Eurographics Workshop on Visual Computing for Biology and Medicine (2019)
K. Lawonn and R. G. Raidou (Editors)

Short Paper

A Visual Analytics Approach for Patient Stratification and Biomarker Discovery

S. Alemzadeh[1,2], F. Kromp[3], B. Preim[1], S.Taschner-Mandl[3], K. Bühler[2]

[1] Department of Simulation and Graphics, Otto-von-Guericke University Magdeburg, Germany
[2]VRVis Zentrum für Virtual Reality und Visualisierung Forschungs-GmbH, Vienna, Austria
[3]Children's Cancer Research Institute, Vienna, Austria

Abstract

We introduce discoVA as a visual analytics tool for the refinement of risk stratification of cancer patients and biomarker discovery. Currently, tools for the joint analysis of multiple biological and clinical information in this field are insufficient or lacking. Our tool fills this gap by enabling bio-medical experts to explore datasets of cancer patient cohorts. By using multiple coordinated visualization techniques, nested visual queries on various data types can be performed to generate/prove a hypothesis by identifying discrete sub-cohorts. We demonstrated the utility of discoVA by a case study involving bio-medical researchers.

Categories and Subject Descriptors (according to ACM CCS): I.3.8 [Computer Graphics]: Applications—

1. Introduction

Modern diagnostics aims at the stratification of individual patients into risk groups and enables the identification of patients that will benefit from certain therapy protocols. Initiatives with the overall aim to use more precise diagnostics to allow the selection of patients for therapies that target particular molecular abnormalities have been launched worldwide. The genome-wide search (often referred to as OMICS) for tumor-specific molecular changes as well as histopathological analyses have led to a better understanding of tumorigenesis, progression and relapse and even treatment failure.

This is a particularly promising approach for the treatment of rare cancers, especially for patients with metastatic and/or relapsed cancer [GCB*17]. Because of unknown characteristics of rare cancers and consequently less effective treatment options, the five-year survival rate for patients suffering from rare cancers is on average lower than in case of common cancers, 47% vs. 65% [GVDZC*11] [KLM*13]. Sub-grouping of tumor types, risk stratification and/or survival prediction typically involve genomics (DNA-based analysis) and gene expression (messenger ribonucleic acid mRNA) analysis which are combined with clinical data and minimal residual disease (MRD) analysis. This work is using the childhood cancer neuroblastoma as a use case for rare cancers which account for 15% of all cancer-related deaths in children [SJLD14].

At present, cancer researchers use various independent tools to navigate the different kinds of data acquired through genome-wide multi-scalar analyses, clinical or MRD data to mine for significant differences between groups of patients, e.g. those with a high vs. low risk of relapse or death or to search for biological similarities. Identification and classification of these patient groups is challenging due to heterogeneity, high dimensionality and sparsity of the data. Besides, not all data types are available for all patients. In our case, the clinical information (62 features) of 170 patients was available, while RNA analysis has only been performed on a subset of all patients as often the biopsy was too small to allow for both DNA and mRNA analysis.

There are various visual sub-cohort identification tools that allow the user to explore the subject's data and visually identify sub-cohorts of subjects [KPS15, ZGP15]. Visual analytics of OMICS data (specially to support precision medicine) in an under-explored topic. Marai et al. [MMB*19] proposed a multiple coordinated views system to allow the exploration of heterogeneous data. The aim of their work is to compute the probability of patient's survival considering other similar patients. They represented the patients' features by Kiviat diagrams along with a Kaplan-Meier plot that shows the predicted survival curve. Lex et al. [LSS*12] proposed StromeX as an integrative tool to explore the correlation of clusters of cancer subtypes across OMICS data. In another work of Streit et al. [SLG*14], StromeX is combined with an exploratory tool to compare the patient groups regarding their clinical, genomic alternations and molecular profiles.

Although our work also allows the user to identify sub-cohorts, it is different from previous works, since discoVA provides coordinated navigation of heterogeneous data and enables the expert to identify sub-cohorts of patients by employing a combination of multi-OMICS and clinical data. In this paper, we propose discoVA,

DOI: 10.2312/vcbm.20191235

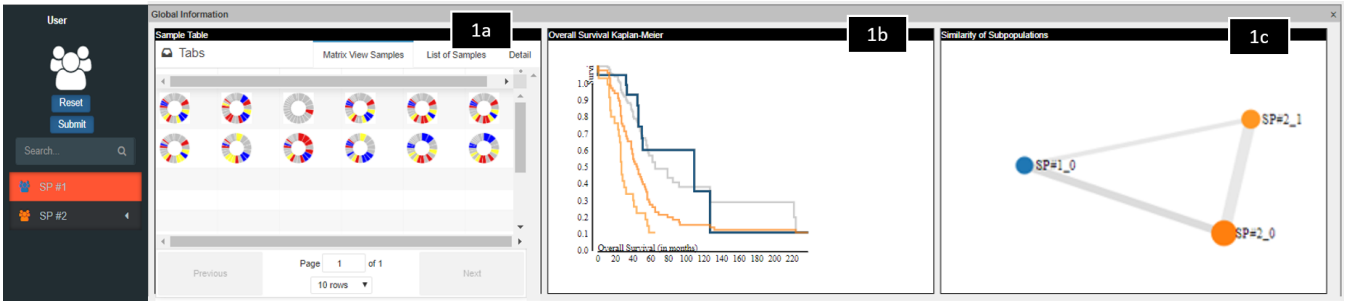

Figure 1: *This panel shows the global information describing the cohort. Component 1a displays the information of the selected individuals belonging to each group. The Kaplan-Meier component (1b) displays the survival curves of identified/brushed group(s). The similarity between identified sub-cohorts based on the shared patients and the brushed sub-cohort is shown in component 1c.*

a Visual Analytic multiple coordinated views tool for refined patient stratification and biomarker discovery. Our contributions include:

- Design and implementation of discoVA to support the expert to identify distinct sub-cohorts using wide-scaled multi-OMICS data along with clinical data in an integrative manner. discoVA functionalities includes:
- A case study of biomedical domain experts using discoVA for joint exploration and identification of a sub-cohort to validate a hypothesis based on previously published correlations in the data.

2. Design Process and Task Analysis

The discoVA design decisions were created in close cooperation between visual analytics experts, data scientists and biology experts. The development of discoVA consisted of three main phases: 1. system design and requirement analysis, 2. system implementation and 3. follow-up. Within each phase of development, multiple meetings were organized to refine the initial system design and requirements. Based on the design considerations, system requirements were defined to allow discovery of discrete sub-cohorts based on the joint exploration of multi-OMICS data along with clinical data. The following describes the tasks in more detail:

T1. Coordinated exploration of heterogeneous patient data.

   a. Clinical and MRD data: It is necessary to visualize the distribution and correlations between variables of clinical and MRD data of a patient cohort. Additionally, the user should be allowed to select a sub-cohort by using these kind of data. The MRD data of bone marrow samples is collected in five time points. The clinical dataset comprises 62 clinical features including metastasis state, blood and urine values.

   b. Genomic data: The user should have an overview and comparison of CNA/genomic intervals of structural genomic aberrations of patient (sub)cohorts to interrogate correlations between patient groups. visualization of individual samples to compare individually. Moreover, having the list of genes located in user-selected chromosomal regions of interest could help to support the generation of new biologically/clinically relevant hypothesis. For our analysis, CNA of 50 samples from 10 patients were available.

   c. mRNA expression data: Heatmaps of genes covering most of the variance in the data can support comparison between different sub-cohorts. As the feature space of mRNA data is large, it is necessary to use techniques to make the interpretation of the data easier. At the time of analysis, mRNA data was available for 50 patients on 20202 genes.

T2. Identification of discrete sub-cohorts using different aspects of the data: Specifying sub-cohorts by iterative queries on heterogeneous data (i.e. MRD, clinical, somatic CNA and mRNA expressions) should be provided to support the hypothesis generation and validation.

T3. Exploration of individual patient data across all data types: To give a more detailed view on the subjects of a sub-cohort.

T4. Comparison of similarity between specified sub-cohorts: The similarity between sub-cohorts regarding the shared members and the deviation survival rates of sub-cohorts from each other should be provided.

T5. Inspection of identified sub-cohorts: The user should be allowed to refer to a previously identified sub-cohort for further investigation.

3. Methods

We propose discoVA as an interactive web-based multiple coordinated views system to support tasks as described in Section 2.

3.1. discoVA Components

The user interface (UI) of discoVA was optimized to support biologists to explore and identify a hierarchy of sub-cohorts using nested visual queries on different data types. At system start-up, datasets are loaded and displayed in all views. Within the following subsections, we describe each of the panels in detail.

3.1.1. Sub-Cohorts Information

This panel provides information on sub-cohorts identified and on individuals within each sub-cohort. It contains three main components:

a. **Samples view:** This component presents an overview on genomic and clinical information of individuals, see Fig. 1(1a). The genomic information of samples is summrized by using

simple circos plots. Sectors represent chromosomes If chromosomes harbor a deletion, a gain or both, the corresponding sector is colored in red, blue and yellow, respectively. Moreover, all chromosomes which do not have any deletion or gain are colored gray. this view presents the user a summary of the genomic information of a sample at a glance (Tasks T1 and T3). This panel contains three tabs:

Matrix view: This tab gives a compressed view of samples in a matrix table for samples with genomic information available.

List View: Gives more information on individuals and contains a list of important features and circos plots of individuals in a cohort.

Detail View: This tab shows the information of individuals by showing the table of all clinical information and a summary of circos plots of samples derived from the same patient (a patient may have multiple genomic samples).

Matrix view and *List view* are filtered for the brushed sub-cohorts. Thus, it lets the expert quickly see the summary of genomic information of patients of the brushed sub-cohort or a selected sub-cohort. Once the user selects one of the samples in these views, the information of the corresponding patient will appear in the *detailed view*.

b. **Kaplan-Meier:** The Kaplan-Meier curve is one of the best techniques to visualize the estimation of the proportion of subjects living for a certain amount of time after diagnosis [GKK10], see Fig. 1(1b). discoVA enables the expert to compare the survival rate of the whole cohort vs. brushed sub-cohort vs. identified sub-cohorts (Tasks T1 and T4).

To start the analysis, the Kaplan-Meier shows the survival curve of the whole population in gray. Once the user brushes a sub-cohort, the survival curve of the corresponding population will appear. Hence, before finalization of a sub-cohort the user can check the survival chance for the brushed population.

c. **Overview of sub-cohorts:** This view shows the similarity of the discovered sub-cohorts and the brushed sub-cohort based on the shared subjects between them (Fig. 1(1c))(Task 4). We model each cohort as a vector, then the pairwise similarity between sub-cohorts is retrieved by cosine similarity. Cosine similarity is already used to calculate the similarity between clinical trial cohorts [LMW17] based on the counties population, but mainly used to show the similarity between documents in the text mining [Hua08].

For this purpose, we generated a binary weighting vector for each sub-cohort by considering a vector with the size of the total number of patients which, by default, is filled with zero. Then, we replaced zero to 1 for involved patients in a specific sub-cohort.

$$A \cdot B = \|A\| \, \|B\| \, cos\Theta \tag{1}$$

After calculating the similarity matrix, the position of sub-cohorts is estimated by multidimensional scaling (MDS) [CC00]. In the overview of sub-cohorts each node visualizes a sub-cohort where the size of nodes represents the size of cohorts and the thickness of line between the nodes represents the pairwise similarity between the cohorts. The expert can refer to a specific sub-cohort by clicking on the node (sub-cohort) (Task T5).

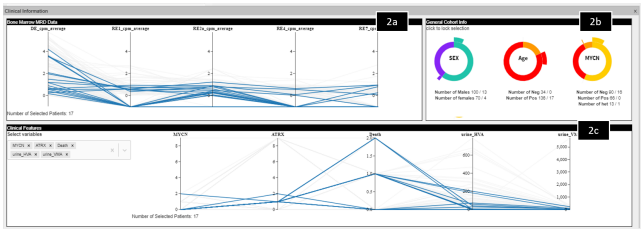

Figure 2: *This panel contains MRD and clinical/genomic manually annotated features. The first parallel coordinates plot shows sequential MRD data in different time points. Some general patient-related statistical information are displayed in component 2b. The last parallel coordinate (2c) shows the clinical and genomic marker information.*

3.1.2. Clinical Information

This panel enables the expert to explore additional clinical information of patients and consists of three components (Task T1).

a. **MRD data:** As shown in Fig. 2(2a), each axis corresponds to one time point of treatment. Each line in the parallel coordinate plot represents a patient and the number of cancer cells in different stages from the diagnosis to after-treatment. The user is allowed to brush specific intervals in any time point to filter the sub-cohort (Task T2).

b. **Statistics:** The sunburst plots show the gender and known risk factors, i.e. these are features which are associated with a high risk of relapse or death, see Fig. 2(2b).

c. **Clinical features:** As displayed in Fig. 2(2c), the x-axis of Parallel Coordinates (PC) plot displays features selected by the user from a drop-down menu and contains time-independent clinical patient information. Because of the high number of features the expert has the option to choose features of interest via a list located in the left side of the plot (Task T1). The expert can apply the brushing of subjects in any desired order.

3.1.3. Genome Information

This panel consists of genomic information of samples in 3 components.

a. **Integrative Genomics Viewer:** The main component of the genomic panel is an embedded IGV which is a well-known tool for biologists to browse genomic information [TRM13](Fig. 3(3a)). The loaded tracks include a cumulative view of genomic CNA intervals, the segmented data, raw data and gene tracks. As shown in Fig. 3(3a) the segmented data track shows the CNA of a sample (with the same color code as in the circos plots).

We enhanced web IGV to adapt to the requirements according to Section 2. We have added a cumulative track of CNA information for the whole genome view (display of all chromosomes) and the whole chromosome view (display of one chromosome) to visualize the total frequency of deletions and gains at each position within the displayed samples (Task T1).

b. **Genome filter management:** As presented in Fig. 3 (3b), this component is linked to the IGV and used for managing the filtering of sub-cohorts by extracting the samples in regions of in-

Figure 3: *The genome information panel comprises three components. The IGV (3a) represents the CNA information of all patients' samples and the raw data of the selected sample. The component 3b is attached to the IGV and is used for the management of regions of interest and settings to specify a sub-cohort. The table of genes (3c) represents the genes within selected locus intervals in IGV.*

terest. To do this, the expert selects specific locus intervals in different chromosomes from IGV and saves the sample lists in each of these regions based on the frequency of deletions or gains. In other words, the expert can save the list of samples which display deletions or gains in a specific region. Then to define/filter a sub-cohort three options are possible: first, getting the union, second, intersect or third, inversion of the samples in the regions of interest. After selection of the desired filter, the operation will be applied to the currently active cohort (Task T2).

c. **Table of genes:** This component shows the table of genes and their relevant information within the intervals (located within the field of view) selected by zooming in IGV, see Fig. 3(3c). These information consist of variant ID, gene name, chromosome, strand and the start and end position of the genes. As the expert selects an interval at a chromosome of interest in IGV, this table will be updated and shows the gene information of the corresponding region. By selection of a certain gene in the table, it will redirect the user to a web page describing the specific gene (https://www.genecards.org).

3.1.4. RNA Expression Data

To show a compact representation of mRNA expression data in discoVA, two components were integrated: hierarchical clustering overlaid to a gene heatmap and a sample similarity plot based on different dimensionality reduction techniques. The components are connected by using the top x genes selected covering most of the variance within the dataset.

a. **mRNA clusters:** As shown in Fig. 4, this component shows a heatmap of agglomeration hierarchical clustering [ML14] of mRNA data for the top selected genes - the top genes can be set by using the threshold slider in the 4b component (Task T1). The results of clustering are shown in a clustergrammer heatmap [FGR*17].

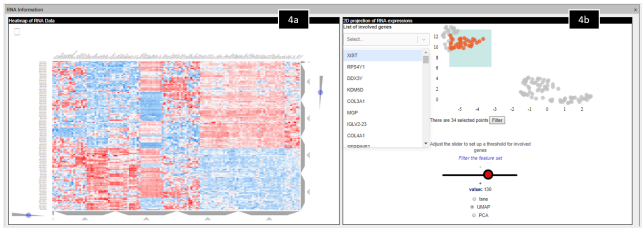

Figure 4: *The mRNA panel consists of two components. The Heatmap (4a) represents the clustered mRNA expression data for the top regulated genes. The second component(4b) shows the similarity between samples for the genes with highest variance. The distance between samples is calcuated by three techniques t-SNE, UMAP and PCA. This component also contains a list of the top genes (the left side). Sub-cohorts can be filtered by brushing data points in this view, where one point corresponds to an individual sample.*

b. **Similarity of samples:** As presented in Fig. 4(4b), this component contains information on the similarity of samples regarding the involved top genes (it is set by the slider) (Task T1). We used three methods to map the samples in 2D space: *t-SNE* [MH08], *PCA* [Fod02] and *UMAP* [MHM18] using python libraries. Thus, it allows the expert to easily compare the results of different techniques.
The user can switch between different dimensionality reduction techniques using radio buttons and to adjust the number of top genes by using the slider. Besides, the expert is enabled to filter the patients by brushing groups of samples of corresponding patients and to generate new sub-cohorts based thereon.

4. Evaluation

To evaluate the power of discoVA, we used the system to investigate a case study. The investigating team consisted of a biologist and a data scientist.

4.1. Case Study

In this case study the analyst team developed strategies to investigate the prognostic relevance, and biological and clinical characteristics associated, of a particular genetic event that has potential relevance in the given cohort of neuroblastoma patients. In many cancer types, mechanisms of telomere maintenance are active, which is not only a major step in tumor development, but is also considered as prognostic factor together with mutations in the TP53 gene and genes of the Ras/MAPK pathway [ACH*18]. Furthermore, it has been demonstrated that in bone marrow metastases certain markers involved in telomere maintenance, e.g. intragenic deletions of the ATRX gene, are frequently associated with copy number aberrations in the 1q and the 19q arm at the time point of relapse [ARB*17].

Therefore the expert user first referred to the clinical features component and selected disease stage and typical genetic features that are either associated with telomere maintenance or that are suspected to occur only in certain sub-cohorts, but not in others. These

were: the status of ATRX intragenic deletion, status of ALT, *MYCN* amplification status and 1p deletion. Then, the expert selected patients with ATRX deletion from the PC of clinical data (Fig. 2(2a)). In the next step, the expert checked the MRD data from the PC of the MRD component.

Second, the expert inspected the detailed genomic CNV information (IGV component) of the selected sub-cohort with ATRX deletion using the cumulative whole genome view in the embedded IGV to see if there are any obvious large scale differences, i.e. segmental chromosomal aberrations. In this step, the expert was interested in aberrations in chromosomes 1 and 19, which have been shown to frequently co-occur [ARB*17]. Thus, the expert took the union list of patients with chromosome 1 and 19 aberrations and submitted it as another sub-sub-cohort. To gain more insights into the characteristics of individual patients in the two sub-cohorts as compared to the total cohort, the expert moved to the Matrix view of samples, then selected one sample and switched to the detail view to see the other samples of the same patient. Next, the expert investigated which genes are located there by using the table of genes component.

Third, the expert referred to the 2D projection of mRNA expression data to investigate the highlighted samples of selected patients. The expert switched between different dimensionality reduction techniques to investigate which ones resulted in a better separation of samples.

5. Conclusion

We developed discoVA, a coordinated multiple views system for integrating and exploring multi-OMICs datasets, to identify potentially new prognositc features and to build new hypotheses. DiscoVA allows the joint exploration of patient-related, clinical, transcriptomic and genomic data of patient cohorts with cancer, using neuroblastoma as a rare pediatric cancer as a use case. Distinct subcohorts can be identified by brushing multiple linked datasets visualized in separate components. To evaluate discoVA, we demonstrated its power by using the system to explore a neuroblastoma case study.

discoVA was considered to satisfy all requirements defined initially. Additional adaptations will be carried out to improve the design and allow a simultaneous view of all components. By increasing the number of samples available in dataset, we plan to implement an unsupervised approach to identify discrete sub-cohorts, hidden relations within the dataset could be revealed .

Acknowledgments

We would like to express our sincere gratitude to Eva Boszaky, Fikret Rifatbegovic, Reza Abbasi, Stefan Fiedler and Ruth Ladenstein for collecting and sharing the data used in this research.

References

[ACH*18] ACKERMANN S., CARTOLANO M., HERO B., WELTE A., KAHLERT Y., RODERWIESER A., BARTENHAGEN C., WALTER E., GECHT J., KERSCHKE L., ET AL.: A mechanistic classification of clinical phenotypes in neuroblastoma. *Science 362*, 6419 (2018), 1165–1170.

[ARB*17] ABBASI M. R., RIFATBEGOVIC F., BRUNNER C., MANN G., ZIEGLER A., PÖTSCHGER U., CRAZZOLARA R., USSOWICZ M., BENESCH M., EBETSBERGER-DACHS G., ET AL.: Impact of disseminated neuroblastoma cells on the identification of the relapse-seeding clone. *Clinical Cancer Research 23*, 15 (2017), 4224–4232.

[CC00] COX T. F., COX M. A.: *Multidimensional scaling*. Chapman and hall/CRC, 2000.

[FGR*17] FERNANDEZ N. F., GUNDERSEN G. W., RAHMAN A., GRIMES M. L., RIKOVA K., HORNBECK P., MA'AYAN A.: Clustergrammer, a web-based heatmap visualization and analysis tool for high-dimensional biological data. *Scientific data 4* (2017), 170151.

[Fod02] FODOR I. K.: *A survey of dimension reduction techniques*. Tech. rep., Lawrence Livermore National Lab., CA (US), 2002.

[GCB*17] GATTA G., CAPOCACCIA R., BOTTA L., MALLONE S., DE ANGELIS R., ARDANAZ E., COMBER H., DIMITROVA N., LEINONEN M. K., SIESLING S., ET AL.: Burden and centralised treatment in europe of rare tumours: results of rarecarenet–a population-based study. *The Lancet Oncology 18*, 8 (2017), 1022–1039.

[GKK10] GOEL M. K., KHANNA P., KISHORE J.: Understanding survival analysis: Kaplan-meier estimate. *International journal of Ayurveda research 1*, 4 (2010), 274.

[GVDZC*11] GATTA G., VAN DER ZWAN J. M., CASALI P. G., SIESLING S., DEI TOS A. P., KUNKLER I., OTTER R., ET AL.: Rare cancers are not so rare: the rare cancer burden in europe. *European journal of cancer 47*, 17 (2011), 2493–2511.

[Hua08] HUANG A.: Similarity measures for text document clustering. In *The new zealand computer science research student conference* (2008), vol. 4, pp. 9–56.

[KLM*13] KEAT N., LAW K., MCCONNELL A., SEYMOUR M., WELCH J., TRIMBLE T., LACOMBE D., NEGROUK A.: International rare cancers initiative (irci). *ecancermedicalscience 7* (2013).

[KPS15] KRAUSE J., PERER A., STAVROPOULOS H.: Supporting iterative cohort construction with visual temporal queries. *IEEE transactions on visualization and computer graphics 22*, 1 (2015), 91–100.

[LMW17] LENERT M. C., MIZE D. E., WALSH C. G.: X marks the spot: Mapping similarity between clinical trial cohorts and us counties. In *AMIA Annual Symposium Proceedings* (2017), vol. 2017, American Medical Informatics Association, p. 1110.

[LSS*12] LEX A., STREIT M., SCHULZ H.-J., PARTL C., SCHMALSTIEG D., PARK P. J., GEHLENBORG N.: Stratomex: visual analysis of large-scale heterogeneous genomics data for cancer subtype characterization. In *Computer graphics forum* (2012), vol. 31, pp. 1175–1184.

[MH08] MAATEN L. V. D., HINTON G.: Visualizing data using t-sne. *Journal of machine learning research 9*, Nov (2008), 2579–2605.

[MHM18] MCINNES L., HEALY J., MELVILLE J.: Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[ML14] MURTAGH F., LEGENDRE P.: Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *Journal of classification 31*, 3 (2014), 274–295.

[MMB*19] MARAI G. E., MA C., BURKS A. T., PELLOLIO F., CANAHUATE G., VOCK D. M., MOHAMED A. S., FULLER C. D.: Precision risk analysis of cancer therapy with interactive nomograms and survival plots. *IEEE transactions on visualization and computer graphics 25*, 4 (2019), 1732–1745.

[SJLD14] SCHLEIERMACHER G., JANOUEIX-LEROSEY I., DELATTRE O.: Recent insights into the biology of neuroblastoma. *International journal of cancer 135*, 10 (2014), 2249–2261.

[SLG*14] STREIT M., LEX A., GRATZL S., PARTL C., SCHMALSTIEG D., PFISTER H., PARK P. J., GEHLENBORG N.: Guided visual exploration of genomic stratifications in cancer. *Nature methods 11*, 9 (2014), 884.

[TRM13] THORVALDSDÓTTIR H., ROBINSON J. T., MESIROV J. P.: Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. *Briefings in bioinformatics 14*, 2 (2013), 178–192.

[ZGP15] ZHANG Z., GOTZ D., PERER A.: Iterative cohort analysis and exploration. *Information Visualization 14*, 4 (2015), 289–307.