

R2 Tutorials

Release 3.3.4

The R2 support team

Apr 08, 2024

Contents

1	Preface	3
2	Using Datasets	5
2.1	Scope	5
2.2	Step 1: Selecting a dataset	6
2.3	Step 2: Advanced selection of datasets	9
2.4	Step 3: Using Dataset favorites	10
2.5	Step 4: Data Scopes	11
2.6	Final remarks / future directions	12
3	One Gene View	13
3.1	Scope	13
3.2	Step 1: Select the View a Gene module	13
3.3	Step 2: Select the gene or reporter	14
3.4	Step 3: Plotting Gene expression	15
3.5	Step 4: Selecting analysis types: View a gene in groups	16
3.6	Step 5: Marking / highlighting samples within a plot	18
3.7	Step 6: Sources for additional information on the selected gene	24
3.8	Step 7: Adapting a plot	25
3.9	Step 8: Selecting subsets	28
3.10	Step 9: Find best track separation with CliniSnitch	29
3.11	Step 10: Finding sample extremes.	30
3.12	Step 11: Probeset verification	31
3.13	Final remarks / future directions	33
4	Multiple Genes View	35
4.1	Scope	35
4.2	Step 1: Viewing multiple genes	35
4.3	Step 2: Viewing multiple genes through track annotation	36
4.4	Step 3: View multiple genes (Bubble plot)	38
4.5	Final remarks / future directions	42
5	Annotation analyses	43
5.1	Scope	43
5.2	Step 1: Relating 2 (categorical) tracks	43
5.3	Step 2: Relating 2 (numerical) tracks	46
5.4	Step 3: Relating a categorical track to a numerical track	46
5.5	Step 4: Annotation plotter and Cohort Overview	47
5.6	Final remarks / future directions	49
6	Differential expression of genes in your dataset	51

6.1	Scope	51
6.2	Step 1: Selecting data and the type of analysis	51
6.3	Step 2: Choose the gene and the annotation track as grouping variable	52
6.4	Step 3: Anova results / adapting plots	52
6.5	Step 4: Finding differentially expressed genes in two groups	56
6.6	Step 5 Setting parameters	57
6.7	Step 5: Correct for paired analysis	59
6.8	Step 6: Find differential expression in multiple groups	62
6.9	Step 7: Inspecting single genes	62
6.10	Step 8: Plot all genes and adapt visualization: Volcano plot etc	63
6.11	Step 9: Using the Enrichr	65
6.12	Final remarks / future directions	66
7	Find genes correlating with your gene of interest	67
7.1	Scope	67
7.2	Step 1: Selecting data	67
7.3	Step 2: Inspecting correlating genes	68
7.4	Step 3: Inspecting correlation between specific genes	69
7.5	Step 4: Relation with Chromosome position	76
7.6	Step 5: Establishing overrepresentation in other domains	78
7.7	Step 7: Gene list in pathway context	79
7.8	Step 8: Further pathways analysis	81
7.9	Step 9: Gene set analysis	82
7.10	Final remarks / future directions	83
8	Working with Kaplan Meier	85
8.1	Scope	85
8.2	Step 1: Selecting the Kaplan Meier module	85
8.3	Step 2: Kaplan Meier by gene expression; the Kaplan Scan	88
8.4	Step 3: Kaplan scan for a group of genes	89
8.5	Step 4: Kaplan scan on your own cohort	92
8.6	Step 4: Cox Regression analysis and hazard ratio	94
8.7	Final remarks / future directions	98
9	Pathway Finder	99
9.1	Scope	99
9.2	Step 1: Selecting data	99
9.3	Step 2: Correlating pathways with a gene	100
9.4	Step 3: Finding pathways relevant to subgroups	102
9.5	Step 4: Determining differentially expressed pathways	103
9.6	Step 5: Verifying a pathway	103
9.7	Step 6: Correlating with the expression of a gene	104
9.8	Final remarks / future directions	104
10	Multiple datasets overview with Megasampler	105
10.1	Scope	105
10.2	Step 1: Selecting multiple datasets	105
10.3	Step 2: Viewing a gene in multiple datasets	108
10.4	Step 3: Stacking subgroups (or datasets)	111
10.5	Step 4: Expression distribution over many datasets	113
10.6	Step 5: Megasearch	113
10.7	Final remarks / future directions	117
11	K-means clustering in R2	119
11.1	Scope	119
11.2	Step 1: Selecting data and module	119
11.3	Step 2: Adapting settings	120
11.4	Step 3: Examining resulting clusters	121
11.5	Step 4: Creating consistent clusters	123

11.6	Final remarks / future directions	125
12	Using signatures	127
12.1	Scope	127
12.2	Step 1: Creating a geneset signature, a Track within R2	128
12.3	Step 2: Determine the activity of a signature	130
12.4	Step 3: Using signature scores	132
12.5	Step 4: Plot signature scores using the relate 2-tracks module.	134
12.6	Step 5: Drawing lines between samples in a XY plot	136
12.7	Step 6: Signature Gene correlations	137
12.8	Final remarks / future directions	137
13	Analysing Time Series	139
13.1	Scope	139
13.2	Step 1: Choosing the time series module and data	139
13.3	Step 2: Finding regulated genes in a time series experiment	142
13.4	Step 3: Using the regulated genes in further analyses	145
13.5	Step 4: Correlate with other datasets	148
13.6	Step 5: In a K-means analysis	149
13.7	Final remarks / future directions	150
14	Using genesets and creating heatmaps in R2	151
14.1	Scope	151
14.2	Step 1: Selecting data and modules; creating a Heatmap	151
14.3	Step 2: Using multiple GeneSets	153
14.4	Step 3: Relating genesets with data annotation	154
14.5	Step 4: Unsupervised hierarchical clustering with a geneset	156
14.6	Final remarks / future directions	159
15	Principle Components Analysis in R2	161
15.1	Scope	161
15.2	Step 1: Selecting data and modules	161
15.3	Step 2: Exploring the principle components	162
15.4	Step 3: Viewing clusters in 3D	164
15.5	Final remarks / future directions	164
16	Sample maps: t-SNE / UMAP, high dimensionality reduction in R2	167
16.1	Scope	167
16.2	Step 1: Selecting t-SNE maps	168
16.3	Step 2: Annotating t-SNE maps	168
16.4	Step 3: Perplexity sweeps for t-SNE maps	171
16.5	Step 4: Creating t-SNE maps	172
16.6	Step 5: Creating groups with the t-SNE lasso tool	174
16.7	Step 6: Creating groups with the t-SNE DBSCAN tool	176
16.8	Final remarks	177
17	Using the R2-Genome browser	179
17.1	Scope	179
17.2	Step 1: Exploring the genome browser	179
17.3	Step 2: Zooming and panning	182
17.4	Step 3: Looking up chromosome regions	184
17.5	Step 4: Working with multiple samples listed within a track	184
17.6	Final remarks / future directions	184
18	DataScopes	185
18.1	Scope	185
18.2	Step 1: Selecting a dataset restrictive DataScope	185
18.3	Step 2: Selecting a Data Scope with a landing page	186
18.4	Final remarks / future directions	188

19	Integrative analysis: ChIP-seq data	189
19.1	Scope	189
19.2	Some concepts	189
19.3	Step 1: Choosing data and modules	192
19.4	Step 2: Exploring genes in a transcriptional context	193
19.5	Step 3: Exploring histone modification patterns	197
19.6	Step 4: Finding active super-enhancers	198
19.7	Final remarks	200
20	Integrative Analysis : Across Platforms	201
20.1	Scope	201
20.2	Step 1: Choosing a combined dataset	201
20.3	Step 2: Correlate two datatypes	205
21	Integrative Analysis : WGS/NGS data	211
21.1	Scope	211
21.2	Step 1: View circos files.	211
22	Target Actionability Literature Reviews : TAR	217
22.1	Scope	217
22.2	Step 1: View a TAR.	217
22.3	Final remarks / future directions	219
23	Adapting R2 to your needs	221
23.1	Scope	221
23.2	Step 1: Adapt your settings	221
23.3	Step 2: How to add data to R2.	222
23.4	Step 3: Create your custom genesets	222
23.5	Step 4: Tracks in R2: create your own data annotation	226
23.6	Step 5: Upload your own tracks	233
23.7	Step 6: Cooperate through R2: sharing tracks, creating communities	237
23.8	Final remarks / future directions	242
24	Exporting data	243
24.1	Scope	243
24.2	Step 1: Using Data Grabber	243
24.3	Step 2: Other formats: TMEV, tab separated, etc	245
24.4	Final remarks / future directions	245
25	R2 Dataset Addition	247
25.1	Scope	247
25.2	What to prepare when you would like to have a dataset added	247
25.3	Who can add datasets to R2	247
25.4	Addition of a public dataset from the GEO database	247
25.5	Addition of personal datasets	248
25.6	Access levels	248
25.7	Preparing the expression data	248
25.8	Preparing the gene annotation	249
25.9	Preparing the sample annotation	249
25.10	Describing your dataset	250
26	Graphs: Adjustable Settings menu versus Repsonsive Settings	253
26.1	Example settings: color the sample maps	253
27	Concepts of R2: did you know..?	255
27.1	R and p-values	255
27.2	Statistical tests	256
27.3	Settings for analyses	257
27.4	Core concepts of R2	259

Dear reader. Welcome to the tutorials for ‘R2: Genomics Analysis and Visualization Platform’. The R2 platform is a (molecular) biologist friendly, web based genomics analysis and visualization application developed by Jan Koster and his team at the department of CEMM in the Amsterdam University Medical Centers (AUMC), location Academic Medical Center (AMC) in Amsterdam, the Netherlands. The idea behind the application is to enable researchers that are not experts in bioinformatics to mine and explore omics data sources, and thereby gain insights for their research. This can be done on publicly available data sets, but also on your own data in a restricted environment. The tutorials have been assembled as guided short stories, that will instruct you how to get things done in the platform by an example. We hope that, by following our examples, you will get familiar with the concepts of R2, and thereby find your way in the platform. Even though many of our examples are illustrated by a neuroblastoma pediatric cancer dataset, R2 has many more (2100+) public datasets to work with, covering nearly any cancer but also other diseases as well as normal reference series (See also Selecting datasets).

If you make use of our platform in manuscripts, then please add a citation that includes the following website: ‘R2: Genomics Analysis and Visualization Platform (<https://r2.amc.nl> <http://r2platform.com>)’. This will help us get the necessary funds to keep on going and in addition allows us to ‘scan’ the literature to keep track of manuscripts that cite our resource.

Copyright (c) 2006-2023 R2 Support Team

Why these tutorials?

The R2 platform (<http://r2platform.com> or <http://r2.amc.nl>) is a genomics analysis and visualization platform that provides a biologist/bio-medical researcher friendly interface to high throughput data. It has initially 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. Currently, we continue the development of R2 in our own group within the department of CEMM. 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 interactive and inter connected tools and visualizations to mine the database.

Even though many people appreciate the concept of R2, getting started with the platform as a new user can be a bit difficult or intimidating with the wide range of options that we provide. With this tutorial book, we hope to help new users getting 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:



Did you know box

These provide additional information not directly related to the tutorial steps

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.

If you don't feel like reading the tutorials, then have a look at our YouTube channel, where team members of the R2 platform demonstrate many functionalities in short videos.

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](#).

Selecting or searching datasets in R2

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 omic profiles such as expression and methylation profiles for more than ~2.500.000 unique individual samples. The samples are grouped in so-called datasets. Each dataset has its own characteristics, such as tissue type, tumor/disease type, or from cell-line experiments. Frequently, new datasets are added the platform.
- 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 but also other datasets will be used a example datasets. Later on, working with multiple datasets will be discussed.

Data set selection - current: Tumor Colon Metastatic Bevacizumab / Cetuximab - Vincent - 575 - D

Species	Data type	Category	Tissue/Tumor
Select	Select Filter	Select Filter	
hs	Expression data	<div><input checked="" type="checkbox"/> (Select All) <input checked="" type="checkbox"/> Allograft <input checked="" type="checkbox"/> Cell line <input checked="" type="checkbox"/> Disease <input checked="" type="checkbox"/> Exp <input checked="" type="checkbox"/> Knock-out <input checked="" type="checkbox"/> Mixed <input checked="" type="checkbox"/> Normal</div>	Bevacizumab / Cetuximab
hs	Expression data		
hs	Expression data		Encyclopedia
mm	Expression data		ous (Sulindac)
hs	Expression data		a (2022-v32)
hs	Expression data		s One
hs	Expression data	Tumor	Neuroblastoma
hs	Expression data	Tumor	Neuroblastoma
hs	Expression data	Tumor	Neuroblastoma prot coding
mm	Expression data	Allograft	Metfusion OE0290
hs	Expression data	Allograft	Neuroblastoma Auranofin Comb (PDX1andT5_Org)
hs	Expression data	Cell line	ALL study1
<div>Confirm selection</div>			

Figure 1: Select types of datasets

2.2 Step 1: Selecting a dataset

- 1. R2 offers a large number of readily available datasets for analysis and visualization. The numbered boxes on the main page will guide you through all the steps necessary to perform a task, starting with the selection of a dataset. In box 1 select **Single Dataset**, in field 2 click on the name of the dataset.

1,409,285 (1,270,489 unique) samples available

- 1** Choose single or multiple dataset analysis
Single Dataset
- 2** Select a dataset for analysis
Tumor Neuroblastoma public - Versteeg - 88 - MAS5.0 - u133p2
- 3** Select type of analysis
View a Gene
- 4** Proceed

Figure 2: Change Dataset on the main page

- A popup window appears that shows all the currently available datasets in a grid: each row represents one dataset, with its main descriptive details split up in the columns. You can simply scroll through the list with the mouse and get more information about a dataset with a click of the mouse on any dataset in this list. If you click on the row of a dataset, extra information about a dataset will appear in an information panel below the grid box. Possible adjustments of the original data by the R2 team, such as data transformations or annotation changes, can be found in the Adjustments section of the information panel. At the bottom you will see links to the original data source and Pubmed resources, if available.

Species	Data type	Category	Tissue/Tumor	Author	N	Normalization	Platform	Comp.	Accession	Release date	R2 date	Access	Flags
hu	Expression data	Exp	Neuroblastoma (Calix modulation)	Torres	15	rma_oligo	hugobio	bulk	GSE7204	2016-03-30	2016-06-07	public	
hu	Expression data	Aliquot	Kidney	Nomura	224	MAS5.0	u133p2	bulk	GSE147089	2019-11-10	2023-02-21	public	✓
hu	Expression data	Cell line	CCLE Cancer Cell Line Encyclopedia	Broad	917	MAS5.0	u133p2	bulk	GSE36133	2012-03-20	2012-04-10	public	✓
mm	Expression data	Exp	Organoids Adrenomedullary (Subline)	Arora	9	custom	genex36	bulk	GSE81836	2016-05-24	2018-05-15	public	✓
hu	Expression data	Exp	KID (PDX02)	Gutierrez	7	MAS5.0	u133p2	bulk		2000-01-01	2017-10-16	restricted	✓
hu	Expression data	HiSeq	Colon Adenocarcinoma (2022-v10)	Ng	512	gen	genecode36	bulk		2000-01-01	2022-06-03	public	✓
hu	Expression data	Tumor	Medulloblastoma RUS One	Kool	62	MAS5.0	u133p2	bulk	GSE13127	2008-09-02	2009-06-17	public	✓
hu	Expression data	Tumor	Neuroblastoma	Bell	97	limn	enah36v8	bulk		2019-12-01	2019-12-15	restricted	✓
hu	Expression data	Tumor	Neuroblastoma	Versteeg	88	MAS5.0	u133p2	bulk	GSE14476	2000-01-01	2008-03-05	restricted	✓
hu	Expression data	Tumor	Neuroblastoma prot coding	Bell	97	limn	enah36v8	bulk		2019-12-01	2019-12-15	restricted	✓
mm	Expression data	Aliquot	Medulloblastoma OES290	Zuckerman	26	gen	genecode12	bulk		2000-01-01	2020-10-28	restricted	✓
hu	Expression data	Aliquot	Neuroblastoma Auronfin Comb (PDX)and(1,Org)	Nunzi	32	deno2	genecode13	bulk		2000-01-01	2022-06-29	restricted	✓

Title: Tumor Neuroblastoma prot coding - Bell - 97 - limn - enah36v8

Summary:

Design:

Available tracks in R2:

- igf1p1_gened_norm_rn_1: 0 - 1
- igf1p1_gened_norm_rn_2: 0 - 1
- igf1p1_norm_rn_rn_1: 0.00 - 15.95
- igf1p1_norm_rn_rn_2: 0.00 - 9.95
- igf1p1_norm_rn_rn_3: 0.00 - 8.99

Adjustments: added 1

Available on R2 since: 2019-12-15

Platform: enah36v8

Species: hu

Number of samples: 97

Source: author ID: Date: 2019-12-01

Pubmed link:

R2 internal identifier: ps_bell_p191207_enah36v8

Figure 3: Select a dataset by clicking on a row

- Every column header provides options to filter the database, e.g. in the Tissue/Tumor text field a keyword such as **medull** can be written to filter for medulloblastoma datasets or, with the Select Filter dropdown under the Data type header you can request an overview of for instance methylation datasets. If you know specific details, other columns can be of help as well: e.g. you can search for an author, or use the N column to search for a datasets of which you know the number of samples. The grid box will display all the datasets

that fulfill the (combined) filter requirements. In the bottom right corner you find the number of (filtered) datasets that are available to you.

File selection - current: Tumor Neuroblastoma Public One - Kind - G2 - MA55.0 - u133p2

Species	Data type	Category	Y Tissue/Tumor	Y Author	Y N	Y Normalization	Y Platform	Y Comp.	Y Accession	Release da.	Y R2 date	Y Access	Y
Select	Select Filter	Select Filter	Years			Select Filter		Select				Select	
Na	Expression data	Tumor	Neuroblastoma public	Versteeg	88	MA55.0	u133p2	bulk	GSE16476	2012-03-08	2010-06-10	public	
Na	Expression data	Cell line	Neuroblastoma (A0)	Versteeg	34	MA55.0	u133p2	bulk		2000-01-01	2010-04-20	restricted	
Na	Expression data	Cell line	Neuroblastoma (A0)	Versteeg	35	MA55.0	u133p2	bulk		2000-01-01	2010-06-23	restricted	
Na	Expression data	Cell line	Neuroblastoma (SK-N-BE-2C)	Versteeg	40	MA55.0	u133p2	bulk		2000-01-01	2010-03-24	restricted	
Na	Expression data	Cell line	Neuroblastoma (MPCN status)	Holmes	10	RNA_platex	hugobank	bulk	GSE64596	2015-12-31	2016-06-09	public	
Na	sGSM data	Cell line	Neuroblastoma (genomewide)	Holmes	21	dbioin	genbeding2	bulk		2012-11-09	2012-11-09	public	
Na	sGSM data	Cell line	Neuroblastoma (single)	Holmes	16	dbioin	gen209728	bulk	GSE209728	2012-11-10	2012-12-21	public	
Na	Expression data	Cell line	Neuroblastoma	Bredner	53	MA55.0 spike	affux	bulk	GSE80149	2016-09-01	2016-09-29	public	
Na	Expression data	Cell line	Neuroblastoma	ForCellData	12	RNA	affux	bulk		2015-03-26	2015-03-26	public	
Na	Expression data	Cell line	Neuroblastoma	Seppanathan	38	custom	ionhugobank2	bulk	GSE19274	2009-12-03	2009-12-04	public	
Na	Expression data	Cell line	Neuroblastoma	Hane	41	tdm	ngp01	bulk	GSE89443	2017-03-07	2017-03-07	public	
Na	Expression data	Cell line	Neuroblastoma	Holmes	20	tpm	genomex3	bulk		2012-10-20	2012-10-20	public	

Confirm selection

Tumor Neuroblastoma public - Versteeg - 88 - MA55.0 - u133p2

Title: Integrated bioinformatic and wet-lab approach to identify potential oncogenic networks in neuroblastoma

Summary: mRNA profiles of thousands of human tumors are available, but methods to deduce oncogenic signaling networks from these data lag behind. It is especially challenging to identify main regulatory routes, and to generalize conclusions obtained from experimental models. We designed the network R2 in parallel with a wet-lab approach of neuroblastoma. Here we demonstrate how R2 facilitates an integrated analysis of our neuroblastoma data. Analysis of the MYCN pathway suggested important regulatory connections to the poly(ADP-ribose) synthesis route, the Notch pathway and the BMP/TGF β pathway. A network of genes emerged connecting major oncogenes in neuroblastoma. Genes in the network carried strong prognostic values and were essential for tumor cell survival.

Dataset: 88 human Neuroblastoma samples were analyzed.

Available tracks in R2:

- age_year: 0 - 13
- expression: 1 - 1
- gene: NA, NA
- cl_name_base_cell_line: (F) 0.00055209 - 0.00474963
- cl_name_base_cell_line: (F) 0.01761796 - 0.01474002
- cell_line_base_cell_line: (F) 0.01400000 - 0.01400000

Adjustments: MA55.0 normalization was performed in OGOS with trimmed mean 96 set to 100 (alpha1=0.04, alpha2=0.06)

Available on R2 since: 2010-05-10

Platform: u133p2


Species: Na

Number of samples: 88

Figure 4: Use the search bar to textually filter the list of datasets with a keyword

1. Click on the table row of the dataset of your choice and click the clue colored button **Confirm selection** in order to use a dataset in box 2 of the main page for further analysis.



 **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` - `chiptype`. The dataset selection grid consists of these informative parts as columns, each with filter options to perform an advanced search through the dataset.

R2: Genomics Analysis and Visualization Platform

433417 (385903 unique) samples available

Choose single or multiple dataset analysis

1

2 **Select a dataset for analysis**

Tumor Neuroblastoma public
Versteeg - 88 - MAS5.0 - u133p2

Category **Additional info** **# Samples** **Platform**

Tumor Neuroblastoma public - Versteeg - 88 - MAS5.0 - u133p2

3 **Tissue / Tumor** **Author** **Normalization**

Select Additional Conditions

elecY	SpeciesY	Data type	Y	CategoryY	Tissue/Tumor	Y	Platform	Y	NormalizationY	#	Y	Author	Y	AccessionY	Release dateY	R2 date
	Select	Select Filter		Select				Select Filter	88			versteeg				
Select	hs	Expression data		Tumor	Neuroblastoma public		u133p2		MAS5.0	88		Versteeg		GSE16476	2012-03-08	2010-06

Go to page: 1 Show rows: 10

Figure 5: The informative parts of a dataset name correspond to the columns of the dataset selection grid.

2.3 Step 2: Advanced selection of datasets

1. The grid itself enables the user to search through datasets using keywords and other filter options as well. The column dropdown functions and textboxes can filter the datasets for specific characteristics, e.g. datasets with a minimal number of samples, or a specific author, platform or publication date. You can easily combine the search functions of the different columns.

As an example, we want to see which large sets are available. First we write part of the word Neuroblastoma in the search box of the Tissue/Tumor column. Next, we use the pull down of the N (sample number) column to order the datasets in descending order.

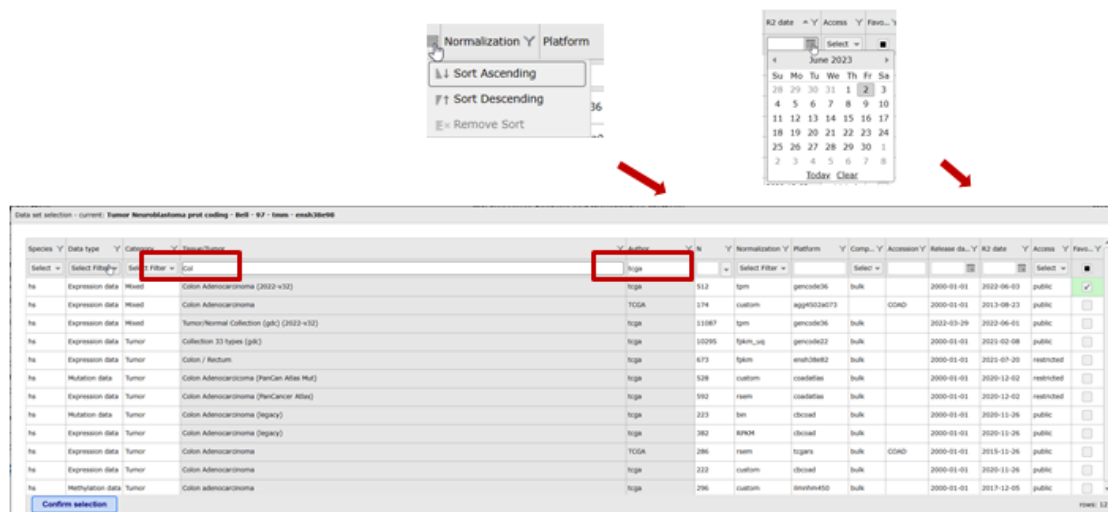


Figure 6: Combine search filters in the grid

2. Again we use the **Confirm Selection** button if we want to continue our analysis with a specific dataset of the grid.
3. Select “Across Datasets” in field 1. Note that in field 2 different options become available compared to the “single dataset” option.

R2: Genomics Analysis and Visualization Platform
726864 (641364 unique) samples available

1 Choose single or multiple dataset analysis
Across Datasets ▼ ⓘ

2 Select an analysis
Select an analysis: MegaSampler (View a gene in more than 1 dataset) ▼ ⓘ

3 Proceed
Next Reset

Figure 7: Selecting across datasets

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

2.4 Step 3: Using Dataset favorites

Since R2 is hosting hundreds of datasets, it could be convenient to store the datasets you often use in a preselection that is easily accessible. In order to maintain favorites, you need to be a registered user. If you did not yet register; accounts can easily be created via ‘Login/Register’ and are absolutely free. Clicking on the dataset name will open the dataset selection grid, where resources can be searched and selected.

3,775,821 (2,658,408 unique) samples available

1 Choose single or multiple dataset analysis
Single Dataset ▼ ⓘ

2 Select a dataset for analysis
Tumor Colon Metastatic Bevacizumab / Cetuximab - Vincent 75 - DESeq2_vst - tpm109neo ▼ ⓘ

3 Select type of analysis
View a Gene ▼ ⓘ

4 Proceed
Next Reset

Figure 8: Change Dataset to access favorites

Within the dataset selection table, you can select and deselect cohorts to add or remove them from your selection of preferred sets. This is done by using the favorite select boxes in the last column on the right side. Favorite datasets will always be represented at the top of your selection table and will be marked with a green background color in the selection column. This makes it very convenient to quickly have access to those.

The screenshot shows a table titled 'Data set selection - current: Tumor Medulloblastoma PLoS One - R2 - MAASL-0 - v133p3'. The table has columns: Species, Data type, Category, Tissue/Tumor, Author, N, Normalization, Platform, Comp., Accession, Release date, R2 date, Access, and Fav... (Favorite). The table lists various datasets, including Medulloblastoma PLoS One, Colon Adenocarcinoma (2022-v10), Neuroblastoma public, etc. Some rows have a green background in the Fav... column, indicating they are favorites.

Figure 09: Managing favorites

2.5 Step 4: Data Scopes

1. R2 can be forced to only display a sub-selection of all the datasets that are available (e.g. only neuroblastoma datasets). These are called data scopes and can be selected from within R2 by the left-hand menu item 'Change Data Scope'. From here you can use one of the preset scopes. This is also the section where you can remove a scope that has been set. An obvious reason why scopes can be convenient, is the focussed view on the available data: to restrict data to a particular subject or as a landing page for a specific publication/subject. Datascope as dedicated landing pages can also be configured to expose additional functionalities and quick jumps to sections in the platform. Just have a look at which ones are accessible given your access rights.
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).

Further details on the use of Datascope can be found in the tutorial Datascope.



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@amsterdamumc.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@amsterdamumc.nl and inquire on the procedure to get your data available in R2 (see also Chapter 22).

2.6 Final remarks / future directions

All the procedures described in this chapter can be executed using the R2: genomics analysis and visualization platform (<http://r2platform.com> / <http://r2.amc.nl>).

If you encounter any quirks or issues, please do not hesitate to contact R2 support at r2-support@amsterdamumc.nl.

We hope that this tutorial has been useful.

Best regards, The R2 support team.

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 that R2 is offering.
- A separate info panel in the one-gene expression level screen provides different types of analyses based on the expression level of the chosen gene.
- Many mRNA expression datasets were generated with Affymetrix profiling arrays and NGS data (rnaseq). In general, the Affymetrix arrays use more than one so-called probeset to measure the expression level of one single gene. With a separate module “Transcript view”, the details of the probesets can be studied. This also holds for multiple RNAseq data in case the chromosomal location of the reporter (Gene) is stored in the R2 database.

3.2 Step 1: Select the View a Gene module

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 and click “Next”.

1,409,285 (1,270,489 unique) samples available

1 Choose single or multiple dataset analysis

Single Dataset ⓘ

2 Select a dataset for analysis

Tumor Neuroblastoma public - Versteeg - 88 - MAS5.0 - u133p2 ⓘ

3 Select type of analysis

View a Gene ⓘ

4 Proceed

Next
Reset

Figure 1: Single gene selection

3.3 Step 2: Select the gene or reporter

1. We will take a look at the expression levels of the samples for the MYCN gene. Type “mycn” in the left “Search by Gene” textfield and click on the first MYCN reporter that shows up in the list of the dropdown. The reporter ID will then be listed in the right “Search by Reporter” box.

In the case of Affymetrix datasets, the term probeset is often used instead of reporter, and more than one probeset can be associated with a gene. As you can see in dropdown list for MYCN in this example, multiple probesets are annotated for the MYCN gene. The term probeset originates from affymetrix arrays, the terms probeset and reporter will be used in this tutorial interchangeably.

By default, *the probeset with the highest average present signal (APS) is annotated as the default probeset in R2*. This APS signal is simply the average of all samples that are considered to express a selected gene (have a present call). After you enter the first letters for the mycn gene in the textfield, you can choose from the available probesets that are listed in a small dropdown. The default R2 probeset will be the first one in the list. 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 by R2 in every dataset, sometimes resulting in a different probeset.

The expression levels of datasets are by default converted to log2 values. This does not apply to datasets that contain ratios or logfolds such as methylation arrays, double labeling arrays, drug data etc etc.

Clicking the *advanced search* button provides a grid where other selection criteria can be applied, such as gene symbol or average signal. Additionally, the sorting option allows for quick checking of genes with a certain expression level. The last column of the grid, named “R2 default”, indicates whether the reporter is set as default in R2 (TRUE) or not (FALSE). This information is not available for each dataset in R2.

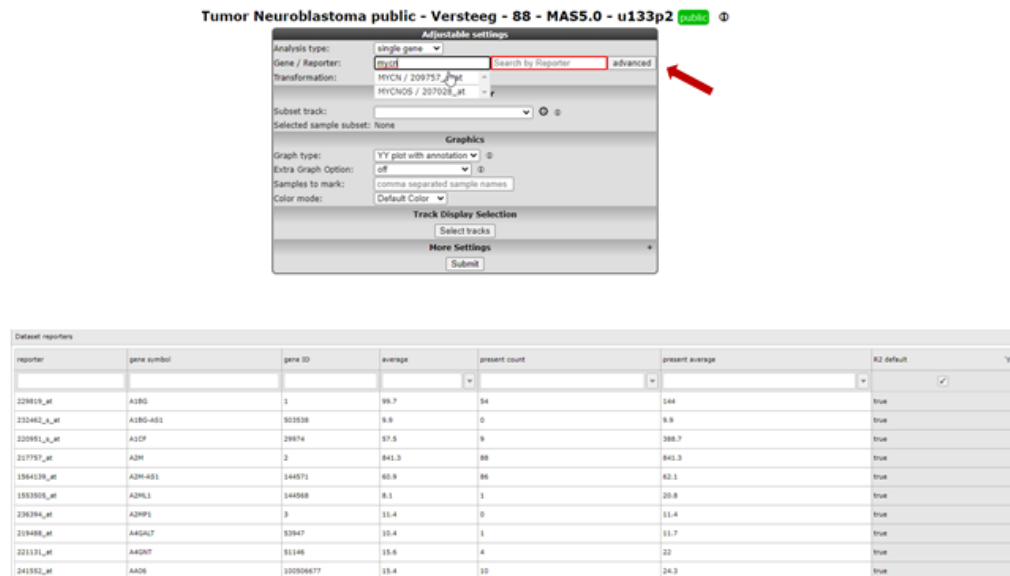


Figure 2: Top, by default the reporter with the highest expression level is selected. Below, the advanced search option with the grid

- To follow the example of this tutorial, use the pre-defined default settings in the rest of the adjustable settings menu, and click 'Next'.

3.4 Step 3: Plotting Gene expression

- 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.

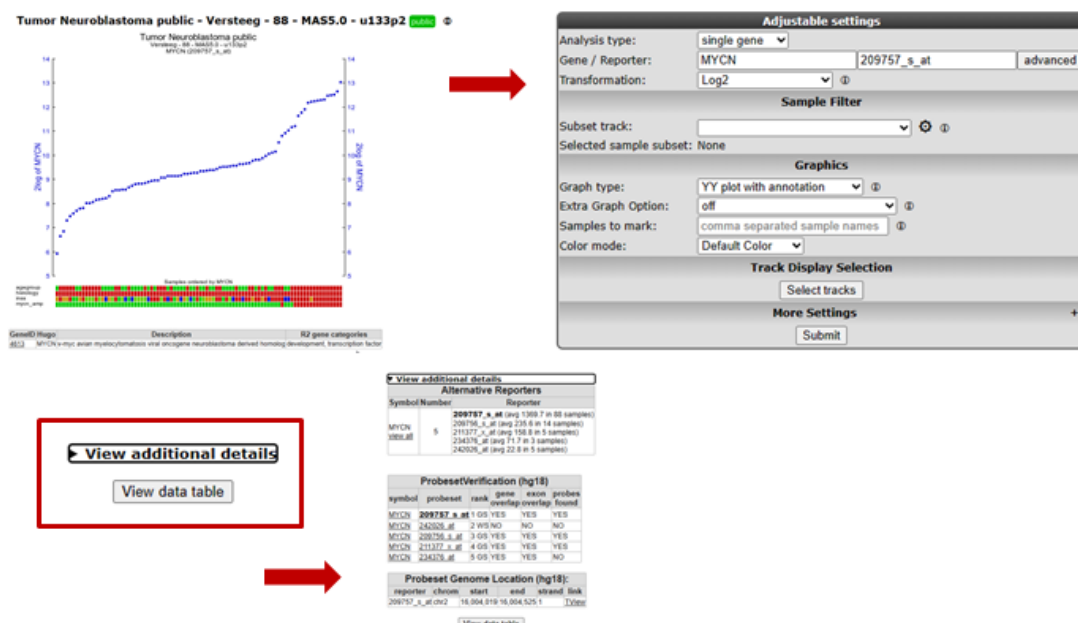


Figure 3: YY plot MYCN expression

- 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 relationship between the expression levels and the tracks for “MYCN amplification” and “INSS-stage”. The display of these tracks underneath the image gives a quick glance at some of the clinical parameters, defined for the dataset. Clicking the triangle at **View additional detail** provides functional information for the reporters and the genome location if available. It is also possible to define your own custom-made tracks, or disable/adapt the settings for default tracks (further explained in the chapter “Adapting R2 to your needs”)

3. Sometimes you get more insight by reviewing the expression levels with other transformations. In order to change the transformation, scroll down to the “Adjustable settings” panel underneath the graph and tracks. In the pulldown menu of the ‘Transformation’ setting (red arrow in Figure 3), choose “none” and then click the button *Submit* at the bottom of the panel.



Did you know that converting expression levels using the “transform” option can help you 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). The MAD is particularly useful in situations where the data may contain outliers or is not normally distributed.
- “center/log2center”: Expression values centered around 0 (on raw values, or log2 transformed values).
- “Rank”: numerical or ordinal values are replaced by their rank when the data are sorted
- “zscore_group”: Coverts 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.
- “Square root”: The square root of a number is the number that gets multiplied to itself to give the product.
- “log2 grouped zscore”: 2log transformed data within each group of a selected track separately and re-merged after calculation

3.5 Step 4: Selecting analysis types: View a gene in groups

In the one gene view plot we have investigated the gene expression of a single gene together with the sample annotation depicted below the graph.

- Thus far, we have been looking at the expression of MYCN ordered by the expression. From the one-gene-view adjustable settings menu there are also other analyses to select. Above the adjustable settings menu you can select different analysis types which are at your disposal as illustrated in Figure 4.

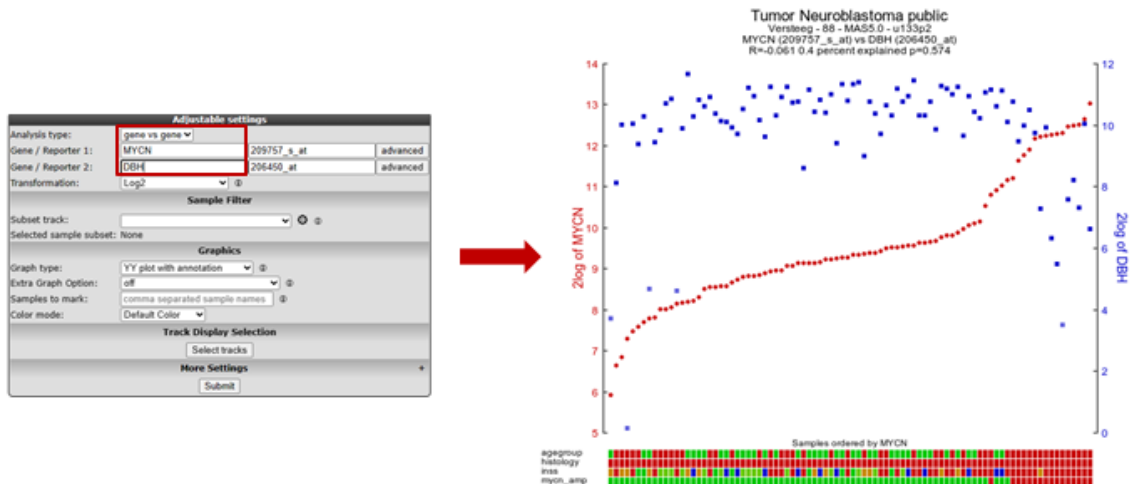


Figure 4: Select other analysis types

- Select in the analysis type menu, the gene vs gene option the adjustable settings will adapt automatically according to the type which has been selected. As illustrated in Figure 4 you simply fill in a different Gene for Gene/Reporter 1 than for Gene/Reporter 2 (red box in Figure 4).

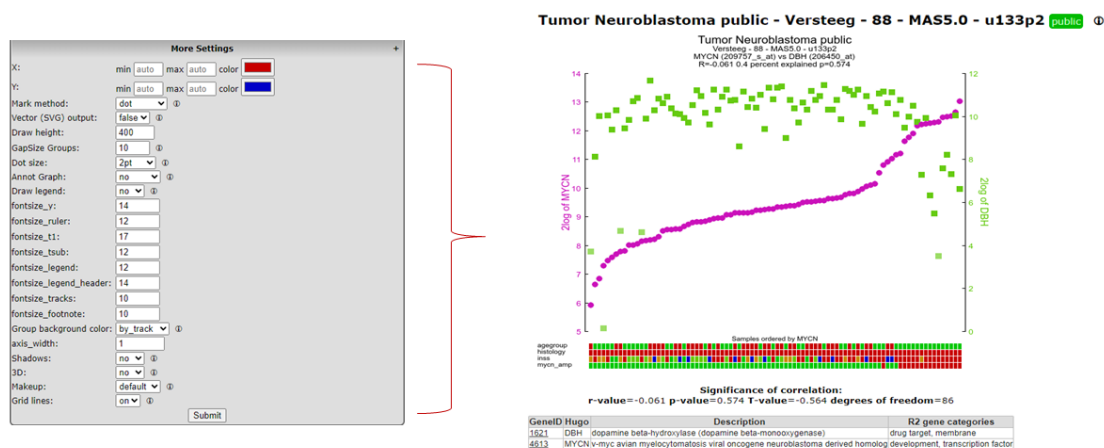


Figure 5: Adjusting the 2 gene plot

3.6 Step 5: Marking / highlighting samples within a plot

1. In the “Adjustable settings” panel, several other settings can be found to change the specific input for the analysis or to adapt the looks of the graph:
 - In the Graphics section you can also select **interactive plot** type which support directly the highlighting/marking of individual samples shown in Figure 6a.

Figure 6a: Marking samples in an interactive plot

- Many layout settings can be adjusted in the *More Settings* option, such as font size, colors and marker type. In order to view the extra options, click at the top on the the ‘More Settings’ sections.
- To highlight specific samples in the graph, you can simply double-click in the graph on the marker-points of the samples that you want to highlight, or you can enter the R2 sample IDs in the field ‘Samples to mark’ from the ‘Adjustable settings’ panel. If you enter multiple IDs, separate them with a comma.
- 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 5, the “red arrow”). Changes in marker type and marker color can be achieved as well with a specific syntax applied in the before mentioned ‘Samples to mark’ text field. Keep reading to find examples of how to use the syntax yourself.

Always click on ‘Submit’ button at the bottom of the ‘Adjust settings’ panel for your adjustments to take effect!

If you enter sample names in the ‘Samples to mark’ entry field, then the appearance of these marked samples can be adapted via the to other fields as indicated above. However, these will be applied to all the marked samples at the same time.

If you would like to use different layouts and or colors for every indicated sample, or for groups of samples, then you can adjust the marker color and type of the samples that you marked with a syntax in the ‘Samples to mark’ field. The syntax for using this strategy is explained below:

Please note that the explanation below for marking samples is redundant and is linked to the static plot option and the static plots will be removed.

The syntax to use for ‘Samples to mark’ is: `sample1,sample2:hexcolor1:method1`.

Lets start with a simple example of manually defining the marking of 2 samples (namely itcc0288,itcc0021). In order to mark those, we add a ‘:’ after the IDs with a color code (formulated as hexcolor, but without a #), and/or add a ‘:’ with a marker type, for example; `‘itcc0288,itcc0021:ff4444:dot’`. Please see the ‘did you know’ box to find out what mark_types ca be used.

To bring attention to different samples, you may want to use multiple colors and types of marking. 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:hexcolor1:method1;sample3,sample4:hexcolor2:method2’`.

Please note, that there is a semicolon ; between the 2 groups of samples here. For example: `“‘itcc0288:ff4444:epicenter;itcc0021:#FF0000:arrow;itcc0013,itcc0132:00ff00:dot’”` creates the markings as shown in the figure below.

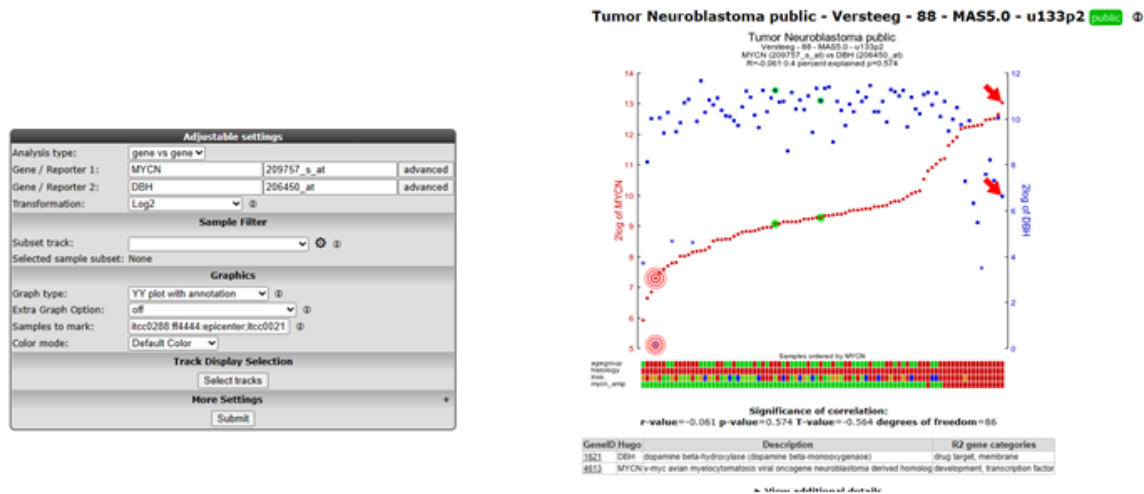


Figure 6: Adjusting the sample mark layout



Did you know that R2 allows you to emphasize samples in the graph with many different marker options?

R2 knows a couple of marker 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
 - 'text': will plot the samplename by the corresponding plot (interactive plot only)
- 'colors': use a hexadecimal color code (many examples of these can be found online)

Note: The dotsize does not scale with 'arrow' and 'triangle' method.

Another often used feature is the **Vector (SVG) output** option. The vector images are often used in manuscripts and to increase the resolution meeting your own demands. Currently R2 is making use of several modules to generate the graphs. So depending on the type of graph type you selected the svg plots can be generated.

- Version A. Click on the top 'More Settings' settings, where you can find a dropdown next to the 'Vector (SVG) output' setting. When the dropdown is set to "True", and the 'Submit' button is clicked, a link appears above the 'Adjustable Settings' panel, that you can right-click to save the vector image to your computer. SVG vector images can be manipulated in any vector graphics software, such as Illustrator, GIMP or Inkscape (left part of fig 7).

- In case you have selected a graph in the interactive section the name in the name you find the save option in the wheel icon directly in the graph.



Figure 7: Obtain a vector (SVG) image of your graph



Did you know that the ‘Adjustable settings’ panel is available under most graphs and analysis results in R2?

Just scroll down the page to find the Adjustable settings box with options to adjust the settings of the analysis or to adjust the looks of the graph. **Don’t forget to press the Submit button at the bottom of the panel in order for your changes to take effect!**

1. When you are in the one gene view plot, a logical step is to investigate the expression levels in correspondence with a group parameter. Select in the analysis type menu “gene vs track” in the track selection menu, the inss track and click “submit”, Figure 8. The cohort is separated accordingly by the patient INSS staging in alphabetical order. It could be that the ordering of the group parameters is not the most convenient representation for your analysis. In order to customize this, you can create your own track as described in (the chapter “Adapting R2 to your needs”) and make your own track with the correct ordering eg: a_stage IV, b_stage 1 etc.

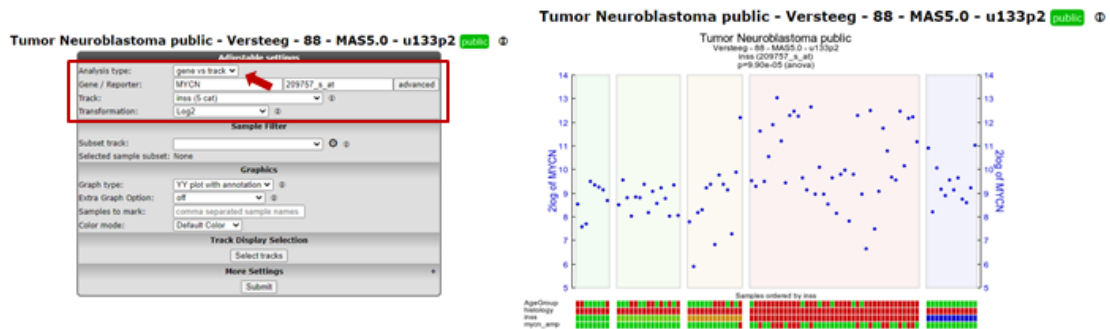


Figure 8: Gene versus track

N.B. The same analysis can also be obtained with the module “View a Gene in Groups” from the main page.

1. The current representation is the most honest way of showing your data, as every single value is visible in the plot. In the adjustable settings you can fine-tune your graph for example by switching on the track and genesort within the groups (Figure 9).

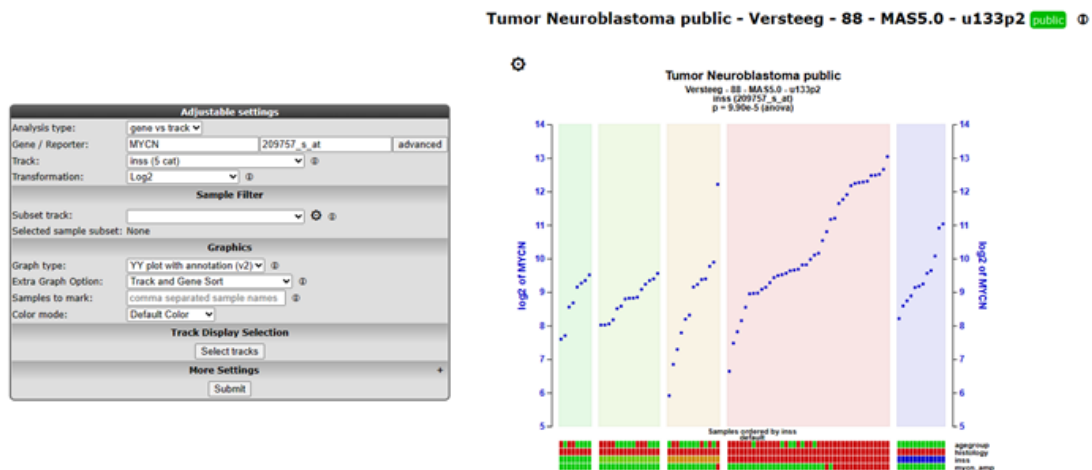


Figure 9: Gene versus track sorted

2. We can also change the graphical representation of the data by selecting another graph type. Select for example ‘boxplot’ from the ‘graphtype’ dropdown and change ‘color by’ to ‘color by track’, such that the inss track is used to color the boxes. Press the ‘Submit’ button again to change the view. We now obtain a boxplot image where the respective groups have been colored according to the inss groups. Adaptations to other graph types can be made in a similar way.

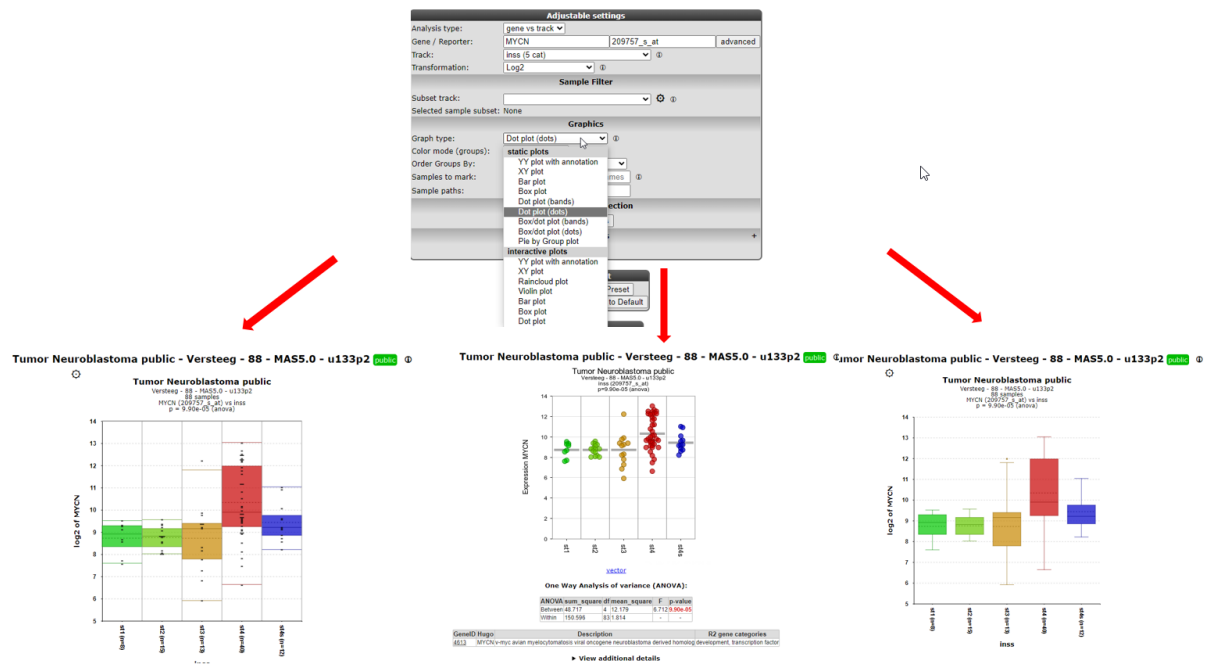


Figure 10a: Making boxplots

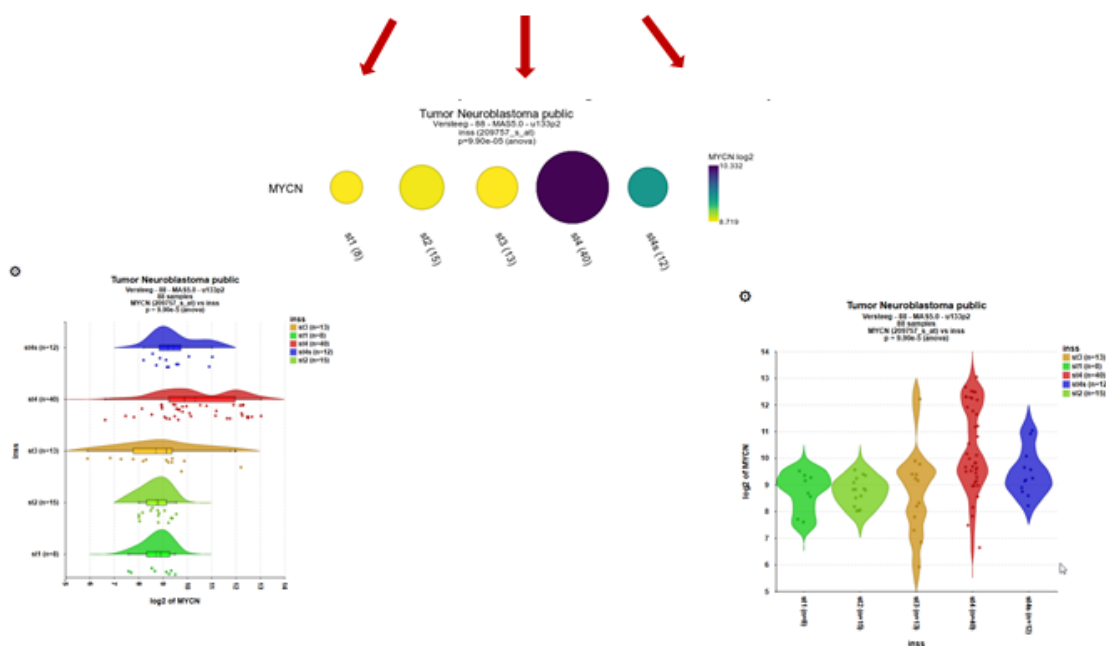


Figure 10b: Making raincloud plots and more

The boxplot and dotboxplot can of course reveal valuable information about statistics and the distribution of data. Another relative new visualisation technic which combines several aspects of the traditional plot such as the dotboxplot and violins plot is the **raincloud** plot combining summary statistics such as median and quartiles and the density estimation of the violin plot. The individual datapoints are represented as points or raindrops along the vertical axis which provides a better understanding of the data distribution (see left graph in Figure 10b).

1. You can also sort the groups by their average or median gene expression and customize your graph in various ways. In Figure 11 this illustrated by the INSS stage sequence according to the median gene expression and enlarged dots of MYCN . And on top of that the individual dots are coloured by the Z-score of the DBH expression. In case the samples have also been profiled for another type of data such as methylation data or drug data, you can also select these datatypes in the pulldown menu when you have selected color by gene.

you can use these values from these corresponding sets in the same graph (Figure 11) combine expression data and other omic data.

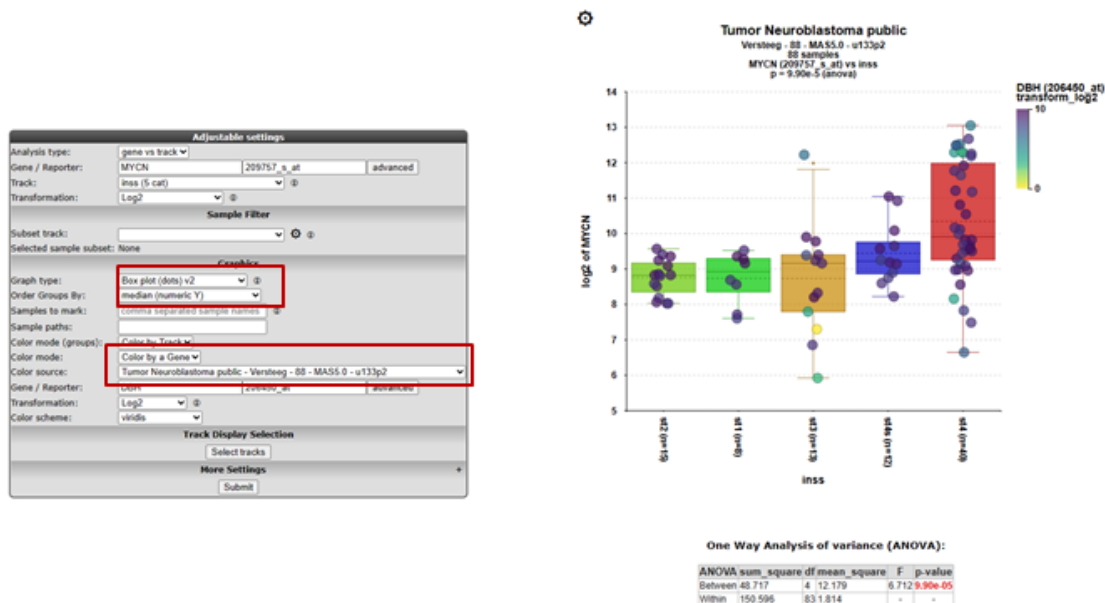


Figure 11: Ordering boxplots

- Also, directly accessible from this menu is the track vs track option described in more detail in Chapter 5: Annotation Analyses; “relate two tracks”. Keep in mind that the analysis type options described in this chapter can also be selected directly from the main menu.

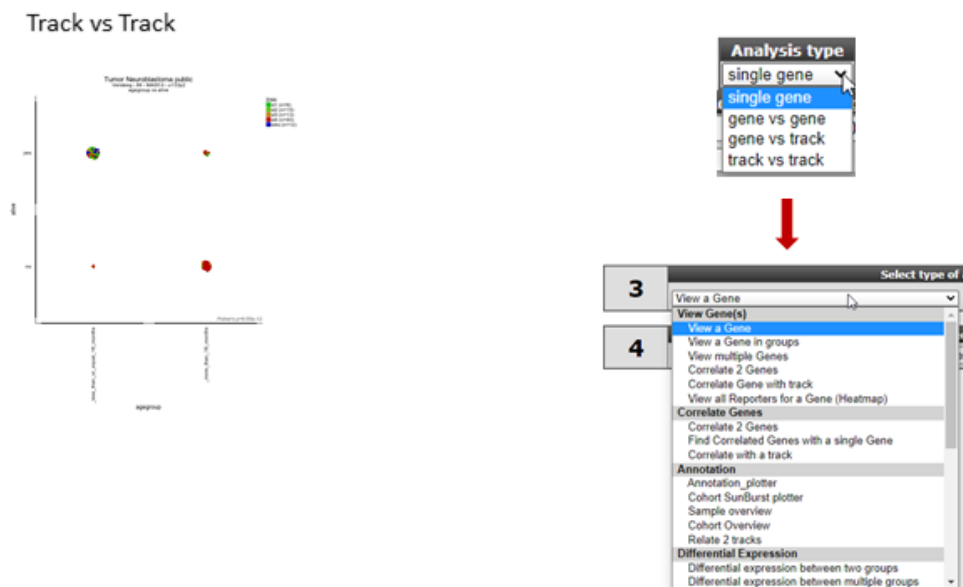


Figure 12: Track versus Tracks / Main menu

- ‘single gene’: In main menu: “view a gene”
- ‘gene vs track’: In main menu: “View a gene in groups”
- ‘track vs track’: In main menu: “Relate two tracks”



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 is greater than 2, R2 will automatically perform brute-force t-testing with Welch correction to identify the combination of 2 groups that have the most significant difference. Just click the triangle at “View additional details” and gain insight into all the tested combinations.

3.7 Step 6: Sources for additional information on the selected gene

1. In all the types of view also a right menu panel is generated.



Figure 12: Left menu panel providing additional info (including link-out) and analyses options

In the right upper menu-panel several options are available to provide you with additional information sources of the MYCN gene and additional analyses. When you click on the link (the name of the gene) under the PubSniffer header 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 “PubReMiner” links redirects you **to Pubmed PubReMiner** which is a tool for PubMed query building and literature mining.

KaplanScan and Time Series analyses will be discussed in separate tutorials. However, keep in mind, in case Kaplan Meier data is available for a given dataset this will always be visible in the right menu for one-gene-view. 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 and other platforms if available.

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). Further, in the Sample Map section pre-generated high-dimensionality reduction maps (t-SNE and umap) can be plotted.

*Did you know that **PubReMiner** 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? PubReMiner is a web-based tool that allows simple text-based query building and information gathering (mining) of the NCBI literature search engine PubMed. PubReMiner 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, PubReMiner 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

3.8 Step 7: Adapting a plot

1. To investigate the values R2 uses for graph generation click on “View Datatable” to open a table with the expression levels for all samples. Within this table you can use filters to restrict samples. By selecting the rows, a second table is generated, that can be copied, create an additional track e.g based on sorted data and subsequently pasted in Excel, for further investigation.

The figure consists of two screenshots of the R2 web interface. The top screenshot shows a 'Data table - current dataset: Tumor Neuroblastoma public - Versteeg - 88 - MASS.0 - u133p2'. It displays a table with columns 'samplenames', '209757_s_at', and '209757_s_at'. The table lists various ITCC samples and their corresponding expression values. A filter is applied to the '209757_s_at' column, showing 'greater than' and '10'. The bottom screenshot shows the same table with a filter applied (greater than 10) and a 'Track display selection' panel on the right. This panel shows a list of samples to be displayed, including ITCC0003, ITCC0015, ITCC0018, ITCC0020, ITCC0021, ITCC0022, ITCC0025, ITCC0026, ITCC0031, ITCC0038, ITCC0056, ITCC0061, ITCC0174, ITCC0189, ITCC0198, ITCC0219, ITCC0383, and ITCC0385. The 'Track display selection' panel also includes a 'Click in the text below to select all' button and a 'Filter' button.

Figure 13: Unfold the datatable

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 One Gene View. Persistent changes to the tracks can be made via the ‘User Options’ menu item, which is present in the main screen (see the tutorial ‘Adapting R2 to your needs’). Note that the Adjustable Settings panel, including the Customize Track parameters, is available throughout R2 for temporary adaptation of Track visibility and other preferences.

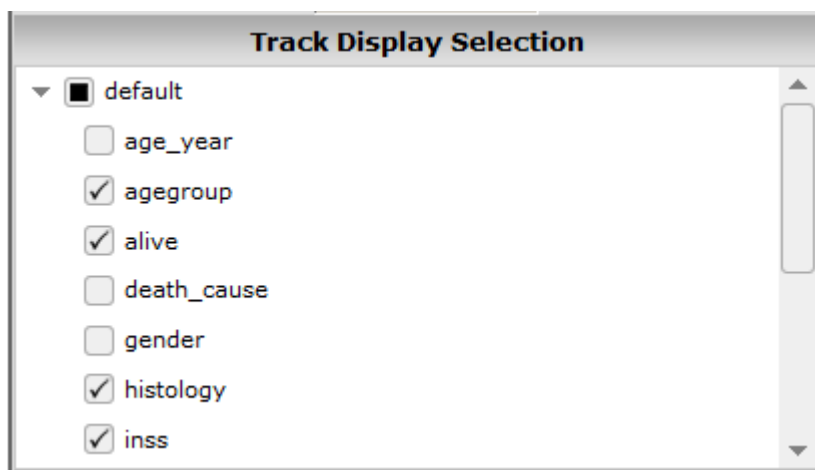


Figure 14: Tick and drag tracks

- 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.

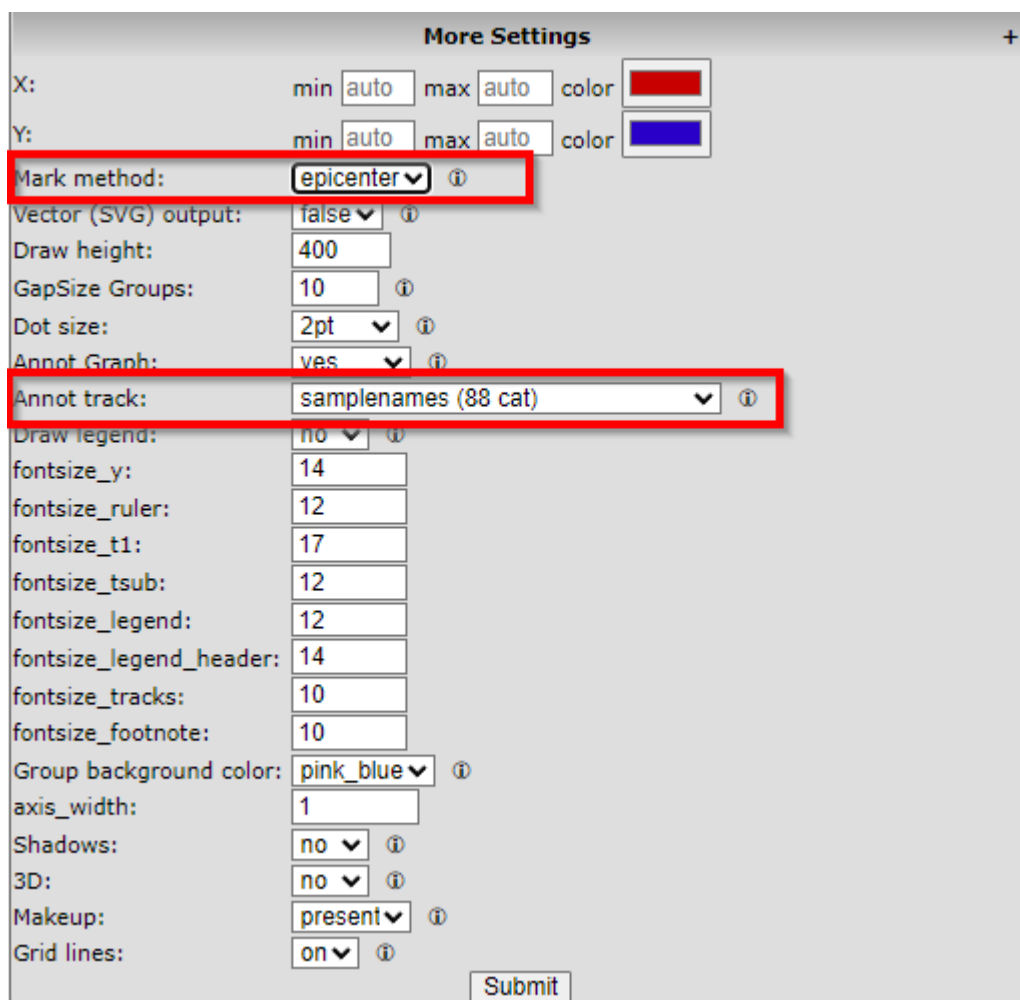


Figure 15: the extra settings Panel

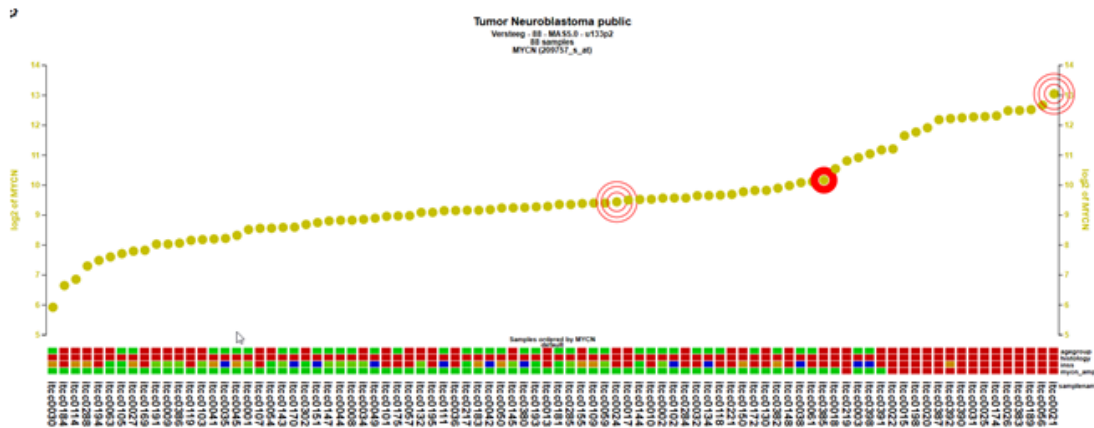


Figure 16: Sample annot added

In Figure 16 sample annotation (“Annot Graph”) and legend (“Draw Legend”) were added. The “Annot Graph” option, adds the information of a selected track to the YY-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 haven 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 a 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. Also by clicking the dots in case you have chosen for the YY-v2 version individual samples can be marked by arrow, circles etc. In Figure 17 an extra coloring for the individual samples has been applied, in this case the alive yes/no track.

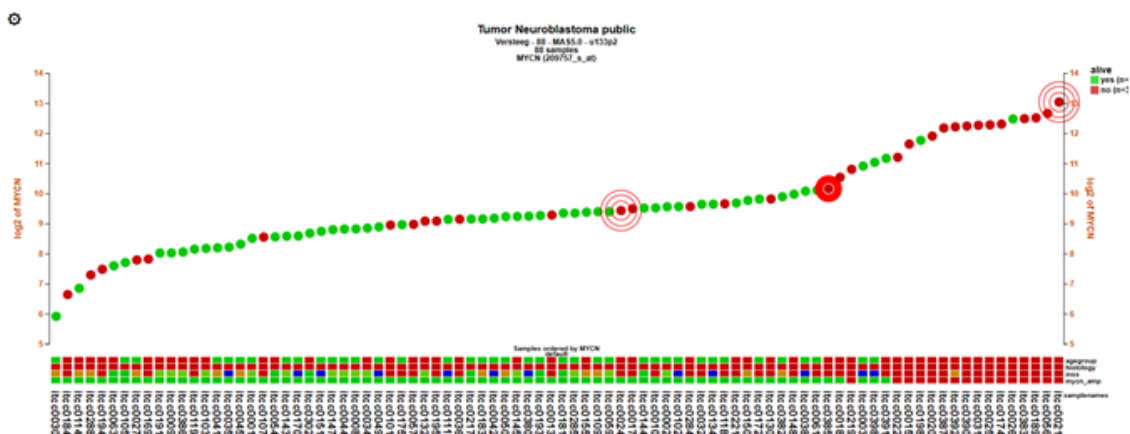


Figure 17: Fonts and Color changed

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 mycn 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 Submit button. These selections can be repeated a couple of times to build your ultimate selection.

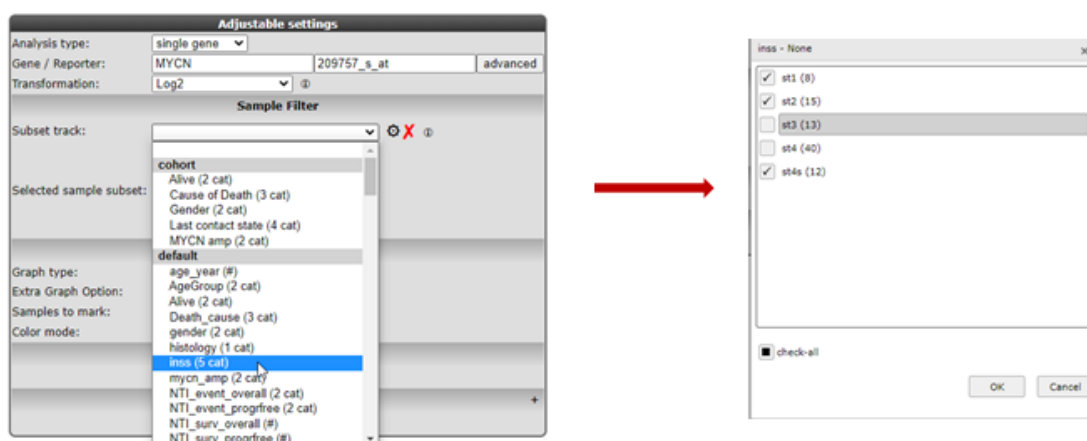


Figure 17: Selecting subgroups

The graphs below were drawn with GraphType BoxDotPlot. All stages are depicted on the right hand side graph and only the lower risk stages on the left.

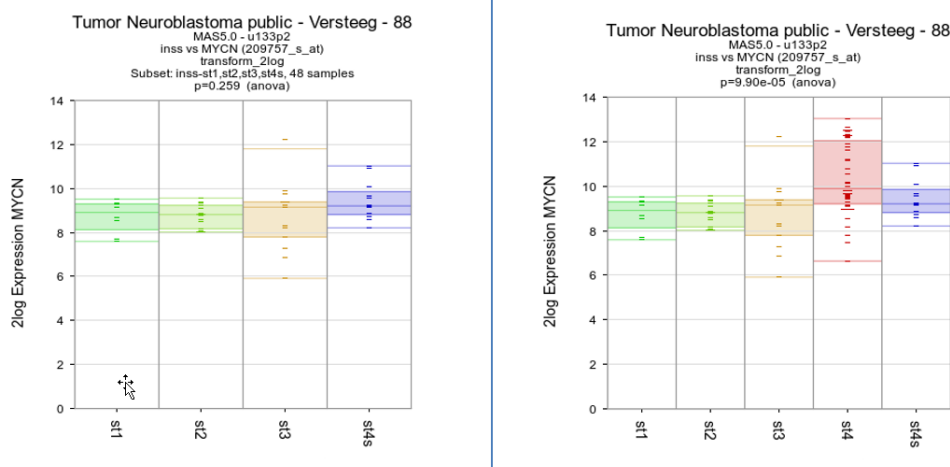


Figure 18: All stages (right) versus lower risk only (left)

Next to the filter dropdown menu, clicking the “wheel” will open a pop-up screen with a grid to create tracks with in or excluding individual samples for the available tracks. From this point also a track can be created and

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

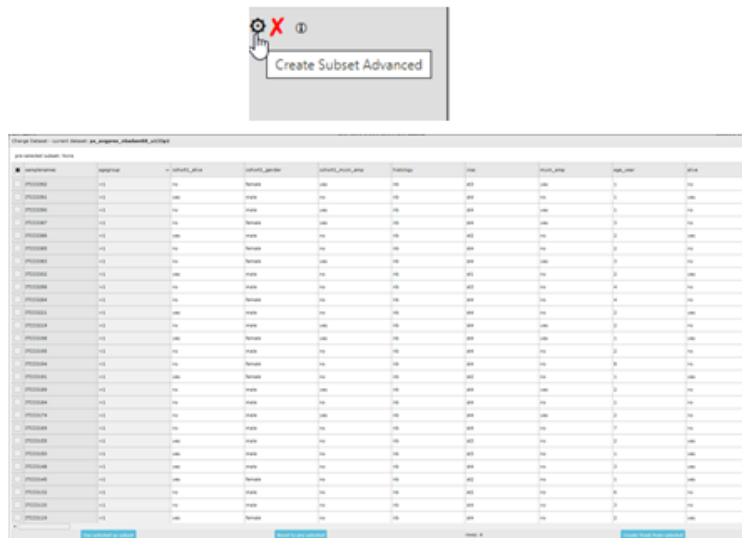


Figure 19: Advanced filtering in the grid

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 test is performed on each track. In addition, tracks are inspected before doing the test, and the test is changed according to the contents: for a numeric vs numeric track the correlation is calculated resulting in an r-pvalue, in other words the correlation between the p-values. categorical vs numerical tracks are tested with an anova test; nonrandom associations for categorical vs categorical tracks are tested with a Fisher's exact test. Furthermore, 'ND' samples are automatically removed, and are not considered a valid group. We can run a CliniSnitch analysis directly from the One Gene View page by clicking on the gene name under 'CliniSnitch' in the upper-left panel. Click on 'MYCN'. Private/Group tracks that you may add to this dataset over time, will automatically be included in these analyses.

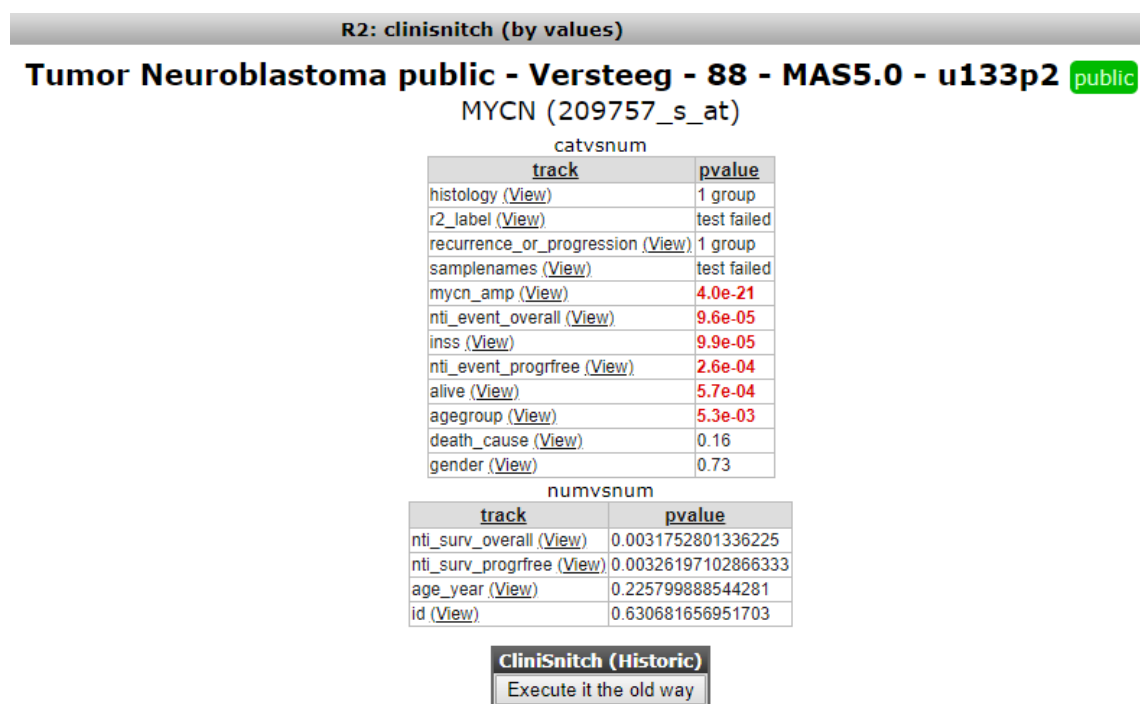


Figure 20: CliniSnitch result for MYCN

- Every test can also be visualized by clicking on the View-link in the table. 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 a significant value. Click on 'View' behind inss to inspect this further.

3.11 Step 10: Finding sample extremes.

In case you wonder whether any unusual expression levels show up for individual samples from a given dataset, you can use the "Find sample extreme" option. In this example we know that sample ITCC0288 harbors a Phox2b mutation which leads to the question: can we find extreme expression values for this sample?

- In the one gene view for this dataset select in the right panel sample itcc0288 in the sample overview section and click 'view', leave all the settings at their default and click next.

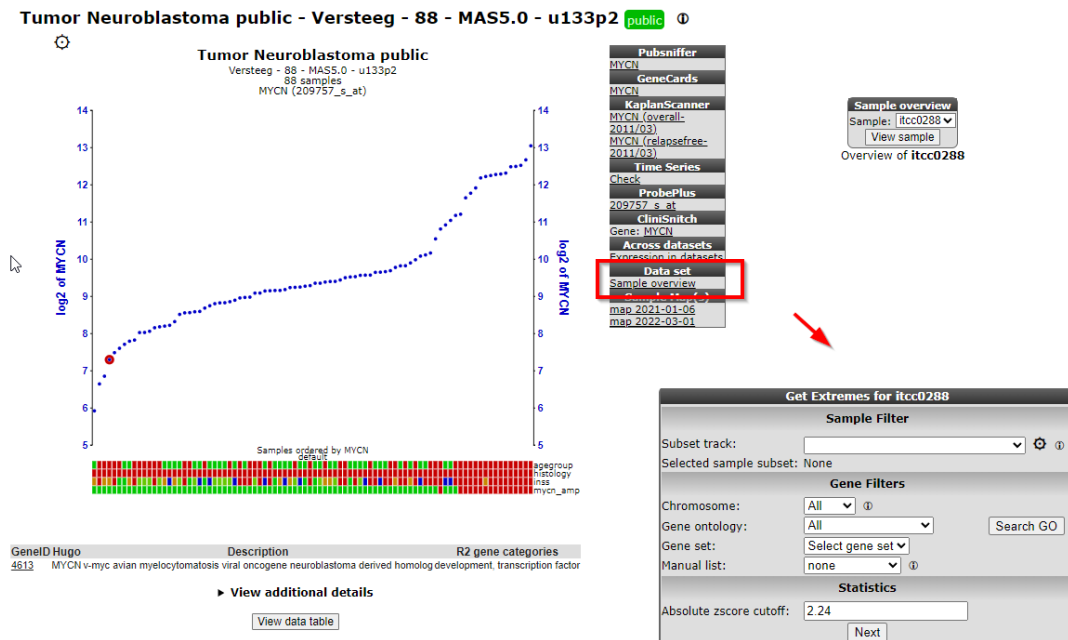


Figure 21: Select your sample to find extremes

2. A table shows the negative z-score (left column) and positive z-score (right column) extremes. In Figure 22, genes which are a part of the Nor-Adrenalin pathway are in the top of the negative z-score list. This suggests that wild-type Phox2b is involved in the up-regulation of the Nor-Adrenalin pathway. You can click on any of the genes listed in the table (here we clicked on “TH”) to obtain the One Gene View of that gene, with your sample marked in the graph.

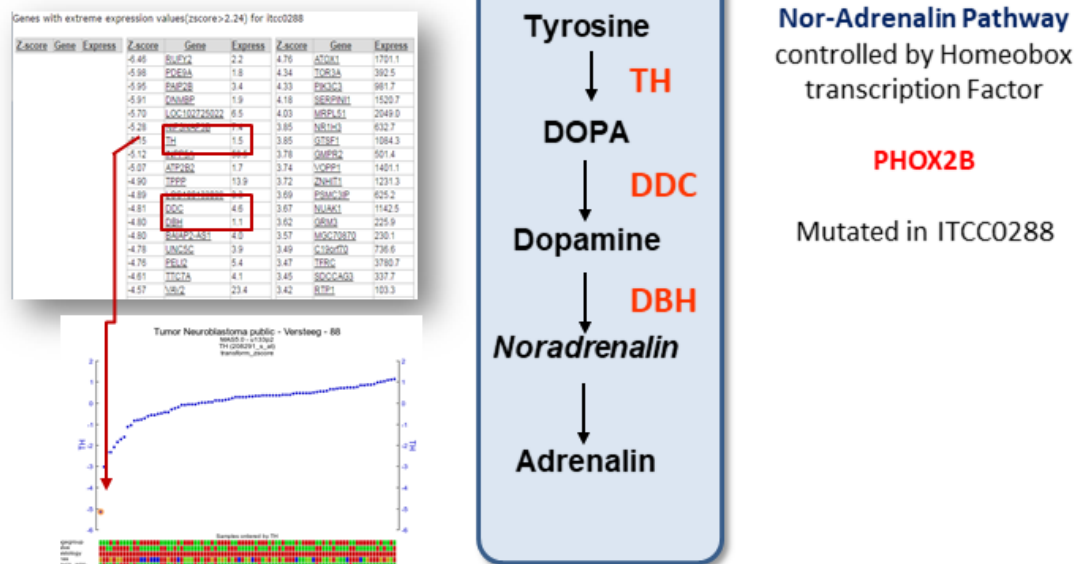


Figure 22: Sample extremes in one sample

3.12 Step 11: Probeset verification

1. Underneath the graph and gene description, click on the down arrow of the text “View additional data”. The table displayed in Figure 23 lists whether the various reporters of MYCN 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). The MYCN reporters with a “NO” indicate there may be an issue with it. Scroll down the page and click on the “Tview” link in the reporter table.

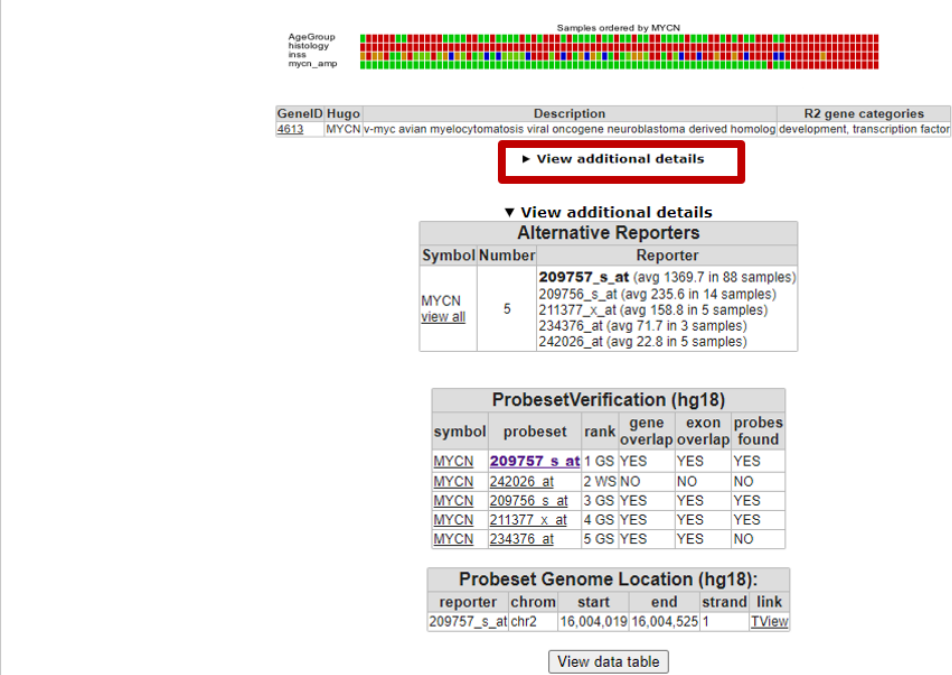


Figure 23: Probeset verification table

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 25). 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 with 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.
2. For many datasets R2 hosts the location of the reporters on the genome so the location and orientation can be inspected as described for the Affymetrix probesets. If the ProbesetVerification table doesn’t appear R2 has no information of the reporter genome location of a given dataset.

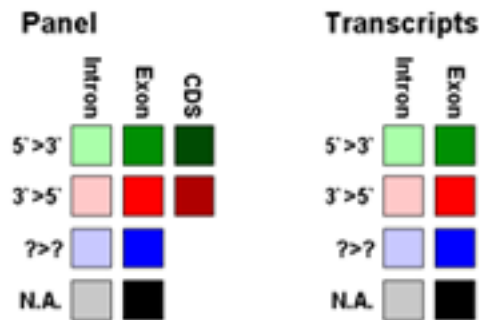


Figure 24: Coloring represents type of transcript

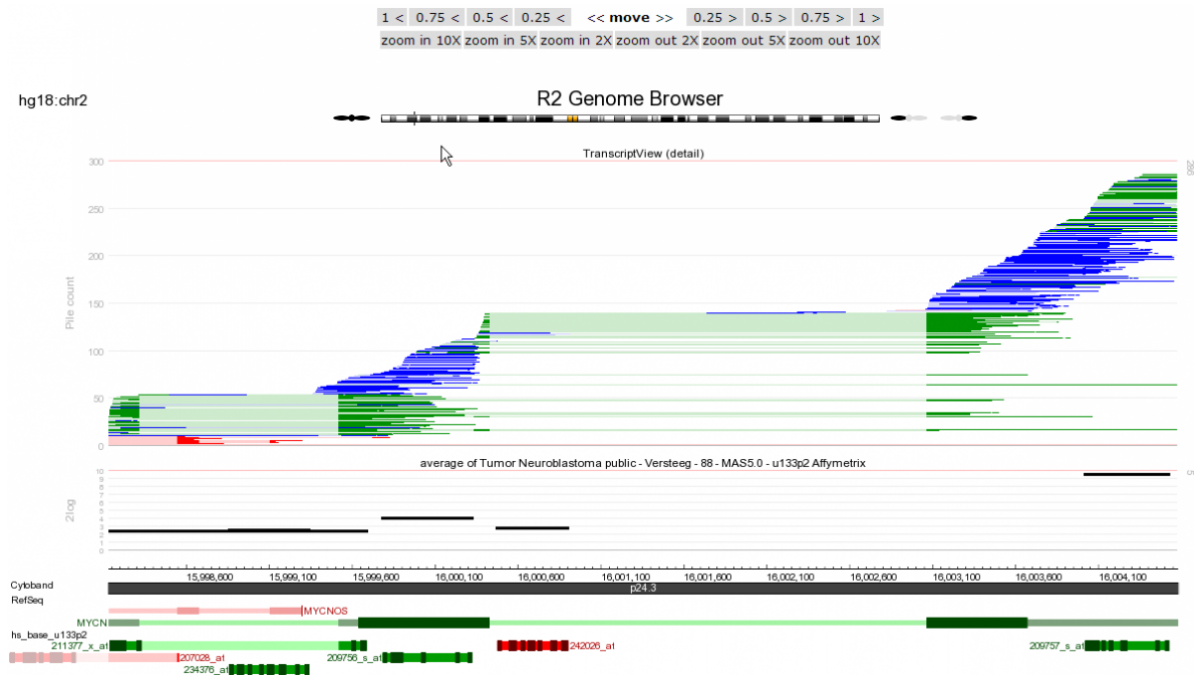


Figure 25: MYCN reporters in Transcript view



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.

3.13 Final remarks / future directions

Some of these functionalities have been developed recently. If you run into any quirks or annoyances, do not hesitate to contact r2 support (r2-support@amsterdamumc.nl).

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

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

1. Use “single dataset” in field 1 and select the “Tumor Medulloblastoma PLoS One - Kool - 62 - MAS5.0 - u133p2” dataset in field 2.
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.

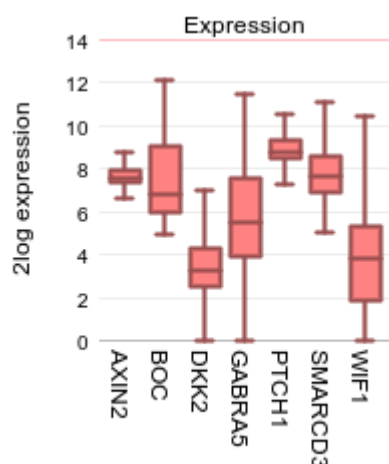
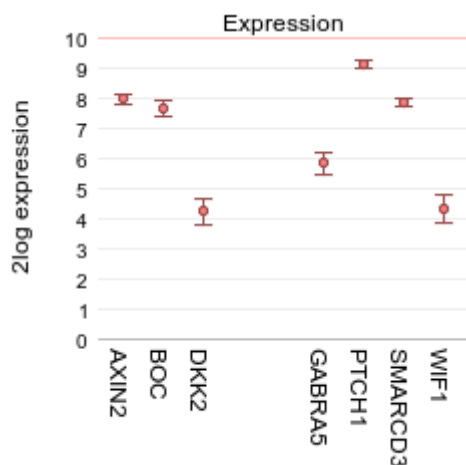


Figure 1: Default multiple genes view.

- Under the graph many settings can be adjusted in the Adjustable settings box. For instance, you can choose a different type of graph. Next to the option to influence the space between the genes, the height of the plot etc, you can add one or more extra spaces between genes of choice by adding one or more extra delimiters. In the GENE/reporter textbox add 2 extra comma's behind dkk2 (AXIN2,BOC,dkk2,,,GABRA5,PTCH1,SMARCD3,WIF1); in the field 'Plot type' select "Average with stderr", set 'Transform' to and click next. This enables the creation of visual subgroups in the gene repre-



sensation.

Figure 2: Add a spacer between genes

4.3 Step 2: Viewing multiple genes through track annotation

- In Figure 1 a selection of gene expression profiles is depicted in one picture in contrast to the one gene view. Figure 2 shows the possibility to make gene subgroups by adding extra spacers to the plot. Now we will look at the option to represent the gene expression separately for each subgroup of a categorical track. In this manner potential relations between subgroups and gene expression can be visualized.
- 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 subgroup of the categorical track "subtypes" of medullablastoma cells, select in the Adjustable settings box "subtype (5 cat)" at 'use track', "lump by group plot gene" at 'handle groups by' and "Track" at 'color by track'. Further set 'Transform' to "none", select "boxplot" at 'Plot type' and click NEXT.

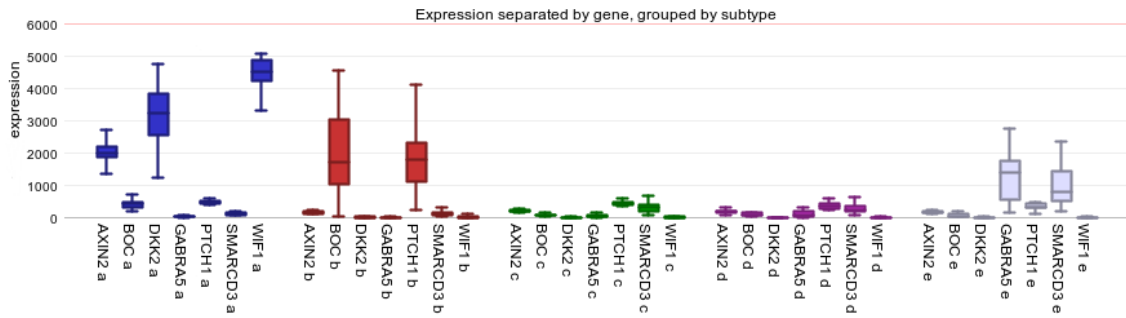


Figure 3: Multiple gene view per subgroup

- Most of the overexpressed genes are part of the WNT (subgroup A) and de SHH (subgroup B) signaling pathway as shown by Kool et al. In Figure 3, the gene names and the subtype labels a,b,c,d and e are concatenated on the x-axis. The Medulloblastoma molecular subtypes clearly show different expression profiles, with an overexpression of genes in subgroup a and b.
- Also try the “lump by gene plot group” which will produce an image where the genes are shown, separated by the subtypes. The “group by panel” option under ‘handle groups by’ will display the same information in a multifaceted graph with separate panels for each sub-group of the chosen track.

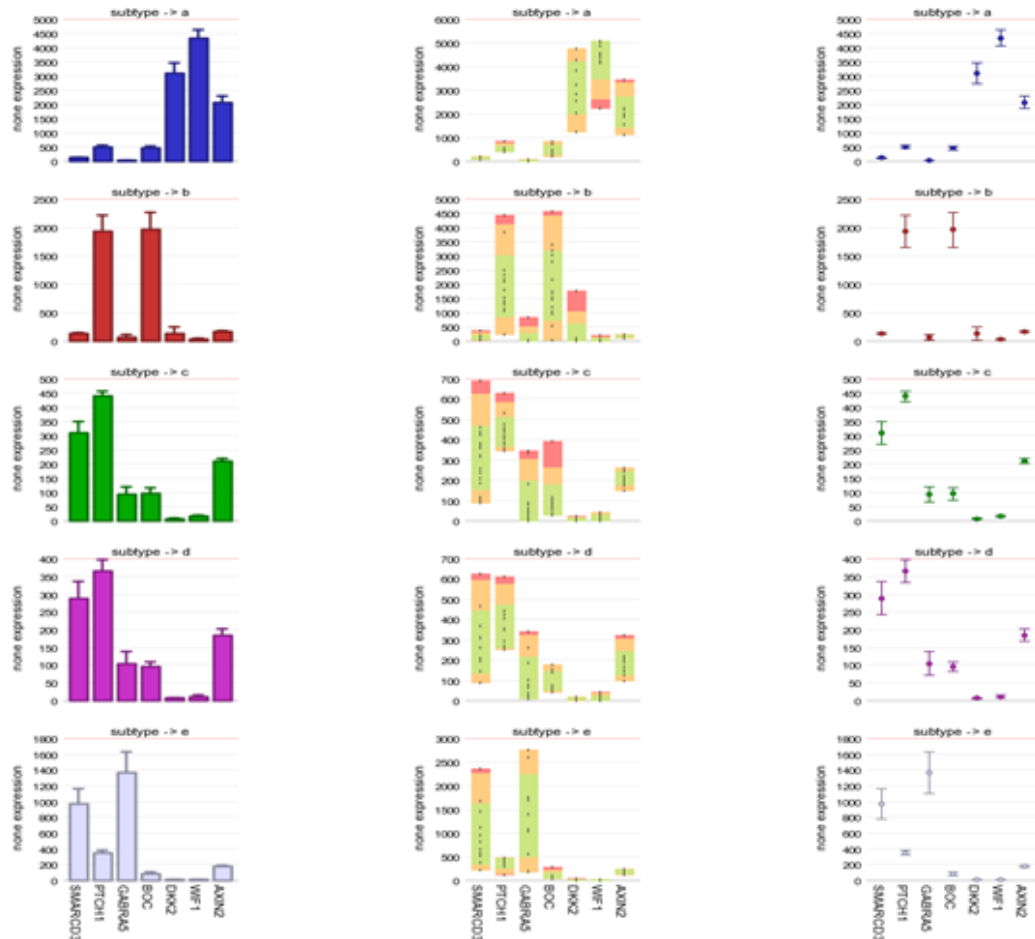


Figure 4: Multiple gene view, panel per subgroup

- The sample filter option allows you to generate a multiple gene view representation, that can also depict parts of the data set using subsets, as is available throughout R2. For example you could choose to exclude one or more of the subgroups, or use another track to make any intersection you would like to use.
- In some situation, you may want to highlight one or more samples or patients from a data set in the multiple

gene view. This is easily achieved, if you use the r2 samplename. Lets try it with one of the patients within this medulloblastoma cohort ('itcc0334'). To mark this sample, we simply paste the samplename in the 'samples to mark' field and update the plot. Using the options provided by the settings you can tweak the result to your personal taste.

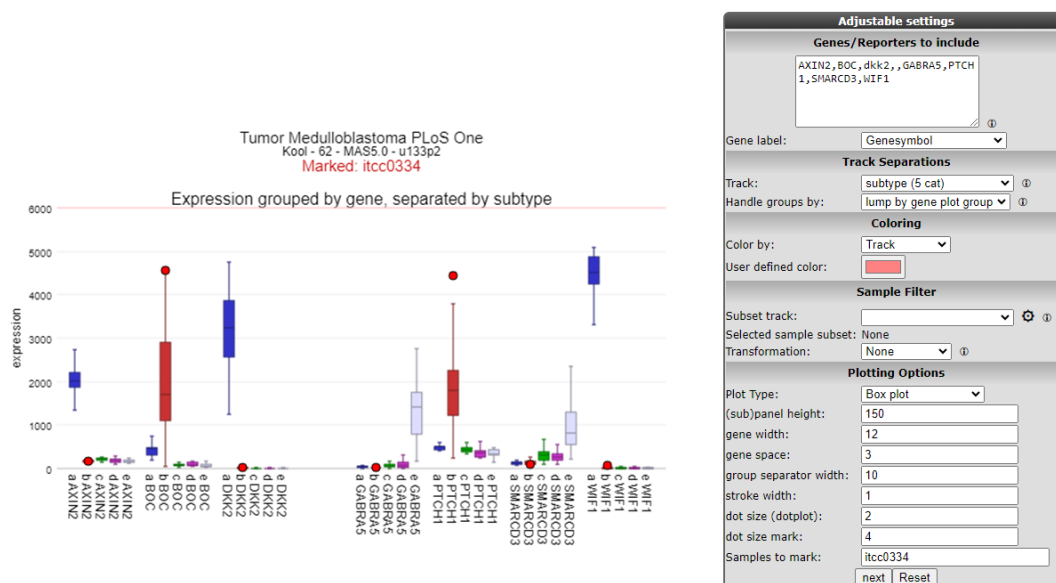


Figure 5: Multiple gene view, mark a sample

- To use the generated image in programs like Powerpoint, you can 'right-click' on the image and 'copy' it to the clipboard. A simple paste will then put the figure in your presentation. You can also store the file as a 'svg' vector file on your computer. Using this route, you are also able to adapt the figure in a program like Adobe Illustrator, or Inkscape.

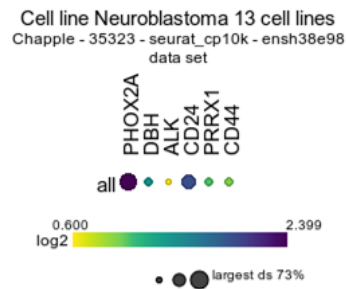
4.4 Step 3: View multiple genes (Bubble plot)

- Up to now, we have used the more classical graphs to look at more than 1 gene at once within a data set. Within R2, you can also depict multiple genes in a graphical format that is often used within single cell analyses, called 'bubble plot'. The concept here shows the average expression of genes (optionally within different groups contained in a track) as a color representation in a grid. Every cell in this grid is then drawn as a circle, such that the surface area reflects the ratio of the samples within such a group that is considered to be expressed (having a Present call within R2).

So let's have a look and explore this feature in a scRNA dataset first. From the main page, we first select a data set called: 'Cell line Neuroblastoma 13 cell lines - Chapple - 35323 - seurat_cp10k - ensh38e98'. This resource contains 13 neuroblastoma cell lines, measured by 10x Genomics technology.

- From the option box 3 we then select 'view multiple genes (bubble plot)' and progress to the next page. Since we will work with a large resource, it may take a few seconds for the page to finish loading the 1st time that you use it within your current session.

At first nothing much is shown. Now paste a few genes in the 'genes' box "PHOX2A,DBH,ALK,CD24,PRRX1,CD44" and press 'next'. We now see 1 row of circles, showing the expression of the different genes in the color representation, and the ratio of cell showing expression as the size of the circles



Adjustable settings

Genes/Reporters to include

PHOX2A, DBH, ALK, CD24, PRRX1, CD44

Gene label: Genesymbol

Track Separations

Track: data set (1 cat)

Bubble width: 15

Scale bubble by: largest proportion

Color scheme: viridis

Sample Filter

Subset track:

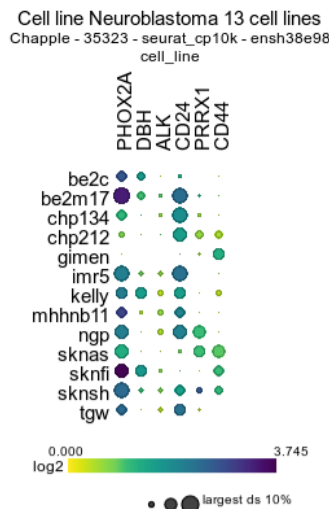
Selected sample subset: None

Transformation: Log2

next Reset

Figure 6: Multiple gene view (bubble), 1

- We can also use a track, to segregate the cells (or samples in case a bulk data set is used). Let's use the 'cell_line' track and press 'next' again. We can now see 13 rows of the circles, one row for every cell line. This starts to look more interesting, as not every gene seems to be equally expressed in the different cell lines.



Adjustable settings

Genes/Reporters to include

PHOX2A, DBH, ALK, CD24, PRRX1, CD44

Gene label: Genesymbol

Track Separations

Track: cell_line (13 cat)

Bubble width: 15

Scale bubble by: largest proportion

Color scheme: viridis

Sample Filter

Subset track:

Selected sample subset: None

Transformation: Log2

next Reset

Figure 7: Multiple gene view (bubble), 2

- We can also reorganize the genes, by putting them in another order. It is also possible to 'group' the genes, by adding an additional ','. Let's try the following order "PHOX2A,CD24,DBH,ALK,,PRRX1,CD44" and press 'next' again. The image looks more organized now.

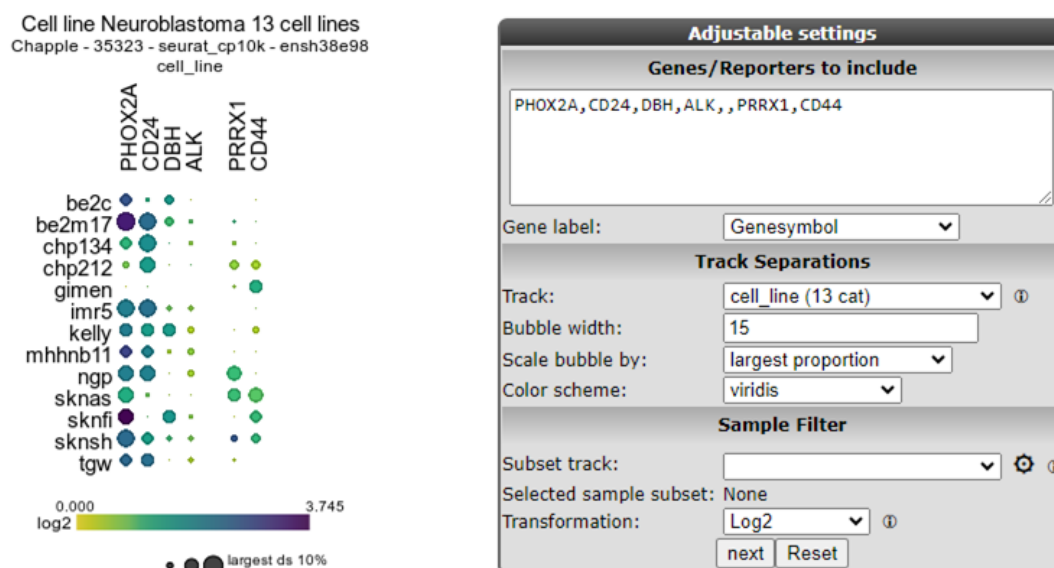


Figure 8: Multiple gene view (bubble), 3

- By default the circle areas are represented as proportion of the complete data set. To not end up with very small circles, the largest identified proportion is chosen as the maximal bubble width, and all the others are scaled proportionally. This allows you to see the expression in those cells that have a signal, but also that not every cell line has equal numbers of cells. If you are not so much interested in the relative numbers between the different cell lines, then we use the 'scale bubble by' drop down and use 'group size proportion'. Press 'next' again. The image now reflects how the different genes are proportionally expressed by row.

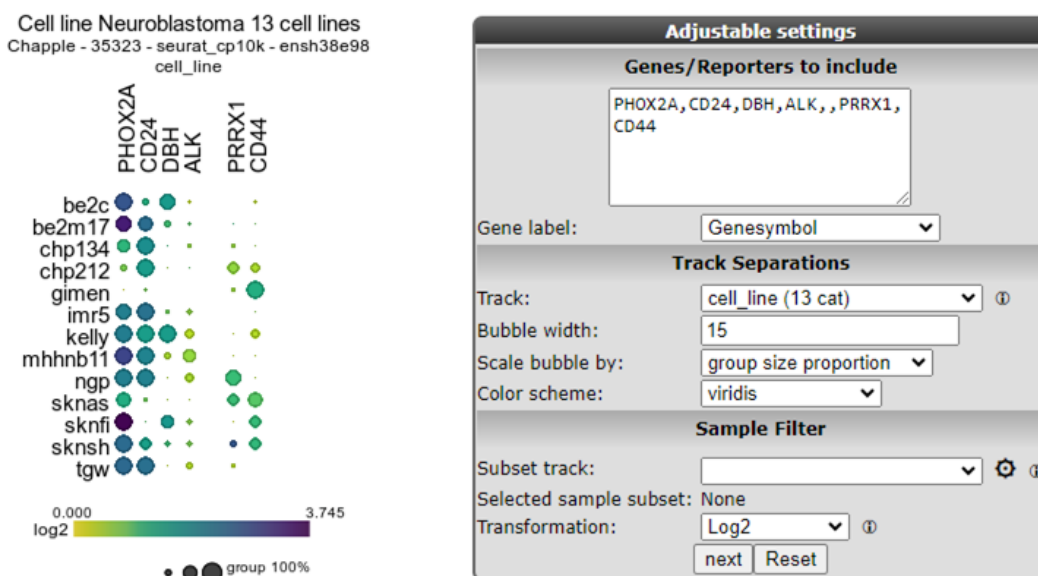


Figure 9: Multiple gene view (bubble), 4

- As with many plots in R2, you can also use different transformations, use other color schemas or make the circles bigger etc. Hovering over the circles will also show you the statistics for a circle, that may be helpful when exploring the data.

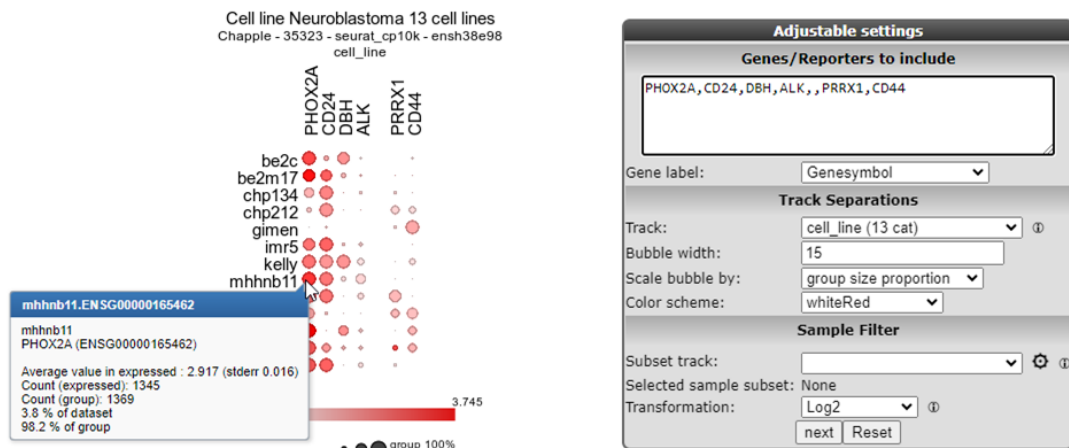


Figure 10: Multiple gene view (bubble), 5

7. As will also be explained in the dimensionality reduction tutorial of the tutorials book, single cell data is often explored in so called 'UMAP' plots as well. These are also available in R2 via the 'menu' at the left side named 'sample maps'. Using that route, you can also explore the expression of a gene, or annotation features in our interactive plots. Combining some of these functions will enable you to test hypotheses, and produce images for your next publication ;)

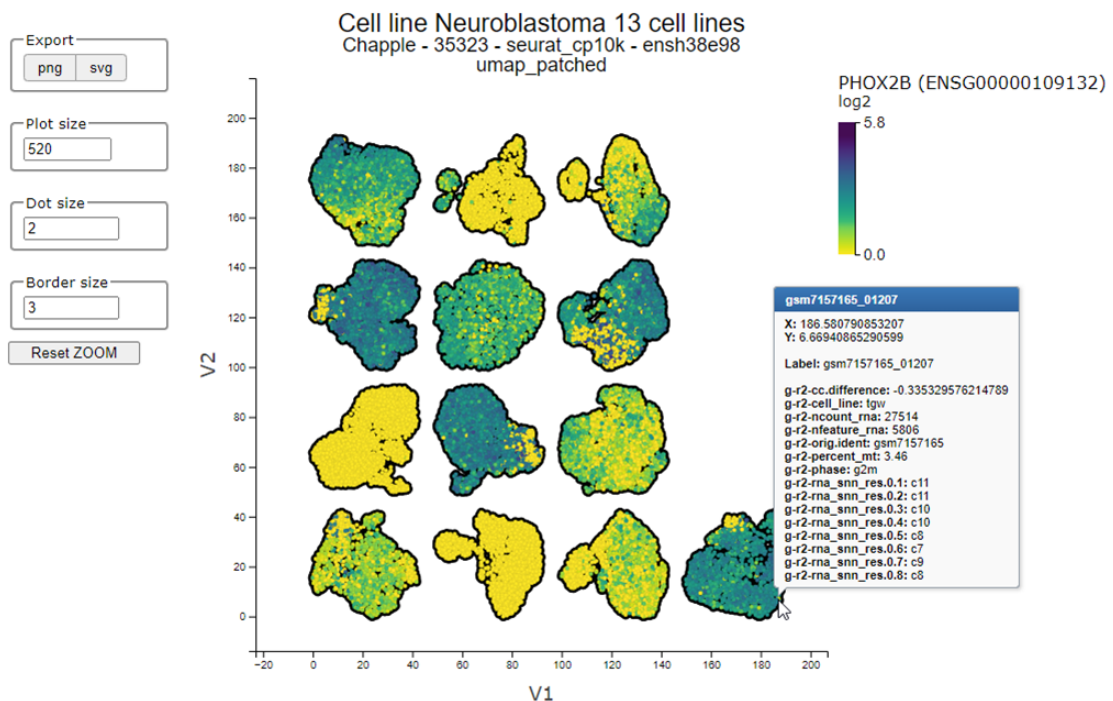


Figure 11: Multiple gene view (bubble), 6

Finally, we have now used these bubble plots in a single cell dataset, where they are very well suited to convey both expression level, as well as 'proportion expressed' information. These plots can be equally informative in many of the bulk expression data sets that are provided within the R2 platform. In many data sets a notion of 'is expressed' is also captured in the 'present call' variable.

4.5 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@amsterdamumc.nl).

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

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

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 is significant overlap 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.
- Sunburst plotter and others; visualization options of tracks within a 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”. In the ‘Select type of analysis’-box (3) select “Relate 2 tracks”, which can be found in the annotation subsection and press ‘Next’.

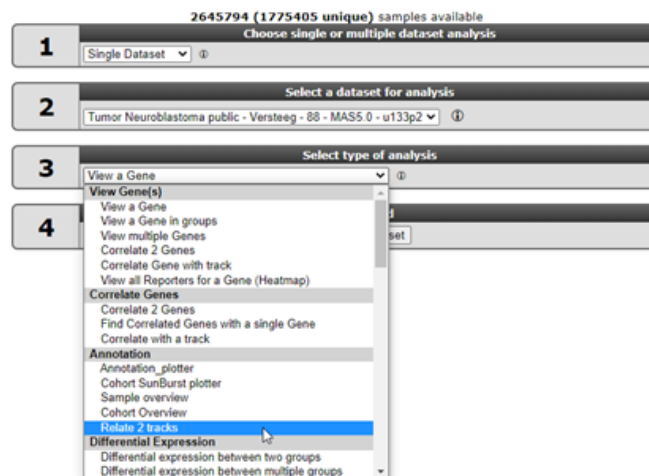


Figure 1: Select “Relate 2 tracks”

- For the different tracks, make sure that you select a categorical one (which can be recognized by (cat)). We 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). Select the Interactive ‘XY’ plot in the graph section. Then press ‘Submit’ to generate the result.

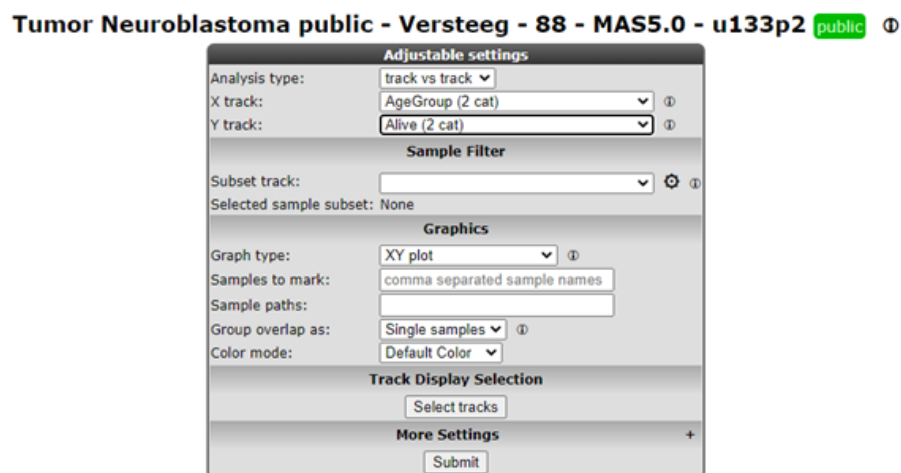


Figure 2: Select Selecting categorical tracks

- 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 “inss” as track. Press the ‘Submit’ button 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 deceased 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.

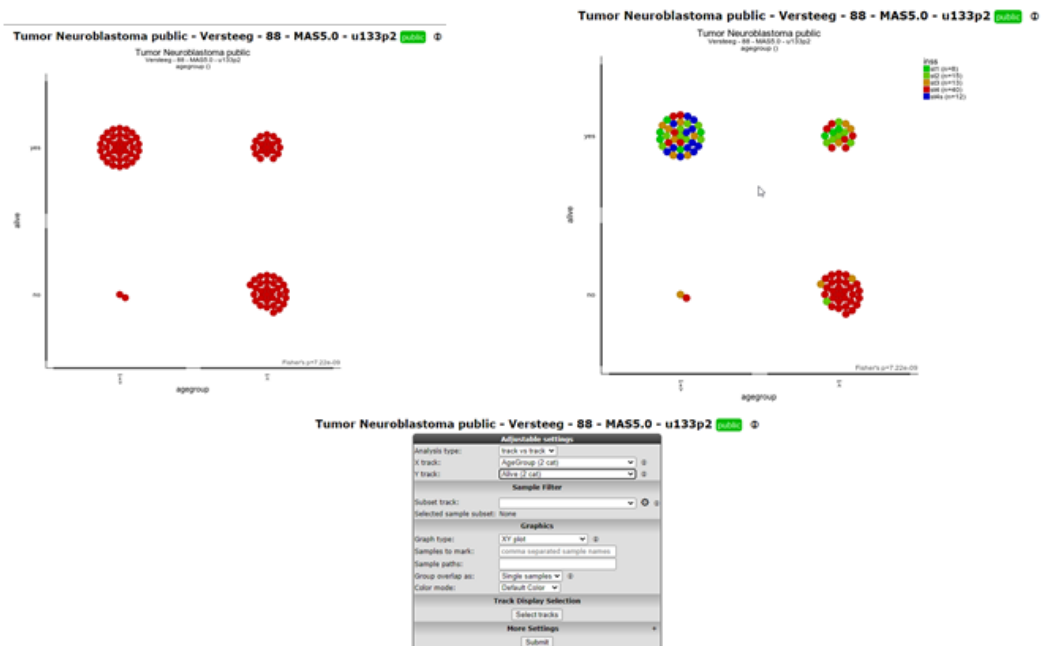


Figure 3: Color samples by track

Figure 3a: Using the interactive plot options

- To compare the absolute or relative shares of track values between subgroups of another track, you can use the “Stacked Barplot” or “Stacked Barplot ratio” respectively. The “Stacked Barplot ratio” option scales every group to 100%, and thereby shows the relative contribution of the different groups. Also the grouped bar plot is a handy visualisation to split your representation for a group parameter.

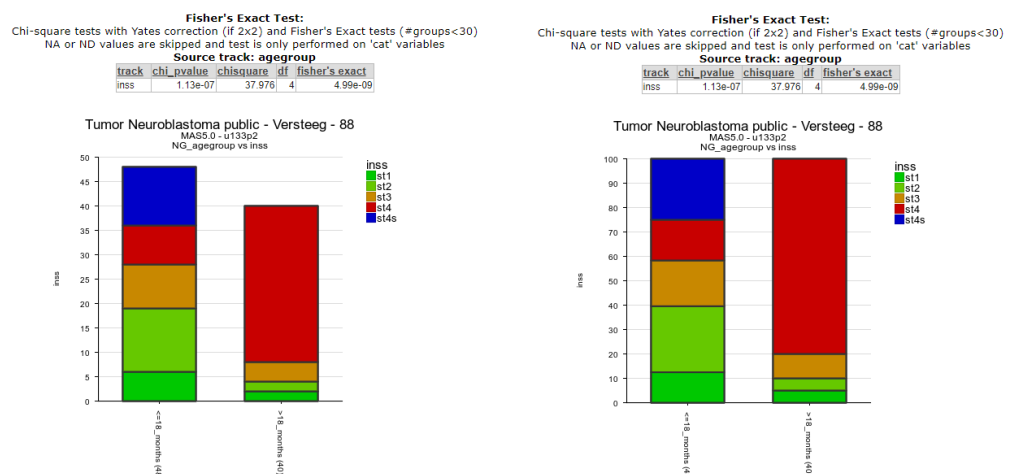


Figure 4: Absolute (left) and relative (right) stacked barplots and in the middle a grouped bar plot

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 click ‘Next’. 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.

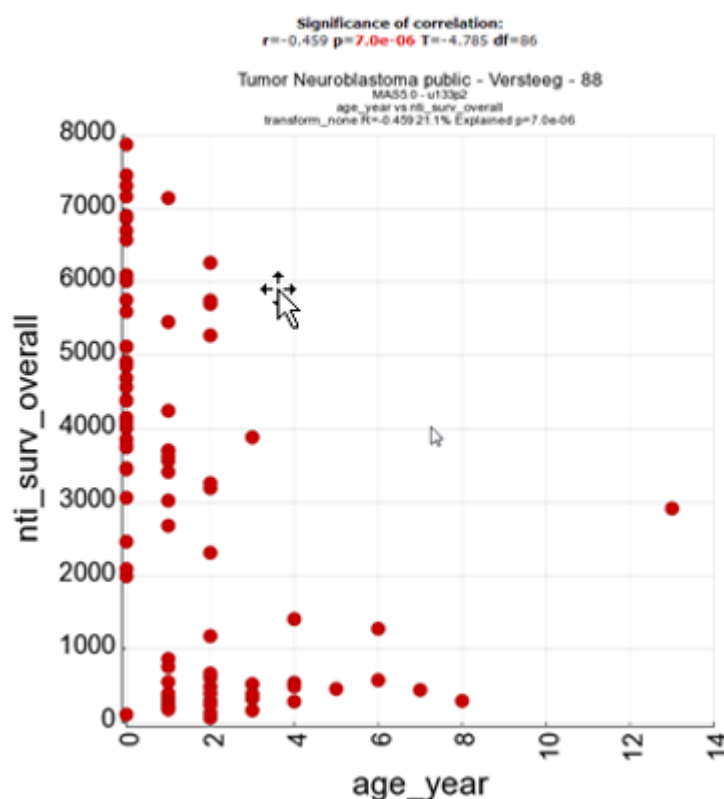


Figure 5: Output of relating numerical tracks

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 stacked box plot, change the colormode 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.

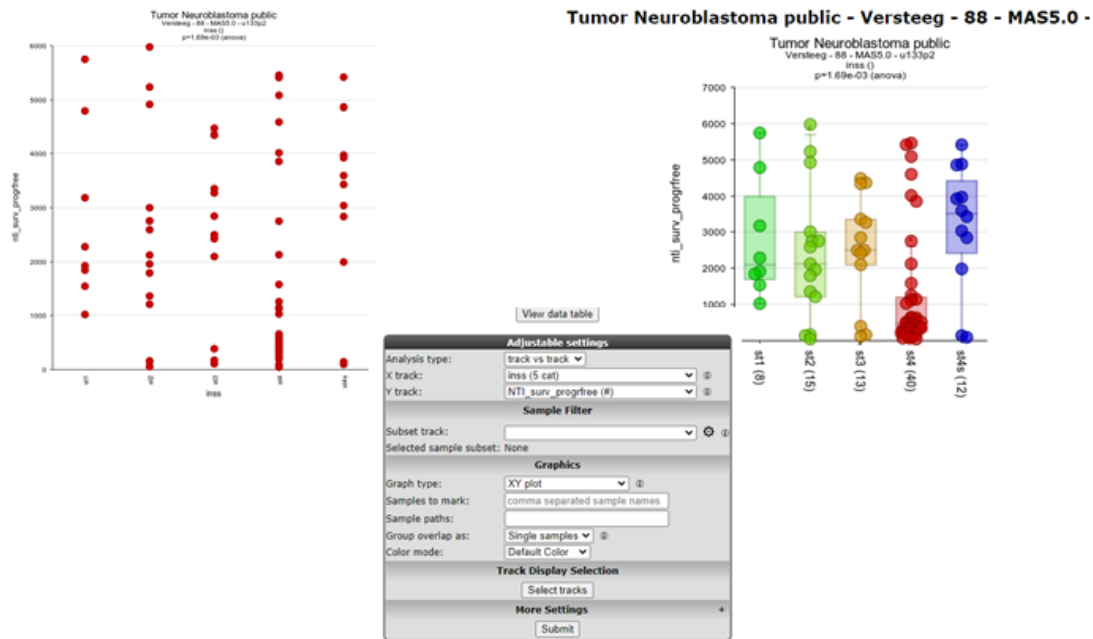


Figure 6: Representing the relation between categorical and numerical tracks

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 three 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.

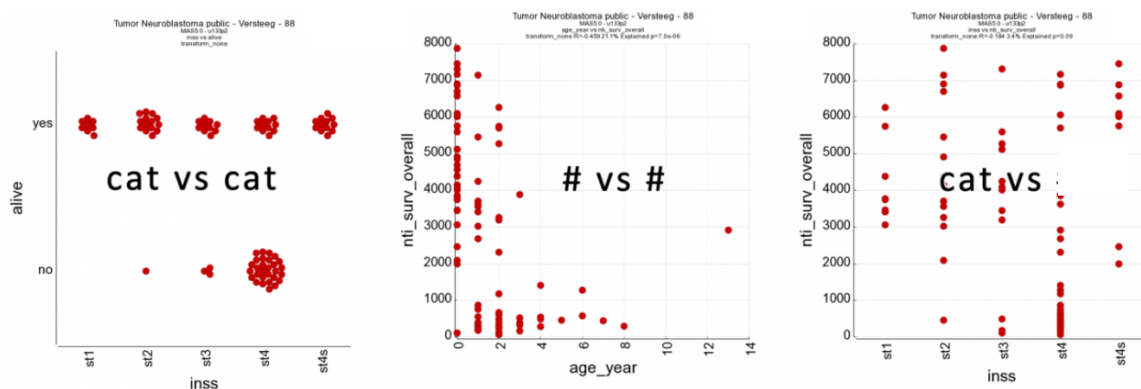


Figure 6: “Representations of relations between different types of tracks in R2

5.5 Step 4: Annotation plotter and Cohort Overview

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 “Select 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 as well as the order in which the samples should be ordered. For removing tracks, drag the trackname out-side the selection box and click submit. Hold en drag the track name to adapt the order in the list of the right selection panel. 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.



Figure 7: Plotting the annotationtracks

4. Another often useful overview is provided by the Cohort Overview. Click on ‘Go to Main’ in the upper left corner, and this time select “Cohort Overview” in box 3 with the “Select type of analysis” dropdown. Then click ‘Next’. Here, pie charts show the shares of the different values of a track. You can visualize tracks with the dropdown menus, hover over the different pie slices, and create a table overview with the values of different tracks of choice for each sample. The ‘Build a track’ button at the bottom of the page conveniently allows you to directly build a track from any selection of available tracks.

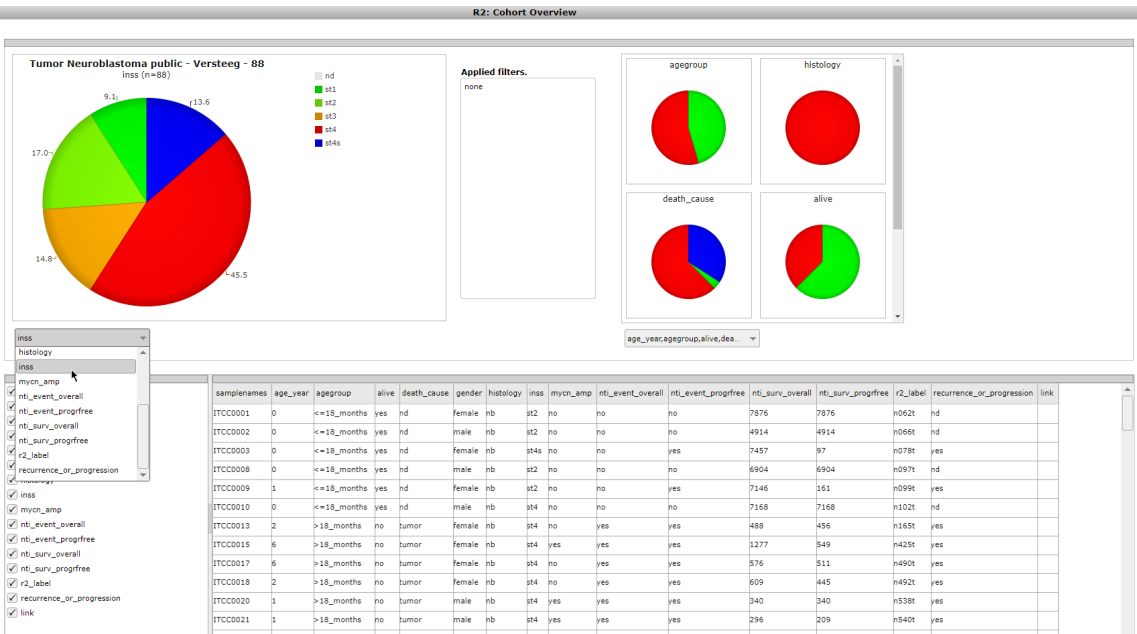


Figure 8: Cohort Overview

5. The sunburst plotter is an often used annotation visualization. In the annotation section of the main menu, select the Cohort sunburst plotter. The sunburst diagram displays a hierarchical structure in a circular shape. The origin of the organization is represented by the center of the circle, and each level of the organization by an additional ring. Additionally, other visualization plot types are implemented as illustrated in Figure 9

(circlepack, treemap and icicle).

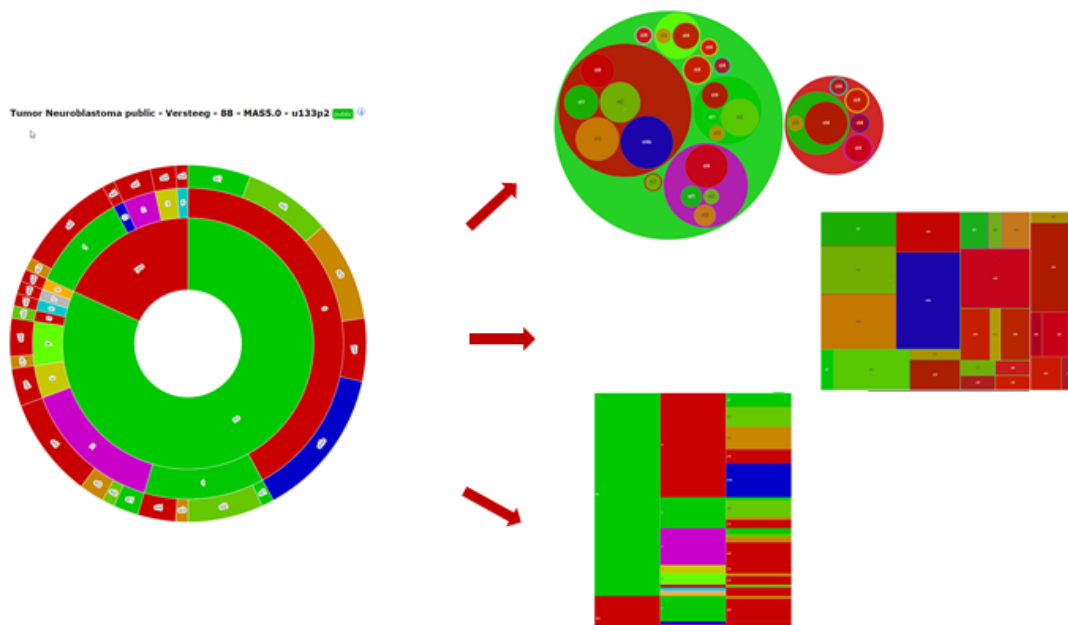


Figure 9: Sunburst plotter

5.6 Final remarks / future directions

Some of these functionalities have been developed recently. If you run into any quirks or annoyances do not hesitate to contact R2 support (r2-support@amsterdamumc.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 (steps 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.
- These settings will be elaborated upon in separate boxes.
- The results of these analyses are presented in adaptable graphics.

6.2 Step 1: Selecting data and the type of analysis

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 - Versteeg - 88 - MAS5.0 - u133p2** dataset is selected in field 2 (see chapter 1 of the tutorial for more information about the selection of a dataset).
3. Choose **View a Gene in groups** in field 3 and click Submit.

6.3 Step 2: Choose the gene and the annotation track as grouping variable

In the next screen you will choose the gene of interest and decide which grouping variable to use to establish the differential expression of your gene of interest.

Tumor Neuroblastoma public - Versteeg - 88 - MAS5.0 - u133p2 public ⓘ

Adjustable settings

Analysis type: gene vs track ▼

Gene / Reporter: Search by Reporter advanced

Track: MYCN / 209757_s_at ▼ ⓘ

Transformation: MYCN05 / 207028_at ▼ ⓘ

Sample Filter

Subset track: ▼ ⓘ

Selected sample subset: None

Graphics

Graph type: YY plot with annotation ▼ ⓘ

Extra Graph Option: off ▼ ⓘ

Samples to mark: comma separated sample names

Color mode: Default Color ▼

Track Display Selection

Select tracks

More Settings +

Submit

➔

Gene / Reporter: MYCN 209757_s_at advanced

Track: Alive (2 cat) ▼ ⓘ

Transformation: ▼ ⓘ

Figure 1: Step-by-step scenario to select ‘View a gene in groups’ on the main page of R2

1. Type **mycn** as gene (see Figure 1) in the first textfield in the Adjustable settings box. Select with a mouse click the first reporter in the popup (the first row). The reporter textfield is automatically filled in.

To view the expression of this gene in groups, you can use dataset specific annotation, the so-called “tracks”, as grouping variable in R2.

1. In the dropdown of the setting *Track* select the track called **Alive (2 cat)**. This track contains survival data of the patients from whom the tumor sample was taken.
2. Note that the other fields can be kept as is, the right choices are already provided. Click **Submit**.

The “one way Anova”/“student T test” test will be performed for data on the selected groups (see explanation in the next step).



Did you know that you can create your own tracks?

Many datasets in R2 contain annotations. You can use these annotation tracks as grouping variable. Another option is to create your own annotation track for any dataset in R2. This is explained in a separate tutorial [Adapting R2 to your needs](#).

6.4 Step 3: Anova results / adapting plots

R2 now performs a one-way Anova statistical test on the fly. More information about which test to choose can be found here: [Statistical test: did you know..?](#).

1. Check the graph and the information that is displayed underneath the graph in the resulting window.

The actual result of the ANOVA calculations is shown in the table under the graph; the difference in average expression between the two groups is significant.

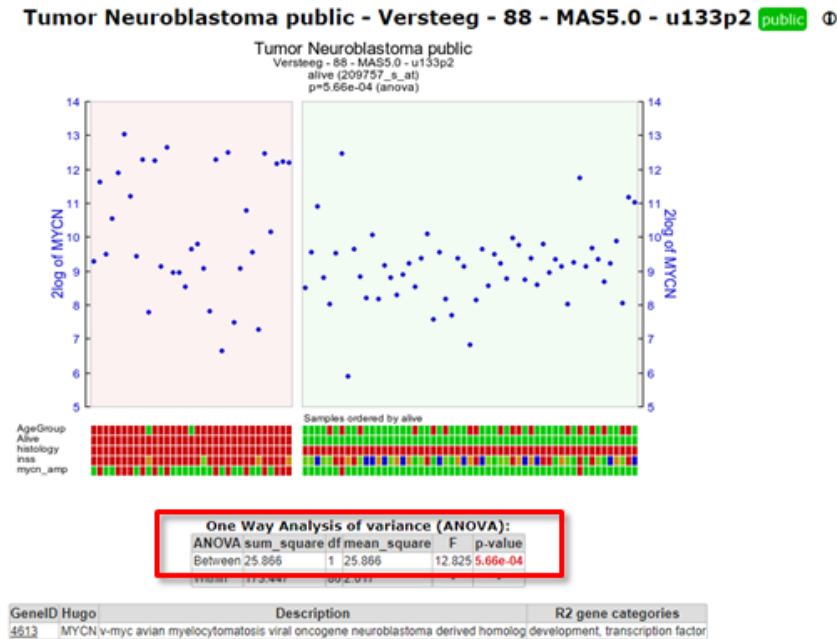


Figure 4: Result of the one-way Anova test for the Neuroblastoma 88 samples.

1. For a nicely ordered graph, you can adjust the settings in the menu at the bottom. Set *Extra Graph Option* to **Track and Gene Sort** (Do not forget to click on Submit 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.

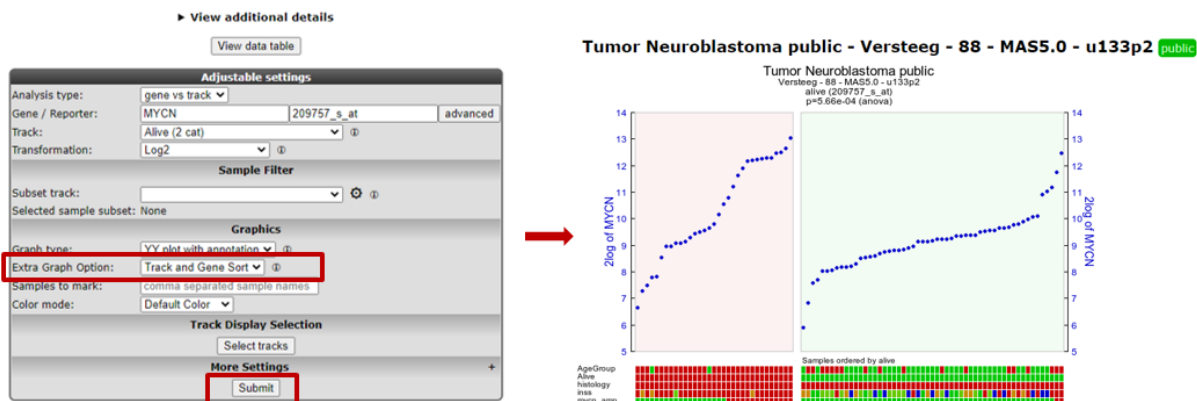


Figure 5: Order the samples per group with increasing expression.

These results can also be shown in different types of plots (Figure 5).

1. Scroll down the window to the Adjustable settings menu.
2. Adapt the selection in the dropdown box *Graph type* to another graph type, e.g. **Dot plot (dots)**, and change *Color mode* to **Color by Track**.
3. Note that you can change the order of the groups with *Order Groups By*, let's take **median (numeric Y)** in our case.
4. Click **Submit**. The resulting graph is adapted accordingly.

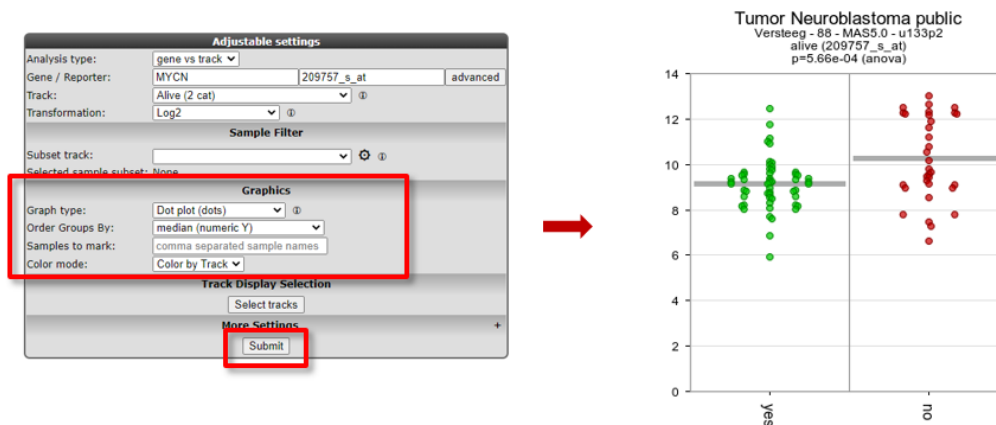


Figure 6: Adapting the Graph type to Dot plot, change order and set Color by Track

The difference in expression between the groups can be shown more dramatically by plotting the data without a log2 transformation. Make sure to use log2 transformation in scientific reports, though, as untransformed mRNA gene expression data is hardly ever normally distributed.

1. In the 'Adjustable settings' menu, set the *Transformation* dropdown to **none** (Figure 7).
2. Click **Submit**.

The resulting 2 graphs also depicted in different types of plots in Figure 6 show the difference between the expression values in the two groups more dramatically.

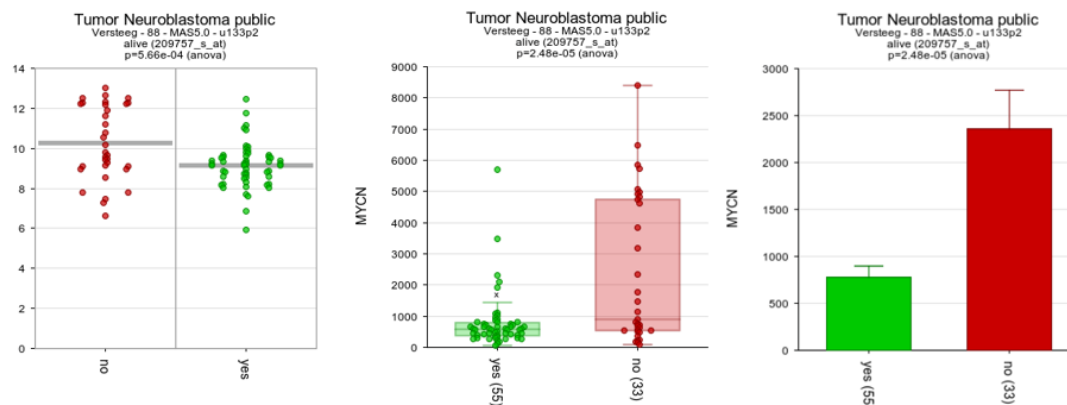


Figure 7: The same data now represented without transformation in bar/box plots



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 on track. First choose a track, then select the wanted subgroups in the track. The analysis will only be performed on the selected subset.

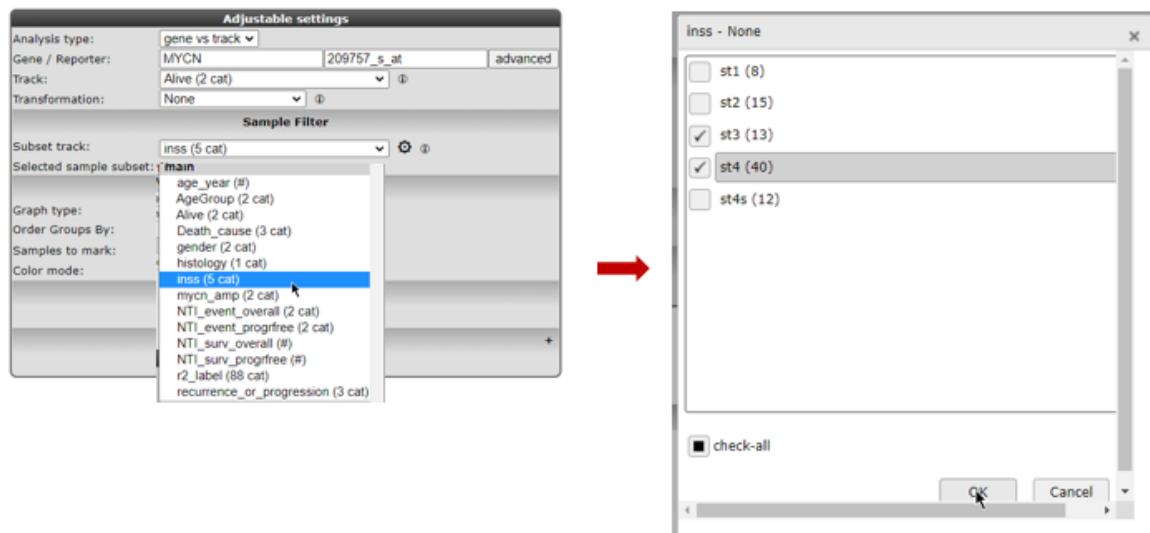


Figure 8: Sample selection with the Sample Filter

Clicking the wheel icon will open a grid supporting all selection combinations of interest

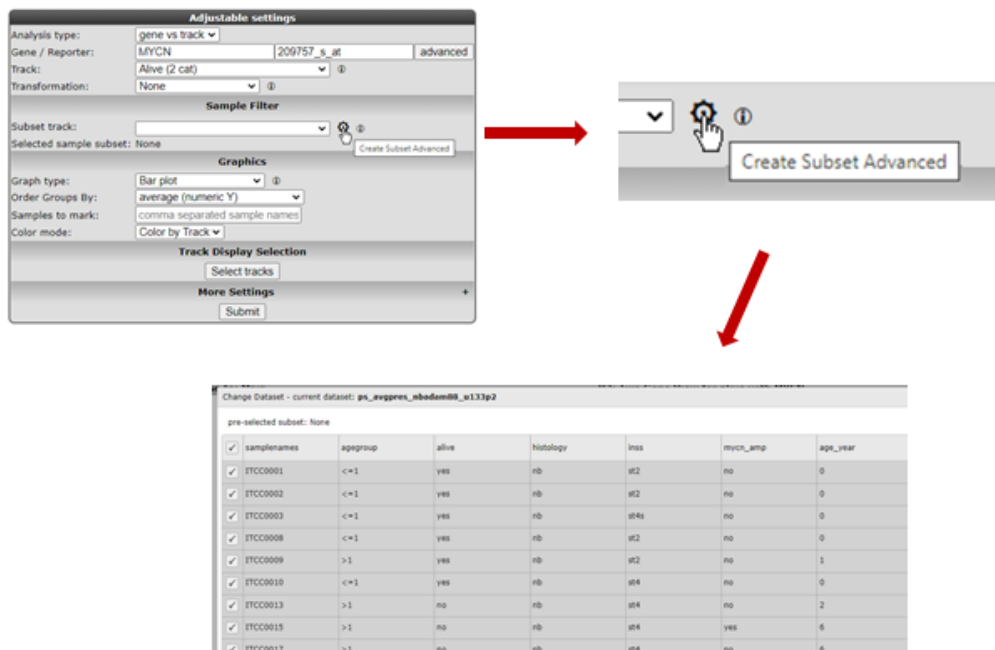


Figure 9: Advanced sample selection

In the samples to mark section, a sample name can be entered that will be highlighted in the resulting graph; ideal for publication purposes. Also note that in the interactive plot section you can mark samples by clicking on the dots.

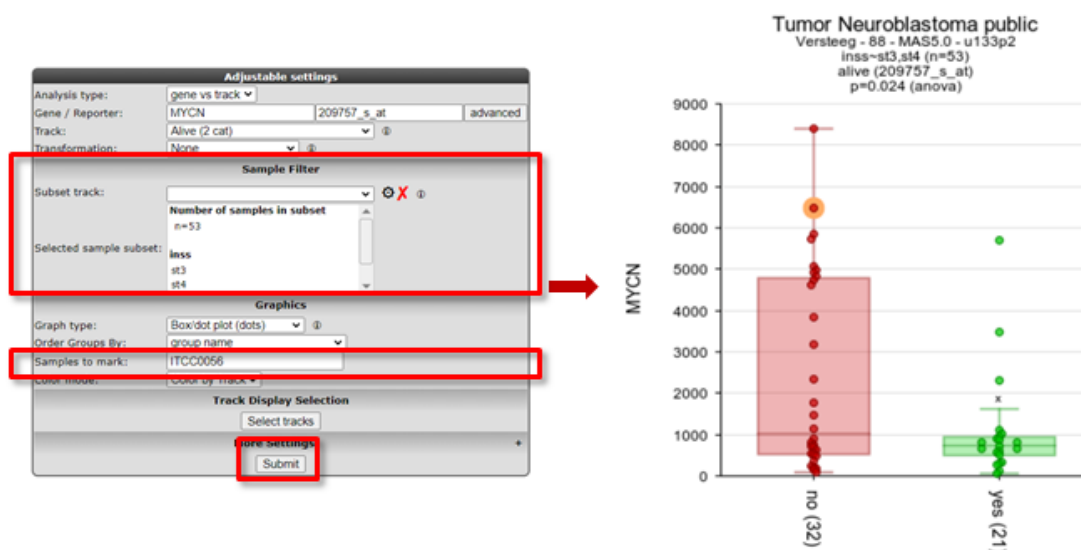


Figure 10: Graph with sample selection INSS 3 and 4 and a marked sample

6.5 Step 4: Finding differentially expressed genes in two groups

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?

1. 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 you find two options to find differential expressed gene lists: 'Find Differential expression between two groups' and Differential expression between multiple groups (Figure 8).
3. Both types of Differential expression modules harbor specific statistical tests. Depending on your chosen dataset, number of groups you want to test and the type of data (RNAseq,microarrays) you can choose from several statistical tests.

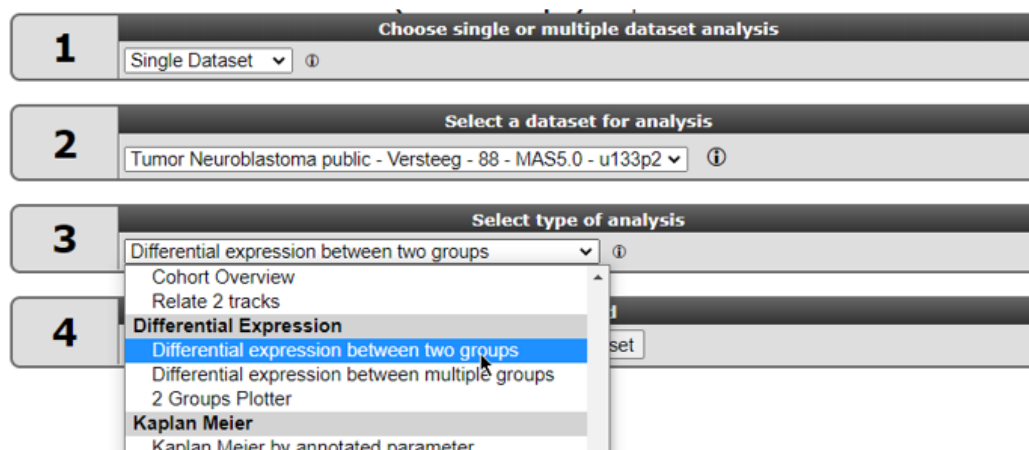



Figure 11: Selecting Find Differential Expression.

4. Select “Differential expression between two groups” and click **Next**.
5. In the next window you can select several types of statistical tests which are present in the selection menu. By default, the **T-test** is selected. We too will use this default test.

Which test is suitable for a given dataset, depends on the normalization of selected dataset and on what kind of data the dataset is build of. Most expression sets are continuous and normally distributed data so the T-test is the most applicable. In case of a dataset which contains categorical data the Mann-whitney test is more suitable.

A special remark for the **DESeq2 algorithm** is at place here. This test is only available for RNAseq data that contains the un-normalised counts. Most of the datasets that have DESeq2_rlog or DESeq2_vst in the name consist of two data parts. A normalised data part is available in case you want to use the T-test and a data part with the counts is available in case the DESeq2 algorithm is preferred. Note that the counts are only used for the test, the values depicted in the graphs etc. are always normalized data. Using the DESeq2 algorithm in case of RNAseq is often appreciated since this is a well established statistical test package dedicated to data such as RNAseq data. In the dataset selection grid box you can search for datasets which have **deseq2_rlog** or **deseq2_vst** as normalization procedure. Datasets with this annotation have three slots, rlog normalized data, deseq normalized data (normcounts) and a counts slot. This counts slot is used when you run the DESeq2 algorithm on the fly for two group comparisons.

Set	Spec	Y	Data type	Y	Categ	Y	Tissue/Tumor	Y	Platform	Y	Normalization	Y	Y	Author	Y	Accession	Y	Release date	Y	R2 date	Y	Access	Y	Perm.	Y
Select	ha	Expression data	Cell line	HN127 PRC2 mutant and WT					gse141437		deseq2_rlog			Kochat		gse141437		2021-06-28		2021-07-26		public			
Select	ha	Expression data	Cell line	Melanoma					erub38e93		deseq2_rlog			Tiffen		gse140673		2019-11-20		2019-11-26		public			
Select	ha	Expression data	Disease	Adipocytes Obese					erub38e83		deseq2_rlog			Strifflers		gse133099		2019-06-21		2020-06-19		public			
Select	mm	Expression data	Disease	Congenital heart					mm38eru99		deseq2_rlog			Garg		gse171237		2019-06-28		2020-06-28		public			
Select	ha	Expression data	Disease	Down Syndrome Liver					erub38e86		deseq2_rlog			Jackson		gse180637		2020-12-04		2020-03-17		public			
Select	ha	Expression data	Exp	Colorectal Pared					gemode32		deseq2_rlog			Russo				2000-01-01		2020-01-23		public			
Select	ha	Expression data	Exp	ES ind. endoderm					gse130340		deseq2_rlog			Peng		gse130340		2021-02-23		2021-03-08		public			
Select	mm	Expression data	Exp	Hepatocytes (Hating-induced changes)					gse121103		deseq2_rlog			Herrig		gse181940		2021-08-30		2021-09-03		public			
Select	ha	Expression data	Exp	Lung COVID19					gse147907		deseq2_rlog			vanDeventer		gse147907		2020-03-23		2020-03-27		public			
Select	mm	Expression data	Exp	Macrophages (LXR pos/Chol)					gse17021		DESeq2_rlog			Uebler		GSE118654		2018-08-17		2018-12-20		public			

Disease Rheumatoid arthritis (Corrona study) - Capila - 128 - deseq2_vst - gpl16791 

Select a test

Test: DESeq2

Group by: rf_status_b1 (2 cat)

Sample Filters

Subset track:

Selected sample subset: none

Submit

Adjustable settings

Group 1: neg (39)

Group 2: pos (89)

Statistics

P-value cutoff: 0.01

Max number of results:

Gene Filters

HugoOnce mode: yes

Min. # Present calls: 1

Minimal maximum value:

Minimal range size: 0

Gene ontology: All

Gene set:

Manual list: none

Search GO

Search GS

Submit

Figure 12: Selecting the DESeq2 test.

6.6 Step 5 Setting parameters

In our case we continue with the Tumor Neuroblastoma dataset and the Differential Expression between two groups analysis with the T-test.

1. Now we also make the choice for the two groups. Select behind *Group by* the track **Alive (2cat)** again. Click **Submit**.

Tumor Neuroblastoma public - Versteeg - 88 - MASS.0 - u133p2 public ⓘ

Test: T-test
Group by: Alive (2 cat) ⓘ

Sample Filters
Subset track: ⓘ
Selected sample subset: None
Submit

Adjustable Settings
Group 1: no (33)
Group 2: yes (55)

Data transformation
Floor value: ⓘ
Transformation: Log2 ⓘ

Statistics
Corr. multiple testing: False Discovery Rate ⓘ
P-value cutoff: 0.01
Max number of results: ⓘ

Gene Filters
HugoOnce mode: yes ⓘ
Min. # Present calls: 1 ⓘ
Minimal maximum value: ⓘ
Minimal range size: 0 ⓘ
Chromosome: All ⓘ
Gene ontology: All ⓘ
Gene set: ⓘ
Manual list: none ⓘ
Submit

Figure 13: Differential expression parameters

2. An extra menu shows up with many options. For now we only adjust the required Group 1 and Group 2 setting; we choose the value **no (33)** for *Group 1* and **yes (55)** for *Group 2*.
3. Click **Submit**.

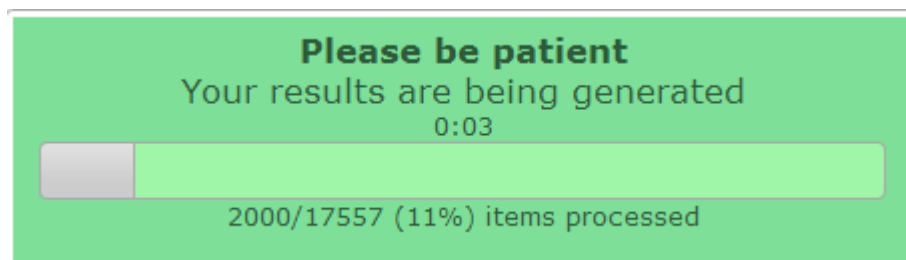


Figure 14: Progress dialog during on the fly calculation

The result is a list of genes that is ordered by the most significant differential expression between the groups that you chose (Figure 15). 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. The generated list can be sorted or filtered by any of the column headers in the grid, such as by the p-value (P) or the difference.

In the right menu numerous modules can be selected to continue the analysis. Also, the generated list can be extracted to continue for further usage outside R2. As indicated in the right menu

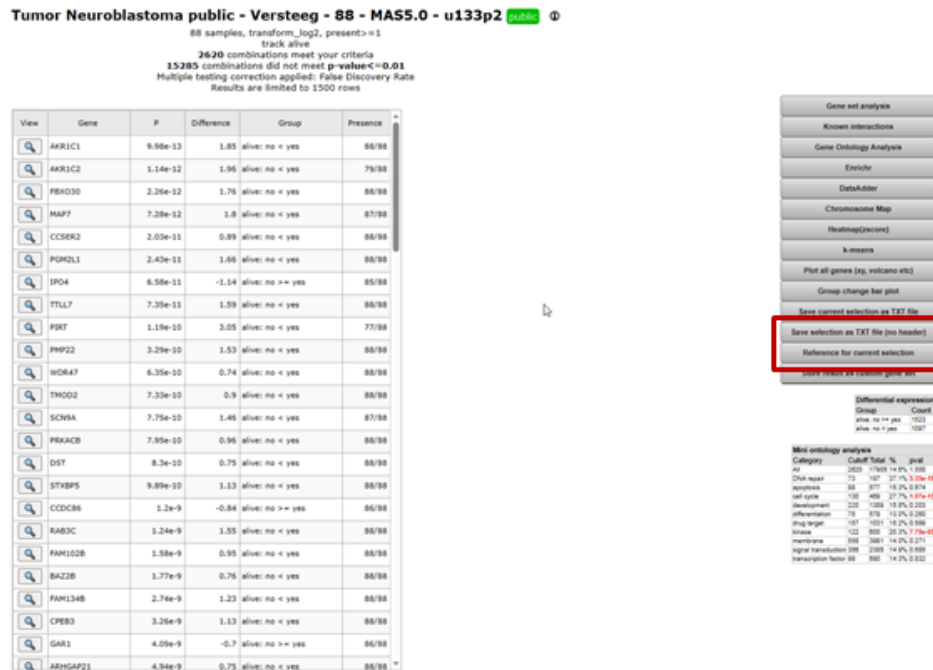


Figure 15: Genes differentially expressed between groups.

6.7 Step 5: Correct for paired analysis

A paired analysis is often performed when observations are natural paired or matched such as in this example.

1. Select in the main screen the following dataset. Exp Neuroblastoma Adrn Mes resistant - George - 12 - tpm - gse165748 and differential expression between two groups, select group by cell_lineage and click submit. Subsequently select adrenergic and mesenchymal for the groups and click the lower submit button. Take a look at the number of found combinations and continue with adapting the settings in the adjustable settings box below the list of combinations.

Exp Neuroblastoma Adrn Mes resistant - George - 12 - tpm - gse165748 public ⓘ

Select a test

Test: Limma

Group by: cell_lineage (2 cat) ⓘ

Sample Filters

Subset track: ⓘ

Selected sample subset: None

Exp Neuroblastoma Adrn Mes resistant - George - 12 - tpm - gse165748 public ⓘ

12 samples, transform_log2, present>=1
track cell_lineage
1885 combinations meet your criteria
16106 combinations did not meet **p-value<=0.01**
Multiple testing correction applied: False Discovery Rate
Results are limited to 1500 rows

View	Gene	P	Difference	Group	Presenc
	TRIM67	1.52e-10	-4.57	cell_lineage: adrenergic...	12/12
	PHACTR1	5.3e-10	-3.97	cell_lineage: adrenergic...	12/12
	DPYSL5	1.36e-9	-6.51	cell_lineage: adrenergic...	12/12
	GNAO1	1.61e-9	-6.02	cell_lineage: adrenergic...	12/12
	FAM57B	2.65e-9	-3.9	cell_lineage: adrenergic...	12/12
	RTL1	2.65e-9	-6.03	cell_lineage: adrenergic...	12/12
	CHRNIB2	3.37e-9	-4.77	cell_lineage: adrenergic...	12/12
	ALK	3.91e-9	-5.8	cell_lineage: adrenergic...	12/12
	UNC79	4.29e-9	-4.82	cell_lineage: adrenergic...	11/12
	SINAP91	4.31e-9	-5.42	cell_lineage: adrenergic...	11/12
	ADAMTSL2	5.14e-9	-4.23	cell_lineage: adrenergic...	11/12
	XKR7	6.41e-9	-5.92	cell_lineage: adrenergic...	11/12
	TMEM151B	9.36e-9	-4.67	cell_lineage: adrenergic...	12/12
	MYT1L	9.49e-9	-4.96	cell_lineage: adrenergic...	11/12
	SH2D3C	1e-8	-4.37	cell_lineage: adrenergic...	11/12
	HS3ST2	1.04e-8	-5.95	cell_lineage: adrenergic...	9/12
	UNC13A	1.31e-8	-6.18	cell_lineage: adrenergic...	12/12
	SCRT2	1.42e-8	-5.24	cell_lineage: adrenergic...	9/12
	DISP3	1.46e-8	-4.21	cell_lineage: adrenergic...	12/12
	PAX5	1.51e-8	-5.32	cell_lineage: adrenergic...	12/12

Figure 14: Genes differentially expressed between groups..

- In the select a test box, select Limma and again for Group by: Cell_lineage and click submit. Select Adrenergic and mesenchymal for the groups and now you can select a track you want to correct for in this case, the genomic_mycn_status, click the lower submit button.

Exp Neuroblastoma Adrn Mes resistant - George - 12 - tpm - gse165748 public ⓘ

Select a test

Test: Limma

Group by: cell_lineage (2 cat) ⓘ

Sample Filters

Subset track: ⓘ

Selected sample subset: None

Adjustable settings

Group 1: adrenergic (6)

Group 2: mesenchymal (6)

Correct for: genomic_mycn_status (2 cat) ⓘ

Exp Neuroblastoma Adrn Mes resistant - George - 12 - tpm - gse165748 public ⓘ

12 samples, transform_log2, present>=1
track cell_lineage
3091 combinations meet your criteria
14900 combinations did not meet **p-value<=0.01**
Multiple testing correction applied: False Discovery Rate
Results are limited to 1500 rows

View	Gene	P	Difference	Group	Presenc
	SLC8A3	2.71e-11	-6.15	cell_lineage: adrenergic...	12/12
	RIMBP2	8.79e-11	-6.37	cell_lineage: adrenergic...	12/12
	NRSN1	1.29e-10	-5.65	cell_lineage: adrenergic...	10/12
	PAX5	2.06e-10	-5.32	cell_lineage: adrenergic...	12/12
	SRRM4	2.06e-10	-5.91	cell_lineage: adrenergic...	10/12
	MMP24	3.41e-10	-3.37	cell_lineage: adrenergic...	12/12
	HPCAL4	3.41e-10	-5.55	cell_lineage: adrenergic...	12/12
	ATP1A3	5.54e-10	-5.38	cell_lineage: adrenergic...	12/12
	SASH1	5.54e-10	5.17	cell_lineage: adrenergic...	12/12
	PHACTR1	9.81e-10	-3.97	cell_lineage: adrenergic...	12/12

Figure 15: Genes differentially expressed between groups with correction.

3. After correction for the genomic_mycn status more genes are found to be significant differentially expressed between the two groups. For example also the PAX5 gene appears higher and more significant in the list which could be a candidate for further investigation.

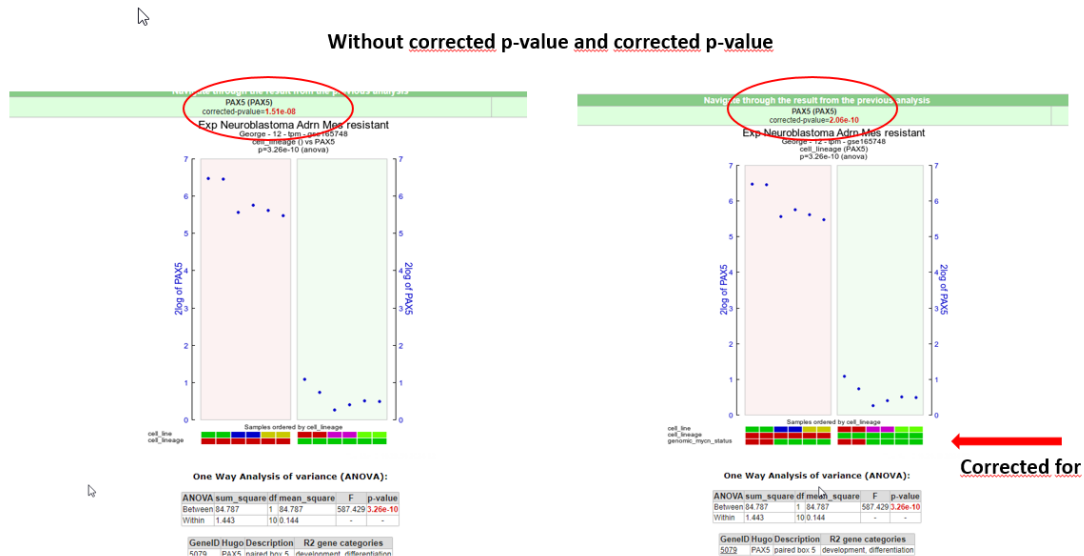


Figure 16: Genes differentially expressed between groups with correction.



Did you know that...

Very useful background information for this tutorial can be found in Chapter 25 Concepts of R2: did you know..?

Check it out:

What were those R and p-values again?: 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. ... Read all about R & p-values in [Chapter 25](#)

You can specify the preferred statistical test and choose a subset of genes? Use any (combination) of the following parameters to adapt the analysis to your needs.

- **Hugo Once (hugoonce):** For most analysis genes should only be reported once in a dataset. R2 uses an algorithm called Hugoonce to choose a single probe-set to represent a gene. Scroll down in [Chapter 25](#) to the Settings section about Hugo Once.
- **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. In addition to 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.
- **Correction for multiple testing:** We are testing a lot of genes here; so we have to correct for multiple testing. Why? Read on about multiple testing in [Chapter 25](#)
- **Gene Filters:** As for many analyses in R2, the gene filters allow you to study a specific subset of genes for differential expression. There are several domains you can choose from. Learn more about gene filters in [Chapter 25](#)

Of course, to actually get familiar with these settings you should not only read about it, but also toy around with them!

6.8 Step 6: Find differential expression in multiple groups

As mentioned above, Find Differential Expression for multiple groups can also be applied with a slightly different “Adjustable settings menu” and including other types of statistical tests. Read further about which test to use in [Chapter 25](#).

1. Go back to the **Main** page by the link in the upper left corner and select again the Neuroblastoma 88 set.
2. Select **Differential expression between multiple groups** and click **Next**
3. Select for *Group* by the value **inss (5 cat)** and leave all the other settings at their default value. Click **Submit**.

Tumor Neuroblastoma public - Versteeg - 88 - MAS5.0 - u133p2 public ⓘ

Adjustable settings

Test: ANOVA

Group by: inss (5 cat)

Sample Filters

Subset track:

Selected sample subset: None

Data transformation

Floor value:

Transformation: Log2

Statistics

Corr. multiple testing: False Discovery Rate

P-value cutoff: 0.01

Max number of results:

Gene Filters

HugoOnce mode: yes

Min. # Present calls: 1

Minimal maximum value:

Minimal range size: 0

Chromosome: All

Gene ontology: All

Gene set:

Manual list: none

Submit

ANOVA

ANOVA

Kruskal-Wallis

Pairwise t-tests (Welch)

Figure 17: Genes differentially expressed between groups.

1. A list of differentially expressed genes between the groups is generated. Of course, now that we have more than two groups, the table no longer contains the Difference column and group order column.

6.9 Step 7: Inspecting single genes

1. Choose one of the genes in the table to inspect further.
2. Hover over the magnify symbol in the list next to the gene name to find a description of the gene.
3. Now click on the magnify symbol. A similar graph is produced as for TF, the differential expression is more pronounced for this gene (Figure 18). In stage 4s, even indicating based on the TF expression that there is possible subgroup within the INSS 4s stage.
4. In the generated picture the samples are not ordered by their gene expression. Go to the adjustable settings menu and select **Track and gene sort** in the *Extra Graph Option* pulldown menu. Click **Submit**.

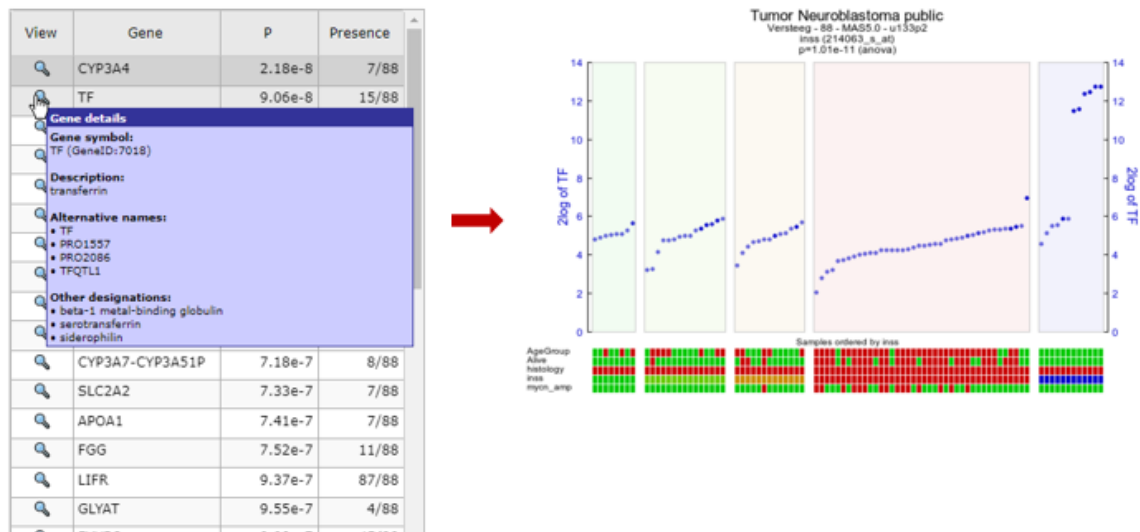


Figure 18: Hover over and click on any gene of interest

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

1. The tab with the list of differentially expressed genes (Figure 15) is still open or perform the analysis again. 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 etc.). We will explore one additional data visualization however to plot all genes of this analysis. In the right menu click see Figure 15 'Plot all genes (xy, volcano, etc.)'.
3. The resulting plot shows all genes of the list in a XY-plot, MA-plot and volcano plot depending on your choice; in case of the XY datapoints above and below the diagonal are differentially expressed. Hovering over the points shows the genesymbol, clicking on the dots will annotate the dots with the gene name., 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. Adapting the plot parameters in the Adjustable settings menu will also adapt the graph on the fly.

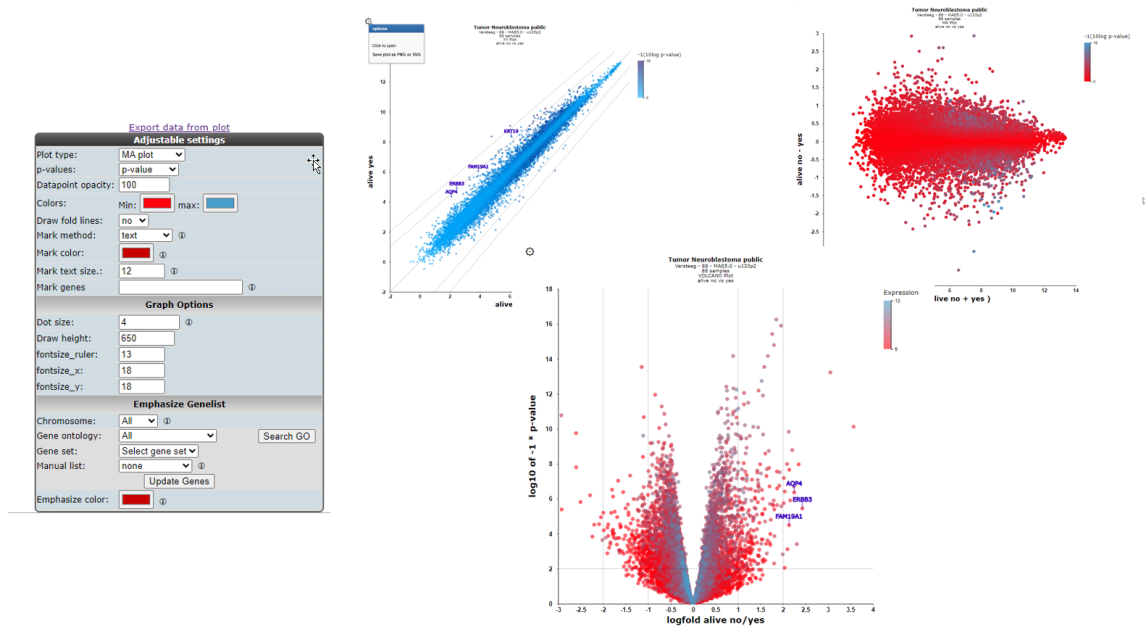


Figure 19: XY, MA , Volcano plot of all genes differentially expressed in the current track;

- Right clicking on the datapoint in the plot opens up a new window showing the expression of the gene in the two groups as a violin plot.

Tumor Neuroblastoma public - Versteeg - 88 - MAS5.0 - u133p2 public ⓘ

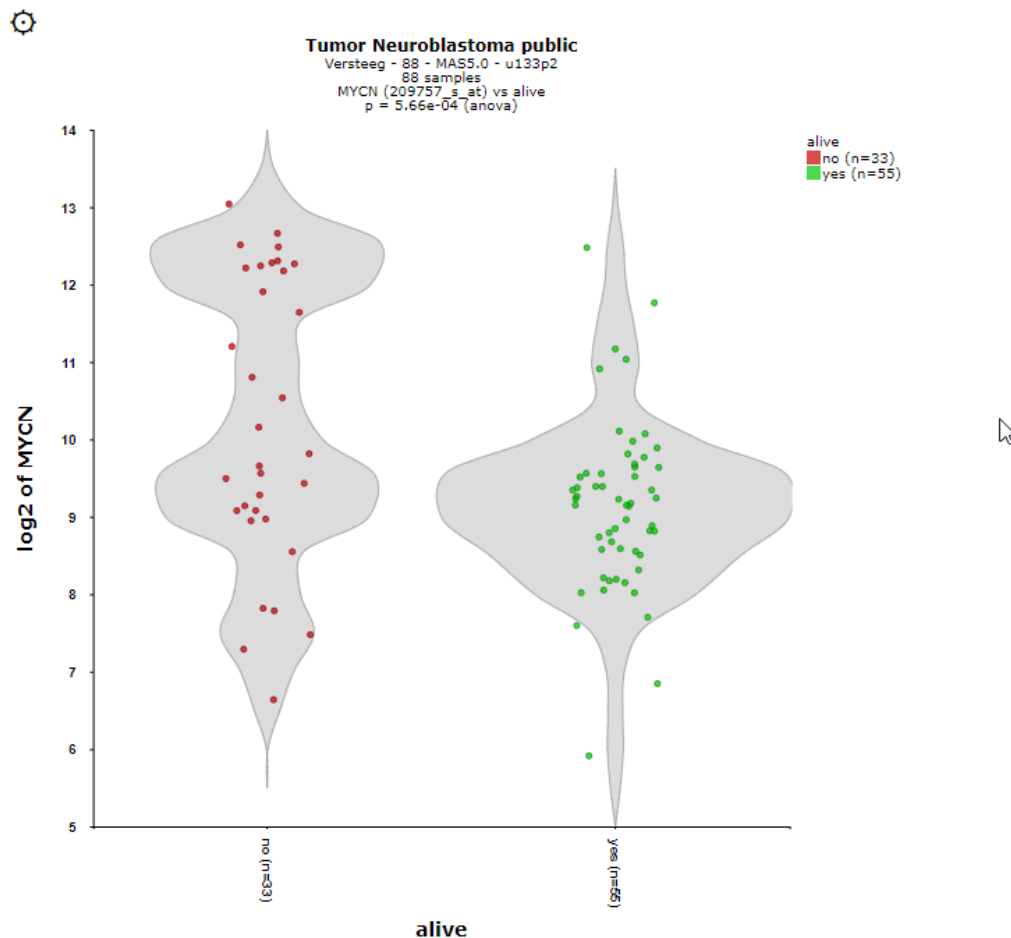


Figure 20 : Differential expression of MYCN

5. The plot has been adapted to show the AKR1C1 genesymbol also DNA-replication genes are highlighted in red. Fold change lines show the regions where differential expression is 1 and 2 fold (Figure 16). Note that most genes of the DNA replication pathway seem to be located below the diagonal.

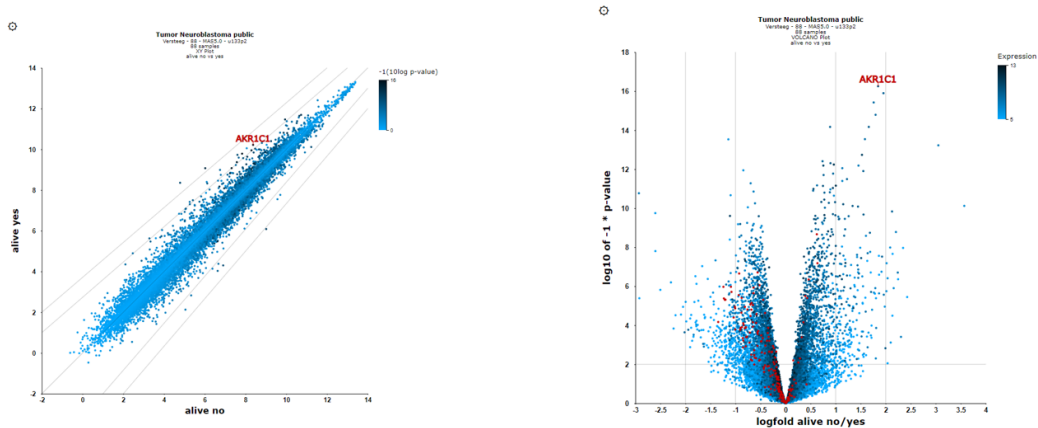


Figure 21: Adjusted visualization of gene expression, hovering over the dots shows the gene name.

6. 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 21). 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
7. In another example in the selected Coloncarcinoma TCGA set, some Ribosomal gene categories which were selected in the gene filter. The KRT16-gene was selected and adapted in the Adjustable settings box.

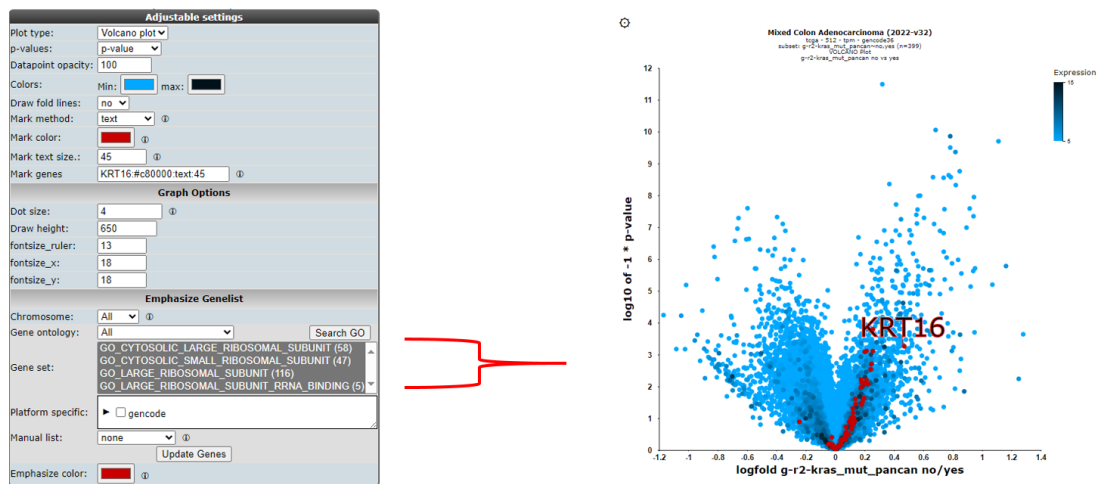


Figure 22 : Adjustable settings for the all genes plot

6.11 Step 9: Using the Enrichr

The right menu also allows you to take your result list of differentially expressed genes (DEG) outside R2 to the public available Enrichr platform. Enrichr (<https://maayanlab.cloud/Enrichr/enrich>) is a web-based platform designed for gene set enrichment analysis (GSEA) and functional annotation of gene lists. It allows you to gain insights into the biological processes, pathways, and functions associated with their gene sets of interest. The Enrichr performs an enrichment analysis by comparing the generated R2-list against a large collection of well curated databases such as Gene Ontology, KEGG pathways and disease-associated gene sets

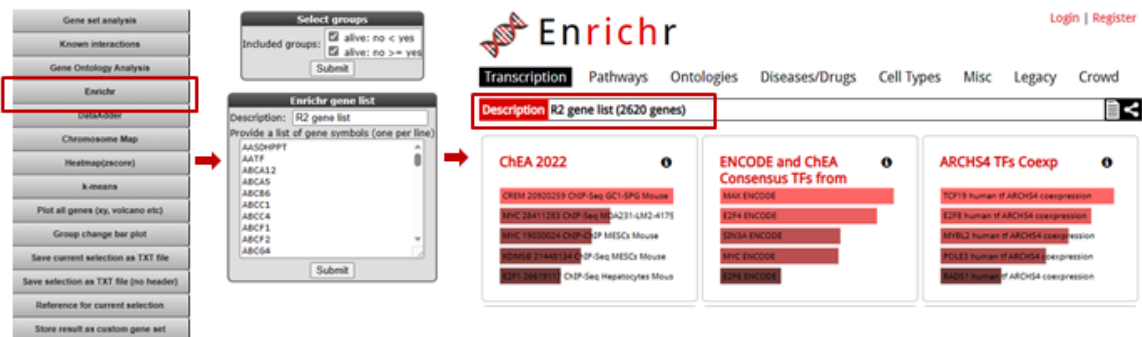


Figure 23: Taking the result to the Enrichr platform.

Figure 23 shows the Enrichr button you can click directly when the DEG list genes is ready, in the next screen you can select just one group when coming from the two group analysis or you want to include and even add or delete genes from list. Hitting the submit button will direct lead to the Enrichr platform.



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.
- Clicking on the dots will annotate the dots with the gene name.
- 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”*

6.12 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” option 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) and click “Next”.
4. In the next screen, type ‘MYCN’ in the gene/reporter field and select the first reporter.

5. Click “Submit”.

Figure 1: Choice of correlation analysis.

6. In the adjustable settings, we set the p-value cut-off to 0.01 and leave the further settings at their default. Note in Figure 1 that you can select for both correlation directions or a single one. The p-value, r-value cut-offs and multiple testing can be adapted. Scroll down the screen and click “Submit”.



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

Just choose ‘Correlate 2 genes’ in the main menu in box 2 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 are 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’ chapter.

7.3 Step 2: 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 ($p < 0.001$) 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, to 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).

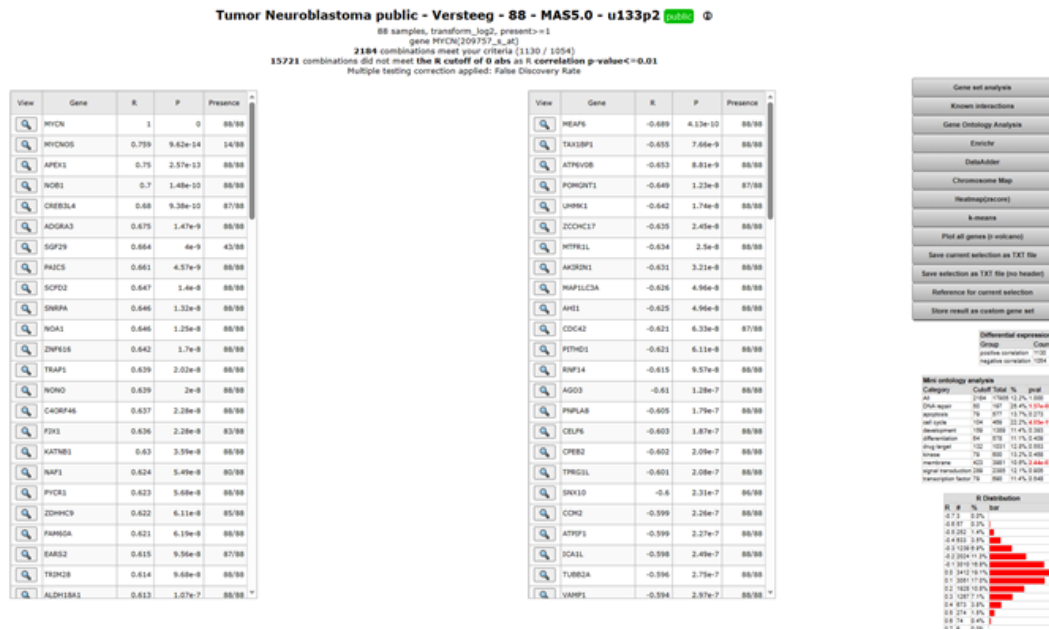
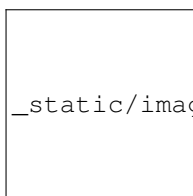


Figure 2: Genes whose expression correlate with that of the MYCN gene in 88 Neuroblastoma tumors

Exact (gene-) numbers listed in the tutorial. Figures such as in this example (2184 combinations) can vary. This is caused by database updates upon a new genebuild release or from a commercial platform such as Affymetrix annotation update.

1. The right table summarizes the genes that show a negative correlation; the expression of MYCN behaves opposite 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. At the bottom of the generated list, the adjustable settings menu is located where the filter options can be adapted.



_static/images/CorrelatingGenes/FindGenes_gotoadjust.png

Figure 3: Adjustable settings

7.4 Step 3: 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.

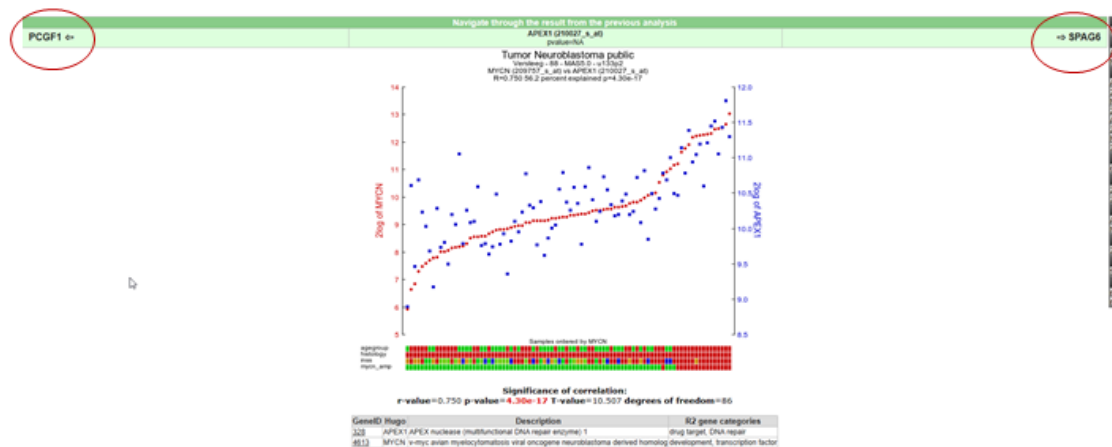


Figure 4: The expression of the MYCN gene correlates with the expression of the APEX1 gene.

- As an example of the opposite, click on one of the top genes in the right column figure 2 ; MEAF6. This produces Figure 6. The original list of results is still open in another tab in your browser, return there.

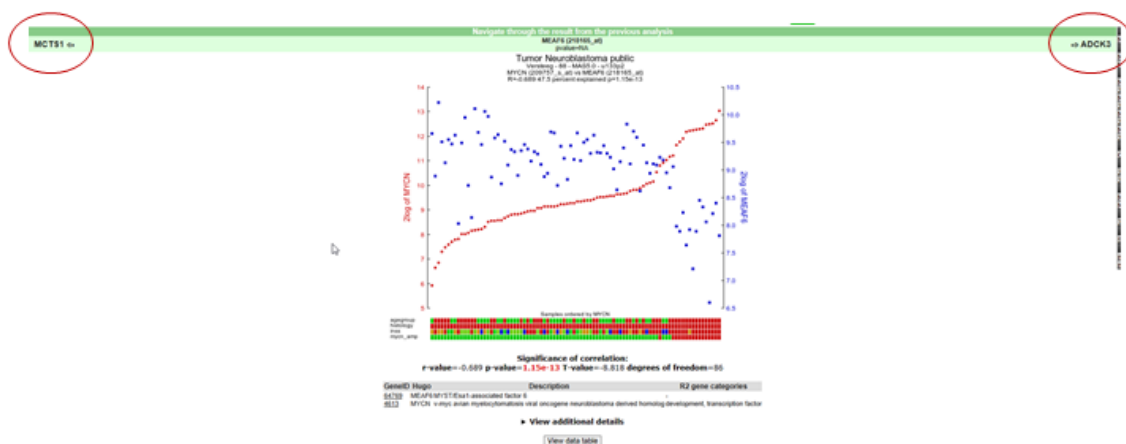


Figure 5: The expression of MYCN has a negative correlation with that of the MEAF6 gene

- To generate a correlation plot where the negative relationship between MYCN and the MEAF6 gene is more clearly visualized, select "XY-plot" as graph type in the graphics section in the Adjustable Settings box and click the "Submit" button. In this correlation plot it is 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, set Histogram to yes, and click on the Adjust Settings button. Now the histogram boxes in the x and y axes show the distribution of the expression levels in the correlation plot, see fig 6. In the green bar located above your graph you can easily click through your generate list instead of going each time to the list to inspect your genes.

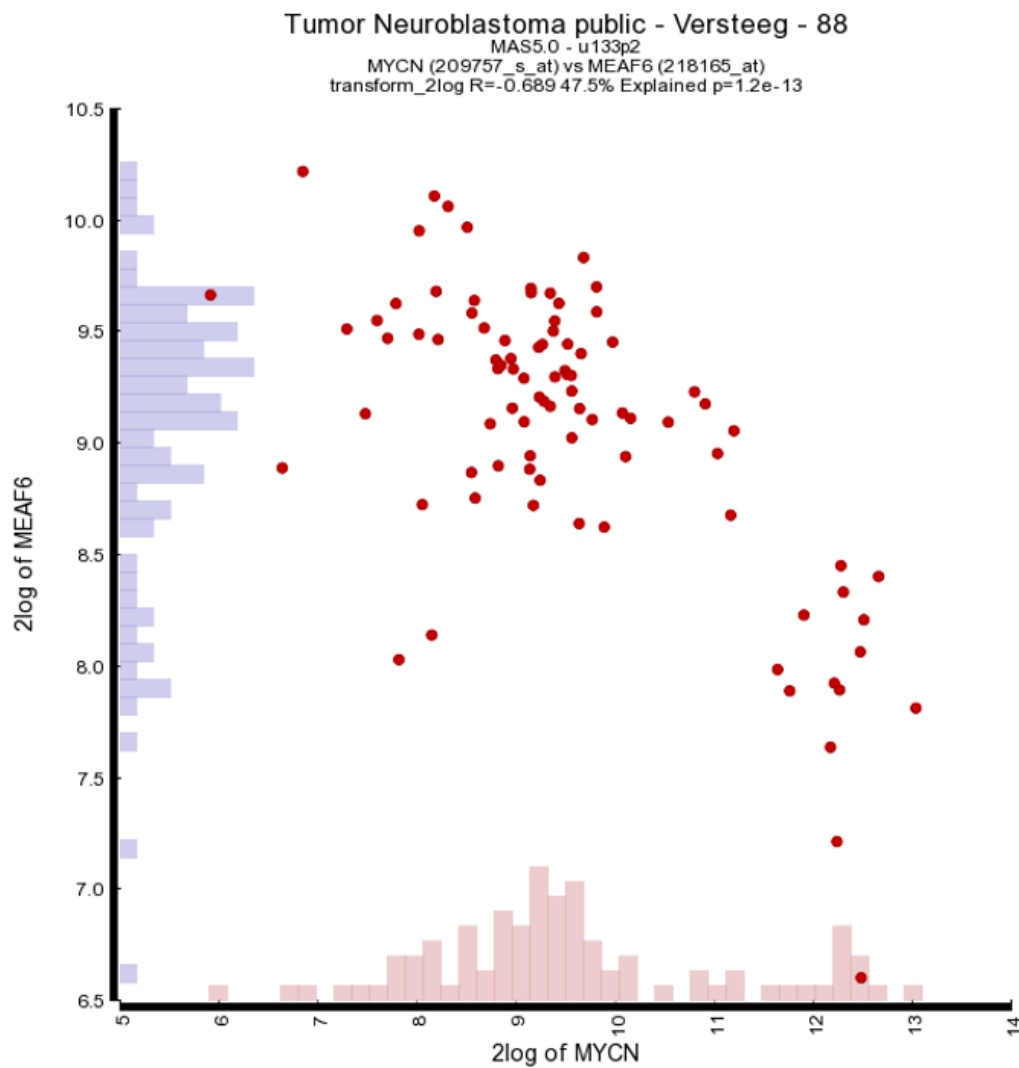


Figure 6: Toggle Histogram

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

Figure 7: Select Color by gene

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 MEAF6 are plotted and subsequently colored by the MYCN expression levels. Of course, you can also enter a third gene for coloring the dots.

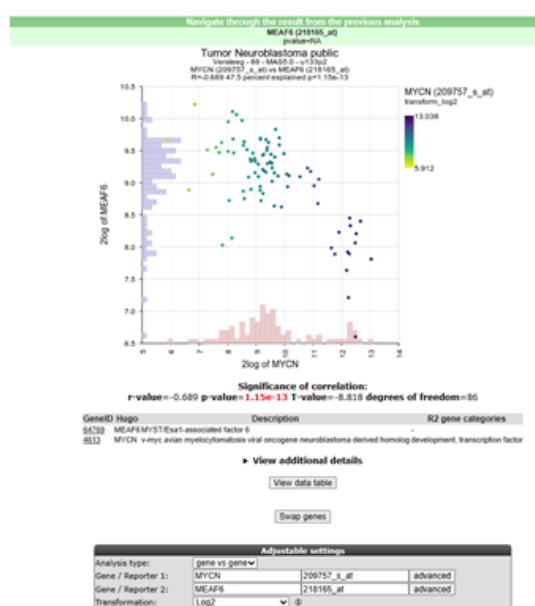


Figure 8: Select Color by gene expression

- Another way to visualize the relationship of the expression correlation in an XY plot is to switch on the linear fit option. In the “More settings section”, turn on “linear fit”.

More Settings

Min (X):

Max (X):

Color (X):

Min (Y):

Max (Y):

Color (Y):

Mark method:

Vector (SVG) output:

Draw height:

Dot size:

Add Boxplot per group:

Linear Fit:

fontsize_y:

fontsize_ruler:

fontsize_t1:

fontsize_tsub:

fontsize_legend:

fontsize_legend_header:

axis_width:

Data point opacity:

Shadows:

3D:

Makeup:

Grid lines:

Figure 9: Select Linear fit

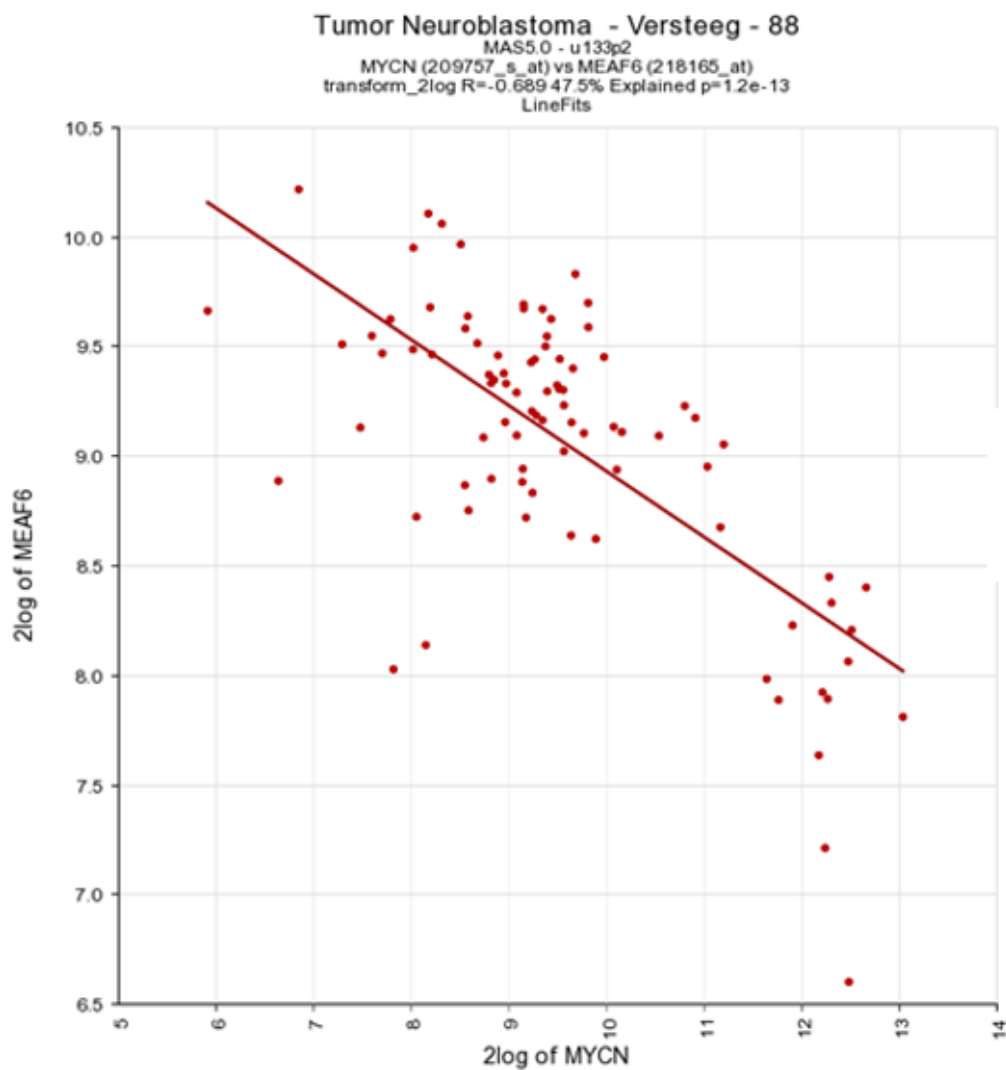
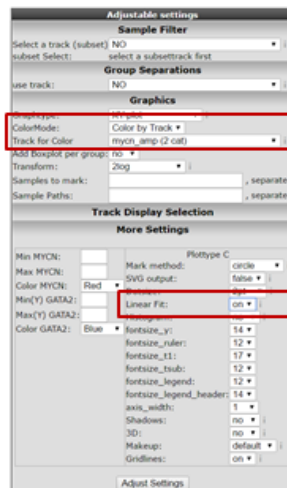


Figure 10: Linear fit

6. 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 the relationship between MYCN and GATA2. In the adjustable settings box enter MYCN and the GATA2 gene and click 'Change Genes'. Change the color mode to "Color by track", select the MYCN track in the graphics section, turn on the linear fit option at 'More settings' as indicated below, and click redraw.

**Figure 11: Linear fit adjust**

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

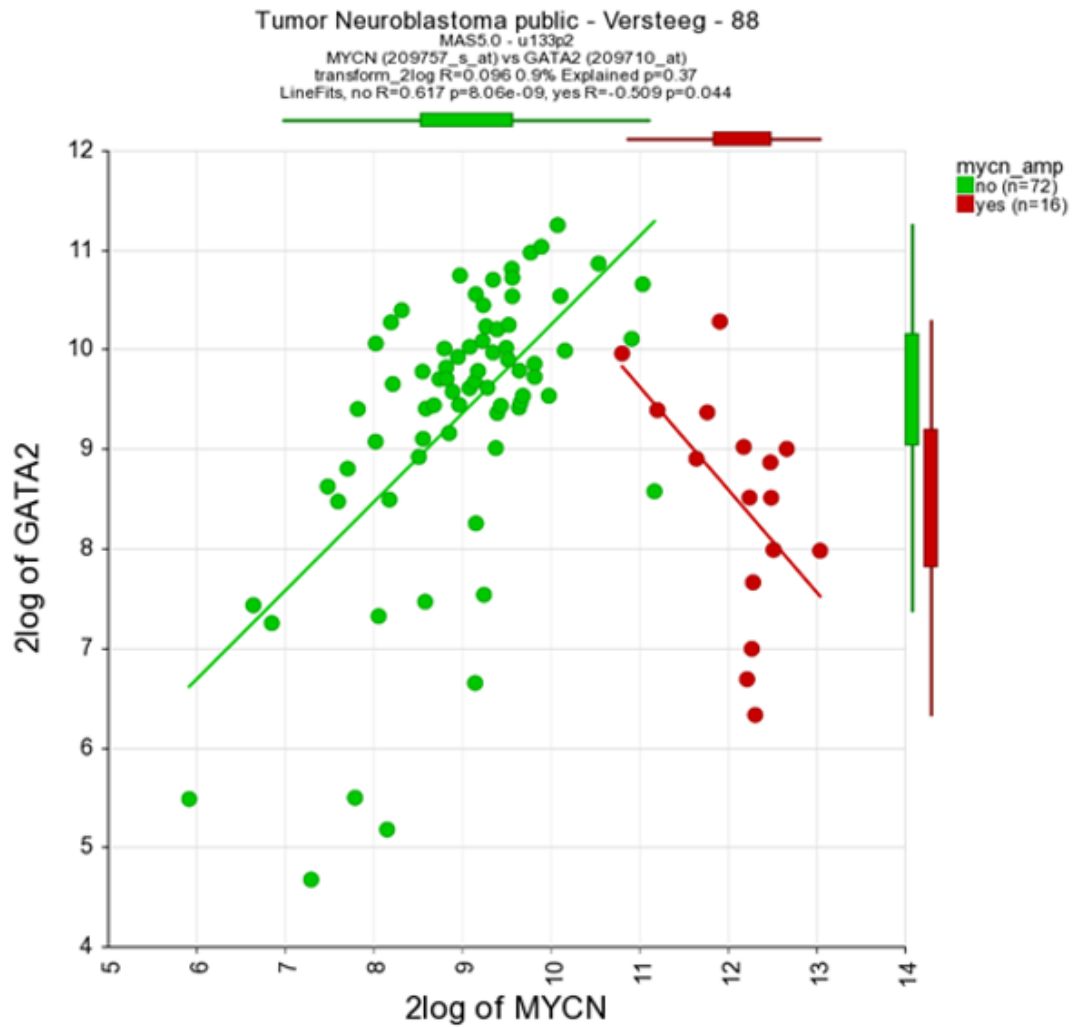


Figure 12: Linear fit per track

8. Through the menu on the right, several additional dataviews and analyses are available. Let us start with different overviews; R2 is able to produce heatmaps of this analysis. Return to the genelist view (Figure 2). Click on the 'Heatmap (zscore)' in the right menu. The gene names are on the y-axis, sample names on the x-axis.

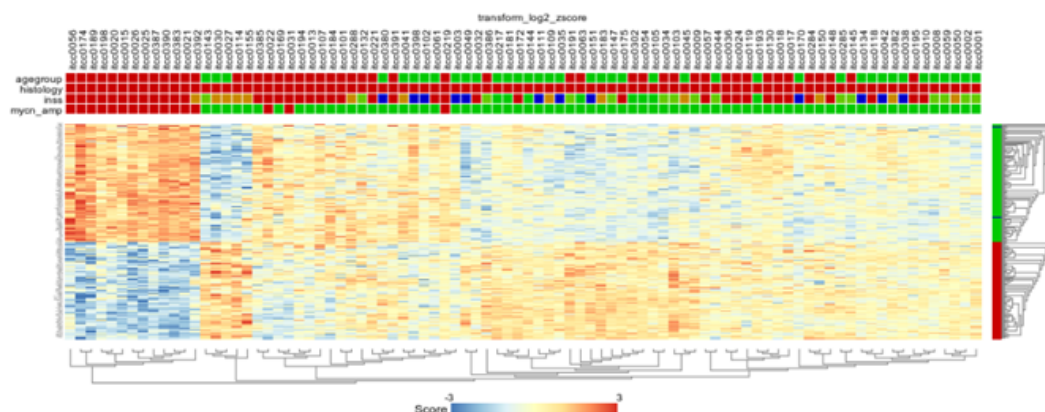


Figure 13 : Heatmap view of the expression of all genes correlating with the expression of MYCN in 88 Neuroblastoma samples.

7.5 Step 4: Relation with Chromosome position

Another possible view is the mapping of these genes on all chromosomes. Click on the ‘Chromosome Map’ menu item again in the menu on the right in Figure 2. In Figure 13, the overrepresentation of genes that correlate with MYCN expression with respect to all genes present on (an arm of) a chromosome is calculated. In Figure 14 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. 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.

chr	Whole chrom					p arm					q arm			
	Total Count	Reg Count	%	pval		Total Count	Reg Count	%	pval		Total Count	Reg Count	%	pval
chr01	2129	300	14.1%	4.6e-08	p	1157	210	18.2%	1.3e-17	q	972	90	9.3%	0.22
chr02	1409	164	11.6%	0.15	p	548	65	11.9%	0.29	q	861	99	11.5%	0.32
chr03	1168	102	8.7%	0.05	p	545	55	10.1%	0.78	q	623	47	7.5%	0.02
chr04	815	88	10.8%	0.76	p	243	24	9.9%	0.76	q	572	64	11.2%	0.57
chr05	942	97	10.3%	0.87	p	174	16	9.2%	0.58	q	768	81	10.5%	0.94
chr06	1153	108	9.4%	0.22	p	663	51	7.7%	0.02	q	490	57	11.6%	0.40
chr07	1056	134	12.7%	0.02	p	347	46	13.3%	0.09	q	709	88	12.4%	0.09
chr08	748	70	9.4%	0.32	p	284	23	8.1%	0.19	q	464	47	10.1%	0.81
chr09	852	60	7.0%	1.1e-03	p	249	19	7.6%	0.14	q	603	41	6.8%	3.3e-03
chr10	820	92	11.2%	0.48	p	192	26	13.5%	0.16	q	628	66	10.5%	0.97
chr11	1207	98	8.1%	7.8e-03	p	416	24	5.8%	1.8e-03	q	791	74	9.4%	0.31
chr12	1078	109	10.1%	0.70	p	296	30	10.1%	0.85	q	782	79	10.1%	0.74
chr13	398	45	11.3%	0.58	p		0	0%	-	q	398	45	11.3%	0.58
chr14	683	61	8.9%	0.19	p		0	0%	-	q	683	61	8.9%	0.19
chr15	658	81	12.3%	0.12	p		0	0%	-	q	658	81	12.3%	0.12
chr16	905	101	11.2%	0.49	p	507	60	11.8%	0.31	q	398	41	10.3%	0.92
chr17	1240	121	9.8%	0.42	p	355	41	11.5%	0.50	q	885	80	9.0%	0.17
chr18	314	25	8.0%	0.15	p	84	6	7.1%	0.32	q	230	19	8.3%	0.27
chr19	1392	159	11.4%	0.24	p	594	85	14.3%	2.2e-03	q	798	74	9.3%	0.27
chr20	624	57	9.1%	0.28	p	220	19	8.6%	0.38	q	404	38	9.4%	0.49
chr21	251	24	9.6%	0.64	p	5	1	20.0%	0.49	q	246	23	9.3%	0.57
chr22	548	36	6.6%	2.9e-03	p		0	0%	-	q	548	36	6.6%	2.9e-03
chrX	824	92	11.2%	0.51	p	332	44	13.3%	0.10	q	492	48	9.8%	0.61
chrY	86	5	5.8%	0.16	p	43	5	11.6%	0.80	q	43	0	0.0%	0.02
Total	21300	2229	10.5%	-	-	-	-	-	-	-	-	-	-	-

Figure 14: Statistics of overrepresentation of genes that have a correlation with MYCN on different chromosomes

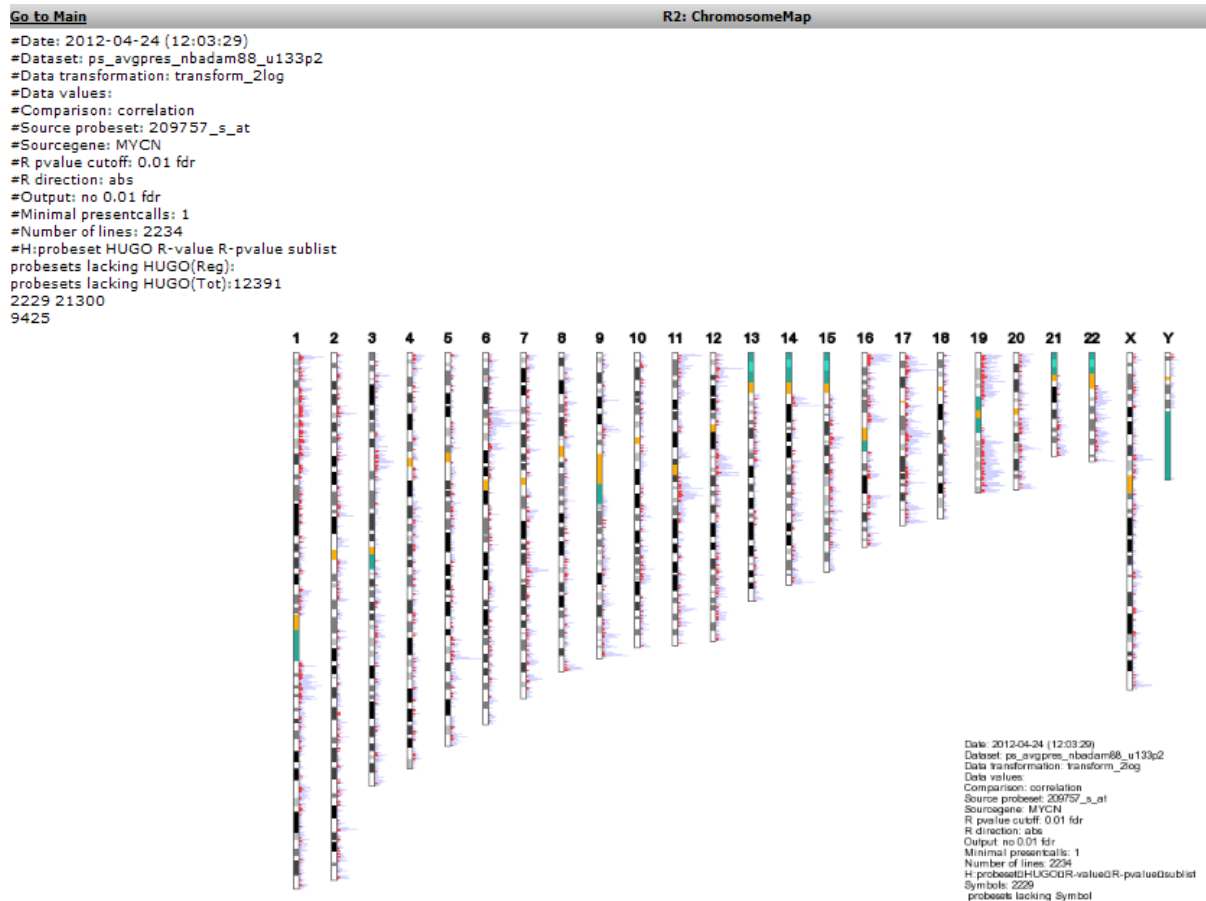


Figure 15: Mapping of the genes correlating with MYCN on allchromosomes

To further explore this set of genes return to the list: Figure 2.

7.6 Step 5: Establishing overrepresentation in other domains

1. Further overrepresentation analyses in other domains can give a first clue of the 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 15). It is possible to sort on p-value by clicking on the column header. Clicking on a pathway ID will open a new screen or tab with the heatmap of the selected pathway.



Figure 16: 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.

- One of the categories where genes of our current set are overrepresented is 'DNA-strand elongation'. All genes in this process have a consistent positive correlation (as can be seen by the green color). Let us take a look to see if we can corroborate this observation in another domain.
- The adjustable panel settings menu allows you to redo the gene-ontology analysis with the up or down regulated genes only.

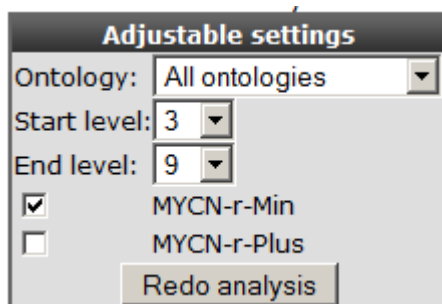


Figure 17: Re-do analysis with genes that are either positively or negatively correlated with MYCN.

7.7 Step 7: Gene list in pathway context

- Return to the gene list Figure 2 and click 'Geneset analysis'. Select the KEGG pathway in the gene set selection pull down.

- 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 in case the p-values are not sorted, to find the most significant ones: Figure 18.

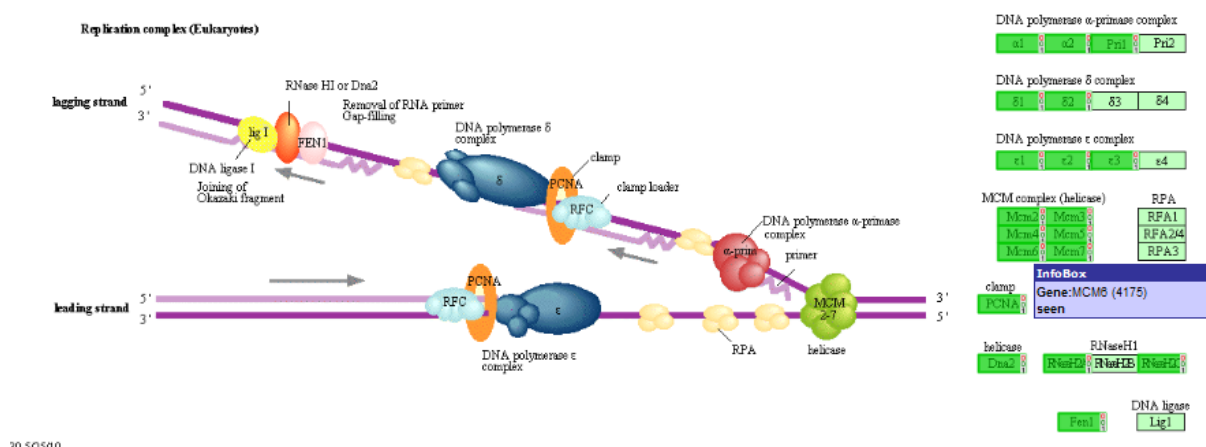
Go to Main **R2: Scavenger**

From your input (n=2234), 567 genes were also present in the current geneset collection (kegg).
 Within the dataset ps_avgpres_nbadam88_u133p2 (), 4493 genes were detected in the current geneset selection (hugoonce=yes and minimal present call=1)
 The table below lists genesets where the number of genes from your list are present more than expected (p<1 from 2X2 contingency table analysis with continuity correction)
 mycn-r-min, mycn-r-plus,

set	R#	#	p value	Genelist
over-representation [A] DNA replication	36	22	1.3e-17	. DNA2, FEN1, MCM2, MCM3, MCM4, MCM5, MCM6, MCM7, PCNA, POLA1, POLA2, POLD1, POLD2, POLE, POLE2, POLE3, PRIM1, RFC1, RFC3, RFC4, RNASEH2A, RNASEH2C
over-representation [A] Pyrimidine metabolism	93	32	4.5e-10	. CAD, CMPK1, CTPS2, DHODH, DTYMK, ENTPD4, NME1, NME4, PNP, POLA1, POLA2, POLD1, POLD2, POLE, POLE2, POLE3, POLR1D, POLR1E, POLR2E, POLR2H, POLR2I, POLR3D, POLR3GL, PRIM1, RRM1, RRM2, TK1, TYMS, UCK2, UMP5, UPP1, ZNRD1
over-representation [A] Spliceosome	113	34	3.4e-08	. AQR, DDX23, DDX5, EFTUD2, HNRNPM, ISY1, LSM7, NCBP2, PPIE, PQBP1, PRPF3, PRPF31, PRPF38A, PRPF8, RBM17, RBM3, SART1, SF3A2, SF3B3, SF3B4, SNRNP200, SNRNP40, SNRPA, SNRPB, SNRPD1, SNRPD2, SNRPE, SNRPF, SNW1, SR140, SYF2, U2AF1, U2AF2, XAB2
over-representation [A] Cell cycle	113	32	7.6e-07	. ANAPC1, ANAPC7, BUB1, BUB1B, CCNB1, CCNB2, CDC25C, CDKN2D, CHEK1, DBF4, E2F1, E2F3, HDAC1, MAD2L1, MCM2, MCM3, MCM4, MCM5, MCM6, MCM7, MYC, PCNA, PKMYT1, PLK1, PRKDC, SKP2, SMAD2, SMC1A, SMC3, TP53, YWHAG, YWHAH
over-representation [A] One carbon pool by folate	17	9	3.3e-06	. ATIC, DHFR, GART, MTHFD1, MTHFD1L, MTHFD2, MTR, SHMT2, TYMS
over-representation [A] Base excision repair	32	13	6.2e-06	. APEX1, FEN1, LIG3, NEIL3, NTHL1, PCNA, POLD1, POLD2, POLE, POLE2, POLE3, UNG, XRCC1
over-representation [A] Purine metabolism	143	34	7.6e-05	. AK5, APRT, ATIC, ENTPD4, GART, GMP5, IMPDH2, NME1, NME4, NUDT5, NUDT9, PAICS, PDE7B, PFAS, PKM2, PNP, POLA1, POLA2, POLD1, POLD2, POLE, POLE2, POLE3, POLR1D, POLR1E, POLR2E, POLR2H, POLR2I, POLR3D, POLR3GL, PRIM1, RRM1, RRM2, ZNRD1
over-representation [A] Non-homologous end-joining	12	6	5.2e-04	. FEN1, LIG4, MRE11A, NHEJ1, PRKDC, XRCC6
over-representation [A] Homologous recombination	28	10	6.6e-04	. BLM, BRCA2, MRE11A, POLD1, POLD2, RAD51, RAD54L, TOP3B, XRCC2, XRCC3
under-representation [A] Neuroactive ligand-receptor interaction	198	9	7.0e-04	. CHRN1, F2R, GABBR1, GABRA5, GABRB3, GABRG, HRH3, NR3C1, PTGER3
over-representation [A] Chronic myeloid leukemia	71	18	2.1e-03	. CBLB, CHUK, CRK, CTBP2, E2F1, E2F3, HDAC1, KRAS, MAP2K1, MAP2K2, MAPK1, MYC, NRAS, PIK3CD, PIK3R5, RAF1, RELA, TP53
over-representation [A] Mismatch repair	22	8	2.4e-03	. EXO1, MSH6, PCNA, POLD1, POLD2, RFC1, RFC3, RFC4
over-representation [A] Oocyte meiosis	106	24	2.7e-03	. ANAPC1, ANAPC7, AURKA, BUB1, CCNB1, CCNB2, CDC25C, CHP, MAD2L1, MAP2K1, MAPK1, PKMYT1, PLK1, PPP2CA, PPP2R1B, PPP2R5B, PRKACB, PRKX, RPS6KA1, SMC1A, SMC3, YWHAG, YWHAH
over-representation [A] Bladder cancer	42	12	3.8e-03	. E2F1, E2F3, FIGF, KRAS, MAP2K1, MAP2K2, MAPK1, MMP2, MYC, NRAS, RAF1, TP53
over-representation [A] Pancreatic cancer	70	17	5.4e-03	. BRCA2, CDC42, CHUK, E2F1, E2F3, FIGF, KRAS, MAP2K1, MAPK1, PIK3CD, PIK3R5, RAD51, RAF1, RALA, RELA, SMAD2, TP53

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

- The DNA-replication pathway pops up as most significant. Note that most genes are similar to those found by the GO process in the former analysis. The pathway will be shown when the blue “A” in front of the pathway name is clicked.
- A hyperlinked KEGG pathway appears: Figure 19.



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

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 2) to show their individual relation with MYCN. However,

it is important to realize that the KEGG pathway interactions are build on Protein-Protein interactions, Enzyme-Substrate Interactions, Protein-DNA Interactions, Protein-Compound Interactions etc etc and the genelist from this module is baed on RNA expression levels only/

7.8 Step 8: Further pathways analysis

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

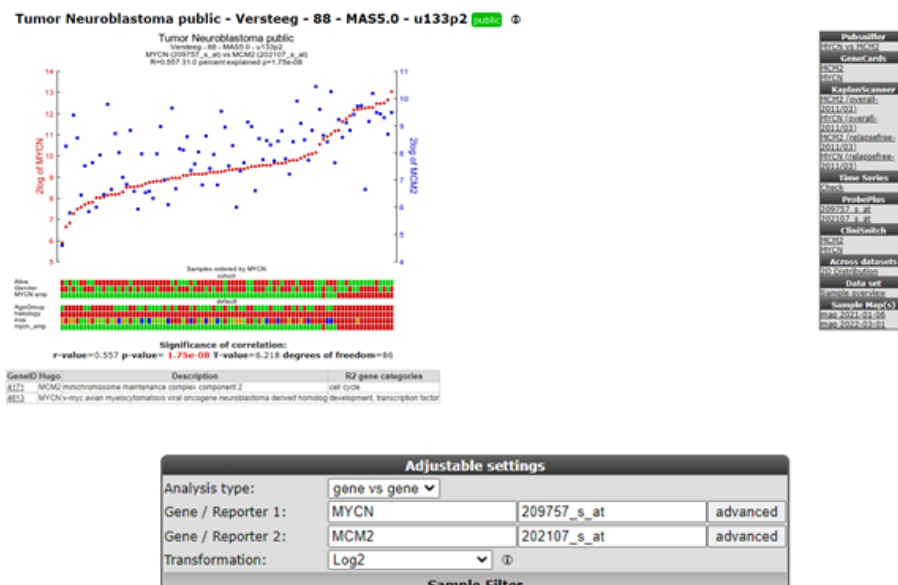


Figure 20: MCM2 expression correlates with MYCN expression.

2. The correlation is significant. In the left upper table there is a link to the Pubsntiffer 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 21.

Checking PubMed for MYCN, MCM2

Query	# Articles	Outlink
'MYCN AND MCM2'	6	Pubreminer
'MYCN'	10637	Pubreminer
'MCM2'	787	Pubreminer
'MYCN AND MCM2 AND cancer'	4	Pubreminer
'MYCN AND MCM2 AND development'	1	Pubreminer
'MYCN AND MCM2 AND neuroblastom*'	1	Pubreminer
'MCM2 AND cancer'	340	Pubreminer
'MCM2 AND development'	104	Pubreminer
'MCM2 AND neuroblastom*'	3	Pubreminer

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: Pubsntiffer 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:

Figure 22: The PubReMiner tool web interface

- 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 “Go to Pubmed with query” button to find the article.
- This article is actually published work by our group where the relation between the MCM genes and MYCN was proven experimentally.

Figure 23: The correlation between MCM genes and MYCN was proven experimentally in this article.

7.9 Step 9: Gene set analysis

- The genelist produced in the beginning of this tutorial (Figure 2) 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.
- The menu to the right gives several possibilities (Figure 2). Some of these have been explored already; we will shortly touch 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 another data analysis tool such as [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”.

7.10 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.

Working with Kaplan Meier

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 the option *Kaplan Meier* 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. For this example, make sure that *Data type* is set to ‘Expression data (H. sapiens)[n=..]’ and that *Datat set* is set to “Tumor Neuroblastoma public “Versteeg “88”. From the dropdown of the setting *Separate by* choose the option ‘a categorical track’. Click *Next*.

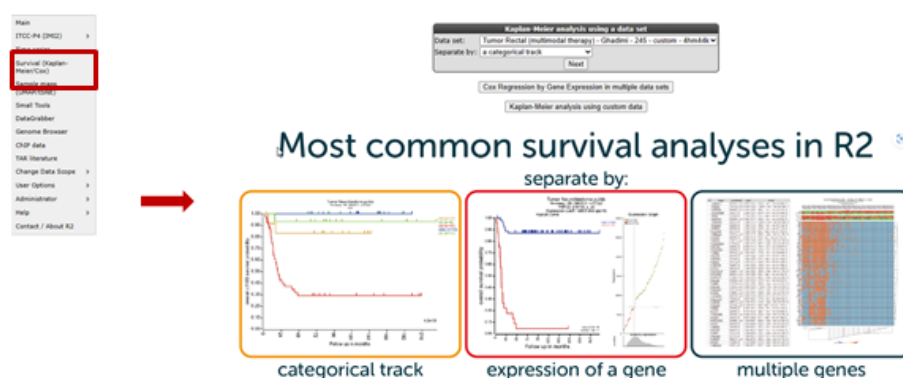


Figure 1: Select a Kaplan Meier option.

- In the adjustable settings menu choose for *Type of survival* the value “overall-c1103”, select “track” at *Separate by* and select “inss (5 cat)” at the *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.

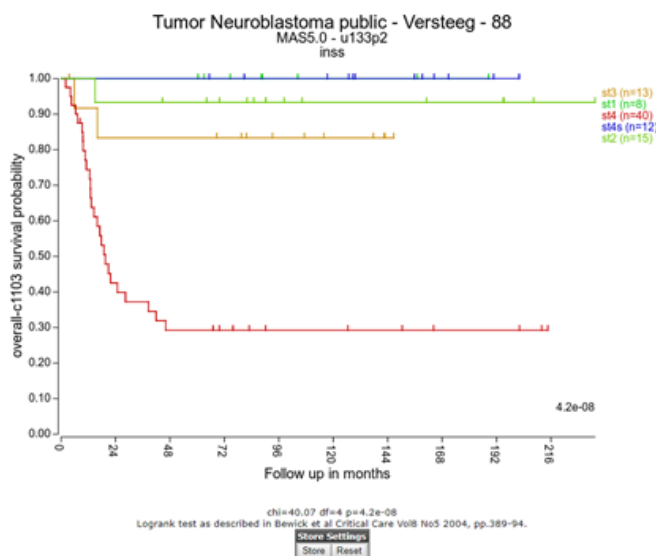


Figure 2: Kaplan Meier by a categorical track

- A handy feature of the R2 Kaplan module is the option to combine two tracks to generate subgroups for the Kaplan Meier analysis. Use the back-button from the browser and select at *Separate by* the option ‘combination of two tracks’. For example, choose for the first track ‘agegroup (2 cat)’ and for the second track ‘mycn_amp (2 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.



For the use of the grid, use the clog icon in the 'Advanced settings box' and hover your mouse over the side of the column of choice to show the filtering options for that track. You can use the filtering options of multiple tracks to make a specific subset. Here we used gender to only select male samples (i.e. deselect female) and death_cause to select tumor and nd samples. A successful subset selection will be shown as a small message indicating the number of samples in the subset, the used track name and the number of groups/categories between brackets

If you instead want to make the subset selection with the pull-down of **Subset track**, click on the track by which you want to define the subset (in the example: gender). In the popup you need to check the box of your preferred subset(s) (here: male). If you want to further narrow down your selection with a different track, click on the same pull-down menu. Select the next track (here: death_cause) that you are interested in and in the popdown, check the preferred subset(s) from that track (here: tumor and nd).

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 the *Kaplan Meier* option or in field 3 of the main menu, and choose *Separate by* 'a single gene'. Make sure that "Tumor Neuroblastoma public Versteeg 88" is selected with *Data type* 'Expression data (H. sapiens)', and click *Next*.
2. In the next screen fill in 'mycn' in the Search by Gene field and use the first reporter in the list of the dropdown that popup. Leave the rest as is 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

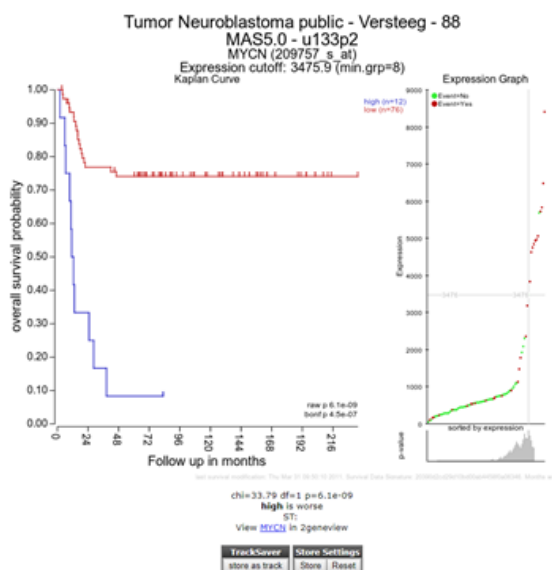
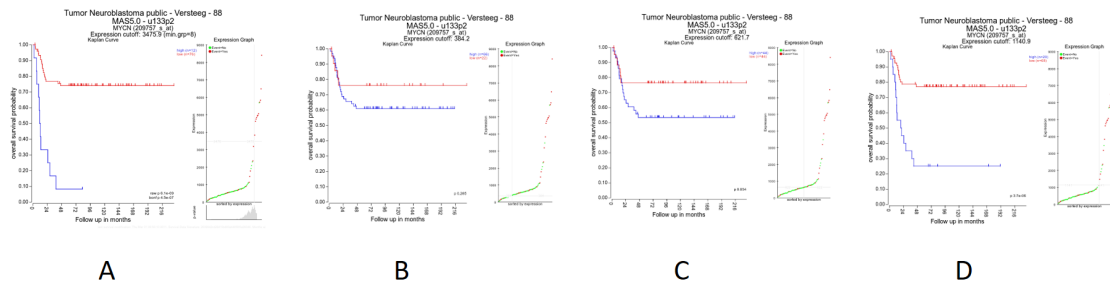


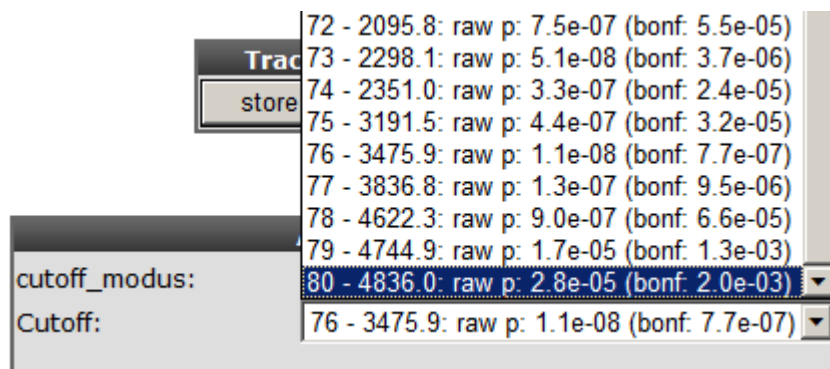
Figure 5: Kaplan Scan for a single gene

- 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 modus. Try to find out whether this is also the case with other cut-off modi.

- 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

- 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”, upperleft corner), choose *Kaplan-Meier* from the left menu and choose *Select by* ‘multiple genes’ and click “Next”.
- In this example search for a *Gene set* with the **Search GS** button. In the popup grid, type ‘apoptosis’ in the text field on top and click on the search button. In the adapted grid, open the KEGG pathways by a click on the arrow scroll and go on till you can check the KEGG ‘Apoptosis’ pathway. Click *Use selected* to go back to the menu, where the ‘GS: Apoptosis (85)’ geneset is now shown to be selected.

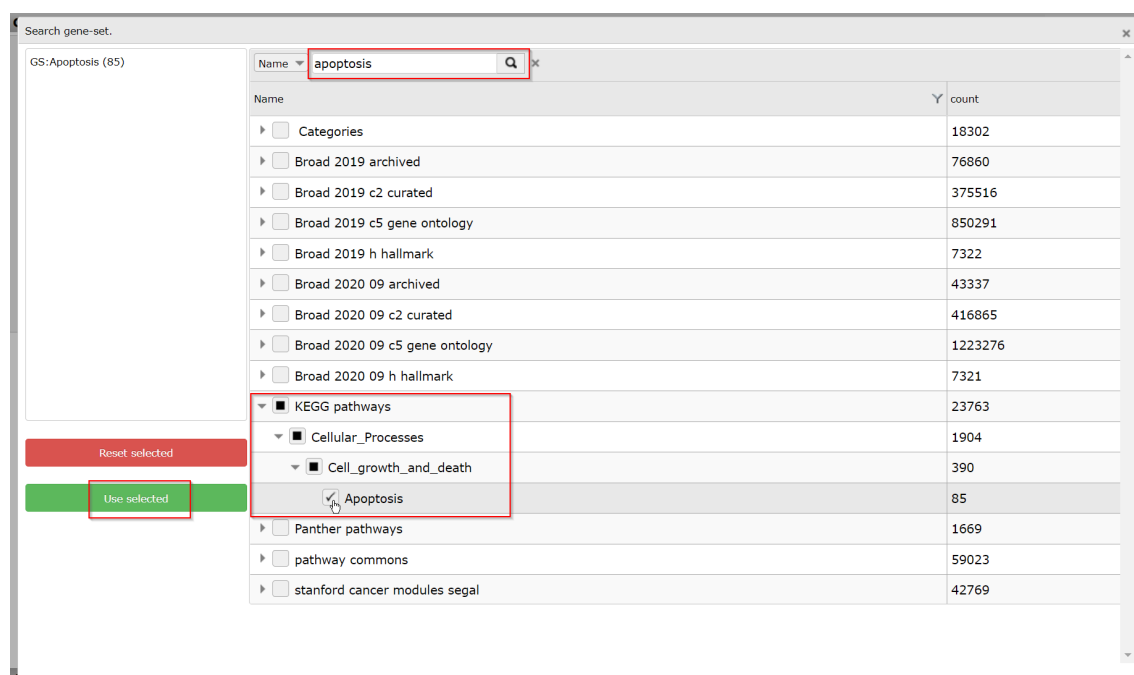


Figure 8A: Select a KEGG gene set

3. Leave the *Type of survival* at 'overall-c1103'. In the statistics panel there are several filtering options possible, leave these options unchanged.
4. In the graphics section select "yes" at *Draw heatmap* and click *Next*.
5. 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.

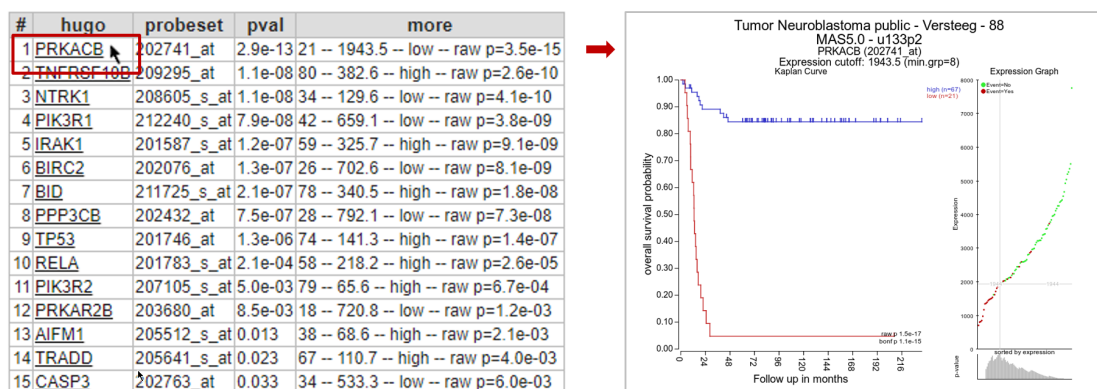


Figure 8B: A list of Kaplan Meier for a group of genes

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

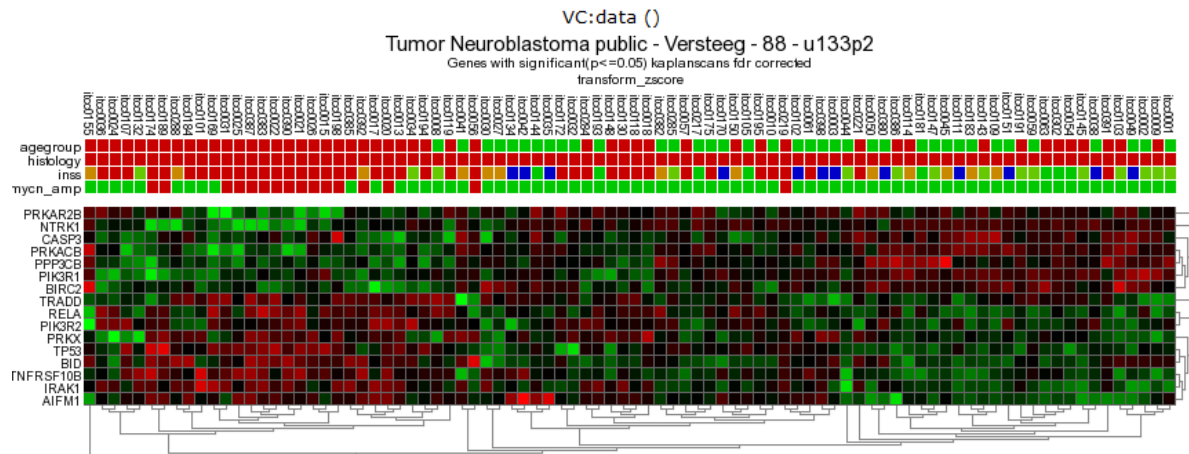


Figure 9: Heatmap of the significant prognostic list of genes.

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, again go to the main page, click on *Kaplan-Meier* from the left menu, select 'by multiple genes' in the *Separate by* option, click next. This time don't select any gene set, but adapt in the Graphic Settings *Draw a heatmap* again to 'yes', choose in *Heatmap data* dropdown 'good_bad (binary)' and click *Next*.

Graphics Settings	
Draw heatmap:	<input type="button" value="no"/>
Heatmap data:	<input type="button" value="good_bad (binary)"/>
Heatmap gene weight:	<input type="button" value="no"/>
<input type="button" value="Next"/> <input type="button" value="Reset"/>	

Figure 10: Select binary heatmap.

Now the heatmap shows a clustering based on the GOOD vs BAD prognoses.

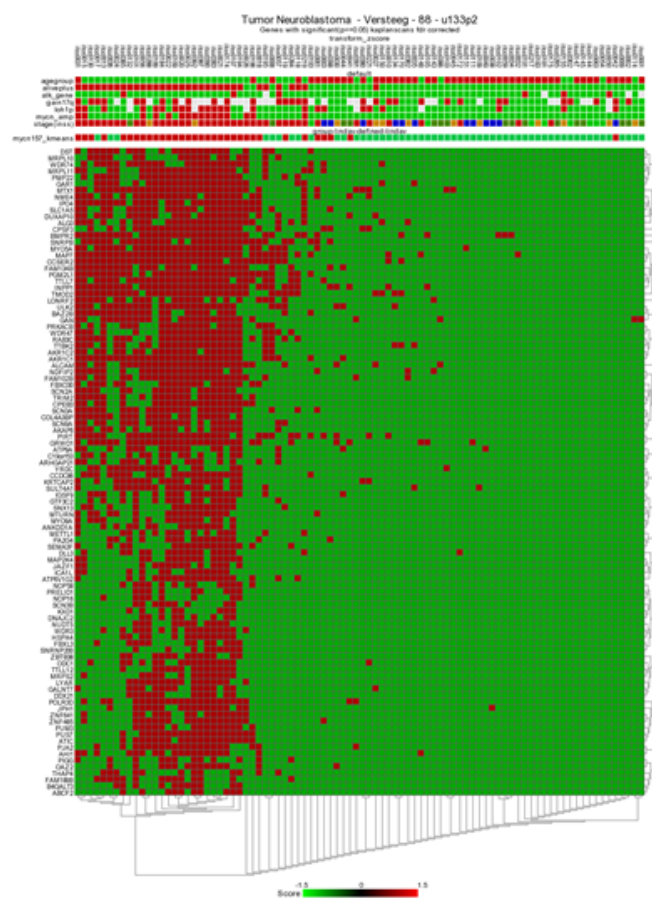


Figure 11 : Binary heatmap.

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* option again from the main page left menu, and click on the *Kaplan-Meier analysis using custom data* button underneath the panel.

R2: Kaplan-Meier

Go to: Main **R2: KaplanScan on UserDefined Data**

- Main
- Time series
- Kaplan-Meier**
- Sample maps
- Small Tools
- DataGrabber
- Genome Browser
- ChIP data
- TAR literature
- Change Data Scope ▶
- User Options ▶
- Help ▶
- Contact / About R2

KaplanScan:

Paste Survival information `samplename; survtime(days); event(1/0/YES/NO); expression_value (tab or ; separated)`

#Copy your data below the #H: line.

#H:samplename;survival_time;event;Gene_x

itcc0001	7876	no	8.505811554
itcc0002	4914	no	9.556122818
itcc0003	7457	no	10.90989308
itcc0008	6904	no	8.819221104
itcc0009	7146	no	8.019590728
itcc0010	7168	no	9.520029304
itcc0013	488	yes	9.280074769
itcc0015	1277	yes	11.64001893
itcc0017	576	yes	9.492253789
itcc0018	609	yes	10.53585856
itcc0020	340	yes	11.90568785
itcc0021	296	yes	13.03792439
itcc0022	394	yes	11.19905882
itcc0024	1408	yes	9.430870664
itcc0025	196	yes	12.2798717
itcc0026	2916	no	12.47555612
itcc0027	108	no	7.785288981
itcc0030	4013	no	5.911691582
itcc0031	866	yes	12.2605447
itcc0032	6872	no	9.636262089
itcc0034	5458	no	8.846744023
itcc0035	6013	no	8.211401637
itcc0036	165	yes	9.141340821
itcc0038	5759	no	10.07293648
itcc0041	5118	no	8.191306019
itcc0042	4693	no	9.173677136
itcc0044	4152	no	8.814742667
itcc0045	4091	no	8.311975314

cutoff_modus: scan Next Reset

Figure 12: Kaplanscan with user defined data

2. 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
3. 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.
4. 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.

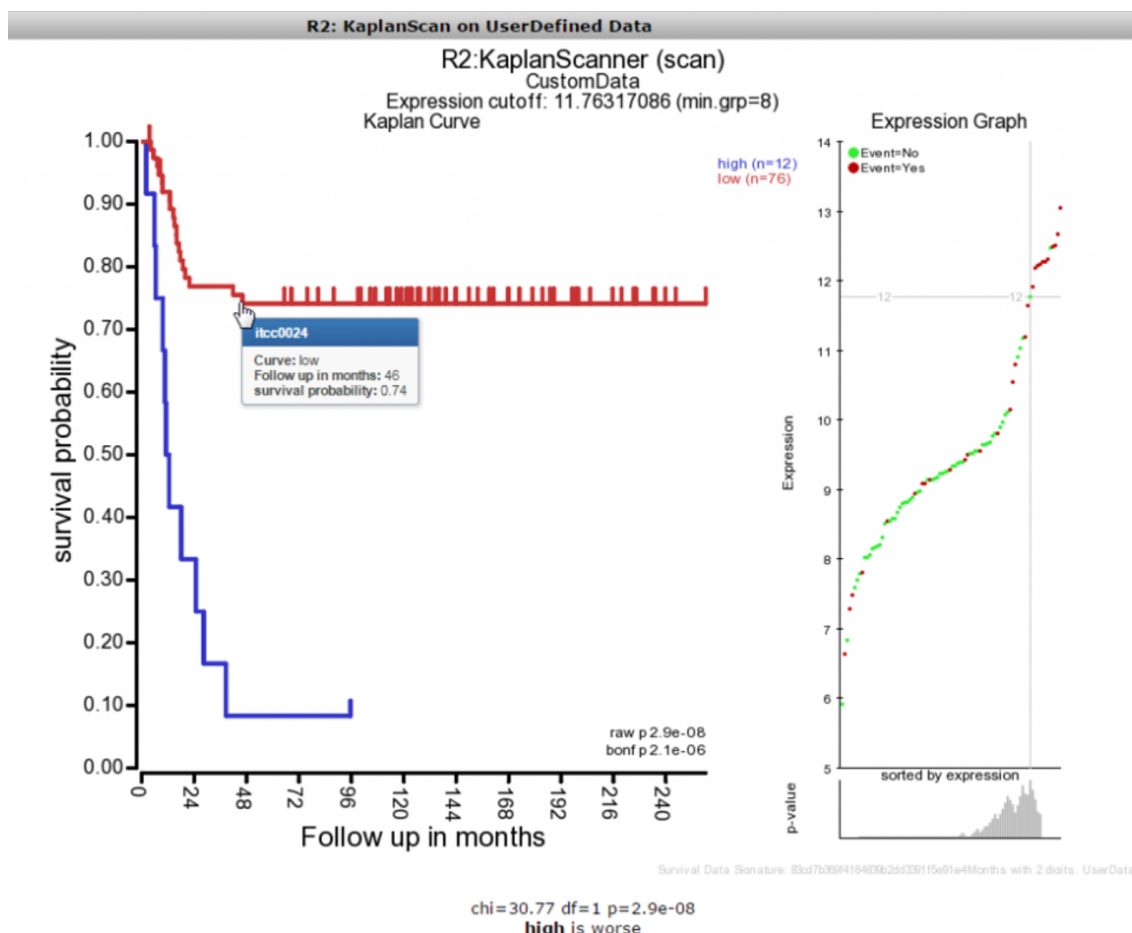


Figure 13: Kaplanscan with user defined data result

- 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 Step 4: Cox Regression analysis and hazard ratio

The Cox regression analysis is a statistical method commonly used in biomedical research to analyze the relationship between gene expression and survival outcomes. It is particularly useful for studying the impact of gene expression levels on patient survival times. In this analysis, gene expression data is combined with survival data to assess whether specific genes or gene signatures are associated with increased or decreased survival rates. In r2 you can identify genes or multiple genes (genesets which may act as potential prognostic markers). In general Cox regression analysis in gene expression provides valuable insights into the molecular mechanisms underlying survival outcomes in various diseases, including cancer.

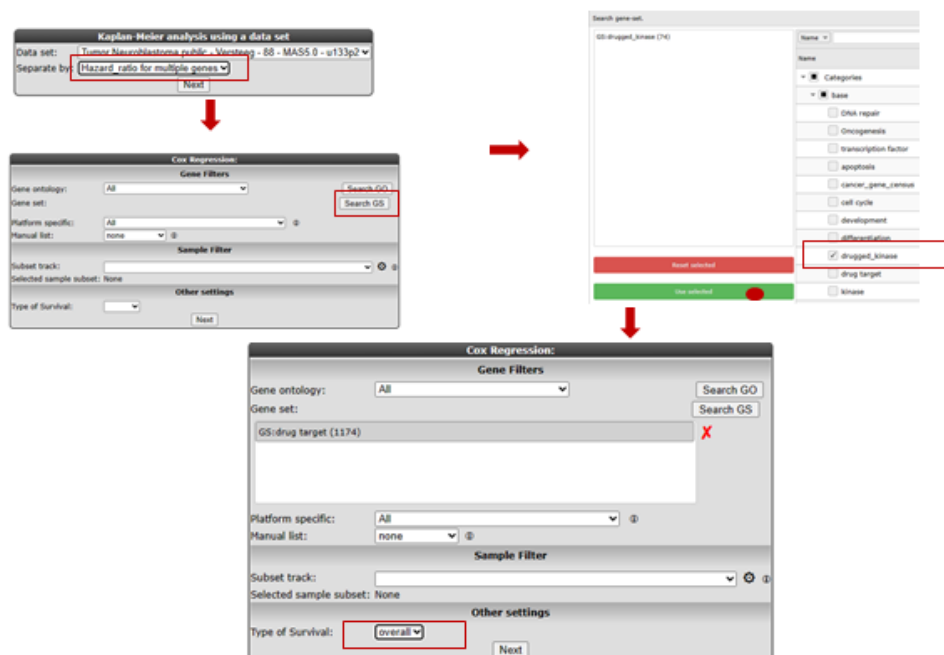
separate by:

categorical track

expression of a gene

multiple genes

The hazard ratio (HR) is a fundamental concept in Cox regression analysis. It represents the relative risk or likelihood of an event occurring in one group compared to another. The hazard ratio quantifies the effect of gene expression levels on the hazard or risk of a specific outcome, typically survival.



In R2 you can scan for significant hazard ratios for a single gene or multiple genes for a single dataset, follow the steps as indicated in Figure 15 and select the drug target geneset and click next. A Hazard ratio quantifies the effect of gene expression levels on the hazard of risk for a typical outcome. When the Hazard ratio is greater than 1, higher expression levels suggest a poor prognosis indicated in blue. Having the opposite, a hazard ratio less than 1 indicates a decreased risk or hazard, suggesting a better prognosis indicated with red. A hazard ratio of 1 implies that there is no difference in risk between the groups being compared.

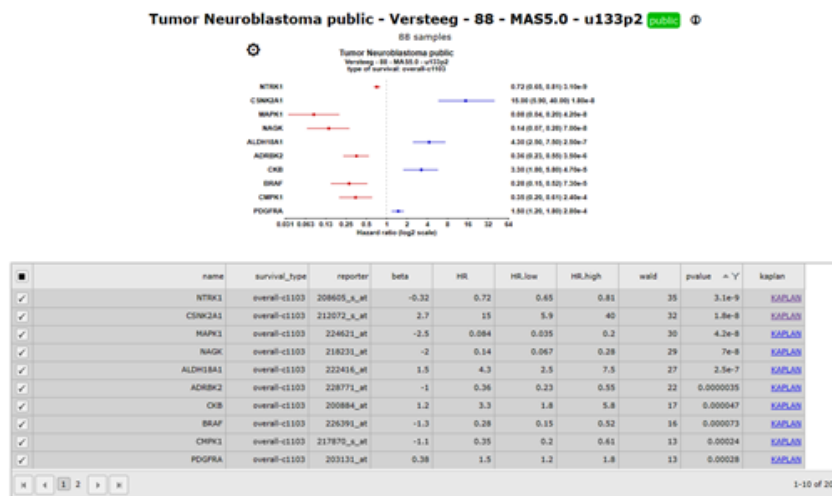


Figure 15: Cox regression and hazard ratios list

The hazard ratio outcome can easily be inspected by clicking on the **KAPLAN** link in the table. As clearly depicted in figure 16 the CSKN2A1 gene with a hazard ratio > 1 in blue show a poor prognosis for a high expression level cut-off while the NTKR1 gene shows a poor prognosis for a low expression level.

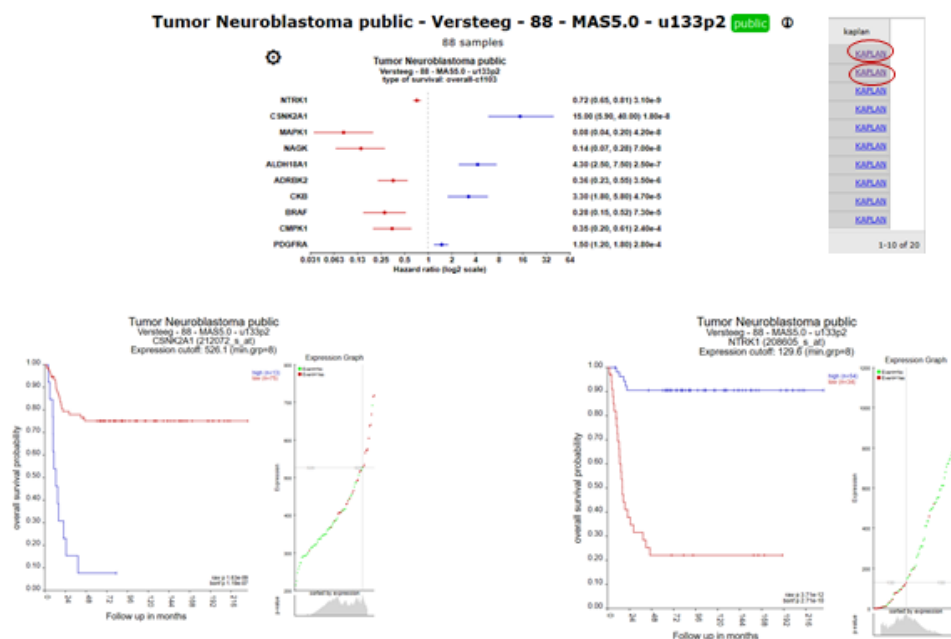


Figure 16: Cox regression and hazard ratios list

After checking the hazard ratio for multiple genes in one dataset you can also check for hazard ratios across multiple datasets. Select via the main menu survival > cox regression in multiple datasets. (Figure 17). Clicking “select datasets” will open the grid where you select the datasets of interest, keep in mind that these dataset are already preselected for those containing survival data.

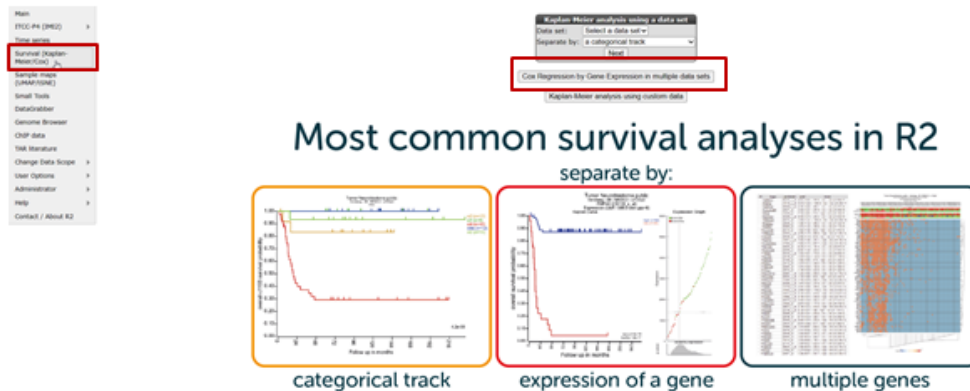


Figure 17: Cox regression for multiple datasets

In Figure 18, the selected datasets show a significant low hazard ratios with a poor survival for the NOTCH2 gene in the low gene expressed group in contrast to the hazard ratios where the group with the high MYCN expression shows a poor prognosis. (Figure 19.)

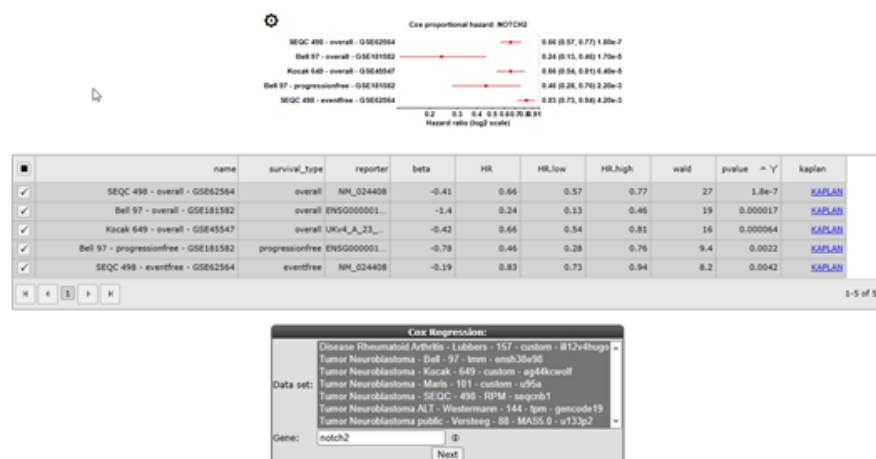


Figure 18: Hazard ratios for the NOTCH2 gene

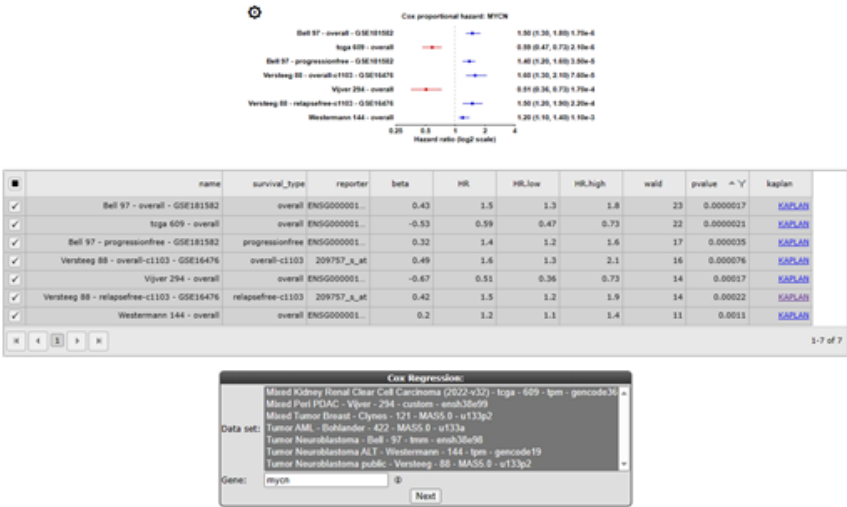


Figure 19: Hazard ratios for the MYCN gene

8.7 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:
 - KEGG Pathway database (<http://www.genome.jp/kegg/pathway.html>)
 - 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 CTNNB1 (catenin) . R2 can find the gene symbol by alternative name also, we'll try 'catenin'; Figure 1.
5. Click 'Submit'

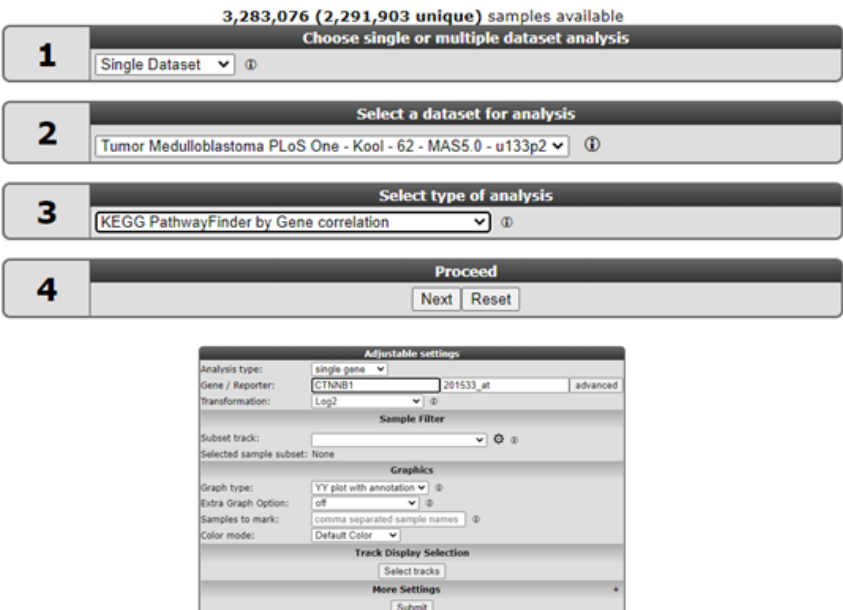


Figure1: Selecting KEGG pathwayfinder by gene correlation forcatenin

9.3 Step 2: 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.

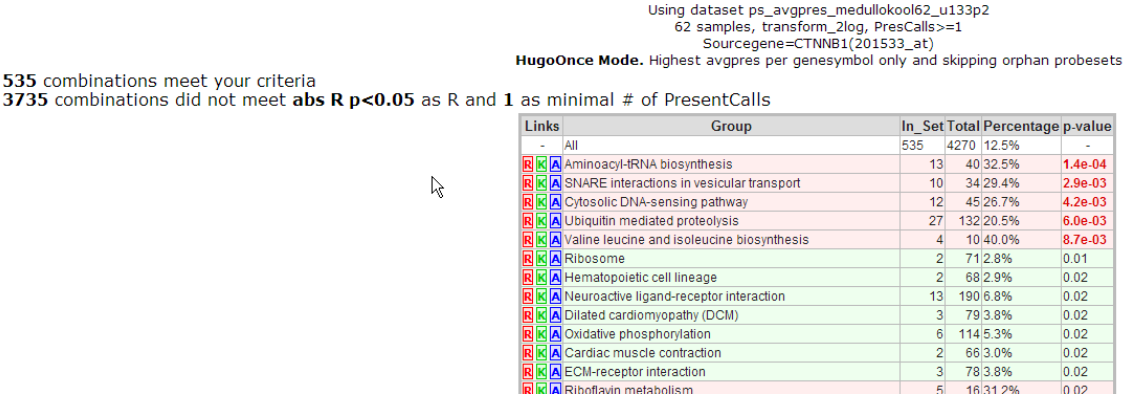


Figure 2: KEGG pathways that have an overrepresentation of genes that correlate with CTNNB1 in this dataset

- 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.

[Go to Main](#)

R2: KEGG Pathway Viewer

MAP: -hsa04130-

Your supplied dataset 1336047169-pathfinder.txt:

#Date: 2012-05-03 (14:12:45)

#Dataset: ps_avgpres_medullokol62_u133p2

#Correlationtype: transform_2log

#Source probeset: 201533_at

#Sourcegene: CTNNB1

#R cutoff: 0.5 abs

#Minimal presentcalls: 1

#Number of lines: 544

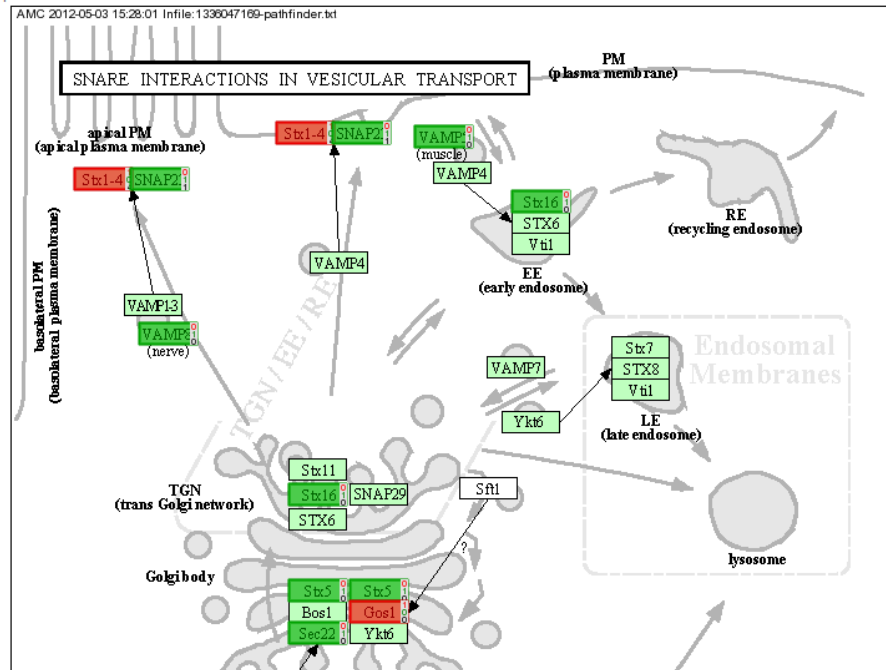


Figure 3: The SNARE pathway; darker green and red are genes correlating with CTNNB1

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.

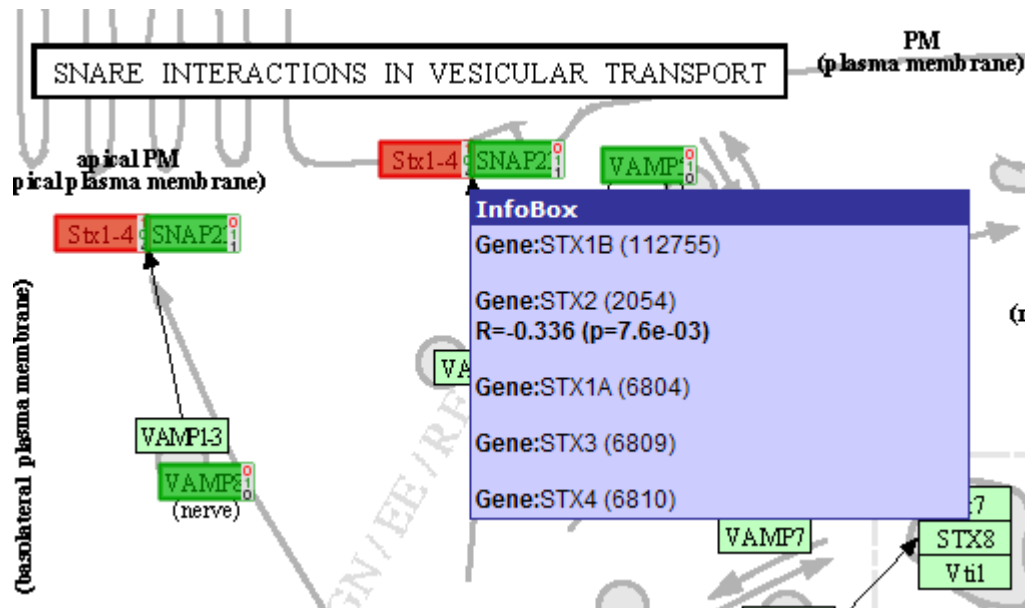


Figure 4: : Hovering over the Stx1-4 box shows that this actually represents 5 genes; only one of them is correlating with CTNBB1.

9.4 Step 3: Finding pathways relevant to subgroups

1. In field 3 on the R2 start page select 'KEGG PathwayFinder by Groups';
2. 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; beat mutation. Select the track, beat_mutation (Figure 5).

3,283,076 (2,291,903 unique) samples available

Choose single or multiple dataset analysis

1 Single Dataset

2 Select a dataset for analysis

Tumor Medulloblastoma PLoS One - Kool - 62 - MASS.0 - u133p2

3 Select type of analysis

KEGG PathwayFinder by Groups

4 Proceed

Next Reset

Tumor Medulloblastoma PLoS One - Kool - 62 - MASS.0 - u133p2 public

Adjustable settings

Track: beat_mutation (2 cat)

Transformation: Log2

Min. # Present calls: 1

min test p-value: 0.0001

Representation: over

Sample filter

Subset track:

Selected sample subset: None

Submit Reset

Figure 5: Selecting the beat mutation track.

1. Click "Submit"

9.5 Step 4: 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.

R2: PathwayFinder

Using dataset ps_avgpres_medullokool62_u133p2
62 samples, transform_2log, PresCalls>=1
Sourcegene=NG_bcat mutation(NG_bcat mutation)

HugoOnce Mode. Highest avgpres per genesymbol only and skipping orphan probesets
607 combinations meet your criteria
3754 combinations did not meet T-test $p < 0.0001$ as R and 1 as minimal # of PresentCalls

Links	Group	In_Set	Total	Percentage	p-value
- All		607	4361	13.9%	-
Wnt signaling pathway		40	136	29.4%	1.8e-07
Protein export		10	23	43.5%	4.2e-05
N-Glycan biosynthesis		14	46	30.4%	1.2e-03
Vibrio cholerae infection		15	51	29.4%	1.4e-03
Basal cell carcinoma		14	47	29.8%	1.7e-03
Systemic lupus erythematosus		3	87	3.4%	4.8e-03
Gap junction		20	82	24.4%	6.2e-03
Melanogenesis		21	88	23.9%	7.0e-03
Oxidative phosphorylation		6	110	5.5%	0.01
Adherens junction		17	70	24.3%	0.01
Ribosome		4	86	4.7%	0.01
- Autoimmune thyroid disease		0	37	0.0%	0.01
Chondroitin sulfate biosynthesis		7	22	31.8%	0.02

Figure 6: 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.

9.6 Step 5: 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;

Tumor Medulloblastoma PLoS One - Kool - 62 - MASS.0 - u133p2

62 samples, transform_log2, present>=1 gene set: Wnt_signaling_pathway
track_bcat_mutation
40 combinations meet your criteria
86 combinations did not meet $p\text{-value} < 0.0001$

View	Gene	P	Group	Presence
	WSP1	5.85e-28	yes > no	25/62
	LEP1	2.6e-34	yes > no	60/62
	DNK2	2.17e-27	yes > no	18/62
	PLCB1	7.49e-23	yes > no	61/62
	NKD1	4.48e-21	yes > no	15/62
	FZD10	4.07e-19	yes > no	28/62
	FZD6	6.75e-15	yes > no	62/62
	NKD2	5.8e-12	yes > no	35/62
	BARH8	7e-12	yes > no	55/62
	TCF7L1	1.78e-11	yes > no	38/62
	PRKACB	3.11e-10	no > yes	62/62
	WNT11	5.69e-10	yes > no	6/62
	FZD2	5.95e-10	yes > no	54/62
	AXIN2	6.52e-10	yes > no	62/62
	NKX	7.72e-10	no > yes	62/62
	FZD7	2.49e-9	yes > no	62/62
	HAPK8	4.28e-9	no > yes	62/62
	LRP5	2.54e-9	yes > no	28/62
	PRKCB	2.74e-9	no > yes	49/62
	WNT3	3.09e-9	no > yes	58/62

Gene set analysis

Map on pathway image

Known interactions analysis

Gene Ontology Analysis

Enricher

Data folder

Relate results

Chromosome Map

Heatmap (Z-score)

K-Means

GeneSetMatrix

Save current selection as TXT file

Save selection as TXT file (no header)

Reference for current selection

Store result as custom gene set

Differential expression

Group	Count
yes > no	25
no > yes	14

Mini ontology analysis

Category	Count	Total %	prob
All	40	100	1.00
Cell cycle	4	10	0.17
Apoptosis	1	3	0.04
Cell death	5	12	0.28
Development	18	44	0.74
Metabolism	18	44	0.74
Drug target	1	3	0.07
Immune	1	3	0.07
Metastasis	11	27	0.68
Signal transduction	32	79	0.98
Transcription factor	4	10	0.25

Figure 7: 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.

9.7 Step 6: 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.

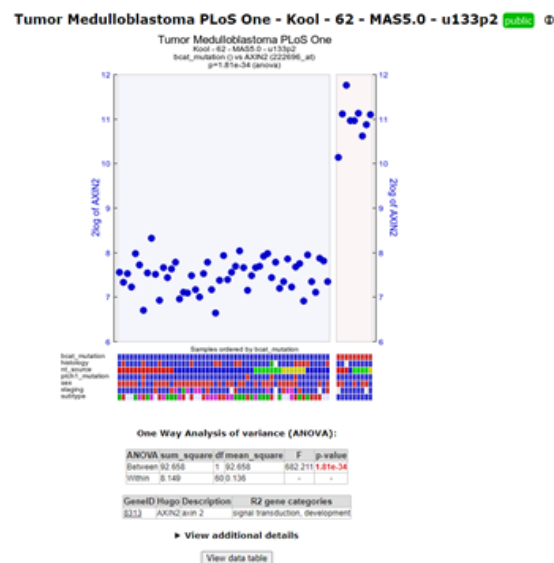


Figure 8:AXIN2 expression correlates with Beta Catenin mutations

9.8 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”.

1,842,385 (1,680,854 unique) samples available

Choose single or multiple dataset analysis

1

Across Datasets

2

Select an analysis

Select an analysis: MegaSampler (View a gene in more than 1 dataset)

3

Proceed

Next Reset

Figure 1: Using across datasets

- 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

Type of data: u133p2, MAS5.0

Use Preset: u133a, MAS5.0
u133b, MAS5.0
u133p2, MAS5.0
u133p2, RMAcr
u95a, MAS5.0

Megasampler only allows you to query multiple datasets if they are of the same chiptype 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”.

R2: MegaSampler

Step1: Select the chip_type and appropriate normalization scheme

MegaSampler

Type of data: **hs, u133p2, MAS5.0**

Use Preset: **NO**

Global Level

XPOsampler

Group Level

- student: practical_markers (u133p2)
- student: Praktikum001 (u133p2)
- r2: Boxdotplot_Big (u133p2)
- r2: final order MYCN 20 tumors (u133p2)
- r2: MYCN_signature (u133p2)
- r2: MYCN_signature_old (u133p2)
- r2: PI3K_signature (u133p2)
- r2: PI3K_signature_old (u133p2)
- student: practical_markers (u133p2)
- student: Praktikum001 (u133p2)

Figure 2: Select a preset

- 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”.

Adjustable settings

Transformation: **Log2**

Graph type: **Boxdot plot (bands)**

Max Y-value: **3**

Adapt: **default**

Vector (SVG) output: **false**

Merge by track groups: **false**

Samples to mark: **comma separated sample names**

Dot size: **2**

Mark method: **dot**

Gene: mycn

- ☒ MYCN (209757_s_at) Avg=92.5
- ☐ MYCN (242026_at) Avg=20.1
- ☐ MYCN (234376_at) Avg=8.0
- ☐ MYCN (209756_s_at) Avg=6.7
- ☐ MYCN (211377_s_at) Avg=3.5

Dataset	Count	Marked
T Breast 351 (EXPO)	1	NO
T Cervix 36 (EXPO)	2	NO
T Colon 315 (EXPO)	3	NO
T Endometrium 209 (EXPO)	4	NO
T Kidney 261 (EXPO)	5	NO
T Lung 121 (EXPO)	6	NO
T Other 296 (EXPO)	7	NO
T Ovary 256 (EXPO)	8	NO
T Prostate 72 (EXPO)	9	NO
T Colon 39 (EXPO)	10	NO
T Colon 38 (EXPO)	11	NO
T Thyroid 34 (EXPO)	12	NO

Paste additional probesets into the box (1 probeset per line)

Submit **Reset**

Figure 3: Megasampler adjustment selection

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. and click “next”.

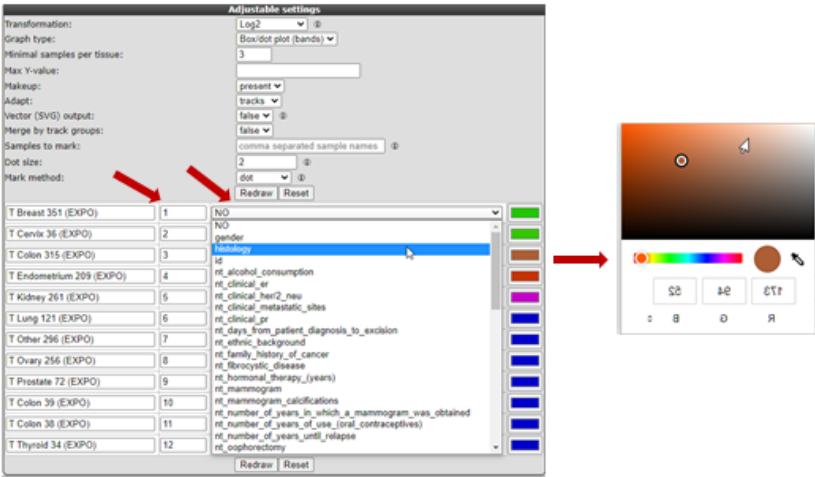


Figure 4: Adjusting the megasampler graph.

- 2. 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.

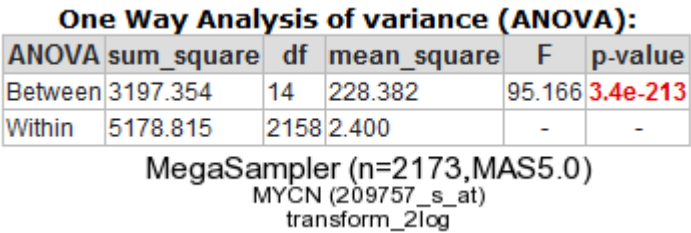


Figure 5: Anova test for the selected datasets.

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.

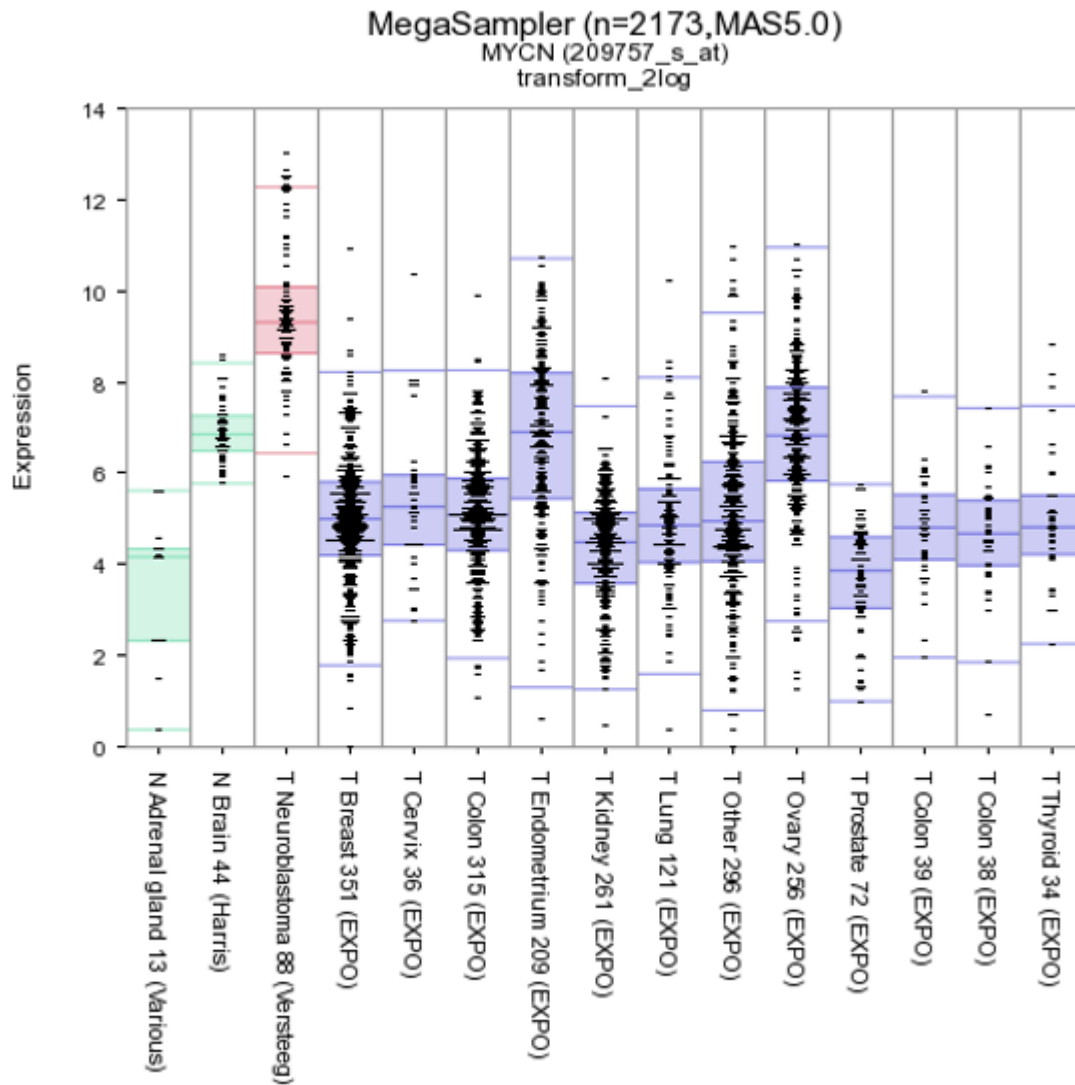
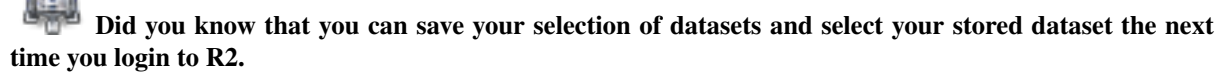
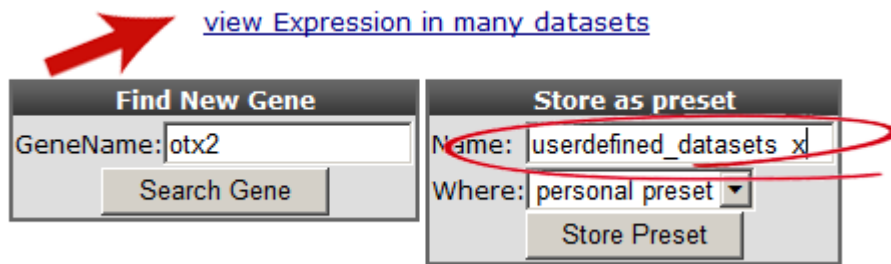


Figure 6: MYCN expression levels in 15 datasets covering 2173 samples.

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.





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).

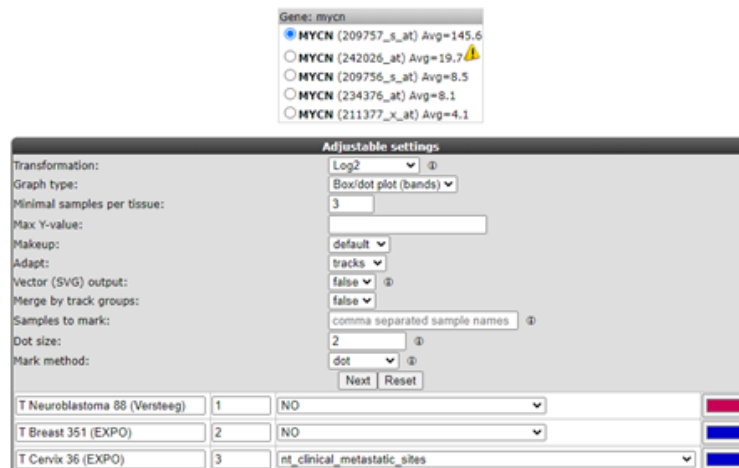


Figure 8: Adjustable settings panel, color groups within adataset.

10.4 Step 3: Stacking subgroups (or datasets)

It could be that you also want to stack subgroups of datasets in one singlebox (or bar etc) in such way that each single box contains one subgroup of multiple datasets for a selected track. Keep in mind that the track name and the corresponding subgroups must have exactly the same spelling since R2 is checking this in the background. To illustrate this we make use of the EXPO datasets which are curated for their annotation. After selection the datasets, make sure that the Merge track by groups is set to **true** and you have selected a track in this case the *histology* track and click submit.

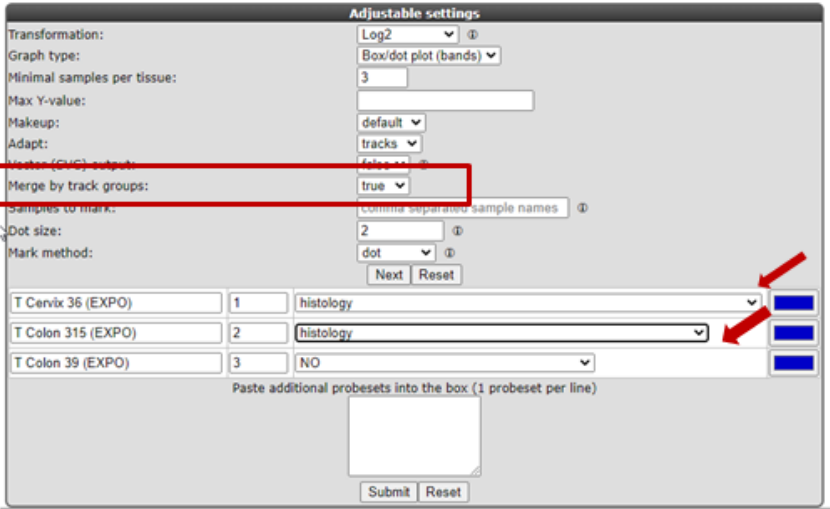


Figure 9: Adjustable settings panel, stacking subgroups.

Now the expression level of the TP53 gene for a single dataset is plotted next to the separate subgroups of the histology track, each box containing the expression levels for single gene of two datasets divided over the subgroups. Of course there is a big chance that you're not so lucky that tracks and their subgroups have the same spelling or you want to stack different subgroups for your research questions. In that case you have to create for each dataset new subgroups with the same spelling for each dataset. You can create these customzid tracks you want to incorporate in the user section of the main page of R2. Once created you can select those in the megasampler section. In case you want to stack complete datasets in one box/bar you have to make a track with a subgroups containing all the samples.

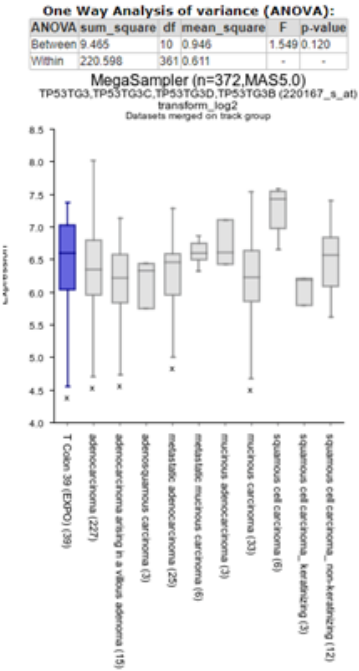


Figure 10: Adjustable settings panel, stacking subgroups.

10.5 Step 4: Expression distribution over many datasets

The blue link [view expression in many datasets](#) 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.

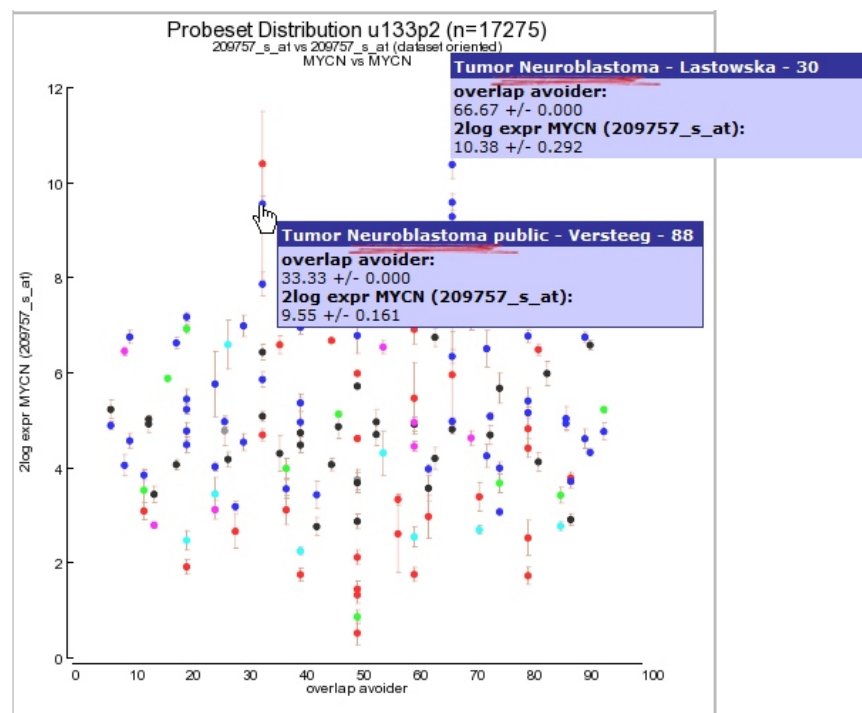


Figure 11: MYCN expression level distribution for all u133-2 datasets in R2.

2. Via 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.

10.6 Step 5: 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.

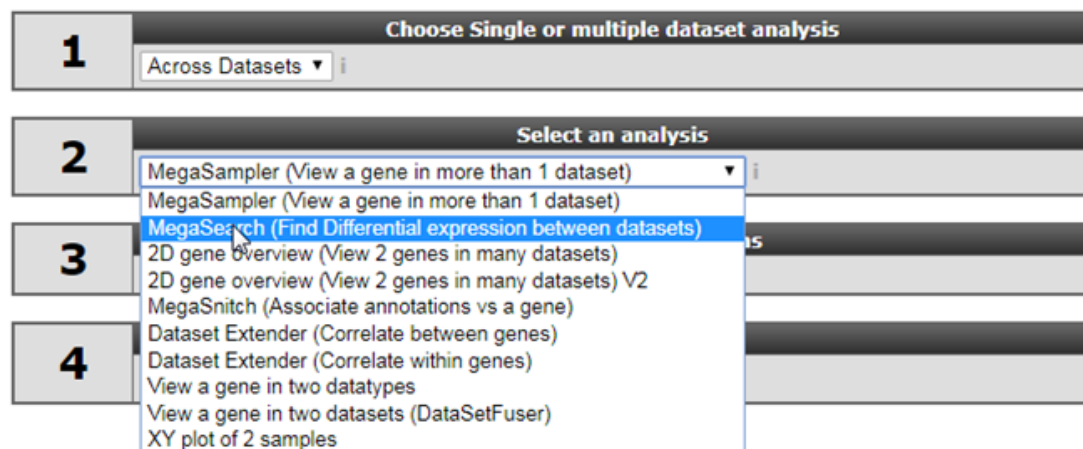
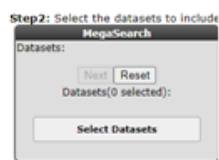


Figure 12: Megasearch select..

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.



Drag a column and drop it here to group by that column																											
	Spec.	Y	Data type	Y	Categ.	Tissue/Tumor	Y	Dataset	Y	Platform	Y	Normalized	Y	N	Y	Author	Y	Accession	Y	Release da.	Y	R2 date	Y	Access	Y	Favo.	
	Select	Y	Filter	Y	Select	Filter	Y	Select	Filter	Y	Select	Filter	Y	Select	Filter	Y	Select	Filter	Y	Select	Filter	Y	Select	Filter	Y	Select	Filter
<input checked="" type="checkbox"/>	na		Expression data	Exp	Cell line	Medulloblastoma		pt_aml_med		u133p2		MASS.0		34		Kool		GSE7579		2007-05-01		2007-05-01		restricted			
<input type="checkbox"/>	na		Expression data	Exp	Medulloblastoma D425			pt_aml_med		u133p2		MASS.0		32		Kool		GSE7579		2007-11-16		2009-11-16		restricted			
<input type="checkbox"/>	na		Expression data	Exp	Medulloblastoma D425 Group			pt_aml_med		u133p2		MASS.0		18		Kool		GSE7579		2007-01-01		2009-09-24		restricted			
<input type="checkbox"/>	na		Expression data	Exp	Medulloblastoma D425			pt_aml_med		u133p2		MASS.0		15		Benson		GSE7579		2007-09-01		2007-09-15		public			
<input type="checkbox"/>	na		Expression data	Exp	Medulloblastoma D425			pt_aml_med		u133p2		MASS.0		9		Versteeg		GSE7579		2007-01-01		2007-01-01		restricted			
<input type="checkbox"/>	na		Expression data	Exp	Medulloblastoma MED44			pt_aml_med		u133p2		MASS.0		10		Kool		GSE7579		2007-01-01		2007-01-01		restricted			
<input type="checkbox"/>	na		Expression data	Tumor	Medulloblastoma			pt_aml_med		u133p2		MASS.0		97		Delattre		GSE7579		2007-01-01		2009-01-09		restricted			
<input type="checkbox"/>	na		Expression data	Tumor	Medulloblastoma			pt_aml_med		u133p2		MASS.0		76		Gilman		GSE7579		2007-01-01		2012-07-15		public			
<input type="checkbox"/>	na		Expression data	Tumor	Medulloblastoma			pt_aml_med		u133p2		MASS.0		62		Kool		GSE7579		2007-01-01		2008-04-17		restricted			
<input type="checkbox"/>	na		Expression data	Tumor	Medulloblastoma			pt_aml_med		u133p2		MASS.0		223		Walter		GSE7579		2017-03-27		2017-03-27		public			
<input type="checkbox"/>	na		Expression data	Tumor	Medulloblastoma			pt_aml_med		u133p2		MASS.0		120		Versteeg		GSE7579		2007-01-01		2009-01-09		restricted			
<input type="checkbox"/>	na		Expression data	Tumor	Medulloblastoma (Dm)			pt_aml_med		u133p2		MASS.0		73		Walter		GSE7579		2014-06-01		2019-10-06		public			
<input checked="" type="checkbox"/>	na		Expression data	Tumor	Medulloblastoma - ATOT			pt_aml_med		u133p2		MASS.0		31		Heath		GSE7579		2015-04-15		2016-08-18		public			
<input type="checkbox"/>	na		Expression data	Tumor	Medulloblastoma Ependymoma			pt_aml_med		u133p2		MASS.0		51		Ben-Bar		GSE7579		2015-10-21		2015-11-12		public			
<input type="checkbox"/>	na		Expression data	Tumor	Medulloblastoma PLUS One			pt_aml_med		u133p2		MASS.0		62		Kool		GSE7579		2008-09-02		2009-02-17		public			
<input type="checkbox"/>	na		Expression data	Tumor	Medulloblastoma public			pt_aml_med		u133p2		MASS.0		97		Delattre		GSE7579		2007-01-01		2012-02-24		public			

Figure 13: 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.

Adjustable settings

Transformation: Log2 ⓘ

Dataset	Value	Action
T AML 179 (Ley)	1	NO
T AML 79 (Metzeler)	1	NO
N Brain 172 (Berchtold)	1	NO
T AML 140 (Bohlender)	1	NO
T Medulloblastoma - ATRT 31 (Hsieh)	2	NO
C Medulloblastoma 14 (Kool)	2	NO

Submit Reset

Adjustable settings

Chromosome: All ⓘ

Gene ontology: All Search GO

Gene set: Search GS

Manual list: none ⓘ

Hugooonce Dataset: ps_amlc_medulloc14_u133p2

medulloc14 (All Samples)

X 1 2 ID

ALL

aml179 (All Samples)

X 1 2 ID

ALL

aml79 (All Samples)

X 1 2 ID

ALL

brainnorm172 (All Samples)

X 1 2 ID

ALL

gse37642geo140 (All Samples)

X 1 2 ID

ALL

gse67851geo31 (All Samples)

X 1 2 ID

ALL

Next Reset

Figure 14: Assign the statistical group for testing

1. In the next adjustable settings menu select at Genecategory 'transcription regulator Act' for gene filtering. In the 'Hugooonce 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 'Hugooonce 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_modarate_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.
- 2.

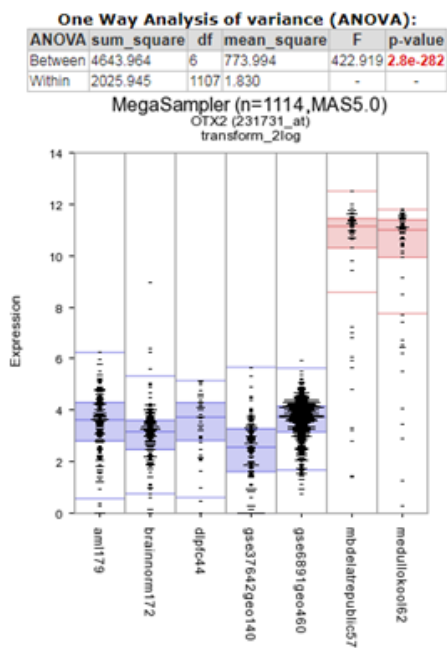


Figure 15: OTX2 BoxDotplot

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.

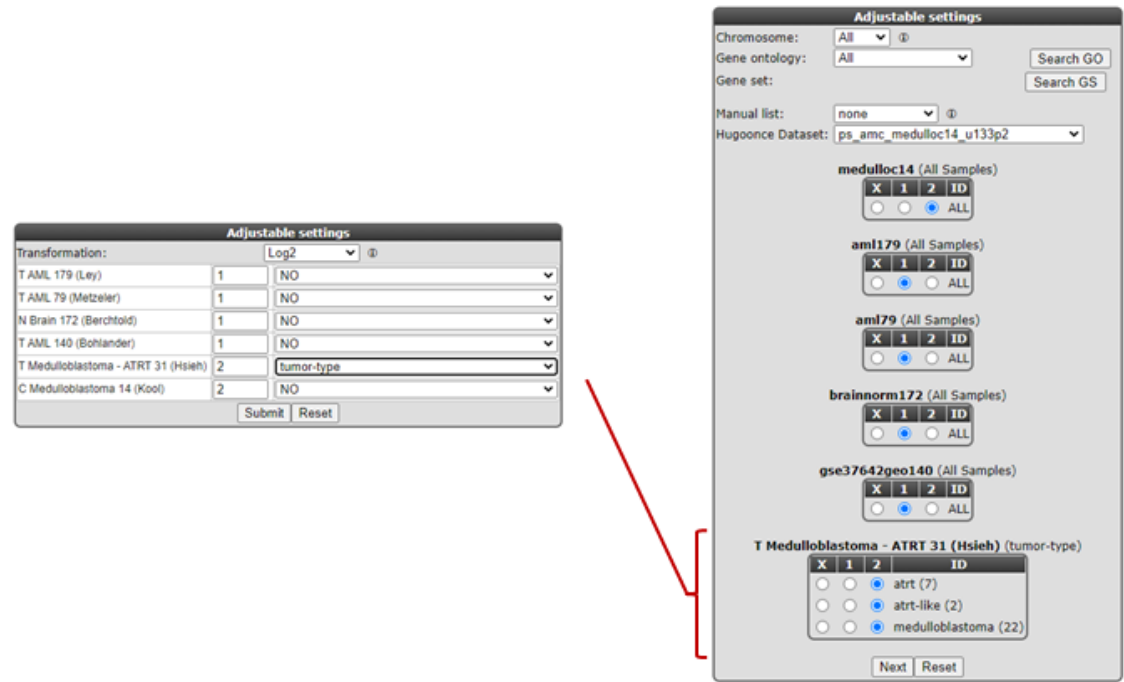


Figure 16: Assign the statistical subgroup for testing



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

10.7 Final remarks / future directions

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

K-means clustering in R2

How to discover novel groups or subtypes in your dataset using k-means clustering

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)

3,119,412 (2,870,389 unique) samples available

1 Choose single or multiple dataset analysis

Single Dataset ▼ ⓘ

2 Select a dataset for analysis

Tumor Medulloblastoma PLoS One - Kool - 62 - MAS5.0 - u133p2 ▼

3 Select type of analysis

K-means ▼ ⓘ

4 Proceed

Next Reset

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

4. Click ‘next’

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.

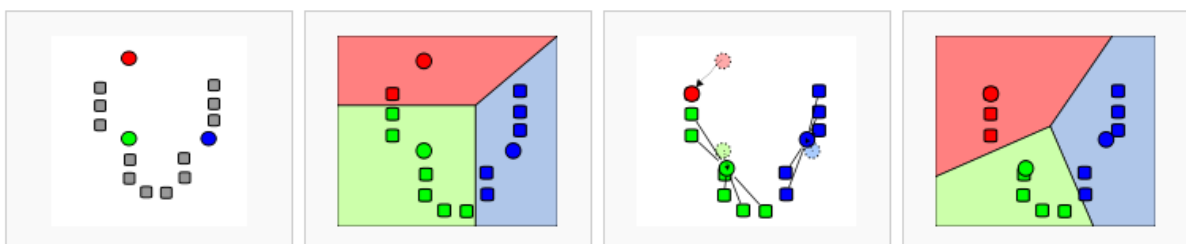
Tumor Medulloblastoma PLoS One - Kool - 62 - MAS5.0 - u133p2 public ⓘ

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).

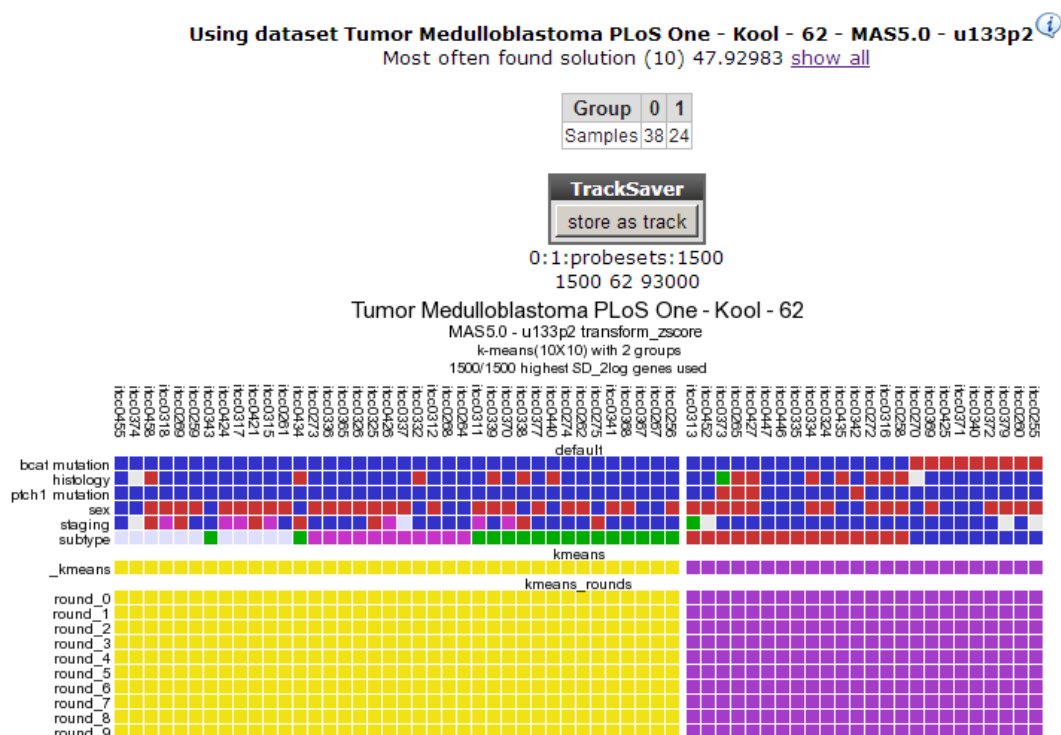


Figure 3: Results for the 10x10 k-means clustering in two groups; two consistent clusters are formed.

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 *k*means 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'.

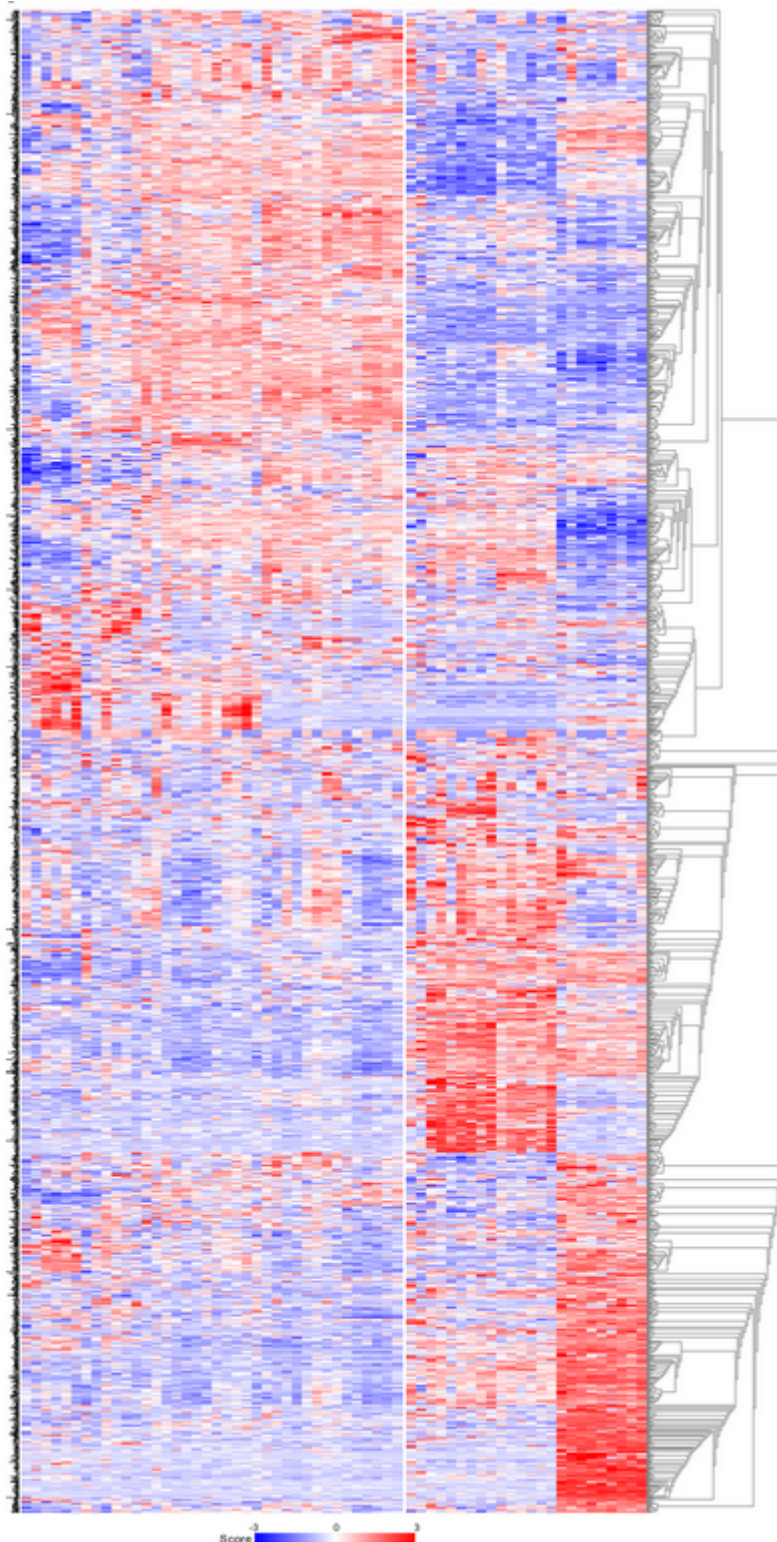


Figure 4: The heatmap for the k-means clustering in 2 groups; it is obvious that the data is represented in the clusters.



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

A solution (lowest sum of distances = 28.63002) was found in 1/10 rounds with 10 passes

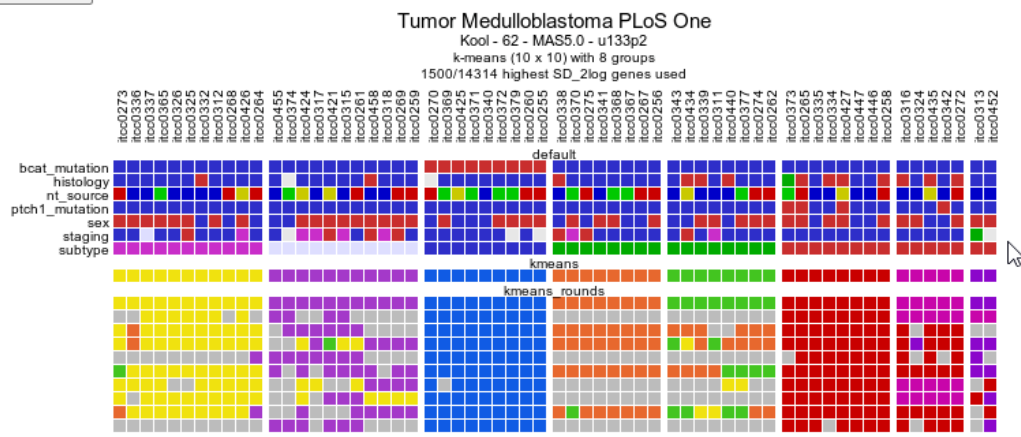
▼ Show all

Round 1: the solution with the lowest sum of distances (29.45377) was found in 1/10 passes
 Round 2: the solution with the lowest sum of distances (29.84505) was found in 1/10 passes
 Round 3: the solution with the lowest sum of distances (29.51494) was found in 1/10 passes
 Round 4: the solution with the lowest sum of distances (29.08444) was found in 1/10 passes
 Round 5: the solution with the lowest sum of distances (30.16343) was found in 1/10 passes
 Round 6: the solution with the lowest sum of distances (29.55448) was found in 1/10 passes
 Round 7: the solution with the lowest sum of distances (29.99220) was found in 1/10 passes
 Round 8: the solution with the lowest sum of distances (29.73602) was found in 1/10 passes
 Round 9: the solution with the lowest sum of distances (30.24882) was found in 1/10 passes
 Round 10: the solution with the lowest sum of distances (28.63002) was found in 1/10 passes

Cluster	Samples
0	11
1	11
2	9
3	8
4	8
5	8
6	5
7	2

Store as track

Tumor Medulloblastoma PLoS One - Kool - 62 - MAS5.0 - u133



A solution (lowest sum of distances = 33.12661) was found in 1/10 rounds with 10 passes

► Show all

Cluster	Samples
0	25
1	15
2	9
3	9
4	4

Store as track

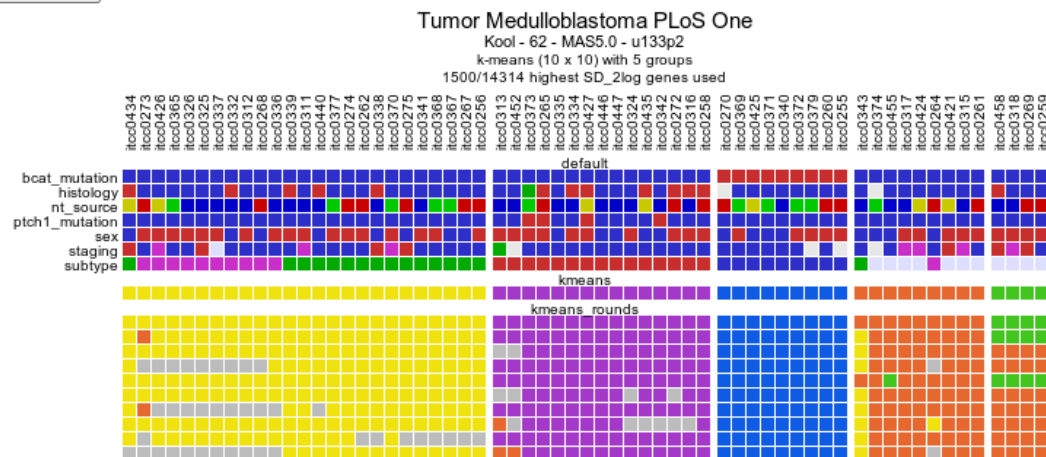


Figure 7: Clustering in 5 groups

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, Mrcic A, Ylstra B, Grajkowska W, Hartmann W, Pietsch T, Ellison D, Clifford SC, Versteeg R.; Integrated genomics identifies five medulloblastoma subtypes with distinct genetic profiles, pathway signatures and clinicopathological features. PLoS One. 2008 Aug 28;3(8):e3088.

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 convention (fixed prefix, followed by `_up` and `_down`; e.g. `mycn_up` and `mycn_down`).

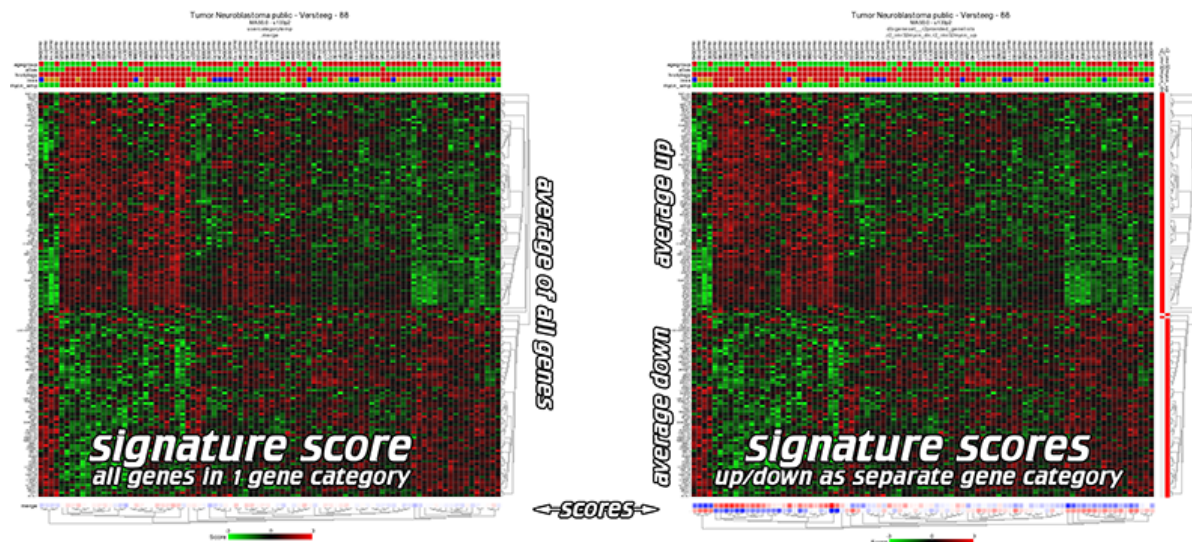


Figure 1: Signature score: one category vs up/downcategory

- 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 Track 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 r2 provided genelist category.
5. Genelists>oncogenomics_valentijn>functional genesignatures the mycn_dn and _up genesets

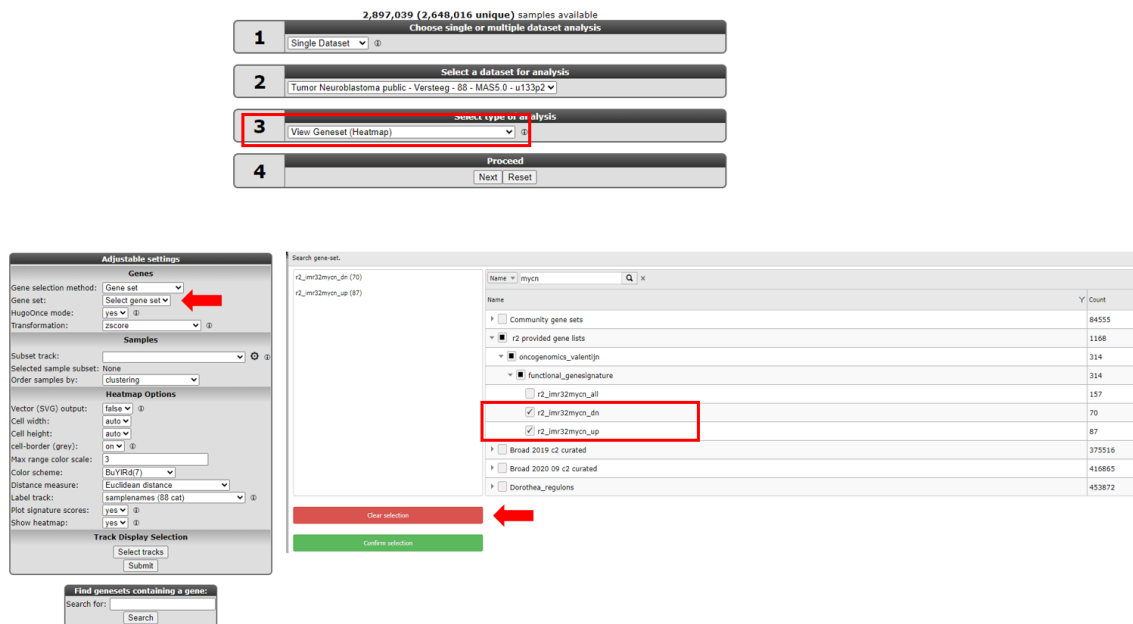


Figure 2 : Generating Geneset signatures

6. Confirm the selection with the Green Button.
7. R2 will produce a hierarchical clustered heatmap image of the selected gene categories. Note that at the right side of the heatmap the red markings indicate in which category a particular gene was represented (Figure 3, box A). In the bottom part of the heatmap (box B in Figure 3), a blue-white-red colorscale is depicted for both gene categories. We can clearly see the opposing effects of the 2 signatures. A third colorscale depicts a weighted score, based on the contributions of both signatures (see point 8).
8. Scrolling down on this page, we will encounter a heading “Gene set values” (Figure 3, the heatmap), 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.

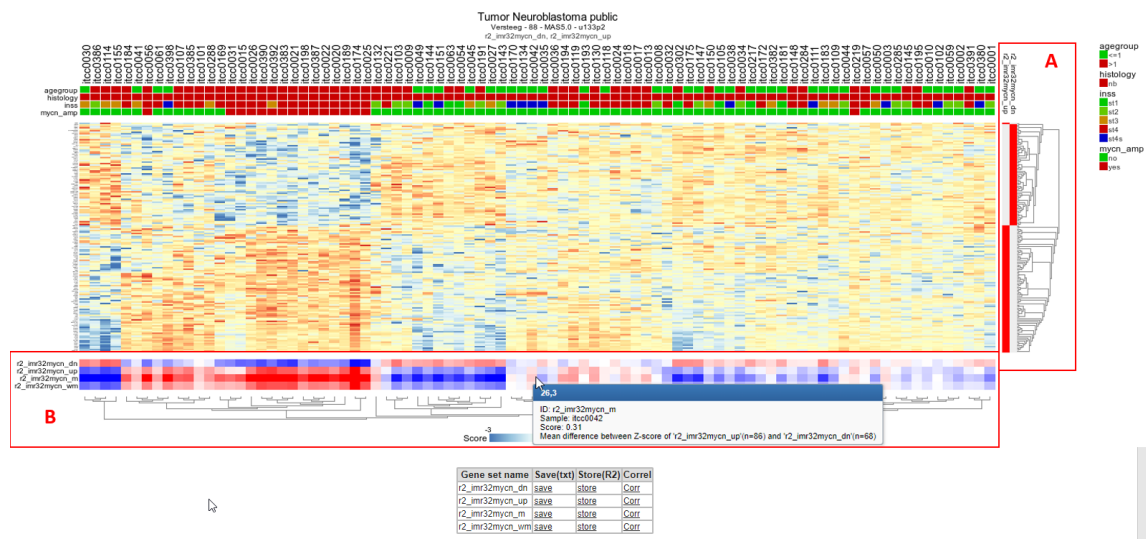


Figure 3: A) Gene set marking per gene; B) Signature score;

9. 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 4).

Adjustable settings

Boxplot statistics




maximum	1.24
median	-0.05
minimum	-1.02
n	88
q05	-0.75
q25	-0.34
q75	0.29
q95	1.01
toomax	1.51
toomin	n

Samples

sample updating suppressed (only numeric and >15 unique values)

Groups

C: / N:74 / U:

min_val:	-1.5	
center_val:	0	
max_val:	1.5	

Track Settings

Track name:

Show as track: ⓘ

Where:

Description (usergroups only):

weighted mean difference between
z-score of
'r2_imr32mycn_up'(n=86) and
'r2_imr32mycn_dn'(n=68)

Figure 4: Generating a Track from a gene set Signature Score

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 “Submit” (Figure 5).

counting... samples available

1 Choose single or multiple dataset analysis

Single Dataset

2 Select a dataset for analysis

Tumor Neuroblastoma public - Versteeg - 88 - MAS5.0 - u133p2

3 Select type of analysis

Correlate Gene with track

4 Proceed

Next Reset



Tumor Neuroblastoma public - Versteeg - 88 - MAS5.0 - u133p2 public

Adjustable settings

Analysis type: gene vs track

Gene / Reporter: MYCN 209757_s_at advanced

Track: r2_imr32mycn_wm_4856 (#)

Transformation: Log2

Sample Filter

Subset track:

Selected sample subset: None

Graphics

Graph type: YY plot with annotation

Extra Graph Option: off

Color mode: Default Color

Samples to mark: comma separated sample names

Track Display Selection

Select tracks

More Settings +

Submit

Figure 5: Gene MYCN vs signature score

- R2 will now produce a plot where the signature score for every patient is related to the MYCN mRNA expression value (Figure 6).
- 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. Also check out other settings, such as the dot size, that become available when you click on “More Settings”. Click “Adjust Settings” to redraw.
- We can now clearly see that MYCN amplified patients have a higher MYCN gene set activity score. The possibilities for numerical tracks are endless with some smart questions (Figure 6).

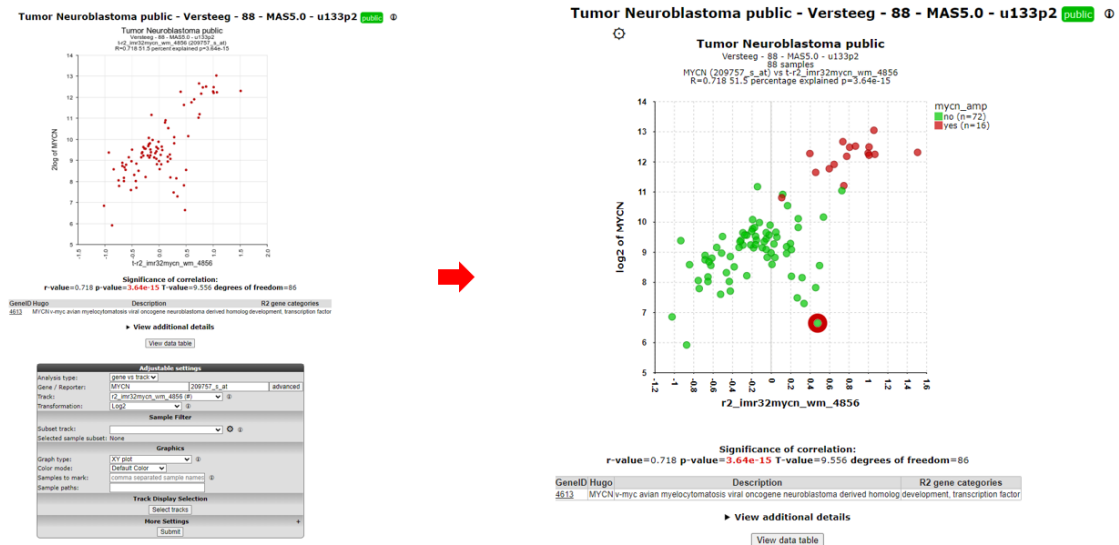


Figure 6: Add group colouring

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 signature gene lists 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" (Figure 7, left).
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_2015_oncogenic' (Figure 7, right). Then click "next".

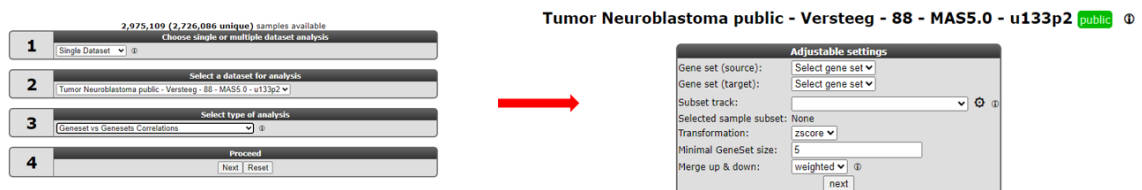


Figure 7: Geneset vs Geneset correlation

3. In the next screen, click the **Select geneset** at Geneset(source). In the pop-up selection grid select at r2 provided genelists > functional_genesignature > oncogenomics_groningen_Natgen_2017 > 'both genesets mes and adrn'. Click the green "Confirm selection button".
4. Follow the same procedure for Gene set (target) and select some genesets of interest like Cellular processes in the KEGG pathway section, again Click the green "Confirm selection button".

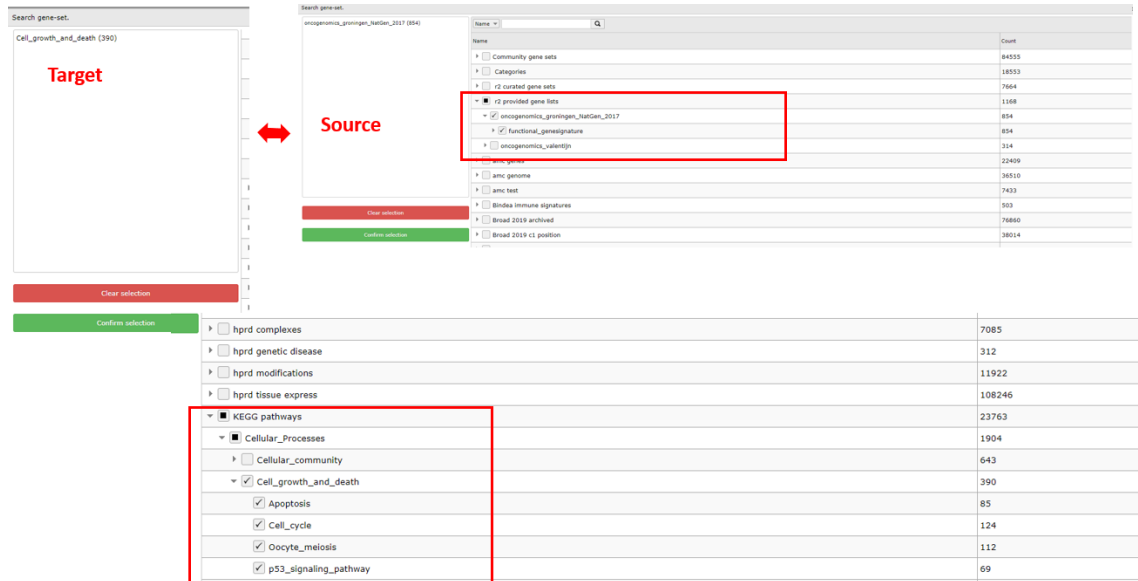


Figure 8: Choose the genesets

- R2 has now generated all the possible correlations for the selected MYCN signature geneset against all the gene lists within the selected KEGG pathway categories. 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 number of genes (same) between two gene lists (source and target). 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.

Tumor Neuroblastoma public - Versteeg - 88 - MAS5.0 - u133p2 public ⓘ

Genesets	r-value	df	t-value	p-value	remark	venn source / same / target	Range (source / target)
r2_mesadrn_mes vs apoptosis	0.639	86	7.711	2.02e-11	xy-plot	481 / 2 / 82	2.452 / 1.095
r2_mesadrn_adrn vs oocyte_meiosis	0.571	86	6.442	6.50e-09	xy-plot	366 / 2 / 105	1.383 / 0.675
r2_mesadrn_adrn vs apoptosis	-0.442	86	-4.568	1.63e-05	xy-plot	366 / 2 / 82	1.383 / 1.095
r2_mesadrn_adrn vs cell_cycle	0.431	86	4.429	2.77e-05	xy-plot	363 / 5 / 118	1.383 / 1.559
r2_mesadrn_mes vs cell_cycle	-0.269	86	-2.593	0.011	xy-plot	482 / 1 / 122	2.452 / 1.559
r2_mesadrn_mes vs oocyte_meiosis	-0.185	86	-1.750	0.084	xy-plot	482 / 1 / 106	2.452 / 0.675
r2_mesadrn_mes vs p53_signaling_pathway	0.158	86	1.485	0.141	xy-plot	481 / 2 / 66	2.452 / 1.230
r2_mesadrn_adrn vs p53_signaling_pathway	0.040	86	0.374	0.709	xy-plot	366 / 2 / 66	1.383 / 1.230

Adjustable settings

Gene set (source): oncogenomics_groningen_NatGen_2017

Gene set (target): Cell_growth_and_death

Subset track: ▼ ⓘ

Selected sample subset: None

Transformation: zscore

Minimal GeneSet size: 5

Merge up & down: weighted ⓘ

next

Figure 9: Correlations and overlap between genesets

- To inspect the correlation in more detail, we can click on the “XY-plot” link.
- 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 10, left side).
- We can also inspect the target signature as a heatmap by clicking on the “View heatmap of”, providing gene-by-gene information (see Figure 10, right side).

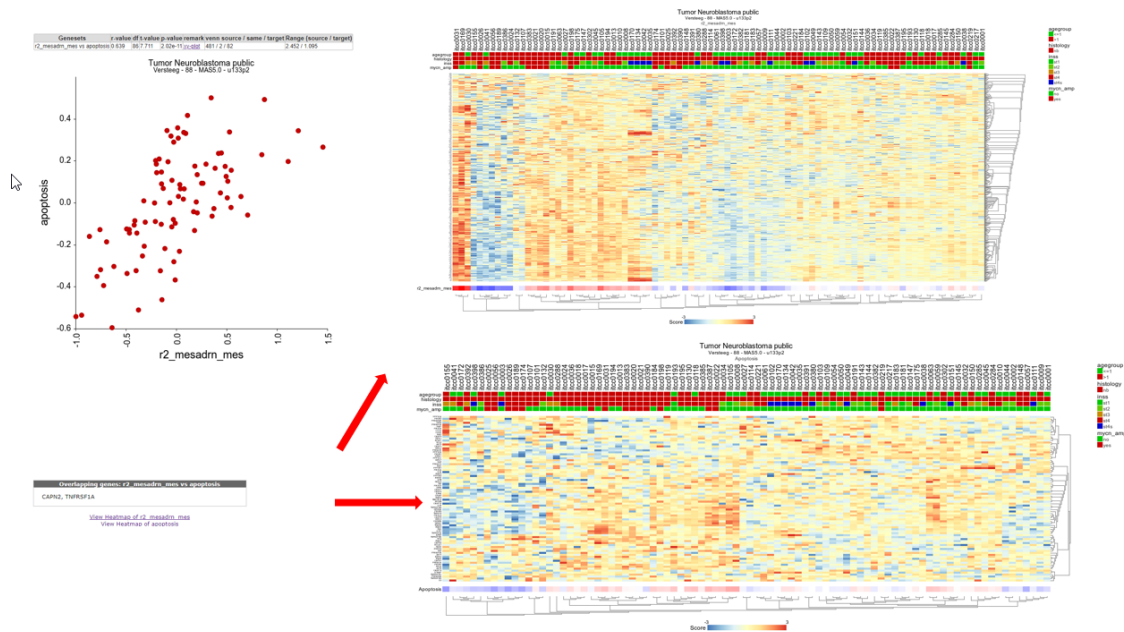


Figure 10: XY signature score plot and heatmap of correlated gene sets

12.5 Step 4: Plot signature scores using the relate 2-tracks module.

In the previous steps we have plotted the genesignature scores directly from a list 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 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 so far.

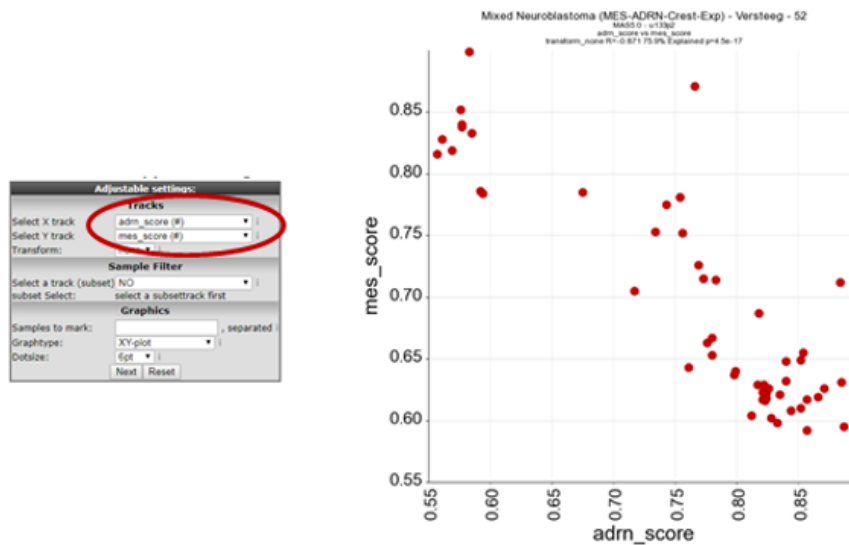


Figure 11: Relate 2 tracks using genesignatures

4. In order to visualise the biological relevance 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.

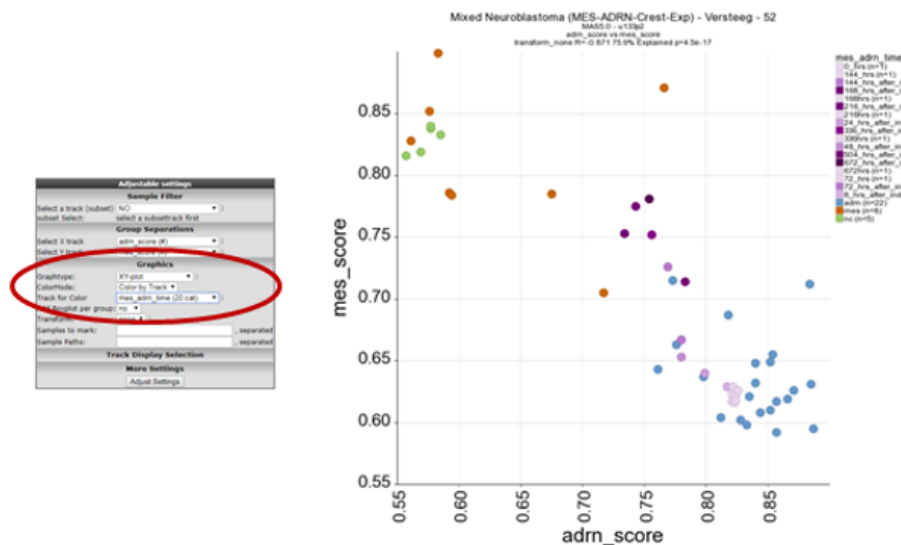


Figure 12: Color by a biologically meaningful track

5. 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"

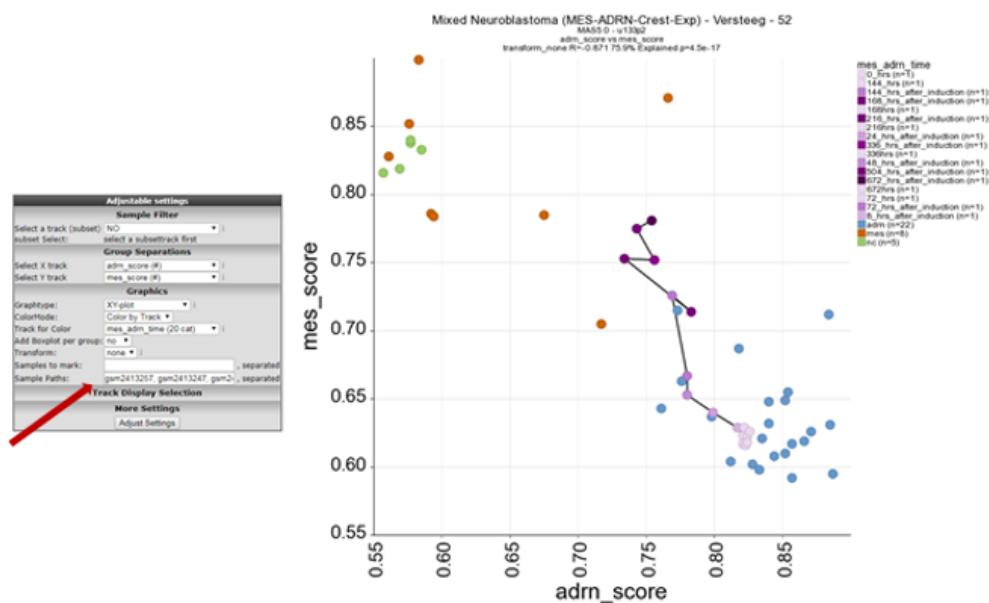


Figure 13: Connecting samples

2. In figure 13 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'

12.7 Step 6: Signature Gene correlations

You can use the gene signature correlation option in order to identify genes which correlate best with the gene signature you are investigating.

1. In the 'Gene set values' table below the Heatmap of Step 1, where you stored the genesignature score previously, this time click the link 'Corr' (Figure 3, box C).

Gene set name	Save(txt)	Store(R2)	Correl
r2_imr32mycn_dn	save	store	Corr
r2_imr32mycn_up	save	store	Corr
r2_imr32mycn_m	save	store	Corr
r2_imr32mycn_wm	save	store	Corr

Figure 14: Select Gene Correlations

2. This option generates a graph where the R-value is ranked from the highest to the lowest correlation for each member of the gene set that you used to generate the signature score. Clicking on a row in the table will generate a XY-plot. The scatter plot shows the gene expression (Y-axis) against the signature score value (X-axis) for each sample.

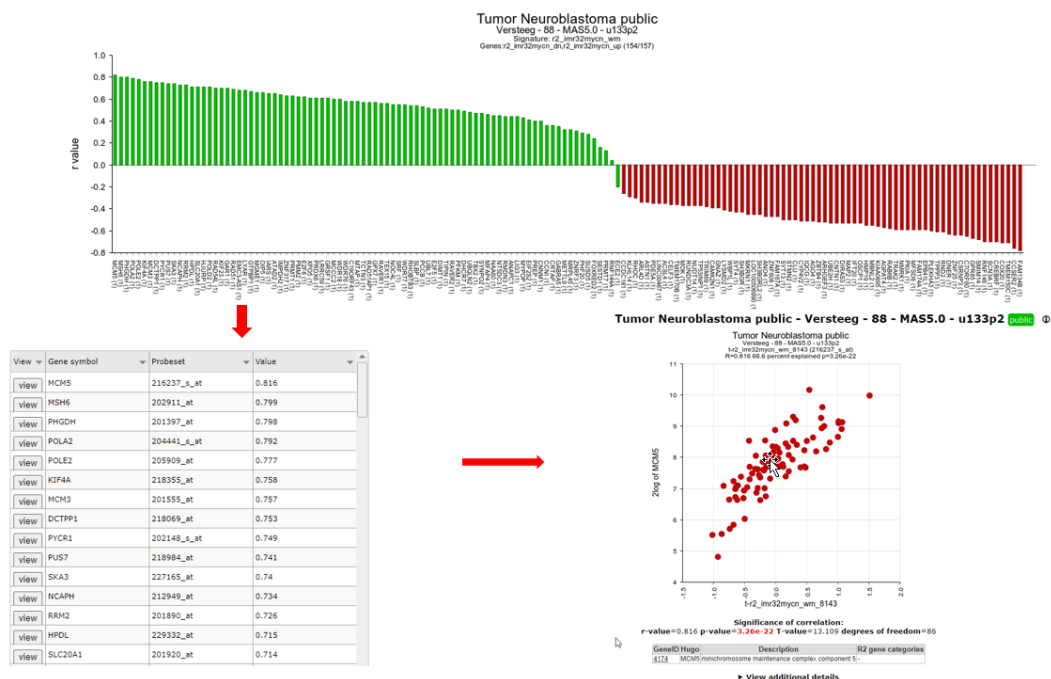


Figure 15: Ordered R-values + XY-plot

You can also select multiple categories to investigate the individual contribution of genes to a signature score. R2 will automatically keep the coloring for the separate gene categories.

12.8 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.

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 access. 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.

1 Select a time series collection
Collection:

2 Select type of analysis

3 Select Additional Conditions
View Genes
View a gene
Create lists
Create a list of genes

4 Proceed

Figure 1: Single selection in the Time-series module

- In field 1, select at 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 a case a publically available Times series data. Select “View a gene; in field 2, type *HMOX1* in field 3*,* and click “next”.
- 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 where 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”.

set_public_u133p2

- ☒ A549
 - ☒ GSE17708
 - ☒ TGFB_1
 - ☒ TGFB_2
 - ☒ TGFB_3
 - ☐ BEC
 - ☐ H929
 - ☐ HCT116
 - ☐ HEPG2
 - ☐ HUVEC
 - ☐ IMR32
 - ☐ MCF7

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 of the upregulated genes after TGF-beta stimulation. It should be noted that a control experiment is missing in this experimental design. Hovering over the individual timepoint reveals additional information.

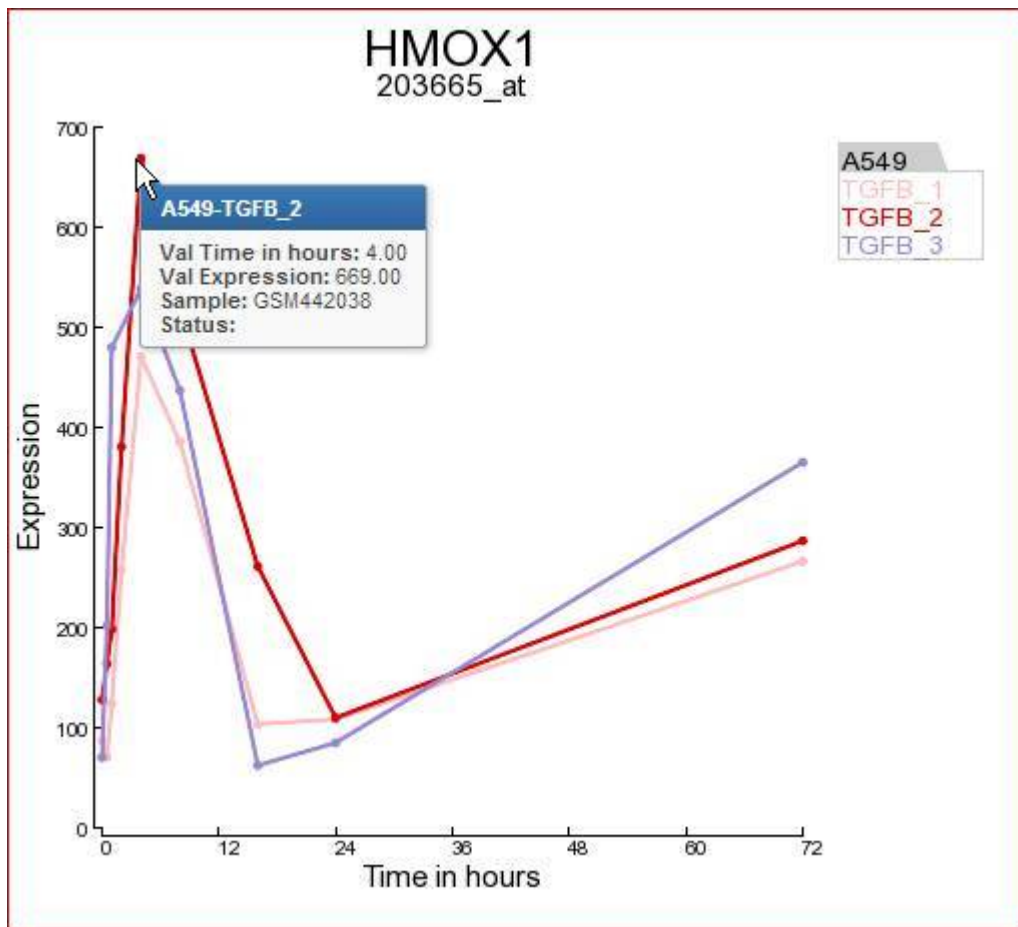
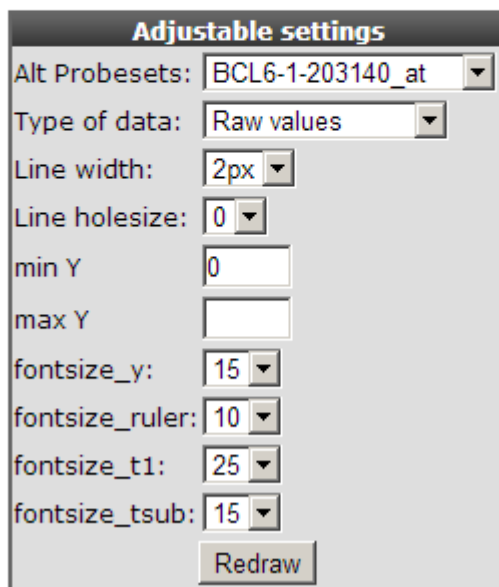


Figure 3: Expression levels of the HMOX1 gene during a time course experiment

- 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".

ProbesetVerification (hg18)

symbol	probeset	rank	gene overlap	exon overlap	probes found	Link
BCL6	203140_at	1 GS	YES	YES	YES	R2 TView
BCL6	215990_s_at	2 GS	YES	YES	YES	R2 TView
BCL6	228758_at	3 GS	YES	NO	NO	R2 TView

A screenshot of a software window titled "Adjustable settings". It contains several configuration options: "Alt Probesets:" with a dropdown menu showing "BCL6-1-203140_at"; "Type of data:" with a dropdown menu showing "Raw values"; "Line width:" with a dropdown menu showing "2px"; "Line holesize:" with a dropdown menu showing "0"; "min Y" with a text input field showing "0"; "max Y" with an empty text input field; "fontsize_y:" with a dropdown menu showing "15"; "fontsize_ruler:" with a dropdown menu showing "10"; "fontsize_t1:" with a dropdown menu showing "25"; and "fontsize_tsub:" with a dropdown menu showing "15". At the bottom right is a "Redraw" button.

Adjustable settings

Alt Probesets: BCL6-1-203140_at

Type of data: Raw values

Line width: 2px

Line holesize: 0

min Y: 0

max Y:

fontsize_y: 15

fontsize_ruler: 10

fontsize_t1: 25

fontsize_tsub: 15

Redraw

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.

The screenshot shows the 'TableBuilder' web interface. It is divided into several sections:

- TableBuilder**: Contains input fields for 'Min # experiments' (2), 'Min highest expression in series' (150), 'Min # present calls in series' (2), 'Min # significant changes in series' (1), 'Min best logfold in series' (2), 'Fold orientation' (All), 'Min lowest change pvalue in series' (0.0001), and 'Force single reporter for hugo?' (yes).
- Gene Filtering**: Contains dropdown menus for 'Chromosome' (All), 'GeneCategory' (All), and 'KEGG Pathway' (All).
- Use Correlation too?**: A section with a single checkbox.
- Graphics settings**: A section with a single checkbox.
- Expert Settings**: A section with a single checkbox.

 At the bottom, there are 'Submit' and 'Reset' buttons.

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 (eg 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 pvalue of 0.00025 should be met. Here you can set the pvalue 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”.

Using dataset set_public_u133p2

9 identical timepoints detected

Counter	Gene	Probeset	A549-TGFB_1		A549-TGFB_3		A549-TGFB_2		hCNT
			max	fold	max	fold	max	fold	
1	IGFBP7	201163_s_at	2199	6.5	1909	9.9	2496	12.9	3
2	IGFL1	239430_at	5656	9.9	4624	8.4	4517	9.8	1
3	IL11	206924_at	2624	6.9	2027	7.5	2467	9.9	2
4	IGFBP5	211959_at	3821	9.5	3190	9.5	3712	5.6	6
5	INHBA	227140_at	536	9.2	448	9.2	650	6.4	3
6	RUNX2	232231_at	349	7.2	342	9.2	460	6.9	5
7	MAF	209348_s_at	2032	7.0	1967	9.1	2550	8.1	4
8	MMP10	205680_at	164	5.5	508	8.8	393	7.5	1
9	ANGPTL4	221009_s_at	2588	6.2	1687	6.6	3358	8.0	2
10	FRMD5	230831_at	238	3.8	371	8.0	294	3.5	2
11	RASGRP3	205801_s_at	284	6.6	274	6.3	239	7.8	2
12	KCNMA1	221584_s_at	NA	NA	754	7.7	893	5.2	4
13	MARCH4	230112_at	577	6.5	556	7.7	478	6.7	1
14	SPOCK1	202363_at	1963	6.1	1278	7.7	1988	7.1	1

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

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.

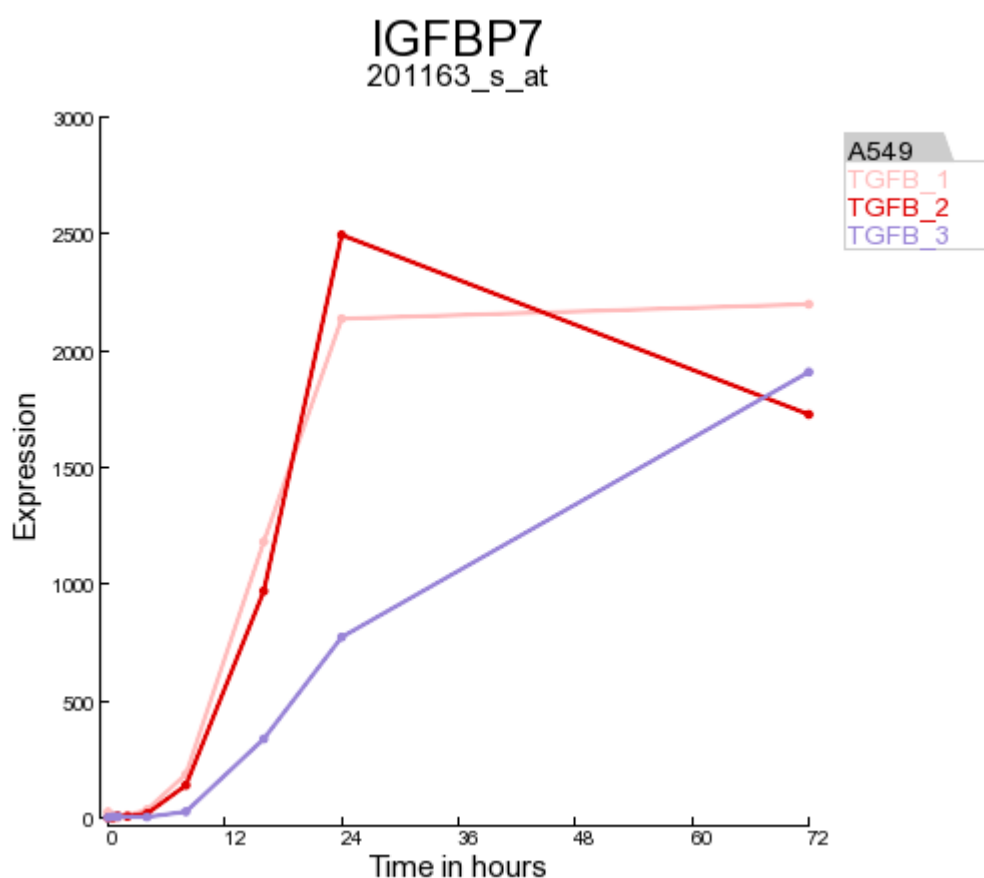
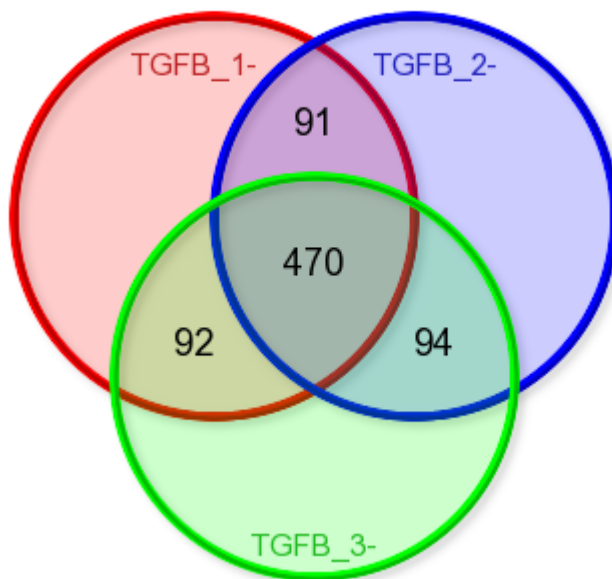


Figure 7: regulated gene.

- 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 genecategories of choice?**

In the ‘User Options’ menu section you can upload text files containing your lists of genes and store them as gene category. Repeating the procedure described above will produce the desired Venn diagrams.

13.4 Step 3: Using the regulated genes in further analyses

- 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 genecategory and add a short description. The genecategory 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 genecategory for the next step.

Custom Category:

CategoryName:

GeneSymbol File:

Species:

Where:

Description

Arraytype: u133p2

Dataset: set_public_u133p2

Date: 2013-05-06 (17:12:31)

Describe: Time series module from R2. Version 1.05

Groups of genes (sublist) to include (none selected=all)

☐ |A549-TGFB_1|A549-TGFB_2

☐ |A549-TGFB_1|A549-TGFB_2|A549-TGFB_3

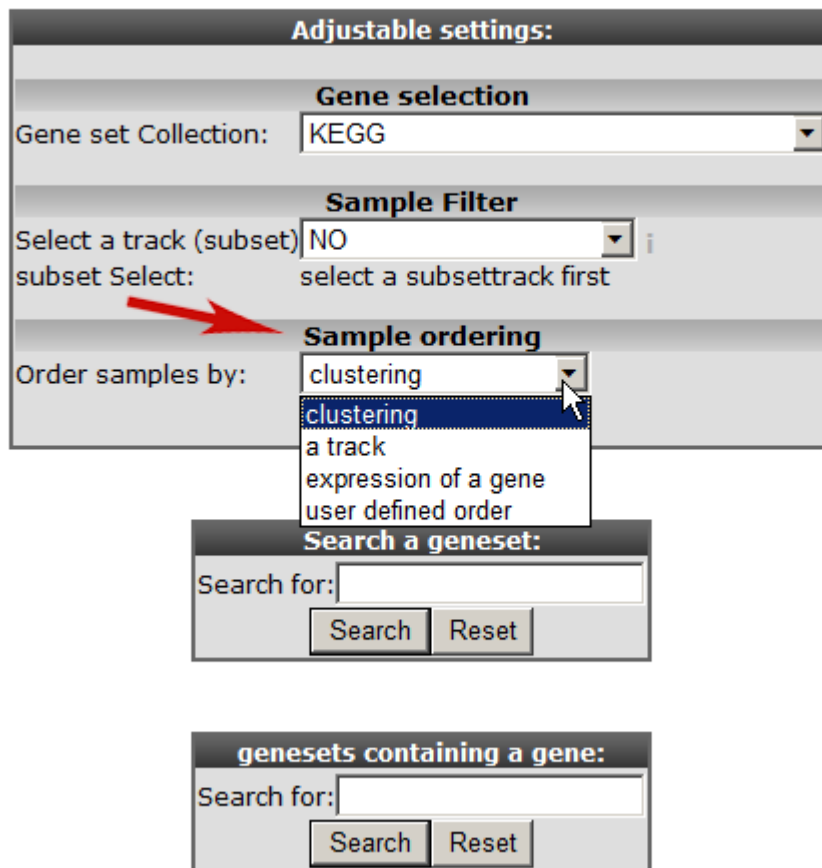
☐ |A549-TGFB_1|A549-TGFB_3

☐ |A549-TGFB_2|A549-TGFB_3

Figure 10: : Store a gene category

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".

Using dataset Tumor Breast - Iglehart - 123 - MAS5.0 - u133p2



Adjustable settings:

Gene selection
Gene set Collection: KEGG

Sample Filter
Select a track (subset): NO
subset Select: select a subsettrack first

Sample ordering
Order samples by: clustering

Search a geneset:
Search for:
Search Reset

genesets containing a gene:
Search for:
Search Reset

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”.

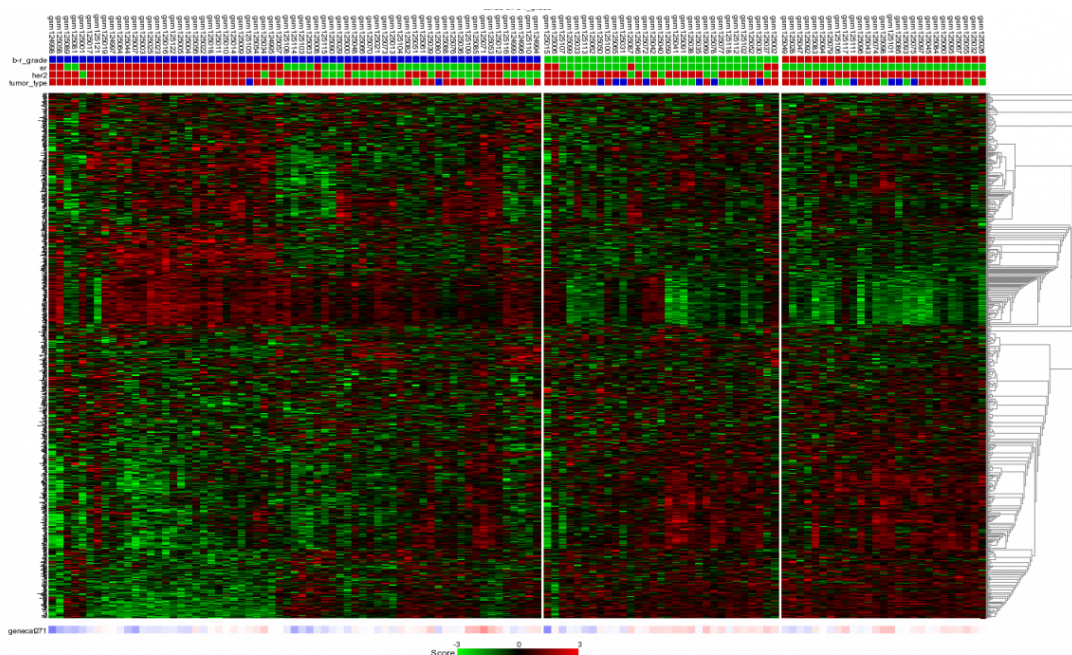


Figure 12: Heatmap of unsupervised clustering within a track of a selected geneset.

- 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 Receptor 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 heatmap rectangle generates a one-gene-view for the chosen gene in the dataset. 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 'User Options' menu item.



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

- 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”.
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-TGFB_1

2: A549-TGFB_2

3: A549-TGFB_3



Foldchange_Pos-group							Foldchange_Neg-group						
		Cell line			Tissue				Cell line			Tissue	
#	Gene Symbol	1	2	3	pvalue	links	#	Gene Symbol	1	2	3	pvalue	links
1	ABAT				1.2e-24	2 View	1	CA12				7.2e-28	2 View
2	CXXC5				5.7e-18	2 View	2	TFF1				9.1e-22	2 View
3	HECTD2				2.3e-17	2 View	3	SLC7A2				4.0e-21	2 View
4	RARA				4.1e-16	2 View	4	CDCA7				2.5e-20	2 View
5	ADAMTS15				6.4e-16	2 View	5	XBP1				3.5e-19	2 View
6	RUNX1				1.2e-15	2 View	6	CDC20				5.7e-19	2 View
7	PPP1R14C				1.5e-15	2 View	7	PER2				2.3e-18	2 View
8	DSC2				1.2e-14	2 View	8	MYO5C				3.5e-18	2 View
9	INPP4B				8.0e-14	2 View	9	MCM6				4.3e-18	2 View
10	NANOS1				1.2e-13	2 View	10	VAV3				6.9e-18	2 View
11	SEMA3C				3.7e-13	2 View	11	AGR2				8.8e-18	2 View
12	TAPT1				5.1e-13	2 View	12	IMPA2				1.2e-17	2 View

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 gene category 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 “Gene category” at the clustering section select “10x10” at “numbers of draw” and click “next”.

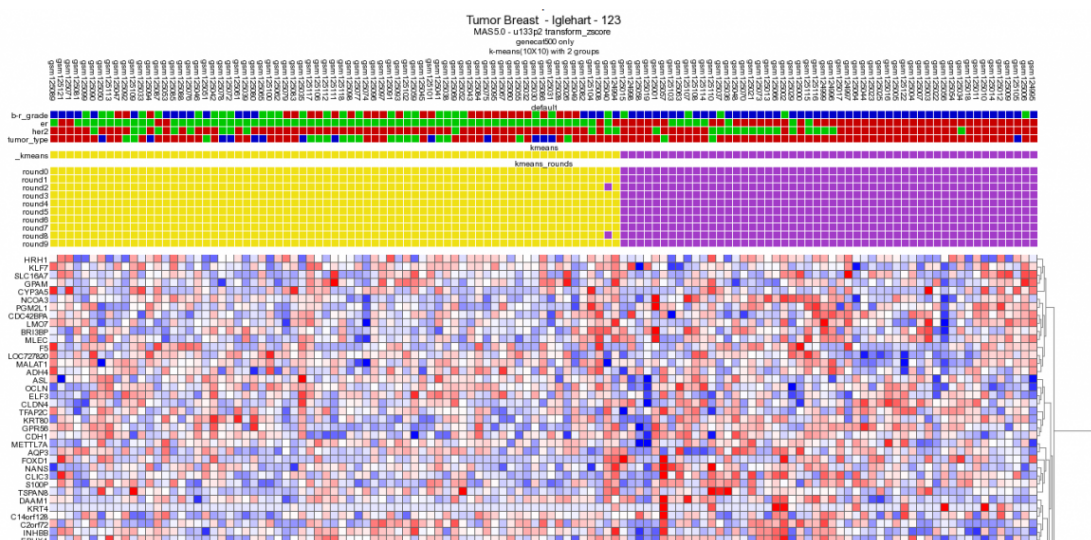


Figure 13: 10x10 Heatmap with the same dataset and gene category as depicted in Figure 9.

Performing a 10x10 K-means clustering via de 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 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.

Using genesets and creating heatmaps in R2

Or how you can generate 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 Toplist 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’.

The figure shows a four-step wizard interface for selecting a dataset and analysis type. The steps are numbered 1 through 4. Step 1 is titled 'Choose single or multiple dataset analysis' and has a dropdown menu set to 'Single Dataset'. Step 2 is titled 'Select a dataset for analysis' and has a dropdown menu set to 'Tumor Neuroblastoma public - Versteeg - 88 - MAS5.0 - u133p2'. Step 3 is titled 'Select type of analysis' and has a dropdown menu set to 'View Geneset (Heatmap)'. Step 4 is titled 'Proceed' and has 'Next' and 'Reset' buttons.

2,754,895 (2,505,872 unique) samples available

1 Choose single or multiple dataset analysis

Single Dataset

2 Select a dataset for analysis

Tumor Neuroblastoma public - Versteeg - 88 - MAS5.0 - u133p2

3 Select type of analysis

View Geneset (Heatmap)

4 Proceed

Next Reset

Figure 1: Select: View a Geneset

2. In the Adjustable settings many choices are available to customize the way the GeneSet data will be presented (Figure 2). With the geneset selection method you can indicate whether you want to pick one of the hundreds of genesets R2 harbors; public or your own genesets. Selecting the manual ordered list enables you

also to manipulate the ordering of the genes on the heatmap. In the adjustable settings many settings can be altered, such as filtering for samples, the sample order and many types of graphical adaptations.

Figure 2: A set can be selected, filtered for subsets and the clustering results will be ordered according to the selection

- 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 will show first what this ordering by clustering looks like. Suppose we want to find a geneset containing Cell Cycle genes. Click on **“Select a geneset”**, Type Cell Cycle in the ‘Search a geneset’ box.

Figure 3: Selecting a geneset

- In the grid box all genesets containing the words Cell Cycle (**Cell_cycle**) in their description are shown. Choose the (KEGG) Cell Cycle collection of Cellular Processes or another category where you find cell cycle. Note: sometimes such as in case of KEGG pathways ‘_’ could be in the naming and click the magnification glass.

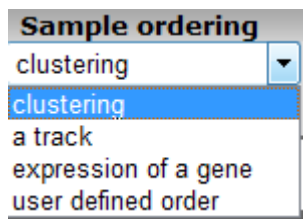


Figure 4: Available ordering domains for samples

5. 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 'User Options' menu item.

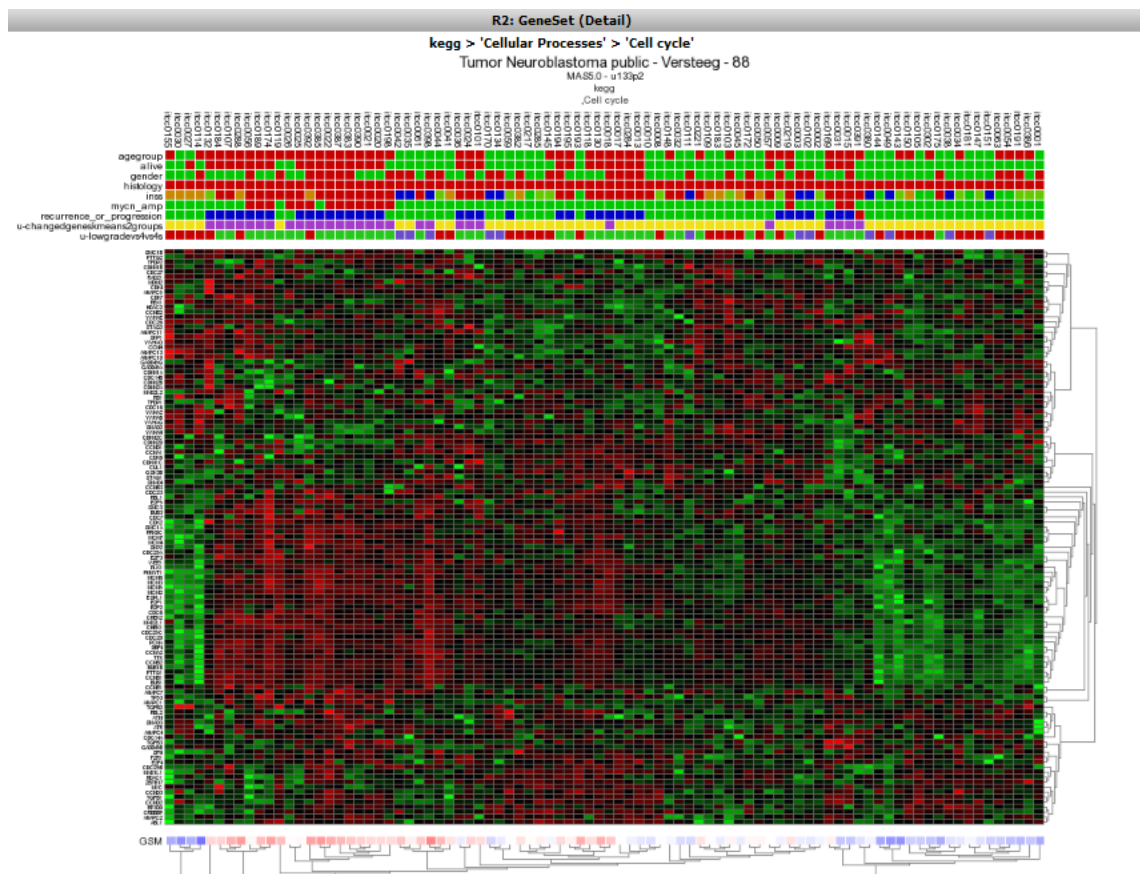


Figure 5: Heatmap view of the Kegg Cell Cycle geneset for the Neuroblastoma 88 dataset; genes and samples are sorted according to the clustering.

14.3 Step 2: Using multiple GeneSets

1. R2 also allows for multiple genesets 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 geneset but look for multiple sets at once.
2. In the next screen click "select geneset", Select 'Cellular Processes' in the KEGG pathway collection of look via keywords or going through all the sub-collections, select the apoptosis path way and click the "Confirm Collection" button.

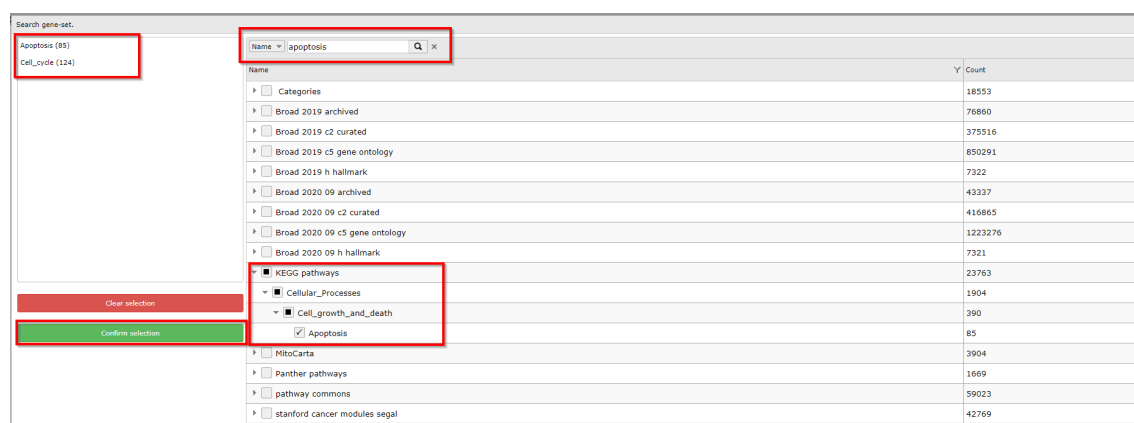


Figure 6: Selection a subcollection from the large collection of pathways

- The resulting heatmap (Figure 7) 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; Using both genesets shows clearly some blocks in the heatmap which could be linked to the annotation plotted above the heatmap

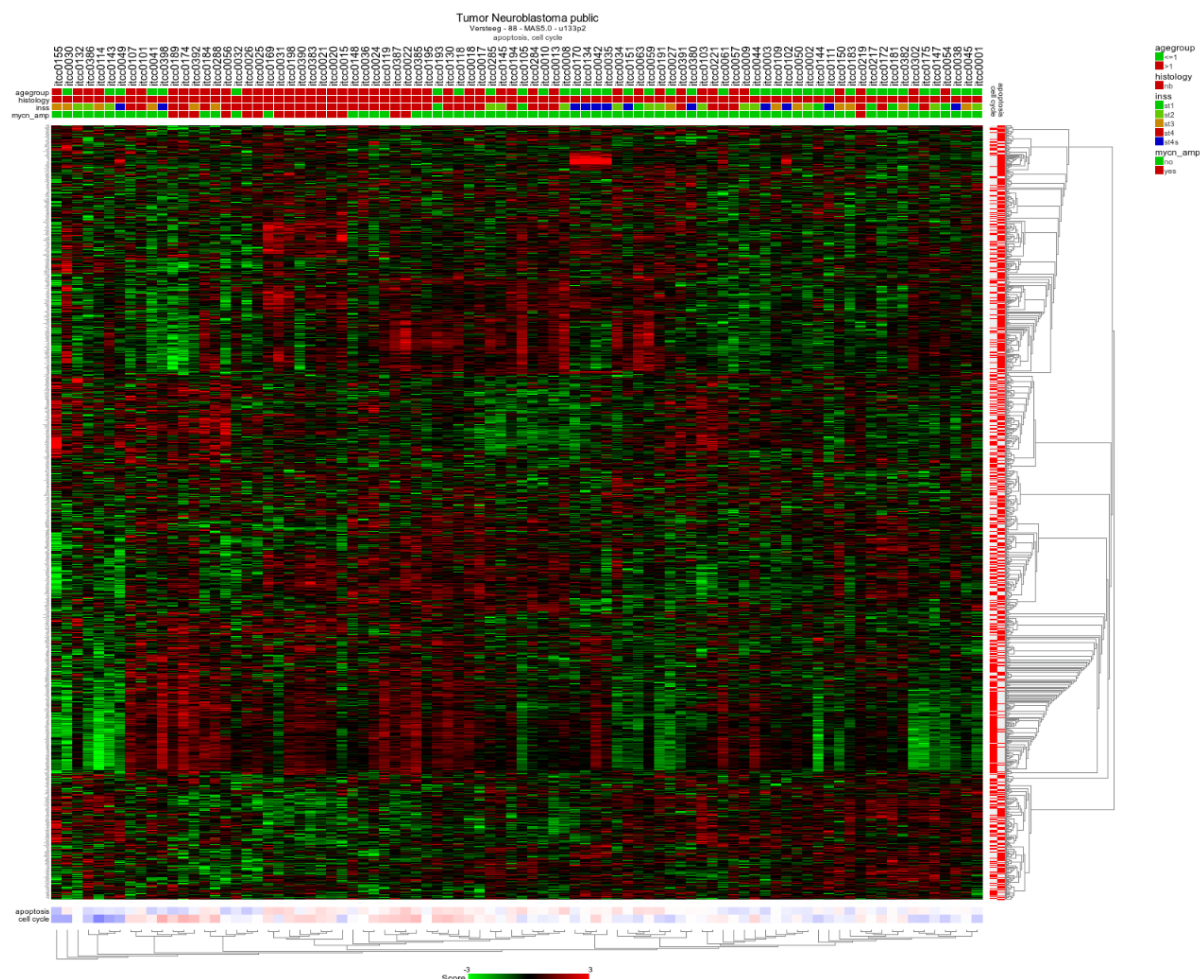


Figure 7: Heatmap view of the Cell Cycle and Apoptosis genesets for the Neuroblastoma 88 dataset.

14.4 Step 3: Relating genesets with data annotation

- 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 some genesets of

interest and choose order samples by track and choose the INSS category. ‘Order samples by a track’ and click “next” (Figure 10). I

Adjustable settings

Genes

Gene selection method: Gene set ▼

Gene set: cell cycle (537) ▲
Apoptosis (85) ▼

HugoOnce mode: yes ▼ ⓘ

Transformation: zscore ▼ ⓘ

Samples

Subset track: ▼ ⓘ

Selected sample subset: None

Order samples by: a track ▼

Track: inss (5 cat) ▼ ⓘ

subgroup gapsize: 4

Heatmap Options

Vector (SVG) output: false ▼ ⓘ

Cell width: auto ▼

Cell height: auto ▼

cell-border (grey): on ▼ ⓘ

Max range color scale: 3

Color scheme: green-black-red ▼

Distance measure: Euclidean distance ▼

Label track: samplenames (88 cat) ▼ ⓘ

Plot signature scores: yes ▼ ⓘ

Show heatmap: yes ▼ ⓘ

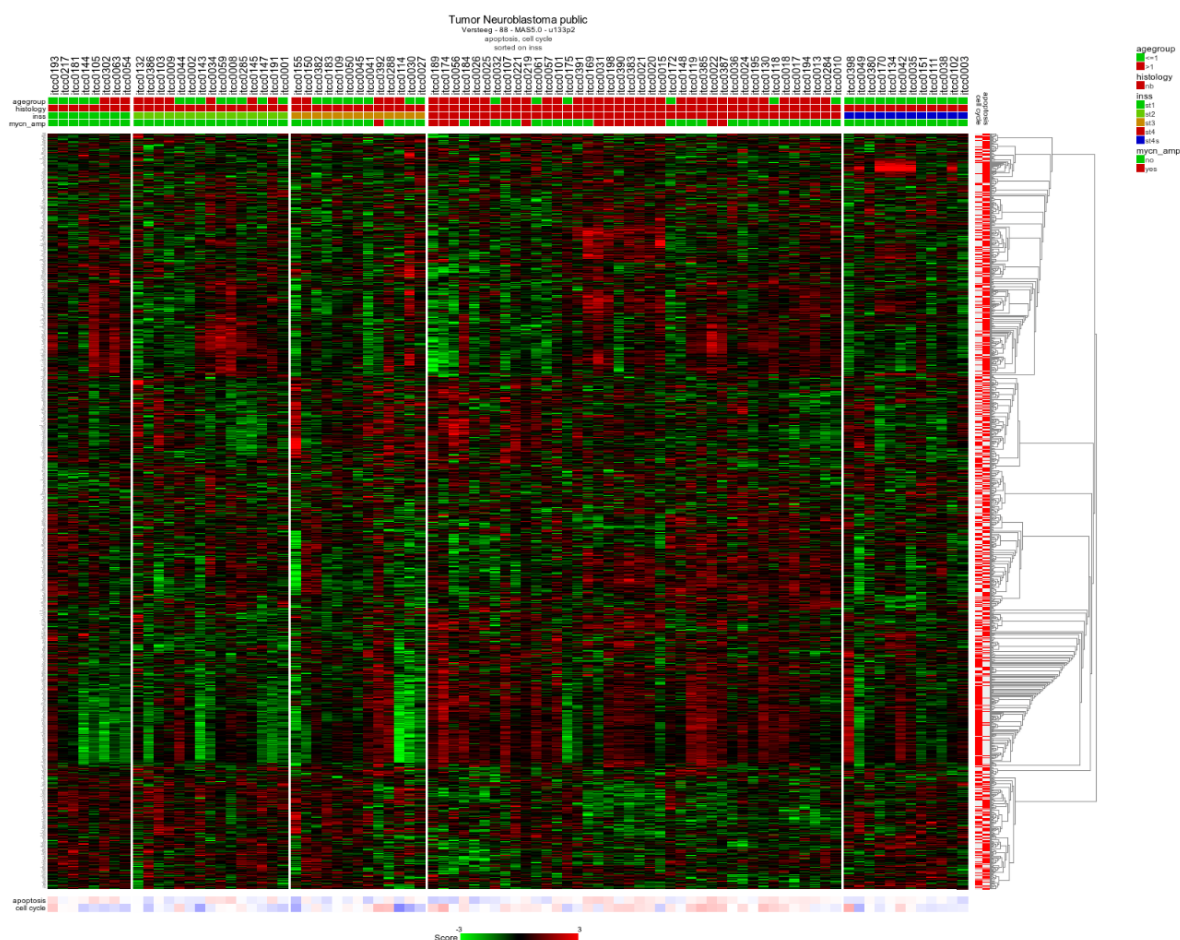
Track Display Selection

Select tracks

Submit

Figure 8: Selecting the track (subgroup) to order samples by: Order by a track

2. In the resulting heatmap it is clear that there is a segment of genes pathways cluster that is consistently upregulated in the stage 4 Neuroblastoma samples (in red in the INSS track).



Figure

9: 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

1. 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 manual order list in the gene selection method and click submit. Here you can paste a list of genes sorted according to your needs.
2. If you want to perform hierarchical clustering with a fixed sample order. Select current of manual order and click submit.

14.5 Step 4: Unsupervised hierarchical clustering with a geneset

It could be that for a given dataset there is no annotation to apply the analysis tools R2 is offering or you're lacking a starting point to further investigate a dataset. A good starting point could be 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, either select "Toplister" from the dropdown menu in box 3 of the main page or go to the menu on the left and select tools> small tools > Toplister. Make sure Data Type is set to Expression data and select Mixed Esophageal Carcinoma - tcga - 174 - tpm - gencode36 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.

Mixed Esophageal Carcinoma - tcga - 174 - tpm - gencode36 public ⓘ

The screenshot shows the 'Adjustable settings' window for the 'Mixed Esophageal Carcinoma - tcga - 174 - tpm - gencode36' dataset. The settings are as follows:

- Which set:** Standard Deviation (SD) ⓘ
- Modus:** normal ⓘ
- How many genes:** 100
- Floor value:** 0 ⓘ
- Transformation:** Log2 ⓘ
- Sample Filter:**
 - Subset track:** (empty dropdown) ⓘ
 - Selected sample subset:** None
- Gene Filters:**
 - HugoOnce mode:** yes ⓘ
 - Min. # Present calls:** 1 ⓘ
 - Minimal maximum value:** (empty) ⓘ
 - Minimal range size:** 0 ⓘ
 - Chromosome:** All ⓘ
 - Gene ontology:** All ⓘ
 - Gene set:** Select gene set ⓘ
 - Platform specific:** (empty) ⓘ
 - Manual list:** none ⓘ

The 'Gene Filters' section includes a 'Search GO' button and a list of gene categories with checkboxes:

- ☐ gencode
 - ☐ biotype
 - ☐ IG_C_gene
 - ☐ IG_C_pseudogene
 - ☐ IG_D_gene
 - ☐ IG_J_gene
 - ☐ IG_J_pseudogene
 - ☐ IG_pseudogene
 - ☐ IG_V_gene
 - ☐ IG_V_pseudogene
 - ☐ lncRNA
 - ☐ miRNA
 - ☐ misc_RNA
 - ☐ Mt_rRNA
 - ☐ Mt_tRNA
 - ☐ polymorphic_pseudogene
 - ☐ processed_pseudogene
 - ☐ protein_coding
 - ☐ pseudogene
 - ☐ ribozyme

At the bottom, there are 'Next' and 'Reset' buttons.

Figure 10: Selecting the track (subgroup) to order samples by: Order by a track

In this case the selected TCGA dataset gene annotation has been assigned to gencode reporters and in this example R2 also has stored gene category information which can be used as a filter. Note that this doesn't hold for all the platforms linked to a dataset.

1. R2 has generated a list of 100 genes showing the highest variation in gene expression. Beneath the list all kinds of settings can be adapted and in the right menu the genelist can be exported or stored in R2 as a geneset.

Mixed Esophageal Carcinoma - tcga - 174 - tpm - gencode36 public ⓘ
 Top 100 standard_deviation normal
 , transform_log2, present>=1

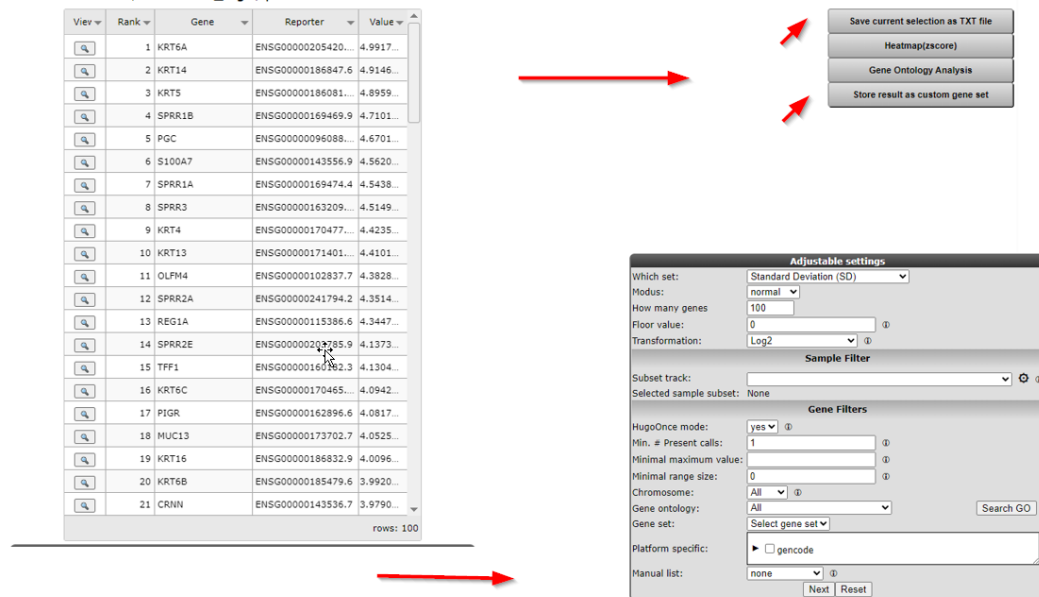


Figure 11: Result of the toplister module.

Clicking on the Heatmap (Zscore) in the right menu will perform an unsupervised hierarchical clustering and plots a heatmap.

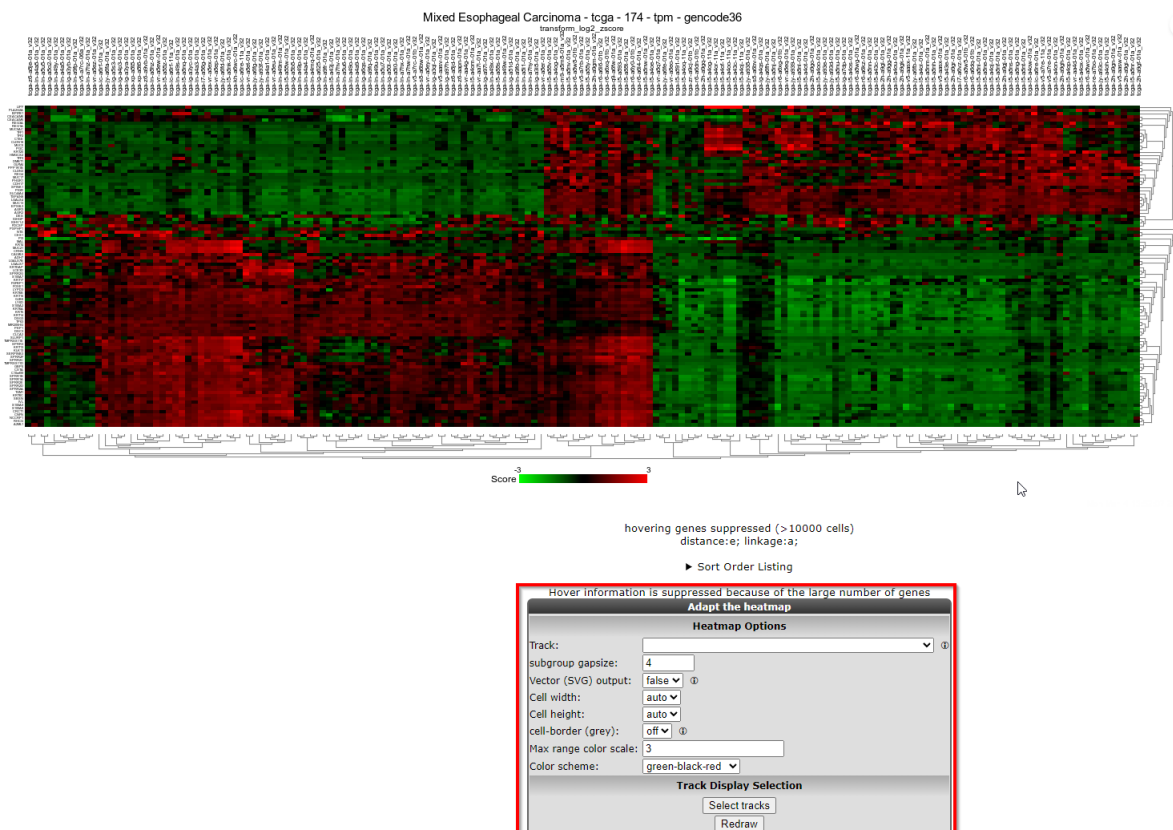


Figure 12: Unsupervised hierarchical clustering revealing subgroups in a Esophagal Carcinoma dataset.

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.

Principle Components Analysis in R2

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.

The screenshot displays a four-step guided interface for selecting data and analysis modules. At the top, it indicates '3,119,412 (2,870,389 unique) samples available'. Step 1, titled 'Choose single or multiple dataset analysis', shows 'Single Dataset' selected in a dropdown menu. Step 2, titled 'Select a dataset for analysis', shows 'Tumor Medulloblastoma PLoS One - Kool - 62 - MAS5.0 - u133p2' selected. Step 3, titled 'Select type of analysis', shows 'Principle Component Analysis (PCA)' selected. Step 4, titled 'Proceed', contains 'Next' and 'Reset' buttons.

Step	Title	Selection
1	Choose single or multiple dataset analysis	Single Dataset
2	Select a dataset for analysis	Tumor Medulloblastoma PLoS One - Kool - 62 - MAS5.0 - u133p2
3	Select type of analysis	Principle Component Analysis (PCA)
4	Proceed	Next, Reset

Figure 1: Selecting Principle Component Analysis

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 of the first 2 principle components. In the adjustable settings box, all the combinations principle components can be selected.
4. In the adjustable setting box select the all PCA-components option to view the several principle components combinations to investigate whether you can distinguish subgroups in your dataset.

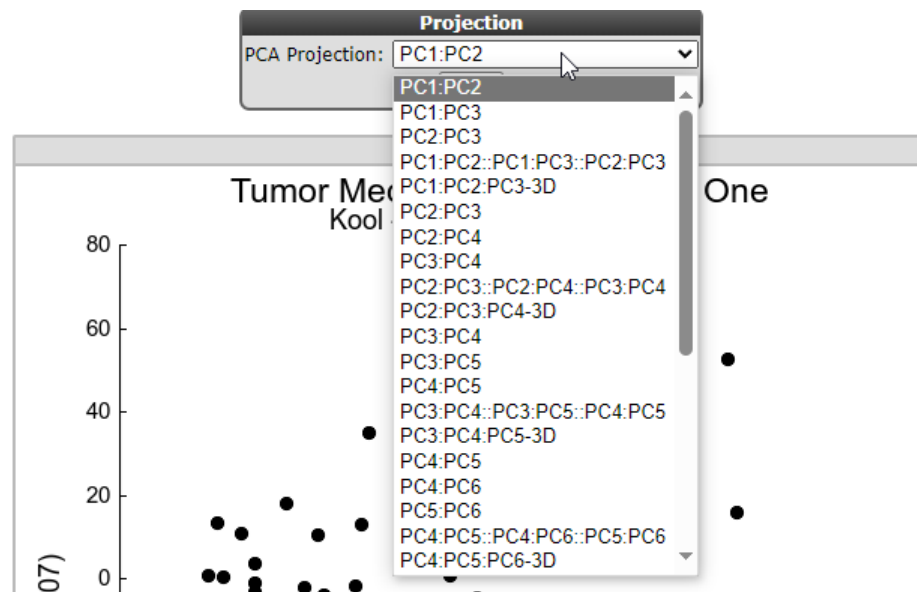


Figure 2: Adjusting PCA settings

Figure 3: Select Tracks

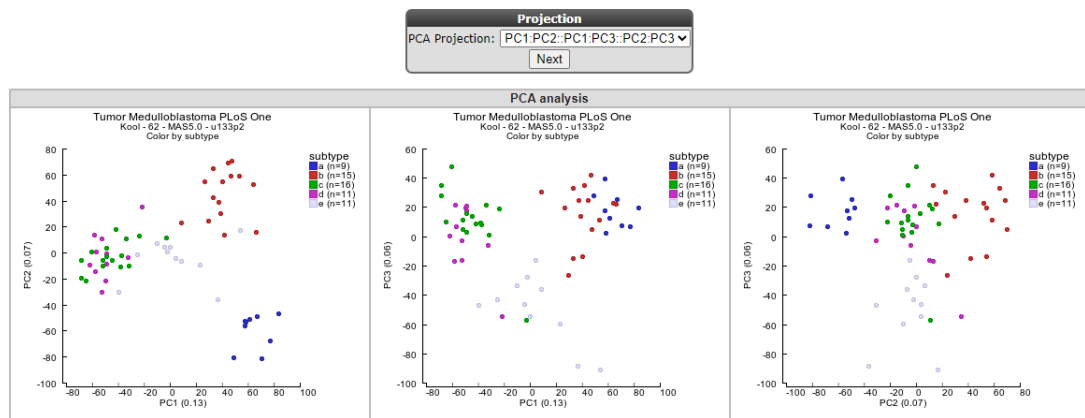


Figure 4: PCA components

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. Also note the variance reported on the axes.



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 covariance 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 reduction of dimensions is achieved 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.*

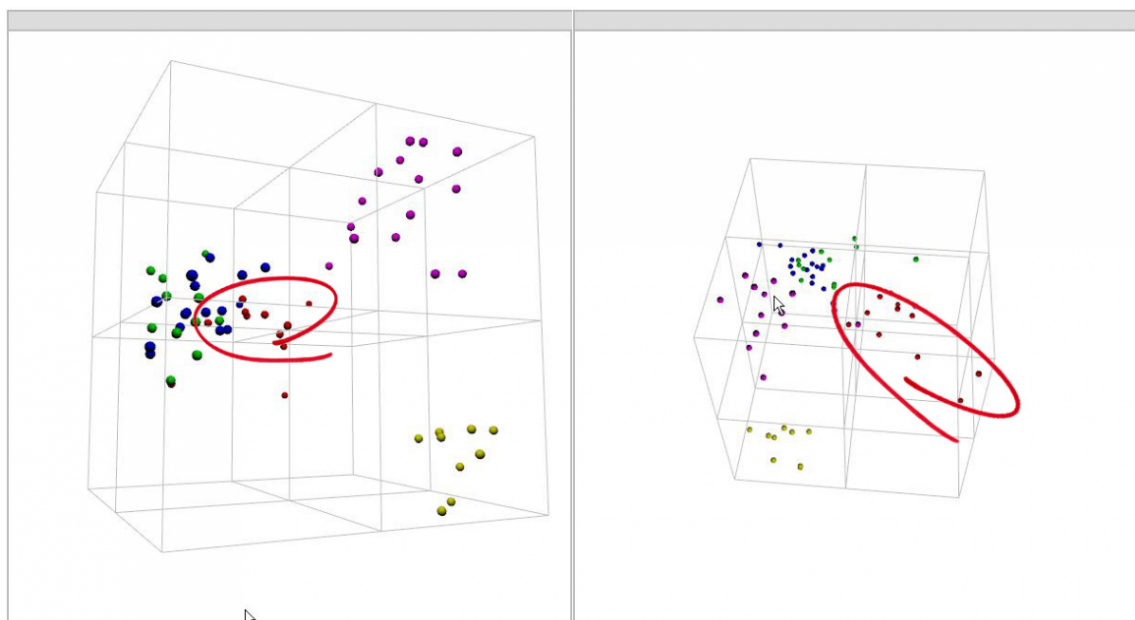
15.4 Step 3: Viewing clusters in 3D

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?

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.

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:

Kool M, Koster J, Bunt J, Hasselt NE, Lakeman A, van Sluis P, Troost D, Meeteren NS, Caron HN, Cloos J, Mrcic A, Ylstra B, Grajkowska W, Hartmann W, Pietsch T, Ellison D, Clifford SC, Versteeg R.; Integrated genomics identifies five medulloblastoma subtypes with distinct genetic profiles, pathway signatures and clinicopathological features. PLoS One. 2008 Aug 28;3(8):e3088.

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.

Sample maps: t-SNE / UMAP, high dimensionality reduction in R2

How to find groups in your dataset using t-SNE or UMAP dimensionality reduction algorithms.

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.

R2 offers several machine learning dimensionality reduction algorithms that are well suited for the reduction of high dimensional datasets to just 2 or 3 dimensions. Samples that have similar expression profiles in a dataset are located closely together on the 2D or 3D map, which enables the user to find clusters of similar samples. One such clustering method that is popular in biomedical research is the so called t-SNE algorithm. t-SNE stands for t-Distributed Stochastic Neighbor Embedding. Another algorithm with similar properties that is gaining more and more popularity is UMAP (Uniform Manifold Approximation and Projection).

Most researchers are already familiar with another dimensionality reduction algorithm, Principle Components Analysis (PCA), which is also available in R2 and is explained in more detail in the Principle Components Analysis tutorial. PCA, t-SNE and UMAP each reduce the dimension while maintaining the structure of high dimensional data, however, PCA can only capture linear structures. t-SNE and UMAP on the other hand, capture both linear and non-linear relations and preserve local similarities and distances in high dimensions while reducing the information to 2 dimensions (an XY plot). While t-SNE is able to preserve local relations, UMAP allows for a better preservation of dissimilarity of samples. Therefore, the distance between clusters of samples in a UMAP plot is more meaningful than in a t-SNE plot.

In the current section, we primarily focus on the t-SNE method, but in the R2 platform you will also encounter maps that have been created by UMAP. 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 datasets the number of perplexities will be less, in case of datasets with more than 1000 samples, only perplexity 50 is calculated. Which perplexity is the best, depends on the structure of the dataset, and is also depended on which display (how the samples are placed) allows for a better interpretation of your biological question. 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 of which the t-SNE has been completed. All users of R2 can explore generated t-SNE maps by coloring with tracks or expression values of a particular gene. Furthermore, the perplexity sweeps can be visualized. Via the ‘Sample maps’ option in the left menu, preprocessed t-SNE and UMAP maps can be viewed and analyzed with different perplexity and coloring settings. Users with collaborator access 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 Sample maps(UMAP/tSNE). You can see that a grid opens that displays the datasets available to you for which sample maps have been created. The headers of the grid show filtering options to search for the dataset that you are interested in. Here we want to search for the dataset ‘Cellline CCLE Cancer Cell Line Encyclopedia - Broad - 917 - MASS.0 - u133p2 ‘.
2. Type ‘CCLE’ in the textfield *Dataset Class*. Multiple sample maps have been generated from this same dataset. Choose the sample map that shows the date ‘2020-11-24’, de tSNE version in the column *Created* by a click on the **Select** button in front of the row.

Select	Info	Dataset Class	Dataset Author	Dataset Samples	Dataset Norm.	Dataset Platform	Map Type	Created	Favourite
		CCLE			Select Filter	Select Filter	Select Filter		
Select	ⓘ	Cell line CCLE Cancer Cell Line Encyclopedia	Broad	917	MASS.0	u133p2	t-SNE	2020-11-24	<input type="checkbox"/>
Select	ⓘ	Cell line CCLE Cancer Cell Line Encyclopedia	Broad	917	MASS.0	u133p2	UMAP	2021-01-08	<input type="checkbox"/>
Select	ⓘ	Cell line CCLE Cancer Cell Line Encyclopedia	Broad	917	MASS.0	u133p2	t-SNE	2017-03-28	<input type="checkbox"/>
Select	ⓘ	Cell line CCLE Cancer Cell Line Encyclopedia	Broad	917	MASS.0	u133p2		2021-02-11	<input type="checkbox"/>
Select	ⓘ	Cell line CCLE gene effects	Broad	1086	custom	depmapgid		2022-09-22	<input type="checkbox"/>
Select	ⓘ	Cell line CCLE Cancer Cell Line Encyclopedia 21q4	Broad	1389	tpm	gencode19a		2022-02-19	<input type="checkbox"/>
Select	ⓘ	Cell line Colon cancer CCLE (CRC)	Broad	69	tpm	gencode19a		2022-02-26	<input type="checkbox"/>
Select	ⓘ	Cell line CCLE Cancer Cell Line Encyclopedia 23q4	Broad Depmap	1495	tpm	ensh38e104	UMAP	2023-12-20	<input type="checkbox"/>
Select	ⓘ	Cell line CCLE Cancer Cell Line Encyclopedia 21q4	Broad	1389	tpm	gencode19a	UMAP	2021-12-16	<input type="checkbox"/>

Figure 1: Select a preprocessed sample map (e.g. t-SNE map) from the grid

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 settings’ from the ‘colormode’ and choose ‘primary site’. Press ‘Set Colors’ to redraw the image.

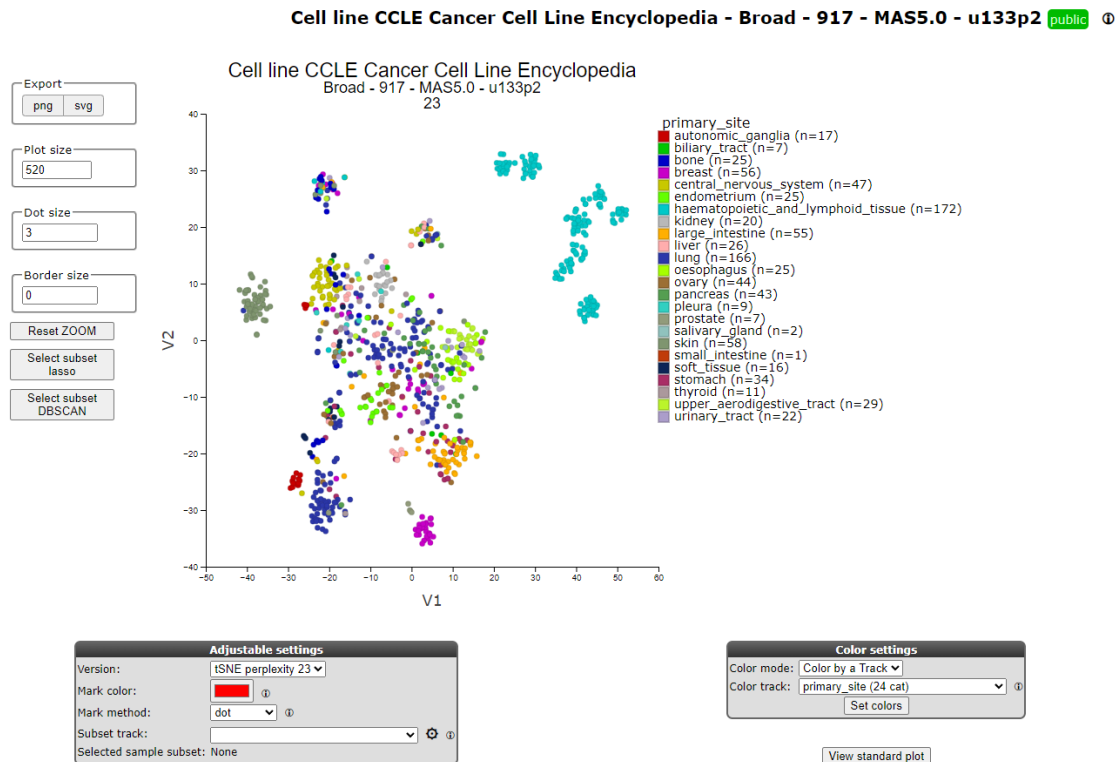


Figure 2: t-SNE preprocessed t-SNE maps

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 *Color mode* to ‘color by gene’.

1. In the ‘Color settings’ box select ‘Color by Gene’ under *Color mode* and subsequently type ‘CLDN3’ in the textfield of *Gene / Reporter*. The corresponding reporter will automatically pop-up (Figure 3), click on it to confirm your choice. The gene selection box autocompletes the proper reporter probeset, but this can take a little bit of time before the gene selection box appears.

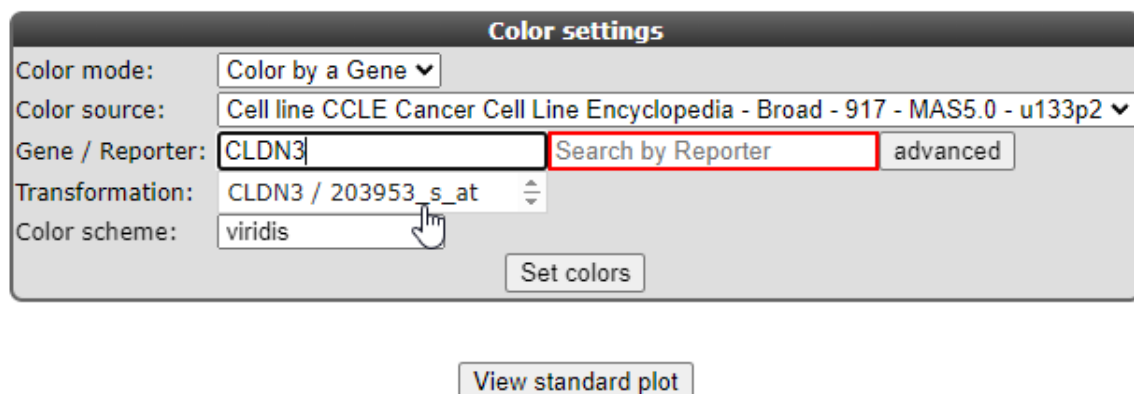


Figure 3a: t-SNE_Color by Gene

1. Again click ‘set colors’ 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 subgroups of the carcinoma samples which have higher level in contrast to the (other) samples.

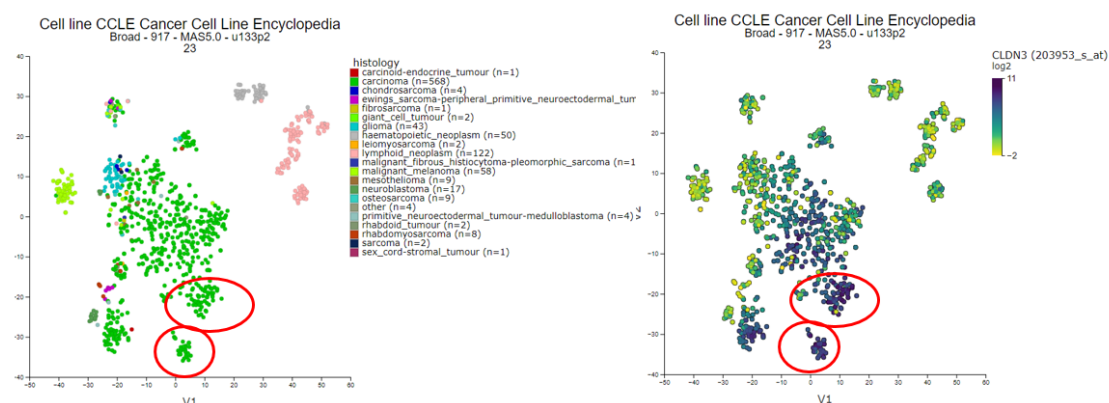


Figure 4a: t-SNE_Color by Gene

1. It could be that you already noticed that the maps are plotted in an interactive fashion allowing you to adapt the graph on the fly. When adapting all kinds of settings such as marking samples, change the dot size, toggle subgroups on and off and much more. As shown in figure 4b.

Figure 4b: Interactively changing the layout

1. Below the color settings box you can also use the standard plot module which is static. 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 (adeno)carcinomas. Another gene which emphasizes the observation in the previous example is the `NR3C1` gene showing an inverse gradient pattern for this subgroup. In this picture below, we also adapted the setting *Color scheme* to 'Fireworks' in the 'Adjustable settings' box.

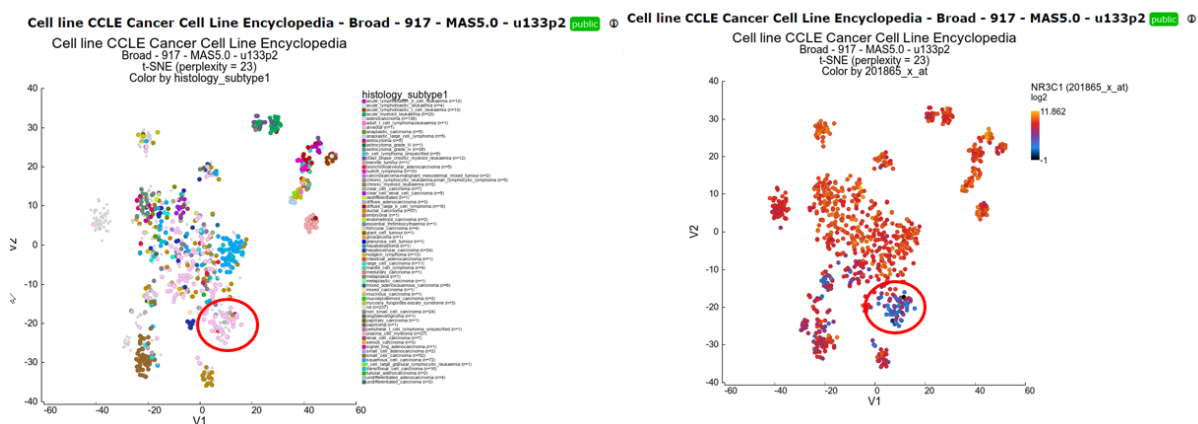


Figure 5: t-SNE_Color by Gene

Adjustable settings

Versions: tSNE perplexity 23 ▾
 Subset track: ▾ ⚙️ ⓘ
 Selected sample subset: None
 Image size: 500 ▾
 Dot size: 3 ⓘ
 Samples to mark: comma separated sample names ⓘ
 Mark method: dot ▾ ⓘ
 Samples paths: comma separated sample names
 Vector (SVG) output: false ▾ ⓘ
 Enable hovering: yes ▾ ⓘ
 x: min: max:
 y: min: max:
 Color mode: Color by a Gene ▾
 Color source: Cell line CCLE Cancer Cell Line Encyclopedia - Broad - 917 - MAS5.0 - u133p2 ▾
 Gene / Reporter: NR3C1 201865_x_at advanced
 Transformation: Log2 ▾ ⓘ
 Color scheme: fireworks2 ▾
 Submit

Figure 6: t-SNE select probeset 2

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 *Perplexity* to 'all' and *Color by Track* modus to eg. 'primary_site' in the 'Adjustable Settings' box.

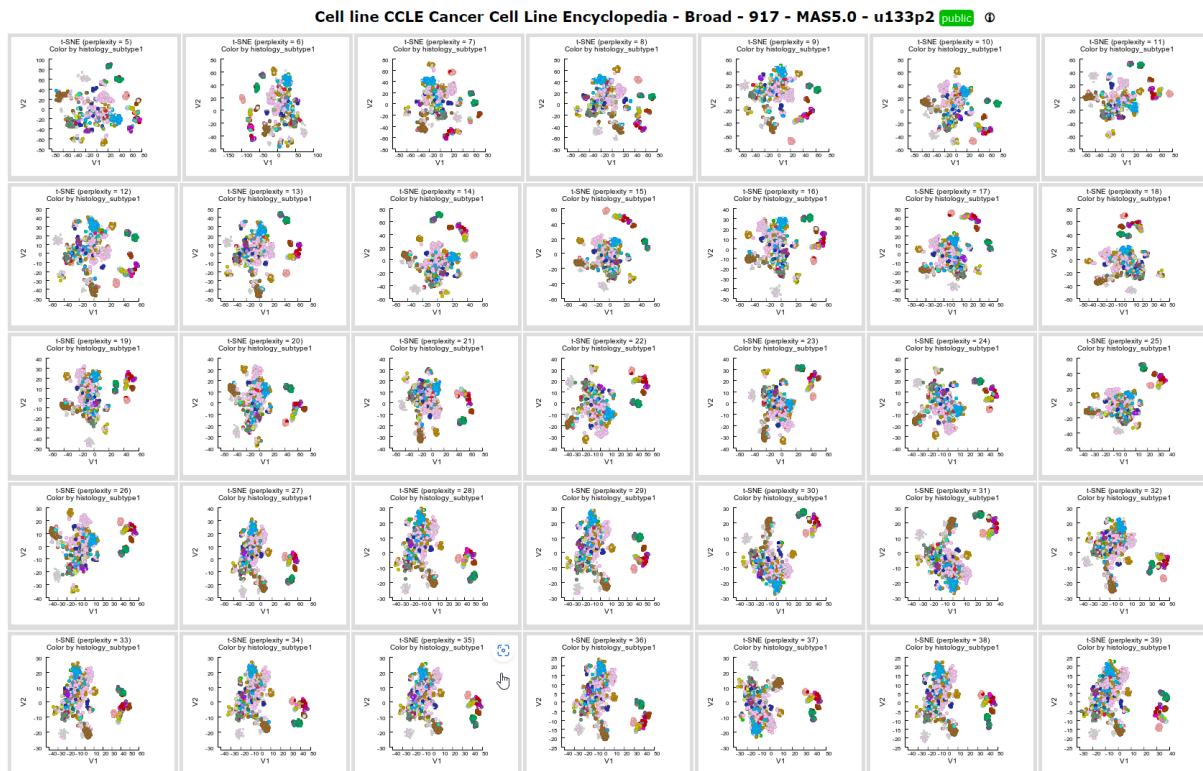


Figure 7: t-SNE: all perplexities

By choosing the perplexity value 'all', miniature tiles will be generated for all perplexities (5-50), where it is still possible to use the color by track mode. Also, you can simply click any of the tiles to generate the map of that particular perplexity in normal large format.

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 is 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 another nice example of an R2 generated t-SNE map: the large dataset of normal tissue expression profiles.

1. In main menu *Change Dataset* to the Normal Tissues GTEx v4 - GTEx - 2921 - RPKM - ensgetxv4 in box 2 (in the Change Dataset grid fill in the number '2921' in the textfield under the *N* column and click *Select*). Then select t-SNE in box 3. Click Next. If the 'default' map has already been calculated, a shortcut button will also appear as shown in figure 8. In the 'Adjustable settings' panel you can adjust several settings, such as sample filtering and specific gene sets

Normal Tissues GTEx v4 - GTEx - 2921 - RPKM - ensgtexv4 public ⓘ

ShortCut option:
Explore this dataset with Default settings

ShortCut

Plot the map default result
Reset

Adjustable settings

Gene Filter

Gene ontology: All Search GO

Gene set: Select gene set

Manual list: none ⓘ

Sample Filter

Subset track: ⚙️ ⓘ

Selected sample subset: None

next
Reset

Figure 8: t-SNE: Menu

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 >6000 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*.

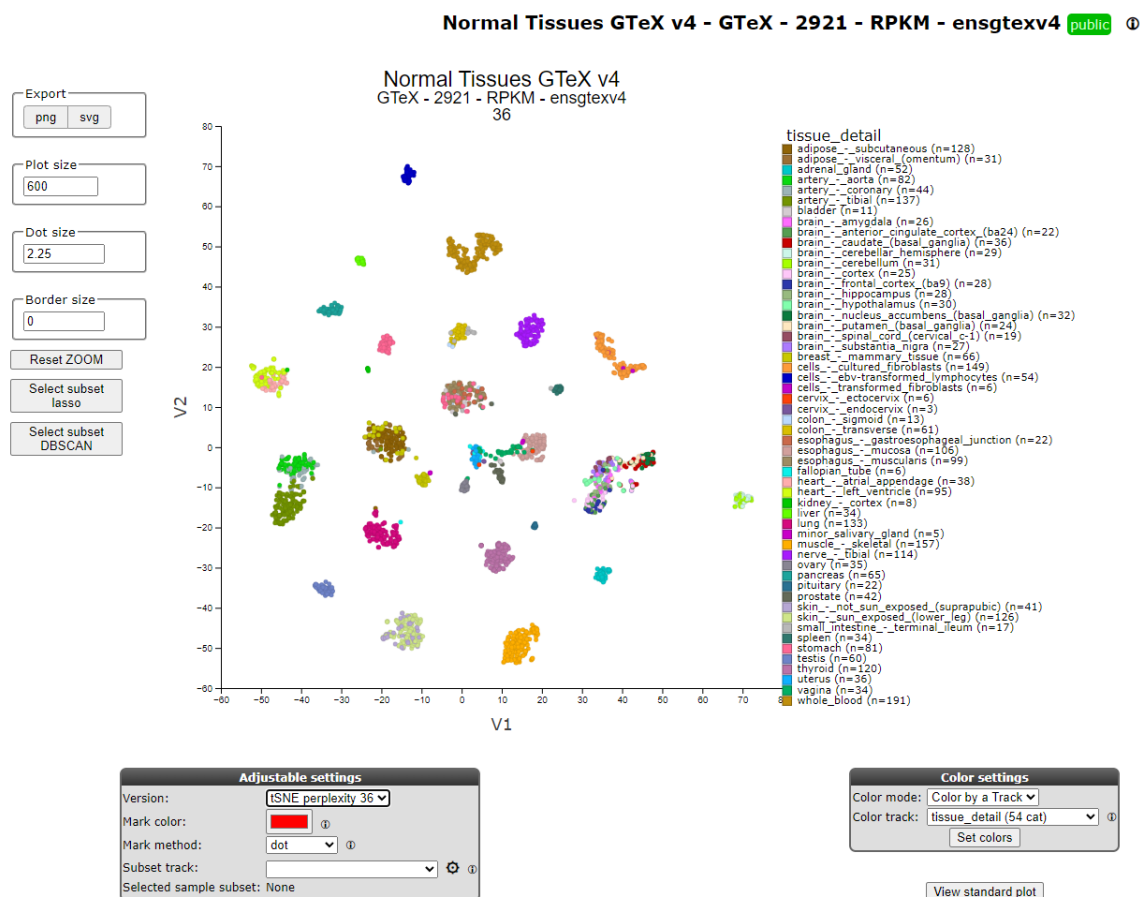


Figure 9: t-SNE: Colored by track

16.6 Step 5: Creating groups with the t-SNE lasso tool

Suppose the t-SNE algorithm produced some interesting sample 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 *Sample maps* and select 'Cellline CCLE Cancer Cell Line Encyclopedia - Broad - 917 - MAS5.0 - u133p2' with the date '2020-11-24' in the *Created* column. Plot the corresponding t-SNE map using perplexity 23 and color the maps by selecting 'primary_site' with *Color track* option. The haematopoietic group can clearly be subdivided in several groups which can be used to investigate these sub clusters in more detail.

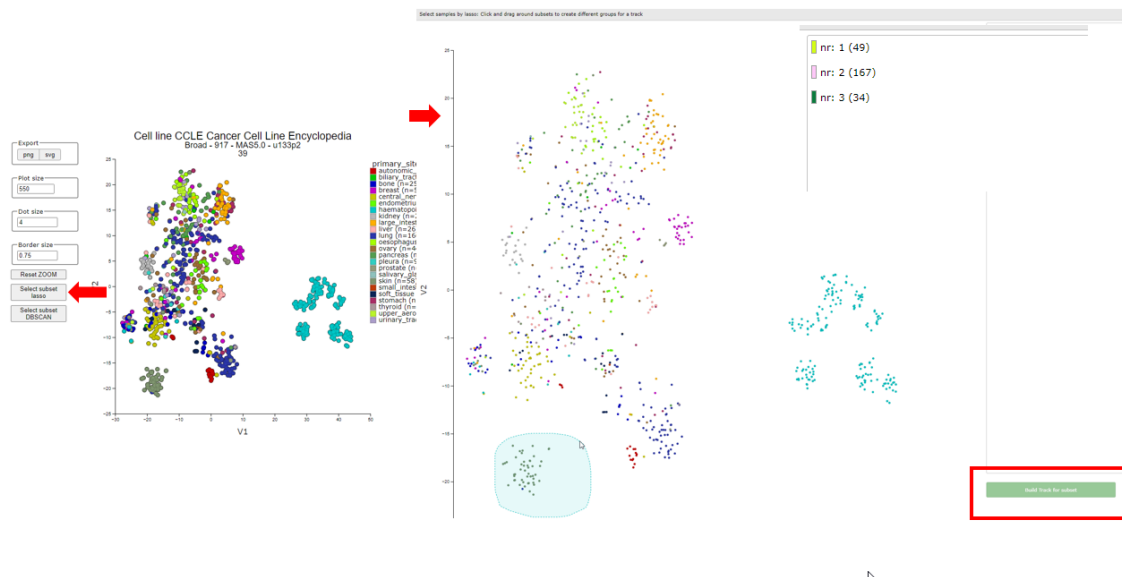


Figure 9: t-SNE: Colored by track

- Click on *Select subset* left from the interactive 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 select as one group. After releasing the mouse button the group with the amount of samples is listed on the right. The samples of the dataset are subsequently annotated with a group id for each lasso selection action. You can select groups up to a number of 10. After you finished the group selections, click below the groups on the “Build Tracks for subset” button.
- 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 (temporary tracks will be deleted after 24 hrs). 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.

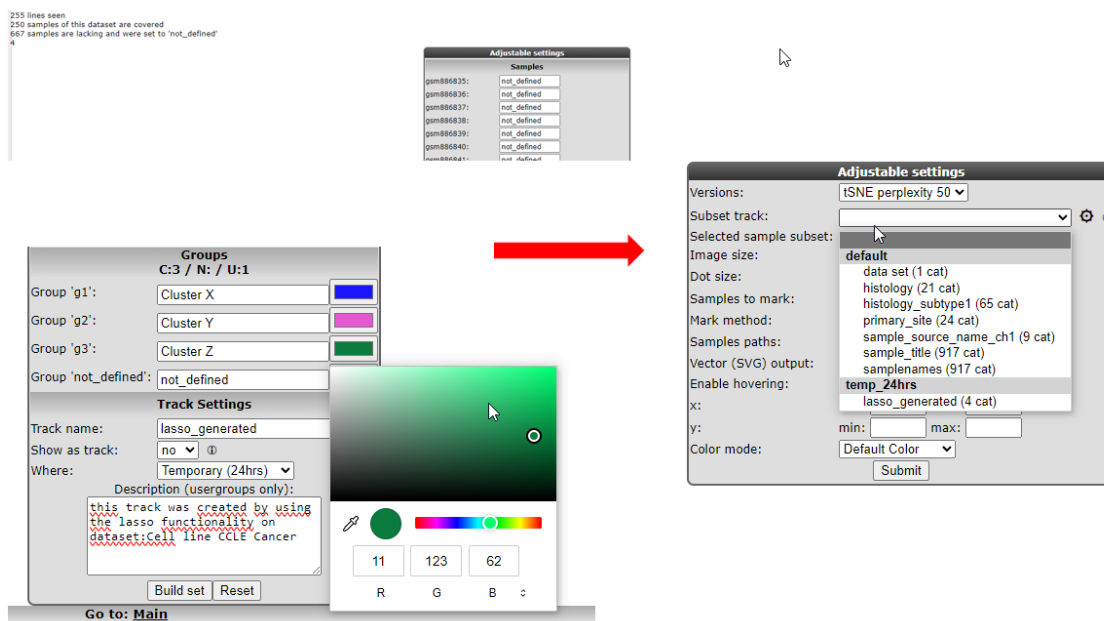


Figure 10: t-SNE: Using the lasso selection

The lasso selection tool is also available for the PCA module;

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 Pts* tool 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.

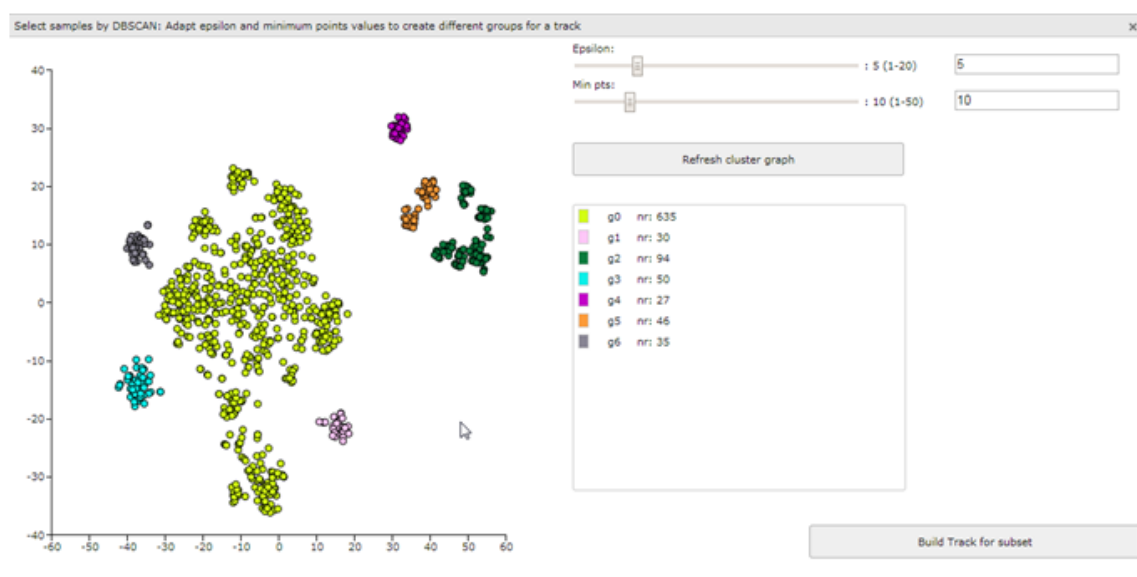


Figure 11: t-SNE: Using the DBSCAN selection

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 pts* 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.

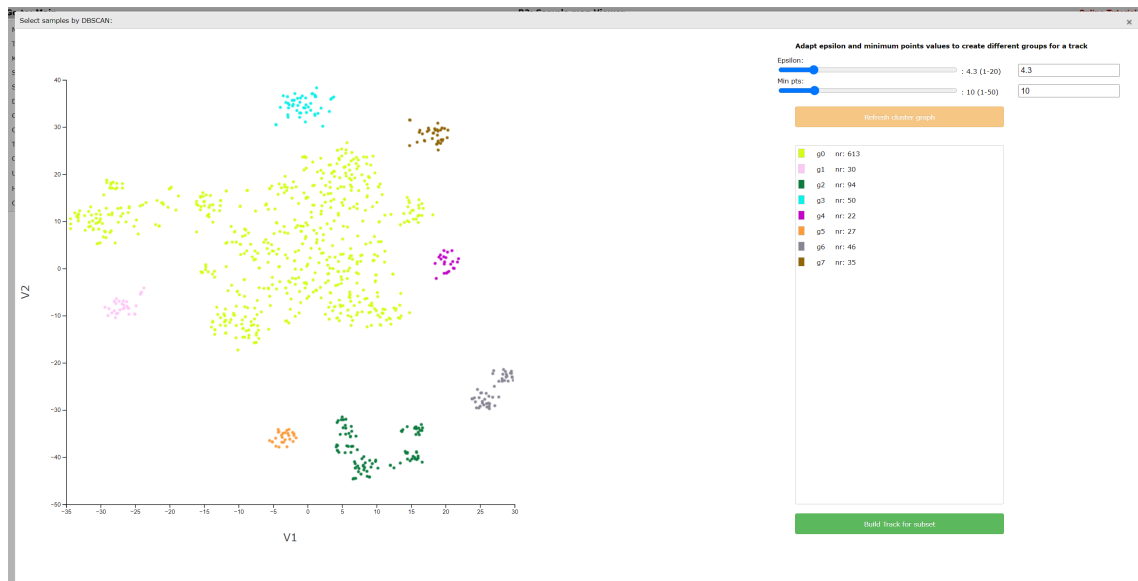


Figure 12: t-SNE: Using the DBSCAN selection

2. 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 available for both t-SNE maps, UMAPS and PCA plots.

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.

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 part of the 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 profiles 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. Note that quite a number of the hundreds of platforms that R2 manages will have genome information. In those cases, the ‘Probeset Genome Location’ box will list those and can bring you to our embedded genome browser by following the R2_Tview link.

ProbesetVerification (hg18)						
symbol	probeset	rank	gene overlap	exon overlap	probes found	Link
MYCN	209757_s_at	1 GS	YES	YES	YES	R2 TView
MYCN	242026_at	2 WS	NO	NO	NO	R2 TView
MYCN	209756_s_at	3 GS	YES	YES	YES	R2 TView
MYCN	211377_x_at	4 GS	YES	YES	YES	R2 TView
MYCN	234376_at	5 GS	YES	YES	NO	R2 TView

Probeset Genome Location:			
209757_s_at	chr2	15,999,496	16,004,577 + R2 TView

Figure 1: Probeset Verification table

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 than 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.

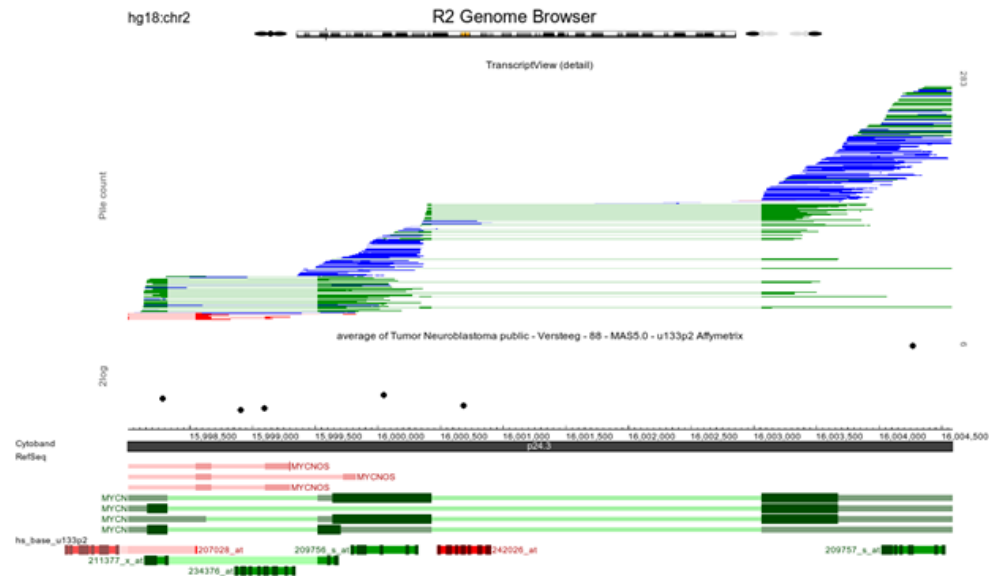


Figure 2 : Genome browser with default tracks.

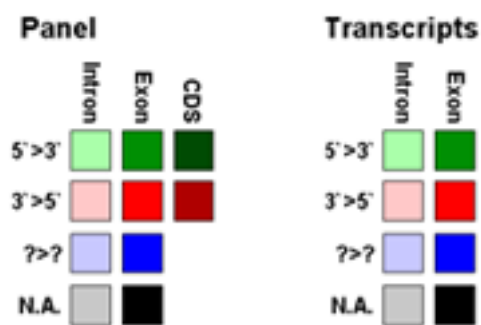


Figure 3: Legend of the color usage

With the default settings the genome browser shows the average expression signal per probeset **for** a chosen dataset **with** their genomic location.

1. 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.
2. 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.

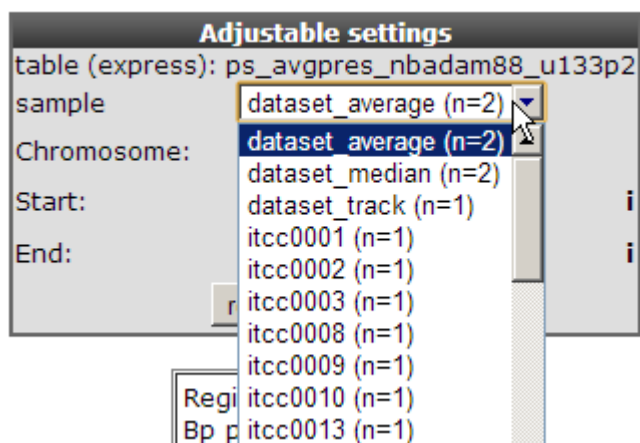


Figure 4: Adjustable settings panel.

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.

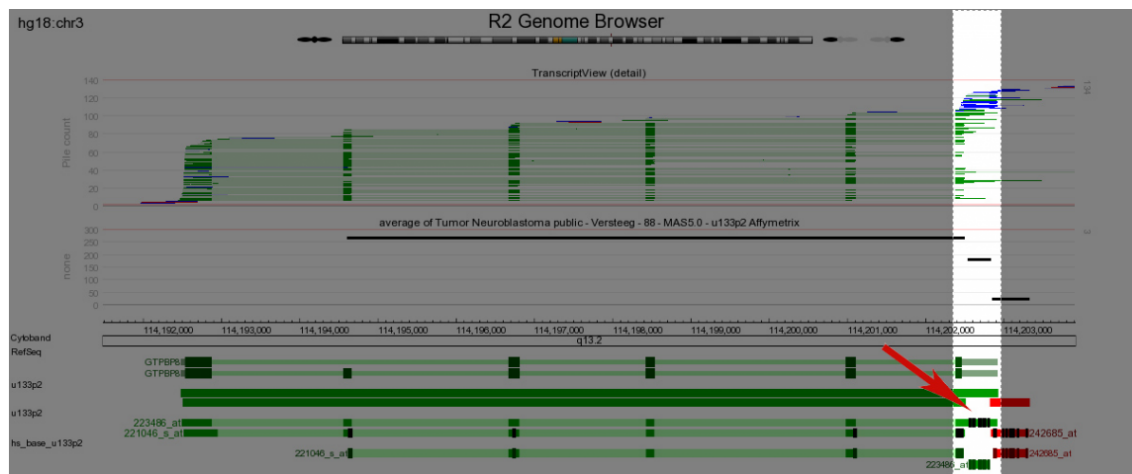


Figure 5: Zoom controls

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.

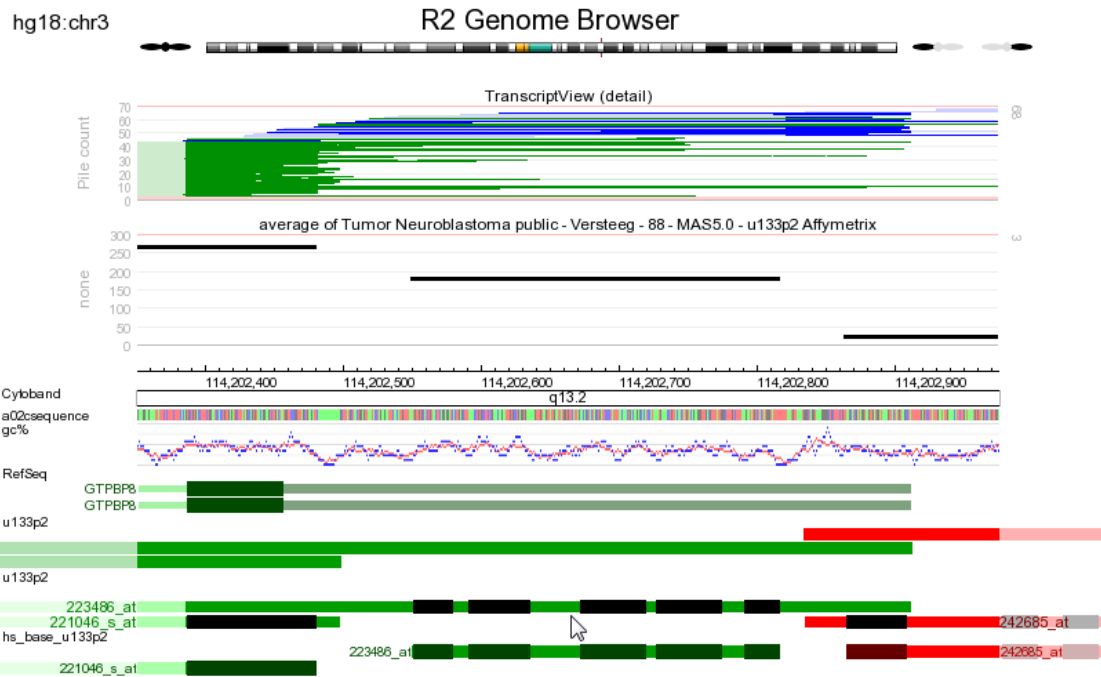


Figure 6: Zoom-in graph

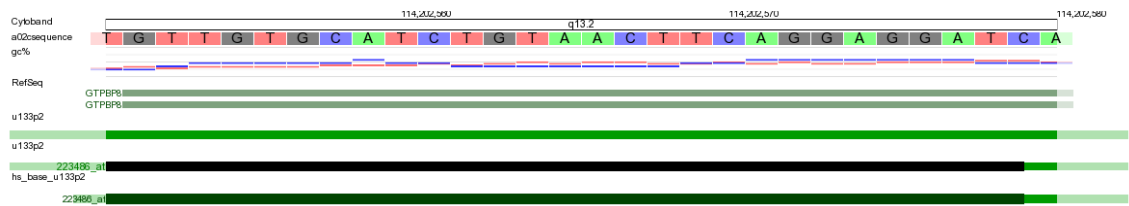


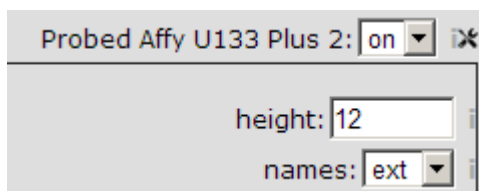
Figure 7: Zoom revealing basepair sequence

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.

- 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.

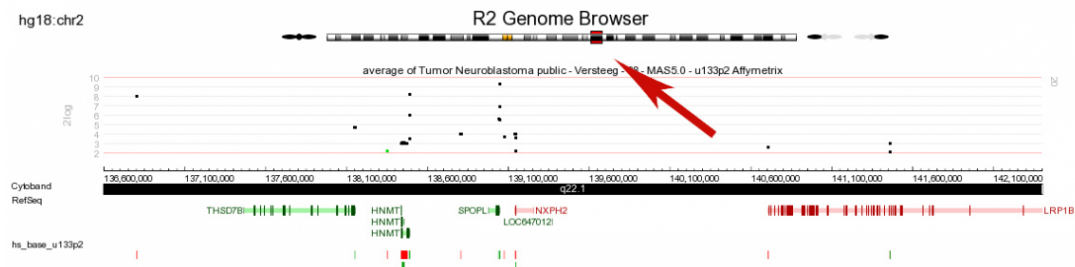


Figure 8: Chromosomal clicking

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.

Datascopees are typically landing web-pages that provide quick jumps to (dedicated) analyses and visualizations in R2. such datascopees are frequently associated with studies that share their data, presented from a manuscript in such a way that you can repeat their analyses, but with other selection criteria etc. Another type of datascope is the ability to limit data sets to a particular selection, such as a tumor entity.

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'. Alternatively you can click on the 'change data scopes' root item and get a tiled display of all the data scopes that you have access to with your account.



Figure 1: Selecting a Data Scope.

2. You may have noticed that the platform now includes 'neuroblastoma' in its title. Additionally, the number of samples available to you has greatly decreased. This is because you will now only be looking at datasets that we have annotated as neuroblastoma. This data scope can be useful if you only want to see datasets related to a specific entity that we have defined. If you want to select a new dataset, you will only be able

to choose from neuroblastoma assets. In analyses such as ‘2D distribution’, this restrictive data scope will also apply.

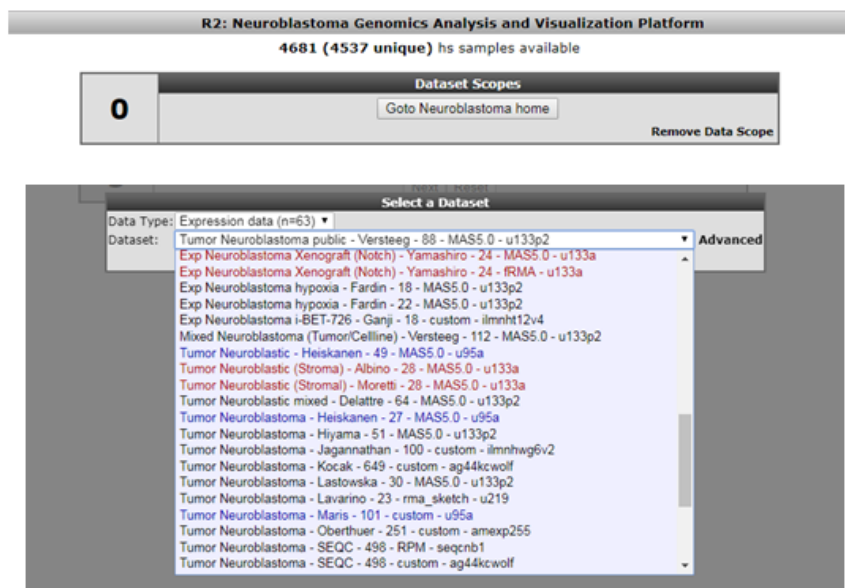


Figure 2: This Data Scope restricts the resources to Neuroblastoma only.

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. These landing pages are essentially ‘quick jumps’ into the tools of the R2 platform, combined with predefined settings. The possibilities offered by the platform are nearly endless (and we are willing to create things that may not be possible yet).
2. 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). Datascope can be reached in 2 ways. Either you select ‘medulloblastoma 500’ from the ‘paper’ group in the Data Scopes menu, or you simply click on the ‘datascope’ main menu item directly, and an overview of all the data scopes within the access of your account will become visible (and click on ‘medulloblastoma 500’).

Welcome to the data scopes index in R2

Data scopes are dedicated 'landing pages' from where predefined jumps into analyses are presented. These data scopes often are parts of projects / consortia, but can also define a focus for a particular tumor entity (the tumor scopes). Please click on one of the tiles available to your access profile to proceed to the respective landing page.

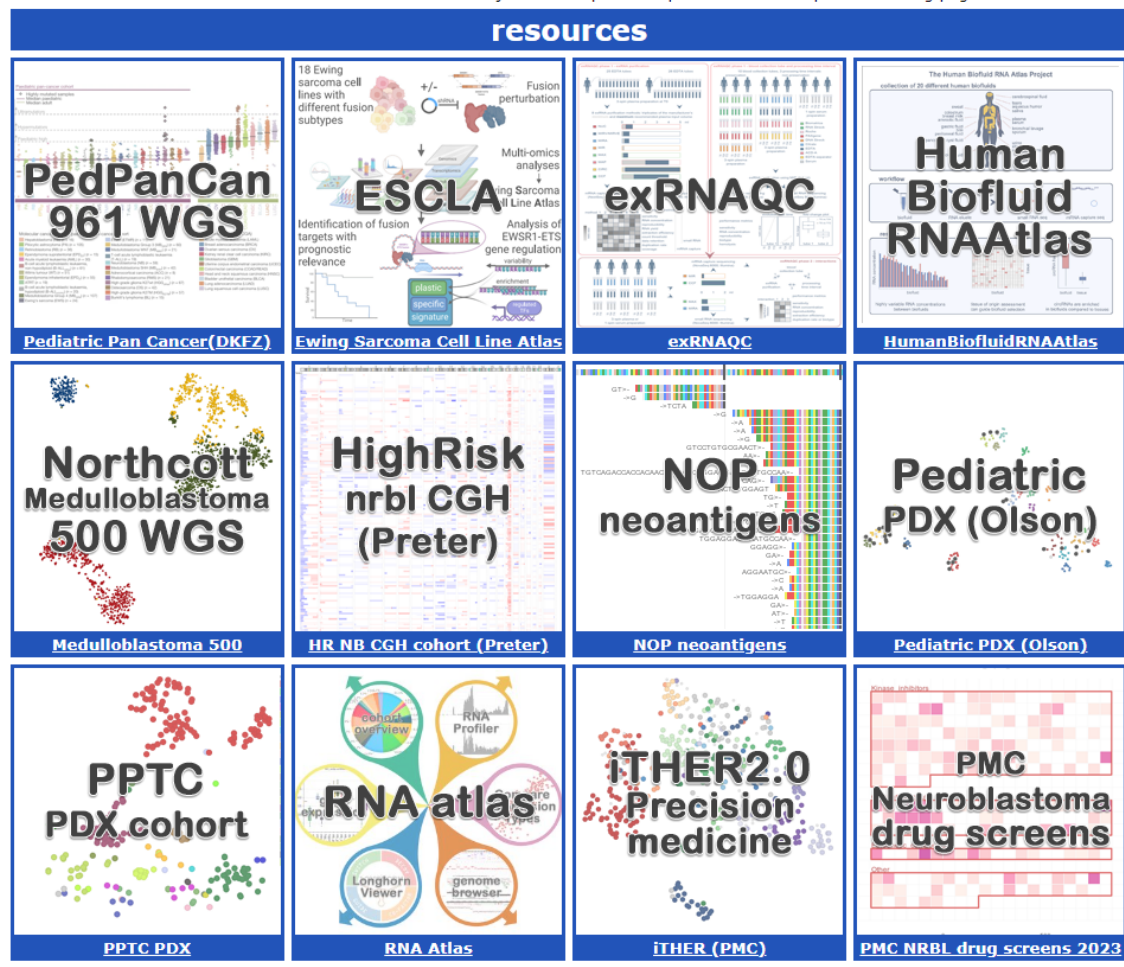


Figure 3: Datascope overview.

- 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.

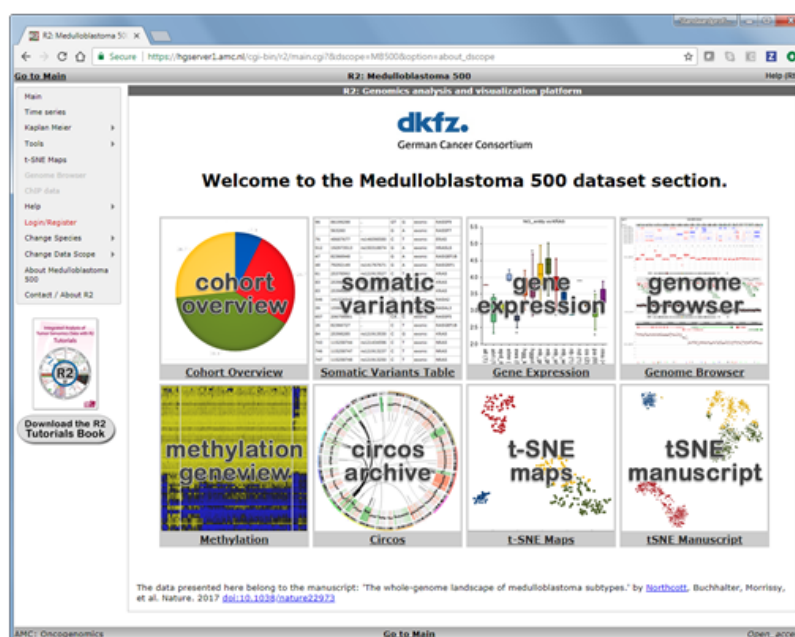


Figure 3: Landing Page for Medulloblastoma 500.

4. Within a landing page ‘tiles’ are defined that will quickly bring you to specific analyses and or visualizations in R2. 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. Another option available here is the ‘circos archive’, which enables you to dive into the events that are observed within a single patient. 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 social media channels where we make announcements on this as well.

[facebook]: <https://www.facebook.com/r2platform/> “FaceBook” [linkedin]: <https://www.linkedin.com/company/72569174/> “LinkedIn” [X(Twitter)]: https://twitter.com/r2_platform “X(Twitter)”

Instagram

Mastodon page

18.4 Final remarks / future directions

Some of the functionalities described in this tutorial have been developed recently. If you encounter any issues or difficulties, please do not hesitate to contact R2 support at r2-support@amsterdamumc.nl. Additionally, if you have ideas for new data scopes or would like to showcase your own data in a similar way, please feel free to get in touch with us through the same support address.

We hope that this tutorial has been helpful.

Best regards,

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.

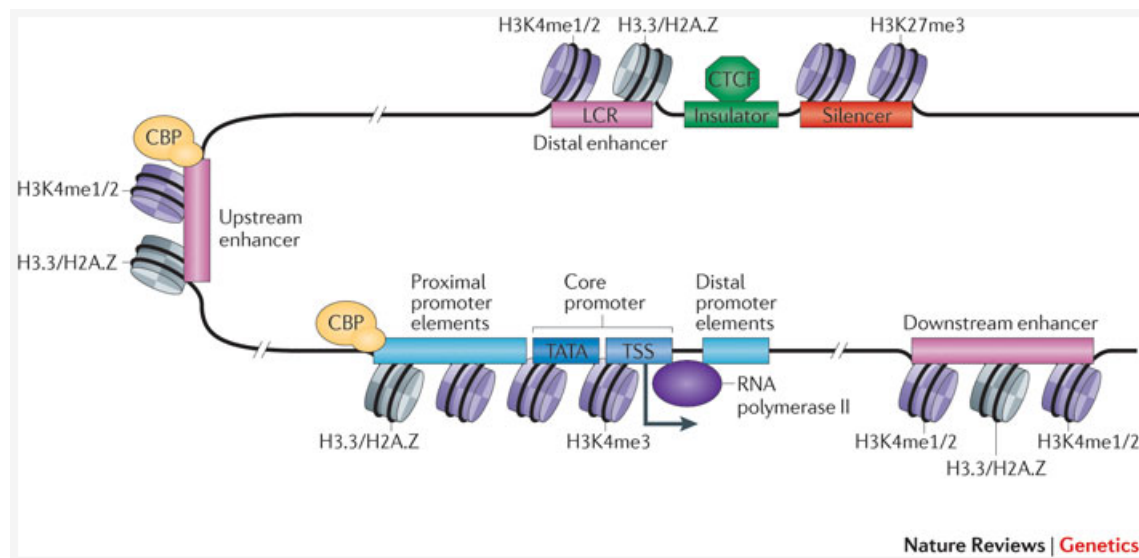
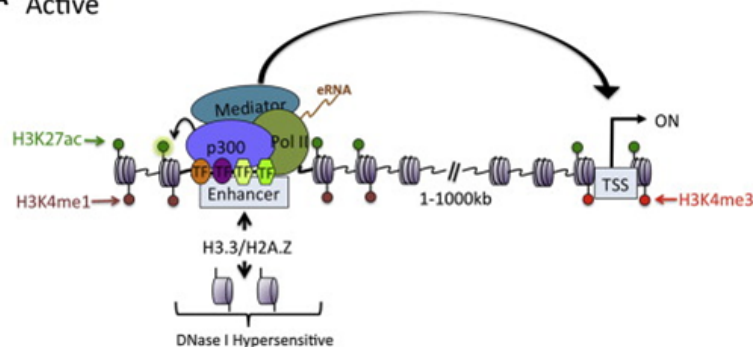


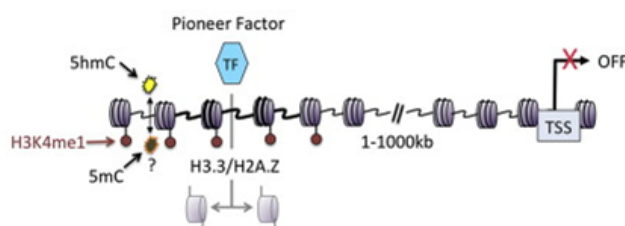
Figure 1: Transcription; taken from *Nature Reviews Genetics* 12, 283-293 (April 2011)

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.

A Active



B Primed



C Poised (mouse and human ESC)

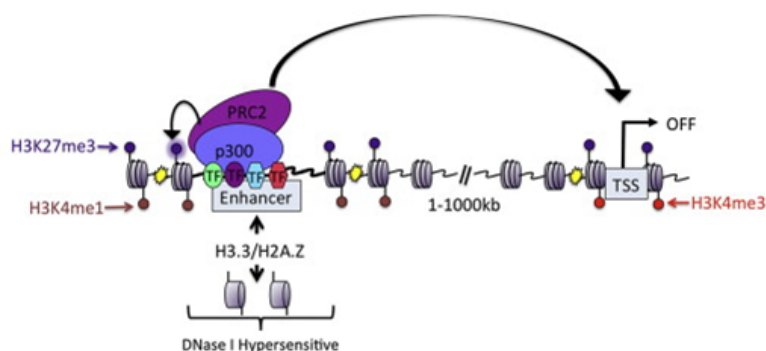


Figure 2: Specific chemical modifications mark specific states of cis-regulatory elements; taken from doi:10.1016/j.molcel.2013.01.038

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. Its 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 is 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).

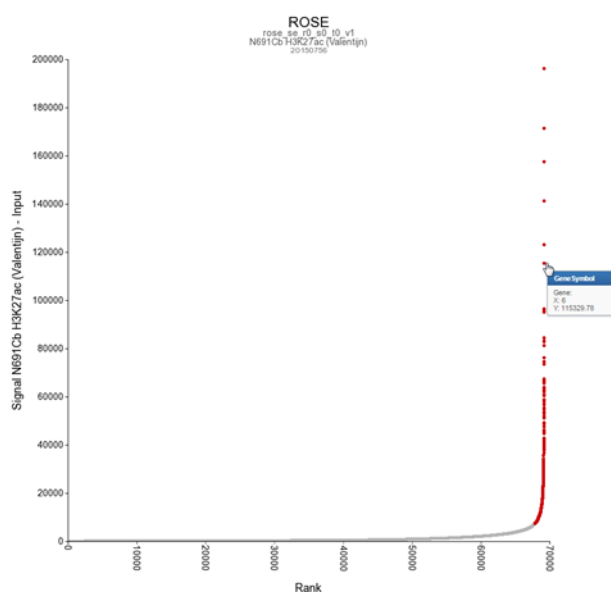


Figure 3: Result of a typical ROSE analysis. Above the inflection point, marked in red, are super-enhancer regions

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. For this example check if the correct dataset has been selected in this case "Tumor Neuroblastoma public-Versteeg -88". Note that you have to select the correct dataset set before starting with the Chip-seq analyses. To enter the ChIP-seq analysis module in R2 select *ChIP data* in Box 3 (Fig 4) and click "Next".

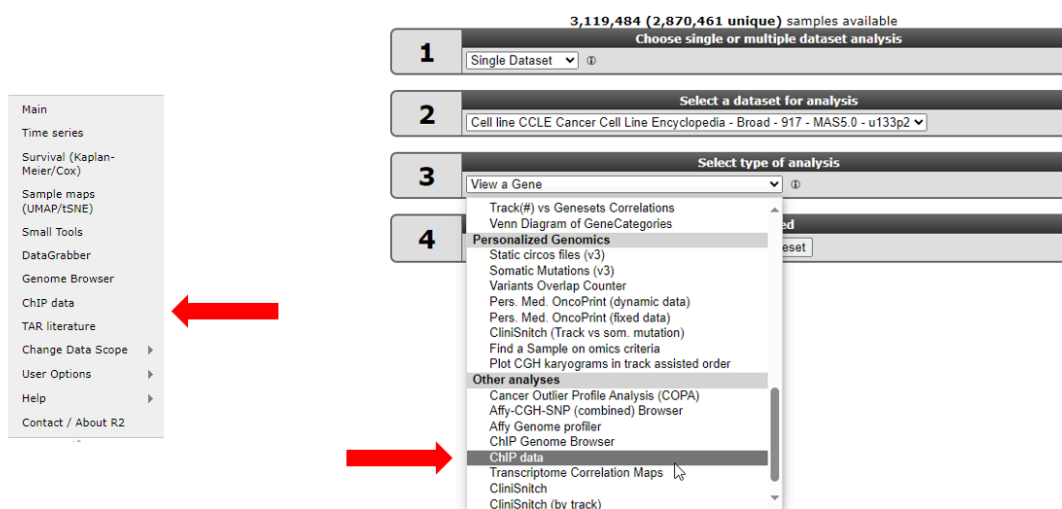


Figure 4: Choose the ChIP-seq module

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 (+GEX)*

ChIP (Transcriptional Start Site)
TSS graphs within experiment
TSS graphs
TSS genesets scan
TSS XY plot 2 experiments
Feature Overlap
ChIPseq
ROSE Super Enhancer Plot
MACS Plot
RSEG Plot
ChIPseq TSS Peak/Coverage Plotter
ChIP Browser
ChIP Genome Browser
ChIP Genome Browser + Gex
Core Regulatory Circuitries
CRC graph
ChIP Utils
Make Gene Groups Prescription (for TSS)

Figure 5: ChIP-seq Menu in R2

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. Check first in the “Adjustable settings” menu if the GenomeBuild is set to **hg19** in the pull down. 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.

The screenshot shows the 'Gene search' panel with 'Gene / Probeset:' set to 'GATA3' and a 'Next' button. Below it is the 'refseq: GATA3 (hg18)' panel, which is a table with columns: View, Locus, Chrom, Start, End, Size, Exons, and CDS. It lists two transcripts: GATA3: trans-acting T-cell-specific transcription factor GATA-3 isoform 2 and GATA3: trans-acting T-cell-specific transcription factor GATA-3 isoform 1, both on chr10 with coordinates 8,135,656 to 8,158,167 and size 22,511. Each row has a 'View' button and buttons for 'Exons' and 'CDS'.

Figure 6: Looking up a single gene in the R2 Genome Browser in ChIP-seq context.

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.

The screenshot shows the 'Available Datasets' panel with a 'Select / Adapt ChIP-Experiments' button and 'nr. selected: 0'. Below it is a table with columns: isoY, chip_group, Y id, Y chip_celline, Y chip_target, Y chip_authorY, colors, and updy. The table lists several experiments, with some rows checked. Below the table are 'Update' and 'show all' buttons, and a 'rows:327' indicator.

Figure 7: Selecting experiments by using grid filtering

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 with H3K27me3.

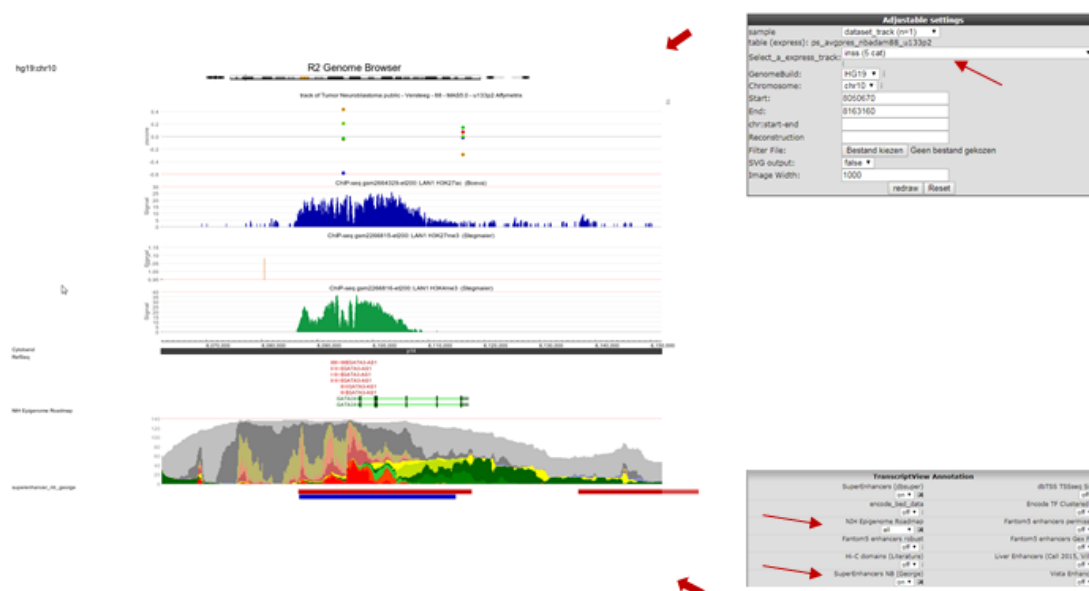


Figure 8: Histone modification profiles around GATA3 location

In the next figure the legend is shown by color what kind of chromatin state the colors represent and also how to toggle the NIH plugin to get detailed information for the tissue types.



Figure 8a: Legend and detailed source information

1. 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. Open a new tab by right click “Go to main” and click chip data in the left menu. Now choose the *ChIPSeq TSS Peak/Coverage Plotter*
2. 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”.
3. Set perspective to peaks if it was not done so already. 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 click “Next”. The Gata binding sites around the genes in the list are shown (Figure 9).

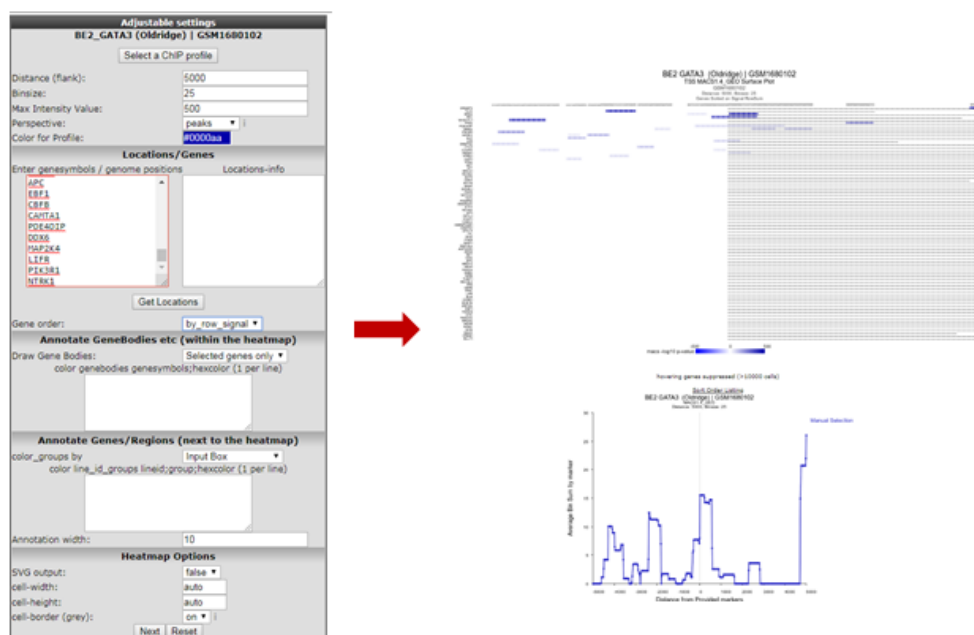


Figure 9: Gata binding site around genes

4. 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).

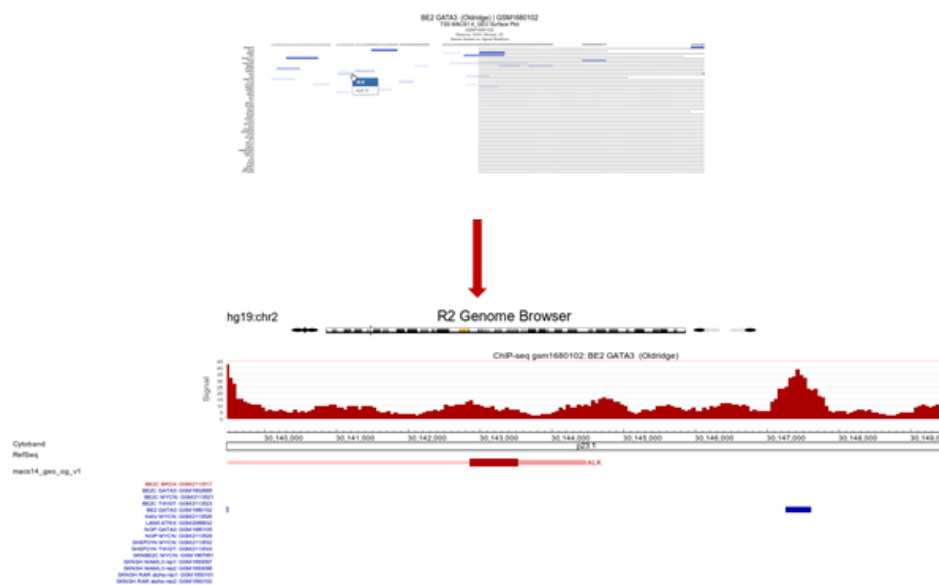


Figure 10: ALK profile within GATA3 ChIP-seq experiment

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. In picture 11, The following experiments were selected: chip_Author:Oldridge -> BE, BE2, Kelly, NGP and SY5Y and chip_author Bernstein -> LAN6. 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.

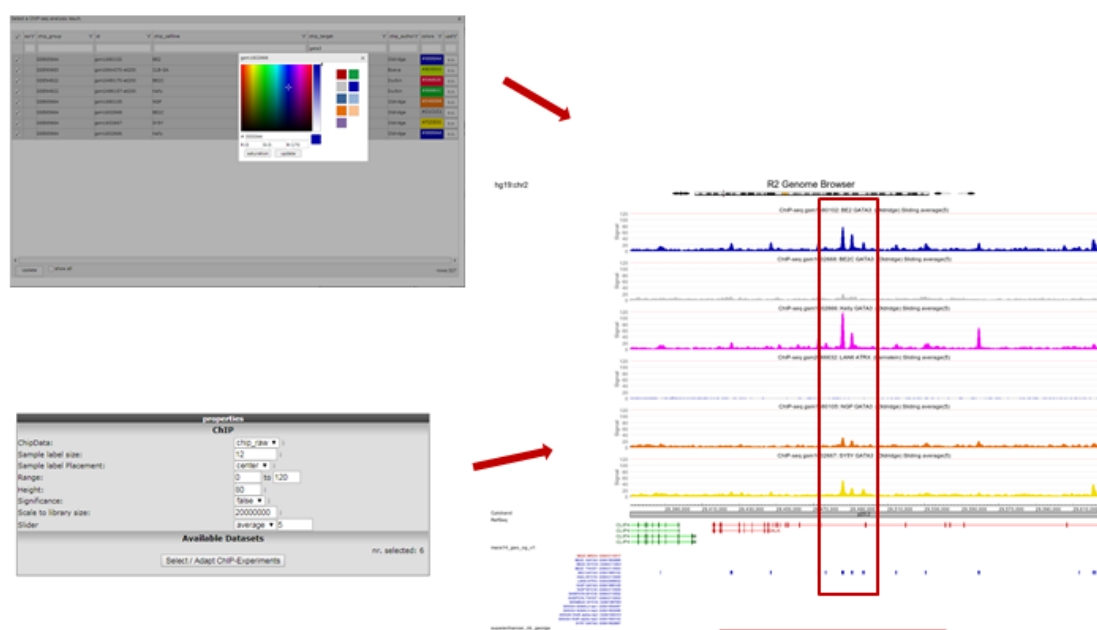


Figure 11: GATA3 binding data around the ALK gene in multiple datasets

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.



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

- 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).



Figure 13: ChIP-seq signals around the TSS of a single gene

19.6 Step 4: Finding active super-enhancers

- 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*
- 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*
- Again select the SY5Y dataset from Oldridge in the next panel

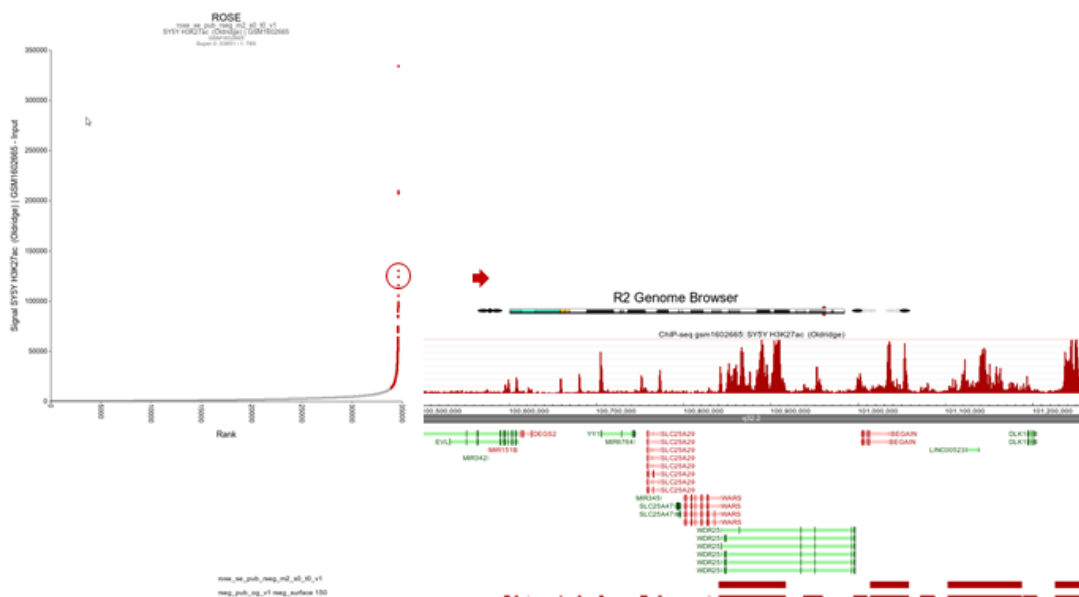


Figure 15: Selecting super-enhancers from an interactive ROSE plot

- 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.

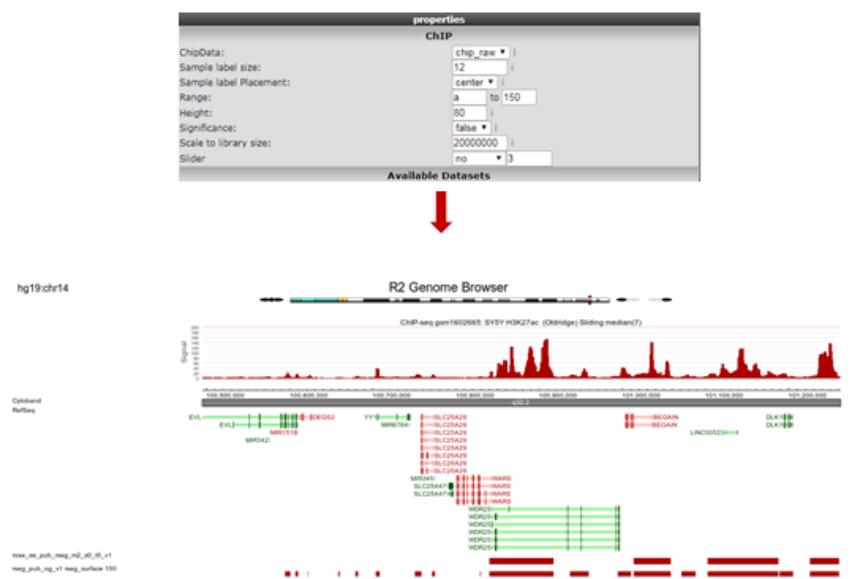


Figure 16: The genomic context of a top-ranking super-enhancer

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.

Integrative Analysis : Across Platforms

Datatypes: Methylation data and expression data

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 - ilmnhm450 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 - ilmnhm450) 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.

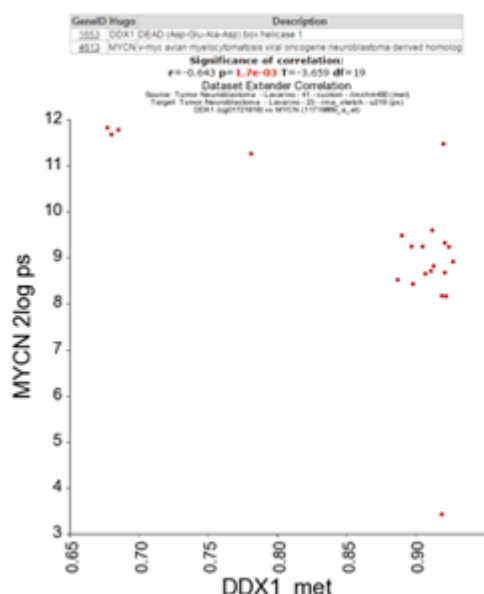


Figure 1a



Figure 1b

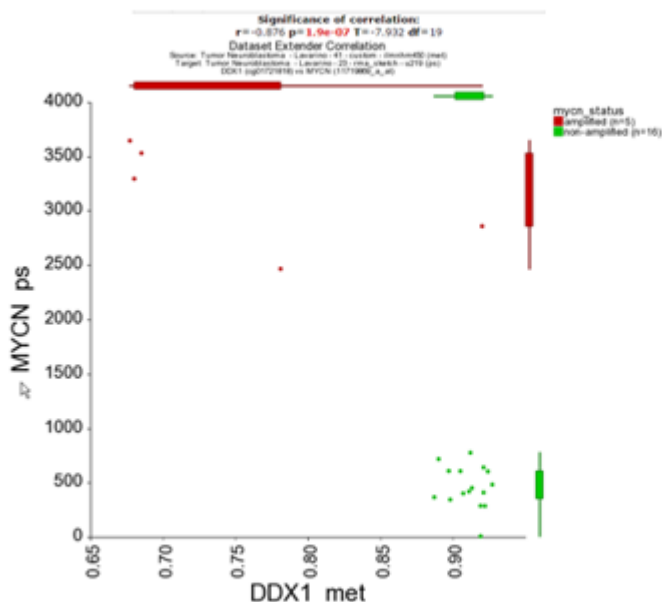


Figure 1c

Figure 1: XY-plot Methylation versus expression

1. 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.

- 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 adjust able settings. (Figure 2).

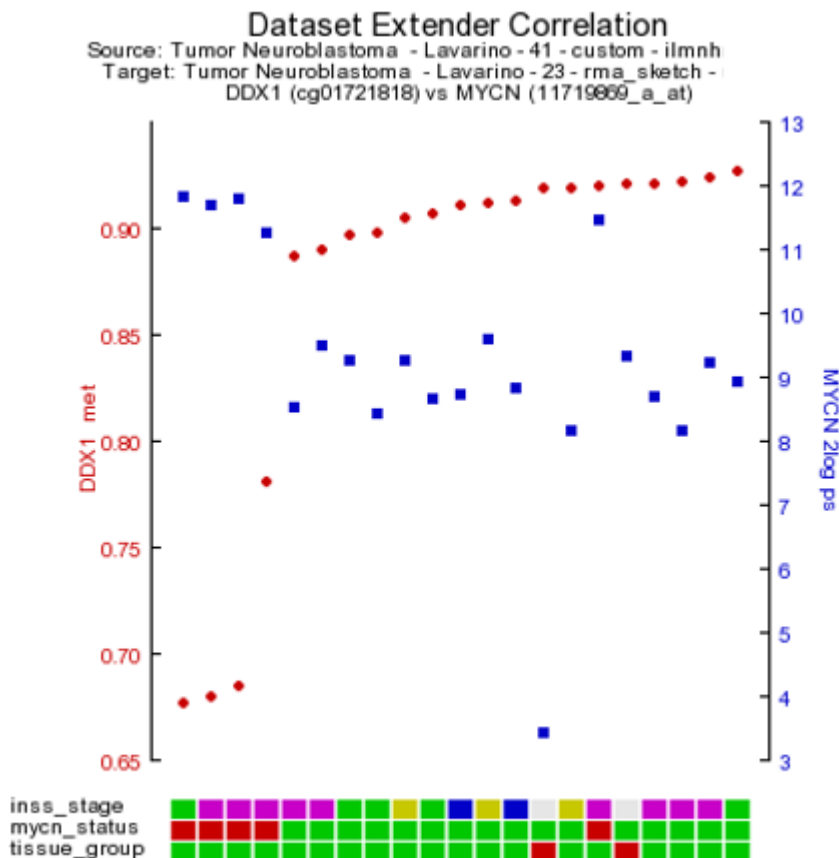


Figure 2: XY-plot Methylation versus expression

- 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'.

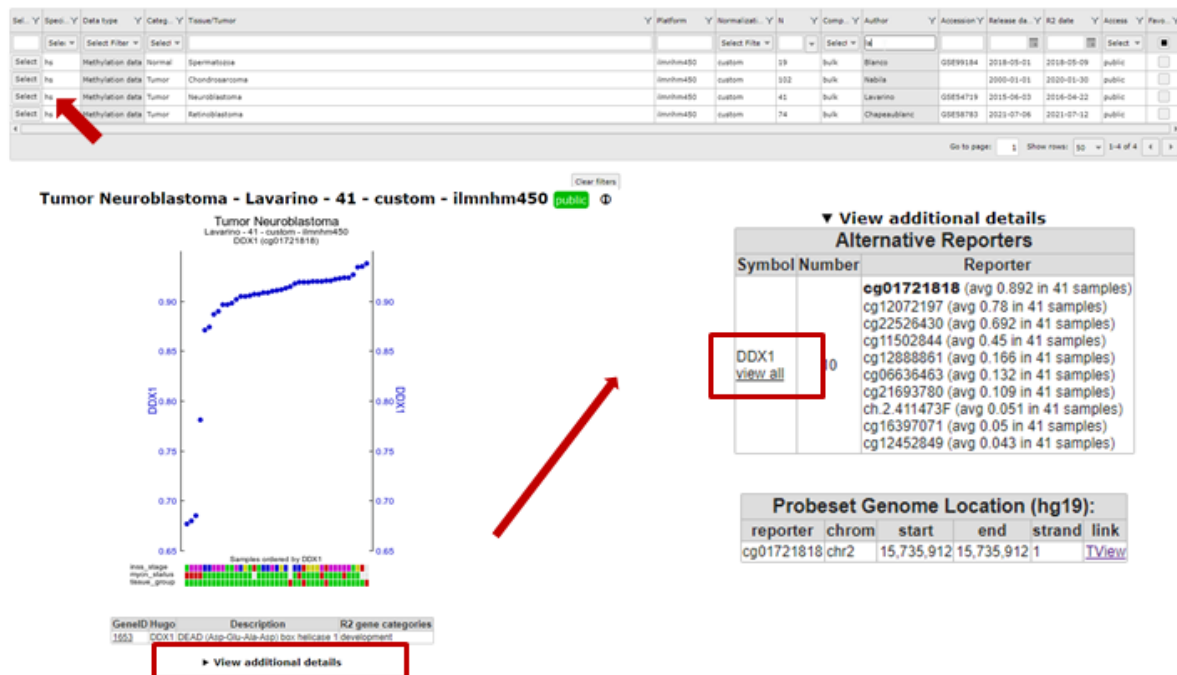


Figure 3:Heatmap select all reporters

1. 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'. Clicking on the Blue link, **View Chr 2, in the genomebrowser**, will open the genome to its full extent with all kinds of functionalities like, zoom in/out, move across the genome browser and adapting all kind of tracks.

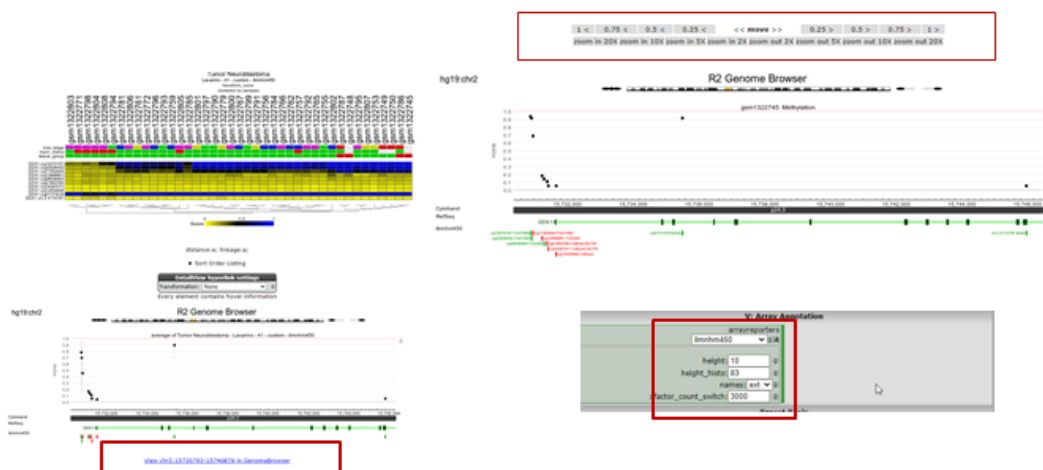


Figure 4:Heatmap select all reporters

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.

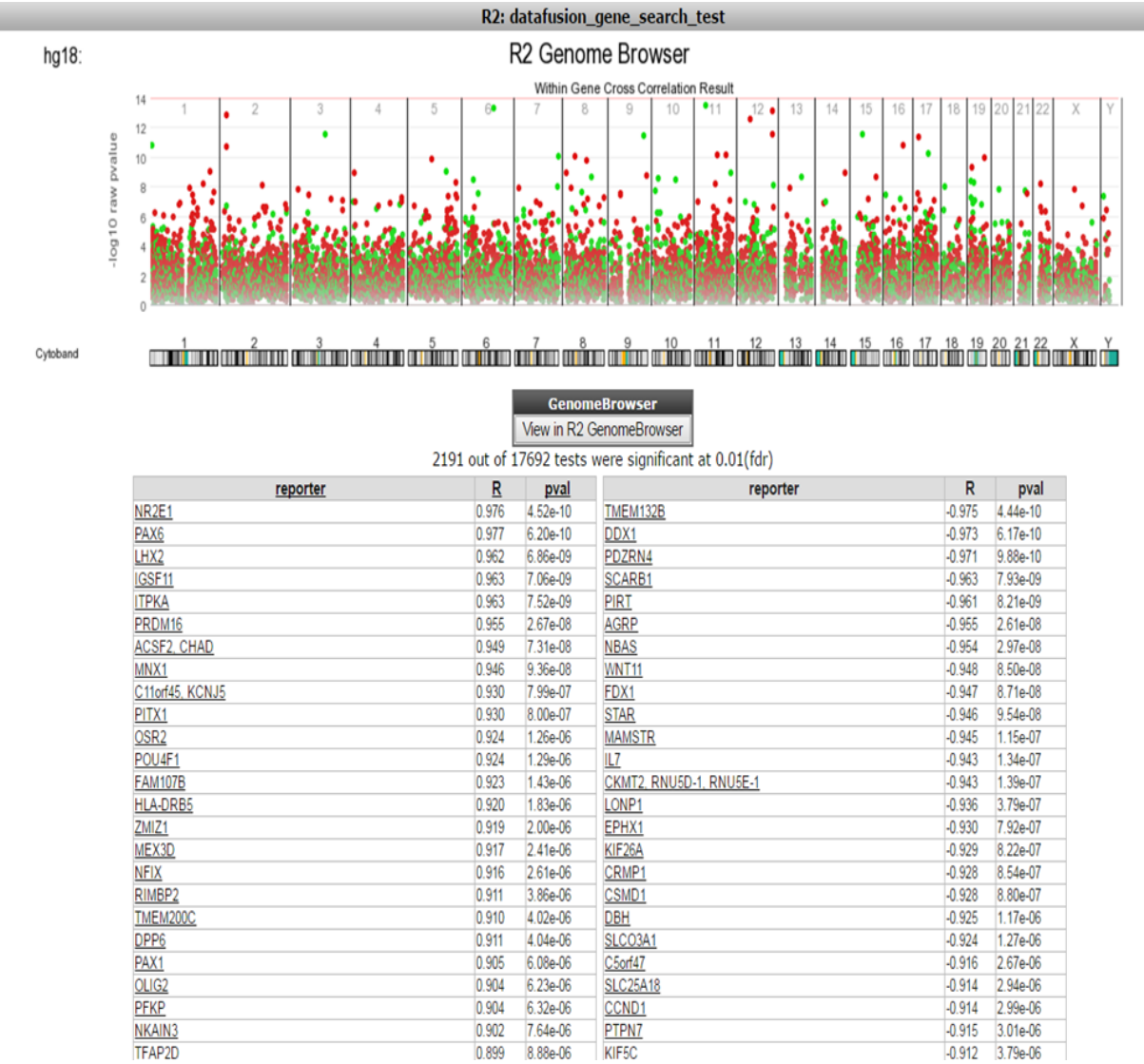


Figure 5:Table of Correlations

1. 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. Hoovering 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).

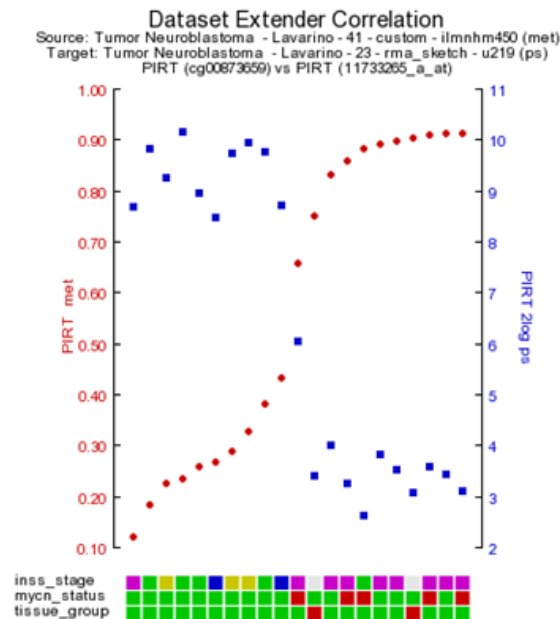


Figure 6:YY from table

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.

2. 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'.

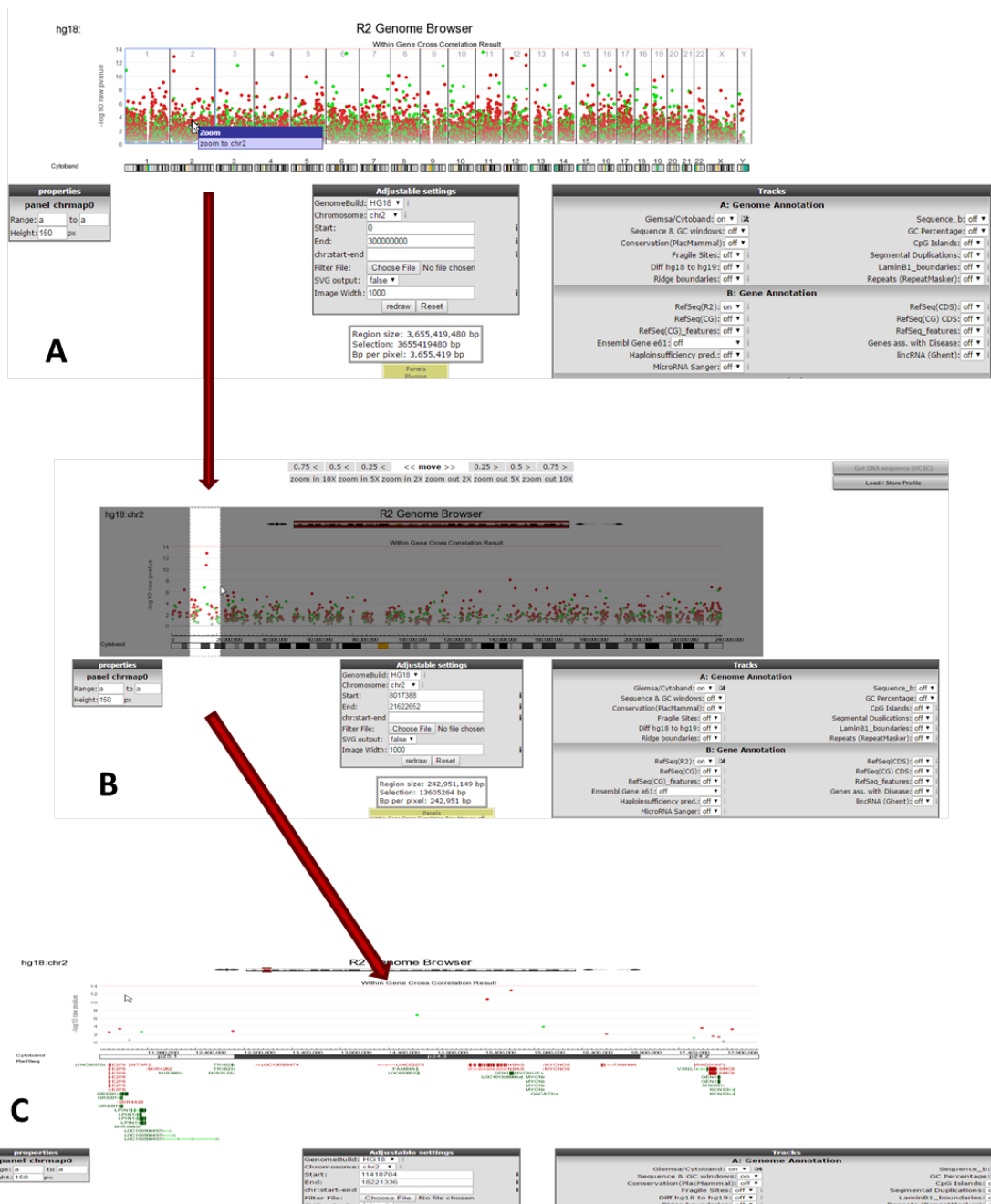


Figure 7: Explore chromosomal regions

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 MYCN methylation reporters and expression reporters.



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*

Integrative Analysis : WGS/NGS data

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).

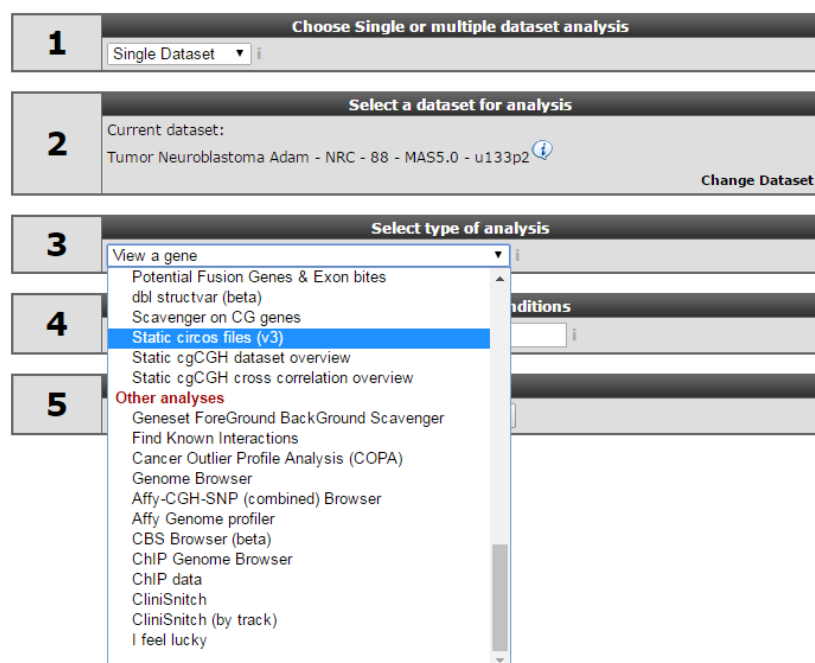


Figure 1: Choose the Circos plot module

2. Select a subset of samples by using the menu on the top.
 - Select *inss (cat 3)* from the select a track (subset) selection box.
 - Select *st2 (1)* from the pop-up selection window and click 'OK'.

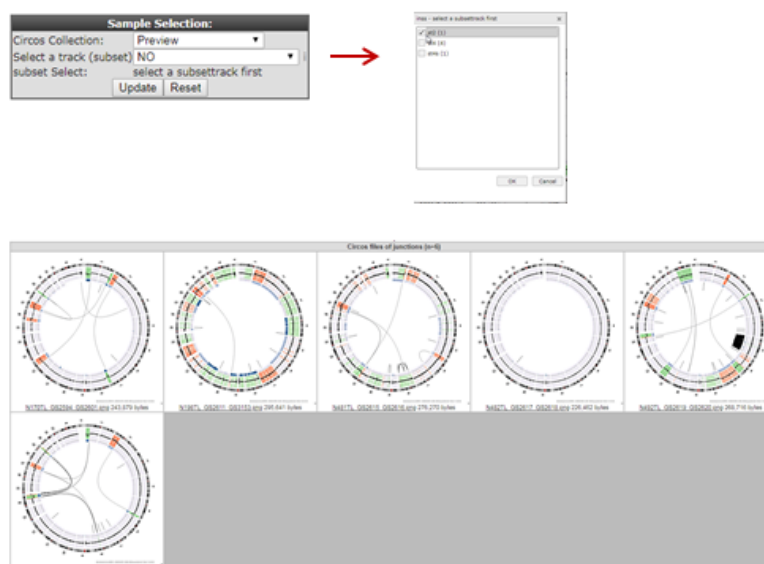


Figure 2: Select a subset of samples

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.

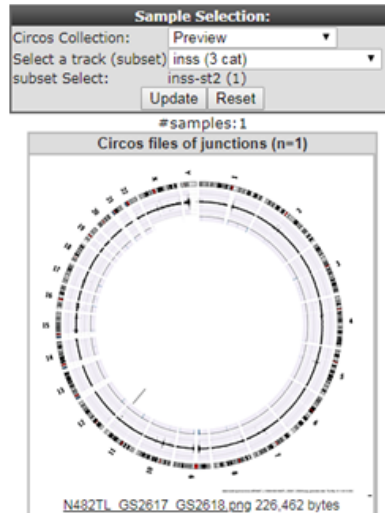


Figure 3: Circos plot

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. High 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.

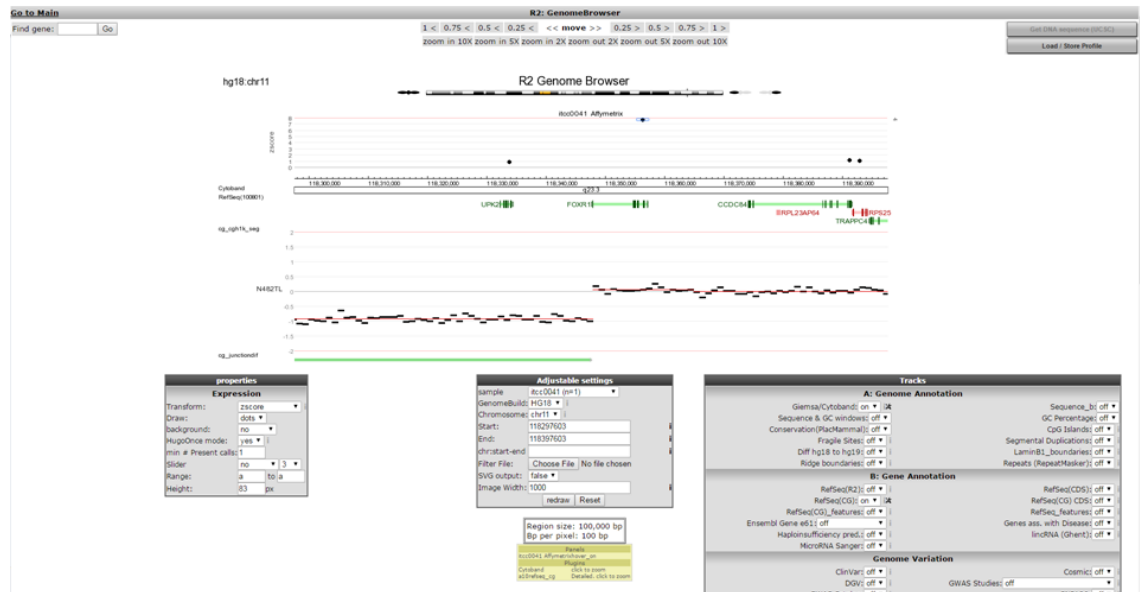


Figure 6: Including Array data in the scatterplot

- 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.

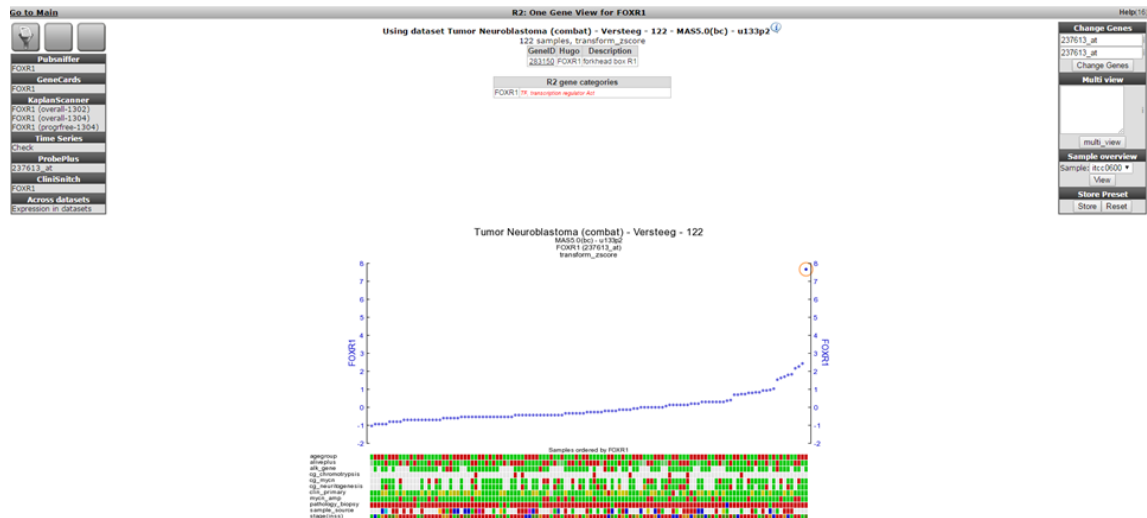


Figure 7: FOXR1 in the tumorseries

Using the MegaSampler tool from within R2 you can also show that the *FOXR1* gene is only expressed in a hand full 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).

Target Actionability Literature Reviews : TAR

Datatypes: Literature review results

22.1 Scope

- In this part R2 is used to provide an overview of manually curated literature data, to support the targeted drug development process for pediatric cancers. This is a specialized topic. Published TARs can be explored by anyone. Many more TARs require specific access or membership in consortia. Next to the presentation of TARS, R2 is also equipped with a TAR creation module, where literature reviews can be created, including an evidence collection tool.

22.2 Step 1: View a TAR.

1. To view a target actionability review in R2, select *TAR literature* from the menu (Fig 1).

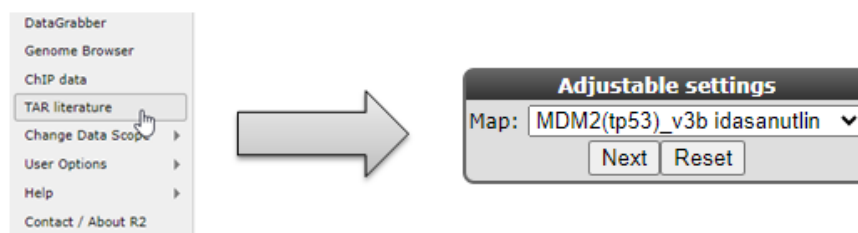


Figure 1: Navigating to the TAR section

1. A TAR typically is a structured stepwise literature review to comprehensively assess proof of concept (PoC) preclinical data. Such a review of published literature is assessed on a particular target gene or pathway and corresponding compounds or drugs, assessed within a collection of pediatric cancers. Such a review then highlights the strength and potential gaps in the current knowledge in the drug target and associated drugs in an informative overview across the assessed malignancies, and may encourage additional preclinical testing in an efficient manner. The TARs can also provide guidance for well-informed decision-making on and prioritization of subsequent further preclinical and clinical evaluation.
- Depending on your account and memberships, you will be able to 'select' a review from the dropdown menu.

- Select the *MDM2* TAR from the selection and click ‘next’.

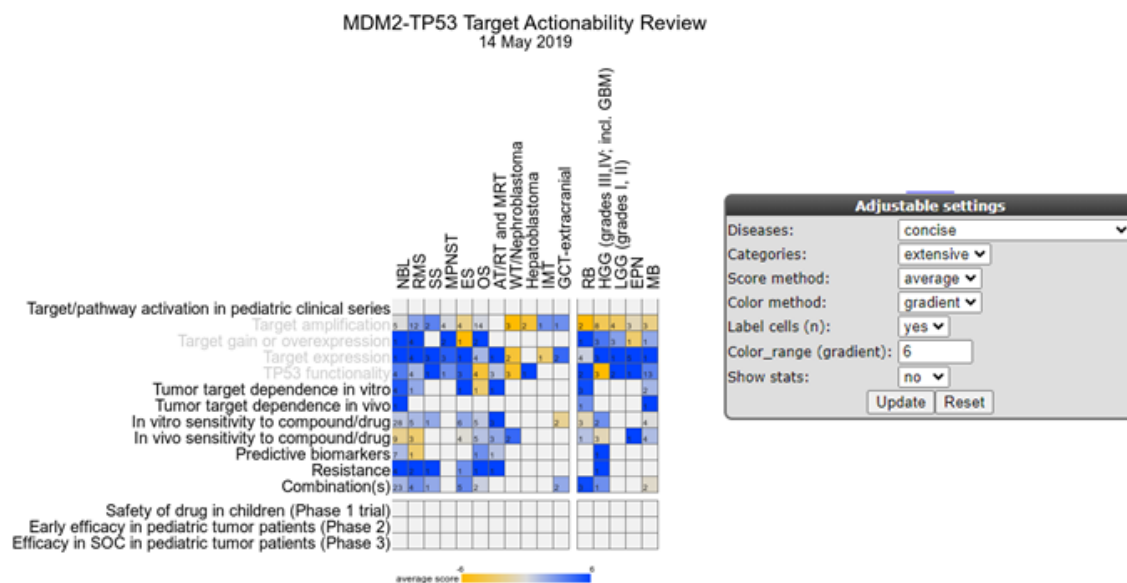


Figure 2: TAR result for MDM2 and Idasanutlin in Pediatric cancer

1. A TAR is populated by ‘evidence items’, which are extracted knowledge from publications that are ‘scored’ for experimental quality (method/sample size) and experimental outcome (support). These scores are then averaged per PoC category and tumor entity and displayed as color intensities in an interactive heatmap.
 - Using the adjustable settings, the TAR can be ‘folded’ and ‘unfolded’ to assess finer sub-groups (if assessed), or focus on a particular tumor entity.

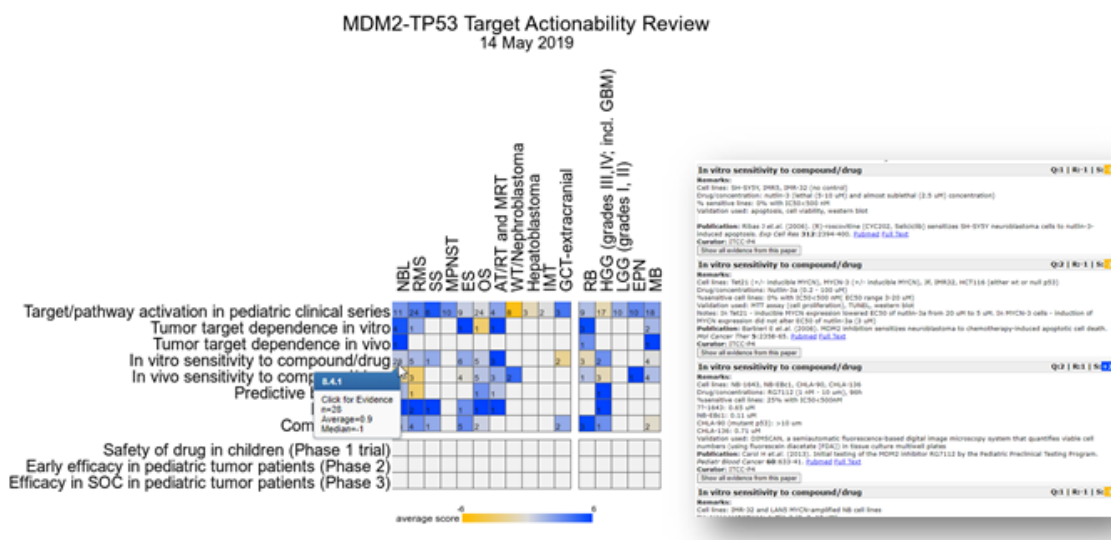


Figure 3: Evidence Item exploration from the TAR

1. All PoC category / tumor entity combinations can be ‘clicked’ to obtain the evidence items that make up the score. These evidence items are little stories explaining the given score. These items also link out to the publication from where the evidence has been extracted.

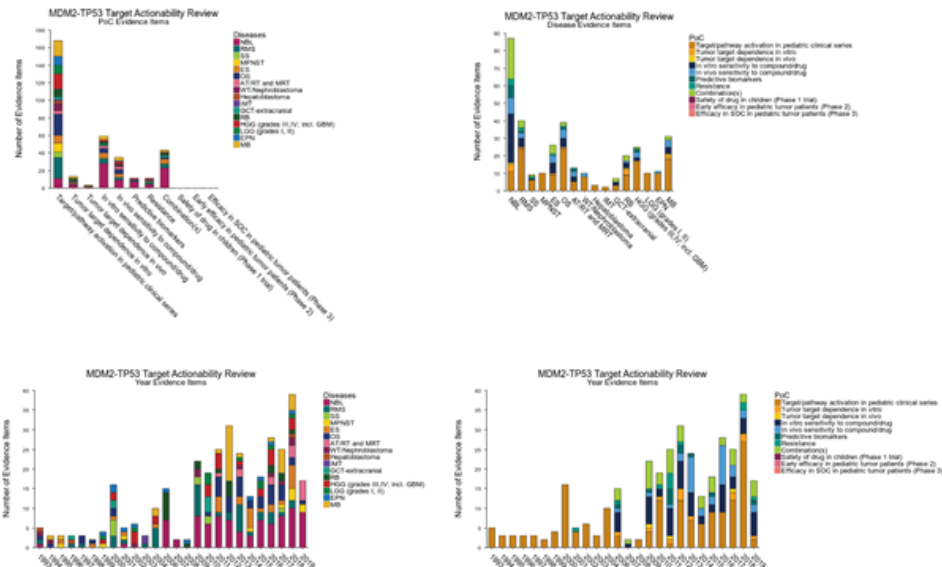


Figure 4: Stats plots for a given TAR

2. You can also visualize some stats for a TAR in comprehensive plots, by switching on ‘show stats’.
3. Depending on your assess and memberships, there can also be the option(s) to perform tasks as a reviewer or TAR administrator. For more information on those functionalities, please get in contact with r2-support@amsterdamumc.nl

22.3 Final remarks / future directions

Some of these functionalities have been developed recently. If you run into any quirks or annoyances do not hesitate to contact R2 support (r2-support@amsterdamumc.nl).

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

Adapting R2 to your needs

Or how you can optimize R2 for your specific data analysis

23.1 Scope

- This tutorial describes the adaptable settings within R2. These are basically all items under the User Options 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

23.2 Step 1: Adapt your settings

1. Personalizing R2 starts with selecting the 'User Options' menu item (Figure 1). When you hover over this item, you see a submenu. In 'Account you can choose a different password or change your personal details. In 'Preferences' you can set some generic R2 analysis and visualization options, click on this item.

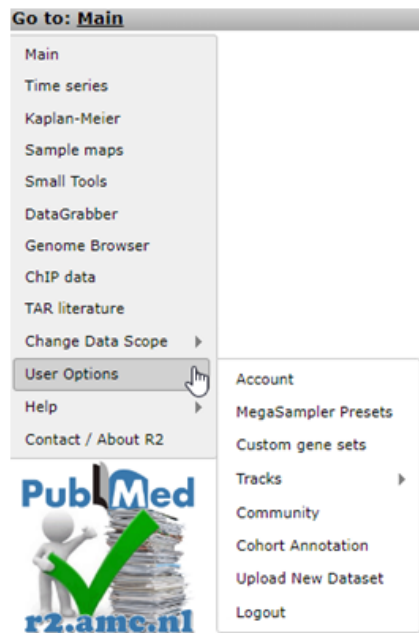


Figure 1: Personalizing R2: the User Options menu-item

2. Next item in the User Options submenu (Figure 1) are the ‘Megasampler Presets’. These are of relevance when you’ve built a specific Preset in an analysis Across Datasets (see chapter: Multiple datasets overview with Megasampler).

23.3 Step 2: How to add data to R2.

1. One of the most appreciated options of R2 is the possibility to add data to R2, be it your own dataset or/and publicly available data that matches your research interest. Due to several reasons, technical as security, it’s almost impossible to automate the process of adding data for standard users. In order to keep the database curated only R2 administrators can add data to R2. In order to do so, the data first has to be processed and uploaded. [chapter 24](#) describes in detail how to prepare your data such that we can process it and upload the data to R2.

(r2-support@amsterdamumc.nl). If you would like to see a publicly accessible dataset in R2, then send an email to r2-support@amsterdamumc.nl with a link to the data, or in the case of a Gene Expression Omnibus dataset, the GSE**** identifier, matrixes in supplemental data and we will take care of the rest.

23.4 Step 3: Create your custom genesets

1. Another powerful functionality to adapt R2 analyses to your specific needs, is by defining gene sets or referend to in r2 as genecategories. Many analyses in R2 can be performed on a subset of genes (see [chapter 14](#) for a tutorial on performing gene set analysis). There are 3 main sources for gene sets. Firstly R2 harbors hundreds of predefined sets of genes (such as KEGG pathways or sets defined by the Broad Institute). Secondly, some analyses will result in gene lists, which R2 allows you to save on the fly such that they can be used for further analyses (e.g. [Toplister analysis](#)) .Next to these two options, you can introduce your own gene sets of interest directly to R2 as well; Hover over the ‘ custom genesets’ sub-item and select the button ‘custom geneset editor’ (Figure 5).

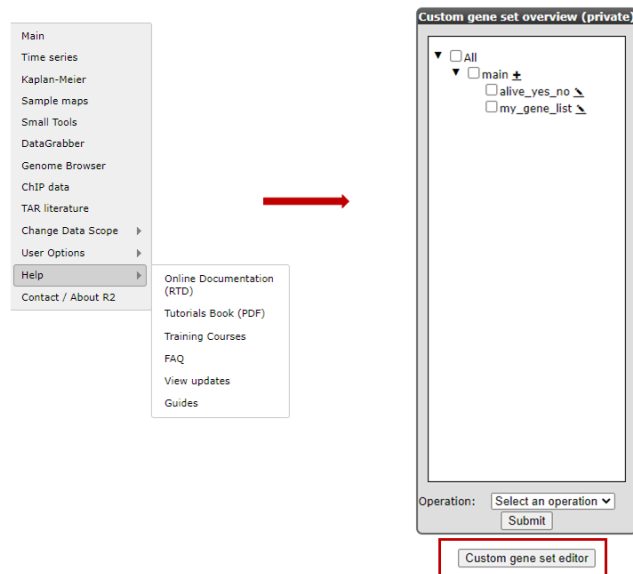
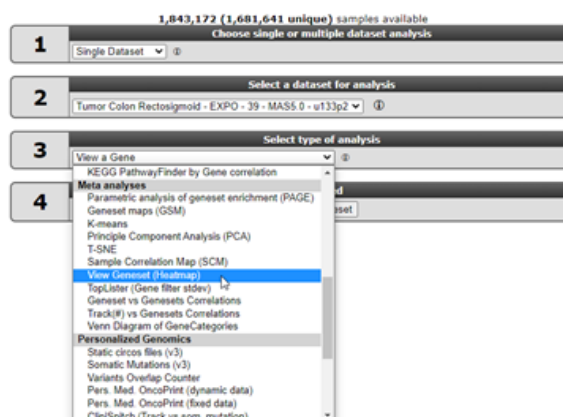


Figure 2: Geneset related menu-items; select Custom genesets to make your own.

2. The 'Custom Geneset editor' window pops up (Figure 3). By default in this window you must provide a unique name for the set. The input box allows you to paste a list of genes to upload as a geneset for use in analyses in R2. In the example a set of genes, specific for ALL tumors are pasted. If you want this gene set to remain available for you, select in the community dropdown, "none" or select a community name for sharing the geneset. The concept "community" is described later in this tutorial. If you just want to store the set temporarily for 24 hours, choose 'yes' in the temporary dropdown. Click "save geneset" to upload the set (Figure 3), you'll receive a message when everything has succeeded. Your set of genes is now available as a geneset for all analyses within R2. Go back to the main page to see where you can use this set.

Figure 3: Using the Input Box to upload your genesets.

3. We're going to lookup your geneset, an example is available in the Gene Set View. In the main menu in Field 3 select 'View a Geneset (Heatmap)' and click "next" (Figure 4)



4.

Figure 4: Using a geneset; select View Geneset

5. In the GeneSetView your custom geneset is privately available for yourself for similar analyses as with any other public gene set present in R2. Select 'My GeneCategories' to choose from your categories.

If you saved your gene set temporarily, choose 'My 24h GeneCategories'. And click Next and click Next again in the following window (Figure 9).

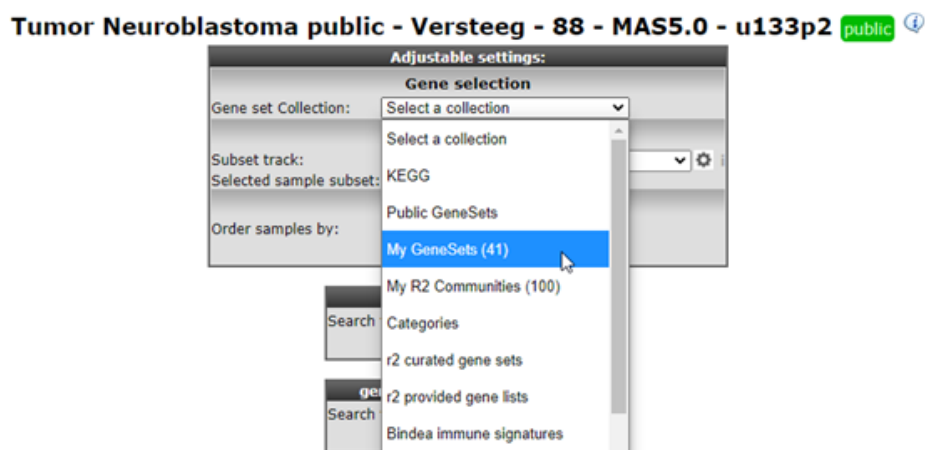


Figure 5: Selecting your genesets

1. Now you can specify which gene set you want to view and how you want the heatmap to be displayed. The geneset 'Changed Genes' we just made above is available (Figure 6), click on it. Also, in the Heatmap Options 'color-scheme(v2a)', select 'green-black-red', or any scheme that you prefer. For now we end here, later on we'll see the geneset again in the context of Tracks.

Adjustable settings

Subset track:

Selected sample subset: None

Gene Set(s)

Gene set:

Transformation

HugoOnce mode:

Transformation:

Heatmap Options

Vector (SVG) output:

cell-width:

cell-height:

cell-border (grey):

Max range color scale:

Color scheme:

Distance measure:

Label track:

Plot signature scores:

Show heatmap:

Track Display Selection

Figure 6 A: Your geneset is available.

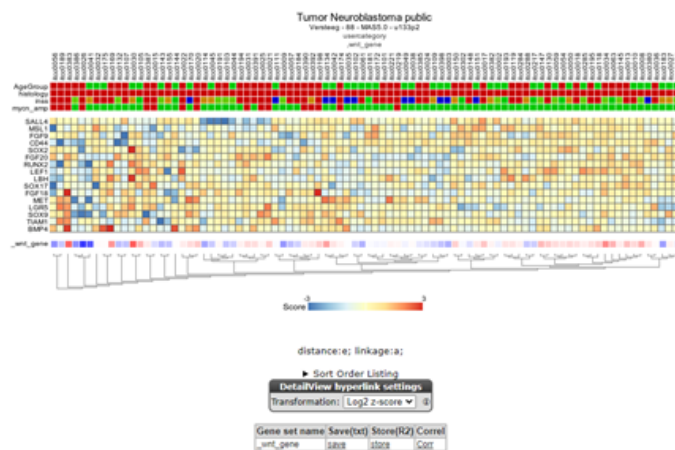


Figure 6 B: Your geneset is used to create a heatmap.

2. We now return to the side menu of the R2 page to find out how we can manage the genesets we just built. From the 'User Options' item in the menu, click Custom geneset. The custom geneset module allow you to organize you custom genesets , assigning the sets to a collection or delete custom sets.

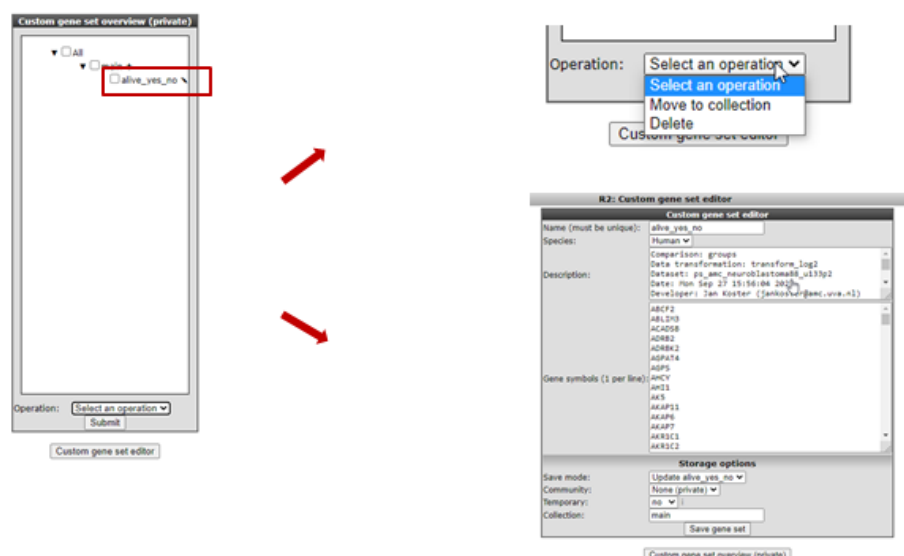


Figure 7: Adapting your genesets

Existing genesets can be adapted, deleted or moved to another collection. New genesets can be based on existing ones. As an example we're going to update the genesets we just made. Click the 'pencil' icon next to the custom geneset in the custom geneset editor. In the next screen you can add or delete genes and provide background information and choose for the update of new geneset option in the pulldown menu.

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

1. Another important feature in R2 that can be adapted to your needs are grouping variables, that we call "tracks" in R2. In R2, the samples can be annotated with sample characteristics, e.g. clinical data or experimental characteristics. Each group of annotated data is called a "Track". Tracks in R2 give you the opportunity to divide your samples in groups with e.g. different phenotypes for comparative or subgroup analysis. They also allow you to restrict your focus on only a part of the samples within a given dataset (when used in the 'subset builder'). For some datasets the annotation that you need may be available already (from the default annotation that was added by the R2 team). For others you might want to add extra sample annotation for analysis such as combining already added tracks, or introduction of new information that you may possess. Tracks can be adapted in multiple ways:

- They can be uploaded with annotation files
- They can be created as a result of analyses within R2 and stored within the platform on the go
- Or you can create a so called Custom Track yourself within R2.

We'll first start with an example of adding a track from the results of an analysis that is performed from within R2. We will illustrate this option using a K-means analysis. Such an analysis results in a division of the samples in two groups (in the case of $k=2$. For more about this analysis see [chapter 14](#)). On the main page of R2 select the K-means analysis in Field 3 (Figure 8)

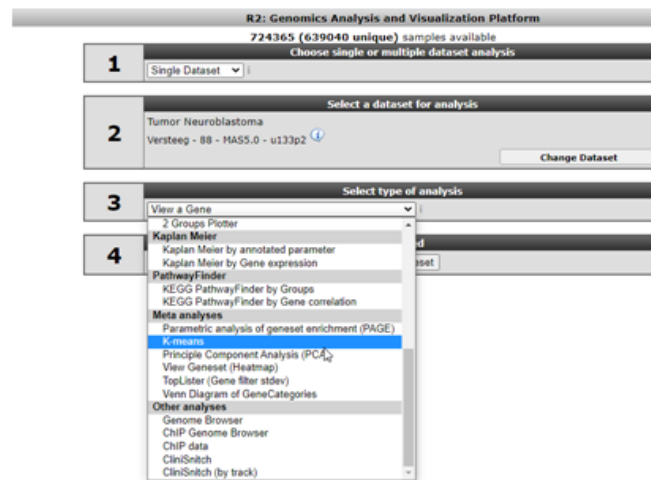


Figure 8: Selecting a K-means analysis

1. In the settings window for the K-means analysis (Figure 9) you can choose the geneset created above to cluster the current set of samples. In our case this is called ChangedGenes. Make sure that the number of draws is set to 10x10, click 'next'

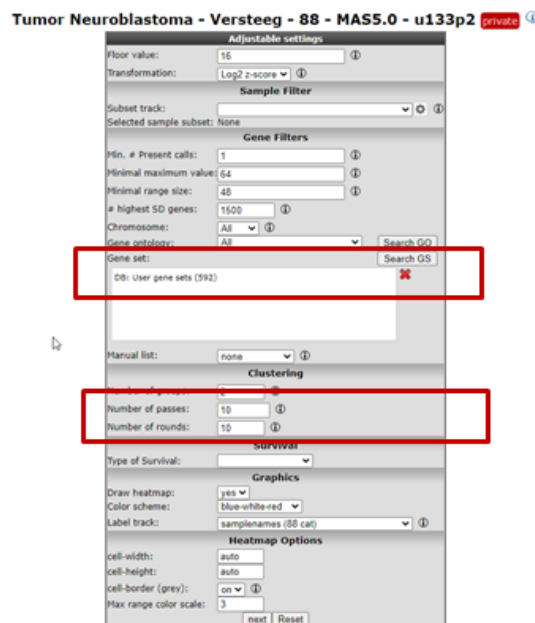


Figure 9: Settings for K-means; the Category built above is available for clustering

2. The resulting clustering in two groups might not be ultimately convincing (Figure 10, your result might look slightly different), but for our testing purposes this is alright. What is important is that the resulting groups can be stored as a new track, personal / available only to your account; click the button 'store as a track'.

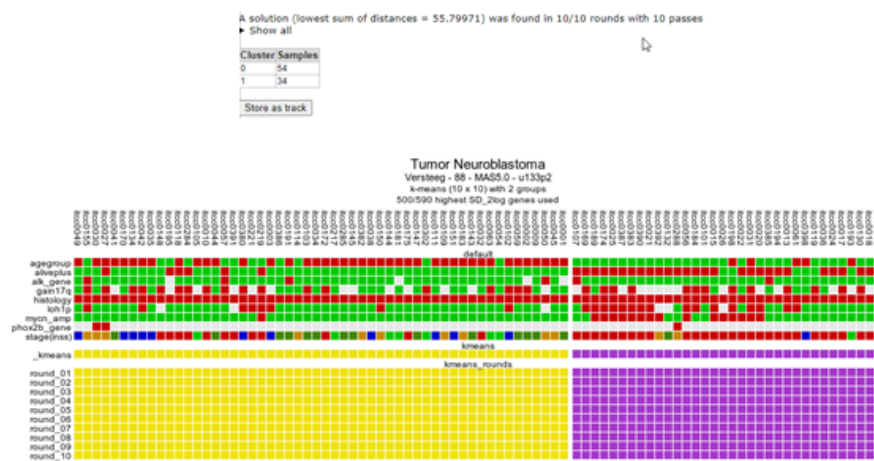


Figure 10: Clustering result of the Neuroblastoma dataset with a geneset built in the former steps

- 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 11). 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.

<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	itcc0221	group 1
<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	itcc0284	group 1
<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	itcc0285	group 0
<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	itcc0288	group 1
<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	itcc0302	group 1
<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	itcc0380	group 0
<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	itcc0382	group 0
<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	itcc0383	group 0
<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	itcc0385	group 0
<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	itcc0386	group 0
<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	itcc0387	group 0
<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	itcc0390	group 0
<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	itcc0391	group 1
<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	itcc0392	group 0
<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	itcc0398	group 1

Adjustable settings

Name for Group1: group 1 A33CC8
Name for Group2: group 0 F0E116
Track name: ChangedGenesKmean
Show as track: no ▾
Where: personal track ▾
Description (usergroups only):
k-means clustering in 2groups
after 10 passes. |
Build set Reset

Figure 11: Storing the current groups as a Track for use in later analysis.

- After optionally changing the parameters, you can click the Build set button to store the track. In the custom tracks manager we can adapt this track again. From the 'User Options' menu select 'Manage Custom Tracks' (Figure 12).

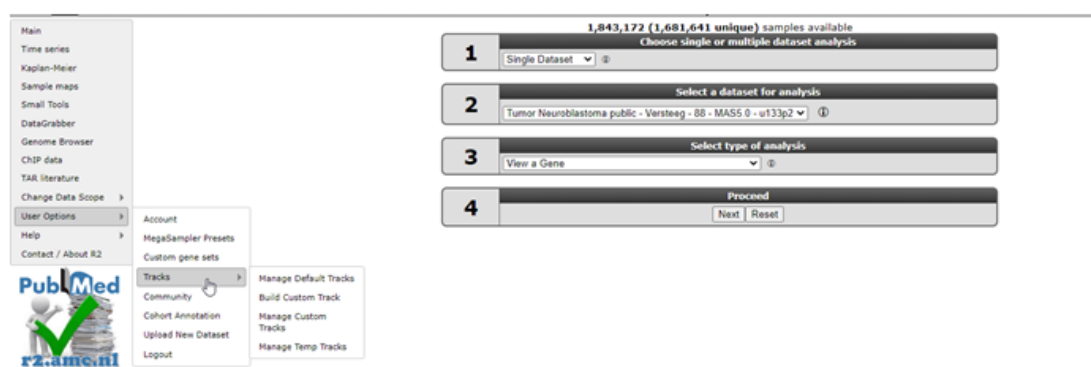


Figure 12: Selecting the Manage Custom Tracks

5. In the next screen keep the default selection, i.e. your current dataset. Tracks are, of course, defined based on a specific dataset; for each dataset you can store your own tracks. Click 'Next'.

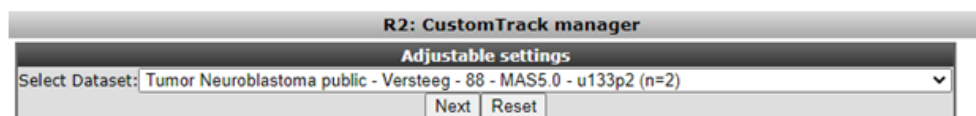


Figure 13: Tracks are defined per dataset; keep the current selection.

6. In the next screen you're able to adapt the Track we just generated. Of interest in here is the option "Show as track: Yes/No", 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. 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!

Custom Track Manager for Tumor Neuroblastoma public - Versteeg - 88 - MAS5.0 - u133p2

Collection	Visible
main	true

Update Collection Reset

Collection	TrackName	Visible	Drawtrack
main	NoNameProvided	true	no
main	mycn	true	no
main	testforkaplan_r2_mesadm_adm	true	no

NEW Collection name:

Update Tracks Reset

Copy/Delete/Rename/Export Tracks

Figure 14: Adapting track parameters.

7. Select Manage Default Tracks from the 'User Options' > 'Tracks' sub-menu (Figure 15)

Go to: Main

- Main
- Time series
- Kaplan-Meier
- Sample maps
- Small Tools
- DataGrabber
- Genome Browser
- ChIP data
- TAR literature
- Change Data Scope
- User Options
 - Account
 - MegaSampler Presets
 - Custom gene sets
 - Tracks
 - Manage Default Tracks
 - Build Custom Track
 - Manage Custom Tracks
 - Manage Temp Tracks
 - Community
 - Cohort Annotation
 - Upload New Dataset
 - Logout
- Help
- Contact / About R2

R2: DefaultTrack manager
Default Track Manager for ps_avyprz_medullokol62_u133p2

Collection	Visible
main	true

Update Collection Reset

Collection	TrackName	Visible	Drawtrack
main	bcat_mutation	true	yes
main	histology	true	yes
main	nt_source	true	yes
main	pitch_mutation	true	yes
main	sex	true	yes
main	staging	true	yes
main	subtype	true	yes

NEW Collection name:

Update Tracks Reset

Reset to defaults

Figure 15: Selecting the Default Tracks Manager

8. 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 16). Basically all annotation provided with this dataset is available as a track. Try out different things here. We'll select additional annotations by changing their Drawtrack value to 'yes': age_year, gender and recurrence will be shown underneath graphs as well in further analyses. Also, we'll set the Collection of age_year and agegroup to NEW. Next to NEW Collection name, we add a befitting name that defines this group of tracks. Here we typed 'Age related'. Be sure to click the 'Update Tracks' button for these changes to take effect.

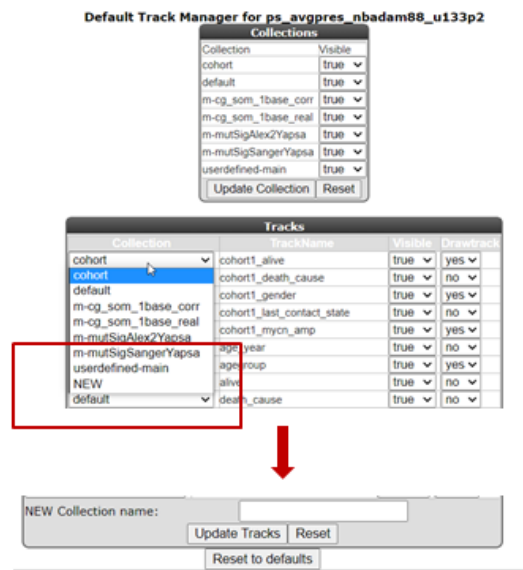


Figure 16 A: Selecting the default tracks for this dataset

When collections of tracks are used, these will show up conveniently as separate groups of tracks under the graph.

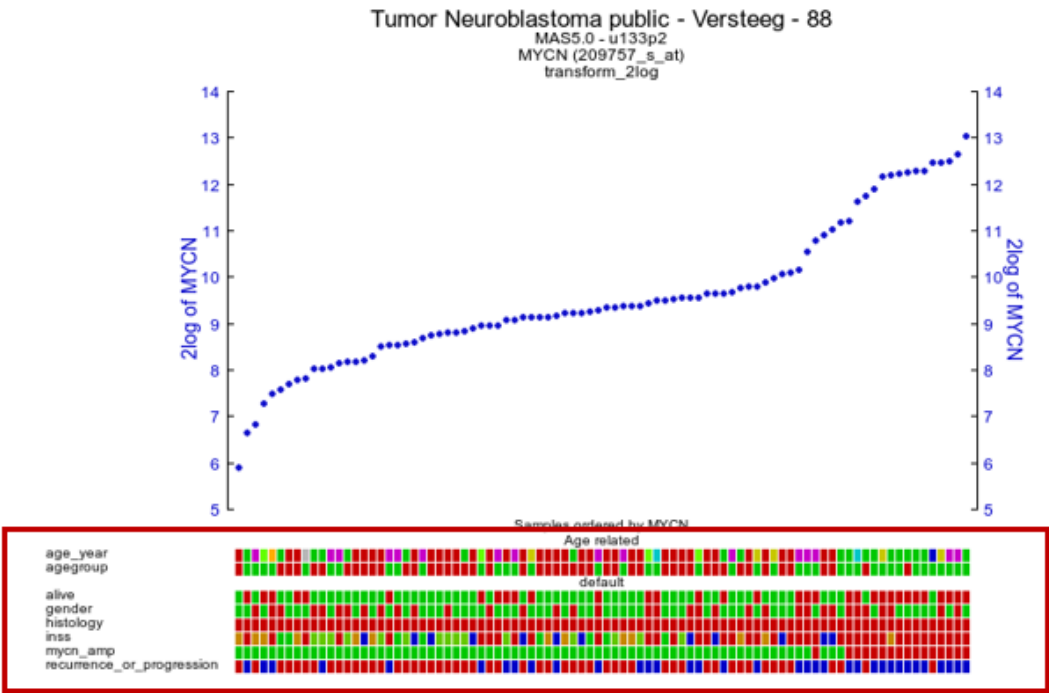


Figure 16 B: ‘Drawtracks’ makes tracks visible under the graph; group Tracks with ‘ Collections’

Use the ‘Reset’ button for Tracks or Collections in the Default Tracks Manager to undo either of the changes, or use the ‘Reset to defaults button’ to go back to the original dataset settings of tracks.

23.6 Step 5: Upload your own tracks

1. R2 also allows you to build your own tracks from scratch. You will be able to assign each sample to a group of your choice. To illustrate this select 'User Options' > 'Tracks' > 'Build Custom Track'. The Custom Track window appears. R2 also provides the possibility to upload a custom track from a prefabricated text file. We will shortly show this route, which is also the most powerful one. Click 'Upload or Paste a Track (txt file)' (Figure 18).

Figure 17: Build a Custom Track: Upload or paste a track.

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 input box. Either of these options provides R2 with the proper assignment of each sample to a specific value. Do note, that the internal identifiers ('samplernames') that are used within R2, need to be provided here. Based on the values you provide in the second column, R2 creates the groups for you. You can create tracks with as many groups as you like. When described in a text file; for each sample a description can also be provided.

```

trackname      testtrackname
group_color    <1      #ff00ff
group_color    >=1     #00ff00
show_as_track  no
description    this is a test track, description on 1 line
ITCC0001      stage4
ITCC0002      stage4
ITCC0003      stage3
ITCC0004      stage3
ITCC0005      stage4
  
```

Figure 18: Uploading a track .

1. If you intend to create a track with a limited number of groups, an easier way is provided through the user interface. We will try that now: click the back button of your browser to return to Figure 12. By default the Custom Track Window (Figure 19) is set to build a track based on a defined number of groups. Underneath you can adjust the number of groups, now change the number to 3 groups and click the Submit button.
2. In the next window a convenient overview of all annotation parameters and their values is available, with check boxes to assign samples with specific annotation values to one of the three groups. In this example we divide the samples by their INSS classification values in 3 groups: the low grade(1,2,3) vs high grade(4) vs special (4s) tumor types. 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 19). 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.

Define your Categories				
NO	G1	G2	G3	Show as col?
age_year	<input type="checkbox"/> 0(37)	<input type="checkbox"/> 0(37)	<input type="checkbox"/> 0(37)	<input type="checkbox"/>
	<input type="checkbox"/> 1(20)	<input type="checkbox"/> 1(20)	<input type="checkbox"/> 1(20)	
	<input type="checkbox"/> 13(1)	<input type="checkbox"/> 13(1)	<input type="checkbox"/> 13(1)	
	<input type="checkbox"/> 2(16)	<input type="checkbox"/> 2(16)	<input type="checkbox"/> 2(16)	
	<input type="checkbox"/> 3(5)	<input type="checkbox"/> 3(5)	<input type="checkbox"/> 3(5)	
	<input type="checkbox"/> 4(4)	<input type="checkbox"/> 4(4)	<input type="checkbox"/> 4(4)	
	<input type="checkbox"/> 5(1)	<input type="checkbox"/> 5(1)	<input type="checkbox"/> 5(1)	
	<input type="checkbox"/> 6(2)	<input type="checkbox"/> 6(2)	<input type="checkbox"/> 6(2)	
	<input type="checkbox"/> 7(1)	<input type="checkbox"/> 7(1)	<input type="checkbox"/> 7(1)	
	<input type="checkbox"/> 8(1)	<input type="checkbox"/> 8(1)	<input type="checkbox"/> 8(1)	
agegroup	<input type="checkbox"/> <=18_months(48)	<input type="checkbox"/> <=18_months(48)	<input type="checkbox"/> <=18_months(48)	<input type="checkbox"/>
	<input type="checkbox"/> >18_months(40)	<input type="checkbox"/> >18_months(40)	<input type="checkbox"/> >18_months(40)	
alive	<input type="checkbox"/> no(33)	<input type="checkbox"/> no(33)	<input type="checkbox"/> no(33)	<input type="checkbox"/>
	<input type="checkbox"/> yes(55)	<input type="checkbox"/> yes(55)	<input type="checkbox"/> yes(55)	
death_cause	<input type="checkbox"/> nd(55)	<input type="checkbox"/> nd(55)	<input type="checkbox"/> nd(55)	<input type="checkbox"/>
	<input type="checkbox"/> toxic(3)	<input type="checkbox"/> toxic(3)	<input type="checkbox"/> toxic(3)	
	<input type="checkbox"/> tumor(30)	<input type="checkbox"/> tumor(30)	<input type="checkbox"/> tumor(30)	
gender	<input type="checkbox"/> female(35)	<input type="checkbox"/> female(35)	<input type="checkbox"/> female(35)	<input type="checkbox"/>
	<input type="checkbox"/> male(53)	<input type="checkbox"/> male(53)	<input type="checkbox"/> male(53)	
histology	<input type="checkbox"/> nb(88)	<input type="checkbox"/> nb(88)	<input type="checkbox"/> nb(88)	<input type="checkbox"/>
id				<input type="checkbox"/>
inss	<input checked="" type="checkbox"/> st1(8)	<input type="checkbox"/> st1(8)	<input type="checkbox"/> st1(8)	<input checked="" type="checkbox"/>
	<input checked="" type="checkbox"/> st2(15)	<input type="checkbox"/> st2(15)	<input type="checkbox"/> st2(15)	
	<input checked="" type="checkbox"/> st3(13)	<input type="checkbox"/> st3(13)	<input type="checkbox"/> st3(13)	
	<input type="checkbox"/> st4(40)	<input checked="" type="checkbox"/> st4(40)	<input type="checkbox"/> st4(40)	
	<input type="checkbox"/> st4s(12)	<input type="checkbox"/> st4s(12)	<input checked="" type="checkbox"/> st4s(12)	
mycn_amp	<input type="checkbox"/> no(72)	<input type="checkbox"/> no(72)	<input type="checkbox"/> no(72)	<input type="checkbox"/>
	<input type="checkbox"/> yes(16)	<input type="checkbox"/> yes(16)	<input type="checkbox"/> yes(16)	
nti_event_overall	<input type="checkbox"/> no(58)	<input type="checkbox"/> no(58)	<input type="checkbox"/> no(58)	<input type="checkbox"/>
	<input type="checkbox"/> yes(30)	<input type="checkbox"/> yes(30)	<input type="checkbox"/> yes(30)	
nti_event_progrfree	<input type="checkbox"/> no(53)	<input type="checkbox"/> no(53)	<input type="checkbox"/> no(53)	<input type="checkbox"/>
	<input type="checkbox"/> yes(35)	<input type="checkbox"/> yes(35)	<input type="checkbox"/> yes(35)	
nti_surv_overall				<input type="checkbox"/>
nti_surv_progrfree				<input type="checkbox"/>
r2_label				<input type="checkbox"/>
recurrence_or_progression	<input type="checkbox"/> nd(53)	<input type="checkbox"/> nd(53)	<input type="checkbox"/> nd(53)	<input type="checkbox"/>
	<input type="checkbox"/> yes(35)	<input type="checkbox"/> yes(35)	<input type="checkbox"/> yes(35)	
samplenames				<input type="checkbox"/>
u-changedgeneskmeans	<input type="checkbox"/> group 0(48)	<input type="checkbox"/> group 0(48)	<input type="checkbox"/> group 0(48)	<input type="checkbox"/>
	<input type="checkbox"/> group 1(40)	<input type="checkbox"/> group 1(40)	<input type="checkbox"/> group 1(40)	

Figure 19: Preselection to make new tracks from existing annotation.

1. Click “next”, all samples appear in a table with check boxes to assign them individually to the appropriate group. Scroll down to adapt the visual characteristics of these groups. Names have been adapted, ‘show track’ is set to yes, the track is set to be stored as 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, fill in a gene of choice (e.g. MYCN) in box 3 and click next twice to see how the data of a gene will be plotted using the new tracks.



Figure 20: Setting the custom track properties.

- Another frequently used approach is to make a track based on bins of gene expression 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. Of course, it is possible to change the names of the bins. Click next to tell R2 to draw the track, change the colors of the track bins and save the track.

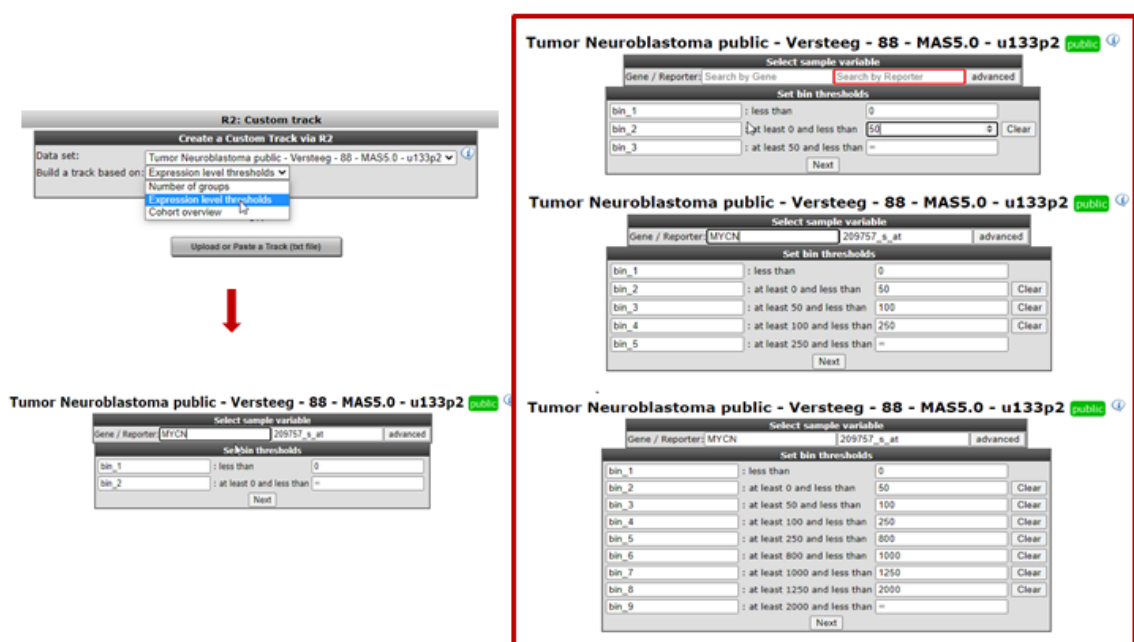
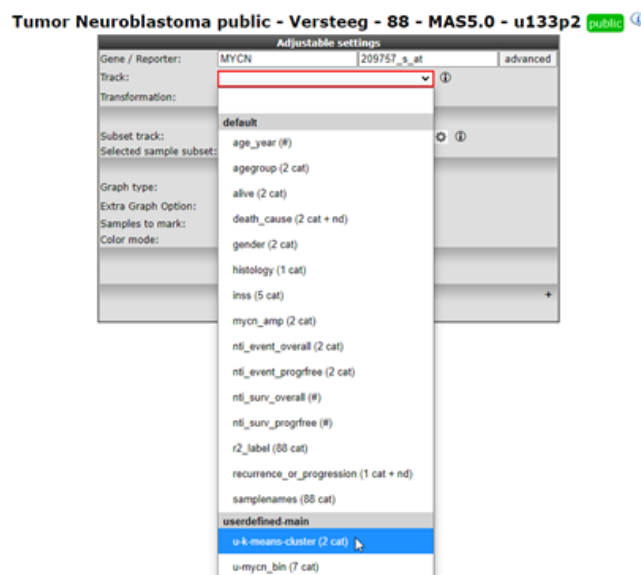
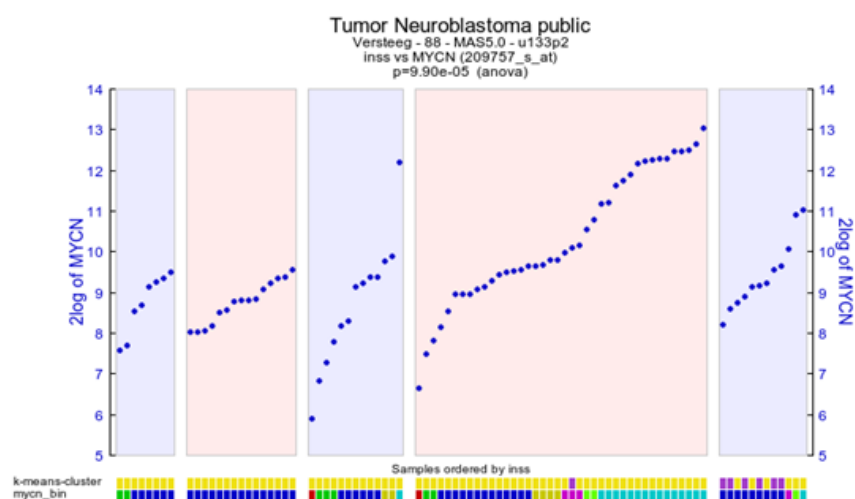


Figure 21: Grouping samples for a track based on gene expression

3. Select View a gene in groups in Field 3 of the main page, Click 'next'. Type MYCN in the gene/reporter box, All the tracks created in the Custom Track manager are available for selection in the 'userdefine'-main group of tracks, as a group separator choose inss stage, in this example we have switched off all the default tracks and selected the two custom tracks we have created.

**Figure 22: Select the Track created in the Custom track manager; u-lowgradvs4vs4s**

4. The expression of MYCN is plotted in the different groups of the inss stage (Figure 26). Extra Graph Option has been set to Track and Gene Sort. Note that the other track of mycn_bin groups the highest bin groups corresponds with the stage 4 group. Which also holds for the lower bin groups which are aligned with stage stage4s, stage 4s is known to have a better prognoses. There is also overlap with the custom created k-means generated tracks.

**Figure 22: Tracks created are visualized underneath the graph**

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 samples when generating a heatmap. Make sure you already have a personal custom track (not a temporary track, 24h). Select 'Manage default Tracks' from the User Options > Tracks menu (Figure 18) 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.

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

1. Cooperation is of great importance in scientific research. You may want to share the tracks created above with other people in your group. For this reason R2 features the Communities feature. Communities are different from user groups, which is important to remember. User groups are granting a user access to datasets and their associated annotation, or may unlock restricted functionalities within the R2 platform. On the other hand, communities are a way by which any user can share grouping variables (tracks), lists of genes (gene categories), megasampler presets or genome browser views with any (group of) other R2 user(s). A user can generate multiple communities and invite other users to share such feature with.

Creating a community is done by clicking 'Community' in the 'User Options' menu (Figure 23).

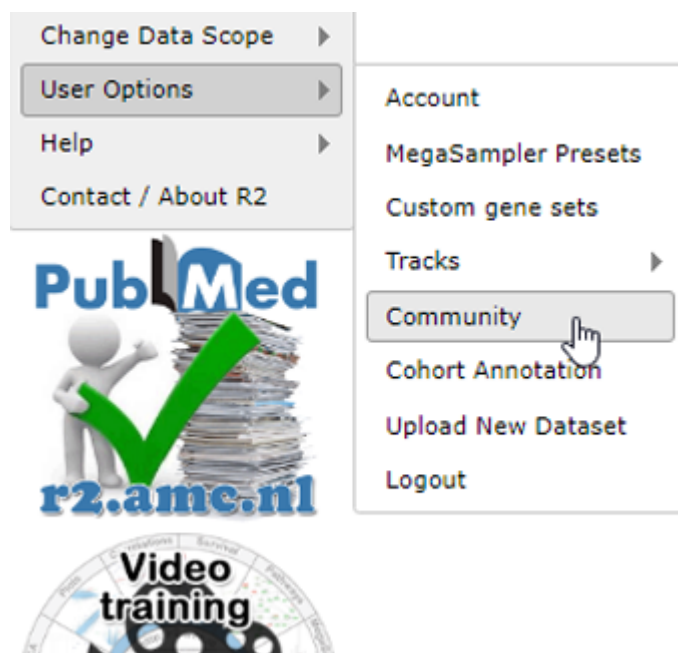


Figure 23: Community in the User Options menu

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 24).



Figure 24: Starting a community

3. In the Community window a name has to be set and a short description for people invited as members for this group (Figure 25). Through a community you can share your own Gene sets, Tracks and Settings.



Figure 25: Setting the Community group name and description.

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 User Options menu (Figure 27). The TestGroup has been created (next to the already existing MyTestGroup for this user). Click the link to start adding users (Figure 30).

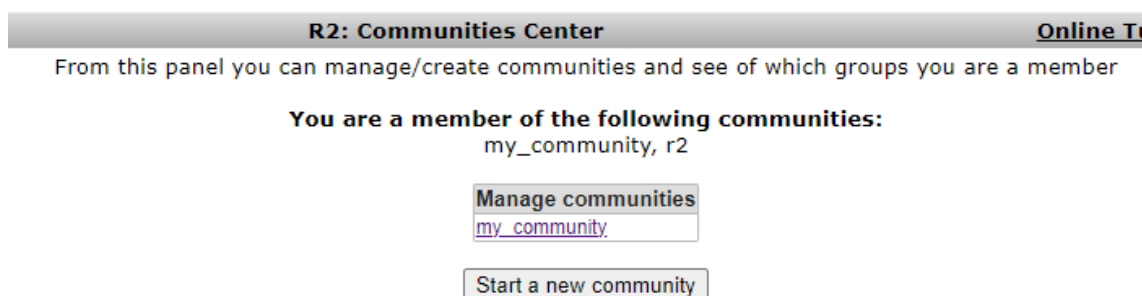


Figure 26: The available Communities for this user

5. Once the community is generated, you can also manage the community from the communities center.

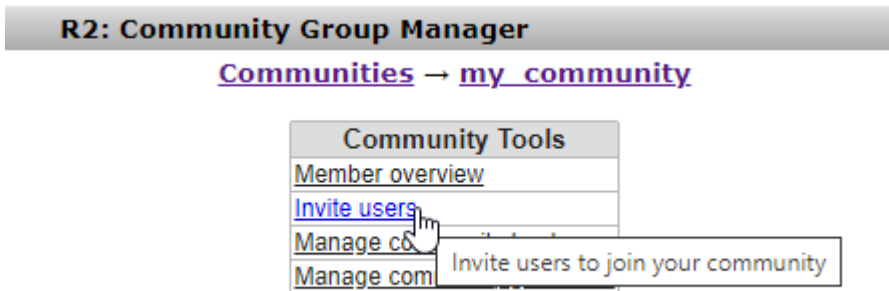


Figure 27: Managing of the community

6. You have to add users; You can invite R2 users to your group by sending them an invitation link. we will create a link with 'create invitation'. For your own personal recollection you can specify a target recipient. Here we have used "my invitee".

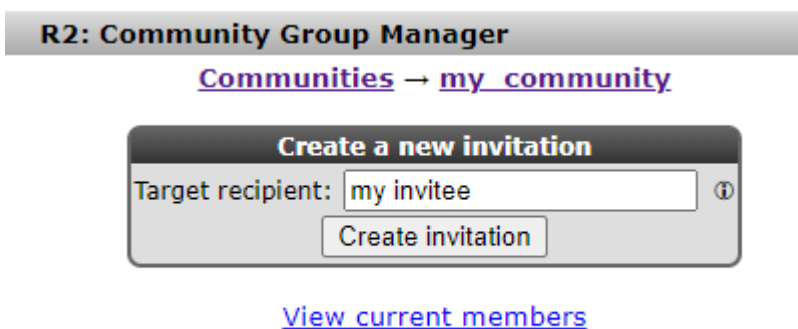


Figure 28: Creation of an invitation for your community

7. Once you create the invitation, then a message will be provided that you can send to the person you want to invite to your group. This you can do by using the template in an email or by sharing the provided link via other means. Please do note that invitations to join your group are valid for a limited time only. How long an invitation is still valid can always be checked in the 'invite users' option. you can always revisit the invitation or extend it using the 'view/renew' button.

R2: Community Group Manager [Online Tutorial](#)

[Communities](#) → [my_community](#)

Invitation created!

Copy the invitation below and send it to "my invitee" (e.g. through e-mail or instant messaging).
 Note: this invitation can only be accepted by one person and will expire after a week.

You are invited to collaborate in the R2 community "my_community".
 Sign in to your R2 account and copy the following web address into your browser to accept the invitation:
https://server2.your-r2.org/r2/main.cgi?operation=community_invite&token=60X13j_L3F4150

Create a new invitation

Target recipient: ⓘ

Pending invitations

	target	days left	view/renew
<input type="checkbox"/>	my invitee	7.0	<input type="button" value="view/renew"/>

[View current members](#)

Figure 29: R2 invited users option

8. The user that has received a link, needs to be logged into their account and can then visit the provided link. This will allow the user to accept the invitation.

R2: Community Invitation

You are invited to collaborate in the R2 community "**my_community**"! ⓘ

The owner(s) of the community will be able to see your [name and email address](#)

[View communities](#)

Figure 30: The invited user will see a message where the user can accept the membership of the group

Once the invitation is accepted by the user, you can see this within your community as well.

R2: Community Invitation

Invitation accepted. You are now a member of the community "**my_community**".

[View communities](#)

Figure 31: Once accepted, the user is a member of the group

9. When the invitation has been accepted the user is available in this community. By default an invitee will become a member, who can only see what you are sharing via your community. Next to the label that you have generated yourself, you will also see the username and email address of the member that accepted the invitation (also to make sure the intended user became member). For any of the members, there is also the possibility to increase the rights of a member, by making the user 'content manager', or (co-)owner.

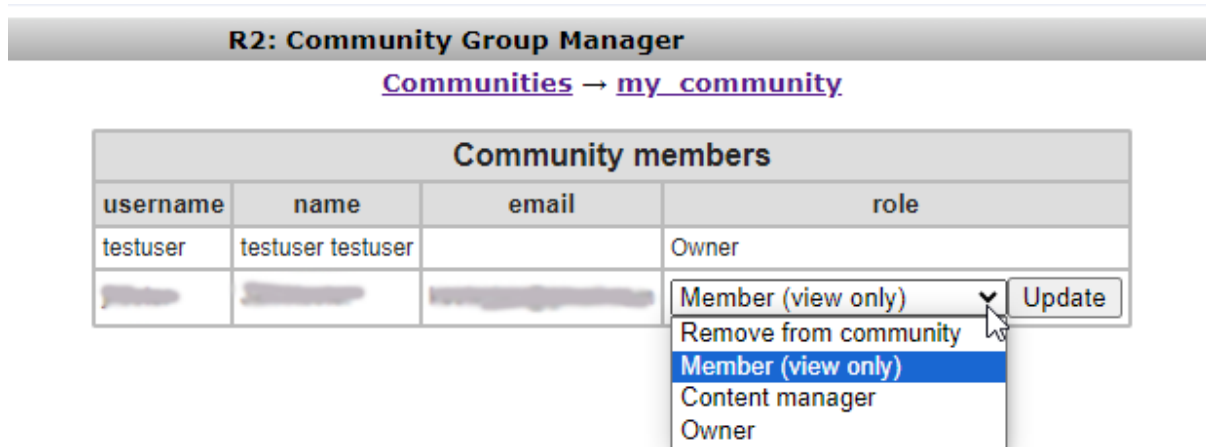


Figure 32: The user is now visible as a member of the group.

1. When you will add a custom gene set, track or preset the next time, it will be possible to make this available to any of the communities that are yours, or where you have been granted access as content manager.

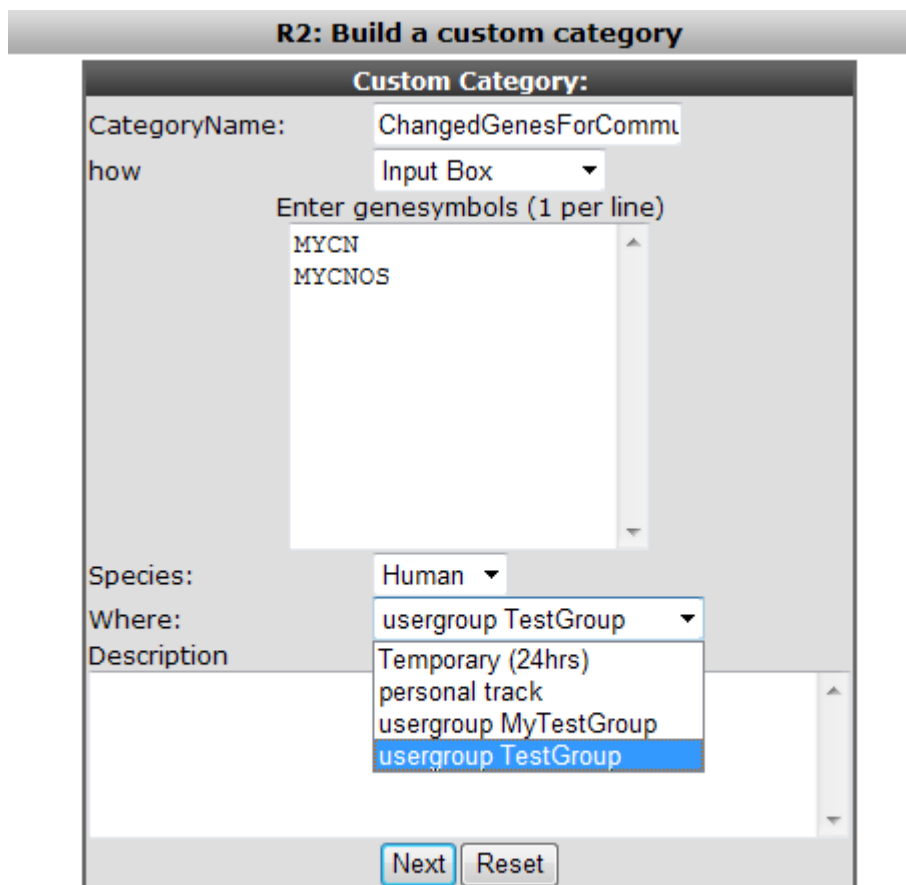


Figure 33: As an example here the creation of a gene set and the assignment to a Community.

2. 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. **my invitee**, as a member of this group, can manage the tracks that have been shared with this user via the default track manager.

23.8 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@amsterdamumc.nl).

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

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

24.1 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).

24.2 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'.

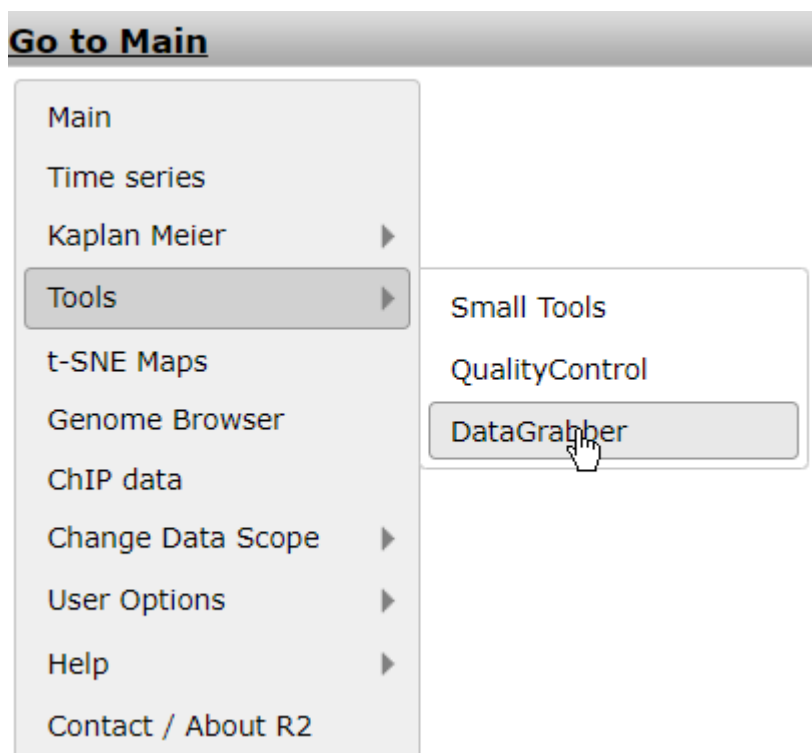


Figure 1: In the main menu “Data Grabber

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’.

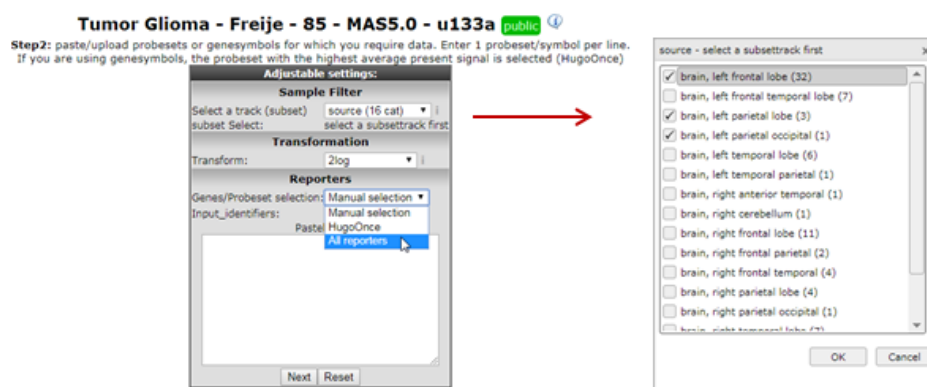


Figure 2: Filter options for exporting data.

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).

24.3 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.

**Figure 3: Continuing your analysis.**

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.

24.4 Final remarks / future directions

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

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

25.1 Scope

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

25.2 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.

25.3 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 administrators of the R2 platform can add new datasets, since they understand the R2 platform architecture and can supervise / guide the upload procedure.

25.4 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>

25.5 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.

25.6 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 sent 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.

25.7 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 sent to us by www.wettransfer.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 1st column. The 1st 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 1st 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.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	probeid	ITCC0001	ITCC0002	ITCC0003	ITCC0004	ITCC0005	ITCC0006	ITCC0007	ITCC0008	ITCC0009	ITCC0010	ITCC0011	ITCC0012	ITCC0013	ITCC0014	ITCC0015
2	1007_at	430.8	226.1	130.6	75	54.9	195.5	124.8	221.5	82.1	57.5	67.1	174	55.8	143.6	110.2
3	1053_at	79.5	95.7	178.8	185.8	144.7	113.1	150.2	157.5	138.4	140.7	198.1	123.7	254.8	130.4	161.5
4	117_at	65.9	14.9	1.8	28.7	4.7	35.2	3	25.5	39.3	17.3	23.7	39.7	4.1	29.8	64.1
5	121_at	203.3	79.4	74.1	60.3	79.7	54.5	60.1	113.3	55.6	85.8	80.1	81.9	72.7	91.9	112.1
6	1255_at	5	129.9	177.8	291.1	192.4	20	235.4	5.6	54.9	56.6	63.8	61.2	300	41.1	40.8
7	1294_at	114.6	57.1	40.1	37.1	44.2	50.9	40.4	28.3	56.3	51.6	38.6	127.8	44	79.4	83.8
8	1316_at	113	92.7	115.1	106.5	63.5	103.6	121.6	97.8	26.1	33.2	27	33.5	24	17.3	52.9
9	1320_at	12.1	17.9	1.7	1	2.2	11.2	17.2	16.1	33.2	58.1	27.4	14.3	46.6	37.9	40.4
10	1405_at	313.8	135.4	26.1	89.8	164.4	223.2	81	15.7	64.1	29	13	406.7	28.9	149.7	233.3
11	1431_at	18.8	13.7	16.2	8.6	8.2	66	10.3	15.1	33	14.2	79.8	12.5	15.1	7.4	28.9
12	1438_at	2.4	51.9	5.2	9.4	22	6.2	8.3	38.6	69.7	193.6	31.5	107.6	59.1	100.6	120.6
13	1487_at	94	88.9	96	52.1	98.3	67.8	92	91.4	63	99.1	160.7	106.4	54.8	84.4	101.6
14	1494_at	30	18.8	33	49	34	33.5	44.9	35.3	26.1	22.9	19.8	18.7	19.8	17	29.1
15	152256_a_at	33.4	55.5	57.9	53.1	52	60.2	63.7	117	182.2	79.1	185.9	187.1	137.1	143.8	251.8
16	1552287_a_at	109.2	78.6	70	83.8	62.7	97.1	73.4	245.7	126.6	113.4	151.9	117.9	112.2	175.1	159.7
17	1552298_at	31.2	8.3	5.3	20.3	7	20.7	35	22.7	22.6	12.2	11.5	24.5	27.3	18.7	16.5
18	1553361_at	45.7	6.6	6.1	14.1	13.7	6.6	18.8	15.7	18.7	33.8	8.8	36.8	17.8	12.8	34.8

Figure1: Example

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.

25.8 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.

25.9 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 1st 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.

samplename	AgeGroup	AgeGroup2	Alive	AlivePlus	Onco-code	Hospital	Age_days	Age_months	event_overall	event_patient	time_point
ITCC0001	<18_months	<1Y	YES	>=5Y	nd	EK2	57	2 no	no	7876	7876
ITCC0002	<18_months	<1Y	YES	>=5Y	nd	VU	43	1 no	no	4914	4914
ITCC0003	<18_months	<1Y	YES	>=5Y	nd	RADBOUT	116	4 no	yes	5210	97
ITCC0004	<18_months	1-2Y	YES	>=5Y	nd	EK2	428	24 no	no	4799	4799
ITCC0005	>18_months	1-2Y	YES	>=5Y	nd	WKZ	742	24 no	no	5750	5750
ITCC0006	>18_months	>2Y	NO	NO	Tumor	RADBOUT	2871	94 yes	yes	322	141
ITCC0007	<18_months	1-2Y	NO	NO	Tumor	EK2	425	14 yes	yes	157	98
ITCC0008	<18_months	<1Y	YES	>=5Y	nd	EK2	280	9 no	no	6904	6904
ITCC0009	<18_months	1-2Y	YES	>=5Y	nd	EK2	391	13 no	yes	7146	161
ITCC0010	<18_months	<1Y	YES	>=5Y	nd	EK2	287	9 no	no	7168	7168
ITCC0011	<18_months	<1Y	NO	NO	Toxic	RADBOUT	182	6 no	no	166	166
ITCC0012	>18_months	>2Y	YES	>=5Y	nd	SKZ	786	26 no	no	5926	5926

Figure2: 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.

#HTRACKS	istrack	isinfo	visible	colors	describe
age_year	no	yes	TRUE		Age at which the patient was diagnosed
AgeDiagnosis	no	yes	TRUE		Age at which the patient was diagnosed
AgeDiagnosis_months	no	yes	TRUE		Age in months at which the patient was diagnosed
AgeGroup	yes	yes	TRUE	ND:E6E6E6<1.00C800>=>1:C80000	Age group in which the patient belongs <1 year vs >=1 year at diagnosis
AgeGroup2	no	yes	TRUE		Age group in which the patient belongs <1 year, 1-2 year, >2 year at diagnosis
Alive	no	no	TRUE	ND:E6E6E6:YES:00C800:ND:C80000	Survival status of the patient, alive or diseased
AlivePlus	yes	yes	TRUE	ND:E6E6E6>=5Y:00C800:2-5Y:66C866:0-2Y:99C899:ND:C80000	Survival status of the patient subdivided in 3 stages of alive
batch	no	yes	TRUE		Affymetrix batch of profiling

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 is not encountered. Finally, you may describe the contents of a track.

25.9.1 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 is 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.

#H:samplenames	EventFreeSurvival	Event
ITCC0600	1341	no
ITCC0601	1236	no
ITCC0602	1036	yes
ITCC0604	886	no
ITCC0607	276	no
ITCC0608	179	no
ITCC0609	614	no
ITCC0611	393	yes

Figure 3: Example3

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.

samplenames	condition	exp	time	r2_ts_series	r2_ts_profile	r2_ts_timepoint
nb420	control	c1	8	control	control_1	8
nb422	control	c1	24	control	control_1	24
nb424	control	c1	120	control	control_1	120
nb429	control	c2	8	control	control_2	8
nb431	control	c2	24	control	control_2	24
nb433	control	c2	120	control	control_2	120
nb010	exp	e1	0	exp	exp_1	0
nb011	exp	e1	8	exp	exp_1	8
nb012	exp	e1	24	exp	exp_1	24
nb013	exp	e1	120	exp	exp_1	120
nb418	exp	e2	0	control,exp	exp_2,control_1	0
nb419	exp	e2	8	exp	exp_2	8
nb421	exp	e2	24	exp	exp_2	24
nb423	exp	e2	120	exp	exp_2	120

Figure : Example 4

25.10 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 use 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).

Graphs: Adjustable Settings menu versus Repsonsic Settings

Adapt and export graphs in R2

R2 often offers the user settings to adapt or enhance the looks of a visualization. Adaptations could be purely aesthetical preferences, but often they are functional, for instance to overlay the plot with extra layers of information. Traditionally, the user can control such settings with an *Adjustable Settings* menu that is placed underneath many graphs in the platform. Adjustments in the settings of these graphs take effect after the button at the bottom of the menu is hit. Increasingly, the platform offers responsive controls as well: when a setting is adjusted, the changes in the menu are directly visible in the graph without the need to click any submit button. Certain graph elements can even be adjusted directly within the plot interface. Plots that offer responsive functionalities can be recognized by a settings wheel in the upper left corner of the graph. Some options may not be immediately evident to users, however, they facilitate swift adaptability. Therefore, this chapter aims to show a few examples as showcases of how to tweak the plots with the different setting types.

26.1 Example settings: color the sample maps

In the Sample maps module, underneath the scatter plots we can find the R2-wide grey Adjustable Settings menu and a Color settings menu. Options listed in these menus require the user to press the “Set [setting name]” button in order for the requested changes to the specific setting to take effect. For instance, here the graph colors can be set to the colors of a track, or they can be set to the expression levels of a gene. This setting requires the user to click on the button “Set colors” in order to take effect. Other settings can be adjusted directly in the plot itself. When your mouse hovers over the legend categories, an information pop up tells you that the respective subgroup of samples can be toggled off and on in the plot with a click on the legend box. Also, a click on the legend title “histology” will invert the selection. If all groups are shown, which is the default situation when you color a map with a track, then the invert option will deselect all groups leaving the plot blank. One more click allows you to subsequently single out a subgroup quickly. With one more click you then toggle one or a few groups on.

Figure 1: Submit buttons (color by track) vs responsive settings (toggle legend subgroups)

In the animation below, we show several responsive settings that you can find in various graphs in R2. In responsive graphs, samples can be marked with a click on the sample in the graph; you can zoom in and out with the scrolling wheel of your mouse and the graph can be repositioned by dragging the plot while holding the right mouse button.

Figure 2: Zoom in / out and reposition the plot, mark a sample

Some graphs offer additional responsive options outside the graph. Below the animation shows several other options in the Sample maps module.

Figure 3: Adapt dot size / add border/ copy paste to powerpoint

Concepts of R2: did you know..?

Did you know. . . that throughout the R2 manual many tips and tricks are provided in small blocks of text containing practical guidance and theoretical background for the analysis at hand? These explanatory blocks start with the phrase “**Did you know..**”. This chapter aims to centralize the information of the most essential concepts and settings of R2, such that it is easy to integrate these options and understandings in your own analysis.

Jump to one of the sections:

- *Statistical terms used in r2 explained?*
- *Statistical tests in Differential expressed genes?*
- *Often used settings for analyses*
 - *Statistics in analyses*
 - *When to use multiple testing*
 - *Hugo Once*
 - *Use subsets of genes with Gene Filters*
 - *Transform expression levels*
- *Core concepts of R2*
 - *Annotation of samples with Tracks*
 - *Gene subsets a.k.a Categories*
 - *Summarize the behavior of a list of genes with a Signature Scores*

27.1 R and p-values

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 $R \sim 0$ 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.

t-statistic: used to calculate the t-statistic for testing the significance of a correlation coefficient (R) between two variables. Here's a breakdown of the components:

- **t**: This is the t-statistic, a measure of how many standard deviations a data point is from the mean. In the context of correlation, it is used to test whether the observed correlation is significantly different from zero.
- **R**: This represents the correlation coefficient between the two variables. It measures the strength and direction of the linear relationship between them. R ranges from -1 to 1, where -1 indicates a perfect negative correlation, 1 indicates a perfect positive correlation, and 0 indicates no linear correlation.
- **r**: This is the sample correlation coefficient, representing the strength and direction of the linear relationship in a sample of data.
- **n**: The number of pairs of data points in the sample.

The formula involves taking the correlation coefficient (R) and adjusting it based on the sample size (n). The denominator is essentially the standard error of the correlation coefficient. By dividing R by this standard error, the formula produces a t-statistic that can be compared to critical values from a t-distribution to determine the statistical significance of the correlation. If the absolute value of the t-statistic is large, it suggests that the correlation is likely to be statistically significant. The degrees of freedom for the t-distribution are given by (n-2) in this context. $t = R / \sqrt{(1-r^2)/(n-2)}$, where R is the correlation value and n is the number of samples. Distribution measure is approximately as t with n-2 degrees of freedom.

27.2 Statistical tests

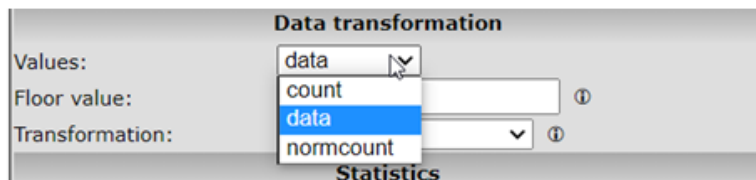
In the modules Differential expression between (two) groups, various statistical tests can be utilized to find the differentially expressed genes between two or more groups.

- two group differential expression.
 - T-test. For normal distributed and continuous data, see ANOVA explanation below.
 - Mann-Whitney-test. The Mann-Whitney U test is used to compare differences between two independent groups when the dependent variable is either ordinal or continuous, but not normally necessarily distributed, such as a skewed distribution.
 - Log2 foldchange: How is the log2 foldchange calculated: $\text{Log2}(\text{untransformed}(\text{group1}) / \text{untransformed}(\text{grp2}))$
 - Limma: Over the past decade, limma has been a popular choice for gene discovery through differential expression analyses of microarray and high-throughput PCR data, more information can be found here [Limma:NCBI](#). [Limma:BioC](#).
 - DESeq2 algorithm: Differential expression analysis based on the Negative Binomial (a.k.a. Gamma-Poisson) distribution. The algorithm uses raw integer read counts for control and e.g. treatment conditions. [DESeq2:BioC](#).
 - Data slot (DESeq2) slots in r2: When using a dataset with DESeq2_rlog / vst normalized, you will find three different slots. The count, data and normcount slot. Within the dataset,
 - * The **'count'** contains the read counts and is used by default for DESeq2 analysis within R2. These are non-normalized counts, not intended for visualization of samples in a data set. However, it can sometimes be informative to be able to see what was the true source data.
 - * The **'normcount'** are the normalized read counts, those are obtained with the DESeq2 package and adjusts for the library size between the samples. This normalization procedure scales the counts by dividing them by size factors, which are estimated based on the median of the ratio of observed counts to the geometric mean for each sample.
 - * While **'data'** refers to the 'standard' normalized data that R2 will use by default for visualizations. This slot will represent itself as the normalization you also see in the dataset title. For DESeq2, the preferred choice for normalized data is rlog values. The DESeq2 aims to stabilize the variance on the raw counts applying a Bayesian Shrink estimator. Some notes about the rlog (data slot), actually what rlog does is adapt the very low gene expression values by elevating them a bit. These very low values tend to result in higher fold changes and occasionally also survive pvalue cutoffs. By elevating those, you reduce this problem, while you preserve the differential expressions at

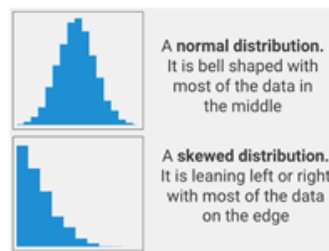
the more highly expressed genes. The calculation of rlog is pretty CPU intensive. for bigger data sets (30-40+) we can also use vst, variance stabilizing transformation, which will add the same value to all measurements, to avoid the low expression values issues.

- * Additional '**slots**' can be defined in some data sets. We typically obtain data matrices from public sources, which allow you to replicate the authors findings in the easiest way. Other ways of representing the data can then also be defined. For example, some also add TPM values to the DESeq2 generated data. These can also be propagated in R2 as separate slots. The concept of using slots is not confined to the DESeq2 data sets in R2, but can be in use wherever it can be useful to represent the data in multiple ways. When multiple slots are available, then there will be a dropdown in the '**Data transformation**' section named '**Values**', where you can select the slot to use.

R2 uses DESeq2_rlog and DESeq2_vst in the dataset name annotation meaning that the selected dataset has been processed with the DESeq2 algorithm but depending on the size of the dataset the rlog value is used or the vst values. Generating the rlog values is a time consuming process so we use a cut-off for $n > \sim 30$ to take the vst values in the dataslot. Both methods have their similarities and some differences.



Dataslots.



Distribution types.

- Differential expression between groups
 - ANOVA: This **AN**alysis **Of** **V**ariance 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.
 - Kruskal-Wallis: You have non-parametric data and more than 2 groups, in this test the data is replaced by their rank position. In case you want to identify the difference between specific groups such as group 1 vs 2, 2 vs 3 you should use the Mann-Whitney test for two group comparisons.
 - Pair-wise tests: T-test is performed for all pair-wise group combinations

27.3 Settings for analyses

R2 provides a range of settings for each analysis. Here we explain the background of some recurring concepts and we provide guidance for the use of the different options.

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: Often we are testing a lot of genes in an analysis such as differential expression between groups (Chapter 6). If this is true for your analysis, you have to correct for multiple testing. Why? Let's look at an 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 $p\text{-value} < 0.05$, 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 9 or 10 heads 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 $(1 - 0.0107)^{100} \sim 0.34$. 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) α , then the Bonferroni correction would be to test each of the individual tests at a significance level of α/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. In R2, the FDR is selected by default, the cut-off value is 0.01; note that the FDR test is also known as the Benjamin-Hochberg test.

Hugo Once (hugoonce): For most analyses genes should only be reported once in a dataset. R2 uses an algorithm called Hugoonce 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 MAS5.0 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.

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.

Transform: Converting expression levels with the "transform" option can help you to gain additional insight.

There are several data transformations available. Note that within R2 most of the data in the resources is stored untransformed, such that you can apply all the transformations within the tools itself.

When to choose which transformation?

- “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.
- “log2 zscore”: 2log transformed data, centered around the average and expressed as the number of standard deviations from the average. This is the most common option.
- “zscore”: 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).
- “zcore_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.

27.4 Core concepts of R2

Tracks: In R2 the samples of a dataset can be annotated with e.g. clinical data or molecular biology parameters, each group of annotated data is called a “Track” in R2. These tracks can be used to filter datasets, to compare groups of samples, to color scatter plots of samples with meta information, or to correlate genomics patterns in your data to e.g. different phenotypes or demographic characteristics. Tracks are sample, and therefore, dataset specific. Most datasets contain default tracks when they are uploaded in R2, but every user can define their own custom made tracks for any available dataset, disable/adapt the settings for default tracks or store their analysis outcome as a track to use in further track-related analyses. User defined tracks are privately visible, but can be shared in user defined R2 communities with other R2 users. To play around with the concept, check out the tutorial on the [One Gene View analysis](#), the chapter on [Annotation Analyses](#), or find out how to make your own tracks in chapter [Adapt R2 to Your Needs](#).

Categories: Gene sets in R2 are called Categories. These categories enable the user to narrow down an analysis to a specific set of genes that is relevant to their biological question (see [Gene Filters](#) in the Settings section above). Users can define custom made Categories or store genes that resulted from an analysis for future use. Follow the tutorial [Using Signatures](#) for a better understanding of the possibilities of Categories; check out the section on Gene Categories in [Adapting R2 to your needs](#) to learn how to create your own Categories.

Signature Score: 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). We like to think of it as a “meta”-gene: a summary expression value of a group of genes for each sample of a dataset. In R2, such scores are automatically generated when one generates heatmaps via the “view a geneset” function. Because each sample of a dataset will be attributed with a value for the signature score, R2 enables the user to store this signature score as a Track for further analyses with the dataset. For a better understanding of Categories as well as Signature Scores, follow the tutorial [Using Signatures](#).