Running the OmixLitMiner tool for literature revival and datamining in the R environment

Omix Literature retrieval and datamining using the OmixLitMiner tool is exemplary visualized on the example of 55 putative protein markers, associated with lymph node metastasis in colorectal cancer. Proteins were selected by Mori et al as significantly differential abundant between CRC with LN metastasis versus CRC without LN metastasis versus normal colonic mucosa and CRC adjacent normal mucosa, from 4000 Proteins, quantified by iTRAQ based LC-MS/MS.[1] .
OmixLitMiner is a tool that can be integrated in the R software environment by any user, without a deep understanding of the underlying programming language. R is an opensource system for statistical computing and graphics. It provides a base system, that can be extended with additional functions, using online available open source packages.
For detailed information about R in general refer to the R FAQ, provided by the Comprehensive R Archive Network project. (CRAN) [2]

Downloading and installing R

For the usage of R for users, not familiar with programming, it is recommended to download the R software environment and additionally obtain the integrated Development Environment software RStudio that provides a clear user interface. RStudio requires the previous installation of the R environment.
To download R, refer to the R-project website. (https://www.r-project.org). The R software environment can be obtained from the CRAN network, which is linked on the website of the R project. (See Figure 1)

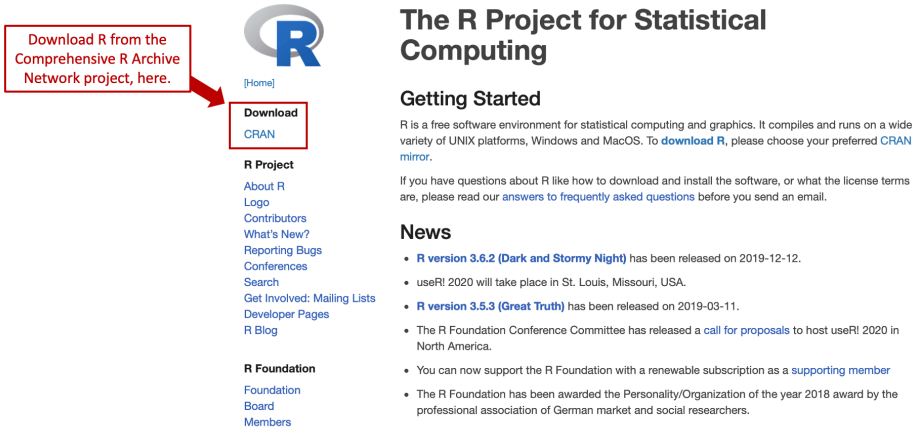


Figure 1: Web design of the webpage for the R Project for Statistical Computing (https://www.r-project.org) and directory for downloading the R software environment from the Comprehensive R Archive Network project. (CRAN)(red).

CRAN is a network of servers and file transfer protocols that provides detailed, up-to-date documentation, source code and binaries for the R environment. [3] After being redirected to the website of the CRAN project, you will get access to different mirror sites. Mirror sites store identical information (websites or files) on distinct servers, at different localizations and are used to optimize the capacity and efficiency of high-volume sited by distributing the workload across different servers (load balancing). To minimize the network load, choose an R mirror provided by

an institution, that is localized close to you. To choose a mirror, click on the Mirrors URL given on the CRAN website. (Figure 2).

Figure 2: Example on available mirrors, mirror specific URLs and server providing institutions located in different countries, hosting downloads for the R software environment.

After being redirected to the mirror specific URL, as exemplary visualized for a mirror hosted by the University of Münster in Figure 3, download links for different operating systems (Windows, OS X (mac) and Linux) will be provided. Download links include precompiled binary distributions for basic R packages and the R base system. In addition, the source code can be obtained. For the usage of R on Windows and Mac systems the usage of precompiled binary distributions is recommended. (Figure 3)

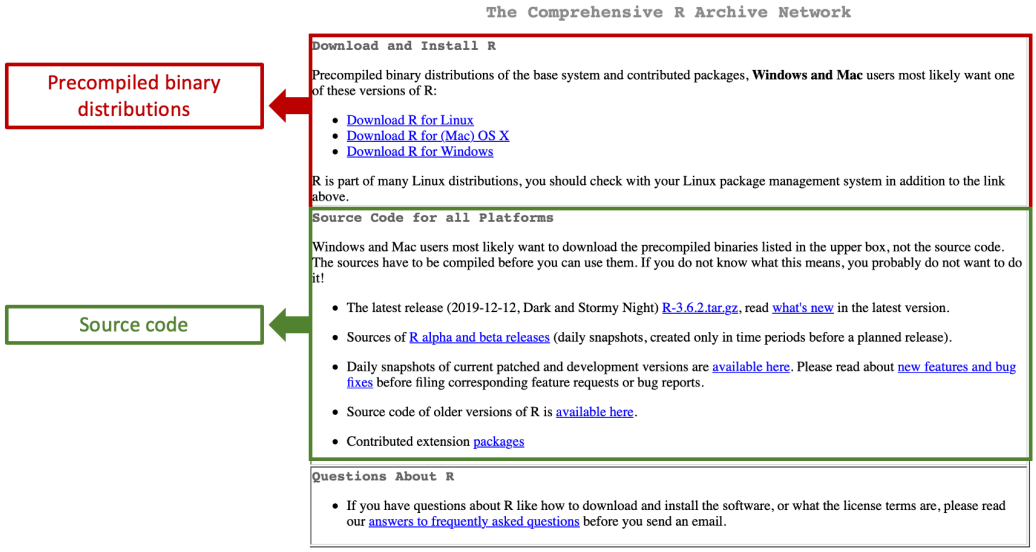
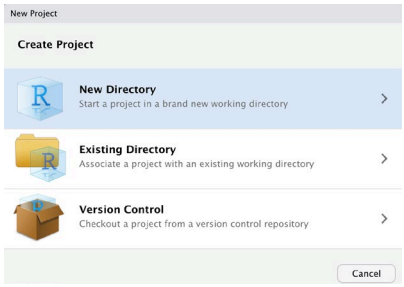


Figure 3: Obtainable precompiled binary distribution and source code for downloading the R statistical computing environment, provided by the University of Münster.

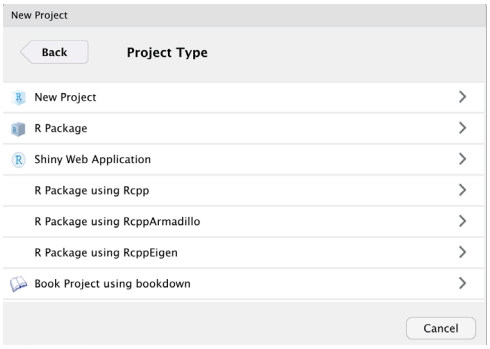
To download R from precompiled binary distributions, click on the provided link and follow the given instructions.
The R consortium additionally developed R Studio, an opensource integrated Development Environment software for the R statistic computing environment. For users, it is highly recommended to download the software, as it offers a clear and user-friendly interface.[4] To download R Studio, refer to the R studio download page (https://rstudio.com/products/rstudio/download/) and follow the given instructions.
Upon downloading RStudio, the software will automatically integrate the most recent R version, stored on your computer.

Creating a new Project in R Studio

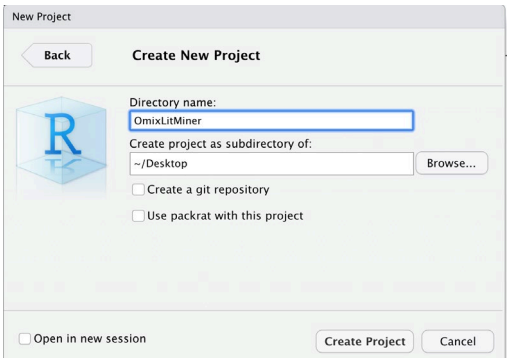
To use the OmixLitMiner tool in R Studio, it is required, to open a New R project. When opening RStudio for the first time, the software will ask you to create a new project that can be stored in a new or an existing working directory. The first R project has to be saved in a new working directory.



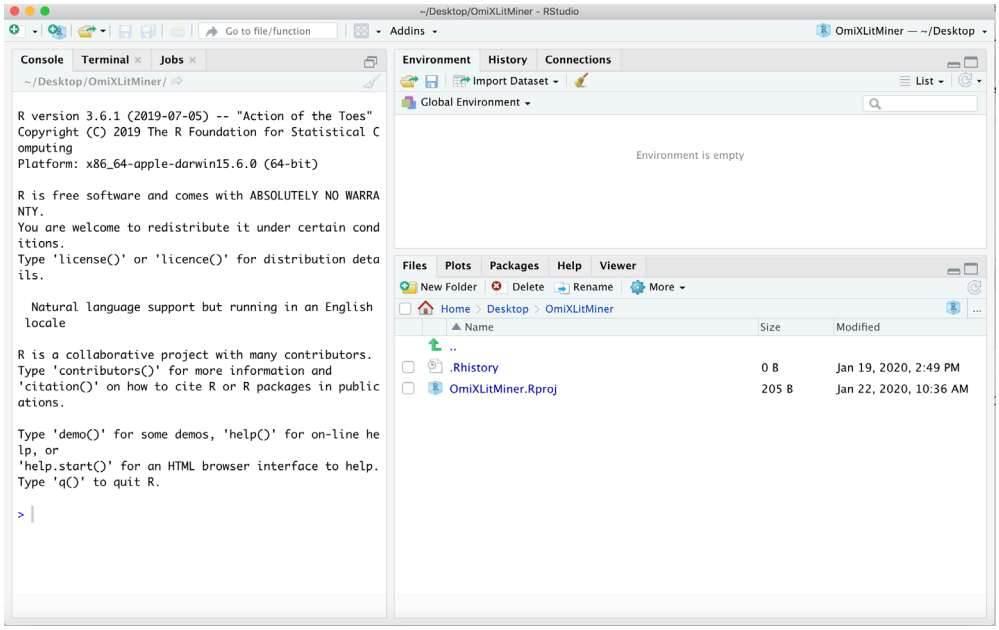
After choosing a working directory type, it is necessary to specify the project type, as R can be used: to use R packages, to design new R packages, to design web applications or to create a book project using bookdown. To use the OmixLitMiner Package, click on the *new project* button.



Following this, RStudio will ask you to specify the directory name and where your R project should be created as a subdirectory (for example at the Desktop of your computer). Fill the required fields and click on the *create project* button. A new Folder for your R project, with the given directory name, will be created as a subdirectory of the chosen storage location.

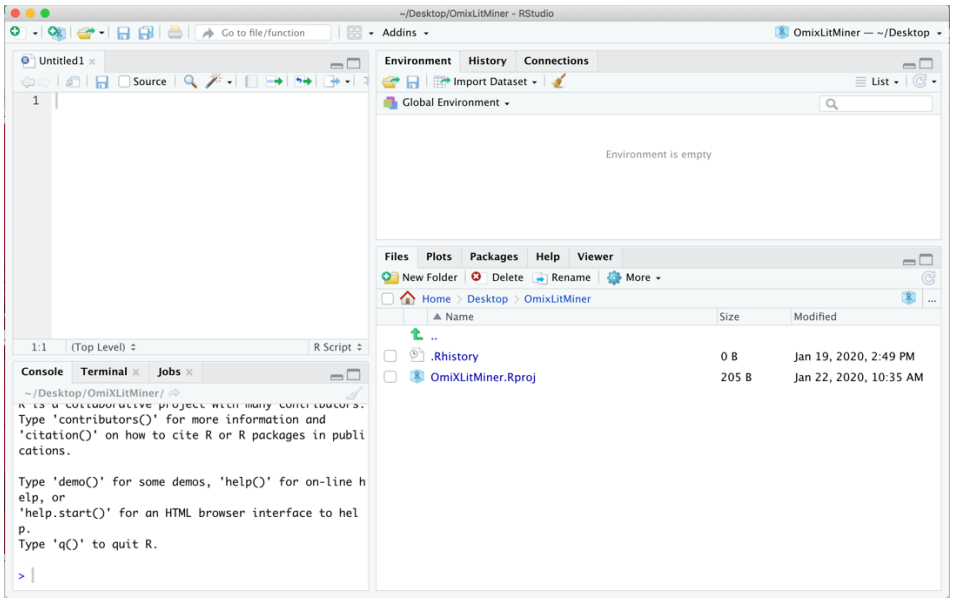


After creating a new R project, the RStudio graphical user interface, will appear. The interface is divided in three different windows: “Console, Terminal and Jobs”; “Environment, History and Connections” and “Files, Plots, Packages, Help and Viewer”.



For the usage of OmixLitMiner the Windows: “Console”, “Environment” and “Packages” are needed. The “Environment” window shows all user-generated data sets, to be used for statistical or graphical analysis using R. All R analyses require matrices as an input, that can be imported from spreadsheet programs like Microsoft Excel or directly created in the R environment. The “Console” window documents all commands executed by R, as well as potentially occurring errors and their cause. The “Package” window can be used to manually add R packages to your software environment, that provide additional functions, not given by the R base system or preinstalled, basic R packages.

To comfortably use tools like the OmixLitMiner, open an additional “Source” Window, that can be used to add commands, upload new input matrices to the environment and to install and load additional R. The additional Source window can be opened by clicking the blank page with a green plus on the top left or by using the keyboard shortcut *Ctrl-Shift-N*.



Preparing Data for Literature retrieval and datamining, using the OmixLitMiner tool

To perform Datamining and Literature retrieval for the sake of investigation certain Proteins and Gens in relation to a predefined Keyword (i.e. Cancer, Metastatsis or other GO Terms) OmixLitMiner requires a .xlsx file as an input, with values for the predefined categories ID, ID Type, TaxID, Keyword and KeywordInTitelOnly for each Protein or Gene of interest. The input data file should be constructed in Microsoft Excel, as exemplary shown in Figure 4 for a List of significant Proteins obtained from Mori et al. [1](Supplement 13, in Steffen P et al., OmixLitMiner A bioinformatics tool for prioritizing biological leads from omics data using literature mining*)*. To obtain an template for an input data file refer to the OmixLitMiners GitHub page (https://github.com/Sydney-Informatics-Hub/OmixLitMiner), click inst/extdata folder and download the file "input_uniprot_keywords.xlsx)

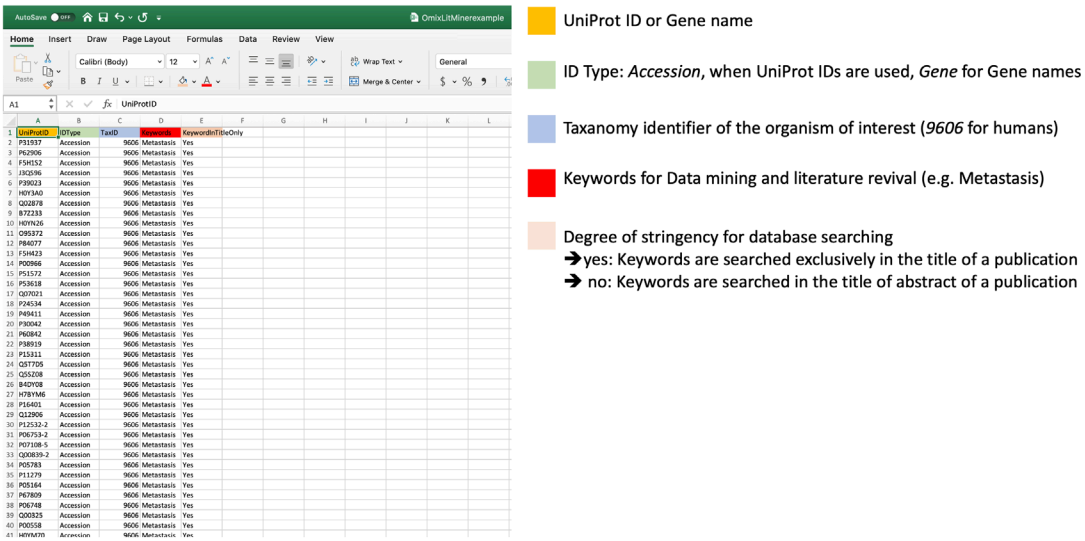


Figure 4: Required properties for .xlxs file, for datamining and literature revival using the OmixLitMiner tool on the example of putative Proteins identified significantly differential abundant between CRC with LN metastasis versus CRC without LN metastasis versus normal colonic mucosa and CRC adjacent normal mucosa.

As an ID value, the UniProt identifier or the Gene Name of each Protein/Gen is accepted. If searching for Proteins/Gens based on UniProt IDs, specify the ID Type using *Accession* as a value in the column IDType. For gene name-based searches, change the IDType to *Gene*.

The Column TaxID is used to specify the Taxonomy of the Organism, a Protein/Gene originated from. OmixLitMiner accepts Taxonomy identifiers, provided by the Uniprot group, that assign each Organism to a combination of digits. For humans, the Taxonomy identifier is *9606*, while the value in the TaxID column should be changed to *10090* (mus musculus), when working with Proteins/Genes of murine origin. To obtain TaxID values for different organisms, please refer to the Taxanomy search option, provided by the UniProt Database (https://www.uniprot.org/taxonomy/).
When associating an ID value with the wrong IDType or TaxID, OmixLitMiner will not obtain results for the candidate.

In the Keyword Column, the user defined Keywords of interest (e.g. Gene Ontology terms) have to be specified. If using more than one Keyword, variables have to be separated by a comma, followed by a space.

OmixLitMiner uses these variables to construct a query term for a PubMed search. This database search can be performed with varying degrees of stringency, that can be specified in the KeeywordsInTitelOnly column. Setting the KeeywordInTitelOnly value to *yes* signifies that publications are only accepted as hits in database searching if the variable terms are found in the title of a publication (strictest search). When setting the KeeywordInTitelOnly value to *no,* the retrieval is expanded to publications, where one or all query terms can be found either in the title or the abstract of a publication (more lenient search).

The prepared excel sheet should be saved in the .xlsx format in the working directory of your created R project folder.

Installing and loading all required R packages for OmixLitMiner

To use OmixLitMiner in RStudio, different R Packages are needed, that require installation, prior to usage.[5] The OmixLitMiner Tool was developed on the basis of the devtools R package, that includes package development tools for R, and therefore requires the installation of the devtools package. Furthermore, the openxlxs R package is needed. Openxlxs is required to write, read and edit .xlxs files.[6] To install devtools, the OmixLitMiner tool and the openxlxs package enter the following command line to the Source window in R studio:

```
install.packages("devtools")
devtools::install_github("Sydney-Informatics-Hub/OmixLitMiner")
install.packages("openxsxl")
```

Press the keyboard shortcut *Shift-Enter* after each command line to run the command or highlight the commands to be executed and click on the *Run* button at the right corner of the window.
If additional tools and packages are needed for your specified R version, you will be asked to install them automatically. If RTools, a collection of software packages used to build R itself or develop R packages when using the Microsoft Windows System, is not installed prior to downloading OmixLitMiner you will be prompted to install it. Follow the given instructions, or download Rtools from the following directory: https://cran.r-project.org/bin/windows/Rtools/. Reinstall OmixLitMiner by using the following command line:

```
install.packages("devtools")
devtools::install_github("Sydney-Informatics-Hub/OmixLitMiner")
install.packages("openxsxl")
```

Installation is only required if OmixLitMiner is used for the first time. Directories for Installed R packages are stored in the “library” directory for R.

For each new R project, installed R packages have to be loaded and attached from the library. To load packages for a new R session the function *library ()* is used. To load all required packages for OmixLitMiner enter the following command line:

```
library(OmixLitMiner)
library(openxlsx)
```

press the keyboard shortcut *Shift-Enter* after each command line, to run the command.

All required R packages for Literature retrieval and datamining with OmixLitMiner are now installed and loaded.

Importing the input xlxs file to RStudio:

You will need to import your prepared data input file. Save it as .xlsx in the same folder as the R Project. To load the .xlsx, the readWorkbook function will be used which reads data from excel files and converts them to R data.frames (a 2-dimensional, matrix like data structure in R) in the R environment. To upload data using the readWorkbook function enter the following command line:

```
dataforomixminer<-readWorkbook("OmixLitMinerexample.xlsx")
```

dataforomixminer→ Name of the data.frame created from the loaded xlxs file. The name can be defined by the user but cannot start with a numeric value.

readWorkbook()→R function, reads from excel file or workbook and converts objects into data.frames

OmixLitMinerexample.xlsx→Name of the input data file, deposited in the R projects working directory. Change this according to your file.

Press the keyboard shortcut *Shift-Enter*, to run the command.

A new data.frame will appear in the environment window (Figure 5).

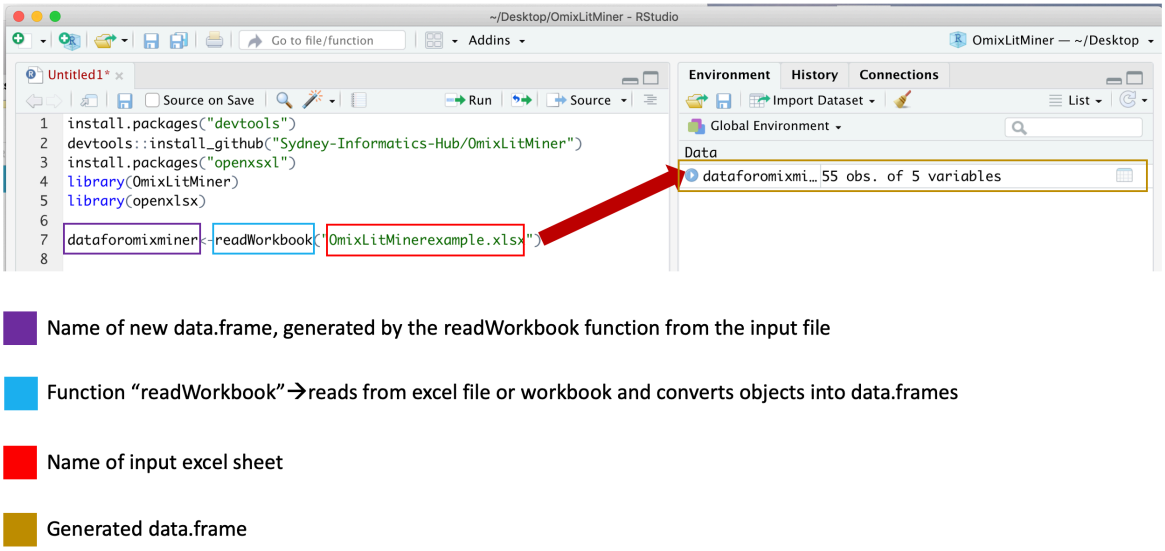


Figure 5: Exemplary visualization of the uploading data and creation of a R data.frame from an xlxs file, using the readWorkbook function.

For further information about the readWorkbook function and possible modifications refer to the associated R documentation page for the open xlxs package (https://www.rdocumentation.org/packages/openxlsx/versions/4.1.4) or enter the following command line:

```
?readWorkbook
```

Press *Shift-Enter* to get directed to the help page for the readWorkbook function.

Troubleshooting: Should your .xlsx not be loaded into R please make sure the file is in the right working directory. Otherwise, you can set the working directory to where the input file is located manually by clicking on the "Files" tab and navigating to the folder. Afterwards, click on the "More" button and select "Set As Working Directory". This will set the working directory of the project to the specified folder. This folder will now be used as default for loading and saving files. Run the command again:

```
dataforomixminer<-readWorkbook("OmixLitMinerexample.xlsx")
```

Run OmixLitMiner in R

The OmixLitMiner tool can be used for literature retrieval and datamining for all proteins/genes and keywords deposited in the uploaded data.frame. To run OmixLitMiner enter the following command line:

```
OmixLitMinerresult<-omixLitMiner(dataforomixminer,output.file =
"omixLitMinerResult.xlsx", plots.dir = "plots")
```

OmixLitMinerresult → Name of result data.frame generated by the OmixLitMiner tool. The name can be defined by the user but cannot start with a numeric value.

omixLitMiner → OmixLitMiner function for literature retrieval and data mining

dataforomixminer → Name of the data.frame created from the loaded xlxs file. (See "Importing the input xlxs file to RStudio" section). This name needs to match the one from the data.frame.

OmixLitMinerresult.xlsx → .xlsx file, containing all results, generated by the OmixLitMiner tool. Will be created in the xlxs format and deposited in the R projects working directory. The name can be defined by the user but cannot start with a numeric value.

Plots.dir="plots" →Plots, generated as results by the OmixLitMiner tool will be stored in a new folder called "plots" in the R projects working directory.

press the keyboard shortcut *Shift-Enter*, to run the command.

OmixLitMiner will perform database searching and create a combined .xlsx format result file, that can be obtained from the folder of the R projects working directory and is stored under the given result file name (*in this example OmixLitMinerresult.xlsx)*

For further information about the OmixLitMiner tool, possible modifications and trial datasets refer to the associated GitHub page (https://github.com/Sydney-Informatics-Hub/OmixLitMiner)

Literature

1. Mori, K. *et al.* Successful identification of a predictive biomarker for lymph node metastasis in colorectal cancer using a proteomic approach. *Oncotarget* **8**, 106935–106947 (2017).
2. Hornik, K. The R FAQ. (2018).
3. R Development Core Team, R. *R: A Language and Environment for Statistical Computing*. *R Foundation for Statistical Computing* (2011). doi:10.1007/978-3-540-74686-7.
4. RStudioTeam. RStudio: Integrated Development Environment for R. (2015).
5. Wickham, H. & Chang, W. devtools: Tools to Make Developing R Packages Easier. (2019).
6. Schauberger, P. & Walker, A. openxlsx: read, write and edit xlsx files. (2019).