detail.

Taken together, the contributions of this work are two-fold : We (i) report on insights from need-finding interviews with biological domain experts and (ii) develop an IA solution for exploring large heterogeneous biological data.

2. Related Work

Before going into the details of our work, we briefly review existing work in the area of biological data visualization.

A widely used tool among biologists is the web-based genome browser JBrowse [SUS*09]. It provides a coherent visual framework for visualizing even larger genome data, such as next-generation sequencing reads. However, JBrowse is for regular desktop displays, which limits the amount of data that can be visible at a time. BactoGeNIE is a genome browser that displays large genome collections on a large display wall, which offers much more space than regular displays [ARJ*15]. The advantages of visualizing data on large displays also come to bear in other areas of biology [RTR*16]. Yet, while big displays can show more data, they are also expensive to install and maintain.

A promising and less costly alternative are VR solutions where the large display space is not physical but virtual. Naturally, a considerable part of biological visualization in VR environments

DOI: 10.2312/eurova.20211092

is related to 3D volumetric protein or molecular data [SBKW18]. For example, VR has been employed to visualize the docking of molecules [AW99], the function and the structure of proteins [MM04], the spatial arrangements of the genome for all their architectural levels [KLK*09], and the behavior and dynamics of molecular assemblies with up to 1.7 million atoms [DPT*13]. More recently, VR has been utilized for viewing, manipulating, and modifying chemical macro-molecular structures [KBL*19].

In summary, recent work in relation to biological visualization in immersive settings has mainly considered volumetric data. In contrast, our work focuses more on abstract tabular biological data to be displayed in the large virtual space offered by VR.

3. Need-finding Interviews

The needs to be addressed by an IA solution have been determined via interviews with biologists. Need-finding interviews are an accepted method of design thinking [BU16].

Interview Participants and Procedure We talked to 12 experienced biological researchers (aged between 25 and 53) working for a crop plant research institute, a plant breeding company and a university that are located in Germany. The researchers are involved in a research project that investigates rape oil seeds (brassica napus). Their backgrounds are in computational biology (4), phytopathology (2), molecular breeding (2), seed technology (1), molecular genetics (2), and integrated mechanistic modeling (1). Apart from one, all researchers were either enrolled in a PhD program (4) or were already holding a doctoral degree (7).

Patnaik et al. [PB99] suggest conducting need-finding in the users' environment and observing natural user behavior. Accordingly, six participants were interviewed in their work places, the other six were interviewed remotely (due to the corona pandemic). Prior to the interviews, we informed the biologists that they will be asked about their research, the visualization techniques they use, and the challenges they face in their data analysis practice.

The actual interviews started with creating a welcoming atmosphere. We encouraged information exchange by advocating that the biologists take their part in leading the conversation. We also encouraged reporting problematic issues and made clear that no interview answer will be right or wrong. Although following a relatively open protocol, the interviewer made sure that the four key topics of need-finding interviews were discussed [BU16]: person (the biologists and their background), gains (pros of current work practice), pains (problems and issues), jobs to be done (hopes and needs with respect to new tools to be developed). From the transcribed interviews, we extracted the following key points.

Relevant Biological Data Types The overarching research aim of our biologists is to develop prolific plants that are resilient against extreme weather conditions due to climate change. To this end, the whole biological life cycle from the expression of genes to the manifestation of phenotypic traits of the plant has to be studied.

Our interviews confirmed that the biologists face a highly interdependent and heterogeneous set of data types, where the genome is a central element. The *genome* is the entirety of all DNA within an organism and becomes transcripted. The result of the transcription is the transcriptome. The *transcriptome* is the sum of all transcribed genes within an organism. Through several intermediate steps involving metabolomics and systeomics, phenotypical traits emerge in an organism. The *phenome* is the set of all phenotypes expressed by a cell, organism, or species.

Thus biological data is highly interrelated. The relations between phenotypic traits and the genome is at the heart of many biologists' research interest. This is where our new immersive analytics solution can step in.

Relevant Analysis Tasks and Needs A major concern is to enable the biologists to identify which anomalies in the genome cause the expression of certain traits of the phenotype. Thus, researchers are searching for consistent relations between genotype and phenotypic traits. The genome browser JBrowse [SUS*09] is often applied for this task. It shows genotypes as horizontal color-coded tracks that are aligned and stacked vertically, which visually results in a tabular representation. While JBrowse offers basic functionality for the visual analysis process, the interviewed biologists brought forward several ideas for improvements.

A critical issue is the linking of information from the diverse data involved in the analysis. So far, this linking is a largely manual process where connections between different tools are made based on alphanumeric IDs. Therefore, the biologists expressed a clear need for on-the-fly access to details and additional information from different datasets (e.g., the phenotypical traits or the environmental conditions an individual was exposed to).

Related to that is the need to get an overview of all relevant information. As the data can be quite large, the biologists also commented on the ability to flexibly navigate between different locations of the genome. According to our interviews, the existing practice is limited in terms of how much data can be seen, and also suffers from substantial latency issues when the data are navigated, which can affect the data analysis adversely [LH14].

The biologists are used to a tabular data representation. They search for systematic anomalies and compare different sections of a genotype. Furthermore, it is also desired to analyze the distribution of differences of a genotype sequence to a selected reference genotype. In line with previous research on tabular representations [PDF14], a much needed functionality is to re-order rows (i.e., the genotypes) of the table and group them to form subsets of the data. This also includes the option to hide rows or row segments that are irrelevant to the analysis focus at hand.

The genotypes visualized together in groups are usually related. A haplotype is a group of genome sequences in an organism that are inherited together from a single parent. The tracing of the inheritance of groups of genome sequences across generations is of high relevance. Such family relations of individuals can be presented with dendrograms. For the biologists its a pressing issue to trace haplotypes across generations.

In summary, our interviews made clear that the existing analysis workflows require substantial manual work from the biologists. The key concern to improve the situation is the flexible integration of different types of biological data. Moreover, there are the (typical) requests to be able to see much data at once and navigate the

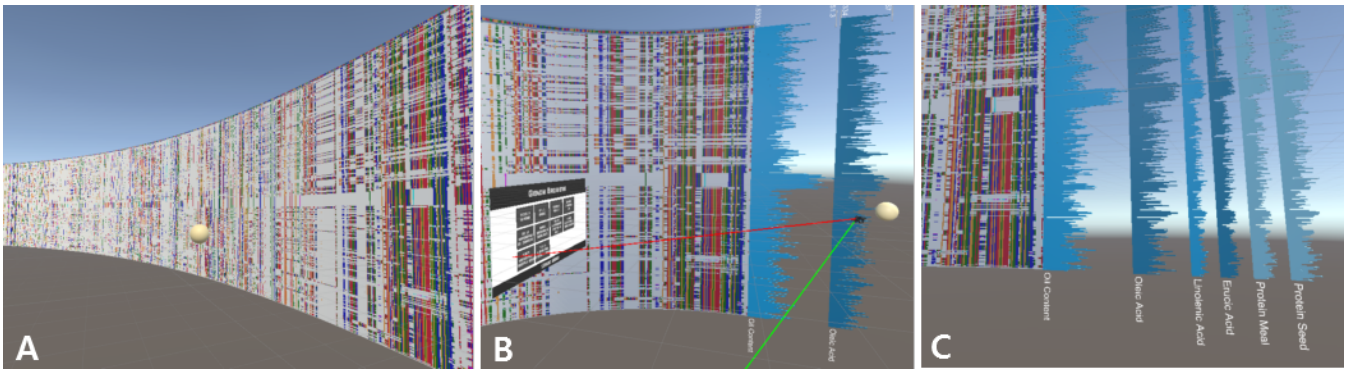

Figure 1: *(A) Overview of genotype data projected as color-coded rows onto a curved virtual wall. (B) + (C) Additional phenotypic traits attached as bar charts to the right allow biologists to relate genotype and phenotype information.*

data with interactive response rates. The following list summarizes the needs identified in our interviews:

N1 Linking different datasets
N2 Overview of larger amounts of data than in JBrowse
N3 Details and additional information on demand
N4 Low-latency data navigation
N5 Comparison of data subsets
N6 Flexible re-ordering and grouping of data
N7 Traceability of inheritance

Facilitating in particular need **N1** - **N4** in virtual reality would yield a relevant advantage in comparison to JBrowse. We hypothesize that the large virtual display space is advantageous for showing large amounts of data. Moreover, immersing the biologists into their data can potentially be helpful for a seamless visual analysis and promote fluid interactive exploration [EMJ*11].

4. Prototype

Addressing the aforementioned needs, we started to develop an IA solution to support the biological data analysis. The goal is to show more data in VR than would be possible on regular displays. The interaction modalities available in VR are to be exploited to facilitate data navigation and reduce the barriers between the diverse types of data involved in the analysis.

Visual Design In accordance with the biologists' current working practice, our prototype is based on a tabular representation. We show a curved wall onto which the genome data are projected for an overview addressing **N2** of our list of needs. The base visualization is designed as follows. A row in the table represents a genotype as a nucleotide sequence of the four base pairs encoded by color: adenine (blue), thymine (yellow), guanine (green), and cytosine (red).

To address the need of comparison **N5**, the visualization is enhanced with a direct encoding of differences [GAW*11]. At the top of the table, a dedicated row shows a selected reference genotype according to which the other genotypes in the table are compared. Per row, sub-sequences that match the reference are colored in white. Sub-sequences that differ from the reference preserve their original blue-yellow-green-red coloring of the nucleotides. As illustrated in Figure 1A, the biologists can immediately see how similar the analyzed genotypes are with respect to the reference.

During the comparative analysis of the genotypes, interesting patterns might be found. One such pattern stands out as rows dominantly colored in white in the center Figures 1A and B and also in Figure 2A. The question for the biologists is how such a potential anomaly relates to the transcriptome and the phenome (**N1**). For integrating the transcriptome, the rows of the table are separated along the vertical axis as in Figure 2B. The created gaps are then filled with transcripts being aligned with the nucleotide sequence. Transcripts can have a length between 18 and 25 nucleotides. The direction in which the transcript is oriented is encoded by color: a lighter color for left-to-right transcription and a darker color for the opposite direction. As the information density can become quite large with the additional transcripts, readability might suffer in VR. Therefore, selected rows of the table can be magnified analog to the focus+context approach of the Table Lens [RC94]. As shown in Figure 2C, this makes it easier for the biologists to study relations between genotype and transcripts in detail (**N3**).

Information about phenotypic traits is embedded in two different ways (**N1, N3**). One option is to show a 3D representation for a selected phenotype (in our case plant seeds). As illustrated in Figure 2A, the seed is displayed in the user's field of view near the ray cursor. The shell of the seed is semi-transparent to reveal its internal tissues. It is possible to select a particular tissue to visualize its transcripts in relation to the corresponding genotype. The second option to access phenotypic information is to attach an additional data table on demand, similar to [BST19]. This table encodes diverse phenotypic traits (typically numerical data attributes) as bars of different length. For example, for the rows that already sparked our interest in Figure 1A, adding the additional table to the right in Figures 1B and C reveals that the underlying genotypes correspond to phenotypic traits associated with higher values, for example, of oil content and oleic acid. The biology experts can immediately utilize this additional information to build new hypotheses or to direct their further investigation of the data.

Overall, our prototype can visualize in real-time the genotypes of four hundred DNA microarrays with each 12.388 base pairs and display numerical phenotypic data, RNA, and a reference genome.

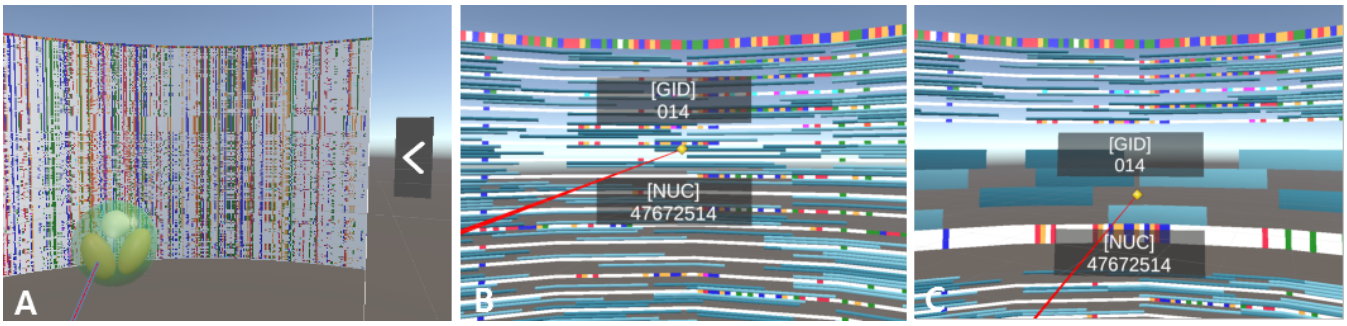

Figure 2: *(A) Genome overview with phenotype information embedded as 3D seed model. (B) Rows of nucleotide sequences are separated to integrated transcriptome information into the created gaps. (C) Focus+context magnification can help biologists inspect the relations between nucleotide sequences and transcripts in detail.*

This amounts to about 5 million values being shown, which should satisfy **N2**, the need to show large quantities of biological data.

Interaction Design To support the data exploration in real-time, our prototype offers several interaction facilities. The virtual scene can be scanned with ray cursors to identify and select particular parts of the data. These selections trigger the display of labels and the magnification of table rows (as shown in Figure 2B and C), and also control the on-demand visualization of transcriptome and phenome data. The linking of the different types of data is aligned with the analysis practice as identified in the need-finding interviews.

Of particular interest is the navigation in the data table via horizontal scrolling (**N4**). By using the joystick of a VR controller, the nucleotide sequences can be scrolled horizontally without noticeable delay. Scanning the entire sequences from left to right takes about 30 seconds, which showed to be unproblematic in initial tests. Alternatively, the user can drag the sequences horizontally using the ray cursor. Buttons on the far left and right of the visualization (see Figures 1B and 2A) also provide scrolling functionality.

When large portions of the visualization move too fast around the user, they are likely to experience motion sickness. The different interaction modalities have different advantages and disadvantages. The buttons reduce the risk of motion-sickness best, but do not support orientation very well. Scrolling by joystick is comfortable, but prone to causing motion-sickness. Dragging with the cursor ray keeps the user oriented and also prevents motion-sickness. Yet, the dragging is more costly to perform, especially when larger distances need to be covered.

For completeness, we mention that our prototype is implemented on the Oculus Rift using the Unity framework. We use a GeForce RTX 2060 in a Ryzen 5 2600X six-core processor system.

5. Conclusion

This paper worked toward enhancing established biological data analysis practice with immersive analytics. Our approach is informed through need-finding interviews with expert biologists. This gives us reason to assume that the developed IA solution can be a useful complement to the currently applied web-based analysis tools. So far, we have addressed needs **N1**–**N5**. The biologists can already easily relate genome, transcriptome, and phenome data (**N1**). Thanks to IA, our solution grants an overview of about 5 million data values (**N2**) and facilitates real-time exploration (**N4**). Various interaction methods provide easy access to details and additional information (**N3**). The comparison of biological data is directly integrated into the visual encoding (**N5**). We consider the developed prototype an initial proof of concept to increase awareness and acceptance of IA technology among expert biologist.

Our solution already is an improvement over the non-immersive data analysis practice of the biologists we interviewed. Although the prototype has not yet been deployed (due to the ongoing pandemic), we discussed the initial results in screen sharing sessions with two biologists. They considered the amount of analyzable data and the integration of different data types to be key advantages of our IA solution. Further pilot testing with two non-biologists confirmed that the software is fully functional and easy to use. Nevertheless the users commented that the picking of single values is challenging. However, we ultimately need to conduct evaluation sessions with the biologists and tune our solution accordingly.

In future work, we aim to utilize the advantages of IA to further enhance biological data analysis. For **N6**, the re-ordering and grouping of data, integrating additional interactions and appropriate sorting algorithms, for example, based on sub-sequence similarity [EST20] appear to be sensible next steps. To support inheritance traceability **N7** a dendrogram needs to become integrated. To further facilitate an integrated data analysis (**N1**), we plan to implement additional views that are also common in the biologists' workflows, such as PCA-based 3D scatter plots, force-directed graphs, segmented seed models, and circos plots. A direct integration of these views into the data table similar to responsive matrix cells [HBS*21] would be interesting. Furthermore we consider the development of smart interaction methods that support precise selection of small elements in virtual space as highly relevant.

Acknowledgements

This work is supported by the German Federal Ministry of Education and Research (BMBF) in scope of the project *Advanced Virtuality and Augmented Reality Approaches in Seeds to Seeds* – AVATARS (FKZ 031B0770A).

References

[ARJ*15] AURISANO J., REDA K., JOHNSON A., MARAI E. G., LEIGH J.: Bactogenie: a large-scale comparative genome visualization for big displays. *BMC bioinformatics 16*, 11 (2015), 1–14. doi:10.1186/1471-2105-16-S11-S6. 1

[AW99] ANDERSON A., WENG Z.: VRDD: Applying Virtual Reality Visualization to Protein Docking and Design. *Journal of Molecular Graphics and Modelling 17*, 3-4 (1999), 180–186. doi:10.1016/s1093-3263(99)00029-7. 2

[BST19] BERGER P., SCHUMANN H., TOMINSKI C.: Visually Exploring Relations Between Structure and Attributes in Multivariate Graphs. In *Proceedings of the International Conference Information Visualisation (IV)* (2019), pp. 261–268. doi:10.1109/IV.2019.00051. 3

[BU16] BRENNER W., UEBERNICKEL F.: *Design Thinking for Innovation*. Springer, 2016. doi:10.1007/978-3-319-26100-3. 2

[DPT*13] DREHER M., PIUZZI M., TURKI A., CHAVENT M., BAADEN M., FÉREY N., LIMET S., RAFFIN B., ROBERT S.: Interactive Molecular Dynamics: Scaling Up to Large Systems. *Procedia Computer Science 18* (2013), 20–29. doi:10.1016/j.procs.2013.05.165. 2

[EBC*21] ENS B., BACH B., CORDEIL M., ENGELKE U., SERRANO M., WILLETT W., PROUZEAU A., ANTHES C., BÜSCHEL W., DUNNE C., DWYER T., GRUBERT J., HAGA J. H., KIRSHENBAUM N., KOBAYASHI D., LIN T., OLAOSEBIKAN M., POINTECKER F., SAFFO D., SAQUIB N., SCHMALSTIEG D., SZAFIR D. A., WHITLOCK M., YANG Y.: Grand Challenges in Immersive Analytics. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2021). doi:10.1145/3411764.3446866. 1

[EMJ*11] ELMQVIST N., MOERE A. V., JETTER H., CERNEA D., REITERER H., JANKUN-KELLY T. J.: Fluid Interaction for Information Visualization. *Information Visualization 10*, 4 (2011), 327–340. doi:10.1177/1473871611413180. 3

[EST20] EICHNER C., SCHUMANN H., TOMINSKI C.: Making Parameter Dependencies of Time-Series Segmentation Visually Understandable. *Comput. Graph. Forum 39*, 1 (2020), 607–622. doi:10.1111/cgf.13894. 4

[FP21] FONNET A., PRIÉ Y.: Survey of Immersive Analytics. *IEEE Trans. Vis. Comput. Graph. 27*, 3 (2021), 2101–2122. doi:10.1109/TVCG.2019.2929033. 1

[GAW*11] GLEICHER M., ALBERS D., WALKER R., JUSUFI I., HANSEN C. D., ROBERTS J. C.: Visual comparison for information visualization. *Inf. Vis. 10*, 4 (2011), 289–309. doi:10.1177/1473871611416549. 3

[HBS*21] HORAK T., BERGER P., SCHUMANN H., DACHSELT R., TOMINSKI C.: Responsive Matrix Cells: A Focus+Context Approach for Exploring and Editing Multivariate Graphs. *IEEE Trans. Vis. Comput. Graph. 27*, 2 (2021), 1644–1654. T. Horak and P. Berger are joint first authors. doi:10.1109/TVCG.2020.3030371. 4

[KBL*19] KINGSLEY L. J., BRUNET V., LELAIS G., MCCLOSKEY S., MILLIKEN K., LEIJA E., FUHS S. R., WANG K., ZHOU E., SPRAGGON G.: Development of a Virtual Reality Platform for Effective Communication of Structural Data in Drug Discovery. *Journal of Molecular Graphics and Modelling 89* (2019), 234–241. doi:10.1016/j.jmgm.2019.03.010. 2

[KLK*09] KNOCH T. A., LESNUSSA M., KEPPER N., EUSSEN H. B., GROSVELD F. G.: The GLOBE 3D Genome Platform - Towards A Novel System-Biological Paper Tool to Integrate the Huge Complexity of Genome Organization and Function. In *Proceedings of HealthGrid Research, Innovation and Business Case* (2009), IOS Press, pp. 105–116. doi:10.3233/978-1-60750-027-8-105. 2

[LBDM19] LEE B., BACH B., DWYER T., MARRIOTT K.: Immersive Analytics. *IEEE Computer Graphics and Applications 39*, 3 (2019), 16–18. doi:10.1109/MCG.2019.2906513. 1

[LH14] LIU Z., HEER J.: The Effects of Interactive Latency on Exploratory Visual Analysis. *IEEE Trans. Vis. Comput. Graph. 20*, 12 (2014), 2122–2131. doi:10.1109/TVCG.2014.2346452. 2

[MM04] MORITZ E., MEYER J.: Interactive 3D Protein Structure Visualization Using Virtual Reality. In *IEEE International Symposium on BioInformatics and BioEngineering* (2004), IEEE Computer Society, pp. 503–507. doi:10.1109/BIBE.2004.1317384. 2

[MSD*18] MARRIOTT K., SCHREIBER F., DWYER T., KLEIN K., RICHE N. H., ITOH T., STUERZLINGER W., THOMAS B. H. (Eds.): *Immersive Analytics*, vol. 11190 of *Lecture Notes in Computer Science*. Springer, 2018. doi:10.1007/978-3-030-01388-2. 1

[PB99] PATNAIK D., BECKER R.: Needfinding: The Why and How of Uncovering People's Needs. *Design Management Journal (Former Series) 10*, 2 (1999), 37–43. doi:10.1111/j.1948-7169.1999.tb00250.x. 2

[PDF14] PERIN C., DRAGICEVIC P., FEKETE J.: Revisiting Bertin Matrices: New Interactions for Crafting Tabular Visualizations. *IEEE Trans. Vis. Comput. Graph. 20*, 12 (2014), 2082–2091. doi:10.1109/TVCG.2014.2346279. 2

[RC94] RAO R., CARD S. K.: The Table Lens: Merging Graphical and Symbolic Representations in an Interactive Focus + Context Visualization for Tabular Information. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (1994), ACM, pp. 318–322. doi:10.1145/191666.191776. 3

[RTR*16] RUDDLE R. A., THOMAS R. G., RANDELL R., QUIRKE P., TREANOR D.: The Design and Evaluation of Interfaces for Navigating Gigapixel Images in Digital Pathology. *ACM Trans. Comput. Hum. Interact. 23*, 1 (2016), 5:1–5:29. doi:10.1145/2834117. 1

[SBKW18] SOMMER B., BAADEN M., KRONE M., WOODS A. J.: From Virtual Reality to Immersive Analytics in Bioinformatics. *J. Integr. Bioinform. 15*, 2 (2018). doi:10.1515/jib-2018-0043. 1, 2

[SUS*09] SKINNER M. E., UZILOV A. V., STEIN L. D., MUNGALL C. J., HOLMES I. H.: JBrowse: A Next-Generation Genome Browser. *Genome research 19*, 9 (2009), 1630–1638. doi:10.1101/gr.094607.109. 1, 2