Rich Analysis of Variable gene Expressions in Numerous tissues

RAVEN

Documentation and user manual

Author:	Julia Sophia Gerke, MSc
Supervisor:	Thomas Grünewald, MD, PhD
Current version:	Raven-v2017.5
Published:	May 2017

Contents

1 Installation

1.1 Requirements

Raven runs on Windows, Linux and Mac.

For Windows and Linux a **64-bit version** is obligatory. A random access memory **(RAM) of at least 8 GB** is necessary. However for big data files as provided from us, 32 GB are recommended for the easy and fast handling. You are not sure if your computer fulfills the requirements? Have a look at Table 1.1 and Figure 1.1 to find out more about your operating system (OS) and RAM or ask your system administrator for help.

Raven is a Java application and requires **Java 8** (see more in the next Section 1.2).

When using a laptop, your screen should have a resolution of at least **1440x900p** to display the application correctly. (Desktop computers normally have a higher resolution)

For the usage of the Peptide Matching Pipeline as described in 4.4 a **stable internet connection** is essential. Make sure you have the permission to let RAVEN send and recieve queries to internet servers. This is sometimes restricted in some institutions with high data security. In this case talk to your IT-administrator.

> **Tip:** This really scares you or the Installation Chapter overburdens you? Ask your IT-administrator for help to check the requirements mentioned above and to install or update Java on your computer. Afterwards you can continue with Section 1.3 to start RAVEN.

[1] https://support.microsoft.com/en-us/help/827218/how-to-determine-whether-a-computer-is-running-a-32-bit-version-or-64-bit-version-of-the-windows-operating-system

You have OS ...	... and use ...	... and look at
🍎	🍎 → 'About This Mac'	Figure 1.1a
🐧	Ctrl + Alt + T → write: `uname -a` → write: `less /proc/meminfo`	Figure 1.1c
⊞ 10 ⊞	⊞ + X → 'system' check out the Microsoft help[1] for older versions	Figure 1.1b

Table 1.1: Keyboard shortcuts depending on your operating system (OS) to open the information panel about your system. Your RAM is shown there. For Windows OS also the version (64-bit or 32-bit) can be seen.

1.2 Installation

Download the current version of RAVENsoftware (`RAVEN_v20XX.X.jar`) from GitHub (https://www.github.com/JSGerke/RAVENsoftware). Click on the green `'download'` button at the right side and choose `'Download ZIP'`

> Tip: You want to analyse a large file and do not know how to get the path to your downloaded file? Easiest way is to save it in your Documents folder → path: `Documents`

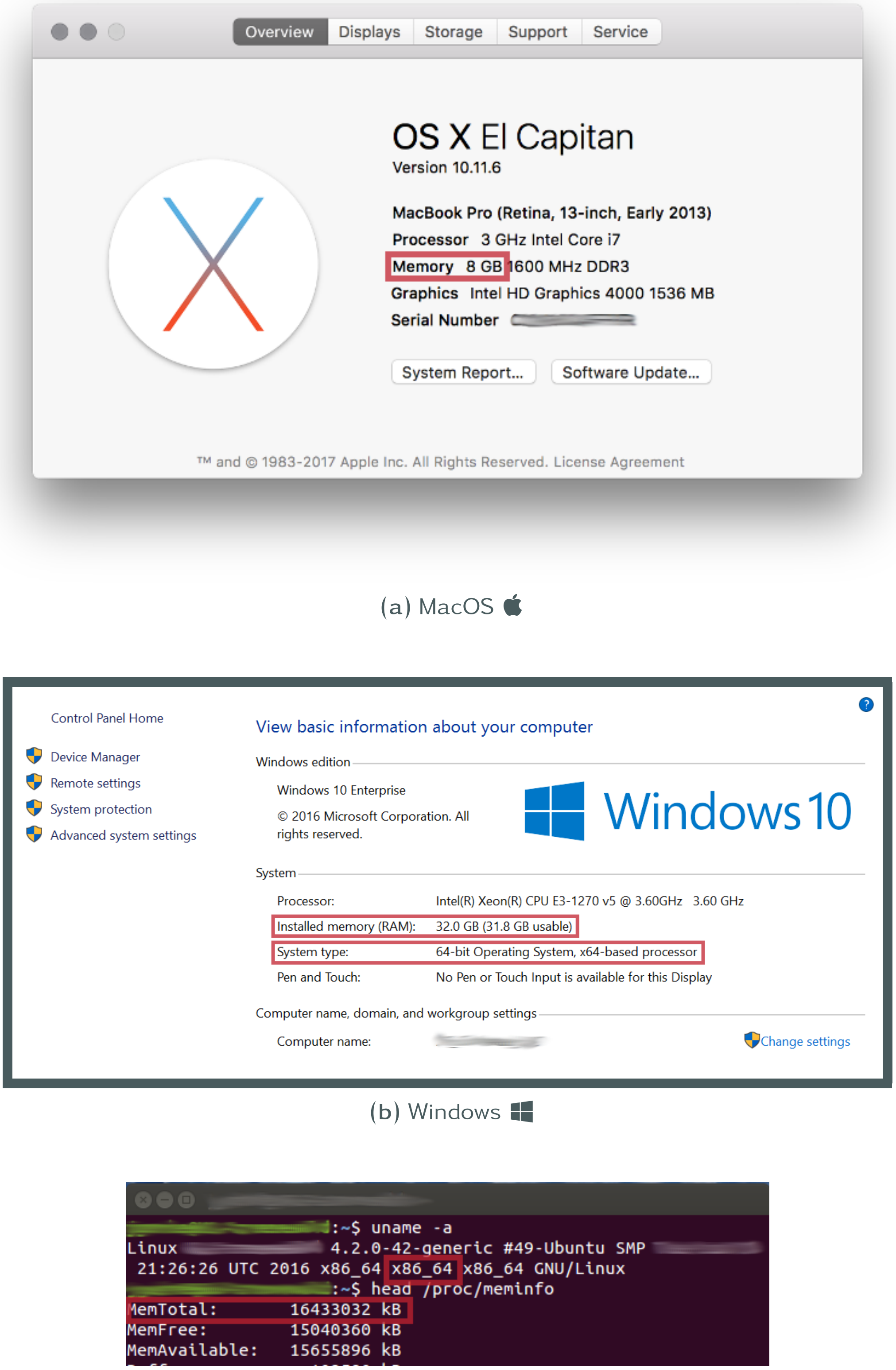
Check your Java version!

Before you run RAVEN the first time make sure Java is installed on your computer and you have the right version of it (> 1.8.0_91 → Java 8). If you have Windows make sure you have installed the 64-bit version. To check this out use one of the following approaches:

- You can display your Java version via commandline. Open your terminal (🍎 , 🐧) or command prompt (⊞) with the keyboard shortcuts shown in Table 1.2 and type the command below:

  ```
  java -version
  ```

  The shown version should be 1.8.0_91 or higher. Additionally, it should include '64-Bit Server VM' both as shown in Figure 1.2. Otherwise you probably have an old version or the wrong system. If you get an error message for unknown



OS X El Capitan
Version 10.11.6

MacBook Pro (Retina, 13-inch, Early 2013)
Processor 3 GHz Intel Core i7
Memory 8 GB 1600 MHz DDR3
Graphics Intel HD Graphics 4000 1536 MB
Serial Number [redacted]

™ and © 1983-2017 Apple Inc. All Rights Reserved. License Agreement

(a) MacOS

View basic information about your computer

Windows edition

Windows 10 Enterprise
© 2016 Microsoft Corporation. All rights reserved.

System

Processor:	Intel(R) Xeon(R) CPU E3-1270 v5 @ 3.60GHz 3.60 GHz
Installed memory (RAM):	32.0 GB (31.8 GB usable)
System type:	64-bit Operating System, x64-based processor
Pen and Touch:	No Pen or Touch Input is available for this Display

Computer name, domain, and workgroup settings

Computer name: [redacted]

(b) Windows ⊞

```
[redacted]:~$ uname -a
Linux [redacted] 4.2.0-42-generic #49-Ubuntu SMP [redacted]
 21:26:26 UTC 2016 x86_64 x86_64 x86_64 GNU/Linux
[redacted]:~$ head /proc/meminfo
MemTotal:       16433032 kB
MemFree:        15040360 kB
MemAvailable:   15655896 kB
```

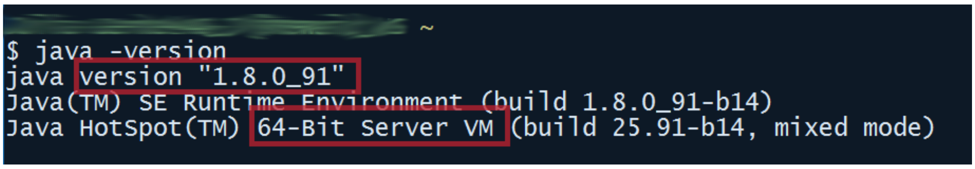
(c) Linux 🐧

Figure 1.1: Where to find your RAM and OS version (both framed red) for different operating systems (, 🐧, ⊞). These information panels can be opened as described in Table 1.1.

command, java is not installed on your computer at all. In both cases update your version or install the right java version as described in the next Section 1.2.

You have OS ...	... and use ...
🍎	⌘ + T
🐧	Ctrl + Alt + T
⊞ 10 ⊞ 7	⊞ + X → 'command prompt' ⊞ menu → search for 'command prompt'

Table 1.2: Keyboard shortcuts depending on your operating system (OS) to open the terminal (MacOS, Linux) or command prompt (Windows)



```
[redacted] ~
$ java -version
java version "1.8.0_91"
Java(TM) SE Runtime Environment (build 1.8.0_91-b14)
Java HotSpot(TM) 64-Bit Server VM (build 25.91-b14, mixed mode)
```

Figure 1.2: Output for 'java -version' on Windows command prompt with red framed java version and its OS dependencies. This output can vary subject to your OS.

- If you are not familiar with the command line, you can follow the instructions from the java website description[2] on your browser to find your installed version of Java.

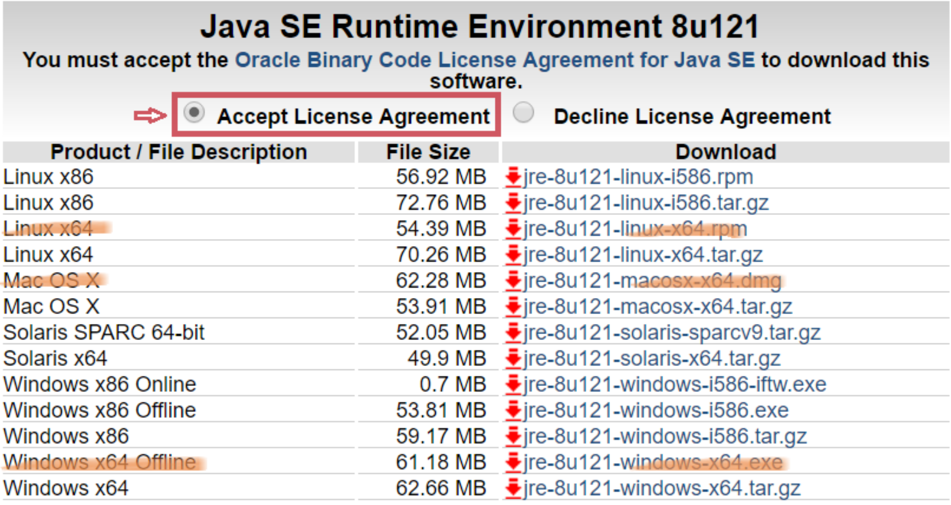
Download Java 8

If you do not have Java 8 already installed on your computer you can download the Java Runtime Environment (JRE) on the website[3] of its developer ORACLE. A detailed view of the website is shown in Figure 1.3, though make sure you accept the license agreement in order to start the download (red framed). Choose the corresponding executables highlighted in orange. After saving the executable application, start the installation process with double click and follow the shown instructions until the installation process is finished. Done!

> **Tip:** It is also possible to download Java from its own website. However, this mostly refers to the Java 32-bit version. Be careful when you are not familiar with it. Rather download java from ORACLE as described above!

[2] https://java.com/en/download/help/version_manual.xml

[3] http://www.oracle.com/technetwork/java/javase/downloads/jre8-downloads-2133155.html



Java SE Runtime Environment 8u121

You must accept the Oracle Binary Code License Agreement for Java SE to download this software.

● Accept License Agreement ○ Decline License Agreement

Product / File Description	File Size	Download
Linux x86	56.92 MB	jre-8u121-linux-i586.rpm
Linux x86	72.76 MB	jre-8u121-linux-i586.tar.gz
Linux x64	54.39 MB	jre-8u121-linux-x64.rpm
Linux x64	70.26 MB	jre-8u121-linux-x64.tar.gz
Mac OS X	62.28 MB	jre-8u121-macosx-x64.dmg
Mac OS X	53.91 MB	jre-8u121-macosx-x64.tar.gz
Solaris SPARC 64-bit	52.05 MB	jre-8u121-solaris-sparcv9.tar.gz
Solaris x64	49.9 MB	jre-8u121-solaris-x64.tar.gz
Windows x86 Online	0.7 MB	jre-8u121-windows-i586-iftw.exe
Windows x86 Offline	53.81 MB	jre-8u121-windows-i586.exe
Windows x86	59.17 MB	jre-8u121-windows-i586.tar.gz
Windows x64 Offline	61.18 MB	jre-8u121-windows-x64.exe
Windows x64	62.66 MB	jre-8u121-windows-x64.tar.gz

Figure 1.3: JRE download panel on the ORACLE website, choose the right one (orange) to your operating system 🍎, 🐧or ⊞.

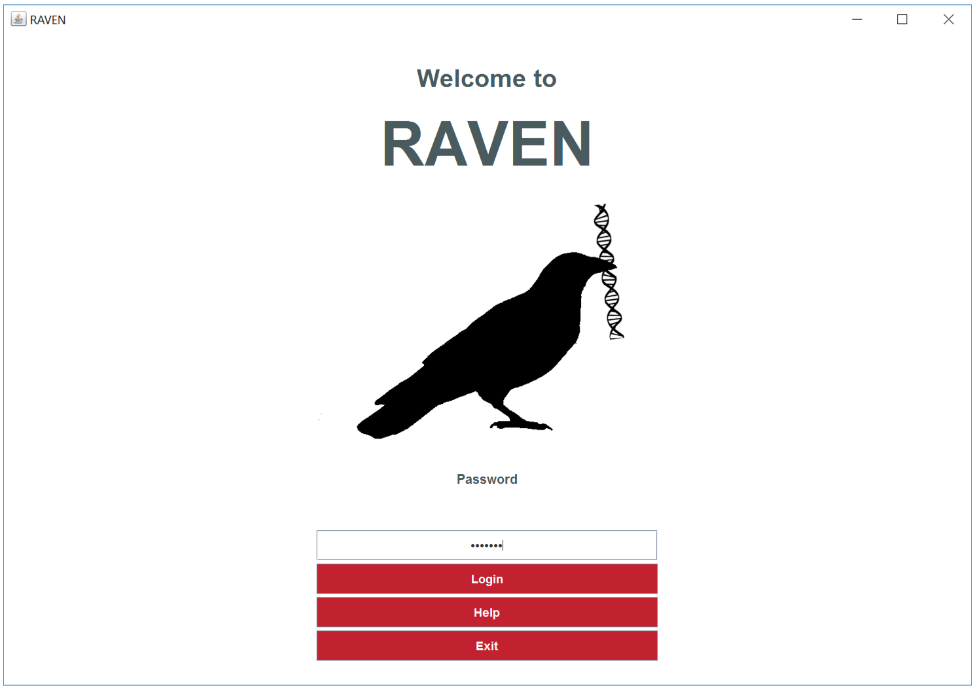


Figure 1.4: RAVEN start window. Enter the password to proceed (section 1.4).

1.3 Start RAVEN

You can easily start RAVEN by double click on the downloaded file. The Login Window as shown in Figure 1.4 will open immediately to welcome you to RAVEN...

If you prefer to run RAVEN from command line navigate into the directory, where you saved RAVEN and start the java application as described below. Replace X with your downloaded version.

```
cd yourDirectoryPath/
java -jar RAVEN_v20XX.X.jar
```

For large input files and depending on your RAM you will have to start RAVEN via command line. Have a look at Table 2.1 to check if this applies to you and start Raven as shown below.

```
cd yourDirectoryPath/
java -Xmx6g -jar RAVEN_v20XX.X.jar
```

1.4 Login

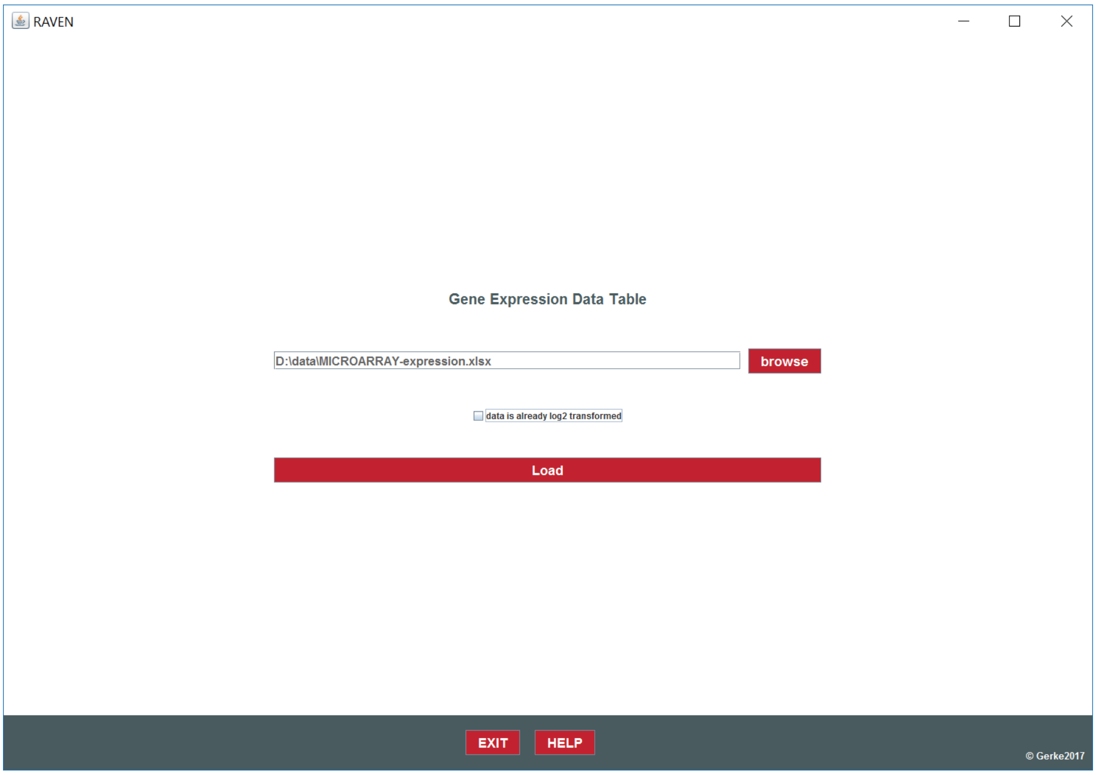
RAVEN can only be accessed with a password (shown in Figure 1.4), which is free for academic use only. Please contact us (thomas.gruenewald@med.uni-muenchen.de) to receive the login password to start exploring RAVEN.

2 Upload Data

2.1 Upload

Enter or browse your input data file as shown in Figure 2.1. For later calculations it is important to know if your data is already log2 transformed or still in its natural scale. Activate the checkbox accordingly before pushing the ′upload′ button. The input data table can be a **text file** (.txt) or in **Excel format** (.xlsx). For more details and requirements regarding your input data have a look at Section 2.3 and 2.4.

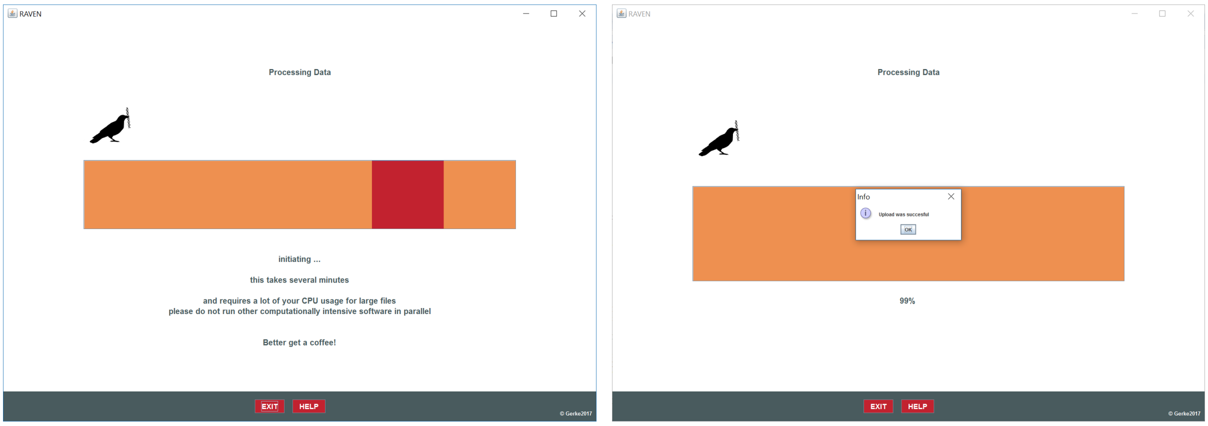
After the upload you will receive a short notification if your upload was successful (Figure 2.2b) or not. A failed upload mostly derives from an formatting error of your data. An error message will notify you about the flaw to help you to correct it. For this, also check out Section 2.3 and 2.4



D:\data\MICROARRAY-expression.xlsx

☐ data is already log2 transformed

Figure 2.1: Gene Expression data upload panel for log2 transformed or natural scaled files in text or Excel format which will directly appear after your login.



Processing Data

initiating ...

this takes several minutes

and requires a lot of your CPU usage for large files
please do not run other computationally intensive software in parallel

Better get a coffee!

(a) View on upload progress ⌛. Watch RAVEN working or get a ☕ break!

Processing Data

Upload was succesful

99%

(b) Info message for successful upload.

Figure 2.2: RAVEN processing your input data.

2.2 Processing Time

Depending on the size of your input file and the available RAM, the processing time can vary from a few seconds to several minutes (Figure 2.2a). In worst case your file is too large and cannot be processed at all. Check out Table 2.1 to see how many samples you can process dependent on your RAM. Make sure your file size does not diverge a lot from those in the table. Otherwise, consider the listed suggestions below.

For large files it can be necessary to start RAVEN differently. They are marked with ! in Table 2.1. In this case follow the description in Section 1.3 on starting RAVEN on command line for large files. For files labeled with ✔ you can proceed as normal.

How to reduce processing time

Your file is to big and cannot be processed or you want to speed up the loading process? Try the following suggestions:

- use text format instead of Excel format
- reduce your decimals to 2-3 digits
- close other parallel running computational intensive software
- change to another computer with higher performance (RAM)

# sample size	RAM			
[file size]	4 GB	8 GB	16 GB	32 GB
500 [62MB]	✓	✓	✓	✓
1000 [124MB]	!	✓	✓	✓
1500 [185MB]	✖	!	✓	✓
2000 [247MB]	✖	!	✓	✓
2600 [330MB]	✖	!	!	✓
3000 [364MB]	✖	!	!	✓
2784 [587MB]	✖	!	!	✓

Table 2.1: Approximate number of samples (including their file size in MB) that can be processed by RAVEN dependent on the available RAM of your computer, assuming that for every sample around ~20000 genes were measured. The files are text files and all measurements (natural scale) are rounded to 2 decimals. Though, the last table row refers to natural scale raw data in Excel format. ✓ file can be processed; ✖ file cannot be processed; ! Raven must be started differently to process file.

2.3 Available Input Files

Together with the RAVEN, we also published two datasets each consisting of several cancer and normal tissue samples from Affymetrix microarray data downloaded from Gene Expression Omnibus (GEO). While for both files the normal tissue samples are the same, they differ in their tumor types. Both files (file names) are already in the right format to be accepted as input for RAVEN with its data already log2 transfomed??. You can download the files as supplementary table or the journals website.

All samples underwent normalization and a quality check described by [Baldauf and Gerke et al].

Normal Tissues

The normal tissues in our files cover 71 different types of in total 929 samples including the most important normal tissues.

Most Common Cancer

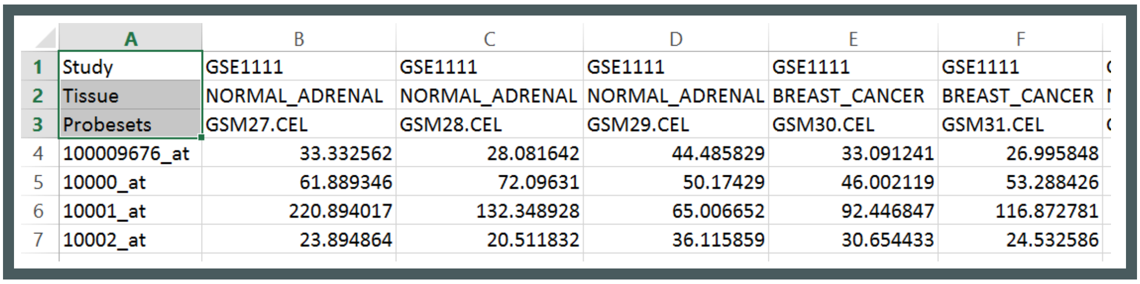
This file contains samples from 16 of the most common cancers in human of in total XXX patients. The tumors include breast, bladder, cervical, colon, endometric, esophagus, kidney, liver, lung, oral, ovarian, pancreatic, prostate, skin, stomach, thyroid cancer.

Pediatric Sarcoma

This file focuses mainly on sarcomas and pediatric cancers. Beside the normal tissues, 1749 samples of 50 tumor entities are assembled in this dataset, such as Ewing sarcoma.

2.4 Creating a new Input File

It is also possible to upload your own data to RAVEN. Make sure your data file matches the specifications and requirements listed below, otherwise your data file upload will fail. An excerpt of an (affymetrix) microarray example file is shown in Figure 2.3.



	A	B	C	D	E	F
1	Study	GSE1111	GSE1111	GSE1111	GSE1111	GSE1111
2	Tissue	NORMAL_ADRENAL	NORMAL_ADRENAL	NORMAL_ADRENAL	BREAST_CANCER	BREAST_CANCER
3	Probesets	GSM27.CEL	GSM28.CEL	GSM29.CEL	GSM30.CEL	GSM31.CEL
4	100009676_at	33.332562	28.081642	44.485829	33.091241	26.995848
5	10000_at	61.889346	72.09631	50.17429	46.002119	53.288426
6	10001_at	220.894017	132.348928	65.006652	92.446847	116.872781
7	10002_at	23.894864	20.511832	36.115859	30.654433	24.532586

Figure 2.3: Example excerpt of an input file for RAVEN. Please note that the first three cells (A1-3) are required to be exactly as shown here!

Input file specification

- the **first line** (or header) starts with the term 'Study', followed by each samples study id (e.g. GSE123). As this is never processed in any method you can also store some sample information in these cells.
- the **second line** starts with the term 'Tissue' followed by the normal tissue or cancer of each sample. Normal tissues **always** have to begin with 'NORMAL_' ! (e.g. NORMAL_BREAST, BREAST_CANCER)
- the **third line** starts with the term 'Probesets' followed by a unique id for each sample. (e.g. GSM3333)
- the forth and **consecutively lines** contain the measured gene expressions (affymetrix) for all samples beginning with the gene id of the corresponding gene in the first column.
- the **gene id** must be an entrezgene id (GeneID) followed by the suffix '_at', as often in affymetrix data. (e.g. 1000_at). Otherwise, check out the description for ID translation below this list.
- all columns (except first one) must be **sorted alphabetically** by their samples' tissue (second line)
- within cells of the first 3 rows no whitespace is allowed. Use _ to join several words to one per cell (COLON_CANCER ✔; ~~COLON CANCER~~ ✖)

- for **decimal mark** the application supports both point (e.g. US, UK) and comma (e.g. D, EU), but no thousands separator for digit grouping (1000.1 ✔ ; ~~1,000.1~~ ✖)
- **no empty cells** within the data matrix/table. Use NA or NaN instead.
- if your file is a text file use **tab separated columns**
- if your file is an Excel file make sure your data table is on the **first sheet**. All other sheets will be ignored.
- if your file is an Excel file it has to **end with** `.xlsx`. Excel files created with older versions than Excel2007 have a different ending and are not accepted.

Tip: RAVEN is optimized to analyze micro array gene expression data. However, you can also run the software with your own RNAseq data, as long as you prepare your data accordingly. Make a quality assessment on your data and preprocess it correctly. Afterwards, normalize your reads and adapt the input file specifications necessary to upload data to RAVEN. However, you probably have to adjust the IDs accordingly to match the specifications (see below)

ID translation

RAVEN only accepts the entrezgene id (GeneID) for genes. If you have another id, you can convert it to GeneID with the ID converter (second tab!) of biodb.jp here[1].

It is also important that your gene ids end with the suffix `'_at'`. If the suffix it missing you can add it via Excel or on command line.

Excel: Open your file in Excel (also possible with a text file but choose the right separator!). Add a new sheet and rename it to `'suffix'`. Copy your first column to the new sheet. There, start with cell B4 and enter the first formula below. Drag this cell (by clicking on the small square in the bottom right cell corner) over the whole column B. The formula will adapt accordingly. Switch back to your original data sheet and override cell A4 with the formula as shown below (second one). Also drag this cell over the whole column A. Save your modifications.

```
=A4&"_at"

=suffix!B4
```

[1] http://biodb.jp/

cmd line: (for Windows only with cygwin installed) Open the terminal and navigate into the directory of your data file as already described in Section 1.2) and write the following. (only working for text files!)

```
cd /yourDirectory/

head -n3 inputData.txt > newInputData.txt

awk -v OFS=$’\t’ ’NR>3 { $1=$1"_at"; print }’
inputData.txt » newInputData.txt
```

3 General Navigation

3.1 Application

RAVEN offers you several statistical analysis methods and bioinformatic pipelines to analyse your gene expression data.

Methods implemented in RAVEN:

- **Extracting Subsets** (see 4.1)
- **Direct Gene Expression Comparison** (see 4.2)
- **Cancer Specific Gene (CSG) Score** (see 4.3)
- **Peptide Matching Pipeline (PMP)** (see 4.4)

Setting Parameters

In general, all methods are based on the same operational procedure. See the stepwise procedure illustrated with black circled numbers in Figure 3.1.

❶ selecting one of the tabs on the left side for the method you want to run (→ tabs are shown at the top)

❷ specify if you want to perform this method on a gene or tissue

❸ enter your favorite gene or tissue

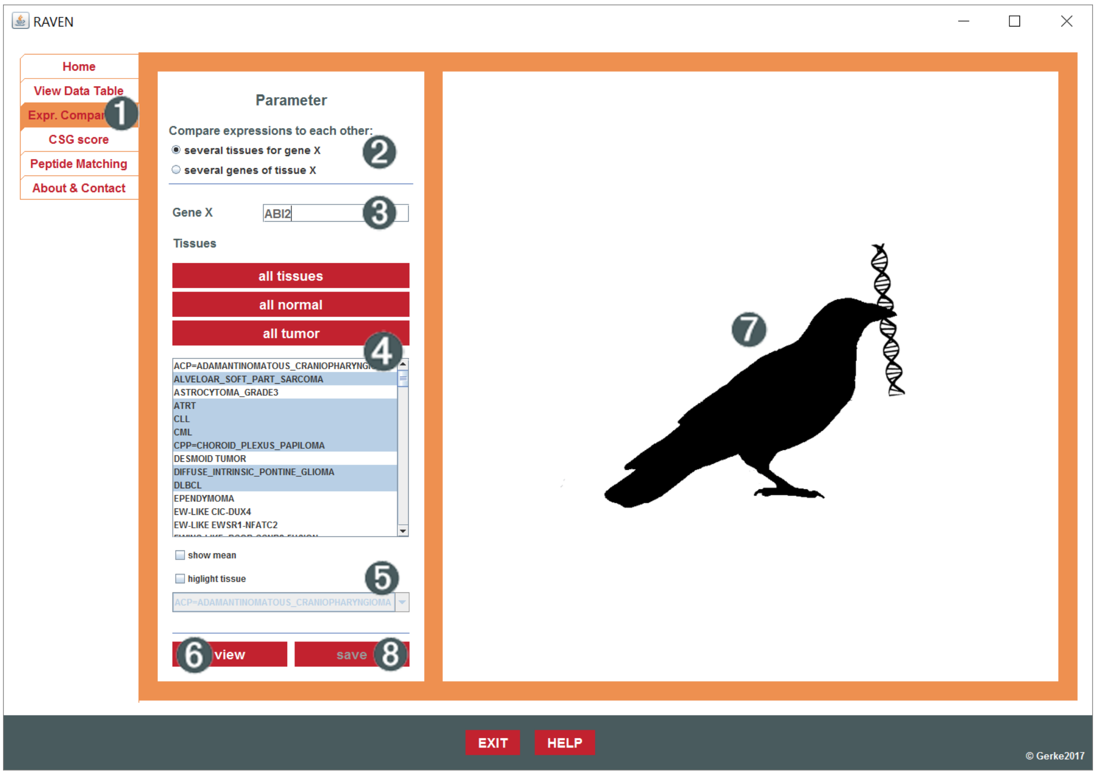
❹ specify the genes or tissues you want to compare/test your favourite against (gene vs tissues or tissue vs gene only!)

❺ select some additional settings

❻ click on ′`view`′ button to run the analysis method

❼ the result will show up here after RAVEN is done with computing

❽ click on ′`save`′ button to save your results or data used for the analysis



Parameter

Compare expressions to each other:
● several tissues for gene X
○ several genes of tissue X

Gene X ABI2

Tissues

ACP=ADAMANTINOMATOUS_CRANIOPHARYNGI
ALVELOAR_SOFT_PART_SARCOMA
ASTROCYTOMA_GRADE3
ATRT
CLL
CML
CPP=CHOROID_PLEXUS_PAPILOMA
DESMOID TUMOR
DIFFUSE_INTRINSIC_PONTINE_GLIOMA
DLBCL
EPENDYMOMA
EW-LIKE CIC-DUX4
EW-LIKE EWSR1-NFATC2

□ show mean

□ higlight tissue

ACP=ADAMANTINOMATOUS_CRANIOPHARYNGIOMA

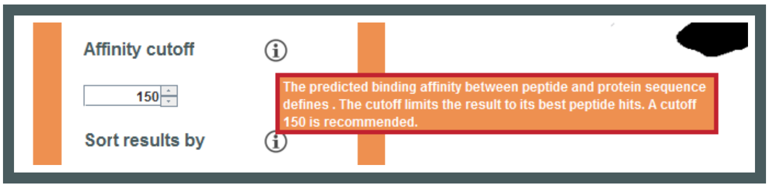
Figure 3.1: Step wise procedure to run an analysis with RAVEN shown on an Example.

While you can add only one gene to the one line text fields in the parameter, you can add several genes in the larger text areas to apply a method with same settings on multiple genes at once. Thereby, you can use the gene symbol as well as the gene ID (entrez) as identifier. However for the peptide matching pipeline you can only use gene symbol. Gene symbols are not case sensitive.

Tissue types cannot be entered, but only selected from the available ones of your input file.

Additional Information

Some parameters or options, which might not be intuitive have a small information symbol ⓘ next to it. A information panel will pop up when you move your mouse pointer onto the symbol (Figure 3.2).



Affinity cutoff

150

Sort results by

The predicted binding affinity between peptide and protein sequence defines . The cutoff limits the result to its best peptide hits. A cutoff 150 is recommended.

Figure 3.2: Inbuilt information with pop up text.

Help

If you need help while using RAVEN and do not have this manual at your fingertips you can open this documentation via the red HELP button in the menubar at the bottom of the application.

3.2 Working directory

Setting a working directory is optional but saves you a lot of time and effort saving multiple files. Browse or type (non existing paths are colored red) your favorite directory where you want to store your data later. To confirm your directory click the 'set directory' button (see Figure 3.3). Then the file manager will automatically open this directory for you when saving data. Only existing directories are allowed. Otherwise, the button will be disabled with its writing fading to grey.

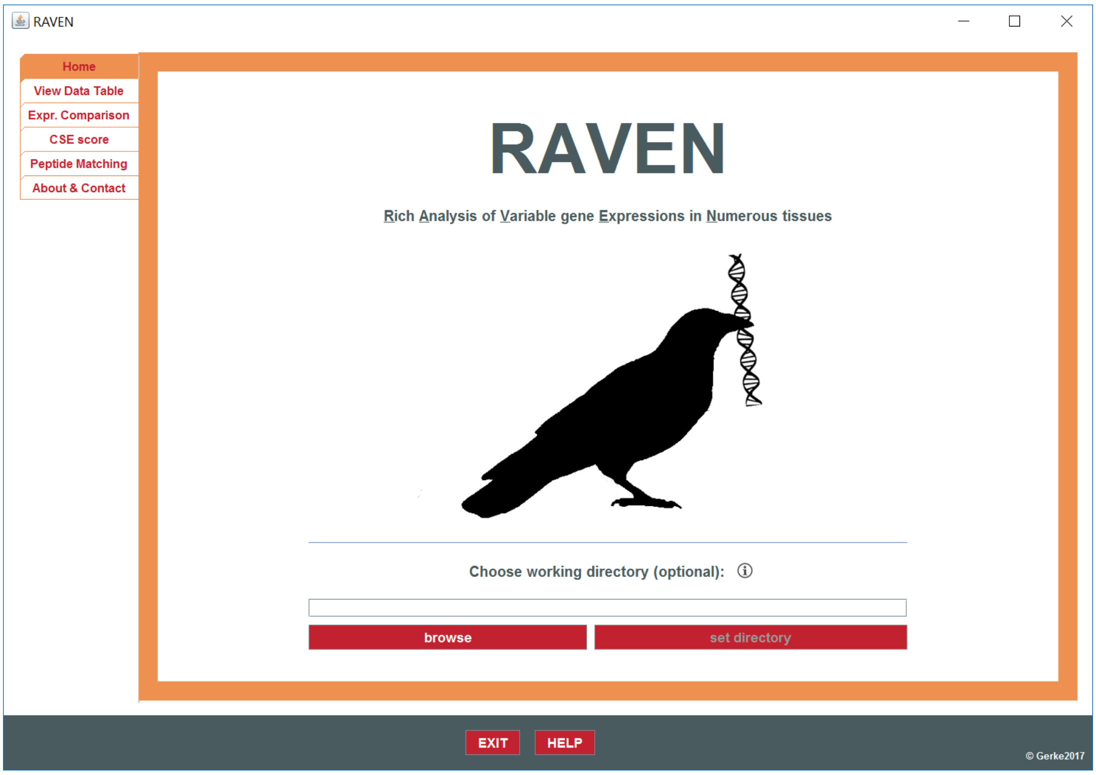


Figure 3.3: Raven Home: set the working directory via the 'set' button (disabled here because of missing/unvalid directory in the textfield)

3.3 Table

The results are always shown or accompanied by a table.

Navigation and Selection

Mostly a table is sorted by the results p-value or most important score. The current sorting column is marked with a small adumbrated arrow, pinpointing to its sorting direction. To resort the table by values of another column, click on the corresponding column header. A second click on the same column changes from increasing order to decreasing or other way around. It is also possible to change the order of the columns. Drag and drop the columns by their header into your desired order. When saving table data the reorder is taken.

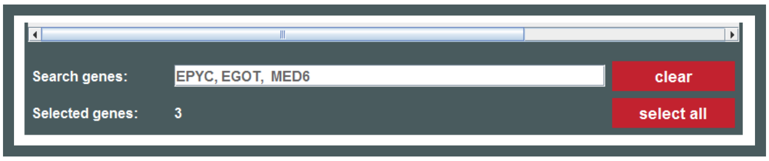
You can select several rows to accentuate them from the remaining ones or to save a subset of data. For an easy navigation you can use the keybord as described in table 3.1. On the search bar (Figure 3.4) below the table you can also see the number of selected rows.

Select...	⊞ / 🐧	
all rows	Ctrl + A	⌘ + A
consecutive rows	⇧ + 🖱 last row to select	⇧ + 🖱 last row to select
scattered rows	Ctrl + 🖱 unselected rows	⌘ + 🖱 unselected rows
deselect a certain row	Ctrl + 🖱 selected single row	⌘ + 🖱 selected single row

Table 3.1: Keyboard combinations to select or deselect your desired table rows dependent on your operations system.

Search

You are looking for a certain gene or tissue? Search for it via the search bar below the table. All rows matching the search word will be filtered in the table. Separate several search terms by comma. If part of the table, you can search for categories like entrezgene id, genesymbol, tissue, tumor, peptide, allele and protein. Be careful when using the `'select all'` button, as this is a simple text search you will also filter other genes containing your search terms (e.g. `SOX1` → `SOX1, QSOX1, SOX12` ...)



Search genes: EPYC, EGOT, MED6

Selected genes: 3

Figure 3.4: Inbuilt information with pop up text.

4 Methods

Here some little mini method Introduction quarter page.

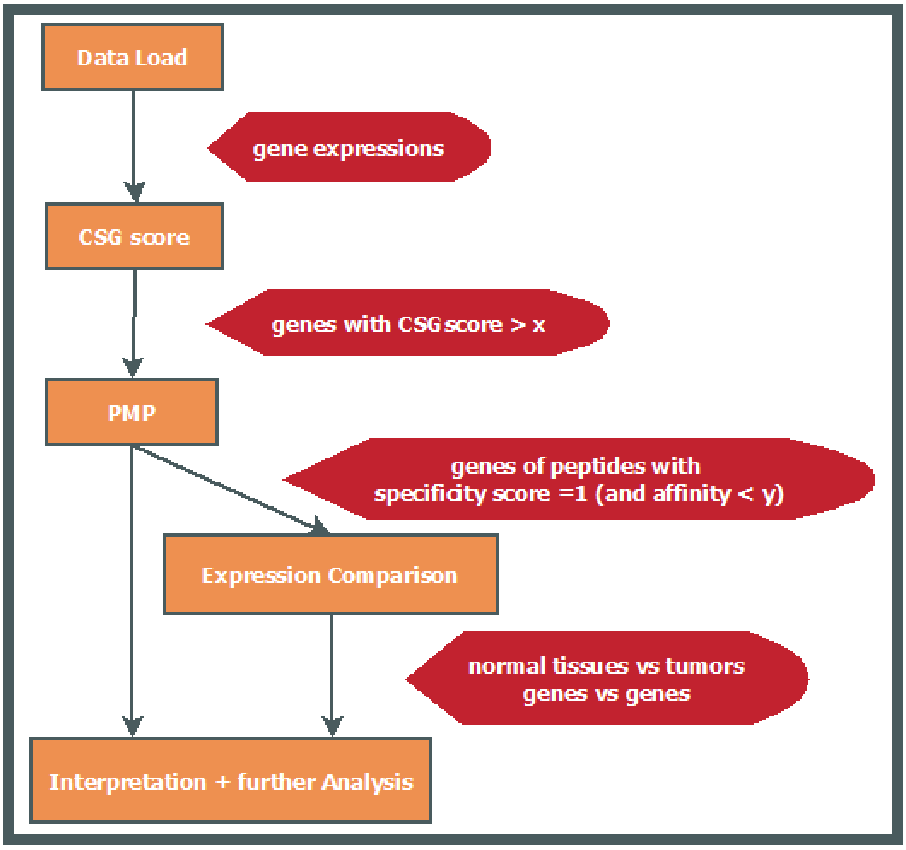
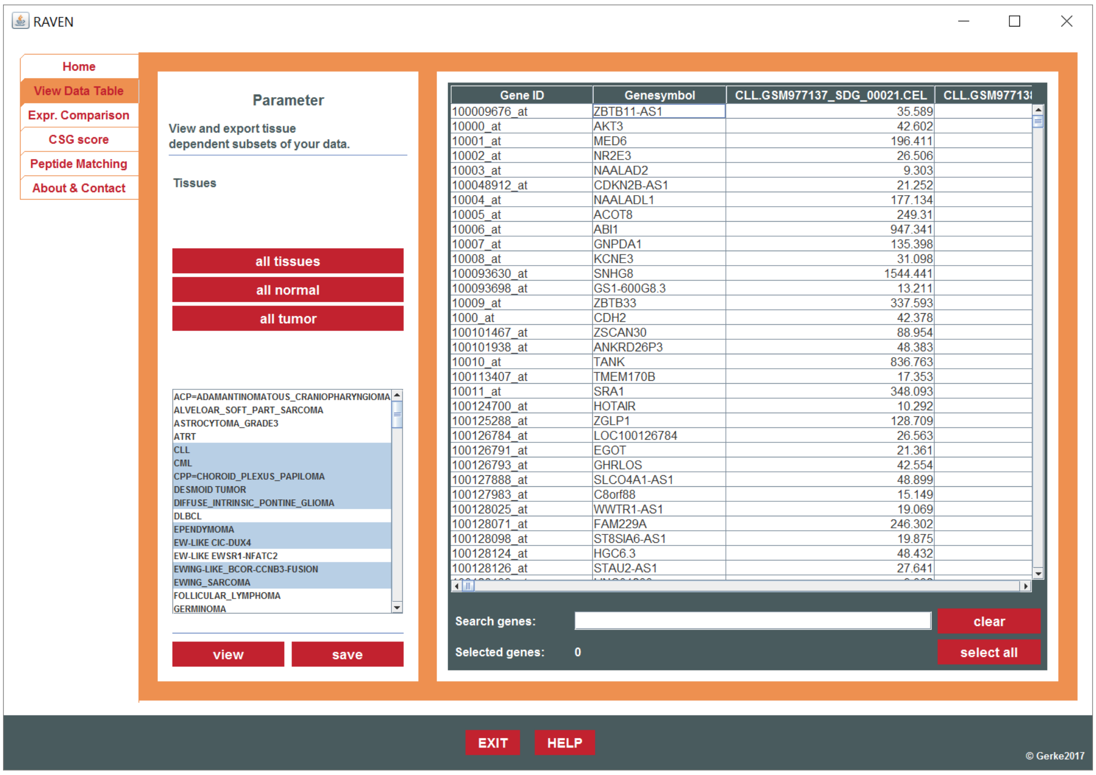


Figure 4.1: blubbl

4.1 Extracting Subsets

Via the 'View Data Table' tab you can display your uploaded data. If your data contains samples from different tissues or tumors you can display a subset of your whole data set. Select all tissue types you want to extract and view the corresponding table. You can also focus on specific genes by searching and selecting them. Save your table via the 'save' button and choose how your data table should look like.

Beside a classical table, you can also save your subset in RAVEN style which enables you to upload your extracted subset into RAVEN saving you the time to manually annotate your table to the standards of RAVEN input files (described in Section 2.4). See in Section 5.2 for more information on saving your data and results.



Parameter

View and export tissue dependent subsets of your data.

Tissues

ACP=ADAMANTINOMATOUS_CRANIOPHARYNGIOMA
ALVELOAR_SOFT_PART_SARCOMA
ASTROCYTOMA_GRADE3
ATRT
CLL
CML
CPP=CHOROID_PLEXUS_PAPILOMA
DESMOID TUMOR
DIFFUSE_INTRINSIC_PONTINE_GLIOMA
DLBCL
EPENDYMOMA
EW-LIKE CIC-DUX4
EW-LIKE EWSR1-NFATC2
EWING-LIKE_BCOR-CCNB3-FUSION
EWING_SARCOMA
FOLLICULAR_LYMPHOMA
GERMINOMA

Gene ID	Genesymbol	CLL.GSM977137_SDG_00021.CEL	CLL.GSM97713
100009676_at	ZBTB11-AS1	35.589	
10000_at	AKT3	42.602	
10001_at	MED6	196.411	
10002_at	NR2E3	26.506	
10003_at	NAALAD2	9.303	
100048912_at	CDKN2B-AS1	21.252	
10004_at	NAALADL1	177.134	
10005_at	ACOT8	249.31	
10006_at	ABI1	947.341	
10007_at	GNPDA1	135.398	
10008_at	KCNE3	31.098	
100093630_at	SNHG8	1544.441	
100093698_at	GS1-600G8.3	13.211	
10009_at	ZBTB33	337.593	
1000_at	CDH2	42.378	
100101467_at	ZSCAN30	88.954	
100101938_at	ANKRD26P3	48.383	
10010_at	TANK	836.763	
100113407_at	TMEM170B	17.353	
10011_at	SRA1	348.093	
100124700_at	HOTAIR	10.292	
100125288_at	ZGLP1	128.709	
100126784_at	LOC100126784	26.563	
100126791_at	EGOT	21.361	
100126793_at	GHRLOS	42.554	
100127888_at	SLCO4A1-AS1	48.899	
100127983_at	C8orf88	15.149	
100128025_at	WWTR1-AS1	19.069	
100128071_at	FAM229A	246.302	
100128098_at	ST8SIA6-AS1	19.875	
100128124_at	HGC6.3	48.432	
100128126_at	STAU2-AS1	27.641	

Search genes:

Selected genes: 0

Figure 4.2: Raven Home: set the working directory via the 'set' button (disabled here because of missing/unvalid directory in the textfield)

4.2 Direct Gene Expression Comparison

The direct gene expression comparison approach enables the direct visualization of gene expression data via box plots. You can decide to analyze differences **between multiple genes** in the same tissue or cancer as well as differences in expression values of the same gene in **between several tissue or/and cancer types**. The bars of the plot are sorted descending by their median. You can spice up your box plot by marking the mean or highlighting your favorite gene, tissue or cancer type in red.

Tip: Especially for box plots with a high number of bars it is recommendable to accentuate the most important bar in the graph.

Based on the data resulting to your shown box plot, RAVEN summarizes each bar of your plot in numbers below the graph.
Beside mean, median, inter quartile range (IQR), normal range, minimum and maximum value of each bar you are also notified of possible outlier.

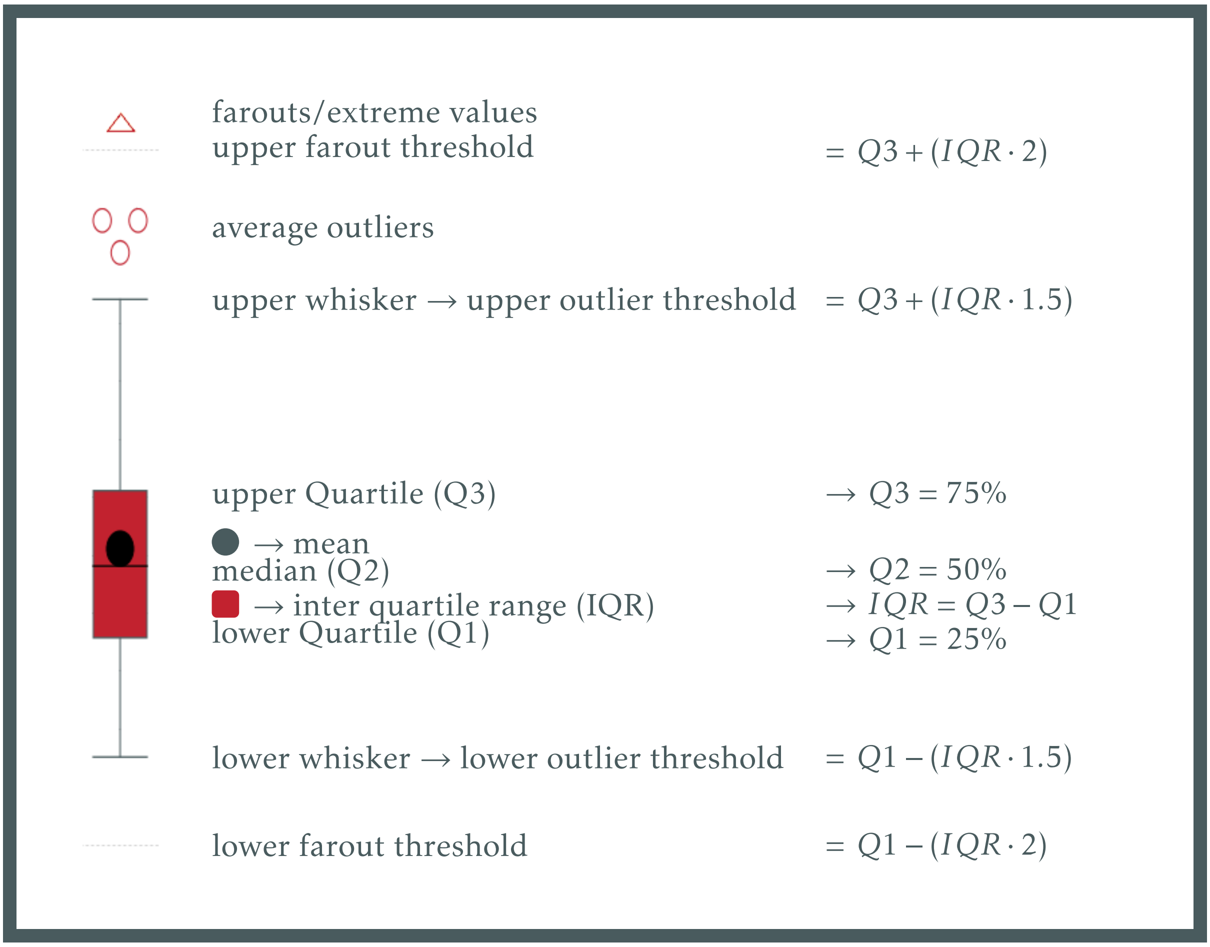
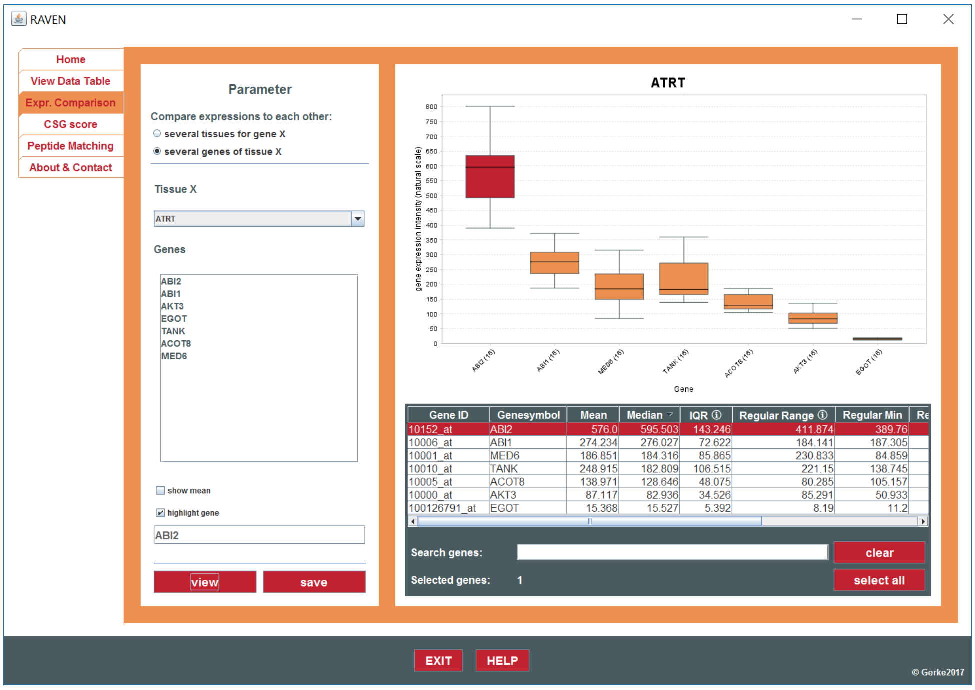


Figure 4.3: Schematic description of the boxplot bar visualized by RAVEN on an artificial example.

4.3 Cancer Specific Gene (CSG) Score

4.4 Peptide Matching Pipeline (PMP)

The Peptide Matching Pipeline (PMP) combines several bioinformatic webservers to predict T-cell binding MHC-I peptides and calculate their corresponding binding affinity. Furthermore, it also identifies all proteins also binding to this peptide which indicates



Parameter

Compare expressions to each other:

○ several tissues for gene X

● several genes of tissue X

Tissue X: ATRT

Genes

ABI2
ABI1
AKT3
EGOT
TANK
ACOT8
MED6

□ show mean

☑ highlight gene

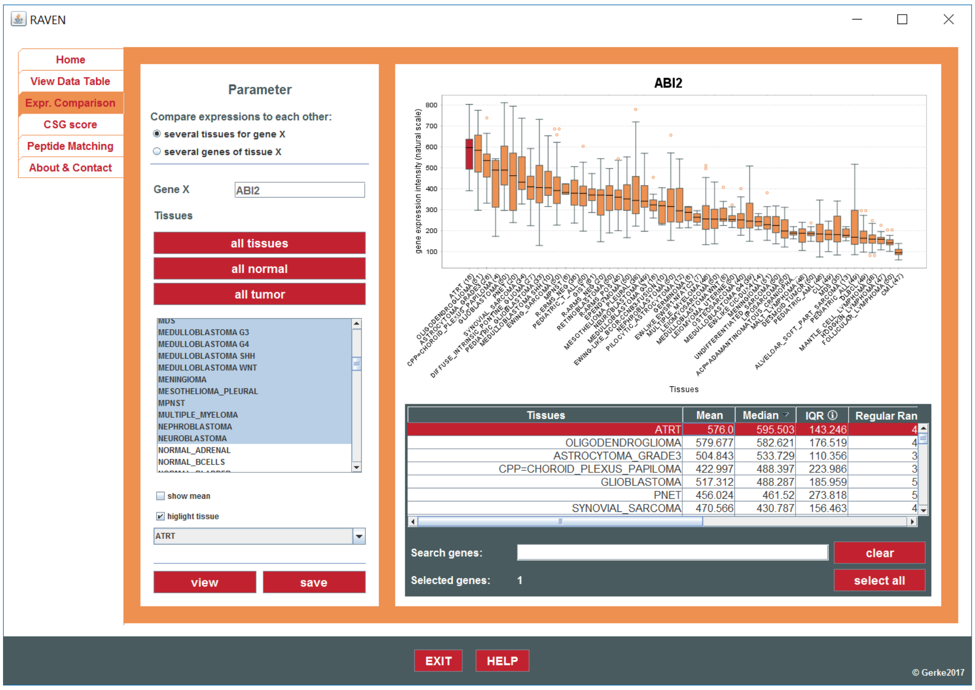
ABI2

Gene ID	Genesymbol	Mean	Median	IQR ⓘ	Regular Range ⓘ	Regular Min	Re
10152_at	ABI2	576.0	595.503	143.246	411.874	389.76	
10006_at	ABI1	274.234	276.027	72.622	184.141	187.305	
10001_at	MED6	186.851	184.316	85.865	230.833	84.859	
10010_at	TANK	248.915	182.809	106.515	221.15	138.745	
10005_at	ACOT8	138.971	128.646	48.075	80.285	105.157	
10000_at	AKT3	87.117	82.936	34.526	85.291	50.933	
100126791_at	EGOT	15.368	15.527	5.392	8.19	11.2	

Search genes:

Selected genes: 1

(a) Comparing expression of multiple genes with focus on ABI2 in a specific cancer (ATRT).



Parameter

Compare expressions to each other:

● several tissues for gene X

○ several genes of tissue X

Gene X: ABI2

Tissues

MDS
MEDULLOBLASTOMA G3
MEDULLOBLASTOMA G4
MEDULLOBLASTOMA SHH
MEDULLOBLASTOMA WNT
MENINGIOMA
MESOTHELIOMA_PLEURAL
MPNST
MULTIPLE_MYELOMA
NEPHROBLASTOMA
NEUROBLASTOMA
NORMAL_ADRENAL
NORMAL_BCELLS

□ show mean

☑ higlight tissue

ATRT

Tissues	Mean	Median	IQR ⓘ	Regular Ran
ATRT	576.0	595.503	143.246	4
OLIGODENDROGLIOMA	579.677	582.621	176.519	4
ASTROCYTOMA_GRADE3	504.843	533.729	110.356	3
CPP=CHOROID_PLEXUS_PAPILOMA	422.997	488.397	223.986	3
GLIOBLASTOMA	517.312	488.287	185.959	5
PNET	456.024	461.52	273.818	5
SYNOVIAL_SARCOMA	470.566	430.787	156.463	4

Search genes:

Selected genes: 1

(b) Comparing expression of the same gene (ABI2) in different cancer types with focus on ATRT

Figure 4.4: RAVEN results of the direct gene expression comparison. Example of analyzing differences in expression values with focus on ABI2 (gene) and ATRT (cancer) compared to other genes, cancer and normal tissues.

the specificity of the peptide. A peptide specifically only binding to single protein can indicate a potential target for immunotherapy. By automatizing this time-consuming and tedious process and therfore minimalizing slips, the PMP of RAVEN supports your research on targeted immunotherapy.

By using both the CSG score (see Section 4.3) calculation followed by the PMP a first draft for research on targeted cancer specific immunotherapy can be modeled.

To use the peptide matching pipeline (PMP), a stable internet connection is required. It is also important that your system allows RAVEN to access the internet to send and receive queries. Otherwise the firewall will block your request and the pipeline will break in the first step. In this case talk to your IT-administrator.

Systematic workflow of the peptide matching pipeline (Figure 4.5)

What RAVEN does for you:

1. Map the symbol of your favourite gene to UniProt to recieve its uniprot accession id (AC)[1]
2. Get the protein sequence in fasta format for your gene from UniProt with the AC[1] [5]
3. Send Query with protein sequence, allele and peptide length to IEDB [6]
4. Receive T-cell binding MHC-I peptides predicted with artifical neural networks (ANN)[4][2]
5. Search each peptide against the UniProtKB database with Apaches Lucene[1] text search provided by Protein Information Resource (PIR) to find all proteins binding to these peptides [7][3]
6. Determine each peptides' specificity by its number of matching proteins

PMP parameter setting

What you have to do:

1. **Enter at least one query gene.** Here, RAVEN only accepts genesymbols (e.g. `MYBL2`, `PTPN18`, `ABI2`). Otherwise, you will receive an error message.
2. **Select your species.** Choose between human and mouse.

[1] https://lucene.apache.org/

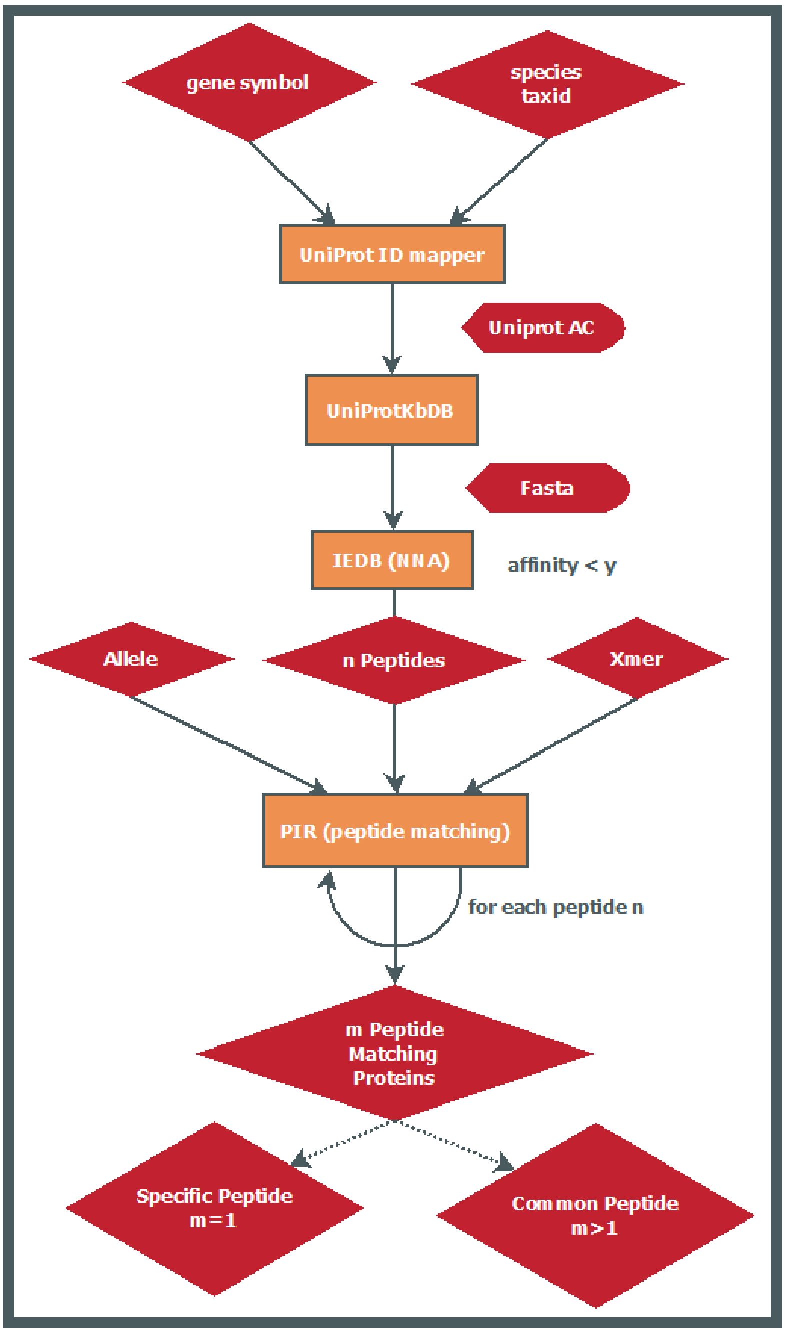


Figure 4.5: Systematic workflow of the peptide matching pipeline.

3. **Choose a database.** You can limit the proteins, which match the peptides, to Swiss-Prot entries or use the whole UniprotKB (Swiss-Prot + Trembl). SwissProt only contains manually annotated and reviewed proteins. Contrary, TrEMBL assenbles automatically annotated and unreviewd proteins. When using both, proteins may be found repeatedly under several different ids, which mostly refer to multiple unreviewed and obsolate sequence versions of an already validated one. To exclude such cases and avoid false positves, the ids have to be revised manually. Thus, we recommend to use Swiss-Prot only.

4. **Specify the peptides.** Chose your favourite allele from the most commen ones

and and confine the peptides length.

5. **Set an affinity cutoff** With this cutoff you limit your results to peptides with a stronger affinity than your set cutoff. The smaller the affinity score the stronger the MHC-I binding.
6. **Sort the results** Sorts the peptides in the result table by this column. While the affinity score indicates the strength of the MHC-binding, the specificity score refers to the number of proteins binding to the peptide.
7. **Submit your query.** Via the 'view' button.

Tip: By sending too many query genes at once, RAVEN has to communicate multiple times with several web services. This susceptible procedure can fail after some time due to overloaded servers or connection issues resulting into a breakdown of the PMP and thereby a loss of already received and processed data.
To avoid this, better split your query into several batches of **not more than 100 genes** to reduce such vulnerability.

PMP result table

What RAVEN tells you:

- **Peptide** → sequences of predicted T-cell binding MHC-I peptides of your query gene(s) (result from the ANN)
- **Query Gene** → gene encoding the protein which binds to the peptide
- **Uniprot ID** → UniProt accession ID of the protein encoded by the query gene.
- **Affinity (nM)** → Binding affinity in nM between the T-cell binding MHC-I peptides and protein (result from the ANN)
- **Specificity Score** → number of proteins binding to this specific peptide
- **Peptide specific proteins** → list of proteins which bind this specific peptide. For convenience to interpret the results we list the proteins by their symbols.

5 Save Data & Results

5.1 Images & Plots

You can save plots and images directly by clicking on it. With a right click of your mouse there will open a small menu which offers you different action such as print or zoom.

Do Right click on the plot 〉 Save as 〉 PNG...

Alterating Plots

You do not like the plot and want to change it? Alter labeling, font, colors etc. via the plot property menu. The color of points of boxes inside the plot cannot be changed.

Do Right click on the plot 〉 Properties

You are still not happy and want to alter more? Save the data on which the plot is drawn and import it in another statistical visualization software such as Prism. Look at the next Section to see how you should save your data to easly import them in other applications.

5.2 Results & Data

Save your results or the used data by clicking on the ′`save`′ button. This is only possible after you clicked on the ′`view`′ button to run and produce the results. Before this the ′`save`′ button is disabled.
The saving menu shown in Figure will pop up for you to define which data should be saved how.

Saving options:

Data to save choose whether saving the result table or the data producing it.

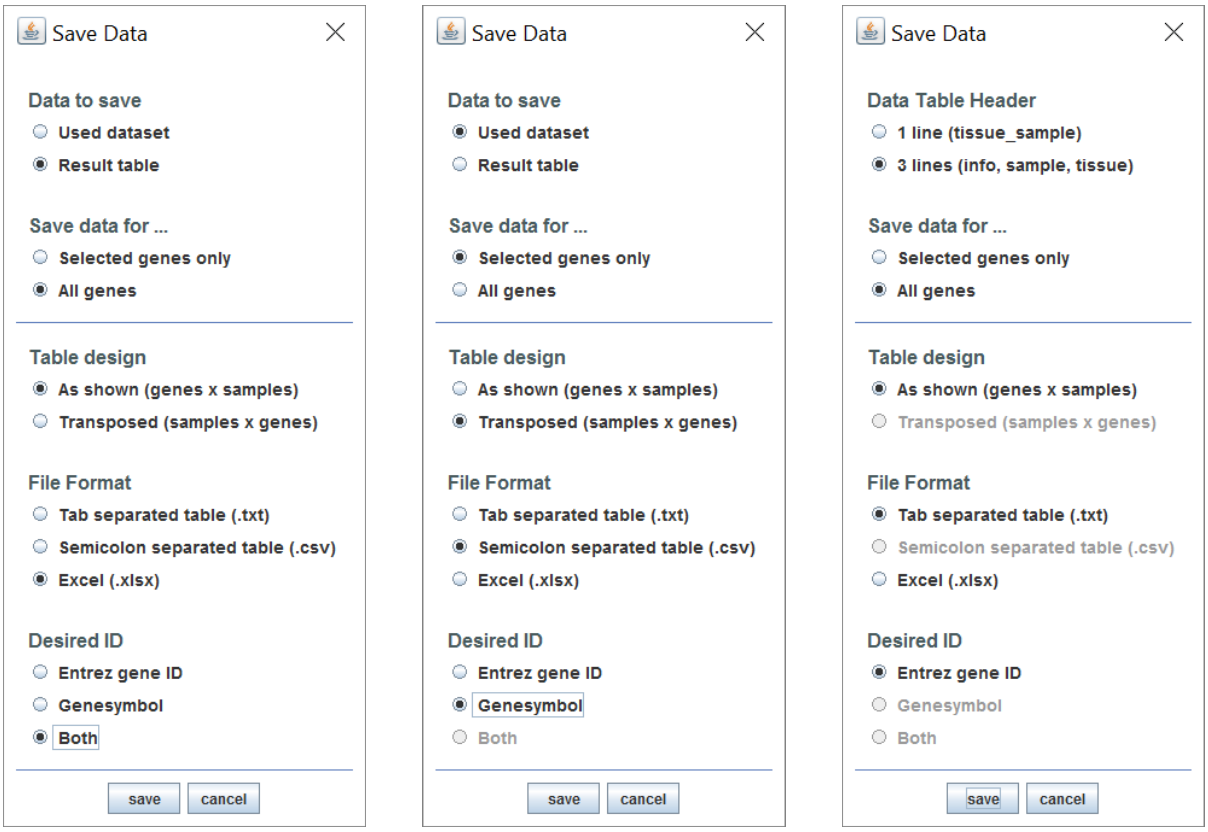
Save data for... do you want to save the whole table or only those you selected!

Table design The first option will save the table as displayed on the screen (or: rows → genes, columns → samples; as in affymetrix data). Choosing the transposed table design, the shown table will be mirrored by its diagonal (or: rows → samples, columns → genes) as favored by GraphPad Prism.

File Format RAVEN offers 3 different formats to save your file. When saving your table to a text file (`.txt`) the columns will be tab separated and can be read by any text editor (e.g. Notepad++) but also in Excel. Saving your table directly as Excel file (`.xlsx`) is also a possibility. Some programs such as GraphPad Prism favor a comma separated text file (`.csv`).

Desired ID When your table corresponds to genes you can choose between entrez gene ID or gene symbol as identifier. The first is easier to processes with other programs the later might be more intuitive for your interpretation. For result tables it is convenient to include both ids.

Table Header This option is only available for the subset creating method. You can export a classical one line header table, where the column name includes both tissue and sample id (e.g. `tissue_sample`). Instead also a header of 3 lines is possible splitting this information into 3 cells per sample (info → tissue → sample id) as required for RAVEN input files.



Save Data

Data to save
○ Used dataset
● Result table

Save data for ...
○ Selected genes only
● All genes

Table design
● As shown (genes x samples)
○ Transposed (samples x genes)

File Format
○ Tab separated table (.txt)
○ Semicolon separated table (.csv)
● Excel (.xlsx)

Desired ID
○ Entrez gene ID
○ Genesymbol
● Both

(a) How to save whole **result table**.

Save Data

Data to save
● Used dataset
○ Result table

Save data for ...
● Selected genes only
○ All genes

Table design
○ As shown (genes x samples)
● Transposed (samples x genes)

File Format
○ Tab separated table (.txt)
● Semicolon separated table (.csv)
○ Excel (.xlsx)

Desired ID
○ Entrez gene ID
● Genesymbol
○ Both

(b) How to save data to import and use with **GraphPad Prism**.

Save Data

Data Table Header
○ 1 line (tissue_sample)
● 3 lines (info, sample, tissue)

Save data for ...
○ Selected genes only
● All genes

Table design
● As shown (genes x samples)
○ Transposed (samples x genes)

File Format
● Tab separated table (.txt)
○ Semicolon separated table (.csv)
○ Excel (.xlsx)

Desired ID
● Entrez gene ID
○ Genesymbol
○ Both

(c) How to save data to upload in **RAVEN**.

Figure 5.1: 'Save Data' windows to choose your saving options for data files and result tables.

Save data to use with Prism

If you want to redo the plots or perform other statistical methods with GraphPad Prism on selected genes you identified with RAVEN methods, you can save the original data in data format to easily import it into Prism. Save your file with the options shown in Figure 5.1b.

How to import into Prism:

open Prism » start new project » choose your 'table & graph' type » select 'empty data table' » create

File » Import... » choose your file (.csv)

Save data to use with RAVEN

Saving a data set which can be uploaded in RAVEN can only be done with the subset creating method in 'View Data Table' (see Section 4.1). Create a subset from your current data set and click the 'save' button. The table options differ from those for the other methods. Save your file with the options shown in Figure 5.1c.

6 About & Contact

6.1 Citation

When using our software, please cite our paper:

Systematic identification of cancer-specific immunogenic peptides with RAVEN

Baldauf and Gerke et al.

Journal ...

PubMed: xxxxxlink

6.2 Authors

Development, Implementation & Design	Julia Sophia Gerke, MSc
Algorithm	Michaela Baldauf
Idea & Supervision	Thomas G.P. Gruenewald, MD, PhD

Visit us...

... online on 👍

www.lmu.de/sarkombiologie

... or in person 🏛

AG Gruenewald
Max-Eder Research Group for Pediatric Sarcoma Biology
Institute of Pathology
Ludwig-Maximilian-University Munich
Thalkirchner Str. 36
80337 Munich, Germany

✉ thomas.gruenewald@med.uni-muenchen.de
☎ ++49-89-2180-73716

6.3 Questions & Error Reporting

If you have problems with RAVEN or unsolved questions, that could not be answered with this manual, feel free to write us. For handling or technical issues please have a look at Section 6.4 first and contact our developer (julia.gerke@med.uni-muenchen.de). With any other questions Dr. Gruenewald (thomas.gruenewald@med.uni-muenchen.de) will be happy to help you.

Error Reporting

When reporting technical issues or problems with the handling please send us the following information regarding your problem if possible. This will help us to solve your problem much faster. Thank you in advance.

- Which operating system do you have? How much RAM do you have? See Table 1.1 or Figure 1.1 (only if you have trouble with the installation or data upload process!)
- What did you do?
- What did you want to achieve with it?
- Did you get an error message? Which one?
- If you started the application via command line you might see an error message written on the command line. If this is the case, please copy it into your email.

6.4 FAQ

I have RNAseq data instead of microarray. Can I analyze it with RAVEN anyway?

RAVEN is optimized to analyze micro array gene expression data. However, you can also run the software with your own RNAseq data, as long as you prepare your data accordingly. Make a quality assessment on your data and preprocess it correctly. Afterwards, normalize your reads and adapt the input file specifications necessary to upload data to RAVEN.

Does RAVEN only accept human gene expression data?

RAVEN is optimzed to work with gene expression data from human tissues and cancers. You can also upload data from other species but have to abandon gene symbols. Gene symbols are only available for humans. As the PMP operates independent from your data set you can use gene saymbols from both human and mouse.

Why are some notes or pop-ups not in English?

Some implemented notes or pop-ups automatically fit to your computer systems language. Change it to English to have a uniform language in RAVEN.

Why did the peptide matching pipeline break early?

The pipeline broke when the red progress bar is not moving anymore.

After sending a query to a web server, RAVEN waits for its response. Sometimes the server needs too long to respond and cancels the job itself after some time. Without the jobs result the pipeline cannot continue and terminates early. In this case try it again later or the even the next day.

Why is the PMP not making progress anymore?

The time RAVEN needs to process a gene via the PMP can vary a lot. Depending on the amount of predicted peptides per gene the time can vary from several seconds to half an hour and more. In this case the red progress bar is still moving but not increasing the displayed percentage. By decrease the affinity cutoff you can reduce the amount of peptides.

However, if your cutoff is less than 150 and the displayed percentage did not increase during the last hour the peptide matching process of the server got stuck. Quit it and try the PMP again. After a restart it should work again.

Why does the program not response anymore?

If you do not have the minimum requirement to run RAVEN, the program most likely will get stuck during the data upload phase. Decrease the size of your input file or increase your RAM.

Why is the entered gene id, name or symbol not accepted?

First, check if you used the allowed ids which are entrez gene id or gene symbol only. Control your spelling. Otherwise, this can occur due to a less common synonym or alias of your gene symbol. You can verify your gene symbol via GeneCards[1]. If all this does not apply to, the gene might be missing from your data set.

Why is my gene missing from the peptide table in the PMP?

A missing gene is either excluded from the beginning (see below) or is dropped during the pipeline process depending on your parameter. If all detected peptides have a higher affinity than your set affinity cutoff, the corresponding gene will be directly excluded due to its missing results. Increase the affinity cutoff to

Why did RAVEN exclude some genes from my query list in the peptide matching pipeline (PMP)?

This can have two different reasons:

- **non-coding gene:** The PMP only works for protein-coding genes. If your query is gene symbol a RNA gene or pseudo gene. The pipeline will not find the protein fasta sequence. Therefore, we excluded them from the query. In this case however, you will get a notification at the end of your processed job, informing you which "genes" were excluded from your query genes in advance.
- **filtering** Missing protein-coding genes were filtered out by your set paramenters. With the affinity cutoff you limit the amount of resulting peptides which should be checked for their specificity. If your cutoff is set too low, no peptide with such a strong binding affinity can be found and thus the gene will not show up in your result table.

Why is the button not working?

This can have different reasons:

- **the button is disabled:** If RAVEN disables a button (→ grey writing) you are not allowed to perform this action at the moment. For example, the 'save' button is disabled if there is no data (table)there which can be saved. Run the method first, before saving its results or used data.

[1] http://www.genecards.org/

- **several methods are running at the same time:** You are running to many jobs in RAVEN at the same time. Wait until RAVEN is done before starting a new job or method. To avoid such an overload run one RAVEN method after another.
- **the previous query was not finished yet:** The current job with your query is still running, RAVEN will not allow you to start a new one until the previous is finished. Wait until RAVEN is finished before sending the next query. You can cancel a running job by closing RAVEN.

Why do I get the following error messages when using the PMP?

- WebServiceConnectionException:
- TimeOutException:

web Service Connection

TimeOut Exception

Why is the header/row name, containing the sample ids, missing when saving the data of Expression Comparison

Each expression value belongs to another sample. Values in the same row/column do not belong to the same sample. Contrary, the samples in the same column/row represent a group. So, the sample ids are not necessary for this group visualization (box plot) and are rather misleading information.

Why did the table not update after changing the parameter?

After changing the parameters, you always have to press the `'view'` button to display the updated table or plot.

RAVEN is not responding anymore! What can I do?

Currently RAVEN seems to be too busy to accept new intake from you. Do not frantically click around and confuse RAVEN even more. Give it some time, it probably will be back in a few minutes. After 15 minutes without response, you can abort and restart RAVEN. In this case it is recommendable to use a computer with more power (see Section 1.1).

6.5 License

Bibliography

[1] UniProt: a hub for protein information. Nucleic Acids Research, 43(D1):D204–D212, oct 2014.

[2] Massimo Andreatta and Morten Nielsen. Gapped sequence alignment using artificial neural networks: application to the MHC class i system. Bioinformatics, 32(4):511–517, oct 2015.

[3] C. Chen, Z. Li, H. Huang, B. E. Suzek, and C. H. Wu and. A fast peptide match service for UniProt knowledgebase. Bioinformatics, 29(21):2808–2809, aug 2013.

[4] Morten Nielsen, Claus Lundegaard, Peder Worning, Sanne Lise Lauemøller, Kasper Lamberth, Søren Buus, Søren Brunak, and Ole Lund. Reliable prediction of t-cell epitopes using neural networks with novel sequence representations. Protein Science, 12(5):1007–1017, may 2003.

[5] S. Patient, D. Wieser, M. Kleen, E. Kretschmann, M. Jesus Martin, and R. Apweiler. UniProtJAPI: a remote API for accessing UniProt data. Bioinformatics, 24(10):1321–1322, apr 2008.

[6] R. Vita, J. A. Overton, J. A. Greenbaum, J. Ponomarenko, J. D. Clark, J. R. Cantrell, D. K. Wheeler, J. L. Gabbard, D. Hix, A. Sette, and B. Peters. The immune epitope database (IEDB) 3.0. Nucleic Acids Research, 43(D1):D405–D412, oct 2014.

[7] C. H. Wu. The protein information resource. Nucleic Acids Research, 31(1):345–347, jan 2003.