R2 Tutorials

Release 3.3.1

The R2 support team

Apr 05, 2019
# Contents

## 1 Preface

## 2 Using Datasets

2.1 Scope ........................................... 5
2.2 Step 1: Selecting a dataset ................................ 5
2.3 Step 2: Advanced selection of datasets ......................... 7
2.4 Step 3: Data Scopes .................................. 9
2.5 Final remarks / future directions ............................... 9

## 3 One Gene View

3.1 Scope ........................................... 11
3.2 Step 1: Selecting a gene .................................. 11
3.3 Step 2: Probesets for a gene ................................. 12
3.4 Step 3: Plotting Gene expression ............................. 13
3.5 Step 4: Probeset verification ................................. 16
3.6 Step 5: Sources for additional information on the selected gene ................. 17
3.7 Step 6: Adapting plot ................................... 18
3.8 Step 7: View a gene in groups ............................... 21
3.9 Step 8: Selecting subsets ................................. 22
3.10 Step 9: Find best track separation with CliniSnitch ................... 23
3.11 Final remarks / future directions ............................. 24

## 4 Multiple Genes View

4.1 Scope ........................................... 25
4.2 Step 1: Viewing multiple genes ............................... 25
4.3 Step 2: Viewing multiple genes through track annotation ............... 26
4.4 Final remarks / future directions ............................. 28

## 5 Annotation analyses

5.1 Scope ........................................... 29
5.2 Step 1: Relating 2 (categorical) tracks ......................... 29
5.3 Step 2: Relating 2 (numerical) tracks ........................ 31
5.4 Step 3: Relating a categorical track to a numerical track ............... 32
5.5 Step 4: Annotation plotter ................................. 33
5.6 Final remarks / future directions ............................. 34

## 6 Differential expression of genes in your dataset

6.1 Scope ........................................... 35
6.2 Step 1: Selecting data and gene ............................... 35
6.3 Step 2: Choose annotation track as grouping variable ............... 36
6.4 Step 3: Anova results .................................. 37
6.5 Step 4: Adapting grouped plots .................................................. 40
6.6 Step 5: Finding all differentially expressed genes .................................. 42
6.7 Step 6: Setting parameters ........................................................... 43
6.8 Step 7: Inspecting single gene ......................................................... 47
6.9 Step 8: Plot all genes and adapt visualization: Volcano plot etc .................. 47
6.10 Final remarks / future directions .................................................... 52

7 Find genes correlating with your gene of interest ................................. 53
7.1 Scope ...................................................................................... 53
7.2 Step 1: Selecting data ................................................................. 53
7.3 Step 2: Verifying settings ............................................................. 54
7.4 Step 3: Inspecting correlating genes ................................................. 55
7.5 Step 4: Inspecting correlation between specific genes .......................... 56
7.6 Step 5: Relation with Chromosome position ...................................... 64
7.7 Step 6: Establishing overrepresentation in other domains ....................... 66
7.8 Step 7: Gene list in pathway context ............................................... 67
7.9 Step 8: Further pathways analysis .................................................. 69
7.10 Step 9: Gene set analysis ............................................................ 71
7.11 Final remarks / future directions .................................................... 72

8 Working with Kaplan Meier .............................................................. 73
8.1 Scope ...................................................................................... 73
8.2 Step 1: Selecting the Kaplan Meier module ....................................... 73
8.3 Step 2: Kaplan Meier by gene expression; the Kaplan Scan .................... 75
8.4 Step 3: Kaplan scan for a group of genes ......................................... 77
8.5 Step 4: Kaplan scan on your own cohort ......................................... 78
8.6 Final remarks / future directions .................................................... 80

9 Pathway Finder ............................................................................ 81
9.1 Scope ...................................................................................... 81
9.2 Step 1: Selecting data ................................................................ 81
9.3 Step 2: Choose the right gene ...................................................... 82
9.4 Step 3: Correlating pathways with a gene ....................................... 83
9.5 Step 4: Finding pathways relevant in groups of genes .......................... 85
9.6 Step 5: Relating to data annotation ................................................. 86
9.7 Step 6: Determining differentially expressed pathways ....................... 86
9.8 Step 7: Verifying a pathway .......................................................... 87
9.9 Step 8: Correlating with the expression of a gene ............................... 87
9.10 Final remarks / future directions ................................................... 88

10 Multiple datasets overview with Megasampler ................................... 89
10.1 Scope ...................................................................................... 89
10.2 Step 1: Selecting multiple datasets ................................................. 89
10.3 Step 2: Viewing a gene in multiple datasets ..................................... 91
10.4 Step 3: Expression distribution over many datasets .......................... 95
10.5 Step 4: Megasearch ................................................................ 96
10.6 Final remarks / future directions ................................................... 99

11 K-means clustering in R2 .................................................................. 101
11.1 Scope ...................................................................................... 101
11.2 Step 1: Selecting data and module ................................................. 101
11.3 Step 2: Adapting settings ............................................................. 102
11.4 Step 3: Examining resulting clusters .............................................. 104
11.5 Step 4: Creating consistent clusters .............................................. 106
11.6 Final remarks / future directions ................................................... 107

12 Using signatures ............................................................................ 109
12.1 Scope ...................................................................................... 109
12.2 Step 1: Creating a geneset signature; a category within R2 .................. 110
These are the tutorials for R2, a biologist friendly, web based genomics analysis and visualization application developed by Jan Koster at the department of Oncogenomics in the Academic Medical Center (AMC) Amsterdam, the Netherlands. For citations, please include the following website: ‘R2: Genomics Analysis and Visualization Platform (http://r2.amc.nl http://r2platform.com)”

Copyright (c) 2006-2018 Jan Koster
Preface

Why these tutorials?

The R2 platform (http://r2platform.com or http://r2.amc.nl) is a genomics analysis and visualization platform that tries to provide a biologist friendly interface to high throughput data. It has been developed within the department of Oncogenomics at the AMC in the Netherlands, where it still serves as a primary entry point for all types of high throughput data generated within the department. The R2 platform consists of 2 parts; a (publicly accessible) database, that stores the data, coupled to a web-interface that provides a set of tools and visualizations to mine the database.

Even though many people like the concept of R2, getting started with the platform as a new user can be a bit difficult or intimidating. With this tutorial book, we hope to get new users started and at the same time demonstrate the diverse set of functionalities to users which may already have experience with R2, but want to get familiar with new additions to the platform.

The setup of these tutorials is as follows: We have divided the different aspects of the R2 platform in a number of chapters which reflect tasks that are often performed within R2. In each chapter, step by step instructions guide you through an analysis, which you can perform online within R2 yourself. During these steps, also features related to the respective chapter, such as additional analyses or visualizations will be introduced, thereby conveying the ease of using the interconnected R2 interface.

The steps will be interspersed with additional information in Did you know boxes:

Chapters 2-10 demonstrate a great number of core functionalities, which you will encounter often and from different angles when working within the platform. The next set of chapters (11-18) dive more into specialized functions and in chapters 19-21 more advanced integrated analyses are shown, that some of you may be interested in. Chapters 22-24 demonstrates how users can adapt R2 to their needs by explaining how you can generate your own grouping variables or store personal lists of genes. These also explain how you can start your own user group and share information with specific other users for collaborative work. It is further elaborated how to add your own data and how to export data from R2 for use in other tools.

We hope that these tutorials will be helpful. If you have any comments or suggestions you’re welcome to contact us through the R2 website.
2.1 Scope

- Working with datasets.
- R2 allows you to perform all kinds of analyses based on a well annotated single dataset or a selection of datasets at the same time. Different analyses are available based on the selection of one of these options in field 1.
- R2 contains mRNA gene expression profiles for more than 70,000 individual human samples. The samples are grouped in so called datasets. Each dataset has its own characteristics such tissue type, tumor type or from cell-line experiments.
- The Tumor Neuroblastoma public - Versteeg - 88 - MAS5.0 - u133p2 dataset will be used as an example dataset to guide you through most of the tutorial. Later on working with multiple datasets will be discussed.

2.2 Step 1: Selecting a dataset

1. In R2 a large amount of datasets are available for analysis and visualization. The numbered items in the main window will guide you through all the steps necessary to perform a task. In field 1 select “single dataset”, in field 2 choose “change”
Figure 1: In the main screen: Change dataset
2. A pull-down menu appears containing a large collection of datasets available for all types of analyses R2 is offering
3. Click on the desired dataset.

Figure 2: Selecting datasets from the pull down menu on the main screen
Did you know that datasets have an informative naming?

Datasets have a structured naming in R2, using the following rules: type_of_dataset - author_number_of_samples - normalization - chiotype. Datasets are listed alphabetically.

2.3 Step 2: Advanced selection of datasets

1. Next to the pull down menu you can also choose for the “advanced dataset selection” tool. The advanced dataset selection facilitates searching through datasets using keywords and other filter options such as the minimal size of a dataset, the date a certain dataset was published etc. An example search would be finding all colon samples which are part of a mixed dataset consisting of normal tissue and tumor samples.

![Advanced selection of datasets](image)

2. Click on the “Advanced” link. A new screen shows a table where the headers can be filled with search entries to fine tune your search for a dataset meeting your search criteria. Enter ‘Neuro’ in the class column and ‘50’ in the ‘#’ column and select ‘greater than’ from the pull down menu. This returns all the datasets containing the search term ‘Neuro’ and having more than 50 samples.

![Advanced selection panel](image)

3. Clicking on ‘Neuroblastoma’ in the class 3 column containing 88 samples reveals a detailed info box containing additional dataset information from the R2 database. When the dataset is publicly available clicking on the GEO ID link redirects to the GEO repository database where RAW data files are available. A Pubmed link is listed in case the dataset is linked to a publication listed in PubMed. Note: Clicking on an exclamation mark also shows detailed dataset information.

![Advanced selection panel](image)
4. Select “Across Datasets” in field 1. Note that in field 2 different options become available compared to the “single dataset” option.

Analysis methods following selecting the “Across Datasets” option in field 1 will be discussed in tutorial “Working with multiple datasets”.

Did you know that clicking on an exclamation balloon provides additional info?

Clicking on the GEO ID link redirects to the GEO repository database where RAW data files are available. A Pubmed link is listed in case the dataset is linked to a publication listed in PubMed.
2.4 Step 3: Data Scopes

1. R2 can also be forced to only display a sub selection of all the datasets that are available. These are called data scopes and can be selected from within R2 by the left hand menu items ‘change data scope’. From here you can use one of the preset scopes. This is also the place where you can remove a scope that has been set. One obvious reason why scopes can be handy, is the focussed view on the available data.

2. Data scopes can be used directly from the internet address line, which can be handy when a referral needs to be made to R2 from a manuscript. For now, you do need to provide a link directly to the server (usually hgserver1.amc.nl/cgi-bin/r2/main.cgi?&dscope=NRBL).

Did you know that the R2-support team is scanning public repositories for interesting datasets to expand the R2-database on a regular basis

In case you want to see a dataset added to R2 please send an email to r2-support@amc.nl Such an email should contain a link to the publicly accessible files, such as a Gene Expression Omnibus number (GSE****). Your own private datasets can also be added to R2 with user/group restricted access. Please send us an email at r2-support@amc.nl and inquire on the procedure to get your data available in R2 (see also chapter 22).

2.5 Final remarks / future directions

Everything described in this chapter can be performed in the R2: genomics analysis and visualization platform (http://r2platform.com / http://r2.amc.nl)

If you run into any quirks or annoyances don’t hesitate to contact r2 support (r2-support@amc.uva.nl).

We hope that this tutorial has been helpful. The R2 support team.
One Gene View

Analyze the expression levels of a single gene within a dataset

3.1 Scope

- Use R2 to investigate the expression levels of all samples from a specific dataset.
- In this example the expression levels of the MYCN gene will be used.
- Adjust several parameters in the advanced settings panel to get a better insight in the expressions levels or adapt your graphic layout.
- In R2, the samples are annotated with e.g clinical data, each group of annotated data is called a “Track” in R2. These tracks can be used to filter data in all types of analyses R2 is offering.
- A separated info panel in the one-gene expression level screen provides different types of analyses based on the expression level of the chosen gene.
- Most of the mRNA expression datasets are generated with Affymetrix profiling arrays. In general these arrays use more than one so called probeset to measure the expression level of one single gene. With a separated module “Transcript view”, the details of the probesets can be studied.

3.2 Step 1: Selecting a gene

1. Use “Single Dataset” in field 1 and make sure that the “Tumor Neuroblastoma public - Versteeg - 88 - MAS5.0 - u133p2” dataset is selected in field 2.
2. Choose “View a gene” in field 3.
3. Type MYCN and click ‘next’.
3.3 Step 2: Probesets for a gene

1. In many cases more than one probeset is reported for each gene. In this example, there are 5 probesets annotated for the MYCN gene. By default, the probeset with the highest average present signal (APS) is selected. This APS signal is simply the average of all samples that are considered to express a selected gene (have a present call). Occasionally other probesets assigned to the same gene could be of interest depending on the structure of the gene (for example a potential splice variant). Also realize that the most informative probeset is re-determined in every dataset, sometimes resulting in a different probeset as the choice of R2. The expression levels are by default converted to log2 values.

2. In the adjustable settings screen, use the pre-defined default settings and click ‘Next’.
Did you know that a reporter with an exclamation marks is an indication there may be something wrong with the reporter (e.g. bad design)?

Hovering over the exclamation mark will inform you on what may be the issue with a specific reporter (probeset). Reporters with an issue will not be used to represent a gene in searches where hugo symbol is used. Hovering over the name of a gene will display concise gene information, such as alternative names for the current gene. In case you are not searching with an official NCBI genesymbol, R2 will also search the alternative names to find your gene of interest.

3.4 Step 3: Plotting Gene expression

1. R2 generates a YY-graph Figure 3 from the MYCN expression levels of all samples with expression levels ordered from left (low) to right (high). Hovering over the dots reveals additional annotation that R2 has stored for the focused sample.

![Figure 3: YY plot MYCN expression](image)

Underneath the X-axis, colored boxes are depicted, representing clinical information of the samples in so called “tracks”. Again, hovering over them will reveal underlying data. For MYCN there is a clear relation between the expression levels and the tracks for “MYCN amplification” and “INSS-stage”. So these tracks underneath the image give a quick glance at some of the clinical parameters, defined for the dataset. It is also possible to define your own custom made tracks, or disable/adapt the settings for default tracks (further explained in “Adapting R2 to your needs”)

3.4. Step 3: Plotting Gene expression 13
3. Sometimes you get more insight by reviewing the expression levels with other transformations. Scroll down and transform the data (Figure 5), choosing “none”, in the ‘transformation’ pulldown menu and click adjust settings. In the “adjustable settings” panel, there are several other settings to adapt the graph R2 generates (like changing font sizes, or adding labels to the datapoints).

Did you know that the Adjustable Settings panel is available under most graphs in R2?

*Just scroll down the page to find the Adjustable Settings box with graph options. Don’t forget to press the “Adjust Settings” button at the bottom of the box in order for your changes to take effect!*

To highlight / mark specific samples in the graph you can enter the R2 sample ID’s in the field ‘samples to mark’ from the ‘adjustable settings’ box. Several marking options can be selected with the ‘Mark method’ that can be found in the ‘More settings’ tab (e.g. ‘epicenter’ and ‘arrow’).

![Figure 4: Adjusting the graph settings](image)

**Did you know that R2 allows you to emphasize samples with marker options?**

*R2 knows a couple of mark options, that you can make use of in the advanced prescriptions:*

- ‘dot’: places a thick border around the sample
- ‘circle’: Places a ring around the sample (diameter 9)
- ‘circle_2’: Places a ring around the sample (diameter 4)
- ‘circle_3’: Places a ring around the sample (diameter 1), effectively a thin border
- ‘epicenter’: Places a set of 3 rings descending in width around a sample
- ‘arrow’: Places a block arrow pointing to the sample
- ‘triangle’: Places a filled triangle under the sample

**Note:** The dotsize does not scale with ‘arrow’ and ‘triangle’ method.
Multiple colors and different markings can be applied to the individual samples you want to highlight. To bring attention to different samples, you may want to use multiple colors and types of marking. In order to do so define the method within the ‘samples to mark’ box by adding another ‘:’ after the color representation. Defining this will overrule the default setting, and thus also enable the use of different markings within the same figure. The skeleton for advanced usage is: ‘sample1,sample2:color1:method1;sample3,sample4:color2:method2’. For example: ““itcc0288:ff4444:dot;itcc0021:009999:arrow;itcc0013,itcc0132:00ff00:epicenter”” creates the markings as shown in the figure below.

Figure 5: Adjusting the sample mark layout

Did you know that converting expression levels using the “transform” option can help you to gain additional insight?

There are several data transformations available

- “none”: Raw untransformed expression values, as they are represented in the R2 database.
- “2log”: logarithmic values with base of 2. Every increment constitutes twice the amount.
- “rank”: Data transformation in which numerical or ordinal values are replaced by their rank when the data are sorted by expression. This transformation is useful for non-parametric statistical tests.
- “zscore”: 2log transformed data, centered around the average and expressed as the number of standard deviations from the average.
- “zscore_nonlog”: raw intensity values, centered around the average and expressed as the number of standard deviations from the average. This transformation is useful when the intensities in R2 are not raw, but for example logfolds as is often the case for aCGH data.
- “mad/mad2log”: Median absolute deviation (on raw values, or log2 transformed values).
- “center/log2center”: Expression values centered around 0 (on raw values, or log2 transformed values).
- “zscore_group”: Converts the expression levels from the zscore within a group (track). Applicable when e.g. technical variation in expression levels is expected. A possible reason could be when samples from the same dataset originate from different centers.

3.4. Step 3: Plotting Gene expression 15
3.5 Step 4: Probeset verification

Figure 6 lists for the various reporters of MYCN whether they are in agreement with the genome position of MYCN reference sequence (RefSeq). If all are stating “YES” then everything appears alright (from the perspective of an automated assessment). For the MYCN reporters “NO” indications indicate there may be an issue with it. Scroll down the page and click on the “Tview” link in the reporter table.

<table>
<thead>
<tr>
<th>ProbesetVerification (hg18)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symbol</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>MYCN</td>
</tr>
<tr>
<td>MYCN</td>
</tr>
<tr>
<td>MYCN</td>
</tr>
<tr>
<td>MYCN</td>
</tr>
<tr>
<td>MYCN</td>
</tr>
</tbody>
</table>

Figure 6: Probeset verification table

1. A new screen (or tab in the browser) appears with TranscriptView. The TranscriptView application depicts the alignment of expressed sequence tags (EST) and mRNA sequences to the human reference genome sequence (Fig 8). The strand orientation of these sequences are indicated by a color (green = positive strand, red = negative strand, blue = strand information is missing). The structure of the reference sequence has also been indicated. Furthermore, the browser shows the alignment of the sequences that were used to generate the reporters on the array (in the case of Affymetrix microarrays). This view can be used to inspect the quality of a reporter. Note, for instance, that the reporter “242026_at” is aligned to the genomic region of the MYCN reference sequence, but that its color is different from the rest (colored in red). In addition, in this particular case the reporter is located in the intronic (light shaded color) region which is another reason not to pick a certain probeset. Indeed, if we compare the gene expression values of this reporter, then its expression is 60 fold lower than R2’s standard pick (22 vs 1369). Below the ESTs the average gene expression of the individual probesets is illustrating that for this example the correct probeset is selected for analysis. NB: Currently probeset verification is only provided for various human Affymetrix array types.

Figure 7: Coloring represents type of transcript
Did you know that you can browse the gene expression values along the genome?

Once you have entered the genome browser with an attached dataset (like above), you can also navigate to / zoom out any other region in the genome. This allows you to look at the neighboring genes in a single go. It can be informative to separate the expression on the basis of a track. This can be achieved by selecting ‘dataset_track’ from the sample dropdown in the middle panel. Finally, within the genome browser, the contents for a panel on the left side can be hidden from a view by setting the height to 0.

3.6 Step 5: Sources for additional information on the selected gene

1. Close the Genome Browser tab or go back to the MYCN One Gene View expression screen.
In the left upper menu-panel several options are available to provide you with additional information sources of the MYCN gene and additional analyses. KaplanScan and Time Series analyses will be discussed in separate tutorials. GeneCards will redirect you to an overview on your gene of interest composed of many different resources. ProbePlus, will provide the sequences probed by the U133 Affymetrix platforms (Will not be shown in other platforms). Across datasets will generate an overview showing the average expression of the gene of interest within all datasets of the same platform/normalization scheme (provided that the normalization supports dataset additions).

When you click on the link (the name of the gene) under the PubSniffer header in the left menu, a new screen opens that lists the number of times your gene of interest is found within the NCBI Pubmed database in combination with dataset keywords. Clicking on one of the “Pubreminder” links redirects you to Pubmed Pub-reminder which is a tool for PubMed query building and literature mining.

**Did you know that Pub re-miner is a helpful tool for literature mining?**

In the large amounts of medical literature, finding information tailored to your needs and interest is becoming more and more complex. Using the right keywords is essential for effective searches, but which ones should you use? Pub re-miner is a web-based tool that allows simple text-based query building and information gathering (mining) of the NCBI literature search engine PubMed. Pub re-miner presents its results, gathered from abstracts, in frequency tables of journals, authors and words, which can be included / excluded in an iterative fashion. Next to building efficient queries, Pub re-miner can also be helpful in other areas: selecting a journal for your current work (by scanning the most often used journals of similar research) Finding experts in a research area (by viewing the authors associated with your query) Determine the research interest of an author (by viewing the keywords associated with an author)

### 3.7 Step 6: Adapting plot

1. To investigate the values R2 uses for graph generation click on “Datatable” to unfold a table with the expression levels for all samples.
2. The “track display selection” section can be opened by clicking on it. In here, you are able to toggle which tracks to display and/or hide within the YY-plots. Do note that these selections are non-persistent and will be forgotten as soon as you leave the xgeneview. Persistent, changes to the tracks can be made via the ‘my settings’ menu item, which is present in the main screen. Note that the adjustable settings panel including the customize track parameters are available throughout R2.

3. Other convenient options are revealed by clicking the “more settings” section. An extra panel unfolds which allows you to adapt your graph to meet for example the requirements of a journal. The appearance of this section will change depending on the kind of graph that you are selecting.

![DataTable](image)

Figure 10: Unfold the datatable

![Track Display Selection](image)

Figure 11: Tick and drag tracks

3.7. Step 6: Adapting plot
In Figure 12 sample annotation ("Annot Graph") and legend ("Draw Legend") were added. The "Annot Graph" option adds the information of a selected track to the Y-Y-plot. This can be helpful for the addition of Sample labels, or cell line names etc. Annotations can be shown in 3 ways; just below/on top of the expression value, as a series below the annotation tracks or at the values for those samples that have been marked. The size of the annotation scales with the setting of the dotsize. Check the More Settings panel for extra options, such as changing the color of the axis or showing link to and SVG output image. The Adjustable Settings menu is available in most of the R2 modules where a one-or two gene view is generated.
3.8 Step 7: View a gene in groups

1. Thus far, we have been looking at the expression of MYCN ordered by the expression. From the current location, we can also inspect the MYCN expression subdivided in groups. To achieve this, we simply scroll to the end of the page and locate the dropdown box in the ‘Group Separations’ section of the ‘Adjustable Settings’. Here we can select a track to separate the cohort accordingly. Select ‘inss’ as a track and press the ‘Adjust Settings’ button at the end of the page. R2 has now separated the patients on the basis of the inss staging track in alphabetical order.

Figure 15: View a gene in groups

1. The current representation is the most honest way of showing your data, as every single value is visible in the plot. We can also change the graphical representation of the data by selecting another graph type. Select ‘boxplot’ from the ‘graphtype’ dropdown and change ‘color by’ to ‘color by track’. Press the ‘Adjust’ button again to change the view. We now obtain a boxplot image where the respective groups have been colored according to the groups. Adaptations to other graph types can be made in a similar way.
Figure 16: View a gene in groups by boxplot

1. You can also sort the groups by their average or median gene expression.

Figure 17: Order groups by the average gene expression value

Did you know that once you separate a dataset in more than 2 groups, R2 will identify the most significant pair?

*If you view a gene in groups within the one-gene-view view page and the number of sub-groups are greater than 2, then R2 will automatically perform a brute-force t-testing to identify the combination of 2 groups that have the most significant difference. Just click on the ‘T-test on combinations’ link underneath the image and gain insight into all the tested combinations.

3.9 Step 8: Selecting subsets

To generate a graph of a subgroup of samples use the ‘Select a track’ pulldown from the ‘sample filter’ section to select a specific group.
In the neuroblastoma field it is well known that the mync expression is strongly correlated with stage 4. But maybe you are also interested in the mycn expression for the lower risk stages.

Go to the Adjustable Settings menu and select in the pull down the INSS stage. In the popup window select the lower risk stages st1, st2, st3 and st4s and click Next. Back in the Adjustable Settings panel click on the Adjust Settings button. These selections can be repeated a couple of times to build your ultimate selection.

![Image of Adjustable Settings menu](image1)

Figure 18: Selecting subgroups

In the picture below all stages are depicted and only the lower risk stages with Graph type BoxDotPlot.

![Image of all stages versus lower risk only](image2)

Figure 19: All stages versus lower risk only

You will encounter the filter option in the adjustable settings box in many modules.

### 3.10 Step 9: Find best track separation with CliniSnitch

1. We could wonder if our gene of interest associates even more with any annotation that is already available for the current dataset (like e.g. age group) than the example in the previous section. For such an analysis
R2 has the CliniSnitch function. Within this functionality a brute force T-test is performed on every possible combination of subgroups within every annotation track. We can run a CliniSnitch analysis directly from the one-gene-view page by clicking on the Gene name under CliniSnitch in the left panel. Click on the ‘MYCN’ gene.

**Figure 20: CliniSnitch result for MYCN**

1. For every track, the most significant test will be highlighted. Every test can also be visualized by clicking on the combination label. Not surprisingly, we can see that MYCN expression is best separated by the MYCN amplification track. If we look at the ‘inss’ track, we can also see that nearly every test containing ‘st4’ has a significant value. Click on ‘st2 vs st4’ to inspect this further.

2. We are now back in the ‘one gene view’, but our dataset has been restricted to only those patients that belong to either st2 or st4.

### 3.11 Final remarks / future directions

Some of these functionalities have been developed recently. If you run into any quirks or annoyances don’t hesitate to contact r2 support (r2-support@amc.uva.nl).

We hope that this tutorial has been helpful, the R2 support team.
**CHAPTER 4**

Multiple Genes View

*Analyze the expression levels of a group of genes within a dataset*

### 4.1 Scope

- Use R2 to investigate the expression levels of more than 1 gene within a specific dataset.
- In this example the expression levels of small groups of genes listed from several pathways will be used showing which genes are differentially expressed per subgroup.
- Adjust several parameters in the settings panel.
- In R2, the samples are annotated with e.g. clinical data. Each group of annotated data is called a “Track” in R2. These tracks can be used to split the group gene expression levels per track.

### 4.2 Step 1: Viewing multiple genes

2. Choose “View multiple genes “ in field 3 and Click Next
3. To illustrate the possibilities of the multiple gene view. Genes identified as classifiers for Medulloblastoma subtypes (Kool et al, Plos one) will be used. In the GENE/reporter textbox type or copy the following genes: AXIN2,BOC,dkk2,GABRA5,PTCH1,SMARCD3,WIF1 and click next.
4.3 Step 2: Viewing multiple genes through track annotation

1. In Figure 1 a selection of gene expression profiles are depicted in one picture in contrast to the one gene view. The multiple gene view enables the option to represent the gene expression separately for each track. In this manner potential relations between subgroups and gene expression can be visualized.

2. The dataset we are using is described in *PLoS One.* 2008 Aug 28;3(8), Kool M et al. Here the classification of 5 medulloblastoma subgroups are reported and annotated as such: A, B, C, D and E. To investigate the expression levels of a small group of genes per sub-category select in the adjustable settings box “subtype (cat)” at use track, “lump by group plot gene” at handle groups by and “Track” at color by track. Further set transform tot “none”, select “boxplot” at Plot type and click NEXT.
3. Most of the used genes are part of the WNT (subgroup A) and de SHH (subgroup B) signaling pathway over-expressed per subtype as shown by Kool et al. These genes are overexpressed in different Medulloblastoma molecular subtypes a,b,c,d and e are plotted together with the gene names.

4. Also try the “lump by gene plot group” which will produce an image where the genes are shown, separated by the subtypes.

5. The sample filter option allows the user to generate a multiple gene view per track.

4.3. Step 2: Viewing multiple genes through track annotation
4.4 Final remarks / future directions

Some of these functionalities have been developed recently. If you run into any quirks or annoyances don’t hesitate to contact r2 support (r2-support@amc.uva.nl).

We hope that this tutorial has been helpful, the R2 support team.
CHAPTER 5

Annotation analyses

Using (custom) annotation tracks as input for analyses

5.1 Scope

As you know by now, annotation of your data is stored in R2 as “tracks”. Within R2, one can easily create new annotation tracks. This can be done either based on results generated within analyses, or completely independent by uploading of tracks. In some cases it is of interest to start comparing one track with another. The type of statistics used to compare the tracks depends on the type of data; either categorical or numerical. One may wonder if there are significant overlaps between 2 tracks (with categorical variables), based on Fisher’s exact test. Alternatively, if there are multiple numerical tracks available; one may wonder if there is a significant correlation between 2 tracks. For these cases, R2 contains the Annotation modules; “relate 2 tracks” and “annotation plotter”.

- Relate 2 tracks (categorical), test significant overlap and view as honeycomb-plot
- Relate 2 tracks (numerical), assess significance of correlation and view as XY-plot
- Relate 2 tracks (Categorical vs numerical), assess differential values between groups
- Annotation plotter. Visualize tracks within sample cohort

5.2 Step 1: Relating 2 (categorical) tracks

1. Make sure that you are on the “main” page of R2, and that the selected dataset is “Tumor Neuroblastoma public - Versteeg - 88 - MAS5.0 - u133p2”. From the “type of analysis” dropdown select “relate 2 tracks”, which can be found in the annotation subsection and press next.
For the different tracks, make sure that you select a categorical one (which can be recognized by (cat)). Let’s investigate whether there is a relation between the neuroblastoma age-group (track=agegroup, flip point being 18 months at diagnosis) and the survival status (track=alive). Then press next to generate the result.

The generated result is now displayed on the screen. As we are testing 2 categorical variables, R2 has tested the relation between the 2 tracks and finds a highly significant Fisher’s exact p-value, indicating that there is a relation between the agegroup and vital status of the patients. The result is also shown in a honeycomb image, where every individual patient is represented as a separate circle, with the annotation as a hover box.

One can add more visual information to the plot, by coloring the patients on the basis of a track. From the adjustable settings at the bottom of the page, set the “colormode” to “color by track” and select the “inss_stage” as track. Press the adapt settings to create an updated image. Now we can clearly see that there is a great over-representation of stage 4 patients in the group of diseased patients who are older than 18 months. As you may appreciate, the combinations that you can make here are virtually endless. We have named this analysis the ‘Visual Fisher’s Exact test’, due to the visual additional insights that it provides over the ‘normal’ p-value that can be interpreted for this test.
5.3 Step 2: Relating 2 (numerical) tracks

1. Just as in the previous example, we select the “relate 2 tracks” option from the main R2 screen and press enter. Now this time, we select 2 numerical tracks, which can be recognized by the (#) sign at the end of a track. Within the Neuroblastoma dataset our options are limited for this example, so we select the “age in years” track vs the “nti_surv_overall” track and proceed to the next screen.

2. In the result page, R2 has detected that 2 numerical tracks were selected, so the correlation between the different tracks is being displayed and tested for statistical significance. Just as in the previous example, we could color the patients by a track if that would be appropriate.
5.4 Step 3: Relating a categorical track to a numerical track

1. The last example for relating 2 tracks, involves the combination of a numerical one to a categorical track. Essentially, this option allows you to test meta-gene data (such as combined expression values of multiple genes expressed as a single value) as well, where you could create a track containing only value information for the patients, and test this track to clinical parameters. We again select “Relate 2 tracks” from the main menu and navigate to the next page.

2. From the track options, we choose a categorical track for X (inss stage), and a numerical one for Y (nti_surv_overall) and navigate to the next page.

3. The result page will now start to look like view a gene in groups, only this time using the data contained in your track. Via the adjustable settings, you can change the representation to another plot type, such as a boxplot, change the color mode to color by track, and you have a nice result here, showing that the survival rate is significantly lower in patients of INSS stage 4.

As a recap for the last 3 tutorial steps, you have used the “relate 2 tracks” option from the annotation methods in R2 and represented different types of tracks with each other to gain new insights from combining 2 tracks. Below the 3-different representations are depicted side by side. Do remember, that this module allows you to use “meta data” tracks that you can assemble either within, but also outside of R2 via the uploading of a track option that will be shown in the “adapting r2 to your needs” chapter.
5.5 Step 4: Annotation plotter

1. In some publications, patient data is represented in slick looking annotation plots, showing the patient characteristics in rectangles. In a sense, these are just like the tracks that are represented underneath YY-plots in R2. To allow users to create these “track” figures, ordered in a user provided order, we have implemented the annotation plotter in R2.

2. Make sure that you are on the “main” page of R2, and that the selected dataset is “Tumor Neuroblastoma public - Versteeg - 88 - MAS5.0 - u133p2”. From the “type of analysis” dropdown select “Annotation plotter”, which can be found in the annotation subsection and press next.

3. The default view for the dataset will be plotted. Now one can change the tracks to display (with track display selection), as well as the order in which the samples should be ordered (track sort order). The order in which tracks are selected for ordering will also dictate the final sort. For some complicated sorts, it may be necessary to create a numeric track that puts the sample in the intended order.
5.6 Final remarks / future directions

Some of these functionalities have been developed recently. If you run into any quirks or annoyances don’t hesitate to contact R2 support (r2-support@amc.uva.nl).

We hope that this tutorial has been helpful, the R2 support team.
Differential expression of genes in your dataset

**Find out which genes make a difference between groups of samples in your dataset**

**6.1 Scope**

- Use R2 to determine whether the expression of your gene of interest is significantly different between groups of samples (step 1 to 5)
- Use R2 to find all genes exhibiting differential expression between groups of samples in a dataset (step 6)
- This is established by use of statistical tests. R2 will guide you through this process in a self-explanatory way
- In order to enable assignment of samples to groups, proper annotation of the dataset is required. In this tutorial a set of neuroblastoma tumors is used that is annotated with several clinical parameters: survival, age of diagnosis, etc.
- All (advanced) parameters can be adapted to your specific needs
- In separate boxes these settings will be elaborated upon
- The results of these analyses are presented in adaptable graphics

**6.2 Step 1: Selecting data and gene**

1. Logon to the R2 homepage using your credentials and make sure the “Single Dataset” field is selected in field 1
2. Make sure the Tumor Neuroblastoma public dataset is selected in field 2 (see chapter 1 of the tutorial for more information about the selection of a dataset)
4. Type MYCN as gene (see Figure 1) in field 4.
5. Click “next” in field 5.
6.3 Step 2: Choose annotation track as grouping variable

Did you know that you can create your own tracks?
This is explained in a separate tutorial Adapting R2 to your needs Datasets are listed alphabetically

1. In the next screen you will decide which grouping variable to use to establish the differential expression of your gene of interest. In R2 the so called “tracks” contain this annotation. For the current dataset there is a track called “alive” containing survival data of the patients from whom the tumor sample was taken. Select this (Figure 2). Note that the other fields can be kept as is, the right choices are already provided. Note also that the proper reporter probeset is already selected.

2. Click “next”
3. In the next window a selection of the groups can be made. Only the selected group(s) will be displayed in the graph; the “one way Anova”/”student T test” test will be performed for data on the selected groups (see explanation in the next step). In this case we want to see both groups so keep the selection as is.

4. Click “next”

6.4 Step 3: Anova results

R2 now performs a one-way Anova statistical test on the fly. This ANalysis Of VAriance is a statistical test that calculates whether the means of variables differ between two or more groups. In the case of 2 groups, this is identical to the student T-test. ANOVA can be considered a sound test when the variables are normally distributed and samples are independent. More information can be found here: http://en.wikipedia.org/wiki/One-way_ANOVA. A simple example calculation can be found here: http://en.wikipedia.org/wiki/F_test#One-way_ANOVA_example.

1. Check the graph and the information that is displayed above in the resulting window. For a nicely ordered graph, you can adjust the settings in the menu below to “Track and Gene Sort” (Don’t forget to click on Adjust Settings afterwards!). R2 displays the mRNA expression of the samples in a splitted plot in which per group the samples are ordered with increasing expression (Figure 4). Note that the “alive” annotation is in the second row (track) beneath the graph. The actual result of the ANOVA calculations is shown above

6.4. Step 3: Anova results 37
the graph; the difference in average expression between the two groups is significant. These results can also be shown in a more conventional bar-plot by adapting the settings and redrawing the graph.

Figure 4: Result of the one-way Anova test for the Neuroblastoma 88 samples.

2. Scroll down the window
3. Adapt the selection in the dropdown box ‘Graphtype’ to ‘Barplot’ and ‘ColorMode’ to ‘Color by Track’
4. Click ‘Adjust Settings’ (Figure 5). The resulting graph is adapted accordingly

Figure 5: Adapting the Graphtype to BarPlot and set Color by Track
Did you know that samples can be filtered and/or marked?

*Under the sub-header “Sample Filter” you can select a specific subset of samples based on the annotation (track). The analysis will only be performed on the selected subset. In the figure the track gender was selected that enables filtering on gender. Be sure to click the red confirm link to set the filter, or make further selections.

Filtering and marking samples: keep in mind that you can repeat the filter procedure on top of the previous one. Don’t forget the red “confirm” link before switching tracks. The extra graph option allows the users to select different graphical representations. In the samples to mark section, a sample name can be entered that will be highlighted in the resulting graph; ideal for publication purposes.*
6.5 Step 4: Adapting grouped plots

The difference can be shown more dramatically by plotting the data without a log2 transformation. Make sure to use log2 transformation in scientific reports, though, as none-transformed mRNA gene expression data is hardly ever normally distributed.

1. Scroll down in the window
2. In the ‘Adjustable settings’ dialog, set the ‘Transform’ dropdown to ‘none’ (Figure 7)
3. Click ‘Adjust Settings’
4. The resulting graph in Figure 8 shows the difference between the expression values in the two groups more dramatically.

5. Note furthermore the menu items to the right and left of the graph. The left panel contains hyperlinks that provide further information about this gene and additional analysis options. The KaplanScanner will be explained in a separate tutorial. The right panel allows you to change your gene of interest immediately, while keeping all the changes that you have made to the graphical representation; just type a gene name and click ‘Change Gene’. The “sample overview” shows the clinical data associated with the samples. Click the view button for the first sample.

6. Figure 9 shows the clinical data associated with that particular sample.

6.5. Step 4: Adapting grouped plots
6.6 Step 5: Finding all differentially expressed genes

1. It would be a pretty tedious job to look for all genes whether they are differentially expressed between groups. Why not let R2 do the job for you? Go back to the Main screen, by clicking the link in the upper left corner of the screen

2. In field 3 of the R2 step-by-step guide select ‘Find Differential expression between groups’ (Figure 10)
6.7 Step 6: Setting parameters

1. In the next window there appear quite a few choices for setting the statistical parameters for this analysis. Luckily only one setting is of real immediate importance; selecting the track of choice (shown in red). The samples will be divided by the groups this track provides. Select the ‘alive’ track again.
2. Click “next”

3. In the next screen click next also; R2 now calculates for all genes (under the parameters set) a one way anova test and also corrects this for multiple testing (Figure 12); an amazing amount of calculations! R2 performs this on the fly.

![Figure 12: Progress dialog during on the fly calculation](Image)

The result is a list of genes that is ordered for having the most significant differential expression between the groups you chose (Figure 13). A short summary of the calculation is given above the table; ~ 2600 genes have met the criteria set by default; their expression exhibits a correlation with the separation in the two groups.
R is the correlation coefficient; it ranges from -1 to +1. If R > 0 the value of two variables tends to increase or decrease together. If R < 0 the value of X increases if that of Y decreases, if R~0 there is no relation. Perhaps the best way to interpret the value of R is to square it. This is the fraction of the variance in the two variables that is shared. For example, if $R^2 = 0.59$ then 59% of the variance in Y can be explained by (or goes along with) variation in X. The p-value for this calculation estimates the probability that this is an observation by pure chance; a p-value of 0.01 you can be 99% sure that this is not the case.

Did you know that you can specify the preferred statistical test and choose a subset of genes?

**Hugo Once (hugoonce):** For most analysis genes should only be reported once in a dataset. R2 uses an algorithm called Hugo Once to choose a single probe-set to represent a gene. For each probe set of a gene, the average expression over all samples with a present call (from the MA55.0 normalization) is calculated (average present signal APS). The probe set with the highest signal is chosen to represent this gene in the analyzed dataset. For every dataset this procedure is repeated, thereby allowing tissue specific selection for probesets to represent a gene. When no call information is available, the average expression of a probeset is used. In platforms other than Affymetrix, we try to generate a similar score if a notion of being expressed is available. Illumina arrays for example can contain such information. For RNA-seq data, FPKM/RPKM/TPM estimates for a particular gene can be 0. These are also defined in R2 to be Absent or not expressed. The expression calls, can also be useful to...
limit the number of tests that need to be performed, and thereby reduce the multiple testing penalties. Genes that are not considered to be expressed in any of the samples in a cohort can be omitted for a valid reason.

Statistics panel: R2 determines p-values for the differential expression of genes by performing either a one-way anova (default setting) or alternatively a brute-force t-test on any combination of groups when the data is untransformed or log2 transformed. For rank-transformed data, a Kruskal Wallis test is performed. Besides these statistical tests, users can also ask for genes with a certain fold change or obtain a top-X list of the genes which are ordered by a user-specified test.

Correct for multiple testing: We are testing a lot of genes here; so we have to correct for multiple testing. For example, one might declare that a coin was biased if in 10 flips it landed heads at least 9 times. Indeed, if one assumes as a null hypothesis that the coin is fair, then the probability that a fair coin would come up heads at least 9 out of 10 times is \((10 + 1)\times(1/2)^{10} = 0.0107\). This is relatively unlikely, and under statistical criteria, such as p-value < 0.05, one would declare that the null hypothesis should be rejected i.e., the coin is unfair. A multiple-comparisons problem arises if one wanted to use this test (which is appropriate for testing the fairness of a single coin), to test the fairness of many coins. Imagine if one was to test 100 fair coins by this method. Given that the probability of a fair coin coming up heads 9 or 10 times in 10 flips is 0.0107, one would expect that in flipping 100 fair coins ten times each, to see a particular (i.e., pre-selected) coin come up heads 9 or 10 times would still be very unlikely, but seeing any coin behave that way, without concern for which one, would be more likely than not. Precisely, the likelihood that all 100 fair coins are identified as fair by this criterion is \((1 - 0.0107)^{100} \approx 0.34\). Therefore the application of our single-test coin-fairness criterion to multiple comparisons would be more likely to falsely identify at least one fair coin as unfair. This occurs in a similar way if we are testing multiple genes in one experiment; we have to correct for this. There are several ways to do so; a conservative approach is the Bonferroni correction. The correction is based on the idea that if an experimenter is testing \(n\) dependent or independent hypotheses on a set of data, then one way of maintaining the familywise error rate is to test each individual hypothesis at a statistical significance level of \(1/n\) times what it would be if only one hypothesis were tested. So, if it is desired that the significance level for the whole family of tests should be (at most) \(a\), then the Bonferroni correction would be to test each of the individual tests at a significance level of \(a/n\). The more sophisticated False Discovery Rate controls the expected proportion of false positives. A FDR threshold is determined from the observed p-value distribution, and hence is adaptive to the amount of signal in your data.

Gene Filters: The gene filters allow you to study a specific subset of genes only. There are several domains you can choose from.

- A specific chromosome can be chosen. Note that when a chromosome is chosen, a specific position range can be defined as well.
- Under GeneCategory some predefined categories can be selected, e.g. some examples are known transcription factors or drugtargets. Here you’ll find the categories you’ve defined yourself also. Kegg pathway selects a set of genes present in the KEGG pathway database.
- Gene Ontology selects a group of genes belonging to a specific Gene Ontology category (www.geneontology.org). Note that if you click a category, further choices deeper down the ontology tree are enabled. Click again on the same dropdown menu to view categories further down the tree.
- Genesets are publicly defined sets or sets you’ve constructed yourself (see for detailed instructions the tutorial “Adapting R2 to your needs”). A convenient search functionality is available to find what you’re looking for. Also in this dropdown feature subsets might be provided once a geneset is selected. Combinations are possible as well; this enables you for example to find the developmental genes on chromosome 1.

Of course: to really get familiar with these settings you have to toy around with them!
6.8 Step 7: Inspecting single gene

1. Click on the hyperlinked top gene in the list.

2. A similar graph as produced for MYCN appears, the differential expression is more pronounced for this gene (Figure 14). In the generated picture the genes are not ordered for their gene expression go to the adjustable settings menu and select “Track and gene sort” in the “Extra Graph Option” pulldown menu. Click “adjust settings”.

3. Figure 14: In the main screen Changedataset.

6.9 Step 8: Plot all genes and adapt visualization: Volcano plot etc

1. The tab with the list of differentially expressed genes (Figure 13) is still open. Click on this tab.

2. Most of the functionalities in the right panel of this window will be explored in more advanced tutorials (K-Means clustering). We’ll explore one additional data visualization however to plot all genes of this analysis. In the right menu click ‘Plot all genes (xy, volcano, etc)’.
3. The resulting plot shows all genes of the list in a XY-plot; datapoints above and below the diagonal are differentially expressed. Hovering over the points shows the genesymbol, in this case the NTRK1 gene (Figure 16). To speed up the graph generation this information is not automatically loaded: click on the “add hovering” button below the graph to add this information. Note: every plot in R2 with larger amounts of datapoints (>5000) will have this “add hovering” button.
4. Clicking on the datapoint in the plot opens up a new window showing the expression of the gene in the two groups as a box plot.

Figure 16: XY plot of all genes differentially expressed in the current track;
5. R2 allows further annotation of the XY plot of all genes; in the XY plot window (still open in your browser) scroll down and adapt the settings; add a genesymbol to mark, eg AKR1C1; choose a KEGG pathway to emphasize, eg DNA replication and set the ‘Draw fold lines’ option to ‘yes’.

6. Click redraw image. The plot has been adapted to show the AKR1C1 genesymbol, DNA-replication genes are highlighted in red. Fold change lines show the regions where differential expression is 1 and 2 fold (Figure 19). Note that most genes of the DNA replication pathway seem to be located below the diagonal.
7. R2 can visualize the same data also as a Volcano plot or an MA plot. In the Adjustable Settings dialog change the Plot type to Volcano or MA (Figure 19). Note that the distinct characteristics of the AKR1C1 gene and the DNA replication are more obvious in the Volcano plot. The DNA replication pathway statistics will be explored in more detail in the ‘Find genes correlating with your gene of interest’-tutorial.

---

Did you know that you can tailor visualization of specific genes in one go?

You can annotate gene names (gene symbols) by providing them in the ‘Mark genes’ field of the Adjustable Settings panel. By default, these will appear in red, size=10, on your plot. You can change the size and/or color of these genes either individually, or in groups.

Please take note of the following rules: to mark groups of genes for which the same criteria apply, first type the genes (comma separated), followed by :s=size, followed by :c=r,g,b for single genes: gene1:s=25:c=0,0,255;gene2:s=20:c=200,0,0 for groups of genes: (gene1,gene2,gene3):s=25:c=0,0,255; (gene4,gene5,gene6):s=20:c=200,0,0”

---

6.9. Step 8: Plot all genes and adapt visualization: Volcano plot etc
6.10 Final remarks / future directions

This tutorial has shown you how to find genes that are differentially expressed in your dataset of choice. Now go ahead and toy around with selecting groups and tracks of choice and see what interesting scientific discoveries might lie ahead!

We hope that this tutorial has been helpful, the R2 support team.
Find genes correlating with your gene of interest

Or how you can find genes that have similar or opposite expression patterns in your dataset of choice

7.1 Scope

• R2 allows you to explore the relations your gene exhibits with other genes in your dataset of choice; correlation statistics is used to calculate this.

• The expression of a set of genes correlating with the expression of MYCN in a series of Neuroblastoma tumors is used to demonstrate that in this tutorial

• The results can be further explored in one-on-one graphs or as a heatmap

• The set of genes can be further explored statistically in several domains as will be shown in this tutorial:
  – In a gene ontology analysis
  – On pathway maps
  – On a chromosome map

• Using this exploratory analysis, new biologically relevant hypotheses can be generated as will be shown in this tutorial by an example concerning MYCN and MCM genes.

• The data can be saved and used in other tools

• Further advanced analysis based on the use of sets of genes can be found in the Kaplan scanner and GeneSets tutorials.

7.2 Step 1: Selecting data

1. Logon to the R2 homepage using your credentials and make sure the “Single Dataset” field is selected in field 1 of the R2 step-by-step guide

2. Make sure the “Tumor Neuroblastoma public dataset” is selected in field 2 (For additional information on these first two steps, consult tutorial : Working with datasets

3. In field 3 select ‘Find Correlated genes with a single gene’ (Figure 1).

4. In field 4 type ‘MYCN’ as gene name
5. Click ‘Next’

![Image of R2 interface]

Figure 1: Choice of correlation analysis.

7.3 Step 2: Verifying settings

1. Further information on the statistics choices and the meaning of the HugoOnce mode you can find in the ‘Differential Expression’ tutorial.

Chapter 7. Find genes correlating with your gene of interest
2. Scroll down the screen and click ‘Next’

**Did you know that you can find the correlation between two genes directly?**

Just choose ‘Correlate 2 genes’ in field 3 if you have a specific gene you want to correlate with your gene of interest. Of course this method would be rather tedious if you want to find new genes, hence we’re exploring exactly this scenario in this tutorial. Another possibility is to correlate your gene with a track (containing numerical data). This essentially tests whether the expression of your gene of interest correlates with the numerical order described in the track. This scenario is further explored in the ‘Differential Expression’ tutorial.

### 7.4 Step 3: Inspecting correlating genes

1. R2 calculates the correlation of the expression of MYCN with the expression of every other single gene in the current dataset. A lot of calculations! The result is presented as two tables (Figure 3) In the header a summary is given: ~ 2200 combinations of MYCN and another gene met the criteria, i.e. having a significant correlation ($p < 0.01$) with the expression of MYCN. ~ 15000 genes did not obey these criteria. The left table represents the genes whose expression correlates positively, or is similar, with that of MYCN in this
dataset. Of course MYCN has a perfect correlation with itself. Some characteristics of the genes are already described in red. R and p-values are given in separate columns (for a short description of their meaning, consult the ‘Differential Expression’ tutorial).

Figure 3: Genes whose expression is correlating with that of the MYCN gene in 88 Neuroblastoma tumors

Exact (gene-) numbers listed in the tutorial Figures such as in this example (2208 combinations) can vary. This is caused by database updates upon a new genebuild release or an affymetrix annotation update.

1. The right table summarizes the genes that show a negative correlation; the expression of MYCN behaves oppositely to that of these genes.

2. A little table to the right summarizes the results of a limited Ontology analysis. More about that in subsequent steps where we also explore the menu items to the right. All gene names are clickable to explore the specifics of the correlation in a separate graph; try and click the APEX1 gene in the left column.

3. In the left upper corner the filter icon is located, this links directly to the “adjustable settings panel” where you adapt the filtering conditions. The filter button is accessible in many analysis modules of R2.

Figure 4: Filterbutton

7.5 Step 4: Inspecting correlation between specific genes

1. The resulting graph depicts the expression of both genes in this tumor series in a graph. The tumor samples are ordered by increasing MYCN expression. Note that the expression of APEX1 follows the expression of MYCN quite good! This is reflected in the R and p-values that are quite significant.
Figure 5: The expression of the MYCN gene correlates with the expression of the APEX1 gene.

2. For an opposite example, click on one of the top genes in the right column; MEAF6. This produces Figure 6. The original list of results is still open in another tab in your browser, return there.
Figure 6: The expression of MYCN has a negative correlation with that of the MEAF6 gene.

3. To generate a correlation plot where the negative relation between MYCN and MEAF6 gene is more clearly visualized select “XY-plot” as graph type in the graphics section in the Adjustable Settings box and click the Adjust Settings button. In this correlation plot it’s also still possible to show expression levels for the samples are distributed. In order to do so click on more settings in the Adjustable Settings box and set Histogram to yes, click Adjust Settings button. Now the histogram boxes in the x and y axes show expression levels are distributed for the samples in the selected dataset see Figure 7.
Figure 7: Toggle Histogram

4. Another nice way to adjust the graphical representation of a XY plot is using the gene expression levels and apply these to a color gradient.

7.5. Step 4: Inspecting correlation between specific genes
5. Select in the “Color Mode” pull down menu the “Color by gene” option. In the next box enter the gene you want to use for coloring the dots. Make sure that after entering the gene name you also select a corresponding probeset and click “Adjust Settings”. In this example the reporters of the MYCN vs MAEF6 are plotted and subsequently colored by the MYCN expression levels of course you can also enter a third gene for coloring the dots.
6. Another option to visualize the relation for the correlation of expressions in XY plot is to switch on the linear fit option. In the “More settings section”. Turn on “linear fit”.

![Figure 9: Select Color by gene](image)

![Figure 10: Linear fit](image)

7.5. Step 4: Inspecting correlation between specific genes
Figure 11: Linear fit

7. Sometimes it could be that you encounter a correlation plot for two genes where you can distinguish two clusters. One group of the samples seems to form a cluster with a positive correlation and a second cluster seems to have an inverse correlation. An example which is not directly listed in previous list of correlating genes is for example the relation between MYCN and GATA2. In the right upper corner enter MYCN and the GATA2 gene and click ‘Change Genes’. Further change the color mode in color by track, select the MYCN track in the graphics section and turn on the linear fit option at ‘More settings’ as indicated below, click redraw.
Figure 12: Linear fit adjust

8. In the next figure the trend line clearly illustrates there is positive correlation for the MYCN non-amplified group and a negative correlation for the MYCN amplified group.

7.5. Step 4: Inspecting correlation between specific genes
9. Through the menu to the right several additional dataviews and analyses are available. Let’s start with different overviews; R2 is able to produce heatmaps of this analysis. Click on the ‘Heatmap (zscore)’ menu item Figure 8. The gene names are on the y-axis, sample names on the x-axis. Return to the genelist view (Figure 3).

![Figure 13: Linear fit per track](image)

![Figure 14: Heatmap view of the expression of all genes correlating with the expression of MYCN in 88 Neuroblastoma samples.](image)

### 7.6 Step 5: Relation with Chromosome position

Another view is the mapping of these genes on all chromosomes. Click on the ‘Chromosome Map’ menu item. In Figure 13 this mapping is depicted in an overview. Sometimes, eyeballing already suggests that some regions seem to be affected. R2 provides a table where the statistics behind this analysis are given: Figure 14. The overrepresentation of genes that correlate with MYCN expression with respect to all genes present on (an arm of) a chromosome is calculated. You can also explore the results in the interactive R2 genome browser, where you can zoom into the results and locate individual genes. To enter this mode, just press the ‘View in R2 genomebrowser’ button.

---

**Did you know that over-representation is explained here?**

*Over-representation quantifies the notion that a subset of genes from a larger set can harbor more genes that have a certain characteristic than you would expect by chance. On the p-arm of chromosome 1 for example, there are 1157 genes located of the grand total of 21300 known genes. From our set of 2229 genes (only slightly more than 10% of the total number) some 210 are present on this arm. This is 18.2% an enrichment above what you would expect by chance. This can be quantified using a 2X2 contingency table with a chi-squared test that produces a p-value to establish whether this difference is significant.*

---
Figure 13: Mapping of the genes correlating with MYCN on all chromosomes

7.6. Step 5: Relation with Chromosome position
Figure 14: Statistics of overrepresentation of genes that have a correlation with MYCN on different chromosomes

1. To further explore this set of genes return to the list: Figure 3

7.7 Step 6: Establishing overrepresentation in other domains

1. Further overrepresentation analyses in other domains can give a first clue for processes that are of importance in this set of genes. A domain is the Gene Ontology; a controlled vocabulary that systematically describes processes, locations and functions in biology. Click ‘Gene Ontology analysis’

2. The resulting categories are presented in a sortable table (Figure 11), sort on p-value by clicking on the column header. Click in on a pathway ID will open a new screen or tab with the heatmap of the selected pathway.
Figure 15: Gene Ontology categories that are overrepresented in the set of genes that correlates with MYCN expression in the current dataset, sorted by increasing p-value of overrepresentation.

3. One of the categories where genes of our current set are overrepresented is ‘DNA-strand elongation’; and what is also obvious that all genes in this process have a consistent positive correlation (as can be seen by the green color). Let’s take a look if we can corroborate this observation in another domain.

4. The adjustable panel settings menu allows you to redo the gene-ontology analysis with the up or down regulated genes only.

Figure 16: Re-do analysis with genes that are either positively or negatively correlated with MYCN.

7.8 Step 7: Gene list in pathway context

1. Return to the gene list Figure 3 and click ‘Map on pathway image’
2. In the next screen a choice can be made for other datasets; we use the KEGG database. Click next.

Figure 17: Choice panel for other datasets

3. A similar overrepresentation analysis is performed on all gene members of the pathways in the KEGG database. Click on the p-value column header again to find the most significant ones: Figure 14

Figure 18: KEGG pathways exhibiting an overrepresentation of genes of the current dataset, ordered by significance.

4. The DNA-replication pathway pops up as most significant. Note that most genes are similar to the GO process found in the former analysis. The pathway will be shown when the blue A in front of the pathway
name is clicked.

5. A hyperlinked KEGG pathway appears: Figure 15

![KEGG Pathway](image)

Figure 19: Mapping of the overrepresented genes (darker green) in the MYCN correlating set on the DNA-replication pathway from the KEGG database. Hovering over the gene shows additional information.

MCM-genes seem to play a role. Go back to list (Figure 3) to show their individual relation with MYCN.

### 7.9 Step 8: Further pathways analysis

1. Scroll down and look for the MCM2 gene, click on the link to show their relationship: Figure 16
Figure 20: MCM2 expression correlates with MYCN expression.

2. The correlation is significant. In the left upper table there is a link to the Pubsniffer tool within R2. This tool performs a live search in the Pubmed literature database for (co-)occurrences of MYCN and MCM2 (and some other keywords). Click the link: Figure 17

<table>
<thead>
<tr>
<th>Query</th>
<th># Articles</th>
<th>Outlink</th>
</tr>
</thead>
<tbody>
<tr>
<td>MYCN AND MCM2</td>
<td>6</td>
<td>Pubreminer</td>
</tr>
<tr>
<td>MYCN</td>
<td>10637</td>
<td>Pubreminer</td>
</tr>
<tr>
<td>MCM2</td>
<td>787</td>
<td>Pubreminer</td>
</tr>
<tr>
<td>MYCN AND MCM2 AND cancer</td>
<td>4</td>
<td>Pubreminer</td>
</tr>
<tr>
<td>MYCN AND MCM2 AND development</td>
<td>1</td>
<td>Pubreminer</td>
</tr>
<tr>
<td>MYCN AND MCM2 AND neuroblastom</td>
<td>1</td>
<td>Pubreminer</td>
</tr>
<tr>
<td>MCM2 AND cancer</td>
<td>340</td>
<td>Pubreminer</td>
</tr>
<tr>
<td>MCM2 AND development</td>
<td>104</td>
<td>Pubreminer</td>
</tr>
<tr>
<td>MCM2 AND neuroblastom</td>
<td>3</td>
<td>Pubreminer</td>
</tr>
</tbody>
</table>

**Co-occurrences of "MYCN AND MCM2" within sentences of abstracts**

*Pubmed(7) MYCN was bound to the proximal promoters of the MCM2 to 8 genes*

Figure 21: Pubsniffer results for gene symbols MYCN and MCM2

3. Apparently there are some abstracts where the two genes are mentioned together, you can view this article directly by clicking the hyperlinked number in the Articles column. The outlink Pubreminer column directs to the PubReminer tool:

![PubMed PubReminer](image)

Figure 22: The PubReminer tool web interface

4. This versatile tool offers quite some functionality to build a literature search query tailored to your needs. That being slightly out of scope of this tutorial, click the “Goto Pubmed with query” button to find the article.

5. This article is actually published work by our group where the relation between the MCM genes and MYCN was proven experimentally.
7.10 Step 9: Gene set analysis

1. The genelist produced in the beginning of this tutorial (Figure 3) can be stored for use in later analyses in R2, or for use in other applications. Return to the page containing the list, this is still open in another tab in your browser.

2. The menu to the right gives several possibilities (Figure 53). Some of these have been explored already; we’ll touch shortly on the rest of them.

   • “Gene set analysis”: use public genesets; this is further explored in the advanced Correlate with DataSet tutorial.

   • “Map on pathway image”, “Chromosome map”, “Gene Ontology analysis”, “Heatmap” have been explored in this tutorial.

   • “MakeMeATable” produces a txt file that is formatted for direct input into the data analysis tool TM4

   • “Save current selection as TXT file” produces a tab separated file containing the current analysis. In the header of such file all information is stored to be able to redo the analyses in the future.

   • Reference for current selection produces a list of probesets and genenames that are considered to be expressed in the current dataset. This is a suitable background set for eg. the DAVID tool DAVID

   • Last but not least the data can be stored as a personal genecategory; this is further explored in the advanced tutorial “Adapting R2 to your own needs”.

Figure 23: The correlation between MCM genes and MYCN was proven experimentally in this article.
7.11 Final remarks / future directions

Based on this tutorial you can further explore R2 in the set of advanced tutorials.

Everything described in this chapter can be performed in the R2: genomics analysis and visualization platform (http://r2platform.com / http://r2.amc.nl)

We hope that this tutorial has been helpful, the R2 support team.
Investigate the prognostic value of a gene or a group of genes

8.1 Scope

- Use R2 to generate a Kaplan graph by “annotated parameter”. Use tracks or combine two tracks to assign the group separation of a specific dataset.
- R2 supports any type of survival data, such as overall survival and relapse free survival.
- Use the Kaplan Scan for a group of genes.

8.2 Step 1: Selecting the Kaplan Meier module

1. Logon the R2 homepage and select Kaplan Meier By Annotated parameter. You can find this option either in the left menu panel on the main screen or in field 3 at the type of analysis pull down menu. Using the Kaplan Meier module via the left menu directly shows from which datasets survival data is available. Make sure that “Tumor Neuroblastoma public “ Versteeg “ 88” is selected and click next.

2. In the adjustable settings menu choose overall survival “overall-c1103”, select “track” at Separate by and select “inss-cat” stage in use track pull down menu. Click “next”. Note that stage st4s en st1 survival curves are overlapping which is in agreement with the clinical outcome of the INSS stages.
3. A handy feature of the R2 kaplan module is the option to combine two tracks to generate subgroups for the Kaplan Meier analyses. Use the back-button from the browser and select at “separate by”, “combination of two tracks”. Choose for example for the first track “agegroup (cat)” and for the second track “mycn_amp (cat)”. And click next.

The combined track agegroup (>18 year) and no mycn application results in intermediate survival probability. Note that there are 3 groups instead of “expected” 4 since there are no patients <= 18 year and a mycn amplification, in this cohort.

Did you know that you can apply a filter to KaplanScan analyze a subgroup of patients for survival?
In addition you can also adapt the graphical representation*

You can select a subset from your samples by one or more tracks. Click on the pulldown of ‘Select a track (subset)’ and click on the track by which you want to define the subset. In the popup you need to check the box of your preferred subset(s). A successful subset selection will be shown as a small message indicating the used trackname, groups and the final number of samples between brackets.

If you want to further narrow down your selection with a different track, click on the same pulldown menu. Select the next track that you are interested in and in the popdown, check the preferred subset(s) from that track. Don’t forget to click on “Redraw Graph” after you made your final selection to redraw the Kaplan Meijer curve.

Nb. If you use the ‘back’ button in your webbrowser, then this selection will be lost and needs to be defined again.

8.3 Step 2: Kaplan Meier by gene expression; the Kaplan Scan

An often used feature of R2 is the Kaplan Scan (KaplanScan), where an optimum survival cut-off is established based on statistical testing instead of for example just taking the average or median. The Kaplan scanner separates the samples of a dataset into two groups based on the gene expression of one gene. In the order of expression, it will use every increasing expression value as a cutoff to create 2 groups and test the p-value in a logrank test. The highest value is then reported, accompanied by a Kaplan Meier picture. So in short, it will find the most significant expression cutoff for survival analysis. The best possible Kaplan Meier curve is based on the logrank test. However, R2 does also allow you to use median, average and more as a cutoff in assessing whether a gene of interest has the potential to separate patient survival. Of course, such analysis is only possible for datasets where survival data is present.

1. Select from the main screen either the left menu or in field 3, Kaplan Meier “by gene expression”. Make sure that “Tumor Neuroblastoma public Versteeg 88” is selected, for analysis choose “Kaplan Scan a single gene”, fill in MYCN and use as cut-off method “scan” and click “Next”.

2. In the next screen use the prefilled settings and click “next”.

3. The Kaplan scan generates a Kaplan Meier Plot based on the most optimal mRNA cut-off expression level to discriminate between a good and bad prognosis cohort.

4. The determined separation in groups can be stored in a track and used in other analyses, click the “store as track” button.
Figure 5: Kaplan Scan for a single gene

5. To illustrate that with the Kaplan Scan more significant biological subgroups can be found, adjust the cut-off mode to “median” in the settings menu and click “Redraw Graph”.

Figure 6: Kaplan plot with multiple cutoffs: A) Scan B) First Quartile C) Median D) Average

It is obvious that with the Kaplan Meier “scan modus” the sample grouping is much more significant compared to the median cut-off mode. Try to find out whether this is also the case with other cut-off modi.

1. Next to the Kaplan plot in Figure 5, a small sub-plot is generated (underlined) which represents a graphical representation of the p-value plotted against the mRNA expression level values. In some cases it could be useful to change the p-value cut-off level and for this reason this graphical p-value plot (which is clickable) could be of help. Alternatively, you could use the “cutoff” field to regenerate a Kaplan curve with that separation.

Figure 7: Adjustable settings menu: change p-value cutoff.
8.4 Step 3: Kaplan scan for a group of genes

1. Instead of using the Kaplan Scan for a single gene you can also analyse a group of genes at the same time. Go back to the main page ("Go to Main", upper left corner), choose Kaplan Meier "by gene expression", select at analysis "Kaplan Scan a group of genes" and click "Next".

2. In this example select the apoptosis route at the Kegg path way pulldown menu. Leave the “type of survival” at overall survival. In the statistics panel there are several filtering options possible, leave these options unchanged.

3. In the graphics section select “yes” at “Draw heatmap” and click next.

4. In the next screen R2 has generated a list of the genes within the apoptosis pathway which have significant prognostic value. A heatmap for this list of genes is generated as well.

Figure 8: A list of Kaplan Meier for a group of genes

In Figure 8, clicking on each gene name in the hugo column will result in a new screen or tab with the corresponding Kaplan plot.

Figure 9: Heatmap of the significant prognostic list of genes.

In this case, the heatmap shows 2 or 3 possible biologically relevant clusters based on this set of genes. Clicking a spot in the heatmap will show directly the gene expression level for all samples via a new one-gene-view screen.

1. To generate a binary heatmap based on the GOOD versus BAD prognoses, click on the back arrow of your browser and in the Adjustable Settings table set the Kegg Pathway under Gene Filters back to “All”.

2. Select in the Graphics panel next to Heatmap data “good_bad (binary)” from the pull down menu. Click Next.
8.5 Step 4: Kaplan scan on your own cohort

1. It may happen that you would like to use the KaplanScan method on a dataset that is not available in R2. Especially for this reason we have made a user defined version within R2, where you can paste your cohort into R2 from e.g. a textfile and run the procedure. To initiate such a user defined Kaplan Scan, select the “Kaplan Meier” > “Kaplan Meier by user provided data” option from the left hand menu.
For the remaining steps to work as intended, we need to take into account a couple of things. You should prepare your data in the following four tab- or semicolon(;) separated columns.

- Column 1 contains a sample identifiers (without spaces)
- Column 2 contains the survival time in days (R2 will convert these)
- Column 3 contains the censoring information (event) and can be yes/no or 1/0
- Column 4 contains the expression value of the gene of interest for the kaplanscan

One can easily prepare this information in Microsoft Excel and paste the selected columns into the large white paste box. Do take care that we use “.” for decimal signs.

After you pasted the dataset information, you make the selection for the cutoff option and subsequently press next. R2 will now calculate the kaplan method that you selected and display the result in an interactive image.
Figure 13: Kaplanscan with user defined data result

5. Once the image has been created, you are able to adapt various parameters to optimize appearance of your result.

Did you know that the survival data used in your scan produces a unique signature?

*R2 will indicate within the image a checksum (MD5 sum) of all the survival information, which can be used to identify whether the same cohort information has been used in different scans that you may perform (this code should remain identical).*

### 8.6 Final remarks / future directions

Everything described in this chapter can be performed in the R2: genomics analysis and visualization platform (http://r2platform.com or http://r2.amc.nl)

We hope that this tutorial has been helpful, the R2 support team.
Which known pathways play a role in your data?

9.1 Scope

• In molecular biology the concept of pathways is important; small molecules, proteins, genes, etc. interact, resulting in specific phenotypic outcomes at all levels in biology.

• Quite a lot of this knowledge is stored as pathways in databases. An extensive resource can be found here http://www.pathguide.org/. To name a few:
  – WikiPathways (http://wikipathways.org)
  – PantherDB (http://www.pantherdb.org)

• R2 allows you to see whether biological pathways might play a role in your dataset of choice.

• In this tutorial we’ll use array data of a set of 62 Medulloblastoma tumors. In some Medulloblastoma tumors the gene beta-catenin is mutated. This specific dataset has clinical annotation for beta catenin mutations. We’re going to investigate this in a pathway context.

9.2 Step 1: Selecting data

1. Make sure that the Single Dataset option is selected in field 1 of the step by step guide.

2. In field 2 locate and select the ‘Tumor Medulloblastoma PLoS One- Kool - 62 MAS5.0 -u133p2’ dataset by clicking ‘Change Dataset’

3. In field 3 select ‘KEGG PathwayFinder by Gene correlation’

4. You might not know the exact gene symbol for beta catenin. R2 can find the gene symbol by alternative name also, we’ll try ‘catenin’; Figure 1.

5. Click ‘Next’
9.3 Step 2: Choose the right gene

1. R2 has found several suggestions with the word catenin, hovering over the gene symbols gives additional information. Based on that information choose CTNNB1, take the probeset with the highest average expression, this is most likely the probeset that best represents mRNA concentration.
Figure 2: Options for alternative name catenin: choose CTNNB1

2. Scroll down, leave the other options as they are, and click ‘Submit’.

### 9.4 Step 3: Correlating pathways with a gene

1. R2 calculates for all genes in the KEGG pathways whether their expression correlates with that of CTNNB1. Next it calculates for all pathways whether they contain a significant number of correlating genes; if the genes correlating with CTNNB1 are overrepresented in that pathway (For an in depth discussion see R2 Tutorial; Find genes correlating with your gene of interest. The result is returned as a list of pathways; Figure 3.)
Figure 3: KEGG pathways that have an overrepresentation of genes that correlate with CTNNB1 in this dataset

2. An overall explanation is printed above the list; of all genes present in all KEGG pathways, ~540 correlate with CTNNB1 with a p value < 0.05. In the table the KEGG pathways are listed ranked by their p-value for overrepresentation (background in red) or under-representation (in green) of these genes. The brightly colored letters in front of the pathway-name are hyperlinked. R links to a list of the genes, K leads to the original KEGG pathway on the Japanese servers, A links to an image of the KEGG pathway that is provided with hover-over information for all genes in the pathway. We’ll discuss the first two later, now click on the A in front of the ‘SNARE interactions in vesicular transport’-pathway.

Figure 4: The SNARE pathway; darker green and red are genes correlating with CTNNB1

3. R2 opens a new window in your browser (Figure 4). In darker green the genes that have a positive correlation with CTNNB1 and in red those having a negative correlation. Hovering over the genes with the mouse pointer presents additional information; some of the gene-boxes represent multiple genes: Figure 5 Although not in this example, it may happen that multiple genes within a box show both positive, as well as negative correlations. In such case the box is proportionally filled with red and green.
4. The result however, is not quite convincing, apparently CTNNB1 expression does not correlate with pathways. We’re going to try it the other way around; which pathways correlate with a catenin mutation.

5. Return to list view (still open in another tab of your browser) and go to the R2 main page by clicking the link in the upper left corner of the screen.

Figure 5: Hovering over the Stx1-4 box shows that this actually represents 5 genes; only one of them is correlating with CTNBB1.

9.5 Step 4: Finding pathways relevant in groups of genes

1. In field 3 on the R2 start page select ‘KEGG PathwayFinder by Groups’;

Figure 6: Selecting PathwayFinder by Groups

2. Click “Next”
9.6 Step 5: Relating to data annotation

1. This set of tumors is annotated with several clinical and molecular biology parameters in so called tracks. One of them is the presence of a beta catenin mutation; bcat mutation. Select this; Figure 7. Pathway_Select


![Figure 7: Selecting the bcat mutation track.](image)

2. Click “Submit”

9.7 Step 6: Determining differentially expressed pathways

1. R2 calculates for all genes in the KEGG pathways whether they are differentially expressed between the groups of tumors having a mutation and those that do not have one. In a subsequent calculation the overrepresentation of these genes in the individual pathways is determined. From the resulting list it is obvious that the Wnt pathway has a strong overrepresentation of genes that are differentially expressed between the two groups.

2. Click on the R link to let R2 create a list of these genes.
Figure 8: The Wnt pathway has a strong overrepresentation of genes that are differentially expressed between the groups of tumors that have and don’t have a beta catenin.

### 9.8 Step 7: Verifying a pathway

1. A list of hyperlinked genes is returned, sort them by descending R-value by clicking on the R-column-header twice;

![R2 PathwayFinder](image1)

Figure 9: Wnt pathway genes correlating with Beta Catenin mutation as a list.

2. Each gene-symbol is hyperlinked to a graph representing the specific results; click the top gene in the list: AXIN2.

### 9.9 Step 8: Correlating with the expression of a gene

1. The graph shows an excellent correlation of the expression of the Wnt pathway gene AXIN2 with tumors having a Beta Catenin mutation. The same goes for a significantly overrepresented set of genes in this pathway. This specific group of tumors is also known as the Wnt-subtype in the Medulloblastoma field.
9.10 Final remarks / future directions

We hope that this tutorial has been helpful, the R2 support team.
Multiple datasets overview with Megasampler

Create an overview of the expression level of genes in multiple datasets

10.1 Scope

- The megasampler is a R2 module to investigate the expression level of a gene in any number of the numerous datasets stored in the R2 database
- Use R2 to compose your selection of datasets to investigate the expression level of a gene
- Use the megasampler “adjustable settings” to adapt the megasampler graphics
- The megasampler allows you to quickly get an overview of the selected gene expression level for all the datasets available in the R2 database
- Go directly from the overview to one-gene view to investigate in detail the expression level in a single dataset.

10.2 Step 1: Selecting multiple datasets

1. Select “Across Datasets” in field 1, by default the “megasampler” option will be selected in field 2 and click “next”.
2. Leave “u133p2, mas5.0” at the “type of data” option and select “XPO sampler” at “use presets”. The meaning of presets will be explained later on.

Did you know that R2 harbours different types of microarray platforms

Megasampler only allows you to query multiple datasets if they are of the same chip type and normalized by the same algorithm.

1. With the “selection preset” option a pre-stored dataset collection with associated settings can be selected. Select “XPO sampler” (Expression Project for Oncology (expO)) to pre-select a series of tumor datasets. Click “next”.

**Step 1:** Select the chip type
2. In the previous screen the preset “XPOsampler” is selected, a collection of datasets is already marked for the megasampler analyses. In Figure 3 clicking the small triangle unfolds the available dataset categories, notice that some of the datasets in the “tumor” section are already marked. In this way this you can adapt your pre-selection of datasets. Unfold the normal and tumor category and select the following datasets. Normal Adrenal gland - Various “ 13, Normal Brain PFC - Harris “ 44 and the “ Tumor Neuroblastoma public - Versteeg “ 88”. Enter MYCN and click “next”.

![Figure 3: Megasampler adjustment selection](image)

Did you know that private datasets linked to a specific user are indicated with a green background color

Add a private dataset to the (pre) selected datasets.

10.3 Step 2: Viewing a gene in multiple datasets

1. In the “adjustable settings” panel there are several options to customize the megasampler graph. For every selected dataset, you can change the order in which they are drawn by adjusting the number in the selection boxes. These are processed first, followed by the dataset names in alphabetical order (so changing the order of 1 or 2 datasets should be sufficient). The pull down next to “dataset ordering pull down menu” enables to split one or more dataset by selecting a track, in this way the chosen dataset(s) will be split according to the numbers of groups of the selected track.

2. For now change the color for the datasets as indicated in Figure 4 and click “next”.

10.3. Step 2: Viewing a gene in multiple datasets 91
3. R2 now performs a one-way Anova statistical test on the fly. This ANalysis Of VAriance is a statistical test that calculates whether the means of the expression levels between the selected datasets are significantly different.

![Figure 4: Adjusting the megasampler graph.](image)

![Figure 5: Anova test for the selected datasets.](image)

By default de megasampler graph is plotted in a so called Boxdotplot representation. The Boxdotplot shows a combined boxplot, on top of which the signals of the separate samples are plotted; a quickly interpretable graph.
Figure 6: MYCN expression levels in 15 datasets covering 2173 samples.

Additional insight can be obtained transforming the data, in this case transform the data to logical values (none) set “graphtype” on barplot and click on “redraw at the bottom of the screen.”
Figure 7: Different Megasampler graphical representations

The plotted graphs for “MYCN” clearly show a high expression level specifically in the Neuroblastoma data sets compared to Normal Tissue and other Tumor datasets. At the bottom of the page it’s possible to adapt dataset coloring, change the order and split datasets in tracks directly.

Did you know that you can save your selection of datasets and select your stored dataset the next time you login to R2.
Storing a preset not only stores the selection of datasets for future use, but will also keep all of the other settings such as order, colors, plot type etc. The same visual representation for any other gene can be generated in this way.

You can use the adjustable panel to adapt the megasampler graph. In case you split one or more datasets according to a specific track in the previous screen, it’s now possible to skip subgroups from your dataset or, more interestingly, apply different colors for groups within a dataset (see Figure 8).

Figure 8: Adjustable settings panel, color groups within a dataset.

10.4 Step 3: Expression distribution over many datasets

The red arrow in the “did you know box” brings you to a handy module to obtain a quick overview of the expression level patterns for most of the datasets R2 contains (providing that the normalization allows comparison).
1. Click “view Expression in many datasets” and a new screen (or tab) appears containing a Probeset distribution graph. The color of the dots represent the different dataset categories (cell line dataset, Tumor or Normal Tissue etcetera). Via this 2D distribution module you can easily detect in what way your probeset of interest is expressed in many other datasets. At the Y-axis the 2log transformed average expression level and the standard deviation is represented. The X-axis “overlap avoider” is simply a way to represent all datasets in the plot without overlap of the circles. Figure 9 clearly shows that the MYCN expression is also high in other dataset which could be of interest and a second Neuroblastoma dataset. Next to the graph 2 tables summarize dataset names and a R-value set to “1. This has no specific meaning in this context but comes of use with the 2D-distribution module where you can quickly scan the correlation between two genes for all datasets of the same platform in R2. This module is discussed in the Correlate Genes tutorial.

![Figure 9: MYCN expression level distribution for all u133-2 datasets in R2.](image)

2. Via the the probeset distribution view you can easily investigate a specific dataset in more detail. Click a preferred colored dataset dot and R2 will generate an one-gene-view graph. The one-gene-view representation is explained in more details in tutorial 2.

### 10.5 Step 4: Megasearch

We have already discussed the ‘find differential expression’ module for a single dataset to find differentially expressed genes. In the across dataset section we can also apply a similar approach, not between groups within single dataset but for a user defined selection of multiple datasets. However, keep in mind that you can only select datasets of the same platform, the most abundant datasets are of the Affymetrix u133p2 platform. As explained before not every platform can be used for the megasearch due to the normalisation procedure which has been used.
Figure 10: MYCN expression level distribution for all u133-2 datasets in R2.

1. Select ‘Megasearch’ in Box 2 and click next
2. At step 1 select the platform you want to use, for now select the default (u133p2).
3. Select the datasets you want to use for the analyses, in this example we have selected Normal Brain, AML and Medulloblastoma datasets, click next.

Figure 11: Dataset selection.

1. For the megasearch module only two groups can be used to find the statistical differently expressed genes. In the settings box assign the proper grouping parameters (1 or 2) leave the pulldown menu at the default setting (‘NO’) for the datasets and click next.
Figure 12: Assign the statistical group for testing 5. In the next adjustable settings menu select at Gene category ‘transcription regulator Act’ for gene filtering. In the ‘Hugonie Dataset’ pulldown menu the first selected dataset will be used as target dataset for probeset usage. For most platforms each gene has multiple probesets, when using this option R2 takes the probeset with the highest average signal. For the megasearch you can not use for each dataset a different probeset for a particular gene. In the ‘Hugonie Dataset’ pulldown menu you can change the target dataset in case you already familiar with one of the selected datasets to make sure that probesets from single datasets analysis are used. In case of OTX2 which is a marker gene for Medulloblastoma two probesets are designed (242128_at and 231731_at) in the most Medulloblastoma datasets depending of the subgroups 242128_at has the highest expression level and will be picked by R2 however in other type of cancers/tissues there is hardly any expression of the OTX2 gene and in that case the other probeset could easily be selected. At the statistics pulldown menu you can select fdr_moderate_t_statistics (Limma , Limma-git) or the standard uncorrected_t_test. The Limma algorithm is specifically designed for the analysis of gene expression data, leave the statistics at moderate_t_statistics and click next. Two tables of genes are generated with the highest significantly expressed genes for group 1 and group 2. In the left genelist (group 2) we find in the top 10, the OTX2 gene which is accociated with medulloblastoma.
1. In the previous Adjustable settings box, where the grouping parameters were assigned, you can also split datasets based on their subgroups (tracks) and incorporate the subgroups in different test groups as illustrated in Figure 14.

![Figure 13: OTX2 BoxDotplot](image)

**Figure 13: OTX2 BoxDotplot**

1. In the previous Adjustable settings box, where the grouping parameters were assigned, you can also split datasets based on their subgroups (tracks) and incorporate the subgroups in different test groups as illustrated in Figure 14.

![Figure 14: Assign the statistical subgroup for testing](image)

**Figure 14: Assign the statistical subgroup for testing**

---

**Did you know that the Megasampler can also be used to investigate through the methylation datasets**

---

### 10.6 Final remarks / future directions

We hope that this tutorial has been helpful, the R2 support team.
11.1 Scope

- In this tutorial expression data of a set of Medulloblastoma tumors will be investigated for the occurrence of groups that have similar expression patterns.
- Affymetrix data will be used in a k-means clustering analysis.

11.2 Step 1: Selecting data and module

1. Make sure that the Single Dataset option is selected in field 1 of the step by step guide.
2. In field 2 locate and select the ‘Tumor Medulloblastoma PLoS One- Kool - 62 MAS5.0 -u133p2’ dataset by clicking ‘Change Dataset’
3. In field 3 choose the ‘K-means’ option (Figure 1)
### 11.3 Step 2: Adapting settings

1. The next window presents a set of fields where specific settings of the clustering algorithm used can be set. There are only a few settings immediately relevant, the others are appropriate for most analyses. For the k-means clustering these are the number of groups and the number of draws. We’ll explain these shortly; for other settings refer to the boxed items.

2. K-means clustering requires a number of groups beforehand, we start with two. To see whether the outcome of the clustering is stable (see boxed text on K-means clustering) we set the number of draws (performing of the calculation) to 10x10.

---

**Figure 1: Selecting K-means clustering on the R2 main page**

4. Click ‘next’
Figure 2: K-means clustering settings

3. Depending on the size of your dataset or geneset you can enlarge or minimize your K-means plot by adapting the size of the rectangles at heatmap option. Click ‘next’

Did you know that K-means is a method of cluster analysis?

In data mining, k-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. This
might sound complicated but is easily illustrated: suppose we have a set of 12 patients where we observe the expression of two genes; expression of gene 1 along the x-axis, gene 2 on the y-axis (in our situation we have much more genes; the calculation will then be done in more dimensions). We’re now going to try to cluster this set of n patients observed in three groups; k = 3. The following steps illustrate the algorithm (1-4 from left to right) # k = 3 initial “means” are randomly selected in the data set (shown in color) # k clusters are created by associating every observation with the nearest mean. This partitions 2-D plane (the so called dataspace) in three areas. # The initial means are moved to the centers of the three areas; the centroids. # Steps 2 and 3 are repeated until convergence has been reached. As is obvious from the end point from this calculation this is a heuristic algorithm, there is no guarantee that it will converge to the global optimum, and the result may depend on the initial, randomly assigned clusters. As the algorithm is usually very fast, it is common to run it multiple times with different starting conditions and compare the outcome. R2 visualizes this in the end result of the calculation.

11.4 Step 3: Examining resulting clusters

1. R2 clusters the samples using the expression of 1500 genes exhibiting the highest standard deviation in this set. The result of 10 sets of 10 calculations each, is shown as colored bars (Figure 3). Below the bars a heatmap is shown of the expression of the genes involved. It is obvious that two consistent clusters are formed; the assignment of the samples to a respective cluster is always the same. Note that figures reproduced by yourself might differ slightly when weaker associations are investigated; k-means is non-deterministic (random initiation).

![Figure 3: Results for the 10x10 k-means clustering in two groups; two consistent clusters are formed.](image)

2. For visualization of k-means clusters, R2 performs hierarchical clustering on the samples for every group of k. Finally a hierarchical clustering is performed on the genes, making use of the information present in all samples. Because this is a large set only part of the map is shown in Figure 4. Below the heatmap, R2 will automatically test the association between the newly created kmeans separation and all of the tracks that are available for the current dataset (Fisher’s Exact Tests). This allows for quick discovery of interesting correlations that may yield biological insights. Depending on the availability of survival information, also
a Kaplan Meier analyses can be added to the analysis (and for example be compared to KaplanScan results for exemplar genes).

3. This dataset has a clinical annotation for BCAT mutations; the upper bar or track in Figure 3. The subgroup having this annotation seems to cluster together in one of the two groups; this is however a subset of one of the two current clusters; more groups are expected. We’re going to use a larger value for k to investigate this. In your browser click the back button and change the number of groups to 8.

4. Click ‘next’.
Did you know that ‘the # highest SD genes is the number of genes with highest Standard Deviation?’

Most of the other options (Sample/Gene filters etc) are explained in former tutorials. The “# highest SD genes” is the number of genes with highest Standard Deviation (genes that ‘make a difference’ in this set) that is used for the K-means analysis. By default this value is 1500.

Did you know that some virus scanners slow drawing of these graphs??

If R2 takes a long time to draw images like these this might have to do with your virus scanner. The graphs are interactive and contain a lot of scripts that are usually scanned by a virus-scanner like McAfee. You can avoid this by disabling Script scanning.

11.5 Step 4: Creating consistent clusters

1. The resulting clustering in 8 groups is depicted in Figure 5

   ![Figure 5: Clustering in 8 groups produces no consistent outcome.](image)

2. The outcome is not consistent (except for the cluster of samples having a bcat mutation)

3. Repeat the procedure for 4 groups.

4. In Figure 6 a clear and almost perfect clustering in 4 groups is established. Kool e.a. (2008) identified 5 subtypes in Medulloblastoma; this set is annotated in the lower track above the k-means tracks; they neatly follow 3 out of 4 clusters; the remaining two subtypes are less clear as is also clear when we try to separate the data in 5 clusters (Figure 7). By choosing subsets (categories) of genes before clustering you might be able to determine more precisely defined groups. An example of this method can be found in the Adapting R2 tutorial.
11.6 Final remarks / future directions

The identification of medulloblastoma subtypes has been published here: Kool M, Koster J, Bunt J, Hasselt NE, Lakeman A, van Sluis P, Troost D, Meeteren NS, Caron HN, Cloos J, Mrsic A, Ylstra B, Grajkowska W, Hart-

We hope that this tutorial has been helpful, the R2 support team.
Find related gene signatures with a specified genelist or novel correlating gene signatures

12.1 Scope

Within the current context, we define a signature as a collection of genes that are defined on a particular basis. This can be the presence within a gene-ontology class, the genomic location of a gene, or perhaps something potentially more meaningful like a functional pathway signature. Functional pathway signatures are mRNA proxies for a particular perturbation, such as the response to the downregulation of a gene, or the consequence of a targeted compound (drug). Especially in this context, the collection of genes may have predictive power for the activity of a process. Of course it becomes cumbersome to assess the activity on a gene-by-gene basis. It would be very handy if we could express the behavior of all the genes in a single value. Within R2, we can convert the behavior of a list of genes into a signature score that can be calculated for all samples within a particular dataset. This signature score is simply defined as the average zscore of a zscore transformed dataset (the standard way of visualizing a heatmap) (Figure 1). In R2, such scores are automatically generated when one generates heatmaps via the “view a geneset” function. With the exception of some exceptional cases, most functional signatures will be composed of both upregulated genes as well as downregulated ones. Using both as a single list may then become problematic, as downregulated genes may counteract the effects of upregulated genes, effectively leveling each other out. To circumvent this problem, we can create 2 separate gene categories, one containing only the upregulated genes, and one containing only the downregulated genes. R2 will recognize couples of gene categories if they follow a specific convection (fixed prefix, followed by _up and _down; e.g. mycn_up and mycn_down).
Figure 1: Signature score: one category vs up/downcategory

- What is a genesignature
- Create a track using the weight scores of a genesignature
- Relate a weighed genesignature track to a single gene
- Find correlating genesignatures with a track

Did you know that you can create gene category couples

R2 can treat particular gene categories in a special way if you follow a simple naming convention. Especially helpful for signature scores are up/down regulated gene couples. Within the “view a geneset” function, you can select multiple gene categories to be used in for the heatmap. If you select 2 categories that contain a fixed prefix, coupled to _up and _down (or _dn), then R2 will treat them as a couple, and will subtract the downregulated signals from the upregulated ones (effectively creating a signature score). We can weigh the 2 separate lists of genes either equally, or weighted as a percentage of the number of genes (the weighted_match / _wm signatures).

12.2 Step 1: Creating a geneset signature; a category within R2

As a start, let’s create the signature scores for a pair of gene categories. In this tutorial, we will make use of a published functional MYCN pathway activity signature that was created on the Neuroblastoma 88 dataset (Valentijn et al 2012). This signature is provided within R2.

1. We start at “Main”. Make sure that the “Single dataset” option is selected in “box 1”.
2. In “box 2” verify that the current dataset is “Neuroblastoma public - Versteeg - 88 - MAS5.0 - u133p2”.
3. In “box 3” we select “View geneset (Heatmap)”. Click “Next”.
4. In the following screen we select the “geneset__r2provided_genesets” Gene set collection and click “Next”.

Figure 1: Signature score: one category vs up/downcategory
5. In the following screen we select the “functional genesignature” subselection and click “Next”.

![Image of a screen showing a list of genes and a heatmap]

**Figure 2: Generating Geneset signatures**

1. We can now see the lists of genes that are represented in this collection. While holding the CTRL key pressed, we can select the following 2 gene categories “r2_imr32mycn_dn” and “r2_imr32mycn_up”. Then click “Next” (Figure 2).

2. R2 will now produce a hierarchical clustered heatmap image of the selected gene categories. Note that at the right end of the image the red boxes indicate in which category a particular gene was represented. In the bottom part of the heatmap, a blue-white-red colorscale is depicted for both gene categories. We can clearly see the opposing effects of the 2 signatures.

3. Scrolling down on this page, we will encounter a heading “Gene set values”, which presents a small table. The links within this table point to the numerical values of the geneset scores. For the 2 gene categories, R2 will create the scores of the 2 separate categories, a matched score (where up and down regulated genes are treated equally (50/50)), and a weighted_matched score (where up and downregulated genes are treated on their contribution (percentage for number of genes)). Click on “store” for the “weighted_matched” signature, so that we can perform additional analyses on it.

4. R2 has now assembled the information into a prescription to generate a track. By default R2 will store the track for 24 hours, which is fine for the current tutorial. Click on “Build set” to store the new track (Figure 3).
12.3 Step 2: Determine the activity of a signature

Now that we have created a signature from our 2 lists of genes, we can start using it as if it was a gene itself. For example we can inspect how the MYCN pathway activity signature correlates to the MYCN gene at the mRNA level.

1. Go back to the “main” page and select “correlate gene with track” from “box 3”. In “box 4” we provide “MYCN” and click “Next”.

2. On the following page, we select our newly created track in the “select a track” dropdown box and click “Next” (figure 4).

3. R2 will now produce a plot where the signature score for every patient is related to the MYCN mRNA expression value (Figure 5).

4. We can make this look a bit prettier by adapting the color for patients on the basis of e.g. MYCN amplification status. To achieve this, we go to the “adjustable settings” at the bottom of the page and select “Color by track” from the “ColorMode” and select “mycn_amp” from the “Track for color” option. Click “Adjust Settings” to redraw.

5. We can now clearly see that MYCN amplified patients have a higher MYCN activity score. The possibilities for numerical tracks are endless with some smart questions (Figure 5).
Figure 4: mycn vs signature score

Figure 5: Add groupcolouring

### 12.4 Step 3: Using signature scores

Now that we have related the signature to a particular gene, it is easy to envision that this can be done as an analysis as well, where the signature is correlated to all genes in the genome (“correlate with a track “ in “box3”). A lot of signatures have been designed and published in literature over the past years. We can convert all of these into signature scores and start searching for relations of these meta-genes with our signature of interest.

1. Go back to the “main” page and select “Geneset vs Geneset correlation” from “box 3” and click “Next”.

---

**12.4. Step 3: Using signature scores**
2. On the next page, select at the input Geneset -> Gene set Collection (source): “geneset__r2provided_genelists”. In the Genesets to Scan (target): select ‘geneset_broad_2012_oncogenic’ (Figure 6). Then click “next”.

3. In the next screen select in the source pull down menu “functional genesignature” and “ALL” in the target pulldown menu leave the remaining settings to their default and click ‘next’.

4. The gene signature consists of two parts. One set of genes which are up regulated by MYCN and one set of down regulated genes. In the next screen select both presented gene lists by holding the Ctrl- button (see Figure 6) and click “next.”

5. R2 has now generated all the possible correlations for the selected MYCN signature against all the gene lists within the broad oncogene category. This results in a table of geneset versus geneset correlations sorted by the p-value. The “venn source/ same / target” column provides insight in overlapping genes between two gene lists. Another informative parameter in the table is the range parameter in the last column. This value indicates the range of geneset scores in gene target signature (see Figure 6).

6. To inspect the correlation in more detail, we can click on the “XY-plot” link.

7. Now R2 has generated an XY-plot of all samples in the dataset. The XY values represent the signature scores for the 2 signatures for every sample. Below the image the overlapping genes in the 2 signatures are listed (see Figure 7).

8. We can also inspect the target signature as a heatmap by clicking on the “View heatmap of “”, providing gene-by-gene information (see Figure 7).
In the previous steps we have plotted the genesignature scores directly from a list of of geneset vs geneset correlations. We can also select and use genesignature scores to plot a XY-plot in the relate 2 track module from R2. In this example we will use MES and ADRN (mesenchymal, adrenergic) genesignature scores generated on a combined dataset of neuroblastoma cell lines and 5 neural crest derived cell lines published by (Groningen, Koster et al 2017).

1. Go back to the “main” page and select the dataset Mixed Neuroblastoma (MES-ADRN-Crest-Exp) - Versteeg - 52 - MAS5.0 - u133p2 in box 2.
2. In Box 3, select the “Relate two tracks” module and click next.
3. In the next screen select in the pull down menu at X-track , adrn_score (#) and at Y-track Y-track the mes_score (#) and click next. Now a XY- plot is generated representing the correlation of the two signature scores. However, a clear significant correlation between the two signatures is shown the biological relevance is less prominent.
Figure 8: Relate 2 tracks using genesignatures

1. In order to visualise the biological relavance of this correlation plot. Select at ColorMode , “color by track” and at track for color the “mes_adrn_time” track in the pulldown menu, click adjust settings.

**Figure 9: Relate 2 tracks using genesignatures color by track **

1. In this new plot, mes defined cell lines cluster together with the neural crest derived lines in the left upper part of the plot (orange and green respectively) and the ADRN lines in blue in the right lower part of the plot. The purple dots represent a time-series experiment where PRRX1 overexpression induces in time a transition towards MES defined cell lines. This is clearly shown by the dark purple colored dots where the light purple colored dots are controls or early time points.
12.6 Step 5: Drawing lines between samples in a XY plot

Sometimes it can be useful to indicate a relation between different samples within a dataset. In this case it could be informative to add a line between samples connecting the shifting samples in time. Let’s give this a try by defining the time series samples within this dataset.

1. Path properties: The appearance of the line can be influenced by providing a color (hex number) and a linewidth. The recipe for these adaptations makes use of ‘:’ and works as follows. sample1, sample2: colorcode: width. In the Sample paths box; Add ‘gsm2413257, gsm2413247, gsm2413248, gsm2413249, gsm2413250, gsm2413251, gsm2413252, gsm2413253, gsm2413254, gsm2413255, gsm2413256: #222222’ and click adjust Settings

![Sample paths box with added lines](image)

Figure 10: Connecting samples

2. In figure 10 now the samples of the time series are connected and follow the transition from ADRN defined cells to MES defined cell lines in this dataset.

Did you know box

*R2 uses a couple of markup options for points in a graph, you can enable these in the advanced prescriptions:*

- ‘dot’: places a thick border around the sample
- ‘circle’: Places a ring around the sample (diameter 9)
- ‘circle_2’: Places a ring around the sample (diameter 4)
- ‘circle_3’: Places a ring around the sample (diameter 1), effectively a thin border
- ‘epicenter’: Places a set of 3 rings descending in width around a sample
- ‘arrow’: Places a block arrow pointing to the sample
- ‘triangle’: Places a filled triangle under the sample

*Note: The dotsize does not scale with ‘arrow’ and ‘triangle’ met
12.7 Final remarks / future directions

Everything described in this chapter can be performed in the R2: genomics analysis and visualization platform (http://r2platform.com / http://r2.amc.nl)

We hope that this tutorial has been helpful, the R2 support team.
CHAPTER 13
Analysing Time Series

Search for up and-down regulated genes in a time series experiment

13.1 Scope

- Time series experiments, where a cell line is manipulated and followed over time, are analyzed and visualized in a separate section in R2. The actual analysis of any time series in R2 has been pre-calculated by the Affymetrix GCOS program. Within GCOS, every timepoint in a series has been compared to time-point zero. GCOS, makes use of the fact that every probe-set of an Affymetrix array is actually a measurement performed 11-16 times. These measurements are used as “independent” inputs to calculate a p-value. Due to the use of GCOS, the time series functionality is only available for the U133 type Affymetrix platforms. Of course, our advice would be to use biological replicates of all experiments.

- In most cases, time series experiments are performed by specific usergroups, whose data is often shielded from public assess. There are however also a few publicly available sets, to illustrate what this part of R2 can do. All the experiments performed on the same platform will most often be stored within a single set.

- Single gene expression levels can be visualized and list of regulated genes can be generated via a range of filtering methods. Please be aware that GCOS uses the probe information (contained within a probe set) from a single array to calculate p-values on. This is a controversial approach, but for many of the performed experiments the only way to obtain statistical values. If a timeseries experiment has been performed 3 or more times, then perhaps a better alternative would be to make use of the functionalities that are available for regular datasets.

- Use the time series module to investigate the expression level of a single gene.

- Use the time series module to generate a list of regulated genes.

- Use a list of regulated genes to analyze a dataset by using the geneset view.

- Use correlate with dataset to optimize your genecategory.

13.2 Step 1: Choosing the time series module and data

1. To view the expression pattern of a single gene from a time series experiment we make use of the “Time-series” module. Logon to R2 and select Time series in left menu panel of R2.
2. In field 1, select collection “u133p2 (public)”. Here “collection” is indicated as a category of Time series experiments. For time series the analysis is limited to the Affymetrix Hu133A or Hu133plus arrays. The “Collection” field not only implies the platform type but may also include another subgroup, in this case a publicly available Time series data. Select “View a gene; in field 2, type HMOX1 in field 3*, and click “next”.

3. In the next screen all the public available time series for the u133p platform R2 is hosting, is listed. In our example we make use of a Time series experiment published in Bioinformatics.2010 Feb 15;26(4):456-63. In this Time series, the experiments are performed in triplo. The A549 Adenocarcinoma cellline is treated with TGF-beta and the expression levels were measured at several timepoints. In Figure 2 click on the (+) sign to unfold the Time-course experiments belonging to the A549 cellline and click “next”. In the adjustable settings menu, leave all the default settings and click “GO”.

Figure 1: Single selection in the Time-series module

Figure 2: Timeseries selection screen.
In Figure 3 the expression levels of the HMOX1 gene are represented in a triplicate Time course experiment after stimulation with TGF-beta. Clearly the HMOX1 gene is an early responder and is upregulated with a maximum at 4 hours. The HMOX1 gene is presented by the authors of the corresponding publication as one the upregulated genes after TGF-beta stimulation. It’s should be noted that a control experiment is missing in this experimental design. Hovering over the individual timepoint reveals additional information.

Figure 3: Expression levels of the HMOX1 gene during a time course experiment

4. Another gene the authors claim to be upregulated by TGF-beta is the BCL6 gene. In the same screen you can quickly generate a time series graph by providing the BCL6 gene in the right upper corner and click “Search Gene”.
Figure 4: Probeset verification and Adjustable Settings.

The probeset verification table lists in the case of BCL6 the 3 probesets designed for the BCL6 gene. By default R2 will select the probeset with the highest average expression level. Clicking on the Tview link opens a separate application “Transcript view” to investigate the reporters in more detail. The Transcript view application is explained in tutorial 2 “one-gene-view”.

In the adjustable settings menu you can customize the Time series graph to your personal needs. Such as fontsize, Line width etc.

Did you know that you can contact the R2-support team to add your Time-series experiments

*Your Time series experiments will be listed as a separate collection and for private analyses only. The R2-support team requires the CEL datafiles provided by your Microarray facility, to generate the result files. (see chapter 20) *

13.3 Step 2: Finding regulated genes in a time series experiment

1. Instead of looking at one single gene, you may most likely want to find novel up and-down regulated genes in your cell-line experiment. Go to the main screen and select in field 2 “Select type of analysis, “create a list of genes” and click “next”.

2. In the next screen select again all the A549 timeseries experiment and click “next”.

3. In following screen you can use the table builder to apply all kind of filtering options to find the novel regulated genes. Some of the options are already set.
Figure 5: The time series table builder.

An explanation of some of the options follows below: **Min #experiments**: Depending on your experimental design select in how many experiments your gene should be regulated according to your filter. **Min highest expression in series**: Set the minimal highest expression in at least one of the samples in your Time Series experiment. In this way the genes which are not expressed above a certain level (e.g. background) will be skipped. **Min # present calls in series**: The affymetrix MAS5.0 algorithm detects whether a gene is significantly detected and receive a “present call”. **Min #significant change in series**: Whenever the expression in a time series is significantly altered compared to time point 0, a “change” call is elevated. At least one change in expression level should be significant (meaning that you also would like to see those results where only in 1 time point a change is observed). **Fold orientation**: determine the orientation here. Up / Down / Both **Minimal best fold in a series**: Set a minimal logfold that has to occur within the series. **Min lowest change pvalue in series**: The MAS5 algorithm provides a p-value of the fold change, before a probeset is considered changed a minimal p-value of 0.00025 should be met. Here you can set the p-value to a more strict level. **Force single reporter for hugo**: It’s possible to analyse all probesets representing a single gene. As explained in another tutorial by default R2 will select the probeset with the highest average expression level. Select and set the options as depicted in Figure 5 and click “next”.

13.3. Step 2: Finding regulated genes in a time series experiment
4. In Figure 6 a part of the up and down regulated genes are shown, listing the fold change and highest expression level for each Time series experiment. Clicking on a probeset link generates a single gene time series plot as shown in Figure 7.

**Figure 6: Up and down regulated genes table sorted on best fold change**

<table>
<thead>
<tr>
<th>Count</th>
<th>Gene</th>
<th>Probeset</th>
<th>A549.TGFB_1</th>
<th>A549.TGFB_3</th>
<th>A549.TGFB_2</th>
<th>hCNT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IGFBP7</td>
<td>201163_s_at</td>
<td>2199 6.5</td>
<td>1309 9.9</td>
<td>2496 12.9</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>KFL1</td>
<td>239430_at</td>
<td>5656 9.9</td>
<td>4824 8.4</td>
<td>4517 8.8</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>IL11</td>
<td>200924_at</td>
<td>2024 6.9</td>
<td>2027 7.5</td>
<td>2497 9.9</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>IGFBP5</td>
<td>211959_at</td>
<td>3821 9.5</td>
<td>3190 9.5</td>
<td>3712 5.6</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>INHBA</td>
<td>227140_at</td>
<td>536 9.2</td>
<td>448 9.2</td>
<td>650 6.4</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>RUNX2</td>
<td>232231_at</td>
<td>349 7.2</td>
<td>342 9.2</td>
<td>450 6.9</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>MAF</td>
<td>208648_s_at</td>
<td>2032 7.0</td>
<td>1967 9.1</td>
<td>2350 8.1</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>MMP10</td>
<td>205680_at</td>
<td>164 5.5</td>
<td>508 8.8</td>
<td>303 7.5</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>ANGPTL4</td>
<td>221009_s_at</td>
<td>2586 6.2</td>
<td>1587 6.6</td>
<td>3358 8.0</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>FRMD5</td>
<td>230331_at</td>
<td>238 3.3</td>
<td>371 8.0</td>
<td>234 3.5</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>RASGRP3</td>
<td>205801_s_at</td>
<td>284 6.8</td>
<td>274 6.3</td>
<td>230 7.8</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>KCMA1</td>
<td>221594_s_at</td>
<td>NA  NA</td>
<td>754 7.7</td>
<td>693 6.2</td>
<td>4</td>
</tr>
<tr>
<td>13</td>
<td>MARCH4</td>
<td>230112_at</td>
<td>577 6.5</td>
<td>556 7.7</td>
<td>478 6.7</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>SPOCK1</td>
<td>202363_at</td>
<td>1963 6.1</td>
<td>1276 7.7</td>
<td>1968 7.1</td>
<td>1</td>
</tr>
</tbody>
</table>

**IGFBP7**

201163_s_at
Figure 7: regulated gene.

5. Clicking on the filter button will open the “adjustable settings” panel to re-adjust the selection options. Clicking on the Venn “diagram button re-direct to the automatically generated Venn Diagram representing the intersection of the genesets.

Figure 8: Top buttons

Time Series Venn Diagram

Figure 9: Time series Venn diagram

*Did you know that Venn diagrams can be created directly from your gene categories of choice?

In the My Settings section you can upload text files containing your lists of genes and store them as gene categories. Repeating the procedure described above will produce the desired Venn diagrams.

13.4 Step 3: Using the regulated genes in further analyses

1. One of the strong points of R2 is that it supports directly further analyses with other modules and datasets stored in the database. Use the table of up and down-regulated genes of step2 to investigate if this list of genes can be of relevance in other datasets. In the left panel click “store results as gene category”. A new screen appears where you can enter an informative name for your gene category and add a short description. The gene category can be stored for 24 hours or stored permanently in your account, after which it is available each time you log in to R2. For now choose the Temporary option at “Where” and remember the name of the stored gene category for the next step.
2. It has been published that the timecourse expression data from the cell experiment used in this example is linked to epithelial-mesenchymal transition (EMT) by TGF-beta induction. It’s also known that this process plays an important role in Breast cancer. We can use this generated genecategory to investigate whether this of any relevance or not.

3. To keep the list of found genes for later usage, right click with the mouse on “Go to main” in the left upper corner and re-open the main screen of R2 in other tab/screen select at “change dataset” the following dataset. Tumor Breast - Iglehart - 123 - MAS5.0 - u133p2. In field 3 select “View Geneset” at “Select type of analyses” and click “next”.

Figure 10: Store a gene category
Using dataset Tumor Breast - Iglehart - 123 - MAS5.0 - u133p2

‘Figure 11: Geneview adjustable settings.

4. At “Gene set Collection” choose “My 24h geneCategories” and select the generated “genecategory” from step 1. Instead of an unsupervised sample clustering you can also cluster samples within a track. Select at “Order samples” by “ordering by track and click next.

5. Click next again.

6. Choose the temporary Genecategory generated via the Timeserie experiments, the track b-r_grade and click “next”.

13.4. Step 3: Using the regulated genes in further analyses
Figure 12: Heatmap of unsupervised clustering within a track of a selected geneset.

7. The samples are unsupervised hierarchically clustered within each group of the selected track and presented in a heatmap. The selected genecategory resulting from the timeserie experiment could be of clinical relevance since the clustering correlates with the Oestrogen Recepter track. At the bottom of the heatmap the z-scores of the selected genecategory is represented.

Did you know that clicking a spot in the heatmap reveals more info

*Clicking on a spot generates a one-gene-view for the chosen gene in the dataset only supported for a genecollection consisting < ~400 genes.*

Did you know that with “sample ordering” (Figure 11) you can manage the way the samples are clustered.

*By choosing “sample ordering “ by “track”, the unsupervised clustering of the samples is applied within the groups of a track. It is even possible to customize the way the samples are ordered by yourself (user defined order).*

13.5 Step 4: Correlate with other datasets

1. The module “correlate” the results with dataset compares the resulting genelist with a dataset of interest. Go back to the screen/tab of the generated gene-list (see figure 6.) Clicking the correlate with dataset button in the left menu redirects to a new screen. Choose the Tumor Breast - Iglehart - 123 - MAS5.0 - u133p2 in field 2, in field 3, choose “relate to differential expression” and click next.

2. Choose “er” at select a track and click “next”.

Chapter 13. Analysing Time Series
3. In the background R2 generates a list of genes based on the module “Find differential expression between groups” and presents the overlay of the results with the list generated in the “time series module”.

Save current selection as TXT file

correlations with p-values lower than 0.01 are colored in red

1: A549-TGFβ_1
2: A549-TGFβ_2
3: A549-TGFβ_3

Figure 14: Part of correlate with dataset genelist.

4. In Figure 14 the overlap is presented between the result from the “time series” module and the “relate to differential expression” option. The list of genes is sub-divided in a positive and a negative correlation list of genes. Clicking the “2-view” link opens a new screen with a combined graph of a one-geneview and the time series experiment.

13.6 Step 5: In a K-means analysis

1. Instead view a “geneset, you can also choose perform a “K-means” clustering based on your stored genecategory described in the steps “above”. Go to the main menu, select “K-means” at “Select type of analysis” and click “next”.

2. In the “Adjustable settings” panel leave most of the default settings but make sure that you select the already stored “Genecategory” at the clustering section select ’10x10” at “numbers of draw” and click “next”.

13.6. Step 5: In a K-means analysis 129
Performing a 10x10 K-means clustering via the main menu supports the potential clinical relevance of the gene category extracted from the timeserie experiment. The K-means cluster module is discussed in more detail in tutorial 10.

*Did you know that you can store the K-means generated track and use this track every time you log in to R2.

You can use this track for further analysis with a custom made track for example by using the “find differential expression between groups”. This approach is explained in more detail in tutorials : Differential Expression Of Gene Between Groups and “Adapting R2 to your needs”

13.7 Final remarks / future directions

We hope that this tutorial has been helpful, the R2 support team.
CHAPTER 14

Using and Creating genesets in R2

Or how you can achieve clear, presentation ready heatmaps of your dataset

14.1 Scope

- In this tutorial the visualization of a set of genes will be explored
- R2 provides a conventional heatmap view
- This heatmap view can be adapted to your needs by sorting the data along the axes according to your wishes.
- Generating your own genelists to analyze using the Toplister function.

14.2 Step 1: Selecting data and modules; creating a Heatmap

1. On the main page of R2 select View Geneset (Heatmap) (Figure 1). Click ‘Next’.
2. In the subsequent window three choices are available to customize the way the GeneSet data will be presented (Figure 2). The first field asks for a collection of genesets (Figure 3); R2 harbors hundreds of publicly available genesets, KEGG pathways being one of them. The second field is a filter for selecting the samples. The third field is the order by which the resulting heatmap will be drawn (Figure 4). By default R2 presents the data in a heatmap where a hierarchical clustering is performed on the genes, making use of the information present in all samples to order the data. We’ll show first what this ordering by clustering looks like. Keep all settings as they are and click next. The search fields below the Adjustable Settings dialog provide a shortcut for finding a GeneSet of choice, we’re going to use this now. Suppose we want to find a geneset containing Cell Cycle genes. Type Cell Cycle in the ‘Search a geneset’ box and click ‘Search’.

Figure 2: The R2 Gene Set view settings; A set can be selected, filtered for subsets and the clustering
results will be ordered according to the selection. We’re now going to search for Cell Cycle gene

Figure 3: Available collections of Genesets in R2

<table>
<thead>
<tr>
<th>Gene selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>KEGG</td>
</tr>
<tr>
<td>1K TFBS refseq</td>
</tr>
<tr>
<td>Public GeneCategory</td>
</tr>
<tr>
<td>My GeneCategories (1)</td>
</tr>
<tr>
<td>geneset_amc_genes</td>
</tr>
<tr>
<td>geneset_amc_genome</td>
</tr>
<tr>
<td>geneset_amc_test</td>
</tr>
<tr>
<td>geneset_broad_2010_compu</td>
</tr>
<tr>
<td>geneset_broad_2010_curated</td>
</tr>
<tr>
<td>geneset_broad_2010_go</td>
</tr>
<tr>
<td>geneset_broad_2010_motif</td>
</tr>
<tr>
<td>geneset_broad_2010_position</td>
</tr>
<tr>
<td>geneset_ctd_toxicogenomics</td>
</tr>
<tr>
<td>geneset_hprd_architecture</td>
</tr>
<tr>
<td>geneset_hprd_complexes</td>
</tr>
<tr>
<td>geneset_hprd_genetic_disease</td>
</tr>
<tr>
<td>geneset_hprd_modifications</td>
</tr>
<tr>
<td>geneset_hprd_tissue_express</td>
</tr>
<tr>
<td>geneset_pathway_commons</td>
</tr>
<tr>
<td>geneset_stanford_cancer_modules_segal</td>
</tr>
<tr>
<td>geneset_stanford_processes</td>
</tr>
<tr>
<td>geneset_stanford_tissues</td>
</tr>
</tbody>
</table>

Figure 4: Available ordering domains for samples

3. In the next window all genesets containing the words Cell Cycle in their description are shown. Choose the KEGG Cell Cycle (in the collection of Cellular Processes) by clicking the ‘View’ hyperlink.
Figure 5: Selecting a geneset out of the large collection of available sets in R2; Cell Cycle was used as a search term.

4. The Affymetrix data for the Neuroblastoma 88 dataset is shown for the genes in the Cell Cycle as a clustered heatmap. Hovering over the heatmap rectangles reveals the sample information stored in the R2 database. Keep in mind that the hovering option is limited to 10000 cells otherwise the graph generation consumes too much time. This limitation can be adapted in the ‘my settings’.
14.3 Step 2: Using multiple GeneSets

1. R2 also allows for multiple gene sets to be shown at once; return to the main page; select View Geneset (Figure 1: Select View a Geneset). We’re not going to search for a gene set but look for multiple sets at once in the KEGG collection. In the next window (Figure 2) leave the default collection to KEGG. Click “next” in the ‘Adjustable Settings’ box.

2. In the next screen a subcollection within the current collection of gene sets has to be defined; Select ‘Cellular Processes’ and click “next” (Figure 7).
3. R2 allows selection of multiple genesets at the same time; CTRL+ select Cell Cycle and Apoptosis and click ‘Next’.

4. The resulting heatmap (Figure 9) has the samples ordered by the result of the clustering of the dataset. On the y-axis the genes are annotated with their membership to both pathways; the upper bar is the Cell Cycle. It is obvious that part of the Cell Cycle member genes are clustering together.
14.4 Step 3: Relating genesets with data annotation

1. We're going to explore that in further detail by sorting the dataset according to the staging. Return to the former choice page (Figure 2) by clicking the back button in your browser. Choose ‘Order samples by a track’ and click “next” (Figure 10). In the next subcollection selection window choose Cellular Processes again (Figure 7) and click “next”.
2. Now select only the Cell cycle subset as GeneSet. Choose the Neuroblastoma staging INSS as track to order samples by and click “next” (Figure 11).

3. In the resulting heatmap it is clear that there is a segment of genes in the Cell Cycle pathway cluster that is consistently upregulated in the stage 4 Neuroblastoma samples (in red in the INSS track).
12: Heatmap sorted by INSS stage, there is a clear relation between the stage 4 tumors (in red in the INSS track) and up-regulation of a subset of genes of the Cell Cycle.

4. In the previous example R2 offers the possibility to set a fixed ordering of samples by track. It’s also possible to perform a clustering and set a fixed ordering of genes. In the “adjustable settings” panel, select user defined order in the geneselection pulldown menu and click next two times. Here you can paste a list of genes sorted according to your needs.

5. If you want to perform hierarchical clustering with a fixed sample order. From the main menu select View a Geneset (Heatmap) , in the sample order section, select user defined order and click next, select a geneset sub collection and click next. Copy and paste your sample order and click next. The export track function see chapter Error: Reference source not found can be handy for this issue.

14.5 Step 4: Unsupervised hierarchical clusterin with a geneset

It could be that for a given dataset there is no annotation to apply the analysis tools R2 is offering or your’re lacking a starting point to further investigated a dataset. A good starting point could be to do is to do a simple hierarchical clustering.

1. In chapter “Differential expression” is explained how to generate genesets describing the difference between tracks (groups) from an annotated dataset. Researchers often want to investigate the presence of subgroups without using annotation information in their dataset or just find a list of genes with the highest variation in gene expression. In case your dataset of interest lack annotation you are still able via the Toplister tool to investigate datasets for biological relevance.

2. For this purpose R2 is hosting a convenient tool to generated lists of genes using a filter. By using this filter you can select groups of genes with the highest, lowest or most variable genes etc etc.

3. To use the toplister tool go to the main menu and select tools>small tools > Toplister. Make sure Data Type is set to Expression data and select Tumor Medulloblastoma PLoS One - Kool - 62 - MAS5.0 - u133p2 and click next. In the adjustable settings screen all kinds of settings and filtering options can be adapted.
We want to know which 100 genes have the highest variation in this case leave “which set” at standard deviation (SD) which is the default settings. Click next.

4. R2 has generated a list of 100 genes showing the highest variation in gene expression.

5. Scrolling down to the bottom of the page shows possibilities to save or export the genelist to use for other analyses and/or usage in other datasets. Clicking on the Heatmap (Zscore) will perform an unsupervised hierarchical clustering and plots a heatmap. In this heatmap the subgroups are clearly clustered together as shown by Kool e.a. (2008).

Figure13: Unsupervised hierarchical clustering revealing subgroups in a Medulloblastoma dataset.

14.6 Final remarks / future directions

Everything described in this chapter can be performed in the R2: genomics analysis and visualization platform (http://r2platform.com / http://r2.amc.nl)

We hope that this tutorial has been helpful, the R2 support team.
How to identify patterns or groups in your dataset using Principle Component Analysis.

15.1 Scope

- In this tutorial expression data of a set of Medulloblastoma tumors will be investigated for the existence of subgroups.
- Principle Component Analysis (PCA) will be used to analyze the tumor samples.

15.2 Step 1: Selecting data and modules

1. Make sure that the Single Dataset option is selected in field 1 of the step by step guide.
2. In field 2 locate and select the ‘Tumor Medulloblastoma PLoS One- Kool - 62 MAS5.0 -u133p2’ dataset by clicking ‘Change Dataset’
3. In field 3 select the ‘Principle Component Option’ option (Error: Reference source not found)
4. Click “next”

15.3 Step 2: Exploring the principle components

1. The next window displays a set of fields where specific settings of the clustering algorithm used can be set. Leave all the settings at their default and click “next”.

2. Click to plot the PCA result.

3. You now see a plot of the first 2 principle components. In the adjustable settings box, all the combinations principle components can be selected. R2 enables visualization of the first three principle components (PC1,2,3) in 2D or 3D graphs.

4. In the adjustable setting box select the all PCA-components option to the several principle components combinations to investigate whether you can distinguish subgroups in your dataset.

Figure 1: Selecting Principle Component Analysis

Figure 2: Adjusting PCA settings
In this example the samples are colored by known groups and fitted with the PCA result. In Figure 4 a clear subgroup, the yellow wnt subgroup is revealed. Hovering over the data points provides the principle component vector #:values which are depicted, as well as additional sample information. This example illustrated that PCA is powerful tool aiding to find possible subgroups in your dataset of interest.

5). Select in the adjustable settings box “Label by Track” under LabelMode. A “Track for Label” pulldown menu unfolds, here you select your track of interest e.g “Samplenames” and click next.
Did you know that PCA clustering is a method that reduces data dimensionality?

[Principle Component Analysis is a method that reduces data dimensionality by performing co-
variance analysis between factors. PCA is especially suitable for datasets with many dimensions,
such as a microarray experiment where the measurement of every single gene in a dataset can be
considered a dimension. It is impossible to make a visual representation of the relation between
genes and their conditions in multi-dimensional matrix. One way to make sense of data is to reduce
dimensionality. Several techniques can be used for this purpose and PCA is one of them. The reduc-
tion of dimensions is archived by plotting points in a multidimensional space onto a space with fewer
dimensions. The reduction is accomplished by identifying directions, so called principle components.]
that describe maximal variation in the data. These principle components can then be used as surrogates to represent each sample, making it possible to visually assess similarities and differences between samples and determine whether samples can be grouped. As the principle components are uncorrelated, they may represent different aspects of the samples and is therefore a powerful tool to identify subgroups in you dataset.

Each datapoint in the graph is now provided with a label annotation.

15.4 Step 3: Viewing clusters in 3D

A very nice feature of the R2 PCA module is the possibility to investigate your data in an interactive 3D-plotted graph. Most recent internet browsers support the 3D visualization.

**Did you know that browser settings might have to be adapted?**

*In Firefox in some cases it’s not possible to rotate the 3D graph in that case adjust the following setting in firefox.: type “about:config” in the URL box, search for webgl and Enable”webgl.force-enabled”: TRUE. The 3-D module also works with Chrome.*

1. In the adjustable settings menu select the “3d” option and click “next”.
2. Click the cube and hold the left mouse button and rotate the picture in order to investigate whether there are any (more) subgroups that become visible.

![Figure 6: Showing a 3D PCA graph from different angles.](image)

3. By rotating the graph more subgroups could be revealed as clearly shown in Figure 6.
15.5 Final remarks / future directions

The identification of medulloblastoma subtypes has been published here:

Everything described in this chapter can be performed in the R2: genomics analysis and visualization platform (http://r2platform.com / http://r2.amc.nl)

We hope that this tutorial has been helpful, the R2 support team.
t-SNE: high dimensionality reduction in R2

16.1 Scope

- In this tutorial several expression datasets will be used.
- t-SNE will be used to find sub groups in datasets.
- t-SNE maps will be annotated with tracks and gene expression.

A clustering method that is gaining more and more popularity in biomedical research, the so called t-SNE algorithm. t-SNE stands for t-Distributed Stochastic Neighbor Embedding and is a machine learning dimensionality reduction algorithm that is well suited for the reduction of high dimensional datasets to just 2 or 3 dimensions.

Most researchers are already familiar with another dimensionality reduction algorithm, Principle Components Analysis (PCA) also available in R2 and explained in more detail in the Principle Components Analysis tutorial. Both PCA and t-SNE reduce the dimension while maintaining the structure of high dimensional data, however, PCA can only capture linear structures. t-SNE on the other hand captures both linear and non-linear relations and preserves local distances in high dimensions while reducing the information to 2 dimensions (an XY plot).

An important parameter within t-SNE is the variable known as perplexity. This tunable parameter is in a sense an estimation of how many neighbors each point has. The robustness of the visible clusters identified by the t-SNE algorithm can be validated by studying the clusters in a range of perplexities. Recommended values for perplexity range between 5-50. Once you have selected a dataset and applied the t-SNE algorithm, R2 will calculate all t-SNE clusters for 5 to 50 perplexities, in case of smaller dataset the number of perplexities will be less. Which perplexity is the best depends on the structure of the dataset, and is also a matter of what you would like to display (how the samples are placed). Before you start analyzing and interpreting the results, it is highly recommended to read about the power and pitfalls of t-SNE in this blog-post. Two important recommendations in this blog are that both size of, and distance between clusters do not have a well defined meaning. The fact that there are clusters has meaning.

Since running the t-SNE algorithm is a time consuming task and can take up to hours of processing time for large datasets, R2 stores the results for every dataset where the t-SNE has been completed. All users of R2 can explore generated t-SNE maps by coloring for tracks or expression values of a particular gene. Furthermore, the perplexity sweeps can be visualized. These options can be accessed via the left menu structure in R2. Users with collaborator, or higher access level are also able to initiate the generation of maps for datasets or subsets within a dataset. These additional options will be available via ‘box 3’ on the main page of R2.
16.2 Step 1: Selecting t-SNE maps

Let’s have a look at a t-SNE result to see what we can learn from this dimensionality reduction algorithm. The analysis is most informative with large datasets, and actually requires more than 16 samples as an absolute minimum (in R2). We will first have a look at the CCLE (cancer cell line encyclopedia) dataset which is comprised of more than 900 cell lines from various cancers.

1. In the left menu click on t-SNE maps and select in the pull down menu , ‘Cellline CCLE Cancer Cell Line Encyclopedia - Broad - 917 - MAS5.0 - u133p2 ’
2. R2 will assess the result and provide a button to start exploring the result.

![Figure 1: t-SNE preprocessed t-SNE maps](image)

1. Click “next”

16.3 Step 2: Annotating t-SNE maps

In this screen the t-SNE result is plotted with the highest perplexity, or a preset value that has been selected upon manual curation. There is no strict rule to select the ‘best’ perplexity. In most cases the highest perplexity is not the best choice to investigate the cluster further. If the perplexity result is something other than 23, then select this perplexity value. We can see structure in the location of the various cell lines. Now we would like to look at the annotations that are available for the cell lines.

1. In the ‘adjustable settings box’ set ‘perplexity’ to the value of 23.
2. Select ‘color by track’ from the ‘colormode’ and choose ‘primary site’. Press ‘next’ to redraw the image.
Another feature that may be informative in the context of a t-SNE map is to ‘overlay’ the expression of a particular gene on the map by coloring the cell lines by the expression values of a dataset, in this case mRNA gene expression. We can have a look at this by changing the ‘colormode’ to ‘color by expression’.

1. In the ‘adjustable settings box’ select ‘Color by Gene’ under Color mode and subsequently type ‘CLDN3’ under Gene for color. The corresponding reporter will automatically pop-up (Figure 3). The gene selection box autocompletes selection the proper reporter probeset, but this can take a little bit of time before the gene selection box appears.

Figure 3: t-SNE select probeset

1. Again click ‘next’ to refresh the view. In this view the samples are not colored by a group annotation (track) but by applying a color gradient which reflects the gene expression level according to a log2 scale. In this sample you can observe a subgroup of the carcinoma samples which have higher level in contrast to the (other) samples.
1. Use the track histology_subtype1 to generate a new t-SNE plot in the ‘Adjustable settings’ menu. It appears that the subgroup which stood out by the color gradient consists mostly of adenocarcinomas. Another gene which emphasizes the observation in the previous example is the NR3C1 gene showing an inverse gradient pattern for this subgroup.

Figure 5: t-SNE_Color by Gene
16.4 Step 3: Perplexity sweeps for t-SNE maps

What perplexity value is the best option for your dataset of interest? This depends on the embedded structure (the subgroups), and even what you personally would like to visualize (the way the samples are layed out). To assess the robustness of the layout as well as the effect that the perplexity parameter has, the R2 platform performs a perplexity sweep. The analysis will be run repeatedly, starting with a value of 5, and stopping at a perplexity value of 50 if the size of the dataset permits.

1. In order to generate an overview of all possible perplexities you have to set the number of perplexities to “ALL” and color by track mode to eg: histology in the “Adjustable Settings”.

![Figure 6: t-SNE: all perplexities](image)

By choosing the perplexity values ‘All’, miniature tiles will be generated for all perplexities (5-50), where it is still possible to use the color by track mode.

16.5 Step 4: Creating t-SNE maps

Depending on your access level in R2, you can create t-SNE maps from any dataset that is represented in R2 (with at least 16 or more samples). The t-SNE module will be located in ‘box 3’ at the main page of R2. You can either run the algorithm on the complete dataset, or focus on a particular sub-section of the samples using the ‘subset’ function.
Let’s take a look at some other nice example of a R2 generated t-SNE maps: the large dataset of normal tissue expression profiles

1. In main menu select Normal Tissues GTeX v4 - GTeX - 2921 - RPKM - ensgtxv4 in box 2 and select t-SNE in box 3. Click Next. If the ‘default’ map has already been calculated, a shortcut button will also appear as shown by the dashed box in figure 7. In the ‘Adjustable settings’ panel you can adjust several settings, such as sample filtering, specific gene categories and expression level restrictions.

![Figure 7: t-SNE: Menu](image)

Keep in mind that after adjusting input settings the t-SNE algorithm will re-run again, even though a t-SNE map already has been generated with the default settings. A note on the execution times of t-SNE: the generation of the maps will take a substantial amount of time to generate, especially for larger datasets (up to a number of hours for datasets >600 samples). Once initiated (showing the message that t-SNE is being calculated), you can close the window. The process will keep on running on the servers and you can view the results later by revisiting the analysis: when you return to the mainpage of R2, select the same dataset, again choose t-SNE in box 3, and click next. In the following window, a shortcut button to plot the requested t-SNE result will appear for your chosen dataset.

1. In our case we just click next

2. In the Adjustable settings box set the color by track on ‘Tissue’ and click next.
Suppose the t-SNE algorithm produced some interesting clusters that you want to explore further. R2 allows you to specifically select any subset of samples from the t-SNE map by using the lasso tool. The subset can be used as track in the other R2 analysis tools. This will be illustrated in the following example.

1. In the left menu click on t-SNE maps and select in the pull down menu, ‘Cellline CCLE Cancer Cell Line Encyclopedia - Broad - 917 - MAS5.0 - u133p2. Plot the corresponding t-SNE map using perplexity 23 and color the maps by selecting “primary site” with the track for color option. The haematopoietic group can clearly be subdivided in several groups which can be used to investigate these sub clusters in more detail.
1. Click on “select subset” below the t-SNE map. In the interactive pop-up t-SNE map you use the lasso tool by clicking on the map and hold the mouse button to draw a shape around the samples you want to cluster. After releasing the mouse button the samples are listed left to the t-SNE map annotated subsequently with a group id for each lasso selection action. You can select groups up to a number of 10. After you finished the selection click below the groups “build tracks for subset”.

2. In a new tab all the samples are listed with the designated and adjustable group label. The samples that were not included in any of the lasso selected subgroups are labeled ‘not_defined’. At the bottom in the “Adjustable Settings” menu you can rename the groups, select a color and store them in your personalized tracks or as a temporary track. Now you can continue with further analysis, for example by using the module “Find differential expression between groups” where you can find your newly created tracks in the selection criteria menu.
For now the lasso selection tool is only available for t-SNE maps module; implementation for usage with other modules such as PCA clustering will be released soon.

16.7 Step 6: Creating groups with the t-SNE DBSCAN tool

Next to the manual lasso tool for sample grouping on the t-SNE map, R2 provides an automated tool as well: the DBSCAN (Density-based spatial clustering of applications with noise). The DBSCAN allows for automatic detection of points that are closely packed together in a plot. A fun and more detailed blogpost about the DBSCAN can be found here.

Starting with an arbitrary point in the plot, the algorithm recursively groups together all the points that are located close to that point and the points within that group. If no more points can be found close to the group, another point on the plot will randomly be chosen and the process repeats itself. With two parameters you can influence the definition of groups: Epsilon and Min pts. Epsilon sets the maximal distance allowed between points to be considered close. Min pts determines the minimal amount of points that are needed in order to be called a group. It is recommended to set the minimal amount of points to 3 or higher. If a point is not within the epsilon distance of any cluster, it’s considered a “noise point”.

Let’s have a look at the DBSCAN and the parameters Epsilon and Min Ptstool in R2.

1. Go back to the tab with the t-SNE scan map which was generated in step 1 before applying the lasso functionality.

2. Click on “DBSCAN select subset” below the t-SNE map. The interactive t-SNE map pops-up. This time you can find slides for the two DBSCAN parameters, Epsilon and Min pts, on the right side. The parameters are set to a default value, which by no means are the best settings for the given dataset.
Figure 11: t-SNE: Using the DBSCAN selection

1. In this example the DB scan tool identifies two clusters which were also quite clear by just observing the t-SNE map. However, adapting the parameters Epsilon and Min pst described above can also aid you to identify less clear subgroups. Play around with these two slides till you find a satisfactory grouping of the samples on the t-SNE map. Don’t forget to click on the button ‘Refresh cluster graph’ after you have changed the values. On the right side an overview is provided that shows the amount of samples in each group.

Figure 12: t-SNE: Using the DBSCAN selection

1. Click ‘Build Track for subset’ to create tracks of these groups, in the same way as described in step 3 of the lasso tool above. The created tracks are stored and can be used as group parameters for further usage. Just as the lasso tool, the DB-scan tool is only available for t-SNE maps but will be implemented in other R2
16.8 Final remarks

Everything described in this chapter can be performed in R2: the genomics analysis and visualization platform (http://r2platform.com / http://r2.amc.nl)

We hope that this tutorial has been helpful, the R2 support team.
CHAPTER 17

Using the R2-Genome browser

Use the embedded R2 Genome browser to verify reporters

17.1 Scope

- In this tutorial we will investigate gene reporters and reveal information R2 (http://r2platform.com / http://r2.amc.nl) is providing based on the genome location.
- Explore gene expression reporters in the genome browser in combination with gene expression profiles (from the one-gene-view).

17.2 Step 1: Exploring the genome browser

1. In the main menu select in field 2 the default dataset “Tumor Neuroblastoma public - Versteeg - 88 - MAS5.0 - u133p2”. In Field 3 choose “View a gene” at “type of analysis”. In field 4: type “MYCN” and click ‘next’.

2. Leave all the settings at their default and click ‘next’. You have now arrived at the “One Gene View”. In this tutorial the main focus is the evaluation of the reporters designed by manufactures such as Affymetrix represented in the R2 Genome browser. To a lesser extent gene expression wprofiles will be looked at as well. When you slide down on the “One gene view” page of the MYCN expression, you encounter the “Probeset verification” table. The Probeset verification table, displays an automated analysis for U133 based Affymetrix platforms, where the reporter-gene relation validity has been verified by their genomic location (also described in more detail in the tutorial on One Gene View). Click on the “R2 Tview” link of the upper probeset and the embedded R2 genome browser will open in a new screen. The genome browser shows the genomic span where the MYCN gene is located together with the 5 MYCN probesets mapped to their genomic position.
When we access the genome browser via the One Gene View page, by default R2 has enabled a number of annotations (Tracks). At the top of the Transcript View display, R2 depicts all known expressed sequence tag (EST) and mRNA sequences aligned to the genome (synchronized with the USCS database regularly). These mappings serve as evidence for the existence of a gene, and are individually hyperlinked to the GenBank database. The EST and mRNA sequences are colored by the orientation of alignments, as determined by exon-intron junctions and poly-A signals to the genome. Here, green alignments indicate a 5'->3' mapping on the positive strand of the genome, while a red mapping represents a 5'->3' mapping on the negative strand of the genome (reverse complement orientation). In sequences where no information on the orientation is encountered, the alignment becomes blue. Underneath the plot the NCBI curated records for transcripts for the gene of interest are shown as reference sequences (RefSeq). The structure of the reference sequence has been indicated. In this MYCN example, the gene as represented by the different isoform refseqs, is green. This tells us that the MYCN gene maps to the positive strand of the genome, and should be read 5'->3' from left to right. The shadings in green, for the separate EST and mRNA mappings, indicate exon (darker) and intronic (lighter) regions (Figure 3 shows a legend to all the different color shades). If we look at the reporters underneath the refseq tracks, we see that most of them are green as well (thus mapping to the same strand as the MYCN gene). However, the 242026_at reporter appears in red (thus mapping to the negative strand), and thus cannot measure MYCN expression. Furthermore, this reporter maps at the intronic region, which is another reason not to use this reporter to represent MYCN. Still, this reporter is annotated to measure MYCN by the Affymetrix company. Below the EST and mRNA mappings, you can see the average gene expression for the reporters in the neuroblastoma dataset that we are investigating. This panel can be handy to check which reporter shows the highest expression (and often is the preferred reporter to use). Since the U133 platforms of Affymetrix are 3' based, you can nicely see that the reporter signals are higher at the 3' end of the gene. The 209757_at reporter is the most informative here, and was also picked by the hugoonce algorithm, embedded in R2. For Affymetrix arrays a probeset by itself is a collection of separate 25-bp reporters. For Affymetrix arrays other then the Hu133-2 and Hu133-a platforms the reporters may vary in the number of basepairs. These measured regions are indicated in the reporter track by very dark shades.
3. The properties and adjustable settings panel allows users to configure the graph display in various ways. In the left properties panel set in the transcriptview section “draw mode” to count and in the expression section “draw mode” to bars. The expression level can also be investigated per sample. The one-gene-view plot shows that ITCC0030 has no MYCN amplification resulting in low MYCN expression levels to illustrate this in the genome browser select in the Adjustable Settings panel, ITCC0030. Click redraw.

4. The picture now shows for one sample the expression levels for all MYCN probe sets in a more simplified fashion with barplots. Note the extra annotation tracks which were selected and hover over the tracks to reveal extra information.
17.3 Step 2: Zooming and panning

1. The R2 Genome browser is a highly interactive application offering several ways to zoom and scroll the genome display. Sometimes it could be useful to zoom into a location such as an aligned probeset. To quickly zoom into a specific region of interest, use the browser’s “drag and zoom” feature. At a desired position click and hold the left mouse button and drag the highlighted window to a second position and release the mouse button. The selected “white” region, can be repositioned (cross mouse indicator). A selection can be cancelled by clicking in the dark regions (Do note however that the positions of the selection were already adapted though ). Also in the track panel set “sequence and GC” windows to on. Click redraw in the middle panel.

2. At a larger magnification certain features such as basepair pair coloring at the sequence annotation track may become visible. Note the black rectangles in the dark green exon region a collection of the probes which form together a probeset . Repeat the same drag and zoom procedure for one probe and click redraw.
Figure 6: Zoom-in graph

Figure 7: Zoom revealing basepair sequence

Now the actual sequence is revealed a single affymetrix probe is matching. Clicking on the refseq bar will automatically zoom out to the genome browser representing the complete gene.

3. Click on the “GET DNA” button to retrieve the DNA sequence directly from the UCSC database (keep in mind this option is available until the region of interest reaches a certain size)

Did you know that the additional settings can be changed in “Tracks Panel”.

Clicking on the tool icon unfolds extra options to configure your graph. For many tracks, this will allow you to increase the size, but the settings may also include options that are specific to a particular analysis.
17.4 Step 3: Looking up chromosome regions

The Genome browser offers various options to zoom into a certain region. In case there is an interest in gene expression levels of a certain region on the chromosome. This can be quickly done by clicking on the chromosome at a certain location.

1. Click at a certain region on the chromosome and a new graph will be generated with average gene expression levels for the selected dataset in that region.

![Figure 8: Chromosomal clicking](image)

2. Furthermore it’s worth mentioning that in order to use the genome browser it’s not necessary to do so via first selecting a dataset. The genome browser can directly be accessed from the main menu including many basic functionalities.

17.5 Step 4: Working with multiple samples listed within a track

In some instances, genome tracks may list a whole array of samples, that can individually be selected for display. If R2 has the ability to also display multiple samples for that specific track, then the items ‘all’ and ‘custom’ will be represented in the sample dropdown list as well. The first does not require further explanation, however the ‘custom’ option probably does. If you select ‘custom’ as the sample to view, then R2 will listen to the ‘custom_id’ field at the bottom of the tracks box. Within this box, you may enter the sample names that you would like to view. samples should be separated by a comma and may include the % sign as a wildcard. In most tracks the samples will be visualized in the provided order. By using the ‘edit’ button next to the ‘custom_id’ field, you can make populating the ‘custom_id’ box a lot easier.

The R2 Genome Browser can be used with different genome builds (versions). Depending on the build, different annotation tracks can become accessible. In addition, depending on your access, additional tracks may be at your disposal.

17.6 Final remarks / future directions

Everything described in this chapter can be performed in the R2: genomics analysis and visualization platform (http://r2platform.com / http://r2.amc.nl)

We hope that this tutorial has been helpful, the R2 support team.
CHAPTER 18

DataScopes

Quick jumps to (dedicated) analyses and visualizations in R2

18.1 Scope

- Use data scopes to restrict data sets to a particular subject
- Use data scopes as a landing page for a specific publication / subject

18.2 Step 1: Selecting a dataset restrictive DataScope

1. DataScopes are represented in the left menu structure in R2 in the item named ‘Change Data Scopes’. Scopes can be grouped by different ‘types’ like ‘paper’ or ‘tumor type’. Depending on your account access more restricted scopes can be represented in your account. We will start by selecting a ‘dataset’ restriction Scope by clicking on ‘tumor’ > ‘neuroblastoma’.

![Figure 1: Selecting a Data Scope.](image)

2. You may notice that the ‘title’ of the platform now includes ‘neuroblastoma’. Furthermore, the number of samples that you have access to has now greatly decreased. That is, because you will now only look at datasets that we ‘annotated’ as neuroblastoma. This data scope can be great if you only wish to see data sets of a specific entity (that we have also defined). If you now want to select a new dataset, then you will note that you can only select neuroblastoma assets. In analyses like ‘2D distribution’, a restrictive data scope will also be applicable there.
18.3 Step 2: Selecting a Data Scope with a landing page

1. In the first example, the DataScope did nothing more than restrict the resources at your disposal, which can be handy for quick selections. A DataScope can become much more interesting if it has also got a ‘landing page’ associated with it. In the next example, we will make use of a datascope that belongs to a manuscript on the analysis of 500 whole genomes of medulloblastoma (Northcott et.al. Nature 2017). Select ‘medulloblastoma 500’ from the ‘paper’ group in the Data Scopes menu.

2. We can go to the ‘landing page’ for this data scope by either clicking on the button ‘Goto Medulloblastoma 500 Home’ in Box 0, or click on the ‘About Medulloblastoma 500’ item near the end in the left menu.

3. Within a landing page ‘tiles’ are defined that will quickly bring you to specific analyses and or visualizations.
These ‘hotlinks’ can be associated with advanced settings, and a such make the usage much easier. In the current Data Scope you can explore the genomics data for the 500 samples by investigating the annotation in the ‘cohort overview’, or check somatic mutation status in the ‘somatic variants’ tile, etc. Most of these analyses can also be found / accessed via the main interface while very specific ones may only be provided in the scope. Here the ‘tSNE manuscript’ would be an example of the latter.

This chapter only serves to explain the basic usage of a data scope. As you may realize, every scope can have many different options and analyses associated with it, which goes beyond what we can document here. Just have a look at the available scopes once in a while to discover what has been added, or make sure to follow our facebook page where we make announcements on this as well.

18.4 Final remarks / future directions

Some of these functionalities have been developed recently. If you run into any quirks or annoyances don’t hesitate to contact R2 support (r2-support@amc.uva.nl). If you may have an idea for another data scope or would like to showcase your own data in this way, then also please get in contact with us via the support address.

We hope that this tutorial has been helpful, the R2 support team.
Integrative analysis: ChIP-seq data

ChIP data visualization can be combined with other types of data

19.1 Scope

- Provide an introduction to the concepts and algorithms used in ChIP-seq data
- Check the properties of binding sites based on methylation and acetylation data
- Relate this to expression data
- Investigate the location of super-enhancers on the genome

19.2 Some concepts

Given the advanced character of this type of data analysis, some introduction on the concepts and algorithms used is in place.

19.2.1 What is ChIP-seq

With Chromatine Immuno Precipitation binding of elements to the genome can be studied. Transcription of DNA to RNA is regulated by the binding of these elements. These can be Transcription Factors, that bind temporarily to start transcription, but also chemical modification of the histones (molecular structures that coil the DNA) by methylation, acetylation, etc. (Figure 1) These modifications change the accessibility of the DNA for transcription.
When a specific antibody is used in the pulldown that recognizes these chemically modified regions, these specific regions can be studied. Regions with H3K27Ac acetylation mark active enhancers and active transcription, H3K4Me3 methylation marks active and poised transcription (Figure 2). Studying the relative contributions of both types of modifications allows a researcher to discern enhancer regions from active transcription sites.

Figure 1: Transcription; taken from Nature Reviews Genetics 12, 283-293 (April 2011)

When a specific antibody is used in the pulldown that recognizes these chemically modified regions, these specific regions can be studied. Regions with H3K27Ac acetylation mark active enhancers and active transcription, H3K4Me3 methylation marks active and poised transcription (Figure 2). Studying the relative contributions of both types of modifications allows a researcher to discern enhancer regions from active transcription sites.
Figure 2: Specific chemical modifications mark specific states of cis-regulatory elements; taken from doi:10.1016/j.molcel.2013.01.038

The assembly of the billions of fragments that result from a ChIP-seq experiment is a challenge. Algorithms to combine and map the reads into a consistent representation are under development. R2 allows you to study the outcome of these computationally intensive calculations through an intuitive visualization. Most default settings are suitable for a first impression of your data. To adapt certain parameters requires some knowledge about the actual computation, so we’ll explain some of the concepts used below.

Peak calling

R2 provides a couple of algorithms to assess significant enrichment ChIP between experiment and control. First is the MACS algorithm; this is often used in ChIP-seq data analyses and publications. In R2 it is used to study the binding of transcription factors. It’s drawback is that it is not very suitable for broad signals.

Some experiments can also be analyzed with the MACS2 algorithm. This version can detect narrow (like transcription factors) or broad (like histone modifications).

Yet another algorithm is RSEG; it is especially designed for histone modification detection. In R2 this used to analyse the histone modification patterns. To distinguish between specific histone modifications (e.g. acetylation vs methylation), R2 allows you to assess the same region in two profiles.

In the sections below, we will briefly explain how you can utilize and visualize the peaks as well as the histogram data (landscapes) that is available for most of the experiments.

Super-enhancers

An enhancer is a short (50-1500 bp) region of DNA that can be bound by proteins (activators) to increase the likelihood transcription will occur at a gene. They can be located up to 1 Mbp (1,000,000 bp) away from the gene, either upstream or downstream from the start site, and either in the forward or backward direction. A super-enhancer is a region of the mammalian genome comprising multiple of these enhancers, collectively bound by an array of transcription factor proteins to drive transcription of genes, often involved in regulation of cell identity. They can be up to 20 times the size of an enhancer.

For identification of super-enhancers R2 uses the Rank Ordering of Super-Enhancers algorithm (ROSE; more on the algorithm here). This takes the peaks called by RSEG for acetylation and calculates the distances in-between to judge whether they can be considered super-enhancers. The ranked values can be plotted and by locating the inflection point in the resulting graph, super-enhancers can be assigned. It can also be used with the MACS calculated data (figure 3).
Figure 3: Result of a typical ROSE analysis. Above the inflection point, marked in red, are super-enhancer regions.

Now that these concepts have been explained we’re going to see how the ChIP-seq data can be accessed through R2.

19.3 Step 1: Choosing data and modules

1. To enter the ChIP-seq analysis module in R2 select *ChIP data* in Box 3 (Fig 4) and click “Next”.

2. See figure 5. Several analysis paths start from here. First we’re going to explore the genomic environment of some genes in context of ChIP-seq data. In the ChIP-seq menu choose the *ChIP-chip Genome Browser*.

19.4 Step 2: Exploring genes in a transcriptional context

You’re now in the R2 genome browser in ChIP-seq context. By default the browser display shows a stretch on the genome around the location of the MYCN gene. Encoding regions of genes are drawn at the bottom of the graph. When in red they’re encoded in the reverse direction, coding exons are colored darker. Zooming and panning is enabled through buttons at the top of the page or by selecting an area. See chapter 17 for a more detailed explanation. The *Properties* panel on the left provides access to ChIP-seq datasets that can be added to the genome browser view. Options in the *Tracks* panel on the right allow for additional public data to be added to the genome browser as so called tracks. In the center panel you control what is being drawn. Always click the “redraw” button in the center panel for any changed settings to take effect.
1. As a first toe in the water we’ll explore the GATA3 gene, but you can choose your own. The genomic location of the gene will be used to map the annotation. Type the name of your gene in the text field of the Find gene textbox located in the upper-left corner and click “Go”.

2. To select the proper transcript in the next screen, click the “View” button.

3. To select one or more ChIP-seq datasets, click “Select/Adapt ChIP-Experiments” in the Properties panel on the left. As an example we write “lan1” in the text field of the chip_celline column. Check the box in front of the preferred experiment(s), optionally change the display colors using the “c.c.” buttons on the right and click “Update” at the bottom. Before we redraw the display, we adjust some additional settings in step 4.

4. In the Tracks panel on the right different annotation settings can be chosen. In Fig 8 we first adjusted the TranscriptView Annotation settings: The NIH Epigenome Roadmap to ‘all’ and the SuperEnhancers NB (George) to ‘on’. Next to the dropdown menus a toolset icon gives access to alternative displays of the information (e.g. a more detailed display per cell line can be chosen for the NIH Epigenome Roadmap information in stead of the cell line aggregated information obtained with ‘all’). An interesting feature of the center panel is the option to show the z-score of the expression data of the chosen dataset for each subgroup of a certain annotation track. This is illustrated in our example by the separate z-scores for each INSS-stage: in the center panel, choose “dataset_track” in the sample drop-down menu and set Select_a_express_track to inss (5cat). Now click on the “redraw” button in the center panel for the changes to take effect. The buttons at the top of the page allow for a further exploration around the gene. Clicking three times the “zoom out 2x” button reveals more binding in front of the MYCN gene.

In the resulting figure the H3K4me3, H3K27me3 and H3K27ac profiles for the GATA3 location in LAN1 neuroblastoma cell lines are shown. Above the genome strand the different histone modifications are depicted. The annotated locations of the Neuroblastoma superenhancer reported by George et al are drawn as colored blocks underneath the genome strand. Furthermore, the epigenetic profiles of the NIH Epigenome
Roadmap project are shown color coded for the chosen cell lines.

In this ADRN type cell line it is clearly shown that active GATA3 is associated with an enrichment of H3K4me3 and H3K27Ac, but not of H3K27me3.

Figure 8: Histone modification profiles around GATA3 location

5. Going through these steps would be a tedious job if you wish to inspect several genes. Suppose we obtained a list of differentially expressed genes from a transcription factor regulation experiment. As an additional requirement we selected for genes with a known cancer association. You can find the list here. Go back to the ChIP-seq choice menu. Now choose the ChIPSeq TSS Peak/Coverage Plotter.

6. We’re going to inspect Transcription Factor binding; Click on “Select a ChIP profile” and filter the grid by typing ‘BE’ in the name textbox, click somewhere in the table row of the BE2 cell line to select the data collected by Oldridge et al. and confirm by a click on the button “Use this experiment”.

7. Copy paste the genes obtained in step 5 or type genes of your interest into the Enter genesymbols / genome positions textbox. In the Gene Order selection box select ‘by_row_signal’ and ‘peaks’ as setting for Perspective, click “Next”. The Gata binding sites around the genes in the list are shown (Figure 9).
8. Since the ordering of the ChIPSeq Peak Plotter lists the genes with the highest signals on top (due to Gene order set to ‘by_row_signal’), we’ll select one of the first listed genes; click on ALK, in a new tab the GATA3 binding signal at the gene location is plotted in the R2 Genome Browser (Figure 10).

Figure 9: Gata binding site around genes

Figure 10: ALK profile within GATA3 ChIP-seq experiment

The view can be adapted by ticking additional datasets; e.g. GATA ChIP-seq experiments in other cell lines. Colors of the data can be adapted on the right side of the grid to easily distinguish them. Remember to always click the “redraw” button in the center panel for any changed settings to take effect. Zooming out produces Figure 11 from which it is apparent that in some specific cell lines there is enriched binding of GATA3 near the Transcription Start Site of ALK. Note that the properties have been adapted accordingly Range ‘a to 120’ and Slider on ‘average’ 5.
19.5 Step 3: Exploring histone modification patterns

Within R2 the regions of histone modification are calculated with the RSEG algorithm. The relative contributions of acetylation and methylation can be used to determine whether a region can be considered to be actively transcribed or to function as enhancer. This assignment can be further corroborated by including actual Transcription Factor binding data.

1. To perform such analyses go back to the ChIP-seq choice menu. Again choose the ChIPseq TSS Peak/Coverage Plotter. Use the grid to filter for the experiment used in this example: the SY5Y cellline that was profiled by Oldridge et al. for H3K27 acetylation, the RSEG peaks of this experiment are additionally scored by ROSE.

2. On the next page indicate the region to show on either side of the TSS; a commonly used value is 100 KB up and downstream; that makes 100000 for Distance (flank) an appropriate setting. Also indicate how many bases are to be collected within a bin. Do note that images are getting very large with small bin-sizes in combination with large regions, 1000 is a proper value in this case. Paste the same set of genes as used above in the genesymbols box. Set the Gene order to ‘by_row_signal’; this will make sure the gene with most enhancers in this region will top the list. Additionally we’re going to color the genebodies of ALK and BRD4: copy-paste without quotation marks “ALK;3BAA3B” and on the next line “BRD4;AA0000” in the color genebodies textbox and choose ‘Selected genes only’ for the setting Draw Gene Bodies.

Did you know that you can provide arbitrary locations on the genome?

Other than GeneSymbols (where R2 will find the most downstream TSS for you) to jump to a different location on the genome, you can also provide genome positions in the center panel, e.g. Chromosome ‘chr1’, Start ‘10020035’, End ‘chr1:10020035’. Or try clicking on a different part of the chromosome strand above the graph section.
1. R2 now shows for all provided genes a 100 Kb region up and downstream of the TSS. Note that the genebody of ALK and BRD4 are colored green and red respectively. Projected on the stretch are the bins that the Rseg-ROSE algorithm considers super-enhancers (Figure 12). Each stretch is clickable and will open a new tab. Click the topmost gene.

![Image of genebody analysis](image)

Figure 12: Histone acetylation around the TSS of a set of genes

2. For the topmost gene the acetylation data is shown on the chosen stretch. To further analyze what’s going on we’ll add GATA3 binding data and methylation data for the same cell line by checking the appropriate boxes. Click “redraw”. Note especially the region to the right where a super-enhancer is located, methylation signal is lower and there is not much GATA binding (Figure 13).

![Image of ChIP-seq analysis](image)

Figure 13: ChIP-seq signals around the TSS of a single gene
19.6 Step 4: Finding active super-enhancers

1. We’re now going to explore the ChIP-seq data the other way around, from the super-enhancer perspective. The selection of histone modified stretches on the genome are judged as super-enhancers by the ROSE algorithm. In R2 the most active regions can be explored through an interactive ROSE plot. Go back to the ChIP-seq choice menu, this time choose ROSE Super Enhancer Plot.

2. In the next screen, change the GenomeBuild to ‘hg19’ and click “Next”; The same algorithm as above is chosen: rose_se_pub_rseg_m2_s0_t0_v1

3. Again select the SY5Y dataset from Oldridge in the next panel.

Figure 14: Selecting super-enhancers from an interactive ROSE plot

4. R2 shows an interactive ROSE plot (Figure 15); dots in red are clickable and represent areas on the genome that ROSE has assigned as super-enhancer. Click one of them. In this example the 5th ranked enhancer was chosen.
Figure 15: Selecting super-enhancers from an interactive ROSE plot

5. R2 opens a new panel showing the location on the genome of the super-enhancer. To further explore which genes might be influenced, rescale the signal to a value of 150 (Range), put Slider to ‘median’ with value 5 and zoom out. The resulting picture shows that there are several genes in the proximity (Figure 16). Also present are other super-enhancers nearby. Feel free to toy around with the settings, and corroborate your findings by showing additional datasets in the same region.

Figure 16: The genomic context of a top-ranking super-enhancer

Note the “Load/ Store Profile” button in the upper-right corner of the page. With this button your chosen settings can be stored as a profile. Such a profile can be retrieved later on, and function as a template to wander on the genome using the same settings.

19.7 Final remarks

Any ChIP-seq dataset can profit from the visualizations provided by R2, just contact us if you want your data added. In R2 the ChIP-seq data visualization is still under development, any suggestions for improvements are welcome. Mail us at r2-support@amc.nl, also if you have any questions or remarks.
20.1 Scope

- For some patients (samples) R2 is hosting multiple types of measurements within the platform.
- Merging gene expression and methylation data.
- The Tumor Neuroblastoma - Lavarino - 23 - rma_sketch - u219 and Tumor Neuroblastoma - Lavarino - 41 - custom - ilmmhm450 datasets will be used for this tutorial.
- Find correlations between methylated sites within a gene and gene expression

20.2 Step 1: Choosing a combined dataset

1. From R2’s perspective, an analysis in which multiple datatypes will be combined is an ‘across datasets’ analysis. Therefore, we need to select this from the main page in box 1.

2. The easiest example of a combined analysis would be to simply plot the contents of 2 different types within a single plot. To do this, we select ‘view a gene in 2 datatypes’ and click ‘next’.

3. Since R2 needs to be instructed that overlapping samples may be identified, we create so called collections, which list datasets with overlapping patients. In this tutorial, we will make use of a public cohort where both mRNA gene expression as well as Illumina 450k methylation bead chip data is available. Select “neuroblastoma_gse54721” from the collection and click ‘next’.

4. Within the current screen you are able to select 2 datatypes to plot against each other. In the current example, only mRNA (Tumor Neuroblastoma - Lavarino - 23 - rma_sketch - u219) and Methylation (Tumor Neuroblastoma - Lavarino - 41 - custom - ilmmhm450) data is available, and within this collection only 1 option can be selected. Select the 2 datasets to combine in the pulldowns and type behind the dataset names ‘DDX1’ in the methylation data box and ‘MYCN’ in the expression data box and click next. R2 will automatically identify overlapping samples within the current selection and create ‘subsets’ for both datasets to only allow the overlapping samples for the plot. From the perspective of both datasets we can now select the reporter to represent the gene(s) that we indicated on the previous page. For now we will keep the preselected reporters. Furthermore, we can select the transformation for both datasets and continue to the actual plot. Click ‘next’ to advance to the image.
5. R2 has generated a XY-plot with the MYCN expression values on the Y-axis against the methylation ratios on the X-axis (Figure 1) with the combined annotation of both datasets. Annotation is being merged on the basis of the name of a track. In a perfect setting this would never result in conflicting data, however sometimes it may happen that the different datasets contain a different annotation. In that situation, R2 will concatenate both values by a semicolon and thus create a new group identifier. If there is an obvious mistake in one of the datasets, then we appreciate a message to R2-support@amc.uva.nl on this, so that we can correct it accordingly.

6. Both sides of the image represent the signals from both dataset perspectives. The image displays the cor-
relation between the 2 datatypes for those patients that were represented in both data sets. In this example plot we observe two main groups, to gain more insight in possible characteristics for the observed groups, we can adapt the visualization in a couple of ways. We can annotate the graph with a track distinction and color all of the circles accordingly. To achieve this, simply select ‘color by track’ and select at track for color “mycn_status” to be used for the coloring (Figure 1b) and click “Adjust Settings...”. Once redrawn, this will also add ‘boxplot’ representations on the sides of the image to represent the signals from both dataset perspectives.

7. Alternatively, we can also represent the image as a ‘YY’ plot, where multiple annotations will be represented underneath the image. In the adjustable settings menu, select ‘yy’ plot and click adjustable settings. (Figure 2).

![Figure 2: XY-plot Methylation versus expression](image)

8. As is customary for a methylation array, multiple reporters are represented around a gene. It could be of interest to inspect the methylation pattern for the other reporters of a gene. To visualize this information, we leave the current analysis by a right mouse click on ‘Go to main’ (upper left corner). In box 2 click ‘change dataset’, select methylation data and the Tumor neuroblastoma Lavarino dataset (Figure 3). Type ‘DDX1’ in box 4 click next. Leave the settings at their default and click ‘next’.

20.2. Step 1: Choosing a combined dataset
Figure 3: Heatmap select all reporters

Now click on ‘view all’ below the gene name. In a new screen (Figure 4) a heat map is generated with the methylation pattern for all samples of a given dataset. Beneath the heatmap the R2 genome browser is plotted with all the methylation reporters for the DDX1 gene plotted against their location on the genome. An alternative route, to the same heatmap representation would be to select ‘view all reporters for a gene’ in box 3 of the ‘Main page’.
20.3 Step 2: Correlate two datatypes

1. Next to simply ‘looking’ at reporters from the different datasets, we can also correlate 2 data types with each other. To achieve this, we first go back to the ‘main page’ by clicking on the upper left link.

2. We would now like to identify which gene has the best association with its methylation status. Therefore we need to correlate every gene with the methylation reporters that are in the annotated (for the the same gene) to belong to a gene.

3. Just like the previous example, we select ‘across datasets’ in box 1 and now select ‘dataset extender (within genes)’ in box 2. This will allow us to identify the best possible combinations where the expression of a gene correlates with the methylation status for the same gene.

4. Again, we need to identify the collection within which R2 will look for the overlapping samples. Select ‘neuroblastoma_gse54721’ and click ‘next’.

5. Leave all settings in the adjustable settings box and click ‘next’.

6. R2 will now perform the search for you. Do keep in mind that the across dataset searches can be quite intensive as all genes are being correlated to all methylation probes. For a simple setup like the current one, more than 2 minutes will be needed to obtain the result. To reduce the strain on the servers and speed up the serving of results, R2 will store the results of an analysis for a couple of days. If you are lucky that someone else has performed the exact analysis that you are interested in in the past few days, then R2 will serve those for you (which reduces the search from 2 min to a mere couple of seconds). This routine is used at multiple places within the platform.
7. Now R2 has generated a list of significant correlations for all the DNA-methylation reporters with the expression value of the corresponding gene (Figure 5). In Figure 5 a chromosomal overview of the significant p-value correlations are plotted, beneath the graph a table is generated divided in a list of genes which have a positive or inverse correlation selected gene expression probesets and their methylation reporter counter parts.

![Figure 5: Table of Correlations](image)

Further the green dots and red dots indicate respectively a positive correlation and negative correlation between the methylation and expression reporters for the same gene. Hovering over the dots will reveal more details such as the correlation values. Here you can investigate further the correlation between gene or methylation by clicking on the gene name. This will generate a YY plot of the gene expression and methylation correlation. (Figure 6).
Next to the YY-plot, more insight can be obtained by using the interactive genome browser. The genome browser is equipped with extended search options to find and explore more regions. The genome browser is a core module within R2 and will be discussed in a separate chapter of this tutorial. Click on the R2 - genome browser button. The genome browser graph is re-generated together with a complete panel of tools to adjust and / or annotate your visualization of the genome region where you are interested in.

8. In the correlation plot there is clearly a high correlation visible between the expression and methylation reporters located at chromosome 2. By clicking on the chromosome 2 area R2 will regenerate a correlation plot zoomed into chromosome 2. In order to investigate the area in more detail you can zoom into this particular region. Press and hold the left mouse button in the designated area (Figure 8) mark the area by dragging the pointer, release the mouse button and click ‘redraw’. 
Figure 7: Explore chromosomal regions

On this zoom-level more known information is automatically plotted in the genome browser. In this example it is shown that the DDX1 is located near to the MYCN gene on the genome. This could explain the high inverse correlation of the MYNC methylation reporters and expression reporteres.

---

Did you know that R2 will determine the overlap between datasets automatically?

*R2 will scan for overlapping samples on the basis of the r2_samplename. Overlap is automatically determined and therefore it can also use cohorts that are not completely overlapping. R2 will simply exclude samples that are only found in 1 of the datasets. In addition, the order in which samples are represented is also accounted for.*
Did you know that the annotation from both datasets is combined?

- On the sides of the image, the combined annotation represents the signals from both dataset perspectives. The image displays the correlation between the 2 datatypes for those patients that were represented in both data sets. From within this view we can adapt the visualization in a couple of ways. When we look at the XY plot, we can annotate the graph with a track distinction and color all of the circles accordingly. To achieve this, simply select ‘color by track’ and select the annotation source to be used for the coloring. Once redrawn, this will also add ‘boxplot’ representations*
Datatypes: Whole Genome Sequencing data and expression data

21.1 Scope

- In this part R2 is used to provide information about how Whole Genome Sequencing (WGS) data can be viewed, shared and analyzed. This is a specialized topic for which you need access to a dataset containing such information. Most likely close collaboration is needed to tailor your own data to your needs.

21.2 Step 1: View circos files.

1. To view circos plots of the sequenced genomes in R2, select Static circos files (v3) in Box 3 (Fig 1).
Figure 1: Choose the Circos plot module

2. Select a subset of samples by using the menu on the top.
   - Select \textit{ins} (cat 3) from the select a track (subset) selection box.
   - Select \textit{st2 (1)} from the pop-up selection window and click ‘OK’.

Figure 2: Select a subset of samples

3. In Neuroblastoma whole and partial chromosome gains and losses are frequent. This is indicated by the red and green colouring of the cgh-like scatterplots.
• One sample doesn’t appear to have large structural defects (N482TL).
• N482 indicates the sample_id, TL indicates that the circos plot shows data of the Tumor compared to the Lymphocytes DNA sequence data.
• Click on the N482TL tile and go to the newly opened tab of your browser.

Figure 3: Circos plot

4. Here we entered the detailed view of the circos plot section. On the right side you can open different information tabs.

1. Sample annotation.
2. Somatic structural variants.
3. Somatic structural variants of a limited size inside or close to genes that could be affected by them.
4. Hight quality non structural somatic variants.
5. A link out to the genome browsers showing a cgh-like plot of the sequencing data of a region of interest.
Figure 4: Cgh-like plot in the Genome Browser

- When you open the Gene Affecting Structural Variants (of limited size) tab you can now see two variants listed and not one as shown in the circos plot. For the circos plot a higher threshold was used for the read pair matches. For the table we show more but less accurate data.

- Click on view inside the link column.

- This opens a double genomebrowser view showing both sides of the selected structural variation.

Figure 5: Structural variant in Genome Browser

5. When the link under a genomebrowser location view is clicked the same location is shown inside the full genomebrowser view. Here you will be able to use the zoom buttons and select extra data to plot with the
cgh-like scatterplot and junction information. For this sample there is also affymetrix gene expression data available. Here the zscore is shown above the cgh-like plot.

Figure 6: Including Array data in the scatterplot

6. The *FOXR1* gene shows a high zscore and by clicking on the dot above this gene you will be taken into the one gene view of r2 showing the expression of this gene inside a Neuroblastoma tumor series. The investigated tumor is highlighted with a red circle.

Figure 7: FOXR1 in the tumorseries

Using the MegaSampler tool from within R2 you can also show that the FOXR1 gene is only expressed in a handful of samples out of thousands, and that these mostly are neuroblastoma cases. Out of the samples that could be checked, all appear to have created in-frame fusions with FOXR1. All of these findings, and additional experiments proving that FOXR1 can serve as an oncogene in neuroblastoma have been published by Santo et al in Oncogene (2012).
Adapting R2 to your needs

Or how you can optimize R2 for your specific data analysis

22.1 Scope

- This tutorial describes the adaptable settings within R2. These are basically all items under the My Settings menu-item. Through these you can personalize the use of R2
- First a couple of regular settings will be treated: changing colors, setting parameters
- Next we’ll delve into the practical adaptation of R2; uploading your dataset, adding your personal genesets (categories), creating/exporting/uploading your own tracks and maintaining a user community

22.2 Step 1: Adapt your settings

1. Personalizing R2 starts with selecting the My Settings menu item (Figure 1). First item under there is the general ‘My Settings’; choose that.
2. In the User Defined Settings window several parameters for analyses in R2 can be adapted to your preferences. For most analyses these are set to appropriate values, but of course you want to set your favorite dataset here!

![User Defined Settings Window](image)

**Figure 1: Personalizing R2: the My Settings menu-item**
Figure 2: The User Defined Settings window; most parameters are appropriate; you want to change your preferential dataset however.

3. Next item in the My Settings submenu are the ‘Megasampler Presets’ (Figure 1), these are of relevance when you’ve built a specific Preset in an analysis Across Datasets. (chapter: Multiple datasets overview with Megasampler).

4. The ‘Upload New Dataset’-, ‘Categories’-, ‘Tracks’- and ‘Communities’-menu-items will be elaborated further upon later in this tutorial. The ‘Timeseries colors’ allows you to set the colors for the graphs of specific experiments in a series (Figure 3).

22.3 Step 2: Upload your data

1. The extensive functionality of R2 can be best experienced when you have your own data available. In order to do so it first has to be uploaded and processed. Click the Upload New Dataset (Figure 1), the upload form appears (Figure 4). This has to be filled out according to MIAME guidelines as (amongst others) used at the gene expression database GEO (http://www.ncbi.nlm.nih.gov/geo/info/MIAME.html). If you have any specific annotations or wishes or microarray samples other than Affymetrix you’d best contact r2 support directly (r2-support@amc.nl). If you would like to see a publicly accessible dataset in R2, then send an email to r2-support@amc.nl with a link to the microarray data, or in the case of a Gene Expression Omnibus dataset, the GSE**** identifier, and we will take care of the rest and send a notification email.
Figure 4: Dataset upload form in R2; fill out the form in a MIAME GEO compliant way.

22.4 Step 3: Create your own genesets; a.k.a. categories

1. Another way in which R2 can be adapted to your specific needs is by introducing your own set of genes, which are called categories in R2; hover over the ‘Category’-item and select ‘Build Category’ (Figure 5)
2. The ‘Custom Category’ window pops up (Figure 6). Default this is set to upload from a text file formatted as a single column of official gene symbols (obtained from NCBI gene). The uploaded data can be stored temporary (24 hrs) or in your personal tracks database. We’re now going to use the input text field for data upload. Select the ‘Input Box’ option in the ‘how’ dropdown box and select ‘personal track’ in the ‘Where’ dropdown.
3. The input box allows you to paste a bunch of genes to upload as a category for use in analyses in R2. In the example a set of genes, mutated in tumors are pasted. Selecting the ‘personal track’ option guarantees that this set will remain available for you (for people in your community, see later in this tutorial). Click “next” to upload the set (Figure 7), you’ll receive a message when everything has succeeded. Your set of genes is now available as a Category for all analyses within R2. Go back to the main page to see where you can use this set.

4. We’re going to lookup your category, an example is available in the GeneSet view. In the main menu in Field 3 select ‘View a Geneset’ and click “next” (Figure 8)
5. In the GeneSetView your Category is privately available for yourself for similar analyses as with any other public geneset present in R2. Select My GeneCategories to choose from your categories (currently only one of course) #: and click Next (Figure 9).

6. In the next window you can specify which geneset you want to view. The Category ‘Changed Genes’ we just made above is now available (Figure 10). For now we end here, later on we’ll see the category again in the context of #:Tracks.

Figure 8: Using a category; select View Geneset

Figure 9: Selecting your genecategories
Figure 10: Your geneset is available as category.

7. We now return to the My Settings menu to find out how we can manage the categories we just build. From the My Settings menu hover over ‘Categories’ and click ‘Manage Categories’.

Figure 11: The Category Manager

8. The Category Manager (Figure 11) allows you to keep track of the categories; sets of genes you’ve stored within R2. Existing Categories can be adapted or deleted. New Categories can be based on existing ones, and to keep track of large amounts of categories these can be stored in collections. As an example we’re going to update the Category we just made. Click the ‘Copy/Delete/Rename/Edit Categories’ button.
9. In the next screen you’ll be asked what type of procedure is needed for your category of choice; in our case we want to keep the category and update it. Select it and click ‘Execute’.
Figure 13: The Category we just built can be adapted in all details.

10. The details of the Category we built in the former steps are available for adaptation. In this way you can keep track and adapt the sets you use for your analyses

22.5 Step 4: Tracks in R2: create your own data annotation

1. Another important feature in R2 that can be adapted to your needs is the Tracks feature. Tracks in R2 give you the opportunity to divide your samples in groups for further analysis. They can be created as a result of analyses within R2 and stored, or you can create them yourself based on sample characteristics. We’ll first start with a K-means analysis that results in a division of the samples in two groups, for more about this analysis see the available tutorials on this subject). On the main page of R2 select the K-means analysis in Field 3 (Figure 14)

Figure 14: Selecting a K-means analysis

2. In the settings window for the K-means analysis (Figure 15) you can choose the Category created above to cluster the current set of samples. In our case this is called stud06-ChangedGenes. Set the number of draws to 10x10, click ‘next’
Figure 15: Settings for K-means; the Category built above is available for clustering

3. The resulting clustering in two groups might not be ultimately convincing (Figure 16), but for our testing purposes this is alright. What is important is that the resulting groups can be stored as a track; click the button ‘store as a track’.
4. R2 now shows all samples as a long table with radio buttons indicating which group each sample belongs to. These can be adapted if you want to. Scroll down the window to find the fields that have to be set in order to store this as a track (Figure 17). You may want to change the group names into something more informative, and potentially also change the name to something you could easily relate to.
Figure 17: Storing the current groups as Track; Track name has been adapted, the track will be stored in our personal track database.

5. Clicking the Build set button will store this as a track. In the custom tracks manager we can adapt this track again. From the My Settings menu select ‘Manage Custom Tracks’ (Figure 18).
6. In the next screen keep the default selection; your current dataset. Tracks are, of course, defined based on a specific dataset; for each dataset you can store your own tracks. Click ‘Continue’.

7. In the next screen you’re able to adapt the Track we just generated. Of interest in here is the option “drawtrack”, which will result in the display of the information underneath the YY-plots. The tracks can also be assigned to collections to make large sets of tracks manageable. We leave the deletion of the track to the imagination of the reader. Now we’ll pay attention to the default tracks for this dataset.
20: The track we just generated can be adapted from here. For a start set the Drawtrack property to ‘yes’; we want to see this track in the graphs we create!

8. Select Manage Default Tracks from the My Settings menu (Figure 21)

9. In the next screen the dataset has to be defined; keep the defaults and click Continue. You’ll end up in the Default Tracks Manager (Figure 22). Basically all annotation provided with this dataset is available as a track, we’ll select the additional annotations of gender and recurrence as visible, they’ll show up in further analyses. Be sure to click the ‘Update Tracks’ button for these changes to take effect. When collections of tracks are used these we’ll show up conveniently as separate groups of tracks under the graph.
22.6 Step 5: Upload your own tracks

1. R2 also allows you to build your own tracks from scratch. You’ll be able to assign each sample to a group of your choice. To illustrate this select ‘Build Custom Track’ in ‘My Settings’ (Figure 21). The Custom Track window appears. R2 also provides the possibility to upload a custom track from a prefab text file; we’ll shortly show this; click ‘Upload a Track’ (Figure 23).
2. In the Upload Custom Track window you can either select a tab delimited txt file built with a tool like Excel, or alternatively paste tab or semicolon delimited text in the paste-box; that provides R2 with the proper assignment of each sample to a specific value. Based on that, R2 creates the groups for you. Using this procedure you can create tracks with as many groups as you like. If you intend to create a track with a limited number of groups, then an easier way is provided through the user interface, which we will now try. Click back button in your browser to return to Figure 23.

3. In the Custom Track window(Figure 23) the number of groups desired has to be given in advance. We try 3. The Gene/PS option allows you to order the samples based on the expression of a gene, and make sub-groups that range in expression. Leave that field open and click Submit.

4. In the next window a convenient grouping of all annotation parameters and there values is available to create groups. In this example we select the low grade(1,2,3) vs high grade(4) vs special (4s) tumor types in the Neuroblastoma dataset as known from the INSS classification. Tick the appropriate boxes in the appropriate group columns. It is also convenient to recapitulate the resulting groups in a separate column so tick that box also (Figure 25).

Figure 23: The Custom Track dataset choice window.

Figure 24: Uploading a track described in a text file; for each sample a description has to be provided.
Figure 25: In the inss row, the stage 1-2-3 tumors are selected to form group 1, stage 4 forms group 2 and stage 4s group 3 in a new track.

5. Click “next”, all samples appear in a table with check boxes to assign them to the appropriate group. Scroll down to create the annotation for these groups (). Names have been adapted, show track is set to yes, the track is set to be stored a personal track and colors per group have been adapted. Click ‘Build set’ to store the set, you’ll receive a message accordingly in the next window. Of course you now finally want to see all our track manipulation in an actual graph. Go to the R2 main page again to see how the data of a gene will be plotted using the new tracks.
Figure 26: Setting the custom track properties.

1. A frequently used approach is using a track based on bins of values. To avoid labour intensive excel usage you can also use the expression threshold option from the pulldown menu. Each time an expression level has been entered, a new box is generated for the next value. Click next to save the track.
1. Select View a gene in groups in Field 3 of the main page, type MYCN as gene name in Field 4 (Figure 27). Click ‘next’.

2. The track created in the Custom Track manager is available for selection: u-lowgradvs4vs4s. Note that the other Track created in the K-means analysis, is also available here. Choose it and click ‘Next’.

Figure 27: Creating bins.

Figure 28: Selecting a gene to view in the newly created tracks.
Figure 29: Select the Track created in the Custom track manager; u-lowgradvs4vs4s

3. The different groups we created as part of our track in the previous steps are available for selection. We want to see them all, keep the settings and click ‘Next’.

Figure 30: The groups created in the track are available for selection

4. The expression of MYCN is plotted in the different groups of the track that was created (Figure 30). Note that the other track of mutated genes has a large overlap with the stage 4 group. There is also overlap with the “recurrence_or_progression” Default Track that we set to visible.
5. Another convenient option from the “custom track manager” is the export function which enables you to manipulate your tracks manually outside R2. This could be of use when you want to share tracks with other users or create new custom tracks. One reason you want to use the export function is to fix the ordering of your sample when generating a heatmap (see Error: Reference source not found). Make sure you already have a personal custom track (Not a temporary track, 24h). Select Manage default Tracks from the My Settings menu (Figure 21) and click next. Here select the dataset of interest, only datasets which have a corresponding personalized track are represented in the pulldown menu. Click the “Copy/delete/ rename/export Tracks” button. Here select the personal track, “export” operation and r2_track at “export as”. Click execute” and download link with the track name can be loaded by clicking the right mouse button.

22.7 Step 6: Cooperate through R2: sharing tracks, creating communities

1. Cooperation is of great importance in scientific research. You probably want to share the tracks created above with other people in your group. For this reason R2 features the Communities. Creating a community is done by clicking ‘Community’ in the ‘My Settings’ menu (Figure 31).
2. Since this will be the first time in the community section, there are no communities yet; click the ‘Start a new Community’ link (Figure 32).

3. In the Community window a name has to be set and a short description for people invited as members for this group (Figure 33). Through a community you can share GeneCategories, Tracks and Settings.
4. Click ‘Next’; you’ll be notified that the group has been created; return to the Communities Center by clicking the Community link again in the My Settings menu (Figure 32). The TestGroup has been created (next to the already existing MyTestGroup for this user). Click the link to start adding users (Figure 35).

5. You have to add users by their R2 username; we’ll add user “pietmolenaar”. He’ll receive a message in the R2 startup page as soon as he logs on the next time. Click “next” to add the user.

6. R2 returns with a message that the user has been invited, he or she has to accept your invitation first before he will see what you are sharing.

7. The perspective of the invited user after logon; he or she can accept the invitation (Figure 37).
Figure 38: The invited user receives a notification on the main page where he or she can accept the membership of the group.

8. When the invitation has been accepted the user is available in this community. When we add a category, track or preset the next time, it is possible to make this available to this community.

You are invited to join TestGroup. accept / decline

9. When a Category is created there is now a possibility to make it available to a Community (Figure 39).

10. Managing the tracks, gene categories and megasampler presets is done in a similar way as has been shown in the user tracks and user categories at the beginning of this tutorial. Pietmolenaar, as a member of this group, can manage the tracks that have been shared with him via his default track manager.

Figure 39: The user is available in the TestGroup.
Figure 40: As an example here the creation of a category and the assignment to a Community.

22.8 Final remarks / future directions

Some of these functionalities have been developed recently. If you run into any quirks or annoyances don’t hesitate to contact r2 support (r2-support@amc.uva.nl).

We hope that this tutorial has been helpful, the R2 support team.
Exporting data

Export (filtered) normalized data for additional analysis outside of R2

23.1 Scope

- Export gene expression data with the ‘Data Grabber’ functionality directly from the main menu of R2 (http://r2platform.com / http://r2.amc.nl)
- Export focussed gene expression data in R2 after an analysis (result).

23.2 Step 1: Using Data Grabber

1. In the main menu select ‘Tools’ > ‘Data Grabber’. A dropdown menu appears from where you can select the dataset of interest and click ‘next’.

![Data Grabber Menu]

Figure 1: In the main menu “Data Grabber”

2. A settings menu appears where several filter options can be applied to the data you want to export. You may optionally select a track (subset) to filter the samples by the annotated groups; in this example select “inss (cat)” and select one or more stage(s). Be sure to click the red ‘confirm’ link to enforce your selection before proceeding.
3. In the ‘reporters’ section, by default, a specific set of reporters (either reporter names or genesymbols) can be selected via copy/paste a set of genes in the “input_identifiers” box. Another option for the reporter selection would be the ‘HugoOnce’, where only a single reporter is chosen for all of the genes annotated within the dataset, and where orphan reporters are omitted. In this example we are interested to perform an additional analysis outside R2 with all reporters for several stages. In the menu select “HugoOnce” and click ‘next’.

A “datagrabber.txt” hyperlink is generated on the fly and appears on your screen. Right-click the link and store the tab delimited file with expression data on your hard drive. In case a dataset contains more than 60,000 reporters and/or exceeds 300mb in size, then R2 only allows up to 60,000 reporters to be exported. This restriction is imposed due to the large file-sizes which would have to be created. You can always drop r2-support a note, if you require these large datasets as a whole. We will then dispatch those via other means (via a file transfer service).

23.3 Step 2: Other formats: TMEV, tab separated, etc

The example listed above shows that R2 facilitates the export of data without performing any analysis. Of course you may want to do the same after analyzing data in R2. In most of the analyses options you can export the data via the right menu.

1. To illustrate this option. Select in the main screen in box 3 “Find Correlated genes with a single gene”.
2. Type “MYCN” in box and click ‘next’, in the next screen leave all the selection criteria at their default settings and click ‘next’.
3. A list of genes significantly correlating (up and down) with the MYCN gene is generated. The right menu provides you with a set of options to continue your analysis including exporting data.
4. Click on “MakeMeATable (TMEV) ready” in a new (tab) screen you can download the table matrix with the corresponding annotation by using the right click of your mouse. The generated table can be used directly in external programs such as the commonly used clustering program as the TIGR Multi experiment viewer (TMEV) which is freely available.

5. Try also the other options listed in this menu, ‘save current …’, ‘reference for …’ and “Store results …”. These are all different formats to export your results for use outside R2 or store in R2 to continue your analyses at a later time-point.

Did you know that you can export data from different types of modules?

Using a different module such as “Time Series” also provides the option to export the results of use outside R2 or at a later time point within R2.

Note: If the requested file size exceeds ~200MB, R2 will terminate the request because of the system load. Of course there is a workaround by mailing the r2-support@amc.nl and request for the dataset expression values of interest.

23.4 Final remarks / future directions

We hope that this tutorial has been helpful, the R2 support team.
CHAPTER 24

R2 Dataset Addition

How to add your own or publicly available datasets for analysis in R2

24.1 Scope

• Learn how to add your own datasets to R2
• Have datasets added that are published in literature.

24.2 What to prepare when you would like to have a dataset added

R2 allows users / groups to have their own (private) human / mouse datasets added to the platform, which enable
them to analyze their own data from anywhere in the world, as long as they have access to the internet. Such
datasets can be added with various access policies, ranging from public to restricted to a single user. The current
document describes all that you need to know about the addition of datasets in R2, and will also show you how
files should be prepared.

24.3 Who can add datasets to R2

Most often, users would like to analyze their data, in combination with datasets already present in the R2 database.
For example, one would like to compare the extend of expression, compared to other tissues (the so called
MegaSampler). One can imagine that such analyses require identical processing for all datasets of the same
platform (provided that the normalization scheme supports this). For that reason, we have decided that only ad-
ministrators of the R2 platform can add new datasets, since they understand the R2 platform architecture and can
supervise / guide the upload procedure.

24.4 Addition of a public dataset from the GEO database

Having a public dataset added to R2 from the NCBI GEO database is by far the easiest. Depending a bit on
whether the array platform is already used within R2, such datasets can be added fairly quickly. In most cases
you only have to send an email to r2-support@amc.uva.nl stating the GEO series identifier GSE***** and one
of the administrators will take care of the rest. One note for your consideration though. Since the R2 platform has initially been designed to work with Affymetrix microarrays, it works best with single channel platforms. We do also incorporate dual channel datasets, but always warn for potentially unwanted behavior. For the addition of Affymetrix datasets, we preferably would like to be able to work with the raw data (CEL files), so please make sure that those are attached in the GEO series.

The GEO database can be browsed from http://www.ncbi.nlm.nih.gov/geo/browse/?view=series

### 24.5 Addition of personal datasets

Within R2, we also house datasets that are provided by authors which are not (yet) available in the public domain. In some cases these are made publicly available, but most often, some form of restricted access is enforced on them.

### 24.6 Access levels

R2 can provide access to datasets on a number of levels. The default access model provides public access to a dataset to anybody using R2. Within R2, datasets can also be made accessible to a group of users. Such a construction is ideal for departments where more people want to make use of the information. Also consortia can make use of this access model. Finally, datasets can also be made accessible to single users. In all of the cases where restricted access is involved, users should create a (free) personal account for R2, and be granted access by an R2 administrator. Requests for access to a group should be send by the owner of a group, or such an owner should at least be cc-ed in the email correspondence. Additional owners of a group or transfer of the ownership to another person can be done achieved by email to r2-support@amc.uva.nl from the current owner. If your research group already has data in R2, then you should already know the name of your user-group.

### 24.7 Preparing the expression data

When the data to be uploaded originates from Affymetrix gene expression platforms, and you would like to have the data added in the standard way, then we prefer to have the original CEL files send to us by www.wetransfer.com or a similar service. We can then add the dataset in such a way that it can be used in conjunction with the publicly available datasets from the same platform and normalization scheme. If normalization schemes, other than the mainstream MAS5.0, RMA, gcRMA, or RMA-sketch are preferred, then you should perform the normalization yourself and send us the normalized data as a matrix (like a tab delimited export from an Excel sheet). A data matrix should be constructed in the following manner: the first row is considered to be the header and should preferably start with the following sign #H: for the 1\(^{st}\) column. The 1\(^{st}\) column should contain the reporter IDs of the platform (such as probe sets, in the case of Affymetrix). If you are not sure which column represents the reporters that can be used in R2, then please ask by email (Especially Illumina arrays contain a number of fields that look usable). The signal values (preferably non-transformed) of the samples can then be added in subsequent columns, where the 1\(^{st}\) row should contain the sample identifier (that needs to be identical to the sample identifier in the annotation file). R2 can perform a range of transformations (such as log2) on the data itself, thereby allowing the most flexible use of your expression data, if provided in a non-transformed fashion. The figure below shows an example for an Affymetrix array.

![Figure1: Example](image-url)
A number of platforms and/or normalizations not only provide a signal intensity, but can also express the likelihood that a reporter is considered expressed (like present calls or detection p-values for Affymetrix U** platforms). Such information may be provided in the matrix file by addition of an additional column (then named sample_pval). Alternatively, every sample may be provided in a separate tab delimited text document, where multiple columns can then be provided for a sample.

### 24.8 Preparing the gene annotation

If a platform has already been added to the R2 database, then there is no need to supply gene annotation again. If a new platform has to be added, then we require at the very least a list of all the reporters, together with their mapping to the genome (for human preferably HG18/NCBI36; for mouse preferably mm9). Furthermore, the relation between reporters and genesymbols as well as gene ids (NCBI Gene) would speed up the process of adding a new platform. In many cases, vendors of the arrays make annotation files available for download. Usually a link to such an annotation will also be sufficient. In case of doubt, please contact us by email.

### 24.9 Preparing the sample annotation

Expression data is not very useful without proper annotation. Annotation is provided in a separate tab delimited text file. Here the 1^st^ column contains the sample names and any subsequent column is treated as an annotation field (termed tracks within R2). Please refrain from using special characters within the annotation. Also, spaces in track naming should be avoided (use _ instead). An example of an annotation file can be seen in the image below. You can add as many tracks as you like / find useful. There are a number of special tracks, which you can make use of, which will now be described.

![Example2](image_url)

**Figure 2: Example2**

Besides providing the annotation for usage in R2, you can also specify how R2 makes use of these annotations, specifically in graphical representations. To make this known, you can prepare a “relate” file for R2. This document comprises of a number of columns that can be provided for the different tracks. Below, you can see a section of such a relate file.

![Example3](image_url)

**Figure 3: Example3**

Please make sure that the header of the relate file is identical to the example, and that the tracknames match to the ones that have been defined in the sample annotation. The “ISTRACK” column tells R2 whether the annotation needs to be drawn as color coded information below YY-plots, and headers of heatmaps. The “ISINFO” column defines whether the information is displayed in the table once you hover over a sample in graphs within R2. “VISIBLE” can enable/disable the use of a track. The “COLOR” column can preset a specific color to groups which are defined within a track. These can be indicated by groupname:hexcolor. The different groups are then separated by the “;” sign. It is not required to supply this information. R2 will color groupnames automatically if such information information is not encountered. Finally, you may describe the contents of a track.
24.9.1 Special sample annotation

Within R2, some well-defined sample annotation labels and/or additional files, enable additional functionalities. Below, an example is described for both.

**Survival information:** When the dataset contains survival information, then R2 can make use of this to draw Kaplan Meier plots. To do so, R2 requires a separate tab delimited text document with a strictly defined header and some rules. Any line starting with # is considered comments, and will be excluded. There is 1 exception to this rule, which if formed by the #H: combination, which will be interpreted as a header row. A survival file should contain a header line that is identical to the example given below, as R2 will then recognize it as such. How an event is defined, may differ (overall / relapsefree etc). This can be expressed in the name of the file that is being provided. For example, the file below would be named “overall.txt”. Subsequent Kaplan curves would get the name “overall survival” on the y-axis.

**Time series graphs:** When the samples are annotated with the appropriate tracks, then R2 can also present datasets as time series. When R2 encounters a column named “r2_ts_timepoint”, combined with either “r2_ts_profile” and/or “r2_ts_series”, then this will enable the option to represent the dataset as a time series (where samples/groups are connected by a line following the time variable). Profiles are intended to connect a single experiment or the following of a single subject in time. Series are intended as the grouping of profiles (for example biological replicates of an experiment), which will also create error bars on the measurements. The “r2_ts_timepoint” annotation should only contain numerical information (the time, in whatever scale you prefer (minutes / hours / days)). The other 1 or 2 annotations should provide a grouping label (which would be useful for you). In case of doubt on the usage of these annotations, do not hesitate to get in contact with us via r2-support.

24.10 Describing your dataset

Within R2, your dataset will get a name, so that you can find it back for analyses. For dataset naming the program makes use of a small number of parts (some of which can be influenced by you). For example, the department of oncogenomics has made its Neuroblastoma dataset available in R2 with the following name “Tumor Neuroblastoma public - Versteeg - 88 - MAS5.0 - u133p2”. The naming is achieved by the following parts:

1. **Dataset Class.** For Human datasets, one can choose from the classes defined below in italic:
   - **Cellline:** Usually used for cell line panels, where no intervention was applied
   - **Disease:**Datasets, where a specific disease has been investigated, other than cancer
   - **Exp:** Experiment datasets. Usually cell line models in which interventions have been applied (Gene transfection, rna interference etc)
   - **Mixed:** If a dataset makes used of multiple items, then it becomes a mixed set
   - **Normal:** The profiling of healthy normal material
• **Tumor**: Datasets which are composed of a specific tumor type belong in this category

2. **Tissue.** Depending a little on the choice of class, usually a description of the tissue / tumor is given in the second part. In the example, this was “Neuroblastoma”, but this could also be “Breast” or “Colon” if such a dataset was described. For experiments, the tissue or tumor type is also often described, to make sure that datasets with the same theme are close together. If we would describe the shRNA knockdown of the MYCN gene in the neuroblastoma cell line IMR32 for example, then this would become “Exp Neuroblastoma IMR32 MYCN shRNA”.

3. **Author.** Finally, you can supply the author / consortium in naming your dataset. This should be self-explanatory.

R2 will add the number of samples within the dataset, a normalization scheme and finally also a code representing the platform which has been used. If you are supplying a dataset other than Affymetrix gene expression arrays (Exon or U***), then you should provide us with the normalization, and platform used as well. The platform, doesn’t have to be the code that R2 uses, but can also be described by the manufacturer and the array ID.

Optionally, you can also describe your dataset in more detail in the following fields (which are also shown if you click on the “i” image next to a dataset). **Title**: 1 line description of your dataset. **Summary**: free text option to describe your dataset in as much detail as you wish (See also GEO for examples). **Design**: free text describing the design of your dataset (See also GEO for examples).

We hope that this document has been helpful in preparing your dataset for inclusion in R2,

R2 support (r2-support@amc.uva.nl).