**Instructions to “oncoClassSurv” software**

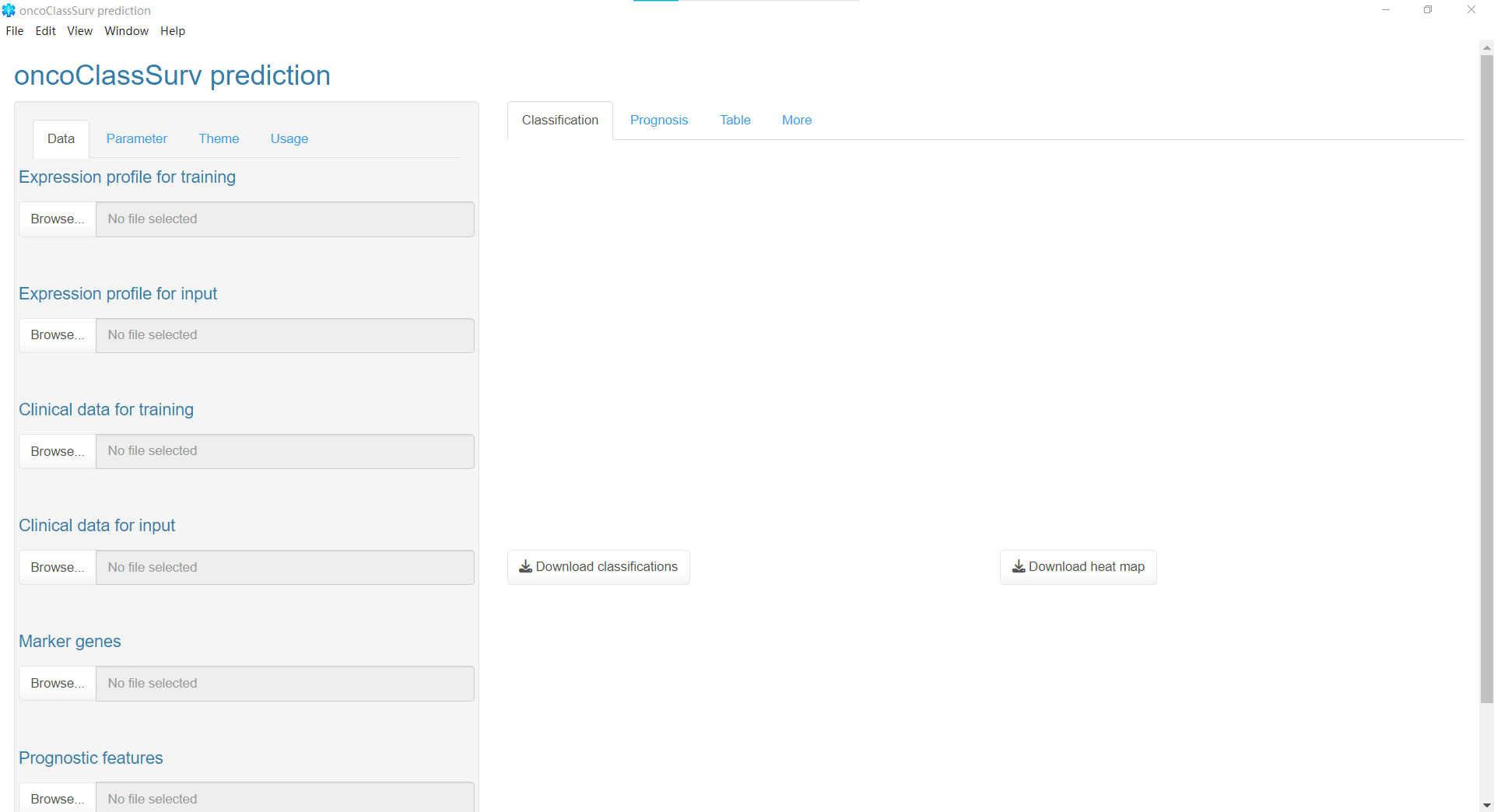
1. **Software functions and data preparation**

The functions and data preparation requirements of the software are detailed in: <https://github.com/OliveryYL/oncoClassSurv>

Data for demonstration: <https://github.com/OliveryYL/oncoClassSurv_Expansion>

1. **Installation and startup of desktop software**

Download the software package “oncoClassSurvApp start win32 x64” directly to any path on a personal computer of Windows operating system, extract the package, and double-click the “oncoClassSurvApp. exe” file with the mouse to start. Of course, users can also send a shortcut to desktop to directly launch programs from the desktop. When starting the software, users may encounter warnings from certain security management software. When it happens, please take it easy and set to trust to continue starting the program. In addition, during the initial startup, there may be a situation where the interface is blank. Users can click “Reload” in the “View” tab of the menu bar, and after the program is reset, it can be used normally; Alternatively, by restarting the software, the interface can usually be loaded normally. The interface of the software after normal loading is as follows:



1. **Panel partitioning**
   1. Menu bar

The menu bar is located in the upper left corner of the software interface. Currently, it includes five tabs: “File, Edit, View, Window, and Help”.

* 1. Left column

The left column provides input of data and parameters, selection of theme style, and easy user manual and help links.

Specifically, it can be divided into four panels: “Data”, “Parameter”, “Theme”, and “Usage”. “Data” provides user-defined data input; The “Parameter” provides personalized analysis parameters for users to conduct data analysis; “Theme” provides multiple personalized theme style options; The “Usage” includes a simple user manual and help links.

* 1. Right main panel area

The right column is the output interface for the analysis results. It consists of four parts: “Classification”, “Prognosis”, “Table”, and “More”.

Among them, “Classification” is the visualization result of predicting molecular typing, including statistical distribution bar plots of different molecular typing; “Prognosis” plots survival curves for individual prognostic risks at different time points; The analysis results of “Classification” and “Prognosis” in “Table” are output as tables respectively; “More” displays the differences between training data and prediction data, as well as principal component analysis plots before and after removing batch effects.

All output data (images and tables) can be saved locally on the computer by clicking the corresponding download button.

1. **Data Input and Analysis Settings**
   1. The “Data” in the left column contains a total of 6 data upload widgets, each with default data (see Table 1 for definitions of default data). Therefore, you can click the “Run” button at the bottom to directly run without inputting customized data. If users need to predict their own data, they need to click on the corresponding widgets to upload their customized data.

Table 1 Definition of Data Upload Widgets in the “Data” Interface

|  |  |  |
| --- | --- | --- |
| **Panel parameter** | **Parameter in the “oncoClassSurv” package** | **Parameter concepts** |
| Expression profile for training | train.exp.path | Expression of the training cohort. The default dataset is the expression profile of malignant samples from TCGA-LIHC. |
| Expression profile for input | input.exp.path | Expression profile of the user”s cohort |
| Clinical data for training | train.clin.path | Known classifications (if the task covers predicting classifications) and or survival data (if the task covers predicting prognosis) in the training cohort |
| Clinical data for input | input.clin.path | If the “task” equals 2 or 3 (predicting prognosis), and clinical covariates (age, stage, etc.) are concerned, please upload the corresponding files. |
| Marker genes | train\_cluster.feature.path | Known classification-specific genes |
| Prognostic features | train\_survival.feature.path | Covariates (genes or clinical factors) involved in the prognostic model |

* 1. “Parameter” in the left column.

(1) Task. Firstly, the analysis task should be confirmed: when “Task” is 1, perform classification prediction. When “Task” is 2, perform survival risk prediction. When the “Task” is 3, both classification and survival risk are predicted.

(2) ClassMethod. For classification tasks, it is necessary to choose an appropriate machine learning algorithm. “ClassMethod” includes two algorithms: RF (random forest) and SVM (support vector machine).

(3) BatchEffect. Whether need to remove the batch effects between the expression matrix of the training cohort and the expression matrix input by the user.

(4) Kernel. A parameter in the support vector machine. Details can be seen in the e1071::svm.

(5) PlotBatchPCA. Whether need to output the plots of batch effects. After removing the batch effects, visualize and return the comparison results. Using the principal component analysis.

(6) PlotSurfCurve. Whether output curves of survival risk over time for individual patients.

(7) ExprFormat. The currently supported standardized formats for expression spectrum data include FPKM and TPM.

(8) TrainNorm. Whether need to normalize the customized training data (transcriptome expression profile data) using the normalizeBetweenArrays function in the limma package.

(9) InputNorm. Whether need to use the normalizeBetweenArrays function in the limma package to normalize the customized input data (transcriptome expression profile data) to be predicted.

(10) RandomSamTrain. Whether to use random sampling method to select the expression spectrum data of some training set samples for model training.

(11) TrainSeed. If the “RandomSamTrain” option is “Yes”, users can further customize the random number seed of random sampling to fix the results of random sampling, with the value range of the random number seed being an integer from 1 to 9999.

(12) SamplingProb. If the “RandomSamTrain” option is “Yes”, users can customize the sampling proportions for model training. The default ratio is 0.7, which means randomly extracting 70% of the sample data for model training.

(13) Ntree. A parameter in the random forest Details can be seen in the randomForest:: randomForest

(14) Nodesize. A parameter in the random forest Details can be seen in the randomForest:: randomForest

(15) Mtry. A parameter in the random forest Details can be seen in the randomForest:: randomForest

(16) Cost. A parameter in the support vector machine Details can be seen in the e1071:: SVM

(17) Predict parts of samples. An input widget. If users want to analyze only a portion of the samples (Expression profile for input), they can save the names of the samples of interest as a table file and upload them through this widget.

(18) Upload Geneset. Input widget. If users desire to display only a specific set of genes when drawing gene expression heat maps, they can save the names of the genes of interest as a table file and upload them through this widget. Kindly note that this feature requires the “Heat Map Geneset” option to be specified as “Upload” at the same time.

(19) Default Geneset. If the user does not specifically specify the gene set displayed on the heatmap, the software will display the default gene set, including two types of default gene sets to choose from: immune-related gene sets (Immune) and four-FS subtype specific gene sets for hepatocellular carcinoma.

(20) Explore interesting genes. Text input box. If users want to explore the expression of certain genes of interest, they can enter the corresponding gene name in this input box (the entered gene name must be included in the predicted gene expression profile data). Kindly note that this feature requires the “Heat Map Geneset” option to be specified as “TextInput” at the same time.

(21) Heat Map Geneset. Select the corresponding gene set for display in the heat map and exploratory research.

(22) Heat Map Options. If the user desires the samples displayed in the heat map to be consistent with the “SurvSample” widget of the survival curves, they can check “Selected Samples”. If the user wishes to display the sample name in the heat map, they can check “Show Sample Name”.

(23) SurvSample. When the initial calculation is completed, the software will automatically load the sample names uploaded by the user for analysis and display them in the “SurvSample” selector. The user can check one or more samples to draw the survival curves of the corresponding samples.

(24) SurvBreak. Survival curve abscissa (time) separation. The default is 12.

(25) SurvUnit. The unit of survival curve abscissa (time). The default is” months”, and other characters can be entered to represent the corresponding time unit.

(26) LegendSize. The size of the legend for the survival curve. If you want to enlarge the legend, you can enter a larger numerical value.

* 1. “Theme” settings in the left column. Clicking on the “Show/Hide Theme Selector” button opens the “Theme” tab, and clicking the button again hides the tab.
  2. “Usage” in the left column. Since this software is based on the R package called “oncoClasssurv” on GitHub (<https://github.com/OliveryYL/oncoClassSurv>), so you can visit the GitHub repository for more information.

Demonstration video of operating the software: <https://v.youku.com/v_show/id_XNjA3MjgzNDM2OA==.html>

**“oncoClassSurv”软件使用说明**

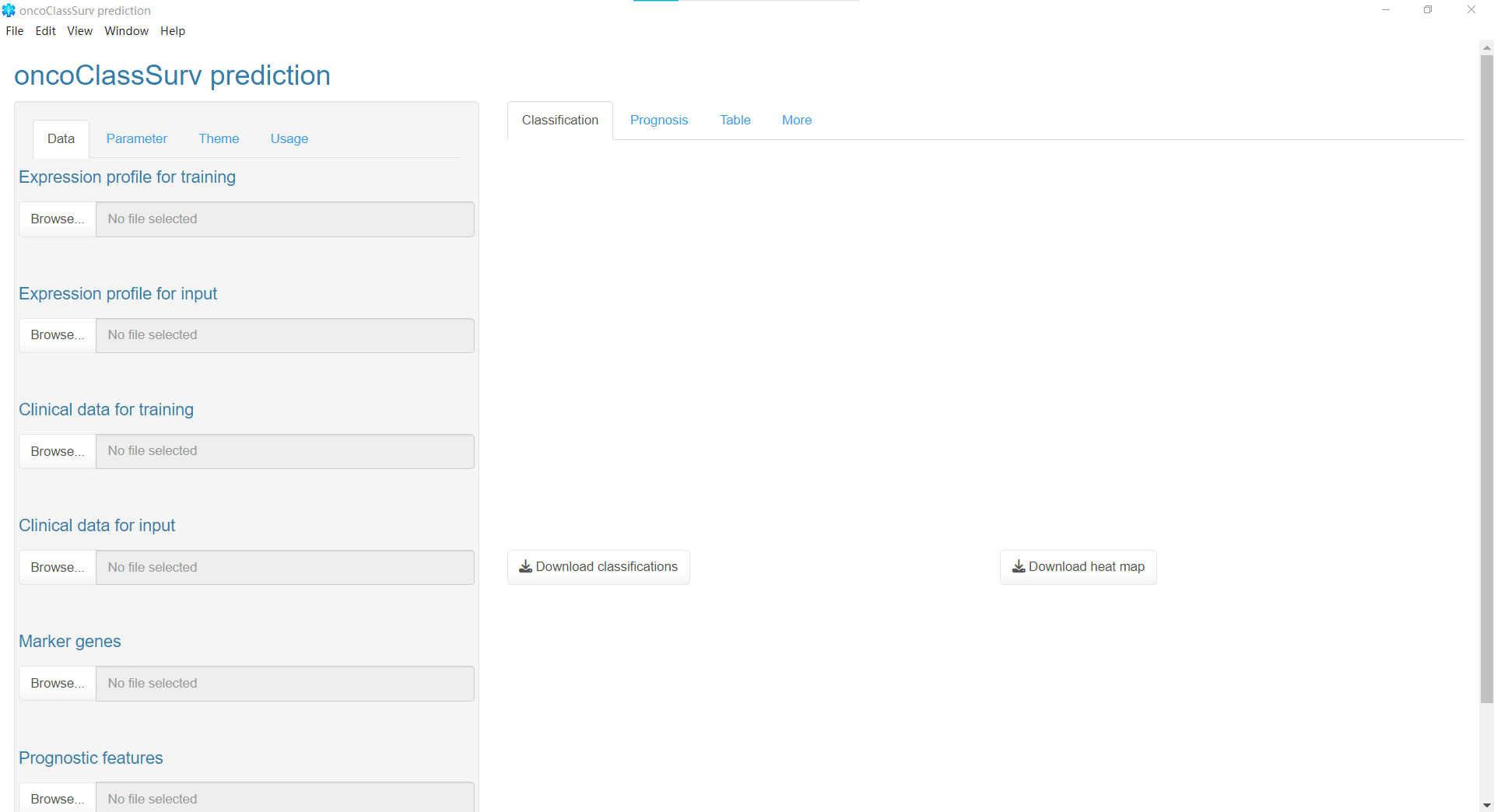
1. **软件的功能与数据准备**

软件的功能与数据准备要求详见：<https://github.com/OliveryYL/oncoClassSurv>

示例数据下载：<https://github.com/OliveryYL/oncoClassSurv_Expansion>

1. **桌面版软件的安装与启动**

直接下载软件包“oncoClassSurvApp-start-win32-x64”到Windows电脑的任何路径，解压软件包后，鼠标双击“oncoClassSurvApp.exe”文件即可启动。当然，用户也可以将发送到桌面快捷方式，以便从桌面直接启动程序。在启动软件时，用户可能会遇到某些安全管理软件的警告，设置为信任即可继续启动程序。此外，初次启动时，可能会出现用户界面为空白的情况，用户可以在菜单栏的“View”选项卡中，点击“Reload”，待程序重置后即可正常使用；或者，重启软件，通常可以正常加载界面。软件正常加载后的界面如下：



1. **面板分区**
   1. 菜单栏

菜单栏位于软件界面的左上角。目前包含了“File、Edit、View、Window、Help”五个选项卡。

* 1. 左侧栏

左侧栏提供了数据和参数的输入，主题风格选择，以及简易的使用手册与帮助链接。

具体可以分为“Data”，“Parameter”，“Theme”，和“Usage”四个面板。“Data”提供用户自定义数据输入；“Parameter”为用户进行数据分析提供了个性化的的分析参数；“Theme”提供了多种个性化的主题风格选项；“Usage” 包含了简易的使用手册与帮助链接。

* 1. 右侧主面板区

右侧栏为分析结果的输出界面。包括四个部分：“Classification”，“Prognosis”，“Table”，和“More”。

其中，“Classification”为预测分子分型的可视化结果，包括不同分子分型的统计分布条形图；“Prognosis”为个体在不同时间点的预后风险绘制生存曲线；“Table”为“Classification”和“Prognosis”的分析结果分别输出为表格；“More”展示了训练数据与预测数据之间的差异，并展示了去除批次效应前后的主成分分析图。

所有的输出数据（图片和表格）均可通过点击相应的下载按钮保存在计算机本地。

1. **数据输入与分析设置**
   1. 左侧栏的“Data”共包含6个数据输入控件，每个控件都有默认数据（默认数据详见表 1），因此可以在不输入自定义数据的情形下点击底部的“Run”按钮直接运行。倘若用户需要预测自己的数据，则需要点击相应的控件输入自定义数据。

表 1 “Data”界面数据输入控件的定义

|  |  |  |
| --- | --- | --- |
| **面板参数** | **“oncoClassSurv”包中的参数** | **参数概念** |
| Expression profile for training | train.exp.path | 训练队列的表达。默认数据集是TCGA-LIHC恶性样本的表达谱。 |
| Expression profile for input | input.exp.path | 用户待预测队列的基因表达谱文件 |
| Clinical data for training | train.clin.path | 训练队列中的已知分类（如果任务涵盖预测分类）和/或生存数据（如果任务涉及预测预后） |
| Clinical data for input | input.clin.path | 如果“任务”等于2或3（涵盖预后分析），并且涉及临床协变量（年龄、分期等），请上传相应的文件。 |
| Marker genes | train\_cluster.feature.path | 预先鉴定的分子分型特异性基因 |
| Prognostic features | train\_survival.feature.path | 预后模型中涉及的协变量（基因或临床因素） |

* 1. 左侧栏的“Parameter”。

（1）Task。首先，应确认分析的任务：当“Task”为1时，执行分类预测。当“Task”为2时，执行生存风险预测。当“Task”为3时，同时对分类和生存风险进行预测。

（2）ClassMethod。对于分类任务，需要选择合适的机器学习算法，“ClassMethod”包括两种算法：RF（random forest，随机森林）和SVM（support vector machine）。

（3）BatchEffect。是否需要去除训练队列的表达矩阵与用户输入的表达矩阵之间的批量效应。

（4）Kernel。支持向量机中的一个参数。详细信息可以在e1071::svm中看到。

（5）PlotBatchPCA。是否需要输出批次效应图。可视化输出去除批次效应前后的对比图。使用主成分分析。

（6）PlotSurvCurve。是否输出患者随时间变化的生存风险曲线。

（7）ExprFormat。当前支持的表达谱数据标准化格式包括FPKM和TPM。

（8）TrainNorm。是否需要对自定义的训练数据（转录组表达谱数据）使用limma包中的normalizeBetweenArrays函数进行normalization。

（9）InputNorm。是否需要对待预测的自定义输入数据（转录组表达谱数据）使用limma包中的normalizeBetweenArrays函数进行normalization。

（10）RandomSamTrain。是否需要使用随机抽样法选择部分训练集样本的表达谱数据用于模型的训练。

（11）TrainSeed。如果“RandomSamTrain”选项为“Yes”，用户可以进一步自定义随机抽样的随机数种子，以固定随机抽样的结果，随机数种子的取值范围为从1到9999的整数。

（12）SamplingProb。如果“RandomSamTrain”选项为“Yes”，用户可以自定义随机抽取特定的样本比例用于模型的训练。默认的比例为0.7，即随机抽取70%的样本数据用于模型的训练。

（13）ntree。随机森林算法中的一个参数。详细信息可以查看randomForest::randomForest。

（14）nodesize。随机森林算法中的一个参数。详细信息可以查看randomForest::randomForest。

（15）mtry。随机森林算法中的一个参数。详细信息可以查看randomForest::randomForest。

（16）cost。支持向量机算法中的一个参数。详细信息可以查看e1071::svm。

（17）Predict parts of samples。输入控件。如果用户希望仅分析部分样本（Expression profile for input），可以将感兴趣的样本名称保存为表格文件，并通过该控件进行upload。

（18）Upload Geneset。输入控件。如果用户希望在绘制基因表达热图时仅展示特定基因集的情况，可以将感兴趣的基因名称保存为表格文件，并通过该控件进行upload。注意，该功能需要同时将“Heat Map Geneset”选项指定为“Upload”。

（19）Default Geneset。如果用户不特别指定热图展示的基因集，软件将展示默认的基因集，包括两类默认基因集可供选择：免疫相关基因集（Immune）与肝细胞癌4个FS分型特异性基因集。

（20）Explore interesting genes。文本输入框。如果用户希望探索某些感兴趣的基因的表达情况，可以在该输入框中输入相应的基因名称（所输入的基因名称必须包含于待预测的基因表达谱数据中）。注意，该功能需要同时将“Heat Map Geneset”选项指定为“TextInput”。

（21）Heat Map Geneset。选择相应的基因集，用于热图展示和探索研究。

（22）Heat Map Options。 如果用户希望在热图中展示的样本与生存曲线中的“SurvSample”保持一致，可以勾选“Selected Samples”。如果用户希望在热图中显示样本名称，可以勾选 “Show Sample Name”。

（23）SurvSample。当初次运算结束后，软件会自动加载用户输入的待分析的样本名称，并显示在“SurvSample”选择器中，用户可以勾选一到多个样本，以绘制相应样本的生存曲线。

（24）SurvBreak。生存曲线横坐标（时间）分隔。默认为12。

（25）SurvUnit。生存曲线横坐标（时间）的单位。默认为“months”，可以输入其他字符以表示相应的时间单位。

（26）LegendSize。生存曲线的图例的尺寸。如果希望放大图例，可以输入较大的数值。

* 1. 左侧栏的“Theme”设置。点击“Show/Hide Theme Selector”按钮可以打开“Theme”选项卡，再次点击该按钮可以隐藏选项卡。
  2. 左侧栏的的“Usage”。由于此软件基于GitHub上名为oncoClasssurv的R包，因此您可以访问GitHub存储库（<https://github.com/OliveryYL/oncoClassSurv>）以获取更多更新信息。

软件操作示例视频：<https://v.youku.com/v_show/id_XNjA3MjgzNDM2OA==.html>