Visual analytics framework for survival analysis and biomarker discovery

Jaka Kokošar[1], Cagatay Turkay[2], Luka Avsec[3], Miha Štajdohar[3], Blaž Zupan[1,3,*]

1 Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia
2 Centre for Interdisciplinary Methodologies, University of Warwick, Coventry, UK
3 Genialis Inc., Boston, Massachusetts, USA
4 Department of Education, Innovation and Technology, Baylor College of Medicine, Houston, Texas, USA

* blaz.zupan@fri.uni-lj.si

Abstract

We introduce a visual analytics methodology for survival analysis, and propose a framework that defines a reusable set of visualization and modeling components to support exploratory and hypothesis-driven biomarker discovery. Survival analysis—essential in biomedicine—evaluates patients' survival rates and the onset of medically relevant events, given their clinical and genetic profiles and genetic predispositions. Existing approaches often require programming expertise or rely on inflexible analysis pipelines, limiting their usability among biomedical researchers. The lack of advanced, user-friendly tools hinders problem solving, limits accessibility for biomedical researchers, and restricts interactive data exploration. Our methodology emphasizes functionality-driven design and modularity, akin to combining LEGO bricks to build tailored visual workflows. We (1) define a minimal set of reusable visualization and modeling components that support common survival analysis tasks, (2) implement interactive visualizations for discovering survival cohorts and their characteristic features, and (3) demonstrate integration within an existing visual analytics platform. We implemented the methodology as an open-source add-on to Orange Data Mining and validated it through use cases ranging from Kaplan–Meier estimation to biomarker discovery. The resulting framework illustrates how methodological design can drive intuitive, transparent, and effective survival analysis.

Author summary

When studying diseases like cancer, it is important to understand how factors such as genes or treatments influence how long patients live after medical intervention. Survival analysis is a well-established statistical and bioinformatics approach for uncovering such insights, but it often requires advanced programming skills, which can be a barrier for many clinicians and life science researchers. To help overcome this, we developed a visual analytics tool that makes the analysis of censored data easier, more interactive, and accessible. Instead of writing code, researchers can use our framework to assemble analysis components into workflows through visual programming. They can explore their data by comparing survival curves, identifying meaningful patient subgroups, and discovering potential biomarkers. The tool is implemented as part of Orange Data

Mining, a free and open-source data analytics platform. We tested it on real-world cancer datasets to demonstrate how easily and quickly researchers can prototype powerful data analysis workflows. Our goal is to make survival analysis more accessible, empowering researchers to generate and test hypotheses without needing to program. Our goal is to make survival analysis more accessible, empowering researchers to generate and test hypotheses without programming, and contributing to the broader democratization of data science.

1 Introduction

Survival analysis is a set of statistical methods used to determine the life expectancy of an investigated population. Survival analysis became particularly valuable in medical research, where researchers use survival data to observe the life expectancy of the investigated population. Notably, in a clinical setting, we refer to survival time as a time to some clinical event of interest, like relapse or progression of a disease [1,2]. In biomedicine, these covariates are understood as potential biomarkers and targets for treatment development (see Fig. 1). To improve survival outcomes, it is essential to understand how patients with different health conditions and genomic predispositions respond to various treatments [3]. Furthermore, the identification of novel markers is crucial for the development of personalized medicine, a rising trend in evidence-based medicine [4].

Fig 1. An example of a Kaplan-Meier survival plot for two gene expression-dependent conditions associated with patient survival. **(a)** The difference between survival curves of gene *CD14* is not evident. **(b)** The survival curve is substantially higher for a group of patients with highly expressed gene *MAP7*. We could say that *MAP7* is hence a better biomarker for survival. In biomarker discovery, one of the tasks is to rank features, e.g. gene mutation status, gene expression values, and protein levels, according to the degree of separation between survival signatures given gene expression.

Survival analysis tools should therefore support the inference of survival, the study of changes in survival over time, and the identification of cohorts of subjects with substantial differences in survival. Cohorts are characterised by features, and tools should support users in finding and exploring the most informative ones. Most of these functionalities are currently available through libraries in scripting environments such as R and Python (e.g. `survival` [5], `scikit-survival` [6], and `lifelines` [7]), where they integrate well with other data visualisation and modelling techniques. However, due to the lack of computer programming experience, these are difficult to access for biomedical domain experts. Moreover, scripting does not lend itself to exploratory data analysis and does not facilitate communication with the domain experts involved.

These limitations of scripting environments for survival analysis call for a complementary visual analytics approach, which combines interactive visualizations, integrated analysis methods, and visual programming. Visual analytics [8] enhances data exploration by allowing users to interactively manipulate and interpret data through graphical interfaces, while interactive visualizations [9] provide dynamic, real-time feedback to reveal patterns and trends. Visual programming (e.g., [10]) enables intuitive, workflow-based analysis, reducing the need for coding and making complex methods more accessible—an essential advantage for medical data analysis, where interpretability, usability, and collaboration between clinicians and data scientists are critical.

Here, we report on our visual analytics approach to survival analysis. We aim to design an intuitive but powerful tool with a smooth learning curve and interactive

interface that motivates exploratory data analysis, explanation, and discovery, supporting a broad scope of potential applications with as few components as possible. Our contributions are the implementation of survival analysis within our existing workflow-based system and the design of interactive elements that support cohort definition, cohort discovery, and discovery of biomarkers. We have designed the components of the new system so that they integrate well with the existing toolbox for visual analytics in the Orange Data Mining software[1] [11, 12], adding to the richness of both available visualizations and complementary procedures for data analysis.

The manuscript begins with an overview of related methods, including survival analysis, visual analytics, and workflow-based approaches to data analysis. We then describe our visual analytics approach, which integrates interpreted workflows, interactive visualisations and the use of computational models. To demonstrate the usefulness of the designed analytical components and tool, we present a set of typical survival analysis and biomarker discovery tasks and show how we can solve them using our proposed visual analytics toolbox.

2 Related work

In application, the data analysis framework we present here relates to concepts from survival analysis and discovery of biomarkers. In the development of exploratory data analysis approach, our work is related to visual analytics and development of pipelines that use workflows and visual programming, and combine machine learning and interactive data visualization. In particular, for the latter, our work borrows from concepts developed within Orange Data Mining framework.

2.1 Survival analysis for biomarker discovery

Biomarkers play a pivotal role in linking a patient's molecular characteristics to clinical outcomes, helping to predict the success of therapies and guide decision-making. In biomedicine, the covariates with a significant impact on patient survival are potential markers, characteristics of a biological system we can measure objectively and use as an indicator of the system's state [13]. In oncology, for instance, biomarkers can distinguish patients likely to benefit from specific therapies, while sparing others from ineffective treatments. Markers may be clinical, related to the patient's symptoms, or biological, related to measurements on the molecular level, like the concentration of a specific protein or expression of a particular gene [14, 15]. Identifying such markers is crucial for the development of personalized medicine that uses data to deliver targeted, patient-specific treatments [16].

2.2 Key visual analytics concepts

Here we discuss a number of key visual analytics principles to facilitate data analysis tasks involving large numbers of data features and multiple data sources. We then highlight key relevant techniques and applications in related domains.

Interplay of computation and human input – Visual analytics is often referred to as a dialogue, an interplay, between the analysts and the data, through the facilitation of computational tools and interactive visualizations [8]. Interaction and integrated computational methods that are triggered by human inputs have been a key facilitator of dialogues that are mechanisms for analysis and reasoning [17]. For instance, such integrated approaches have been used to cluster large collections interactively [18], or for high-dimensional data analysis [19]. Through flexible programmable visual analytics

---

[1]https://orangedatamining.com

workflows, our approach in this paper facilitates such interactions in a more structured manner and makes opportunities for the analysts to interact more explicit.

Input variation & result comparison – A key mechanism often used in visual analytics methods is to explore the variation in the results in response to variation of data and parameters that control an analytical operation or a model [20]. Visual analytics approaches have shown to effectively support comparative analysis for modelling in the context of classification [21], or in dimension reduction [22]. In this work, we build on this notion of input-output variation exploration within the context of survival analysis.

Visual analysis workflow & interactive propagation – A number of visualization approaches and systems have made the analysis workflow explicit and available for further interaction starting with the seminal VisTrails work by Bavoli et al. [23]. More recently, similar visual programming and visual workflow methods have been used in generating visualization design variations [24]. While these methods are primarily focused on visualization generation, they pave the way for more general, flexible and transferable data analysis workflows where interactive alterations at an intermediate stage of the analytical process can be propagated further through the analytical tools and visualizations that communicate the results.

2.3 Visual analytics for survival analysis

The challenges around survival analysis have been approached by visual analytics methods in a number of studies. In their Caleydo system, Lex et al. demonstrates how Kaplan-Meier (KM) plots could be associated as visual indicators to characterize sub-cohorts in a visual analysis of cancer subtype stratification process [25]. Corvo et al. makes survival analysis a core analysis focus and demonstrates how KM plots could be generated in response to interactive selections in their SurviVIS tool [26]. Marai et al. employ KM plots and Nomograms in their visual analytics system designed to explore and detect similar head and neck cancer patients [27]. However, these solutions primarily focus on a pre-defined tasks (or disease types) and do not support user-defined analytical pipelines. Our work takes these methods further by integrating a combination of analytical methods and interactive visualizations within a visual programming paradigm.

Several web-based tools with a focus on survival analysis have been proposed to assist biomedical researchers in genome-based biomarker discovery [28, 29]. These tools provide a graphical user interface combining genomics data, clinical data and analysis tools designed specifically for mining survival-related gene markers across different data sources and tumor types. For instance, Maller et al. used *Kaplan-Meier Plotter* to evaluate the association between *PLOD2* gene expression values and relapse-free survival [30] and Ingebriktsen et al. used it to evaluate a six-gene signature concerning recurrence-free survival in breast cancer patients [31]. While the proposed web-based solutions can make exploring data simpler, they do not fully implement the key concepts of visual analytics (discussed in Sec. 2.2) and this dashboard-based approach often struggle to adapt to real-time insights or changing hypotheses. In addition, as discussed by Idogawa et al. [32], there are concerns about the data integrity of the proposed systems and more importantly, the underlying analysis is often closed-source. This ultimately leads to a lack of trust within the research community and hinders the reproducibility of research results.

2.4 Orange data mining framework

We have implemented the visual analytics approach to survival analysis we describe here as an add-on to the Orange data mining framework [10, 12, 33]. Orange is a

general-purpose data analysis tool. Data analysis pipelines in Orange are assembled through visual programming and consist of widgets—components with a graphical user interface—that perform distinct data analysis or visualization tasks. We entirely borrow the visual analytics concept from Orange. The principal innovation of the work presented here is designing components for survival analysis so that they can collectively be applied to a wide range of tasks and seamlessly integrate with other existing Orange components. Notice that we could develop a similar survival analysis framework in different general purpose workflow-based data analysis systems, including popular RapidMiner[2] or KNIME[3]. We choose Orange, though, since it uses interpreted workflows, where any change in the settings of the components or any user's action such as data or feature selection immediately propagates through the pipeline, a requirement for the implementation of visual analytics. Orange is also different from the two mentioned workflow systems in supporting interactive visualizations. Despite all these differences, our choice for Orange is biased: several authors of this report are its principal designers and contributors.

3 Key analytical approaches to survival analysis for biomarker discovery

The literature shows that current practice uses a limited range of computational and data visualization methods for the analysis of survival data. These methods can be used in combination to address many different task in biomarker discovery where the overreaching goal is to identify or validate data features that are best associated with patient survival. By far the most widely used and established methods of survival analysis are the Kaplan-Meier estimator and Cox regression model. Their adaptation and importance are evident in their status among the most cited articles [34] as well as more focused reviews and guidelines on their practical use in biomedical cancer research [1, 35–39].

3.1 Characterization of key goals and tasks

Based on the above-mentioned reviews and guidelines, we have externalized the common goals (**G**) and tasks (**T**) typically employed in different analytical scenarios. While we do not delve into the details and definitions of survival analysis methods, we do, however, outline their use in the context of biomedical application and how they can be applied across diverse analytical scenarios in oncology.

G1 – **Understand the distribution of survival times.** Survival data is most often presented as a survival curve, which illustrates how survival prognosis changes over time. The most fundamental step in survival analysis is the construction of the survival curve.

> T1 **Estimate and plot the survival curve.** The most common approach to survival curve estimation is the Kaplan-Meier method [40]. The calculated survival probabilities are then plotted against time to provide a visual summary of survival probabilities over time.

> T2 **Summarize key statistics.** Kaplan-Meier curves can be thought of as a form of descriptive statistics for survival data [41]. To add context to the visual representation, Kaplan-Meier curves are often juxtaposed by additional

---

[2]https://rapidminer.com/
[3]http://knime.com

statistics, such as total number of samples, number of samples with observed events, median survival time, and confidence intervals for survival probabilities, among others. 189 190 191

G2 – **Analyze the differences in survival between different cohorts.** Survival 192 curves provide a basis for comparing the survival experience of different sample 193 groups. 194

T3 **Stratify the data by variables of interest.** Kaplan-Meier analysis does 195 not directly account for the impact of various variables on survival. However, 196 it is often used to conduct a univariate analysis for comparing survival curves 197 of groups defined by different categorical variables. For continuous variables 198 researches have to first define boundaries defining different groups. 199

T4 **Compare survival curves of different groups.** To inspect differences in 200 survival between groups, researchers typically plot multiple survival curves on 201 the same graph [41]. A visual inspection can provide an initial indication of 202 whether a particular group has, for example, a more favorable prognosis 203 compared to others. Visual comparison is further supported by statistical 204 tests, predominantly log-rank test, to quantify this difference [42]. 205

G3 – **Interpret the effect of data features on survival outcomes.** Survival 206 analysis can be extended to measure the impact of different data features on 207 survival rates. 208

T5 **Develop a prognostic model of patient survival.** Researchers often 209 model the impact of data features on survival in either univariate or 210 multivariate settings using Cox regression analysis [43]: 211

T5.1 **Univariate cox regression analysis** is often used a feature screening 212 technique or to observe how each data features affects survival 213 independently. 214

T5.2 **Multivariate cox regression analysis** is used to observe how the 215 significance of features shifts when they are adjusted for other data 216 features. 217

T6 **Model interpretation and explanation.** By examining the hazard ratios 218 and confidence intervals derived from the prognostic models, the researchers 219 assess the strength and direction of associations between data features and 220 survival outcomes, identifying which factors have significant prognostic value. 221

T7 **Model application and visualization.** Researchers often present 222 prognostic models graphically as Nomograms to enable individualized survival 223 predictions [44]. Nomograms translate prognostic models into an accessible 224 visual format that aids in their interpretation and makes it easier for 225 clinicians to apply them in practice. 226

G4 – **Develop personalized risk-based patient stratification.** Patient 227 stratification exemplifies the practical application of prognostic models, allowing 228 identification of high-risk groups that may, for example, benefit from targeted 229 interventions. 230

T8 **Assign risk scores to individual patients.** A common approach is to 231 utilize the coefficients derived from the developed prognostic model to 232 calculate the risk score (sometimes referred to as the prognostic index) for 233 each patient [45]. These risk scores are then used to stratify patients into 234

different risk groups. This stratification process is crucial for identifying subgroups of high- and low-risk patients with distinct prognoses that can inform personalized treatments or biomarker development and validation.

G5 – **Validation of developed prognostic or risk-based stratification models.** In biomarker discovery, models with strong discriminatory power can help validate the clinical relevance of the identified biomarkers, showing that they provide meaningful insight into patient outcomes [46].

    T9 **Evaluating the discriminative power.** It is common practice to visually inspect the separation between the curves for high- and low-risk groups, which indicates good discriminatory power. More analytical another common approach is to use the concordance index (C-index) [47], which quantifies how well the model distinguishes between individuals who experience the event of interest and those who do not.

3.2 Typical analytical pipelines

This goal-oriented abstraction can generally be used for analyzing any type of survival data. However, typical biomarker discovery pipelines tend to be more complex and multifaceted. Also, how survival analysis is used can vary; survival data can be the primary concern of the analysis or just as a validation step, where the results are evaluated in relation to clinical outcomes. In either case, this goal-oriented abstraction can be applied across various analytical scenarios; each goal can function as a standalone analysis or be combined in different ways to suit specific analysis needs.

For example, in their clinical trail, Lawrence et al. [48] set out to evaluate the relationship between expression of the DNA mismatch repair (MMR) gene *MLH1* and survival outcomes in patients with resected pancreatic cancer who received adjuvant chemoradiation. They set out to analyze survival distributions of study population, evaluate the differences in survival of different cohorts, and assess the prognostic impact of potential gene marker (**G1**, **G2**, **G3**). Their data analysis pipeline includes stratifying patients based on median expression of *MLH1* (**T3**), estimating survival curves for each cohort (**T1,2**), evaluating how significant is the differance in survival between cohorts (**T4**) and estimating the prognostic value of gene *MLH1* in univariate and multivariate setting (**T5,6**). We can observe similar data analysis pipelines in several different studies for biomarker discovery in oncology [49–52].

Expanding on this concepts, in their analysis pipeline, Long et al. [53] developed and validated a multi gene-based prognostic index to stratify patients into low- and high-risk groups (**G4**, **G4**, **G5**). They first screened survival related genes (**T5.1**) and selected a four-gene prognosis model (**T5.2**) that assigns risk scores to individual patients. Their analysis includes a validation of newly developed prognosis model in relation to established clinical factors (**T6**) and construction of a predictive nomogram (**T7**). They also validate a risk-based patient stratification by splitting patients into low- and high-risk groups (**T3**), calculating survival curve for each group (**T1**) and test if there are significant differences (**T4**) in survival rates (**G1**, **G2**). Similary as before, we can observe this kind of analytical approaches across various different studies [54–56].

4 Integrating survival analysis, workflows and visual programming

Our approach to support survival analysis consists of decomposing the key analytical approaches (**T**) discussed in Sec. 3 into independent and reusable components, each

addressing a specific analytical or exploratory function. These components can then be assembled into diverse workflows through visual programming, allowing users to construct pipelines that meet the needs of their specific analytical scenarios (**G**). By grounding our design in this task-oriented approach, we have created a survival analysis framework that users can use to build custom data analysis pipelines ranging from simple exploratory analysis of survival curves to complex pipelines for multi-gene biomarker discovery, all within a visual programming environment. Below, we describe this general framework and list the components we have implemented for survival analysis.

4.1 Data analysis framework

The following elements from Orange were critical in supporting our design:

Widgets are Orange's components in data analytics workflows that receive data on input, process or visualize the data, and transmit the results or user selections to the widget's output. Widgets encapsulate data processing and visualization functions and allow us to combine tasks (**T**) and goals (**G**) into independent, reusable components:

  Kaplan-Meier plotter (T1-4) is a component that we can use to analyse survival function (**G2**) and assess the differences of survival between different cohorts (**G2**).

  Cox Regression (T5-6) component allows us to estimate the effect different data features have on survival of our patients (**G3**).

  Rank Survival Features (T5.1) is a very specialized component that ranks data features based on univariate cox regression analysis.

  Cohorts (T8) is what we can use for the development of prognosis model and assign risk score to each patient (**G4**).

  Survival Nomogram (T7) enables graphical representation of developed prognostic model and supports its interpretation (**G3**).

  We then combine these components into workflows using visual programming.

Workflows are ordered collections of widgets that combine different analytical tasks and goals into user defined analytical pipelines, like the ones discussed in Sec 3.2. For example, a workflow from Fig. 2 starts with the widget for loading the data and then branches into pipelines for (1) exploring biomarker significance, (2) risk-based patient stratification, and (3) exploring similarity between data instances. This exemplifies how a visual programming paradigm supports iterative assembly of workflows, allowing users to test and refine their hypothesis in real time. Orange workflows are read from left to right, with inputs to widgets on the left and outputs on the right. Note that in this workflow *Kaplan-Meier Plot* widget is reused to display the difference in survival of either the feature-defined cohorts, cohorts defined by risk-based stratification model or between group of patients selected in the t-SNE visualization.

Interactive interface. Orange widgets are interactive. The first and most obvious type of interaction is the selection of display options in the widget's control area, as in the *Kaplan-Meier plot* widget from Fig. 3a. More important for visual analytics, however, are interactive displays of data. For example, in the *Distributions* widget, the user can select any part of the distribution (Fig. 3b) and instruct the widget to output only the subset of the selected data instances or pass selection indicator with the data on the output (Fig. 3c).

Fig 2. An example workflow to explore cohort formation based on clustering and biomarker discovery. The workflow starts by loading the data (*Datasets* widget). The workflow progresses from **(1)** stratification of patients based on specified cutoffs value of continuous features (using *Distributions*, *Kaplan-Meier Plot*) to more complex **(2)** cohort formation based on several data features (the branch with *Rank Survival Features*, *Cohorts*). For comparison, **(3)** we plotted the map of patients (*t-SNE*), where patients are grouped by similarity of their feature-profiles. The survival curve for the cluster of patients selected in t-SNE is then compared to all other patients (*Kaplan-Meier Plot (1)*).

Fig 3. We zoom in on the top branch in workflow from Fig. 2. We use the *Distribution* widget to plot the distribution of the tumor size feature. The distribution plot is interactive, for example, we select data instances where the tumor size is less than 35mm **(b)**. This selection is passed to *Kaplan-Meier plot* and provides an interactive browser to characterize the survival of a selected group **(c)**. Different analysis options are available to the user, for example plotting of the confidence interval **(a)**.

Signal processing. To support interactive visual analysis, widgets emit their output whenever their data selections or parameters change. They communicate user's selections to downstream pipeline, where components instantaneously react to changes immediately. In Fig. 2, for example in the middle branch, a combination of *Rank Survival Features*, *Cohorts* and *Kaplan-Meier Plot* provide a interactive way for users to select top-ranked data features that are then used for construction of risk-based model, split patients into low- and high-risk cohorts and plot the survival curve of each cohort (see Fig. 4).

Fig 4. Focusing on the middle branch of workflow from Fig. 2. We use the *Rank Survival Features* to prioritize features most relevant **(a)**. Data that is passed to *Cohorts* widget contains only two selected features to construct a risk-based model and stratify patients into low- and high-risk cohorts **(b)**. This information is then passed to the *Kaplan-Meier Plot* that estimates, plots and assess how significantly different are the two survival curves **(c)**.

5 Evaluation Through Case Studies

To evaluate our approach, we designed a series of case studies based on recently published research in survival analysis and biomarker discovery. These case studies test the ability of our approach and developed software to address key challenges in the field, assessing whether the provided components are sufficient, whether they can be effectively combined to solve relevant analytical tasks, and whether the approach is practical for real-world applications. Through these examples, we demonstrate how interactive visual analytics facilitates exploratory data analysis, hypothesis generation, and discovery of survival-related biomarkers.

5.1 Data

In all case studies, we use data from the METABRIC database. The METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) dataset comprises genomic and clinical data from nearly 2,000 breast cancer patients, including gene expression profiles, clinical features, and survival outcomes [57]. The dataset is publicly available and widely used in studies of breast cancer prognosis and biomarker discovery.

We use this dataset as a representative example of a large, complex dataset with multiple data types and survival outcomes, which is typical of many studies in oncology and survival-related biomarker discovery. An example of the METABRIC data as provided in Orange is shown in Fig. 5.

Fig 5. A small sample of the METABRIC dataset available in Orange, where each row represents a primary breast cancer patient characterized by (**1**) clinical features (such as age at diagnosis and tumor stage), (**2**) gene expression values (e.g., *KRAS*, *RERE*, *PHF7*), (**3**) and clinical outcomes (overall patient survival).

5.2 Exploring prognostic role of gene KRAS

Hwang et al. [50] have recently investigated the prognostic role of the mRNA expression of a proto-oncogene *KRAS* in breast cancer with the application of survival analysis on two different datasets (TCGA [58] and METABRIC [57]). The authors set out to find if the expression of *KRAS* is prognostic in breast cancer. *KRAS* is a well-known driver of lung, prostate, and colon cancer. To investigate the prognostic impact of *KRAS*, Hwang et al. stratified patients into high- and low-expression groups based on median expression levels (**T3**). They plotted Kaplan-Meier survival curves (**T1**) for overall survival of both cohorts and assessed statistical significance using log-rank test (**T4**). They incorporated multivariate Cox proportional hazards models to assess the independent prognostic value of *KRAS* while adjusting for key clinical variables, such as age, tumor stage, and receptor statuses (**T5-6**). The analysis showed significantly worse outcomes for the high-expression group and confirmed high *KRAS* expression as an independent indicator of poorer prognosis across both datasets.

Fig 6. Evaluation of the prognostic potential of individual genes. This simple workflow effectively reproduces the study of Hwang et al. [50]. The METABRIC data set is pushed through the *Distributions* widget to allow for an interactive selection of gene expression threshold, and survival probability is visualized in the *Kaplan-Meier Plot* widget. Different gene expression thresholds can be explored at a click of a button.

In the workflow in Fig. 6 we show how can we readily reproduce steps of their analysis with our proposed survival analysis components. Workflow starts with the *Genes* widget, a part of Orange's Bioinformatics add-on, we selected a *KRAS* gene. The *Genes* widget also shows information from the Entrez Database—thus simplifying gene exploration within Orange. We use *Discretize* widget to specify the median expression as the threshold between high- and low-expressors. In the upper branch of the workflow we use *Kaplan-Meier Plot* widget to visualize and compare survival curves. The lower branch of the workflow performs *Cox regression* analysis for *KRAS* expression levels adjusted for other known clinical factors. Users can compare the results of both analysis side by side. In this workflow we focused only on METABRIC dataset, however, the analysis can be easily repeated with different dataset by loading the data with *File* widget. The analysis above focused only on *KRAS* gene, but how would we identify related genes that may be just as important?

5.3 Exploring the related genes

The RAS pathway is a central regulator of cell growth and differentiation in normal tissues, and perturbations of the pathway frequently result in cancer or developmental syndromes (RASopathies) [59]. It is natural to ask the question of the relative

importance of other genes in the RAS pathway. Finding such genes can be easily accomplished using a combination of existing and our new widgets (Fig. 7). 386 387

In Fig. 7a, we use *Genes* and *Gene Sets* widgets to pull in the RAS pathway from the KEGG pathway database and keeping only set of genes from the RAS pathway. Like before, we use the *Discretize* widget to set a threshold between high- and low-expressor groups for all the genes in the gene set (**T3**). We use the *Rank Survival Features* widget for ranking genes in the entire gene set and prioritize ones that best separate cohorts according to survival (**T5.1**). We select five top-ranked genes and observe that the gene that is best at separating patients by survival is *FLT3*. 388 389 390 391 392 393 394

The workflow continues with Fig. 7b, we pass the selected genes back into the *Genes* widget to get more context around these genes. For example, we can see that the protein product of *FLT3* expression functions as a receptor tyrosine kinase normally expressed on hematopoietic stem cells. Its mutations are well studied in acute myeloid leukemia [60]. In addition, the upregulation of *FLT3* expression has also been observed in lymph node metastases of patients with primary breast cancer [61]. We observe this also in our exploratory analysis (*Kaplan-Meier plot* in the top branch), where we observe that overexpression of *FLT3* leads to a favorable prognosis (**T1-4**). However, individual genes seldom function in isolation. Understanding the broader molecular circuitry (beyond single-gene biomarkers) can help us better understand the complex nature of diseases such as cancer. The bottom branch of this workflow uses the *Cohorts* widget to develop a risk score model based on all of the selected genes (**T8**). We then use the *Kaplan-Meier Plot* widget to plot and compare survival curves of high- and low-risk groups (**T1-4**). The resulting workflow thus provides a browser of genes and sets of genes that we can evaluate individually or in combination. 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409

Fig 7. (**a**) By using *Gene sets*, we have reduced the initial data to include only those genes that are participating in *Ras signalling pathway* from KEGG pathway databse. (**b**) We conduct standard survival analysis tasks on those genes through *Rank Survival Features* and *Cohorts*. Visualizing survival curves in *Kaplan-Meier plot* allows us to explore newly defined cohorts.

5.4 Cohort formation based on similarity of data instances 410

Our exploratory analysis of METABRIC dataset above focused heavily on survival analysis components. We would like to demonstrate that the goals and tasks discussed in Sec. 3 can also be used interchangeably with different analysis goals. For example, one might want to explore the similarity of data instances (patients) based on how similar their gene expression profiles are. But then, based on identified clusters, the goal might be to describe and compare different clusters in terms of their survival experiences (**G1 and G2**). The key question is: what characterizes those clusters? 411 412 413 414 415 416 417

In Fig. 8 we demonstrate the utility of *Hierarchical Clustering* to group patients into three distinct clusters. User selected clusters are propagated to *Kaplan-Meier Plot* where we explore the difference of survival functions of patients in different cohorts (**T1-4**). Interestingly, we observe that for the cluster of patients in red survival rates drop rapidly compared to other clusters at the beginning. To investigate further, we extend the workflow and use *Box Plot* to reveal features that characterize patients of each cluster. The exploration of our dataset revealed that patients in the red cluster are defined by the overexpression of gene *PSAT1*. This finding is interesting, since the gene *PSAT1* is indeed a subject of an active research in breast cancer with similar reports [62, 63]. 418 419 420 421 422 423 424 425 426 427

Again, notice that any change in interactive cluster selection in *Hierarchical Clustering* propagates through the workflow and updates the content of both *Box Plot* 428 429

Fig 8. Identifying groups of patients on the METABRIC dataset. The workflow computes distances between patient profiles (*Distances* widget) and clusters the patients accordingly in *Hierarchical Clustering*. The user can use the dendrogram in *Hierarchical Clustering* to choose the clusters by brushing or defining a cut-off distance. We display survival curves of identified clusters of patients in the *Kaplan-Meier Plot*. The workflow uses *Box Plot* to characterize given clusters by providing a sorted list of features, which can help in the formation of new hypotheses. In the particular case displayed in the figure, we observe that expression levels of gene *PSAT1* are highest in cluster C2.

and *Kaplan-Meier Plot* and thus enable on-the-fly hypothesis formation. 430

6 Discussion 431

Our goal was to design a visual analytics system explicitly tailored to survival data 432
analysis tasks. We aimed to provide comprehensive coverage of key visual analytics 433
concepts (see Sec. 2.2) and also to test the hypothesis that a small, well-defined set of 434
reusable visual analytics components could enable domain experts to effectively address 435
a broad range of survival analysis tasks. 436

Focus of our approach is the flexibility afforded by visual programming, enabling 437
users to construct and adjust data analysis pipelines dynamically. This adaptability 438
supports a nuanced data exploration through comparative analysis and variation of 439
inputs. Changes in the outcomes of the computational analysis may be observed 440
interactively by changing any method's settings, selecting a different data or feature 441
subset, or designing different analysis paths in the same workflow. Our system's design 442
inherently supports the seamless interplay between computational processes and human 443
input and demonstrates how users can be actively engaged in various analytical tasks. 444
In particular, this includes the selection of features and cohorts. Our framework 445
supports various means of human intervention, ranging from computationally driven 446
decisions, such as feature ranking, to domain knowledge-driven decisions, such as 447
selecting genes of interest. 448

Support for visual analytics in our proposed framework is primarily a consequence of 449
developing survival analysis components within Orange data mining framework. From 450
Orange, we benefit from interpreted workflows, interactive visualizations, and the ability 451
to encapsulate methods of choice within widgets that act on their own, but through a 452
responsive exchange of data collectively, constrained within a workflow, form an 453
exploratory data analysis environment. This design ensures that the analytical 454
environment—characterized by its exploratory nature and the synergy between different 455
components—is entirely shaped by the user's interactions and decisions. 456

In summary, our work contributes to the field of survival analysis by providing a 457
novel visual analytics framework that enhances user engagement through intuitive 458
design and visual programming. 459

7 Conclusion 460

Survival analysis is concerned with data where the outcome variable of interest is time 461
to an event. Its results can help us understand the effects of factors that determine a 462
patient's lifespan or design treatments for disease. With advances in technology to 463
profile patients, phenotypes, drugs, and treatments and to collect the data related to 464
survival, the democratization of survival data analysis and visual communication of 465
results becomes essential. 466

In this paper, we have described a tool and an approach for visual analysis of survival data. Our main contributions are the definition and implementation of analytic components for survival analysis workflows and the demonstration that these are sufficient and convenient to solve standard survival analysis tasks. We have also shown that these workflows can bring humans into the loop and thoroughly implement key visual analytics concepts, involving users in both experimentation and interpretation of results.

Data and Software Availability

The toolbox we developed is available as an open source add-on "Survival" to Orange (`http://orangedatamining.com`). Interested readers are welcome to review or help improve the code at `https://github.com/biolab/orange3-survival-analysis`. The workflows from the case studies are available at `https://orangedatamining.com/examples` under the section "Survival Analysis".

Acknowledgements

References

1. Fleming TR, Lin D. Survival analysis in clinical trials: past developments and future directions. Biometrics. 2000;56(4):971–983.

2. Pazdur R. Endpoints for assessing drug activity in clinical trials. Oncologist. 2008;13(2):19.

3. Van't Veer LJ, Bernards R. Enabling personalized cancer medicine through analysis of gene-expression patterns. Nature. 2008;452(7187):564–570.

4. Chin L, Andersen JN, Futreal PA. Cancer genomics: from discovery science to personalized medicine. Nature Medicine. 2011;17(3):297–303.

5. Terry M Therneau, Patricia M Grambsch. Modeling Survival Data: Extending the Cox Model. New York: Springer; 2000.

6. Pölsterl S. scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. Journal of Machine Learning Research. 2020;21(212):1–6.

7. Davidson-Pilon C. lifelines: survival analysis in Python. Journal of Open Source Software. 2019;4(40):1317.

8. Andrienko N, Andrienko G, Fuchs G, Slingsby A, Turkay C, Wrobel S. Visual Analytics for Data Scientists. Springer International Publishing; 2020.

9. Ward MO, Grinstein G, Keim D. Interactive Data Visualization: Foundations, Techniques, and Applications. 2nd ed. CRC Press; 2015.

10. Curk T, Demsar J, Xu Q, Leban G, Petrovic U, Bratko I, et al. Microarray data mining with visual programming. Bioinformatics. 2005;21(3):396–398.

11. Demšar J, Curk T, Erjavec A, Črt Gorup, Hočevar T, Milutinovič M, et al. Orange: Data Mining Toolbox in Python. Journal of Machine Learning Research. 2013;14:2349–2353.

12. Demsar J, Zupan B. Orange: Data mining fruitful and fun—A historical perspective. Informatica. 2013;37(1):55.

13. Group BDW, Atkinson Jr AJ, Colburn WA, DeGruttola VG, DeMets DL, Downing GJ, et al. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clinical Pharmacology & Therapeutics. 2001;69(3):89–95.

14. Sotiriou C, Piccart MJ. Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? Nature Reviews Cancer. 2007;7(7):545–553.

15. Henry NL, Hayes DF. Cancer biomarkers. Molecular Oncology. 2012;6(2):140–146.

16. Collins FS, Varmus H. A new initiative on precision medicine. New England Journal of Medicine. 2015;372(9):793–795.

17. Endert A, Ribarsky W, Turkay C, Wong BLW, Nabney I, Blanco ID, et al. The State of the Art in Integrating Machine Learning into Visual Analytics. Computer Graphics Forum. 2017;36(8):458–486.

18. Andrienko G, Andrienko N, Rinzivillo S, Nanni M, Pedreschi D, Giannotti F. Interactive visual clustering of large collections of trajectories. In: 2009 IEEE Symposium on visual analytics science and technology. IEEE; 2009. p. 3–10.

19. Krause J, Dasgupta A, Fekete JD, Bertini E. Seekaview: An intelligent dimensionality reduction strategy for navigating high-dimensional data spaces. In: 2016 IEEE 6th Symposium on Large Data Analysis and Visualization (LDAV). IEEE; 2016. p. 11–19.

20. Sedlmair M, Heinzl C, Bruckner S, Piringer H, Möller T. Visual parameter space analysis: A conceptual framework. IEEE Transactions on Visualization and Computer Graphics. 2014;20(12):2161–2170.

21. Choo J, Lee H, Kihm J, Park H. iVisClassifier: An interactive visual analytics system for classification based on supervised dimension reduction. In: 2010 IEEE Symposium on Visual Analytics Science and Technology. IEEE; 2010. p. 27–34.

22. Chatzimparmpas A, Martins RM, Kerren A. t-visne: Interactive assessment and interpretation of t-SNE projections. IEEE Transactions on Visualization and Computer Graphics. 2020;26(8):2696–2714.

23. Bavoil L, Callahan SP, Crossno PJ, Freire J, Scheidegger CE, Silva CT, et al. Vistrails: Enabling interactive multiple-view visualizations. In: VIS 05. IEEE Visualization, 2005. IEEE; 2005. p. 135–142.

24. Yu B, Silva CT. VisFlow-Web-based visualization framework for tabular data with a subset flow model. IEEE Transactions on Visualization and Computer Graphics. 2016;23(1):251–260.

25. Lex A, Streit M, Schulz HJ, Partl C, Schmalstieg D, Park PJ, et al. StratomeX: Visual Analysis of Large-Scale Heterogeneous Genomics Data for Cancer Subtype Characterization. Computer Graphics Forum. 2012;31(3pt3):1175–1184.

26. Corvò A, Caballero HG, Westenberg MA. SurviVIS: visual analytics for interactive survival analysis. In: 10th International EuroVis Workshop on Visual Analytics; 2019. p. 73–77.

27. Marai GE, Ma C, Burks AT, Pellolio F, Canahuate G, Vock DM, et al. Precision risk analysis of cancer therapy with interactive nomograms and survival plots. IEEE Transactions on Visualization and Computer Graphics. 2018;25(4):1732–1745.

28. Chen X, Sun X, Hoshida Y. Survival analysis tools in genomics research. Human Genomics. 2014;8:1–5.

29. Zheng H, Zhang G, Zhang L, Wang Q, Li H, Han Y, et al. Comprehensive review of web servers and bioinformatics tools for cancer prognosis analysis. Frontiers in Oncology. 2020;10:68.

30. Maller O, Drain AP, Barrett AS, Borgquist S, Ruffell B, Zakharevich I, et al. Tumour-associated macrophages drive stromal cell-dependent collagen crosslinking and stiffening to promote breast cancer aggression. Nature Materials. 2021;20(4):548–559.

31. Ingebriktsen LM, Finne K, Akslen LA, Wik E. A novel age-related gene expression signature associates with proliferation and disease progression in breast cancer. British Journal of Cancer. 2022;127(10):1865–1875.

32. Idogawa M, Koizumi M, Hirano T, Tange S, Nakase H, Tokino T. Dead or alive? Pitfall of survival analysis with TCGA datasets. Cancer Biology & Therapy. 2021;22(10-12):527–528.

33. Godec P, Pančur M, Ilenič N, Čopar A, Stražar M, Erjavec A, et al. Democratized image analytics by visual programming through integration of deep models and small-scale machine learning. Nature Communications. 2019;10(1):1–7.

34. Van Noorden R, Maher B, Nuzzo R. The top 100 papers. Nature News. 2014;514(7524):550.

35. Mallett S, Royston P, Waters R, Dutton S, Altman DG. Reporting performance of prognostic models in cancer: a review. BMC Medicine. 2010;8(1):1–11.

36. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Medical Research Methodology. 2014;14:1–11.

37. Phung MT, Tin Tin S, Elwood JM. Prognostic models for breast cancer: a systematic review. BMC Cancer. 2019;19(1):1–18.

38. Altman DG, McShane LM, Sauerbrei W, Taube SE. Reporting recommendations for tumor marker prognostic studies (REMARK): explanation and elaboration. BMC medicine. 2012;10(1):1–39.

39. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Annals of Internal Medicine. 2015;162(1):W1–W73.

40. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association. 1958;53(282):457–481.

41. Cupples LA, Gagnon DR, Ramaswamy R, D'Agostino RB. Age-adjusted survival curves with application in the Framingham study. Statistics in Medicine. 1995;14(16):1731–1744.

42. Harrington DP, Fleming TR. A class of rank test procedures for censored survival data. Biometrika. 1982;69(3):553–566.

43. Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological). 1972;34(2):187–202.

44. Balachandran VP, Gonen M, Smith JJ, DeMatteo RP. Nomograms in oncology: more than meets the eye. The Lancet Oncology. 2015;16(4):e173–e180.

45. Mehta HB, Mehta V, Girman CJ, Adhikari D, Johnson ML. Regression coefficient–based scoring system should be used to assign weights to the risk index. Journal of Clinical Epidemiology. 2016;79:22–28.

46. Harrell Jr FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine. 1996;15(4):361–387.

47. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statistics in Medicine. 2011;30(10):1105–1117.

48. Lawrence YR, Moughan J, Magliocco AM, Klimowicz AC, Regine WF, Mowat RB, et al. Expression of DNA repair gene MLH1 correlates with survival in resected pancreatic cancer receiving adjuvant chemoradiation: NRG Oncology RTOG 9704. Cancer. 2018;124(3):491.

49. Nodin B, Fridberg M, Jonsson L, Bergman J, Uhlén M, Jirström K. High MCM3 expression is an independent biomarker of poor prognosis and correlates with reduced RBM3 expression in a prospective cohort of malignant melanoma. Diagnostic Pathology. 2012;7:1–10.

50. Hwang KT, Kim BH, Oh S, Park SY, Jung J, Kim J, et al. Prognostic role of KRAS mRNA expression in breast cancer. Journal of Breast Cancer. 2019;22(4):548–561.

51. Chen L, Shi Y, Yuan J, Han Y, Qin R, Wu Q, et al. HIF-1 alpha overexpression correlates with poor overall survival and disease-free survival in gastric cancer patients post-gastrectomy. PloS one. 2014;9(3):e90678.

52. Goncalves A, Finetti P, Birnbaum D, Bertucci F. The CINSARC signature predicts the clinical outcome in patients with Luminal B breast cancer. NPJ Breast Cancer. 2021;7(1):48.

53. Long J, Zhang L, Wan X, Lin J, Bai Y, Xu W, et al. A four-gene-based prognostic model predicts overall survival in patients with hepatocellular carcinoma. Journal of Cellular and Molecular Medicine. 2018;22(12):5928–5938.

54. Lee H, Kwon MJ, Koo BM, Park HG, Han J, Shin YK. A novel immune prognostic index for stratification of high-risk patients with early breast cancer. Scientific Reports. 2021;11(1):128.

55. Yuan J, Duan F, Zhai W, Song C, Wang L, Xia W, et al. An aging-related gene signature-based model for risk stratification and prognosis prediction in breast cancer. International Journal of Women's Health. 2021; p. 1053–1064.

56. Zhang M, Wang X, Chen X, Zhang Q, Hong J. Novel immune-related gene signature for risk stratification and prognosis of survival in lower-grade glioma. Frontiers in Genetics. 2020;11:363.

57. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486(7403):346–352.

58. Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, et al. The cancer genome atlas pan-cancer analysis project. Nature genetics. 2013;45(10):1113–1120.

59. Downward J. Targeting RAS signalling pathways in cancer therapy. Nature Reviews Cancer. 2003;3(1):11–22.

60. Quentmeier H, Reinhardt J, Zaborski M, Drexler H. FLT3 mutations in acute myeloid leukemia cell lines. Leukemia. 2003;17(1):120–124.

61. Chen R, Wang X, Fu J, Liang M, Xia T. High FLT3 expression indicates favorable prognosis and correlates with clinicopathological parameters and immune infiltration in breast cancer. Frontiers in Genetics. 2022;13:956869.

62. Gao S, Ge A, Xu S, You Z, Ning S, Zhao Y, et al. PSAT1 is regulated by ATF4 and enhances cell proliferation via the GSK3β/β-catenin/cyclin D1 signaling pathway in ER-negative breast cancer. Journal of Experimental & Clinical Cancer Research : CR. 2017;36.

63. Feng M, Cui H, Tu W, Li L, Gao Y, Chen L, et al. An integrated pan-cancer analysis of PSAT1: A potential biomarker for survival and immunotherapy. Frontiers in Genetics. 2022;13.
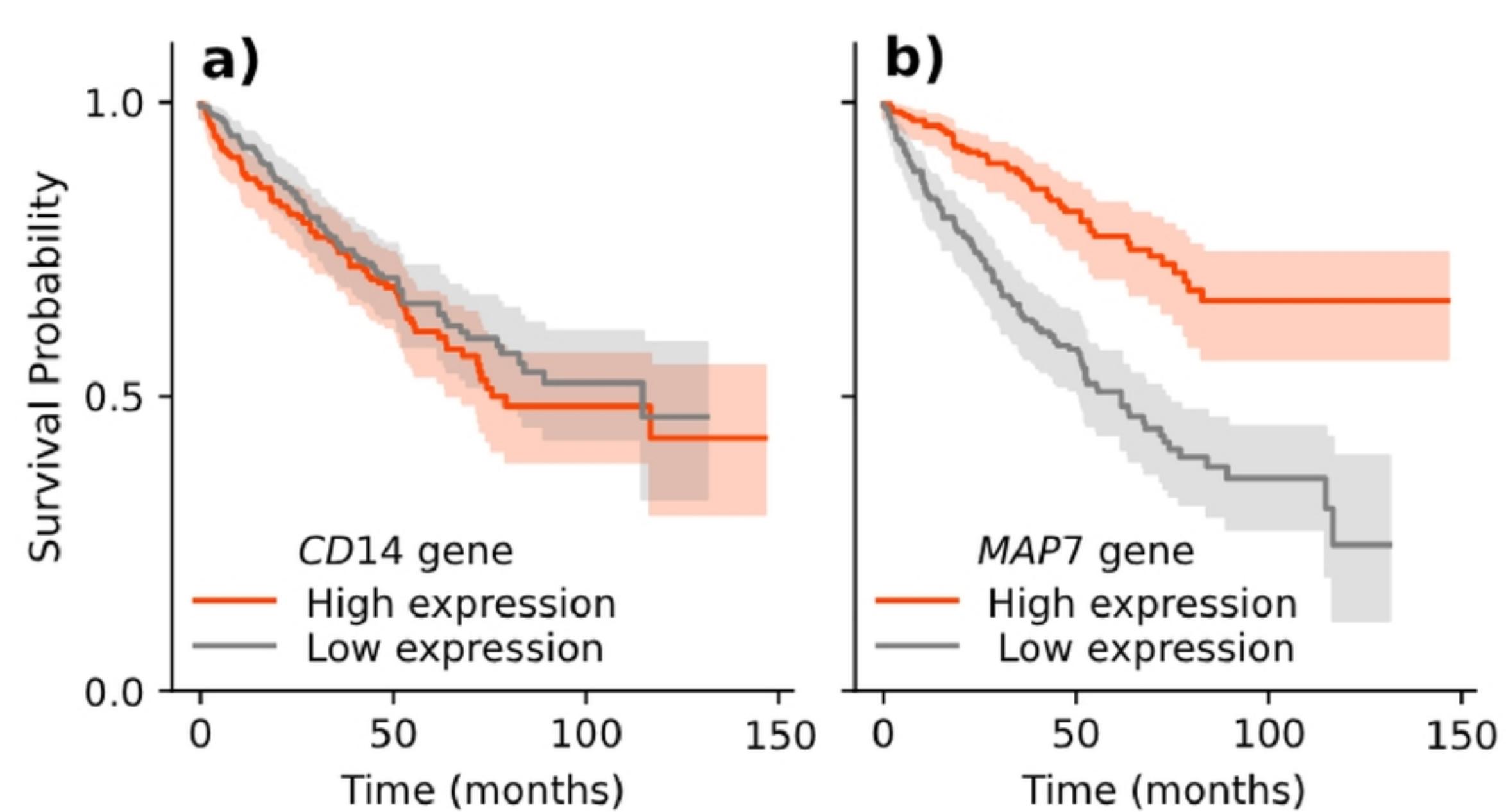
Figure 1
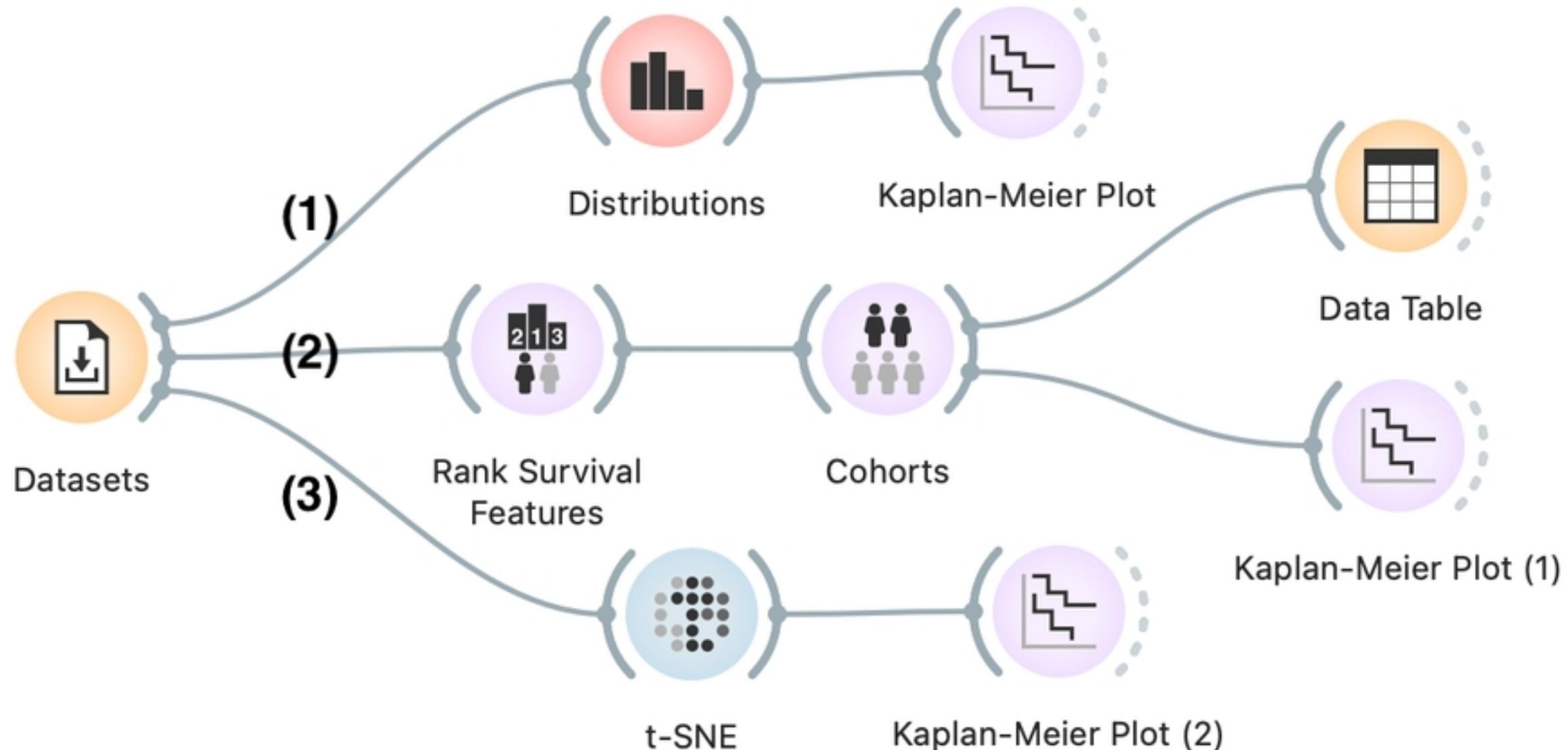
Figure 2
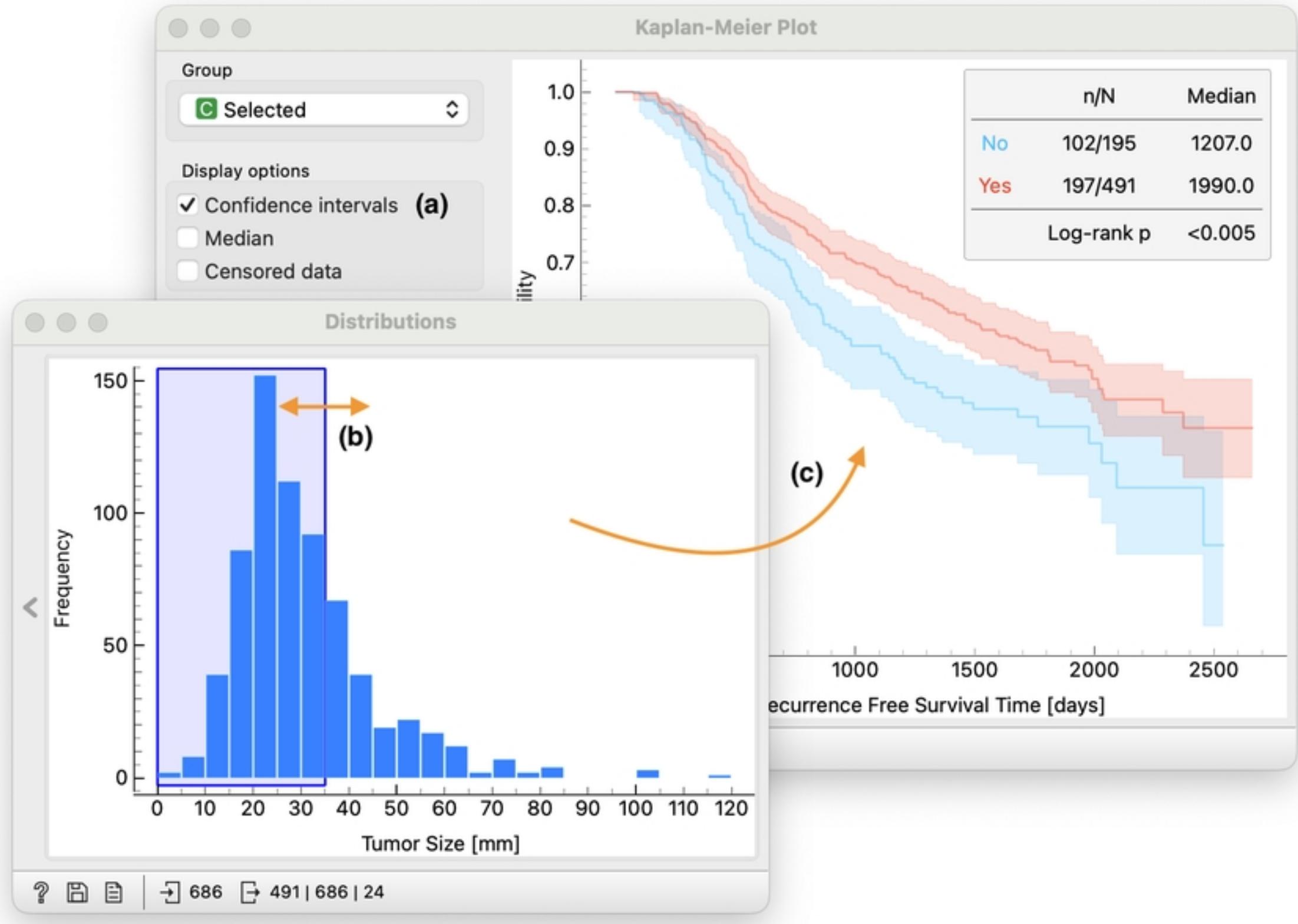
Figure 3

Rank Survival Features

		Log-Likelihood Ratio test	p ^	FDR
1	N Number of Positive Nodes	50.0	1.51e-12	1.21e-11
2	N Progesterone Receptor [fmol]	34.0	5.38e-09	2.15e-08
3	C Tumor Grade	20.0	7.80e-06	2.08e-05
4	N Tumor Size			
5	C Hormonal Th			

(a)

Cohorts

Stratify on

Cox Risk Score
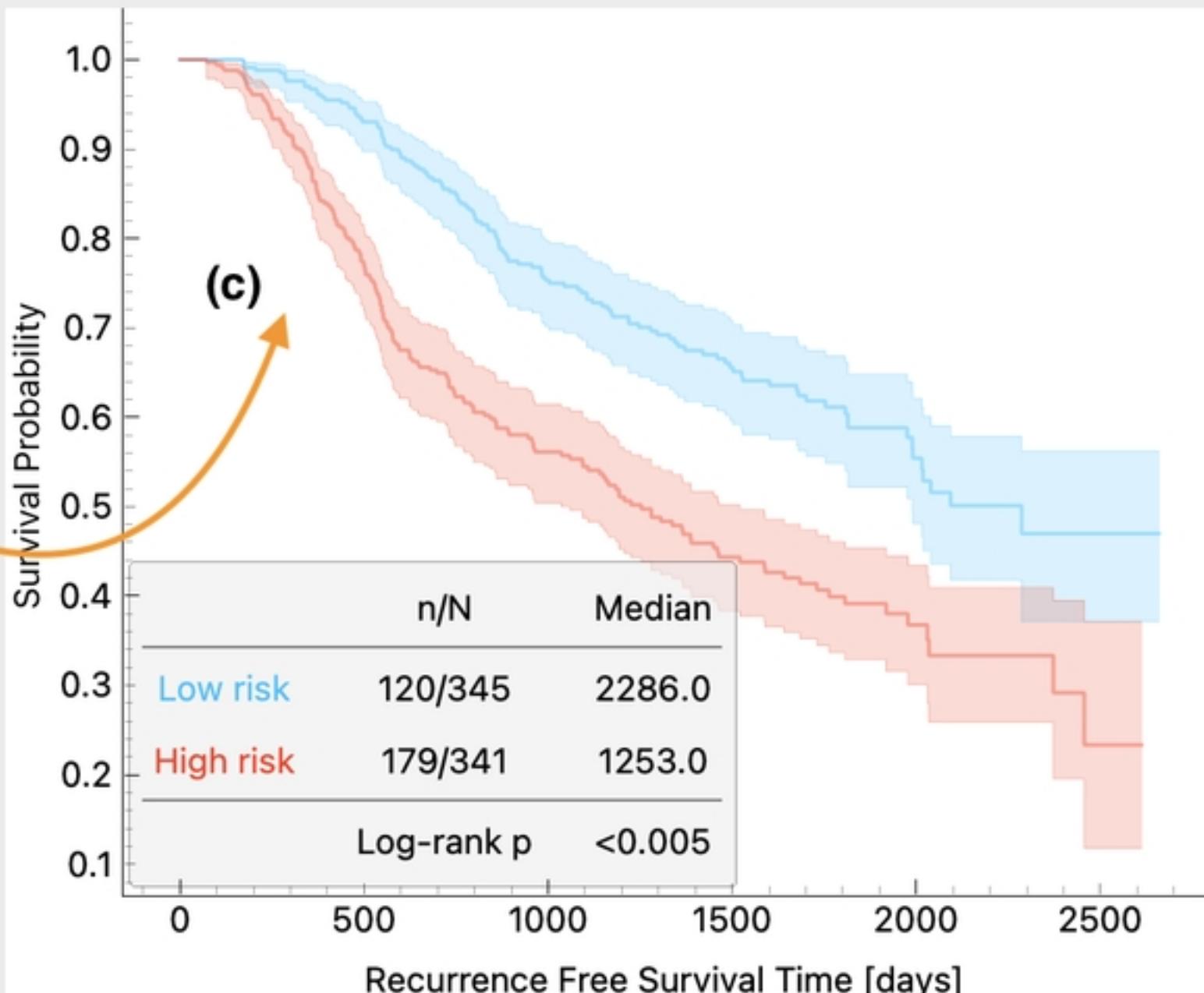
Splitting Criteria

● Median
○ Mean
○ Log Rank Test

(b)

☑ Commit Automatical

? | → 686 | - → 686 | 2

Kaplan-Meier Plot
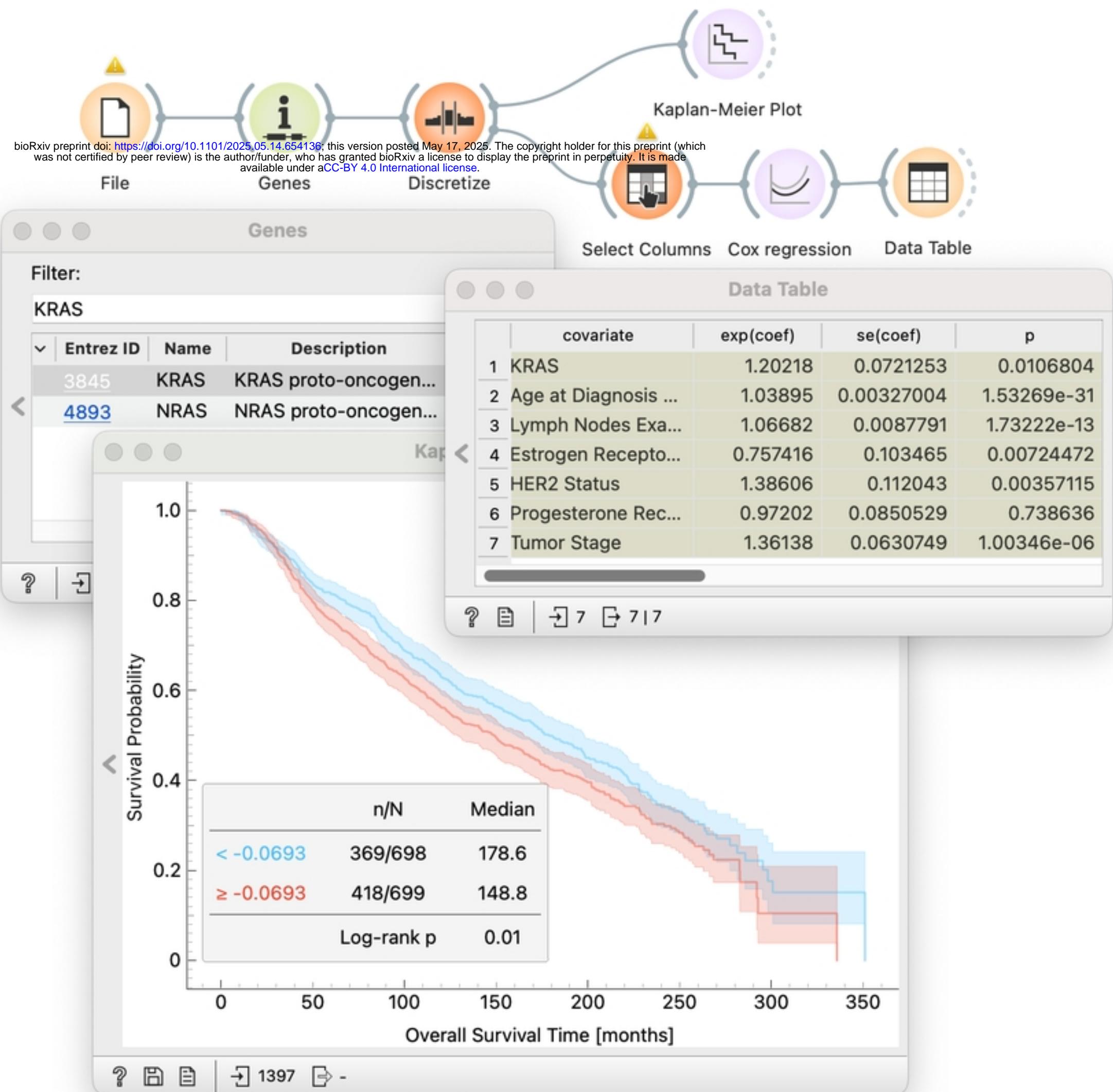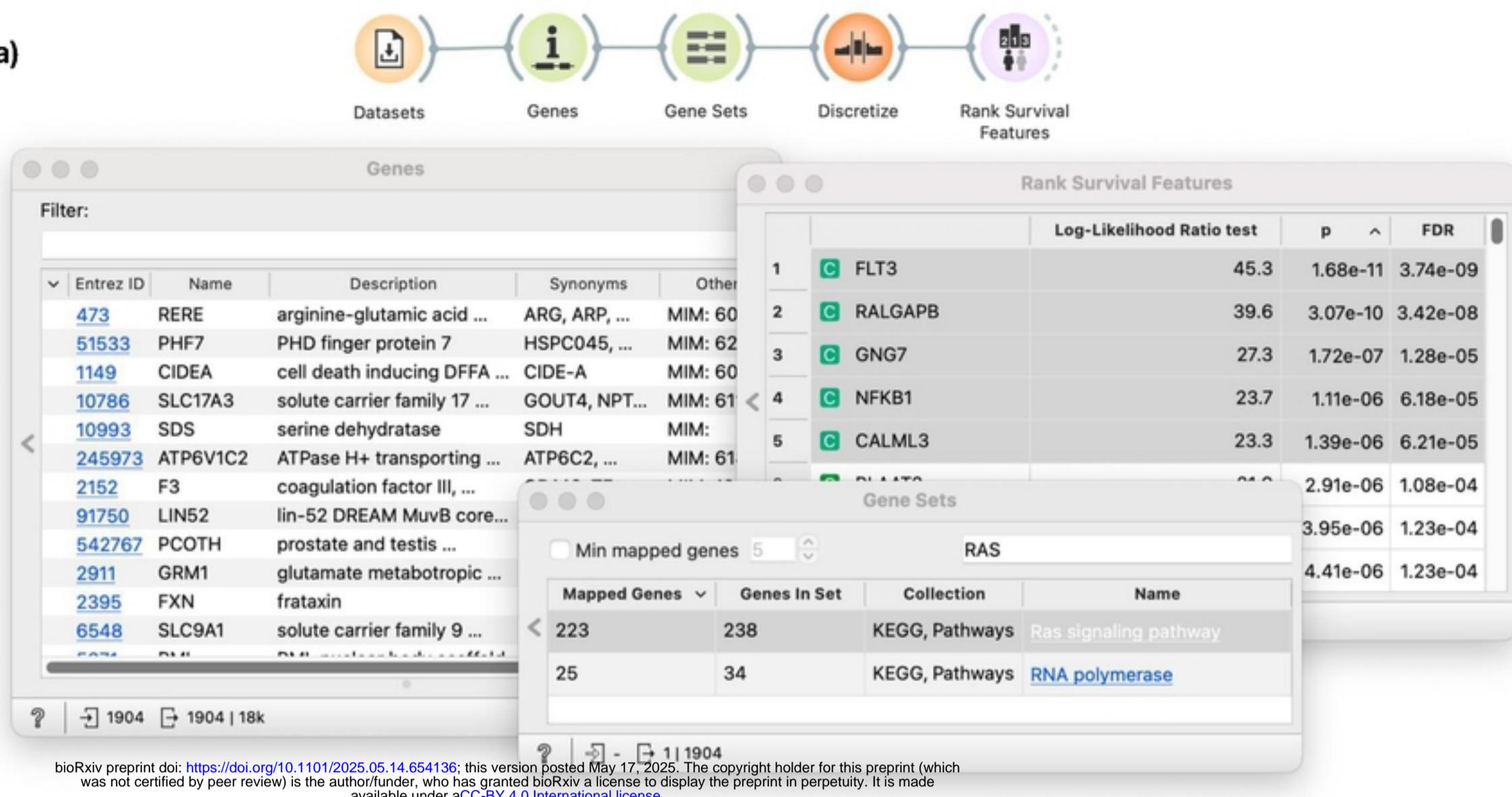
(c)

|	n/N	Median
Low risk	120/345	2286.0
High risk	179/341	1253.0
Log-rank p		<0.005

? 🖫 🖹 | → 686 ↦ -
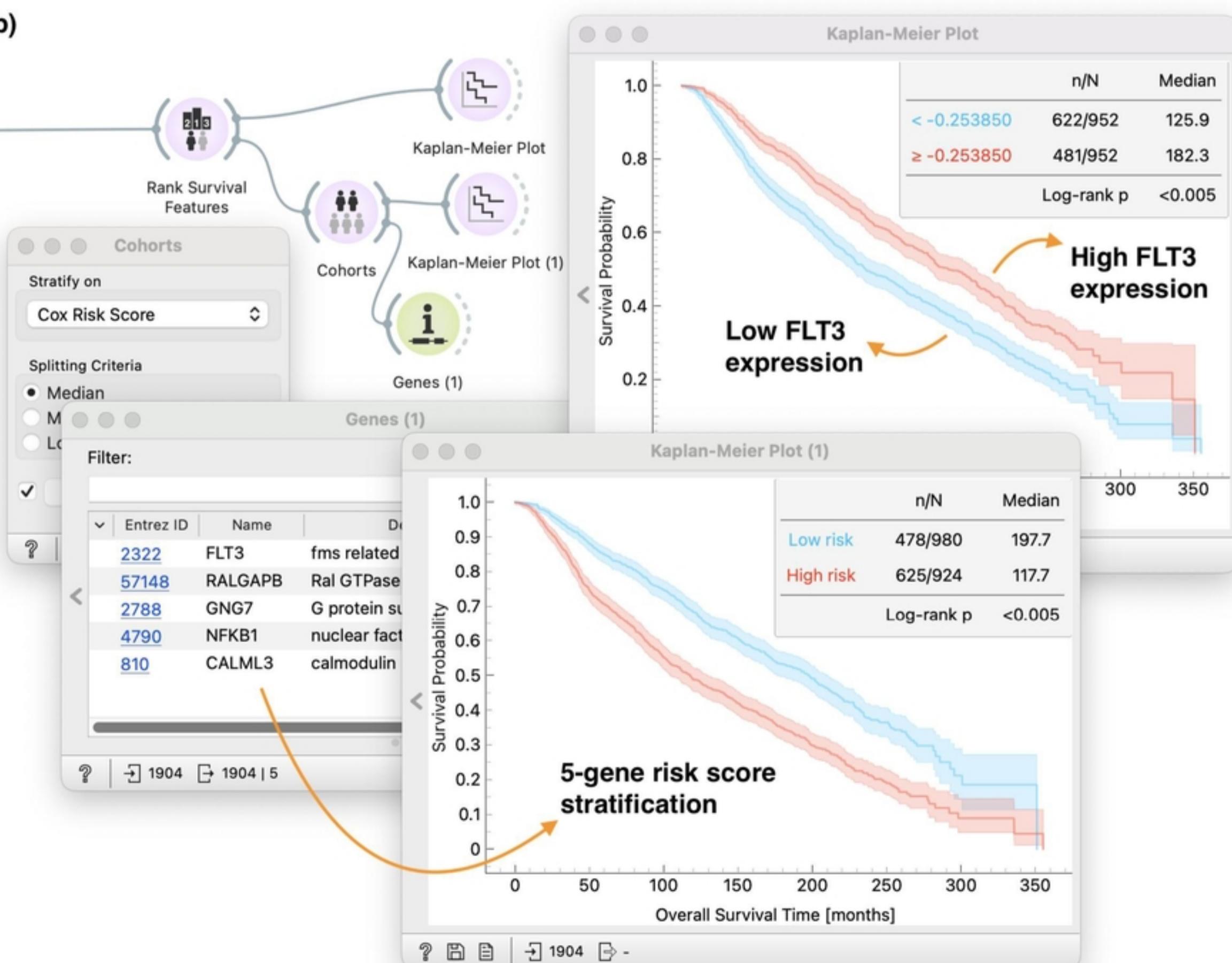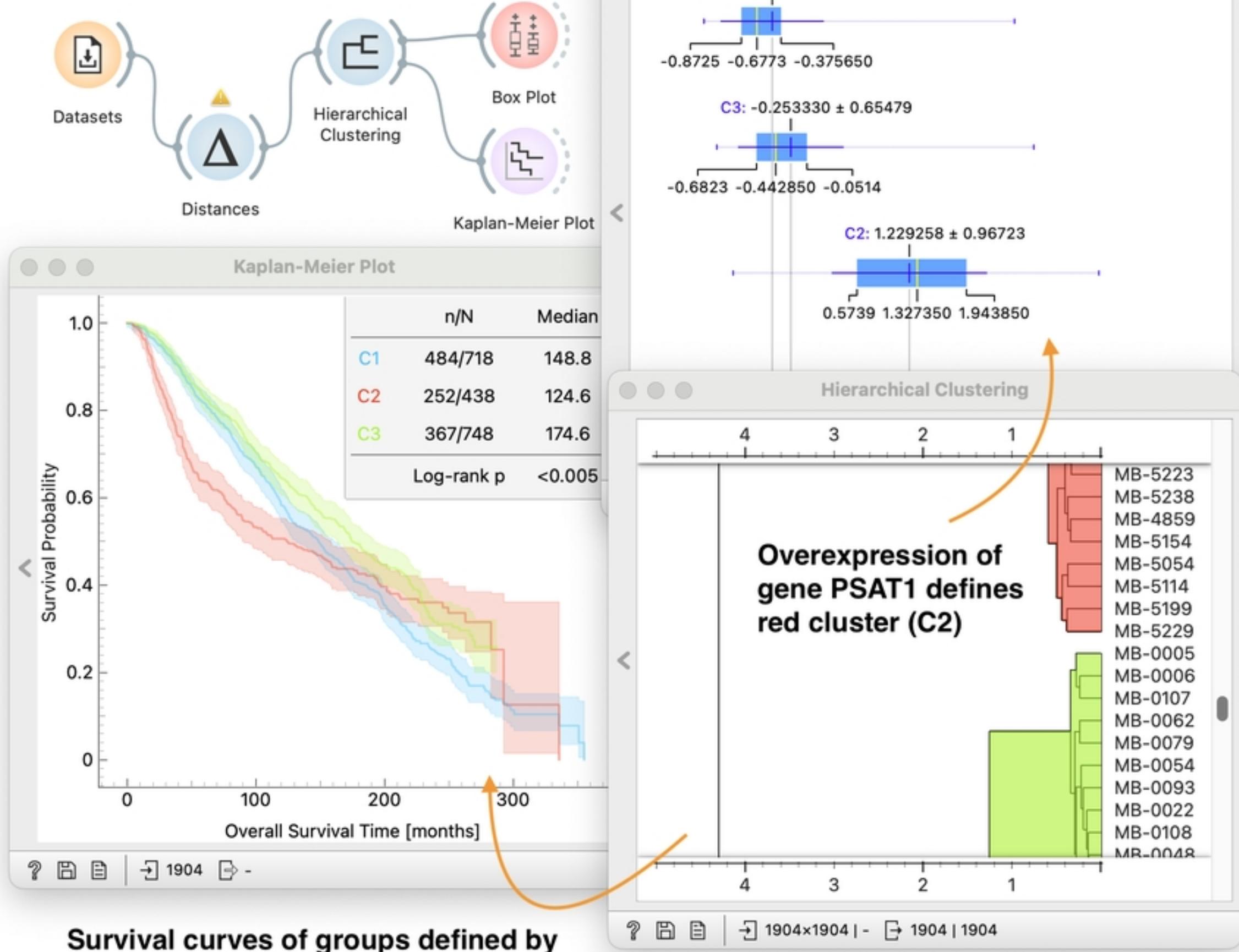
Figure 4

Figure 6

Figure 7

Figure 8

	Survival Time	Event	Age	Tumor Stage	HER2 Status	KRAS	RERE	PHF7	CIDEA	SLC17A3
1	140.5	0	75.65	2	Negative	-2.3162	1.3762	-0.9217	3.8334	0.3946
2	84.6333	0	43.19	1	Negative	-0.6302	-0.0226	-1.0580	-0.1394	1.3907
3	163.7	1	48.87	2	Negative	0.5484	-2.2425	-0.46...	0.1429	-0.1929
4	164.933	0	47.68	2	Negative	0.5877	-1.7706	-0.4199	-1.0699	-0.3918
5	41.3667	1	76.97	2	Negative	-0.6389	-2.0498	-0.0172	2.0533	0.3816
6	7.8	1	78.77	4	Negative	0.3136	-1.3204	0.3796	-0.7636	-0.3045
7	164.333	0	56.45	2	Negative	0.1411	-0.7563	-0.07...	1.8458	-0.8656
8	99.5333	1	89.08	2	Negative	0.4336	-1.2133	-0.45...	2.5638	-0.1929
9	36.5667	1	86.41	2	Negative	-0.1940	0.3691	-0.84...	-0.7042	0.3554
10	36.2667	1	84.22	2	Negative	0.2829	1.1416	1.5271	-1.0327	1.0685
11	132.033	1	85.49	4	Negative	-0.0299	0.7249	2.0576	0.6257	-0.3075
12	163.533	0	70.91	1	Negative	-0.6801	-0.8197	1.4398	-1.0413	1.3313

1904   1904 | 1904

Figure 5