GenePattoole:
A mutation-pattern browser for discovery of the severity-driven SNPs in genetic sequences

Mahshid Z. Baraghoush[*]

SFU, Interactive Arts and Technology Department

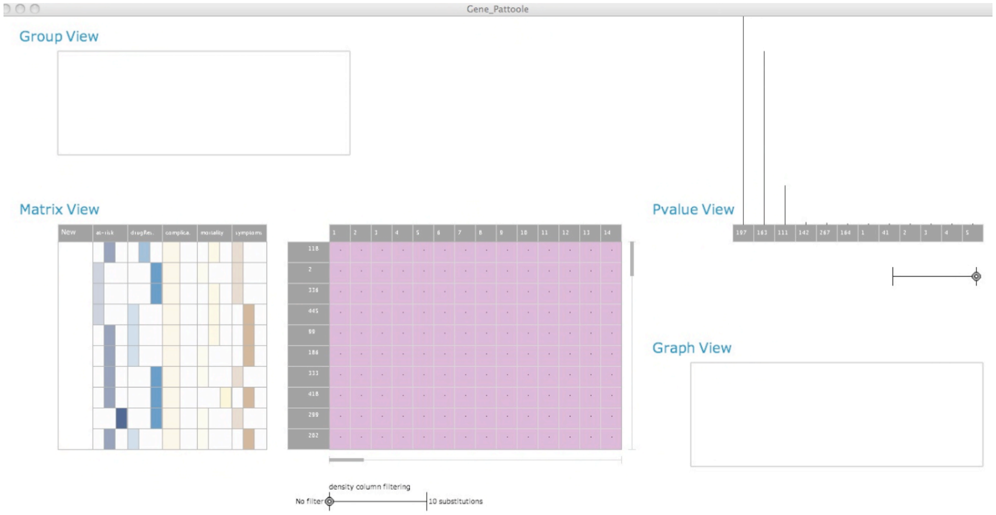


Figure1. GenePatoole is a visualization tool that supports browsing multi-aligned genetic sequences as well as their metadata. The tool helps the analyst to sort the strains based on values of a metadata, define an aggregated metadata from the existing ones. It lets the user filter the positions in the multi-aligned view to being able to look at less information at a time. The tool also helps the user to find interesting pattern of mutations defined by a metric. The user could find relationships between the different positions of the strains and make groups of positions for further investigations. This snapshot contains the *vast challenge 2010* dataset.

Abstract

GenePattoole is a tool that supports the typical visualization of multiple sequence alignments as well as their associated metadata. This tool lets the user sort all the sequences on values of a metadata, aggregate different metadata's using an operation between any of the existing ones to make a new user-defined metadata. This tools supports different filtering options to filter a number of positions in the sequences to help minimizing cognitive memory load in the comparison tasks. The tool has different linked views to support the user in finding interesting mutations, correlations and anti-correlation between positions containing mutations and making groups of positions for further investigations. GenePattoole is part of a general framework for discovery of knowledge in genomic information called IMAS [13]. This prototype tool aims to redesign IMAS multi-alignment view based on well-tested information visualization principles. An informal evaluation of the design convinced us that this visualization tool should become usable for biological researchers with the interest of finding disease-driven SNPs in the genetic sequences. .

Keywords: Information Visualization, vast challenge

1 Introduction and Background

Viruses cannot replicate or evolve without the use of a living-cell's machinery. Once a virus infects a host, it makes copies of itself, growing the population of virus within the same host and eventually spreading to others. During the viral replication process, its gene sequence has to copy and transmit the exact same sequence of about 1000 nucleotides to its child cells. "Consider the fact that in life (literally), nothing is perfect, typically some mistakes made and as a result, some changes appear in genetic sequence"[3]. A gene sequence consists of a sequence of single nucleotide that coded as A, T, C, or G. Each substitution replaces an existing nucleotide in the gene sequence with a different one (for example, A changing to C). These changes in genetic sequence make the virus stronger or weaker and result in more or less dangerous disease.

Health investigators are typically interested to know how they could relate the gene substitutions with disease characteristics. This visualization aims to help an analyzer understands which substitutions in evolved viral strains make a virus cause more or less severe disease. The dataset comes from the vast challenge 2010 biological dataset that contains strains of a particular original virus and their metadata, which is about different disease

[*] email: mzeinaly@sfu.ca

characteristics. The strains are the result of spearing of a disease over time to different infected people. Each of these strains has a gene sequence with one or more nucleotide changes from the original virus's sequence.

The Vast Challenge goal is to invite visual analytics researcher to solve challenge's problems, and evaluate their tool using benchmark datasets. I participated in the challenge using IMAS tool [13] and I was able to understand our tool's strengths and pitfalls in solving the challenge problems. In this project, I will present how I used InfoVis design guidelines and other participant's solutions in the context of IMAS to redesign the multi-alignment view of this tool. Finally I implemented the paper prototypes to be able to evaluate the designs with bioinformatics analyzers using digital prototypes.

I am enhancing IMAS with the hope that it helps analysts to make sense of such gene sequences datasets. My objective of presenting the process of how we improve IMAS tool is to help novice visual analytics researchers to relate the information visualization concepts and guideline presented in by a practical example.

I have broken down the contributions of this visualization design based on a nested four level framework [7]. This work mainly contributed to the level 3 and 4: the operations and data abstraction and encoding and interaction design. There is a little discussion on the domain problem and data characterization (level 1) and the algorithm (level4) under the discussion title of this paper. Figure 2 shows the broken down layers for GenePattoole project.

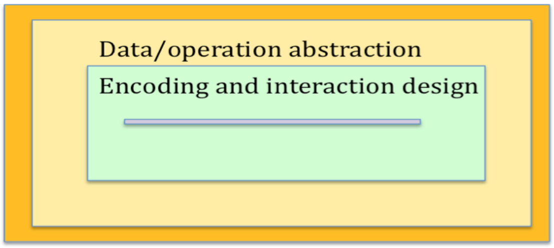


Figure 2. The nested layer for GenePattoole: The size of each box shows the amount of contribution of this project in that level. From the outer to the inner layer: Orange: domain problem identification, Yellow: data/task abstractions, Green: encoding and interaction design, Purple: algorithm design.

2 Previous Work

Biological visualization tools seek to help the analysts browse and manipulate the data to be able to interpret and making sense of the datasets. The tool and visualizations of the biological data and in particular genomic datasets still could be further developed to match domain expert needs and the tool features.

Historically an interactive matrix viewer provides a visualization of multi-aligned sequences. This view generally uses color for each type of nucleotide to highlight variations in the sequences [9]. These views usually accompanied by another metadata matrix view that is a kind of table lens where each columns contains information about one metadata and each row represent the value of that metadata for each strain. The rows are sort-able according to the values of each column [10].

Some of the works done in improving these visualizations focused on the scale of the data. TreeJuxtaposer is one of these tools that uses stretch and squish navigation with the aim of represent dense datasets in a single viewpoint. This tool guarantee visibility for all the data even when some part of it is under the focus of the user [5]. Other scalable applications such as iHat access to a database system in order to scale the data and still keep the program efficient [14].

3 Data and Task

The dataset/task of this work comes from the Vast Challenge 2010, mini-Challenge 3 dataset/challenges 3 and 4.

The dataset is synthesized data of different mutations of a disease.

3.1 Data Set

The dataset consists of two tabular data, which one of them contain information about each strain's genetic sequence and its value for different disease characteristics. The other table contains disease characteristics information for each strain. There are around 100 strains in the dataset each is about 1000 nucleotides long. There are 5 disease characteristics that each has ordered data-type information for each strain. Table 1 shows the disease characteristics table columns values for one sample strain.

Sequence ID	Symptoms	Mortality	Complications	Drug Resistance	At-Risk Vulnerability
32	Mild	High	Minor	Resistant	High

Table 1. Disease characteristics values for strain ID 32

Table 2 contains definitions of each characteristics comes with the dataset.

Characteristics	Definition
Symptoms	What a patient experiences (e.g., pain, sore throat, vomiting, swelling, tremors)
Mortality	Number of deaths as a result of disease
Complications	Unfavorable evolution of illness (e.g. deafness, spontaneous abortion)
Drug Resistance	Mutant vulnerability to anti viral drugs
At-Risk Vulnerability	Disproportional effect on certain risk groups (e.g. children, elderly)

Table 2. Disease characteristics definitions

3.2 Tasks

- Task 1: identify the top 3 mutations that lead to an increase in symptom severity (a disease characteristic). The mutations involve one or more base substitutions.
- Task2: identify the top 3 mutations that lead to the most dangerous viral strains. The mutations involve one or more base substitutions.

Addressing these questions requires the ability to sort the strains on a characteristics and filtering the columns until finding the answers. The filtering decisions require finding interesting patterns, comparing columns and finding their relationships.

4 GenePattoole

GenePattoole is a visualization system composed of multiple linked views [12] including the main view, the matrix view, the P-value view, the graph view and the group view. The name comes form Genomic Pattern Recognition Tool +e. Tool+ e is a Persian word meaning a child. I have used it to show this prototype still needs to be developed and has a long way to become a tool!

5 Implementation

GenePattoole is implemented in the Processing programming language [11]. The structure of the source code and some of the definitions follows Multeesum open-source project [4].

5.1 Main View

Figure 3 (b) shows the screen shot of this view. Each row corresponds to one sequence and each column encodes one position in the genetic sequence. The original dataset contains a categorical data type for each cell in the dynamic range between "A", "C", "T" and "G". I defined a derived variable that its value is a categorical data in the dynamic range of "changed" or "unchanged" from the original sequence in each cell.

This view uses hue as an extremely powerful channel for categorical data to encode in each cell. The ranking of channels to encode the attributes comes from the effectiveness of a visual channel [8]. The purple color with a dot shows there is no change in comparison with the original sequence in the cell happens. The yellow color shows the letter that is changed in comparison with the original sequence. The color choice has been checked in http://www.vischeck.com/.

5.1.1 Interactions

The view supports horizontal and vertical navigations to browse the data. There is a domain constrain that does not allow reordering of the columns as it changes the intuitive order of positions in a sequence. There is also a slider at the bottom that filters the columns with a number of changes less than the value of the slider. (For example when the slider shows 0, all the columns with no changes occurring in them will be filtered). This interaction designed to reduce redundancy in the data and enables the user to see more of their interesting positions in a single viewpoint and release the cognitive memory in the comparison tasks.

There is another interaction that lets the user choose part of the positions and hide them. This does not filter the columns and there will remain a sign for click to unhide them again. This filtering options designed for the situation that there is not a predicted criteria for filtering the columns, but the biologist based on their own experience make a choice to get rid of some of the columns temporarily. The view is using animated transitions [2] when hiding/unhanding is requested in the purpose of remaining the context as much as possible and avoiding a jump to the next stage.

5.2 Matrix View

Figure (a) represent the matrix view with 5 characterises, each represented by a column. Each cell contains the ordered data as a value of a particular characteristic in a particular strain. The last column is the "new" column that could be defined by the analyst, using simple operations (+) in between of the existing columns. This could be useful when a derived characteristic is essential to aggregate the information form the existing columns. This function is defined to address the general need of making the "overall severity" in task 2.

The matrix view uses position as the most powerful channel to encoding the ordered data [8]. I divided each cell in a particular column into the number of dynamic range of a particular characteristics associated with that column. From the right to the left it encodes the higher to the lesser values. It also uses the saturations of a hue for each column, to redundantly encode the same information.

5.2.1 Interactions

The matrix view sorts all the rows according to values of a selected characteristic if the user click on that. Using the scroll vertical bar of the main view also activates scrolling down and up in to this view to match the values with their corresponding strains.

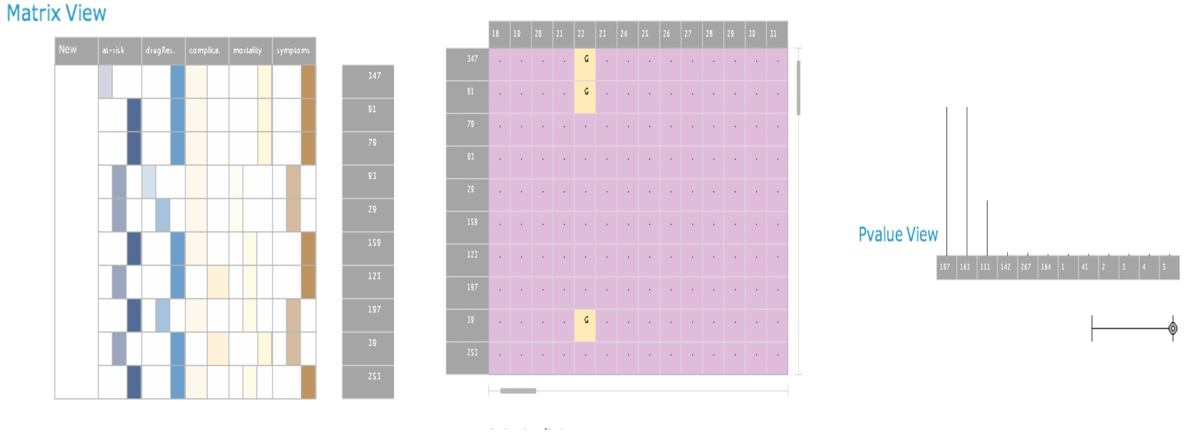
5.3 P-value View

There is a pattern in some of the columns that makes them interesting candidates regarding the answers. The pattern is shows in figure 4. As humans are not good in detecting tasks [6], we can not rely on them to finding this pattern in columns. In fact we need a metric to guide the users in this searching task.

The Noblis team [15] suggested using the reverse of the Mann-Whitney U test's P value as a metric for guiding the user finding interesting positions. I have used the length channel as the next powerful channel after the position channel for encoding the ordinal values of the P-value metric. Figure (c) shows the sorted positions on their P value. The reason of designing a separate view comes from the mentioned constrains in the domain that does not allow sorting the columns.

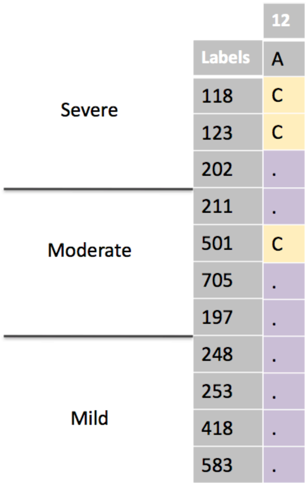
5.3.1 Interactions

This view is linked to other views when the mouse rolls over in each position it highlights (and jump to a window that contains that position) in the main view. There is also a slider that filters positions with less length than the value of the slider. This filtering affects all the view and filters the corresponding column.



(a) (b) (c)

Figure 3 . Screen shots from the system. (a) Matrix view, (b) main View and (c) the P-value view

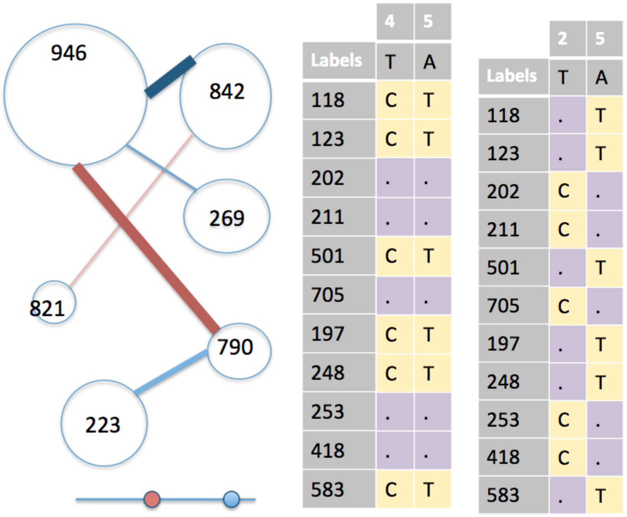


	12
Labels	A
118	C
123	C
202	.
211	.
501	C
705	.
197	.
248	.
253	.
418	.
583	.

Figure 4. The visual representation of the interesting pattern

5.4 Graph View

I have used the Correlation formula to calculate the co-relationship between any two columns. +1 means they are highly correlated and -1 means they are highly anti-correlated. The figure 5 shows an example of each of this relationship. A slider filters column based on the degree of any of these relations.



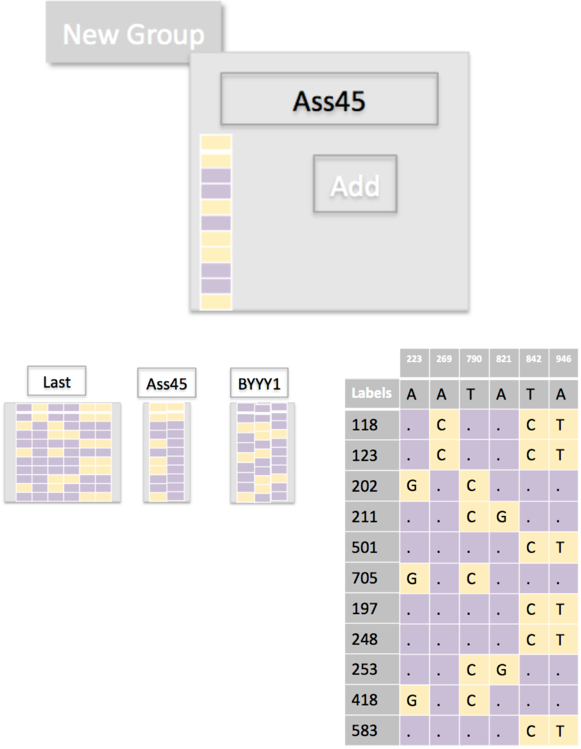
	4	5
Labels	T	A
118	C	T
123	C	T
202	.	.
211	.	.
501	C	T
705	.	.
197	C	T
248	C	T
253	.	.
418	.	.
583	C	T

	2	5
Labels	T	A
118	.	T
123	.	T
202	C	.
211	C	.
501	.	T
705	C	.
197	.	T
248	.	T
253	C	.
418	C	.
583	.	T

Figure 5. The Graph View: the node size is encoding the P-value to make it easier fining those positions in this view. The right pair of columns shows the highly anti-correlations and the middle pair shows the high correlation between two columns. The links width encodes the degree of these relations and the hue is coded for each relationship: Blue for correlated columns and red for anti correlated columns.

5.5 Group View

The group view idea comes from the fact that in finding the relationship between different column, some columns are common between two sets of other columns. This view allows the user to combine columns and make groups for further investigations. The overview shows the name of the group and the general pattern of the member of that group. Figure 6 shows the process of creating a group, which is by “add” columns. A “Last” group saves the last state of the data (filtered positions will be filtered) in to that particular group to enable the user to always come back to the last stage if they select any group and loaded their data in all the views



	223	269	790	821	842	946
Labels	A	A	T	A	T	A
118	.	C	.	.	C	T
123	.	C	.	.	C	T
202	G	.	C	.	.	.
211	.	.	C	G	.	.
501	.	.	.	.	C	T
705	G	.	C	.	.	.
197	.	.	.	.	C	T
248	.	.	.	.	C	T
253	.	.	C	G	.	.
418	G	.	C	.	.	.
583	.	.	.	.	C	T

Figure 6. The group view creation process is to select a name for the group and assign columns. Once a group created, an overview of the pattern of the column would be created and the “last” group is constantly updating after each filtering of the data

6 Evaluation

In this section I briefly discuss the details of my qualitative evaluation with a postdoctoral researcher at the BC Cancer Agency’s Genome Science Centre with the purpose of refining the task abstractions and design ideas. The session was one hour meeting and I used interview protocol and recording the answers by making handwritten notes [1].

7 Results

The scenario was the same (Sorting the rows, filtering the columns until finding the answer.)

7.1 Suggestions for improving each views:

- The size of the main view needs to be bigger.
- The main view needs to have the zooming feature.
- The sliders are very unique and useful.
- The matrix view “new” feature show implemented in a way that considers having different numbers of each level in each column.
- The p-value view could be integrated with the main view and allow the sorting of columns as long as the user could always come back to the original order. This will reserve space.

- The graph view and the group views seem useful as well.

The synthesized dataset has some limitations in comparison with the real world datasets (discussed under the limitation section).

- The P-value and the correlation metrics seem believable to work for this project, but in reality biologist have their own standard metrics for fining the patterns in the data.

8 Discussion

8.1 Strengths

The strength of this work is its multiple filtering options and also the grouping option. Also the abstraction and analysis of each design adds more value to this tool.

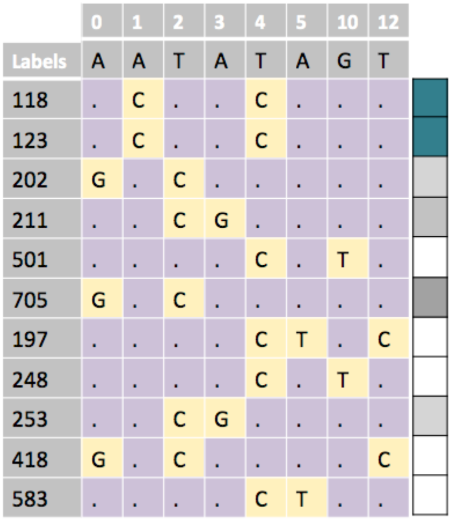
8.2 Limitations

Some of the limitation of this work comes from the synthesized dataset:

1. The missing information about genes.
2. Information about whether or not a mutation occurs in the both chromosomes or just one of them.
3. Information of confident call is missing.
4. All the SNPs are the same across a column. In reality they differ and this creates the other patterns and demand other metrics for finding them.
5. The scale of the data is more. There should be scaling management strategies to address that.

8.3 Future work

For future, I would complete the implementation of this work and evaluate the design and translations with more qualitative studies. There is also a missing part in this paper that is about filtering and hiding rows that I did not have a time to update that part in the current version of this paper. Figure 7 shows how a filtering indicator would look like. Also I will update this paper in with the new content and also will edit its writing. The current implementation is too slow for the original dataset. I would change the data structures and algorithms and if necessary use a database system to manage the data scale in the back. All the current work and the future work would be used in my master thesis.



	0	1	2	3	4	5	10	12	
Labels	A	A	T	A	T	A	G	T	
118	.	C	.	.	C	.	.	.	■
123	.	C	.	.	C	.	.	.	■
202	G	.	C	.	.	.	.	.	▨
211	.	.	C	G	.	.	.	.	▨
501	.	.	.	.	C	.	T	.	□
705	G	.	C	.	.	.	.	.	▨
197	.	.	.	.	C	T	.	C	□
248	.	.	.	.	C	.	T	.	□
253	.	.	C	G	.	.	.	.	▨
418	G	.	C	.	.	.	.	C	□
583	.	.	.	.	C	T	.	.	□

Figure 7.Indicatore guides the user to how to delete a row. Clusters of hue shows similar rows. Within each claster, saturaions of a color shows how they are near of far from the centre of the cluster.

8.4 Lesson learnt

- Collaborations make these interdisciplinary works different.
- This dataset is good for a starting point. In order to making a more effective tool, one could search for synthesized or real dataset in the scale of the vast challenge that simulates the real world detests. After all the evaluations of the design and the usability studies, then it could be a good time to address the scale problem. The ultimate goal is to integrate such research tools with existing real world software that experts use.
- Although these kinds of works seem to be practical but some design guidelines and patterns for the specific domain of biology could be extracted from them.
- The use of hue should be limited to around 6 and maximum 12. Figure 8 shows how I want to limit it by picking just one hue. The reason is because the number of hue in the matrix view could be increased by having more disease characteristics or using the "new" column.

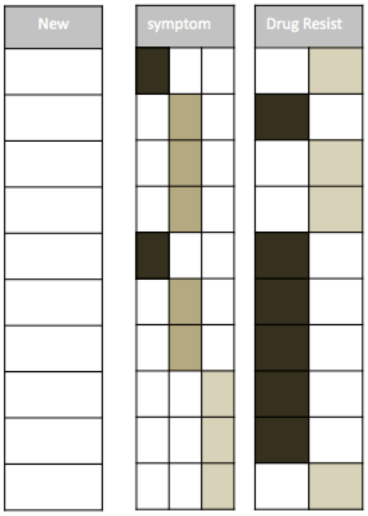


Figure 8. Using one hue for the matrix view

9 Acknowledgement

I would like to acknowledge Chris Shaw my senior supervisor at SFU for the initial inspiration of this work, and for continued feedback throughout the different phases of this project. I thank Tamara Munzner for her very insightful comments related to this project that makes a difference in analysis and the design. I thank the genome research centre biology expert for her time and valuable comments for the case study. I also thank Abtin Rasoulian and Ji-Dong Yim for their useful comments on the paper prototype of the project.

References

1. Creswell, J.W. Research design: qualitative, quantitative, and mixed methods approaches. SAGE, 2009.
2. Heer, J. and Robertson, G.G. Animated transitions in statistical data graphics. IN IEEE INFORMATION VISUALIZATION (INFOVIS, (2007).
3. L., P. DNA replication and causes of mutation. Nature Education, (2008).
4. Meyer, M., Munzner, T., DePace, A., and Pfister, H. MulteeSum: A Tool for Comparative Spatial and Temporal Gene Expression Data. IEEE Transactions on Visualization and Computer Graphics 16, 6 (2010), 908-917.

5. Munzner, T., Guimbretiere, F., Tasiran, S., Zhang, L., and Zhou, Y. TreeJuxtaposer: scalable tree comparison using focus+context with guaranteed visibility. ACM Trans. Graph. 22, (2003), 453-462.
6. Munzner, T., Guimbretière, F., Tasiran, S., Zhang, L., and Zhou, Y. TreeJuxtaposer: scalable tree comparison using Focus+Context with guaranteed visibility. ACM Trans. Graph. 22, 3 (2003), 453–462.
7. Munzner, T. A Nested Process Model for Visualization Design and Validation. IEEE Transactions on Visualization and Computer Graphics 15, (2009), 921–928.
8. Munzner, T. Information Visualization:Principles, Methods, and Practice. .
9. Procter, J.B., Thompson, J., Letunic, I., Creevey, C., Jossinet, F., and Barton, G.J. Visualization of multiple alignments, phylogenies and gene family evolution. Nature Methods 7, 3s (2010), S16-S25.
10. Rao, R. and Card, S.K. The table lens: merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence, ACM (1994), 318–322.
11. Reas, C. and Fry, B. Processing: A Programming Handbook for Visual Designers and Artists. The MIT Press, 2007.
12. Roberts, J.C. State of the Art: Coordinated & Multiple Views in Exploratory Visualization. Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization, 2007. CMV '07, IEEE (2007), 61-71.
13. Shaw, C.D., Dasch, G.A., and Eremeeva, M.E. IMAS: The Interactive Multigenomic Analysis System. IEEE Symposium on Visual Analytics Science and Technology, 2007. VAST 2007, IEEE (2007), 59-66.
14. Vehlow, C., Heinrich, J., Battke, F., Weiskopf, D., and Nieselt, K. iHAT: the interactive Hierarchical Aggregation Table. IEEE VGTC, 2011-33 (2011).
15. Noblis VAST 2010 team. http://www.noblis.org/VAST.