Ordino: visual analysis tool for ranking and exploring genes, cell lines, and tissue samples

Marc Streit[1], Samuel Gratzl[1,2], Holger Stitz[1], Andreas Wernitznig[3], Thomas Zichner[3,*] & Christian Haslinger[3,*]

[1] *Institute of Computer Graphics, Johannes Kepler University Linz, Linz, Austria.*
[2] *datavisyn GmbH, Linz, Austria.*
[3] *Department of Pharmacology and Translational Research, Boehringer Ingelheim RCV GmbH & Co KG, Vienna, Austria.*
** The last two authors should be regarded as joint last authors.*
Email: marc.streit@jku.at, christian.haslinger@boehringer-ingelheim.com

A common approach in data-driven knowledge discovery is to prioritize a collection of items, such as genes, cell lines, and tissue samples, based on a rich set of experimental data and metadata. Applications include, for instance, selecting the most appropriate cell line for an experiment or identifying genes that could serve as potential drug targets or biomarkers. This can be challenging due to the heterogeneity and size of the data as well as the fact that multiple attributes need to be considered in combination. Advanced visual exploration tools – going beyond static spreadsheet tools such as Microsoft Excel – are needed to aid this prioritization process. To address this task, we developed **Ordino** (https://ordino.caleydoapp.org), an open-source, web-based visual analysis tool for flexible ranking, filtering, and exploring of cancer genomics data (**Fig. 1**).

In Ordino the user starts the prioritization by defining a set of items. The item set can be determined by manually entering a list of identifiers (e.g., a list of gene symbols), by selecting a previously saved or predefined list of items, or by uploading a comma-separated file (**Supplementary Fig. 2**). Users can interactively add (i) raw experimental data or metadata stored in a database, like the expression data for a single cell line or the biotype of all listed genes, (ii) dynamically computed scores, such as the average gene expression of tissue samples from a specific tumor type, and (iii) uploaded custom data attributes. We preloaded mRNA expression, DNA copy number, and mutation data from The Cancer Genome Atlas (TCGA) (https://cancergenome.nih.gov) and the Cancer Cell Line Encyclopedia (CCLE) [1], as well as two depletion screen data sets from McDonald et al. [2] and Meyers et al. [3] (**Supplementary Table 1**). A description of the data pre-processing can be found in the supplementary material (**Supplementary Notes**).

The tabular data is visualized using an extended version of our interactive ranking technique LineUp (http://lineup.caleydo.org) [4] (**Fig. 1a and 1c, Supplementary Fig. 3**). Users can rank the table by a single column or by interactively created weighted combinations. The combined column is then shown as a stacked bar highlighting the contribution of individual attributes to the total score. More advanced combinations can be defined interactively or via a scripting interface. The exploration is supplemented with filtering features such as setting cutoff values for numerical attributes or specifying one or more categories in categorical attributes.

Users can select one or more items in the table to explore them using a collection of detail views (**Fig. 1b, Supplementary Notes**). Detail views can be (i) specialized visualizations (e.g., a co-expression plot for comparing multiple genes, an expression vs. copy number plot, or an OncoPrint), (ii) another ranked table (e.g., a list of all tissue samples plus their expression, copy number, and mutation data for the selected genes), or (iii) embedded external resources (Ensembl, Open Targets, etc.). We demonstrate the use and effectiveness of Ordino in two case studies (**Supplementary Notes, Supplementary Figs. 2–10, and Supplementary Video 1**).

Acknowledgements

We thank Christian Lehner for contributions to the implementation of the tool as well as Daniel Gerlach, Markus Bauer, and Anita Steiner for their contributions to data preparation and data handling. This work was supported by the Austrian Science Fund (P27975-NBL), the State of Upper Austria (FFG 851460), and Boehringer Ingelheim RCV.

Author Contributions

M.S., S.G., T.Z., and C.H. jointly conceived the project. T.Z., S.G., and M.S. designed the software and methods. S.G. and H.S. implemented the software. A.W. and T.Z. worked on the data integration. C.H. A.W. and T.Z. developed requirements and use cases. M.S. and T.Z. wrote the manuscript, with contributions from S.G., H.S., A.W., and C.H. M.S., T.Z., and C.H. oversaw the project.

Competing Financial Interests

The development of Ordino was supported financially by Boehringer Ingelheim RCV as part of a research collaboration with Johannes Kepler University Linz. M.S. and S.G. are shareholders of datavisyn GmbH.

References

1. Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-607 (2012).
2. McDonald III, E. R. et. al. Project DRIVE: A Compendium of Cancer Dependencies and Synthetic Lethal Relationships Uncovered by Large-Scale, Deep RNAi Screening. *Cell* **170**, 577-592.e10 (2017).
3. Meyers, R. M. et. al. Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. *Nature Genetics* **49**, 1779–1784 (2017).
4. Gratzl, S., Lex, A., Gehlenborg, N., Pfister, H. P., Streit, M., LineUp: Visual Analysis of Multi-Attribute Rankings. *IEEE Trans Vis Comput Graph* **19**, 2277-2286 (2013).

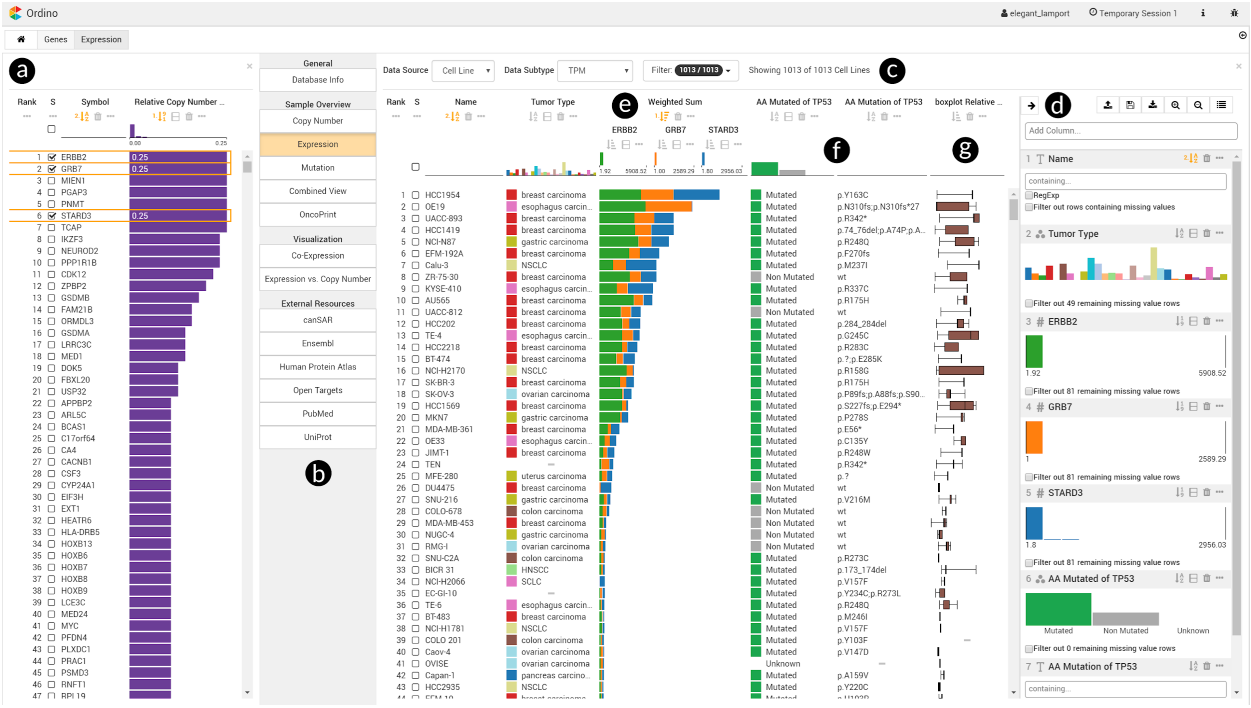

Figure 1. **Ordino state showing genomic alteration and gene expression data of breast cancer cell lines.** In the left panel (a), all human protein-coding genes are ranked by their relative amplification frequencies in a set of about 60 breast cancer cell lines. The researcher selects three of the most frequently amplified genes (*ERBB2*, *GRB7*, *STARD3*) and opens a detail view (b) on the right, displaying the expression of these genes across a set of over 1,000 cell lines (c). The side panel (d) that is shown on demand enables the user to define a ranking hierarchy and to set filters. Combining the three gene expression columns to stacked bars (e) allows cell lines in which one or more of these genes might play an important role to be identified. Next, the researcher adds two columns (f) that represent the mutation status and actual mutations of the cancer gene *TP53*. Furthermore, a column visualizing the distribution of copy number values across ~15 frequently amplified breast cancer genes is loaded (g). Based on the added information, the researcher gains various insights, including that the cell line with the highest expression of the three genes of interest is HCC1954, which has a p.Y163C *TP53* mutation.

Link to Ordino state shown in this figure: http://vistories.org/ordino-teaser-figure