Visualization Challenges of Variant Interpretation in Multiscale NGS Data

E. Ståhlbom[1,2], J. Molin[2], C. Lundström[1,2,3] and A. Ynnerman[1]

[1] Division of Media and Information Technology, Linköping University, Sweden
[2]Sectra AB, Sweden [3] Center for Medical Image Science and Visualization, Linköping University, Sweden

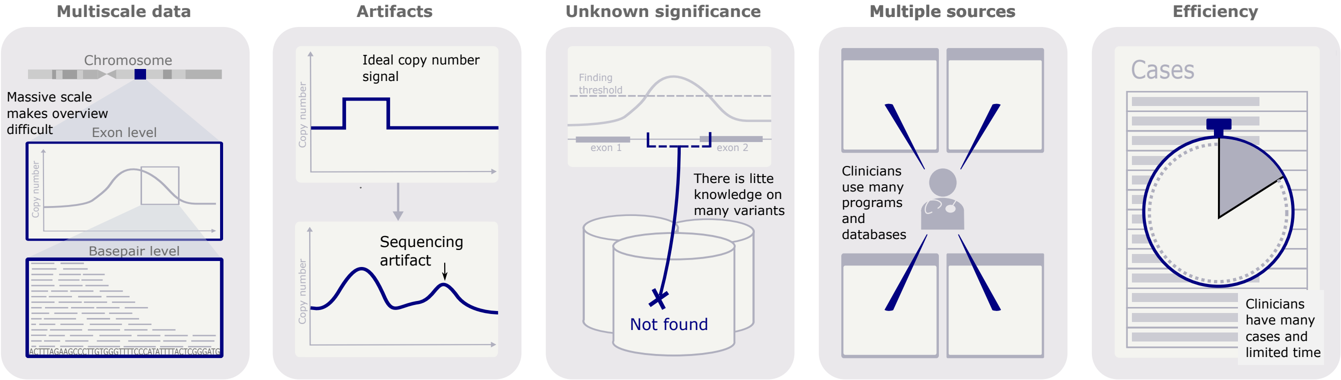

Figure 1: *Five challenges for visualizing genomics data in the clinic identified in this study. Challenges are the multiscale nature of the data, artifacts introduced by the sequencing, unknown significance of findings, a multitude of sources of information, and a need for efficiency.*

Abstract

There is currently a movement in health care towards precision medicine, where genomics often is the central diagnostic component for tailoring the treatment to the individual patient. We here present results from a domain characterization effort to pinpoint problems and possibilities for visualization of genomics data in the clinical workflow, with analysis of copy number variants as an example task. Five distinct characteristics have been identified. Clinical genomics data is inherently multiscale, riddled with artifacts and uncertainty, and many findings have unknown significance, so it is a challenging visual analytics domain. Moreover, as in other clinical domains, high efficiency is key. This characterization will form the basis for follow-on visualization prototyping.

CCS Concepts

• Human-centered computing → Information visualization; Visualization design and evaluation methods; • Applied computing → Genomics;

1. Introduction

The movement towards precision medicine [Hod16] will heavily rely on DNA sequencing, for purposes such as treatment selection for patients with cancer or hereditary diseases. As it gets cheaper and more accessible to sequence a patient's genome, more and larger panels will be utilized, leading to a vastly increasing amount of data to review. Simultaneously, healthcare providers are under economic pressure and have difficulties increasing the spending per patient. Our work is directed towards a new generation of visual analytics tools for clinical genomics, that work with the clinical protocols in place, and scale to manage the growing challenges.

Next generation sequencing (NGS) has decreased the time and money required to sequence DNA [Wet21]. In NGS the DNA is broken into fragments which are often amplified before being sequenced. The sequenced fragments are then mapped to a reference genome, to piece together the sequence of the sample.

In this work we use analysis of copy number variations (CNVs) to probe the visualization needs. Having a CNV means that a large

DOI: 10.2312/evp.20221132

section of the genome is either deleted or copied any number of times [PRB*21]. When there are multiple copies of a section it increases the number of fragments obtained from that region, which is one way to detect CNVs in NGS data [ZWW*13]. Current cancer research shows that CNVs are important for cancer development [BMP*10] as well as in some genetic diseases, but their impact is far from fully understood [PRB*21]. Many CNV visualization efforts in the literature are based on graphs or scatter plots of the copy ratio [KSC*14, Ull17, MYR*19, TSBB16, DVP18].

The main contribution of this work is a domain characterization of the clinical practices for visual review of genomics data. The findings made constitute a complex combination of prerequisites that pose interesting challenges for future research on, and development of, visualization tools.

2. Methods

This paper reports on results from two deep semi-structured interviews with one variant analysts each, site visits to three hospital genomics laboratories, and a shorter interview with one of the variant analysts and a geneticist. In total we engaged with 11 experts throughout the study.

This work uses a design study approach [SMM12], where the communication with domain scientists and end users plays a large role. The domain characterization is part of a larger effort that also includes prototyping of novel visualization tools. This meant that we were in a position to elicit insights from the discussions also through end user feedback on sketches and other design artifacts.

3. Results

Variant review in the clinic is based on call lists, where a computational algorithm locates possible variants and the clinician reviews them and their importance for the patient case. The reviewer can review the data in a genome browser [NHG19] but can also just trust the call list. Including CNVs in the analysis is quite new so there are no clear routines defined yet. Though not exhaustive, five main types of challenges were identified (see Figure 1):

Multiscale data: There is often important information both on the chromosome level, on a more zoomed in regional level, and on basepair level. This multiscale characteristic presents a significant visualization challenge. Study participants expressed that it is not plausible for clinicians to scroll through data at the gene level, therefore they rely on algorithms to identify the interesting areas. Specifically for the CNV detection task, candidate CNVs can also be found by looking for outliers in a scatterplot of the estimated copy numbers along a section of the genome. If more information is needed, basepair level review can be done in a genome browser.

Artifacts: Genomics data are full of artifacts and uncertainties, which the analysts have to consider and assess in their interpretative work. In the interviews, multiple subjects emphasized the importance of distinguishing between artifacts and true findings in the data. It is important to these users that they can review the data to exclude the possibility that their potential finding is erroneous. One user expressed that they can not report a finding unless all evidence agree or they find a reliable explanation for the deviation. The same group of users expressed that knowledge of the chemical and biological processes involved in data extraction was crucial for understanding the data, and that leaving this information out would be irresponsible. On the other hand, they acknowledged that this type of genomics data will be analyzed also by end users not having the same experience and knowledge as they.

Unknown significance: When a finding is made, the available literature must be searched for evidence that the finding has clinical relevance. Much is not yet known about the genome and how different variants change the phenotype of the patient, which study participants expressed as a limiting factor. This is particularly pertinent for CNVs, so many findings are not actionable from a clinical perspective and need to be disregarded.

Multiple sources: Evaluating a finding involves comparing it to other samples and specifics of the lab processes, such as targeted sections. Finding the clinical relevance involves consulting multiple databases for evidence and comparing to gene annotation data. Since this data is often found in different places, the interviewees had to use many different channels (tabs, windows and programs).

Efficiency: The messages were clear from all users and domain experts that time is limited when reviewing a case, so there will not be time to closely investigate every finding in the clinical setting.

4. Discussion

The use of large-scale DNA sequencing in clinical practice is still in its early stages. Even though there are efforts made to create guidelines for assessing and reporting variants [RAB*15, RAC*20], much ambiguity seems to remain about what to do in practice. This is further complicated by the need for efficiency, which hinders the close review of each suspected variant. This is common in medical fields, where limited resources and time is a constraining factor.

A fundamental challenge, especially for CNVs, is that there is important information at multiple zoom levels. This could perhaps be helped by multiple linked views or tracks [SJM*14], and a details-on-demand [Shn96] approach, to enable closer review of selected regions. There is also much metadata in the genome browser, which could be presented in more zoomed-out graphs to indicate sections that might contain artifacts. Yokoyama et al. [YSS*19] also illuminate the challenge of distinguishing nested CNVs.

It is potentially useful to make room in the visualization environment to express data hunches [LAML21], since the users know the data contains artifacts, similar to the situation described by McCurdy et al. [MGM19]. This would provide visualization provenance [XOW*20], by allowing users to follow each other's work.

5. Conclusions

To conclude, the area of genomics visualization presents interesting challenges for multiple reasons, in particular when considering it in the clinic where additional constraints apply.

6. Acknowledgments

This research has been funded by The Swedish Foundation for Strategic Research (grant ID20-0092) and the Knut and Alice Wallenberg foundation (KAW 2019.0024).

References

[BMP*10] Beroukhim R., Mermel C. H., Porter D., Wei G., Raychaudhuri S., Donovan J., Barretina J., Boehm J. S., Dobson J., Urashima M., Mc Henry K. T., Pinchback R. M., Ligon A. H., Cho Y.-J., Haery L., Greulich H., Reich M., Winckler W., Lawrence M. S., Weir B. A., Tanaka K. E., Chiang D. Y., Bass A. J., Loo A., Hoffman C., Prensner J., Liefeld T., Gao Q., Yecies D., Signoretti S., Maher E., Kaye F. J., Sasaki H., Tepper J. E., Fletcher J. A., Tabernero J., Baselga J., Tsao M.-S., Demichelis F., Rubin M. A., Janne P. A., Daly M. J., Nucera C., Levine R. L., Ebert B. L., Gabriel S., Rustgi A. K., Antonescu C. R., Ladanyi M., Letai A., Garraway L. A., Loda M., Beer D. G., True L. D., Okamoto A., Pomeroy S. L., Singer S., Golub T. R., Lander E. S., Getz G., Sellers W. R., Meyerson M.: The landscape of somatic copy-number alteration across human cancers. *Nature 463*, 7283 (2010), 899–905. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7283 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Cancer;Cancer genetics;Mutation Subject_term_id: cancer;cancer-genetics;mutation. URL: https://www.nature.com/articles/nature08822, doi:10.1038/nature08822. 2

[DVP18] Dharanipragada P., Vogeti S., Parekh N.: iCopyDAV: Integrated platform for copy number variations—detection, annotation and visualization. *PLOS ONE 13*, 4 (2018), e0195334. Publisher: Public Library of Science. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0195334, doi:10.1371/journal.pone.0195334. 2

[Hod16] Hodson R.: Precision medicine. *Nature 537*, 7619 (2016), S49–S49. URL: https://doi.org/10.1038/537S49a, doi:10.1038/537S49a. 1

[KSC*14] Kim H., Sung S., Cho S., Kim T.-H., Seo K., Kim H.: VCS: Tool for visualizing copy number variation and single nucleotide polymorphism. *Asian-Australasian Journal of Animal Sciences 27*, 12 (2014), 1691–1694. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4213679/, doi:10.5713/ajas.2014.14143. 2

[LAML21] Lin H., Akbaba D., Meyer M., Lex A.: Data hunches: Incorporating personal knowledge into visualizations. *arXiv:2109.07035 [cs]* (2021). URL: http://arxiv.org/abs/2109.07035, arXiv:2109.07035. 2

[MGM19] Mccurdy N., Gerdes J., Meyer M.: A framework for externalizing implicit error using visualization. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (2019), 925–935. URL: https://doi.org/10.1109/TVCG.2018.2864913, doi:10.1109/TVCG.2018.2864913. 2

[MYR*19] Markham J. F., Yerneni S., Ryland G. L., Leong H. S., Fellowes A., Thompson E. R., De Silva W., Kumar A., Lupat R., Li J., Ellul J., Fox S., Dickinson M., Papenfuss A. T., Blombery P.: CNspector: a web-based tool for visualisation and clinical diagnosis of copy number variation from next generation sequencing. *Scientific Reports 9*, 1 (2019), 6426. URL: http://www.nature.com/articles/s41598-019-42858-8, doi:10.1038/s41598-019-42858-8. 2

[NHG19] Nusrat S., Harbig T., Gehlenborg N.: Tasks, techniques, and tools for genomic data visualization. *Computer Graphics Forum 38*, 3 (2019), 781–805. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13727. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13727, doi:10.1111/cgf.13727. 2

[PRB*21] Pös O., Radvanszky J., Buglyó G., Pös Z., Rusnakova D., Nagy B., Szemes T.: DNA copy number variation: Main characteristics, evolutionary significance, and pathological aspects. *Biomedical Journal* (2021). URL: https://www.sciencedirect.com/science/article/pii/S2319417021000093, doi:10.1016/j.bj.2021.02.003. 2

[RAB*15] Richards S., Aziz N., Bale S., Bick D., Das S., Gastier-Foster J., Grody W. W., Hegde M., Lyon E., Spector E., Voelkerding K., Rehm H. L., ACMG Laboratory Quality Assurance Committee: Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. *Genetics in Medicine: Official Journal of the American College of Medical Genetics 17*, 5 (2015), 405–424. doi:10.1038/gim.2015.30. 2

[RAC*20] Riggs E. R., Andersen E. F., Cherry A. M., Kantarci S., Kearney H., Patel A., Raca G., Ritter D. I., South S. T., Thorland E. C., Pineda-Alvarez D., Aradhya S., Martin C. L.: Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the american college of medical genetics and genomics (ACMG) and the clinical genome resource (ClinGen). *Genetics in Medicine: Official Journal of the American College of Medical Genetics 22*, 2 (2020), 245–257. doi:10.1038/s41436-019-0686-8. 2

[Shn96] Shneiderman B.: The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages* (1996), pp. 336–343. ISSN: 1049-2615. doi:10.1109/VL.1996.545307. 2

[SJM*14] Scales M., Jäger R., Migliorini G., Houlston R. S., Henrion M. Y. R.: visPIG - a web tool for producing multi-region, multi-track, multi-scale plots of genetic data. e107497. Publisher: Public Library of Science. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0107497, doi:10.1371/journal.pone.0107497. 2

[SMM12] Sedlmair M., Meyer M., Munzner T.: Design study methodology: Reflections from the trenches and the stacks. *IEEE Transactions on Visualization and Computer Graphics 18*, 12 (2012), 2431–2440. Conference Name: IEEE Transactions on Visualization and Computer Graphics. doi:10.1109/TVCG.2012.213. 2

[TSBB16] Talevich E., Shain A. H., Botton T., Bastian B. C.: CNVkit: Genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS computational biology 12*, 4 (2016), e1004873. URL: https://escholarship.org/uc/item/8487v9vx, doi:10.1371/journal.pcbi.1004873. 2

[Ull17] Ullmann R.: GenomeCAT: a versatile tool for the analysis and integrative visualization of DNA copy number variants. Publisher: figshare. URL: https://figshare.com/collections/GenomeCAT_a_versatile_tool_for_the_analysis_and_integrative_visualization_of_DNA_copy_number_variants/3660305, doi:10.6084/m9.figshare.c.3660305.v1. 2

[Wet21] Wetterstrand K. A.: DNA sequencing costs: Data, 2021. Accessed on 2022-04-07. URL: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data. 1

[XOW*20] Xu K., Ottley A., Walchshofer C., Streit M., Chang R., Wenskovitch J. E.: Survey on the analysis of user interactions and visualization provenance. *Comput. Graph. Forum* (2020). doi:10.1111/cgf.14035. 2

[YSS*19] Yokoyama T. T., Sakamoto Y., Seki M., Suzuki Y., Kasahara M.: MoMI-g: modular multi-scale integrated genome graph browser. 548. URL: https://doi.org/10.1186/s12859-019-3145-2, doi:10.1186/s12859-019-3145-2. 2

[ZWW*13] Zhao M., Wang Q., Wang Q., Jia P., Zhao Z.: Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics 14*, 11 (2013), 1–16. Number: 11 Publisher: BioMed Central. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-S11-S1, doi:10.1186/1471-2105-14-S11-S1. 2