StratomeX: Guided Visual Exploration for Tumor Subtype Identification in The Cancer Genome Atlas

Marc Streit[1], Alexander Lex[2], Christian Partl[3], Dieter Schmalstieg[3], Peter J Park[4], Nils Gehlenborg[4,5]

[1] Johannes Kepler University Linz, Austria; [2] Harvard School of Engineering and Applied Sciences, Cambridge, MA, USA; [3] Graz University of Technology, Austria; [4] Harvard Medical School, Boston, MA, USA; [5] Broad Institute, Cambridge, MA, USA

Caleydo StratomeX (Lex *et al.* 2012, http://stratomex.caleydo.org) is an open-source visual data exploration system developed to support analysts in the identification and characterization of tumor subtypes in large patient populations such as those studied by *The Cancer Genome Atlas* (*TCGA*). Here we report recent additions to the visualization framework and the creation of a data repository for the system that greatly simplifies access to *TCGA* tumor datasets.

Multiple heterogeneous molecular (mRNA, miRNA, protein, copy number, gene mutations) and clinical datasets can be loaded into *StratomeX* to efficiently generate and confirm hypotheses about tumor subtypes as well as their functional and clinical effects. Patient stratifications (groupings) are visualized as columns where each group is represented by a "brick". Associated data, such as gene expression matrices or copy number states, can be visualized in the context of the stratification using heatmaps, profile plots, histograms or other visualization techniques.

How different stratifications relate to each other can be explored by loading multiple stratifications as columns into the visualization. *StratomeX* uses ribbons of varying width between the columns to encode the magnitude of overlap between groups. Wide bands indicate a strong overlap and thin or absent bands indicate only a small or no overlap. *StratomeX* also provides features that allow analysts to explore clinical and functional differences between potential subtypes. The former is achieved by using dynamically generated Kaplan-Meier plots for each of the candidate subtypes, while for the latter the software integrates pathways and overlays molecular data for the individual candidate subtypes.

These techniques are very well suited for the exploration of relationships between a known set of candidate subtypes. In order to help analysts to identify promising candidate subtypes, *StratomeX* has been extended with computational methods to rank stratifications and identify stratifications that provide corroborating evidence for candidate subtypes. Combined with the ability to rapidly investigate many candidate stratifications, this is an effective approach to get a deeper understanding of large amounts of data.

To simplify access to *TCGA* data, we generate comprehensive data packages for *StratomeX* based on the results of the monthly standard data and analysis runs of the Firehose (http://gdac.broadinstitute.org) data processing pipeline at the *Broad Institute*. A public data repository (http://compbio.med.harvard.edu/tcga/stratomex) allows users to load these data packages into *StratomeX* for visualization and analysis with just a few mouse clicks, thereby making visual analysis of *TCGA* data easily accessible to a wide audience. *StratomeX* also features a data import wizard that enables users to directly load their own datasets and analysis results into the software for comparison with *TCGA* or other public datasets.

Based on initial feedback from collaborators within *TCGA*, we anticipate that the cancer biology community will greatly benefit from the availability of *StratomeX* since the software allows biologists to explore and interpret potential tumor subtypes within the vast *TCGA* dataset more efficiently and in greater detail than with previous approaches.

Acknowledgements

We thank Hans-Jörg Schulz for his contributions and the *National Cancer Institute* (U24 CA143867), the *FWF* (P22902) and the state of Styria (GZ:A3-22.M-5/2012-21) for funding.

References

A Lex, M Streit, H-J Schulz, C Partl, D Schmalstieg, PJ Park and N Gehlenborg, “StratomeX: Visual Analysis of Large-Scale Heterogeneous Genomics Data for Cancer Subtype Characterization”, *Computer Graphics Forum (EuroVis 2012)*, **31**:1175–1184 (2012).