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

Marc Streit[1,2], 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

Abstract

Ordino is a web-based analysis tool for cancer genomics that allows users to flexibly rank, filter, and explore genes, cell lines, and tissue samples based on pre-loaded data, including The Cancer Genome Atlas (TCGA) and the Cancer Cell Line Encyclopedia (CCLE), as well as manually uploaded information. A core component of Ordino is an interactive tabular data visualization that facilitates the user-driven prioritization process. Detail views of selected items complement the exploration. Findings can be stored, shared, and reproduced via the integrated session management.

Availability and Implementation

Ordino is publicly available at https://ordino.caleydoapp.org.
The source code is released at https://github.com/Caleydo/ordino.

1 Introduction

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 and Google Spreadsheets — are needed to aid this prioritization process. However, powerful general-purpose tools such as Tableau and Spotfire are insufficient, as they can be too difficult to use for a non-expert and have limitations with respect to the aggregation and visualization of genomics data. To fill this gap, we developed Ordino, an open-source, web-based visual analysis tool for flexible ranking, filtering, and exploring of cancer genomics data (**Fig. 1**).
Furthermore, we demonstrate the use and effectiveness of Ordino in two case studies (**Supplementary Notes**, **Supplementary Figs. S2–S10**, and **Supplementary Video**).

2 Software Description

The main interface of Ordino is a tabular visualization that presents multi-attribute data (shown as columns) for a set of items (shown as rows). The basic workflow comprises three steps. (1) Select or define a list of items consisting of genes, cell lines, or tissue samples. (2) Interactively add data columns and rank, filter, and explore items based on them. (3) Obtain detailed information about one or more items of interest by selecting them in the table and opening various detail views.

2.1 Item List Definition and Adding Data

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 (**Fig. S2**).

Afterwards, users can interactively add (i) raw experimental data or metadata stored in the Ordino 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 the Ordino database with 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; Barretina, 2012), as well as two depletion screen data sets from McDonald (2017) and Meyers (2017) (**Supplementary Table S1**).

2.2 Interactive Visualization of Rankings

The tabular data is visualized using an extended version of our interactive ranking technique LineUp (http://lineup.caleydo.org; Gratzl, 2013) (**Fig. 1a and 1c**, **Fig. S3**).

Users can change the visual representation of columns on demand. Single numerical attributes, for instance, can be visualized using bars, varying brightness, or as circles whose sizes are proportional to the data values. Columns containing dynamically computed scores, which aggregate data from multiple entities, can be additionally shown as row-wise box plots or heat maps. The exploration is supplemented with filtering features such as setting cutoff values for numerical columns or specifying one or more categories in categorical columns.

Furthermore, users can rank the table by a single column (e.g., the value of numerical column or the median of box plot column) or by interactively created weighted combinations of two or more columns. 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.

2.3 Detail Views

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.). Newly opened detail views appear on the right side of the interface, causing the previously active view to be shown in a more compact format on the left.

2.4 Reproducibility and Sharing

A core feature of Ordino is its ability to let users store, revisit, and share findings at any time during an analysis session. To achieve this, Ordino requires users to log in before using the system. To avoid a tedious registration process for creating the login credentials, the system auto-generates temporary accounts. By default, analysis sessions are temporary, which means that they are only stored in the local cache of the browser. Users can deliberately make sessions persistent, which moves the sessions to the database on the Ordino server. Persistent sessions can be shared by simply copying the URL shown in the browser. When opening a link to a persistent session, the system will restore the exact state of the analysis, including the history of all previous steps (Gratzl, 2016). Note that the states shown in **Fig. 1** as well as **Supplementary Figs. S2-S10** can be reproduced by following the links provided below the figures.

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.

Funding

This work was supported by the Austrian Science Fund (P27975-NBL), the State of Upper Austria (FFG 851460), and Boehringer Ingelheim RCV. 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).
5. Gratzl, S., Lex, A., Gehlenborg, N., Cosgrove, N., Streit, M., From Visual Exploration to Storytelling and Back Again. *Computer Graphics Forum* **35** (3), 491–500 (2016).

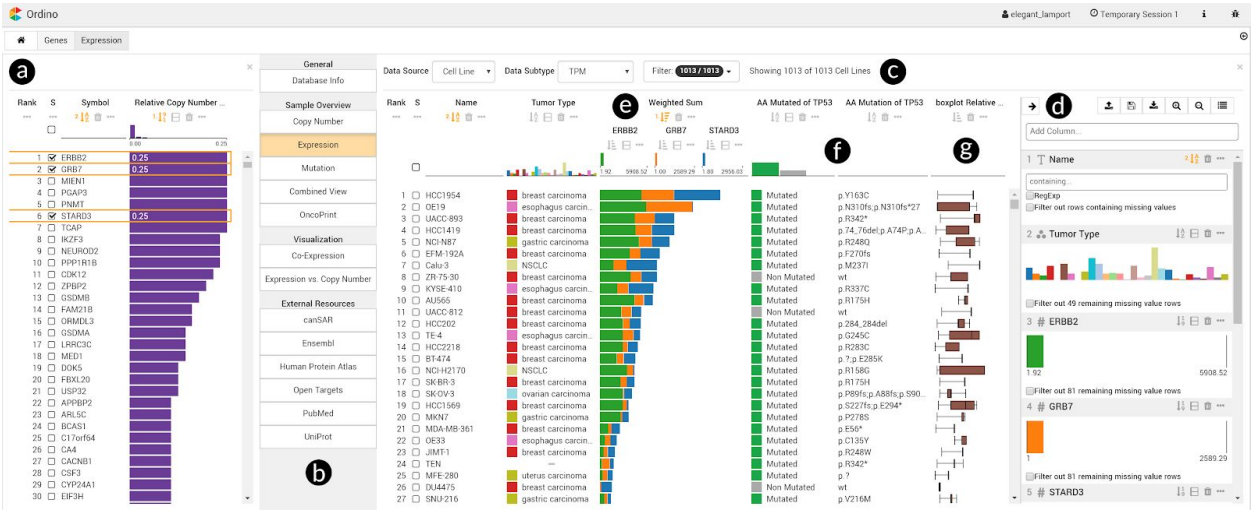

Figure 1. **Ordino state showing genomic alteration and gene expression data of breast cancer cell lines.** In the left panel (a), all human protein-coding genes are ranked by their relative amplification frequencies in a set of about 60 breast cancer cell lines. The researcher selects three of the most 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), which 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).

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