Vistories: Reproducibility, Collaboration, and Communication for Exploratory and Visual Analyses

Nils Gehlenborg
Harvard Medical School - Boston, MA, USA

Alexander Lex
University of Utah - Salt Lake City, UT, USA

Samuel Gratzl
Johannes Kepler University Linz - Linz, Austria

Marc Streit
Johannes Kepler University Linz - Linz, Austria

February 29, 2016

A proposal submitted in response to
NIH/Wellcome Trust/HHMI Open Science Prize Competition.

MOTIVATION

Open Science has the potential to transform how scientific discoveries are made. When scientific findings are fully accessible to all audiences and when any interested party is able to reproduce, re-analyze, extend, and learn from published data, methods, and results, discoveries can be scrutinized, discussed, and disseminated much more widely. A side effect of openness is increased pressure on scientists to undertake and publish their science in a way that enables reproducibility, the lack of which has been identified as a major bottleneck for scientific progress [1, 9, 15].

The Open Science community is increasingly successful in promoting and enabling the publication of data, analysis tools, and computational workflows associated with studies and scientific publications. The number of products and services in this space—often referred to as tools for "reproducible research"—is growing rapidly. Most important in this context are public data repositories (e.g., those hosted by NIH and EBI), source code repositories (Github, Bitbucket, Sourceforge), and solutions to capture and communicate the analysis steps of a published study (e.g., Arvados, Galaxy, Taverna, GenePattern, Jupyter Notebooks, Refinery Platform). The analysis steps essentially represent the "data provenance" of a study, which describes the origins of each data artifact.

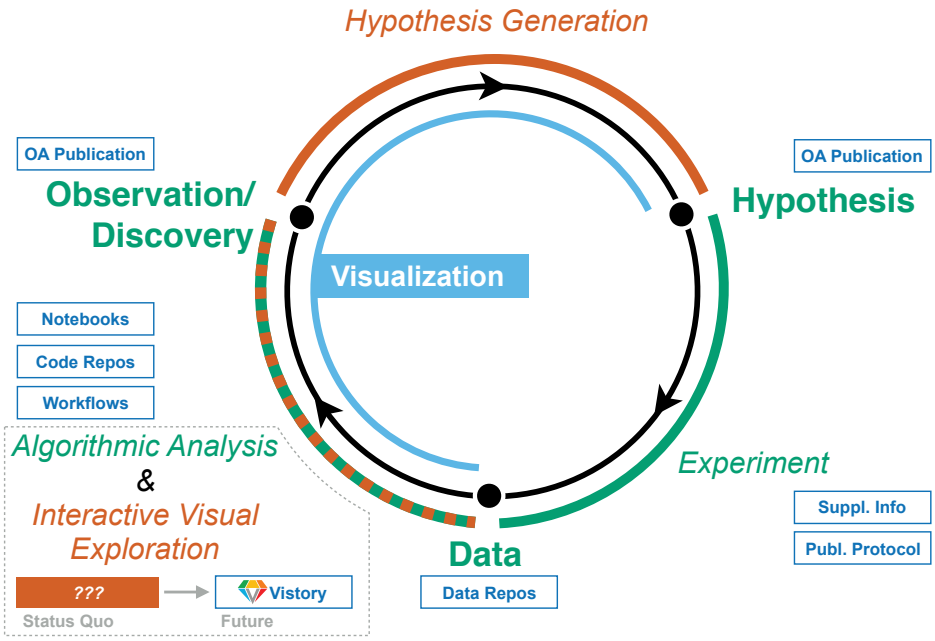


Figure 1: Processes and artifacts of scientific discovery and the application of data visualization (modified from [11]). Processes are set in italics, outcomes in bold. Blue boxes indicate where information about processes and artifacts is captured and disseminated. Green text indicates that solutions to capture and communicate the corresponding processes or outcomes exist, while red text indicates that such solutions are absent.

These solutions, however, address only part of the scientific discovery process (see Figure 1). In any data-driven scientific endeavor **the process of interactive data exploration is essential to guide the discovery process**. Any omissions or mistakes, whether unintentional or intentional, in interpreting the data and generating hypotheses will have an impact on the outcome of a study (see *algorithmic analysis and interactive visual exploration* in Figure 1). While the results might not necessarily be wrong, they will at least be incomplete. Therefore, **capturing, annotating, and communicating the full provenance of data interpretation is a high priority for Open Science and reproducibility**. We call this "exploration provenance", which comprises both data provenance (i.e., algorithmic analyses in workflows) and interaction provenance (i.e., visual exploration).

PRELIMINARY WORK

Our team combines strong expertise in data visualization and computational tool building with bioinformatics expertise. We have jointly developed the Caleydo Visualization framework (`http://caleydo.org`) and many visual data analysis solutions for the sciences based on it. All of our prototypes and the underlying visualization frameworks have consistently been published under permissive open source licenses (see `http://github.com/Caleydo/`). Caleydo also provide easy access to open datasets, including those published by *The Cancer Genome Atlas*.

The applicants also have a long-standing and successful academic collaboration with more than 30 joint papers published in the major data visualization and bioinformatics venues.

Preliminary scientific work relevant to this proposal falls into two categories: (1) the Vistories approach, a method for enabling reproducibility, collaboration, and communication of results of exploratory visual analyses, which is at the core of this proposal, and (2) visualization tools for large-scale scientific data sets that illustrate the breadth of our expertise.

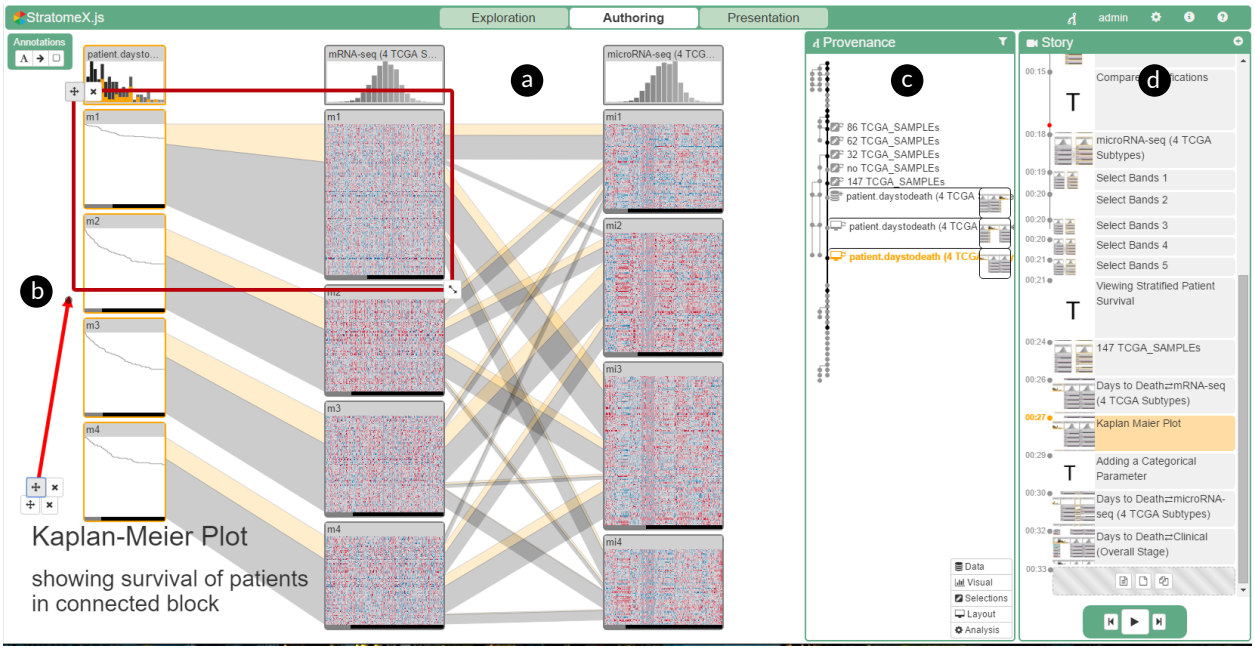


Figure 2: A version of StratomeX fully compatible with the Vistories approach. StratomeX (left) shows the relationship of mRNA-seq and microRNA-seq clusters, and the survival data for the mRNA-seq subtypes. The provenance graph and the interface to interactively curate stories are also shown (right). **Vistory (best with Chrome)**: `http://vistories.org/stratomex.html`

VISTORIES

To address the challenges of reproducible, collaborative, and open research, we developed an approach [4] (accepted for publication at EuroVis 2016) that instruments visualization software to capture semantically meaningful operations in a visualization framework, and enables users to annotate their analysis actions. We can thus capture both the full provenance of the visual analysis, and via annotations also selected insights and intentions of the user. We use this provenance data not only for the purposes of reproducibility, but also to communicate results: in a second step users can choose which parts of an analysis was relevant, curate the process, and create a "Vistory" (see `http://vistories.org`), i.e., an interactive, curated visualization that shows how a discovery was made and is connected to the underlying provenance information. "Vistory" is

a portmanteau of visualization, history, *and* story. Because they also allow consumers to switch from a passive viewer role to an active explorer role, Vistories go significantly beyond the idea of an "interactive figure" or a "visual story" and have deeper implications for Open Science. We have implemented this infrastructure for StratomeX (see below, Figure 2), and for a tool inspired by Gapminder (`http://www.gapminder.org/`), a well-known public health / open data visualization tool (see Figure 3).

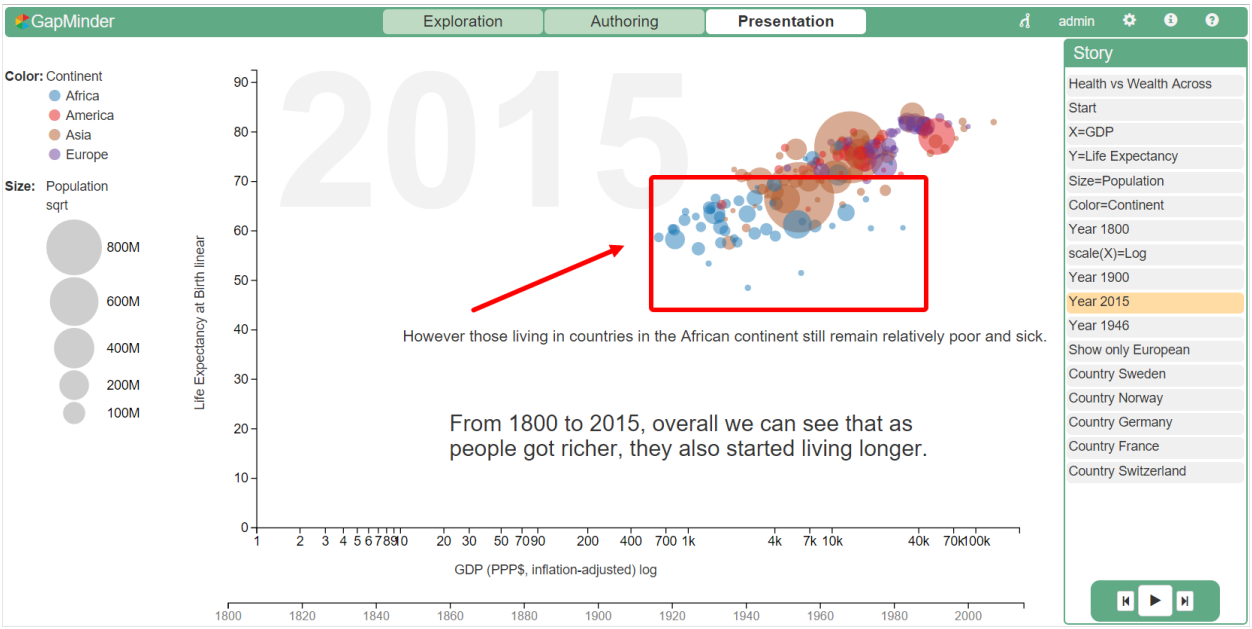


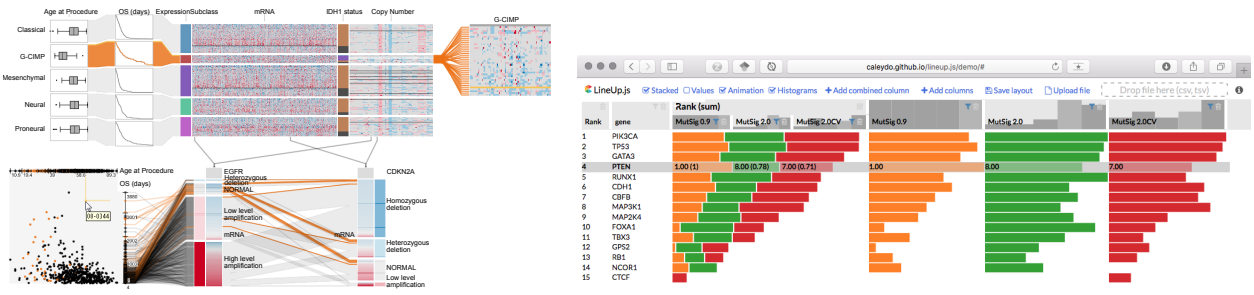
Figure 3: A Vistory enabled version of Gapminder. The scatterplot illustrates the relationship between GDP and life expectancy among different countries over time. **Vistory (best with Chrome)**: `http://vistories.org/gapminder.html`

VISUALIZATION TOOLS

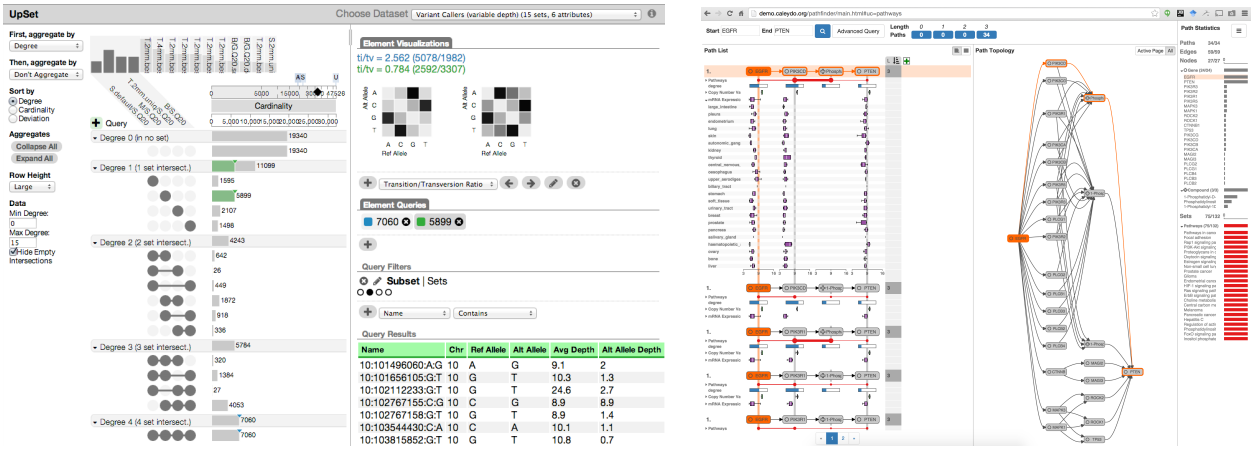
StratomeX [6, 18, 22], shown in the left part of Figure 2, is a visual analysis technique developed to address challenges related to the generation and testing of hypotheses about cancer subtypes. We have use StratomeX for the analysis of subtypes in the TCGA Kidney Carcinoma data set [20], others have used it for example to study subtypes in the thyroid and melanoma cohorts of TCGA [10, 21]. **Domino** [3] (Figure 4a generalizes the approach of StratomeX, allowing arbitrary arrangement and combinations of subsets of data sets that share one or two common identifiers (e.g., patient ids and gene ids).

We have also developed tools for visual exploration of pathways in the context of large omics data sets [7, 12, 13], and for general networks based on path queries (**Pathfinder** [14], Figure 4d).

Vials [19] was designed to analyze and compare alternative splicing data on a large scale. **LineUp** [2] (Figure 4b) is an interactive ranking technique for rankings composed of multiple attributes. We also developed **UpSet** [8], an interactive visualization technique for set data with heterogeneous attributes (Figure 4c) that scales beyond the trivial cases where Venn diagrams are appropriate [5].



(a) Domino **(b)** LineUp



(c) Upset **(d)** Pathfinder

Figure 4: A selection of visualization tools built by the team. See `http://caleydo.org/tools/` for more information.

SOLUTION

The aim of this project is to integrate our existing Vistories approach with algorithmic analysis tools and to add data provenance to the Vistories approach.

BRIDGING ALGORITHMIC AND VISUAL ANALYSIS

In a first step we will integrate algorithmic methods (scripting) and visual methods. We will extend Jupyter Notebooks [16] (`http://jupyter.org/`) with interactive visualization capabilities that read data from the data structures used in the code (which is currently supported in Jupyter), but also feed data back into data structures, which then can be used for further processing. We chose Jupyter Notebooks as they are excellent at tracking analytical provenance and are also widely popular. To achieve this integration of visual and algorithmic analytics, we will take two distinct approaches:

1. We will provide **simple visualizations** (scatterplots, histograms, line charts, etc.), similar to those created with matplotlib, ggplot, bokeh, etc., and display them in line with the notebook. In contrast to current approaches, however, they will not only be interactive, they will also **allow users to execute operations, such as filters, selections, aggregations (groupings), and use the results in subsequent analysis steps**.

2. We will **allow users to launch full-fledged, Vistory-enabled, multiple coordinated view visualization systems and integrate them with notebooks on the data level**. Examples for such tools are network visualization tools (e.g., Cytoscape (`http://www.cytoscape.org`) or Pathfinder [14]), multidimensional data visualization tools (e.g., Domino [3]), or specialized visu-

alization system for scientific data (e.g., a genome browser, a cancer subtype analysis tool). We argue that certain types of visual analysis approaches are simply impossible within the confines of a notebook. For example, analyzing large networks requires rich query and interaction capabilities and appropriate user interfaces. Again, operations on the data within the visualization tool will be accessible in data structures for further algorithmic analysis in the notebook. In addition, we will also make functions defined in the notebook available in the visual interface. E.g., a clustering algorithm written in a notebook cell will be available to run from the graphical user interface to execute on data (sub)sets selected with the visual tool. As the output of such a visual analysis, we will show a representative figure of the interactive visual tool, captured using our Vistory interface, in addition to the derived data (filters, selection, aggregations) that is fed back into the notebook.

To choose which visualization tool to integrate in our prototype, we will run a community survey to identify the most useful tool for open science data analysis applications.

INTEGRATING ANALYTICAL WITH VISUALIZATION PROVENANCE

We will **extend our Vistories approach and Jupyter Notebooks to be able to capture the algorithmic analysis process**. Jupyter is based on a sequence of cells that contain, among others, text that document the process, and source code, that executes an analysis. We will treat each cell in a notebook as a distinct state in the provenance graph and use either text cells or markup in the source code to summarize the cell in the provenance graph. When cells are run with different datasets during an interactive analysis *each data/cell combination will be recorded as a distinct state*. This will allow us to create analysis trees, and enable users to easily switch between branches of the tree and track the outcomes of an analysis given different input data.

In later stages of the project, we will also explore managing code in cells. This will make it possible to run diverging analysis, where each analysis branch is appropriate for the data (sub)set under investigation.

STORYTELLING FOR ALGORITHMIC ANALYSIS

A key component of our Vistories approach is the **bridge between exploration and storytelling**: by authoring which story to tell, we can enable analysts to create compelling and interactive visuals for, e.g., publication in interactive journals, which can also serve as a gateway to the underlying analysis. While there is no equivalent for public communication of source code, we will allow analysts to explain their analysis steps based on the provenance information, e.g., by semi-automatically generating flowcharts for inclusion in a Vistory.

VISTORIES DEVELOPER KIT

To make Vistories accessible to a wider audience, it is essential that we **enable developers to integrate other open tools easily**. To address this goal, we will extend and refine our Vistories developer kit. The developer kit handles the storage of the provenance information and provides the visual interfaces to create and edit Vistories based on the captured provenance (see Figure 2).

The developer kit will provide an API that pushes actions, such as data operations, user selections of data items, or changes in a visualization, into the provenance graph. Our approach is intended for web-based tools, but is independent of the specific technologies employed, how a tool manages its data, or how a user interface is designed. To make tools compatible to the Vistories approach, they only need to trigger appropriate events describing an action, and need to be able to undo and redo these actions. We believe that this flexible mechanism to extend existing tools with Vistories capabilities will foster the integration of the Vistories mechanism into a wide range of tools from various scientific domains. In order to motivate developers to use the Vistories approach with their tools, we will provide step-by-step tutorials together with detailed API documentation.

VISTORIES COMMUNITY PLATFORM

In addition to the Vistories developer kit, we will create **a web platform that enables sharing, exploring, and commenting on published Vistories**. In contrast to existing platforms such as figshare (`https://figshare.com`) or Dryad (`http://datadryad.org`) that allow researchers to share and describe static figures, our platform will provide the full feature set that Vistories offer. We will fill the platform with a collection of Vistory examples created using our prototypes. To ensure that links to vistories can be cited and remain accessible in the future, we will rely on DOIs (`http://www.doi.org`) and repositories such as Zenodo (`http://zenodo.org`).

VISTORIES SCALABILITY

A key component of Vistories is the **tracking and visualization of the provenance graph**. In complex real-world analysis sessions, the provenance visualization needs to scale to hundreds or even thousands of actions. We currently use level-of-detail approaches to achieve scalability. For larger graphs, however, we will need to employ more sophisticated methods. We will apply hierarchical and motif based aggregation techniques to automatically collapse and aggregate parts of the graph. In recent work [17], we successfully applied these strategies for handling large data provenance networks in the Refinery platform (`http://www.refinery-platform.org`).

IMPACT

Making exploration provenance readily available as Vistories will have a broad impact on how we conduct science. Here we describe four of many possible scenarios in which exploration provenance captured in Vistories can directly influence the outcomes:

Collaboration In any team research setting Vistories can be used to collaboratively explore data sets. For example, collaborators who join a project can review earlier explorations, comment on them, and add their own explorations.

Review Vistories will enhance the peer-review process by providing editors and reviewers with accessible and detailed information about the hypothesis generation process. This information can be applied to confirm the soundness of a study under review or to identify potential weaknesses and omissions.

Publication We envision that in future scientific publications Vistories will play an essential role in communicating results and key steps in the analysis processes. Vistories will both replace text plus figures in some sections and be included as supplemental material. Readers will be able to branch off the published Vistory at any time and start their own explorations within the context of the publication and published data.

Education Vistories will be an excellent tool for teaching exploratory data analysis. Students will be able to retrace and discuss the steps that lead to strong hypotheses in published Vistories. Teachers will be able to use the Vistories created by students to identify weaknesses in their problem-solving approaches that would otherwise remain hidden.

BENCHMARKS

In the first phase of the project we will develop a Vistories prototype that is integrated with Jupyter Notebooks and that supports embedded interactive visualizations. We will also publish and document the Vistories developer kit that can be used by the community for integration with other tools. Finally, we will launch the Vistories community platform where users can browse through, refine, and comment on Vistories.

If the prototype is successful, we plan to improve and harden our tools, create advanced visual exploration tools for Jupyter that support sophisticated visualizations, and address any scalability issues. We will also build a user and developer community around Vistories and the associated software ecosystem. For example, we intend to hold workshops and give talks at conferences. We also hope to involve publishers and Open Science service providers in our efforts.

BUDGET

We will use the budget of the first phase to pay PhD students and developers implementing the prototypes, and to pay for hosting on a cloud service provider. The budget will be split 2:1:1 between Linz, Harvard, and Utah.

LICENSING AND DISSEMINATION

All software developed will be in public GitHub repositories and will be published under a BSD License. Documentation will be published under the Creative Commons Attribution license (CC BY). Any publications resulting from this work will be published using an Open Access option or in an Open Access journal.

REFERENCES

[1] C. G. Begley and L. M. Ellis. Drug development: Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533, 2012. ISSN 0028-0836. doi: 10.1038/483531a. URL http://www.nature.com/nature/journal/v483/n7391/full/483531a.html.

[2] S. Gratzl, A. Lex, N. Gehlenborg, H. Pfister, and M. Streit. LineUp: Visual Analysis of Multi-Attribute Rankings. *IEEE Transactions on Visualization and Computer Graphics (InfoVis '13)*, 19 (12):2277–2286, 2013. doi: 10.1109/TVCG.2013.173.

[3] S. Gratzl, N. Gehlenborg, A. Lex, H. Pfister, and M. Streit. Domino: Extracting, Comparing, and Manipulating Subsets across Multiple Tabular Datasets. *IEEE Transactions on Visualization and Computer Graphics (InfoVis '14)*, 20(12):2023–2032, 2014. ISSN 1077-2626. doi: 10.1109/TVCG.2014.2346260.

[4] S. Gratzl, A. Lex, N. Gehlenborg, N. Cosgrove, and M. Streit. From Visual Exploration to Storytelling and Back Again. *Computer Graphics Forum (EuroVis '16)*, 2016. to appear.

[5] A. Lex and N. Gehlenborg. Points of view: Sets and intersections. *Nature Methods*, 11(8): 779–779, Aug. 2014. ISSN 1548-7091. doi: 10.1038/nmeth.3033. URL http://www.nature.com/nmeth/journal/v11/n8/full/nmeth.3033.html.

[6] A. Lex, M. Streit, H.-J. Schulz, C. Partl, D. Schmalstieg, P. J. Park, and N. Gehlenborg. StratomeX: Visual Analysis of Large-Scale Heterogeneous Genomics Data for Cancer Subtype Characterization. *Computer Graphics Forum (EuroVis '12)*, 31(3):1175–1184, 2012. ISSN 0167-7055. doi: 10.1111/j.1467-8659.2012.03110.x.

[7] A. Lex, C. Partl, D. Kalkofen, M. Streit, S. Gratzl, A. M. Wassermann, D. Schmalstieg, and H. Pfister. Entourage: Visualizing Relationships between Biological Pathways using Contextual Subsets. *IEEE Transactions on Visualization and Computer Graphics (InfoVis '13)*, 19(12): 2536–2545, 2013. doi: 10.1109/TVCG.2013.154.

[8] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister. UpSet: Visualization of Intersecting Sets. *IEEE Transactions on Visualization and Computer Graphics (InfoVis '14)*, 20(12): 1983–1992, 2014. ISSN 1077-2626.

[9] A. Nekrutenko and J. Taylor. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nature Reviews Genetics*, 13(9):667–672, Sept. 2012. ISSN 1471-0056. doi: 10.1038/nrg3305. URL http://www.nature.com.ezp-prod1.hul.harvard.edu/nrg/journal/v13/n9/full/nrg3305.html.

[10] T. C. G. A. R. Network. Integrated Genomic Characterization of Papillary Thyroid Carcinoma. *Cell*, 159(3):676–690, 2014. ISSN 0092-8674. doi: 10.1016/j.cell.2014.09.050. URL http://www.sciencedirect.com/science/article/pii/S0092867414012380.

[11] C. B. Nielsen. Visualization: A Mind–Machine Interface for Discovery. *Trends in Genetics*, 0(0), 2016. ISSN 0168-9525. doi: 10.1016/j.tig.2015.12.002. URL http://www.cell.com.ezp-prod1.hul.harvard.edu/article/S0168952515002139/abstract.

[12] C. Partl, A. Lex, M. Streit, D. Kalkofen, K. Kashofer, and D. Schmalstieg. enRoute: Dynamic Path Extraction from Biological Pathway Maps for In-Depth Experimental Data Analysis. In *Proceedings of the IEEE Symposium on Biological Data Visualization (BioVis '12)*, pages 107–114, 2012. doi: 10.1109/BioVis.2012.6378600.

[13] C. Partl, A. Lex, M. Streit, D. Kalkofen, K. Kashofer, and D. Schmalstieg. enRoute: Dynamic Path Extraction from Biological Pathway Maps for Exploring Heterogeneous Experimental Datasets. *BMC Bioinformatics*, 14(Suppl 19):S3, 2013. doi: 10.1186/1471-2105-14-S19-S3. URL http://www.biomedcentral.com/1471-2105/14/S19/S3/abstract.

[14] C. Partl, S. Gratzl, M. Streit, A. M. Wassermann, H. Pfister, D. Schmalstieg, and A. Lex. Pathfinder: Visual Analysis of Paths in Graphs. *Computer Graphics Forum (EuroVis '16)*, 2016. to appear.

[15] F. Prinz, T. Schlange, and K. Asadullah. Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, 10(9):712–712, Sept. 2011. ISSN 1474-1776. doi: 10.1038/nrd3439-c1. URL http://www.nature.com/nrd/journal/v10/n9/full/nrd3439-c1.html.

[16] M. Ragan-Kelley, F. Perez, B. Granger, T. Kluyver, P. Ivanov, J. Frederic, and M. Bussonier. The Jupyter/IPython architecture: a unified view of computational research, from interactive exploration to communication and publication. *AGU Fall Meeting Abstracts*, 44, 2014. URL http://adsabs.harvard.edu/abs/2014AGUFM.H44D..07R.

[17] H. Stitz, S. Luger, M. Streit, and N. Gehlenborg. AVOCADO: Visualization of Workflow-Derived Data Provenancefor Reproducible Biomedical Research. *Computer Graphics Forum (EuroVis '16)*, 2016. to appear.

[18] M. Streit, A. Lex, S. Gratzl, C. Partl, D. Schmalstieg, H. Pfister, P. J. Park, and N. Gehlenborg. Guided visual exploration of genomic stratifications in cancer. *Nature Methods*, 11(9):884–885, 2014. ISSN 1548-7091. doi: 10.1038/nmeth.3088. URL http://www.nature.com/nmeth/journal/v11/n9/full/nmeth.3088.html.

[19] H. Strobelt, B. Alsallakh, J. Botros, B. Peterson, M. Borowsky, H. Pfister, and A. Lex. Vials: Visualizing Alternative Splicing of Genes. *IEEE Transactions on Visualization and Computer Graphics (InfoVis '15)*, 22(1):399–408, 2016. doi: 10.1109/TVCG.2015.2467911.

[20] The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, 499(7456):43–49, July 2013. ISSN 0028-0836. doi: 10.1038/nature12222. URL http://www.nature.com/nature/journal/v499/n7456/full/nature12222.html#supplementary-information.

[21] The Cancer Genome Atlas Research Network. Genomic Classification of Cutaneous Melanoma. *Cell*, 161(7):1681–1696, June 2015. ISSN 1097-4172. doi: 10.1016/j.cell.2015.05.044.

[22] C. Turkay, A. Lex, M. Streit, H. Pfister, and H. Hauser. Characterizing Cancer Subtypes Using Dual Analysis in Caleydo StratomeX. *IEEE Computer Graphics and Applications*, 34(2):38–47, 2014. ISSN 0272-1716. doi: 10.1109/MCG.2014.1.