Amsterdam UMC
University Medical Centers

R2 Introduction Workshop

Finding causes in Neuroblastoma genomics data

Answers

Lieke Hoyng , on behalf of the R2 Team

Dept. CEMM | Location AMC

Amsterdam University Medical Centers (AUMC)

University of Amsterdam, the Netherlands

R2 Support: **r2-support@amsterdamumc.nl**

Jan Koster: **jankoster@amsterdamumc.nl**

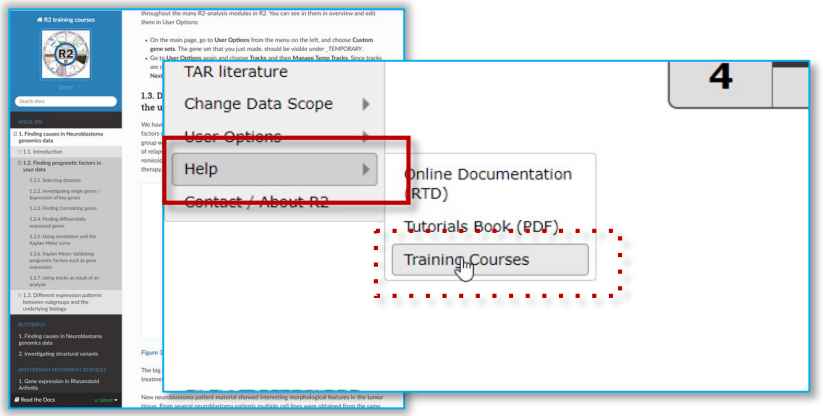
r2platform.com

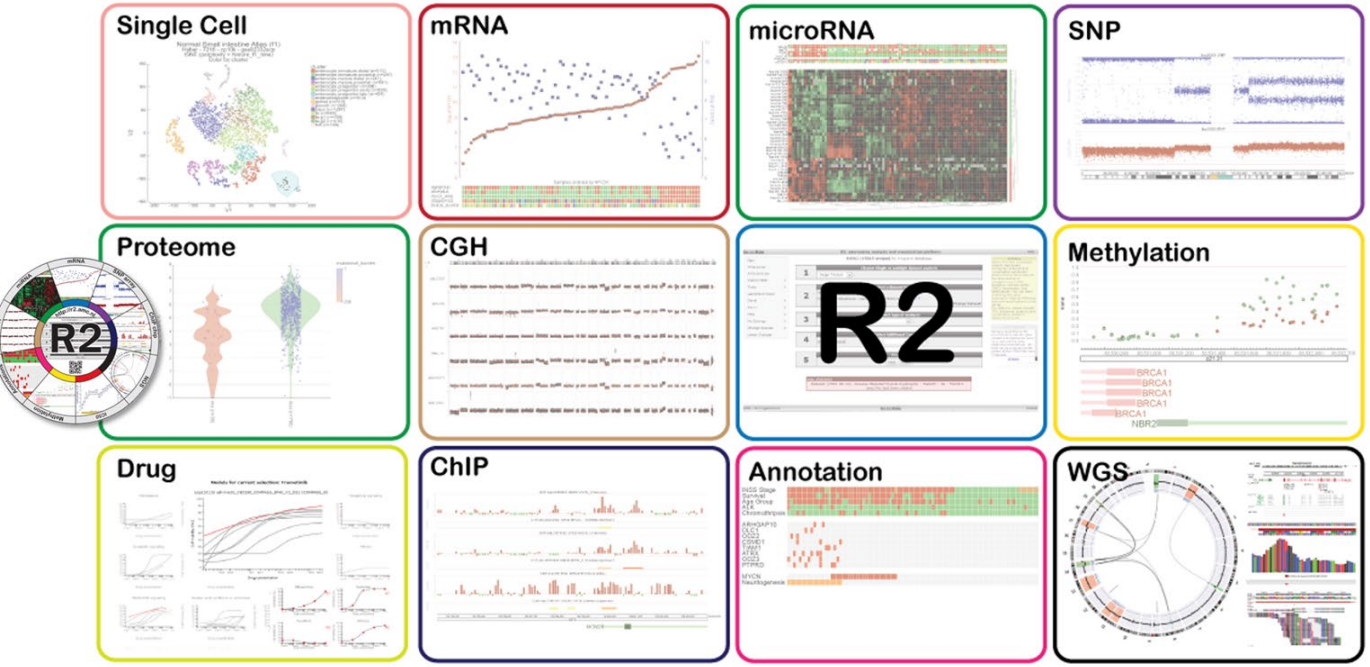
R2-Platform

Webbased Course Resource

- The course for now can be found at:

https://r2-training-courses.readthedocs.io/en/latest/R2IntroductionWorkshop2024_day1.html



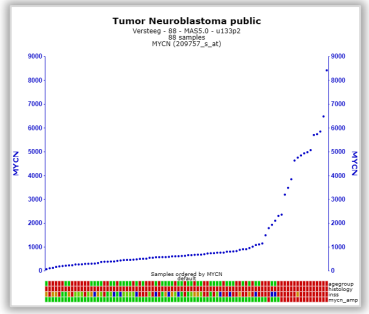


Contact us at r2-support@amsterdamumc.nl

- If this document can no longer be found at this location
- If you want your dataset or a dataset from a public resource uploaded to R2
- If you have any questions related to R2

1.2.2 Investigating single genes / Expression of key genes

Want to know more?
https://r2-tutorials.readthedocs.io/en/latest/One_Gene_View.html



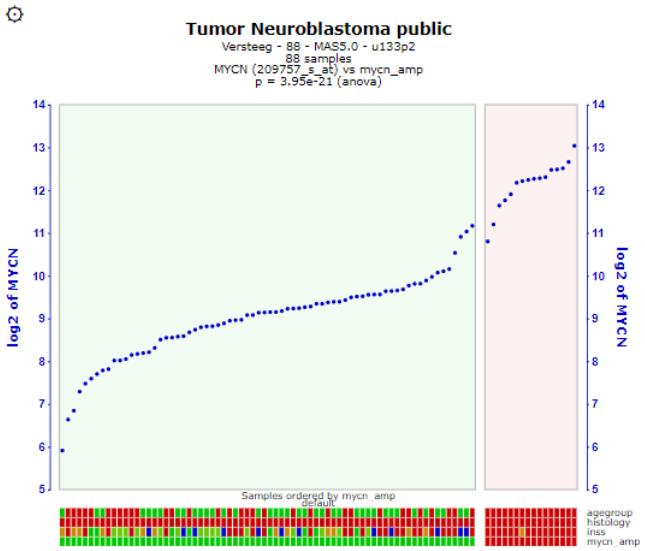
- **What is the relation between the track MYCN amplification and MYCN expression?**
  - Patients that were found to have clinically proven mycn amplification show high mycn expression, as you can see from the red colored mycn amp annotation that are positioned on the right side of the graph, with the highest values of mycn expression
  - In R2 the expression values in this module are depicted with a log2 transformation. This ensures a more compact view in which extreme values are less dominant and smaller values are better visible. It is good though, to realize that this transformation might create a skewed view. To see the untransformed values in the graph, change the *Transformation* setting underneath the plot to **None** and you will find the difference between the expression values of the mycn_amp:yes samples and the rest of the dataset even more apparent.
- **What do you notice about inss staging versus mycn amplification when you look at the annotation underneath the graph?**
  - All but one mycn_amp patients were diagnosed with high risk inss stage 4 (inss red color), one patient was diagnosed with inss stage 3 (orange color).
  - We also see that it is a *subgroup* of the stage 4 patients that was found to have a mycn amplification.

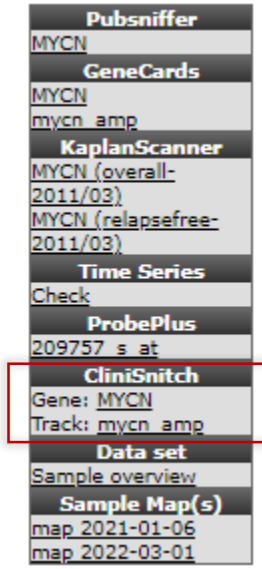
N.B. To see the values of samples for the tracks, you can either hover your mouse over dots in the plot, hover your mouse over the annotation squares underneath the plot or draw the legend by changing the setting to yes in the menu that opens with the settings wheel

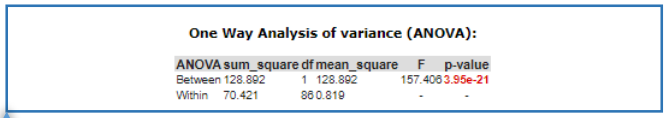
1.2.2 Investigating single genes vs tracks | description

Want to know more?

https://r2-tutorials.readthedocs.io/en/latest/One_Gene_View.html?highlight=anova#step-9-find-best-track-separation-with-clinisnitch





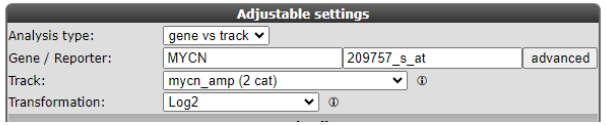


One Way Analysis of variance (ANOVA):

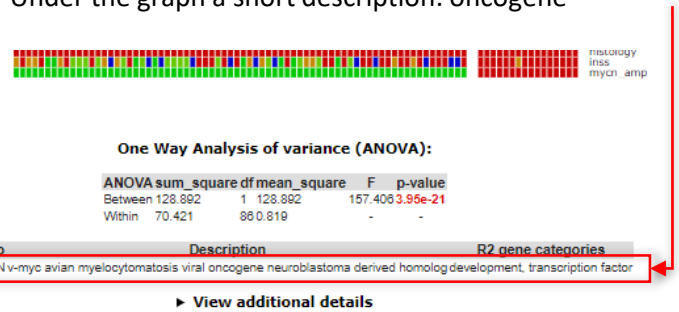
ANOVA	sum_square	df	mean_square	F	p-value
Between	128.892	1	128.892	157.406	3.95e-21
Within	70.421	86	0.819	-	-

N.B. To assess whether the averages of the mycn_amp track groups are statistically different from each other in mycn expression:

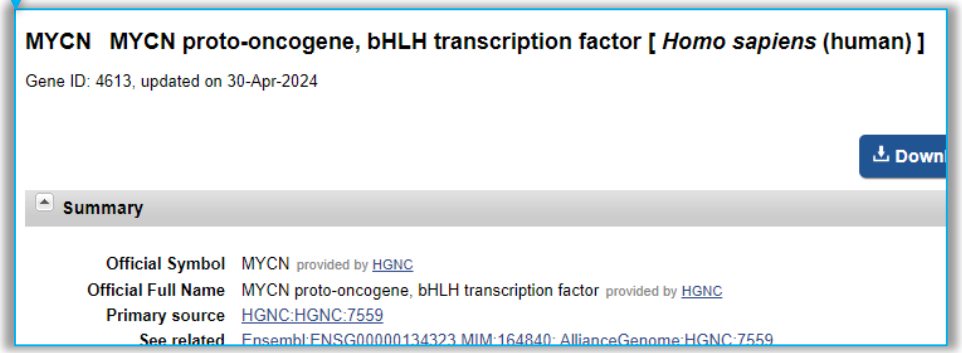
- Use **CliniSnitch** from the options on the right
  - https://r2-tutorials.readthedocs.io/en/latest/One_Gene_View.html#step-9-find-best-track-separation-with-clinisnitch
- Or you can change the Analysis type in **gene vs track**. You find the p-value of the association of the mycn expression values with the mycn_amp track values.



- **What is MYCN?**
  - Under the graph a short description: oncogene



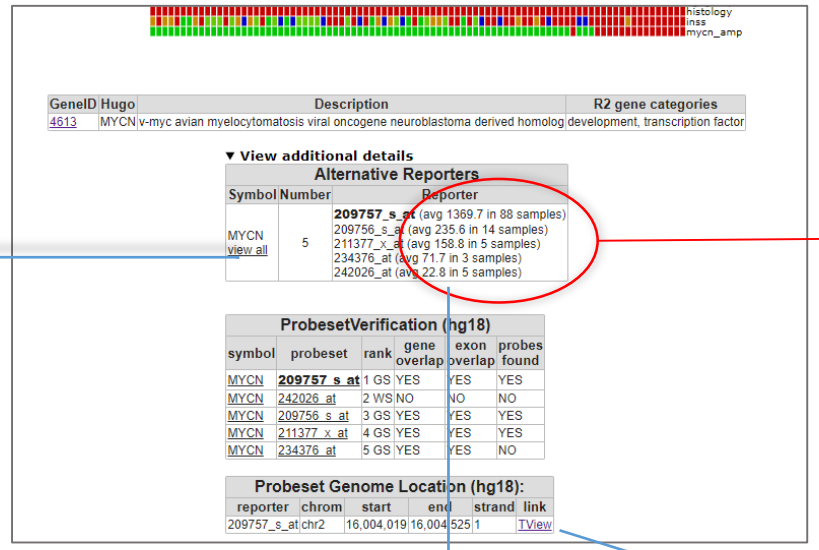
  - Link to National Center for Bioinformatics database:
    more in-depth description of the proto-oncogene MYCN.
    - N.B. A proto-oncogene proto-oncogene is a normal gene that plays a key role in the regulation of cell growth, differentiation, and survival. An oncogene is a mutated or overexpressed version of a proto-oncogene that drives uncontrolled cell proliferation and contributes to the development of cancer.

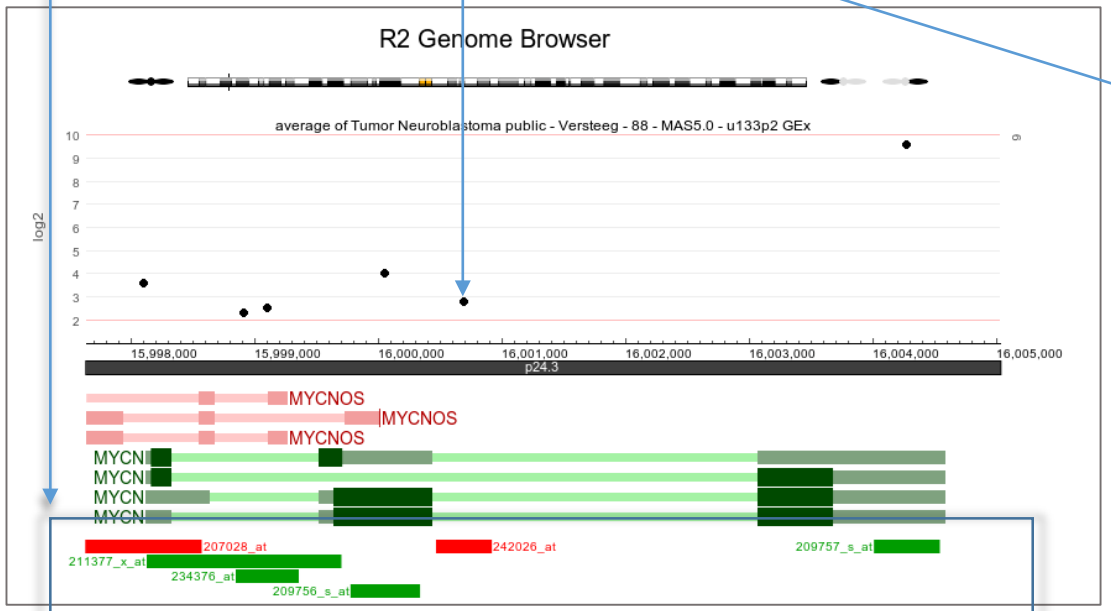


1.2.2 Investigating single genes > reporters in Genome Browser

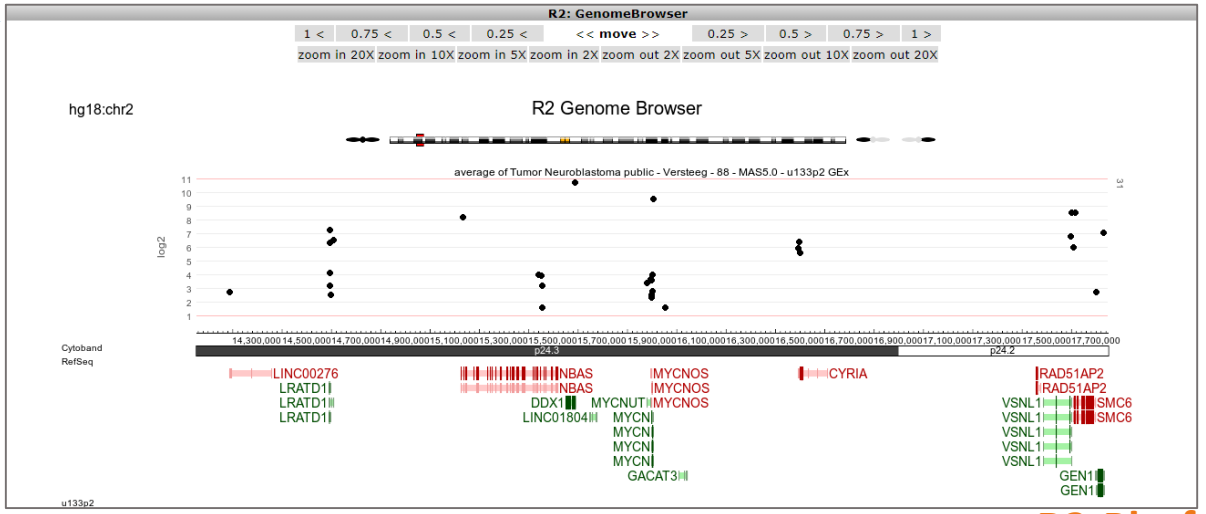
Want to know more?

https://r2-tutorials.readthedocs.io/en/latest/Using_The_Genome_Browser.html?highlight=genome%20browser





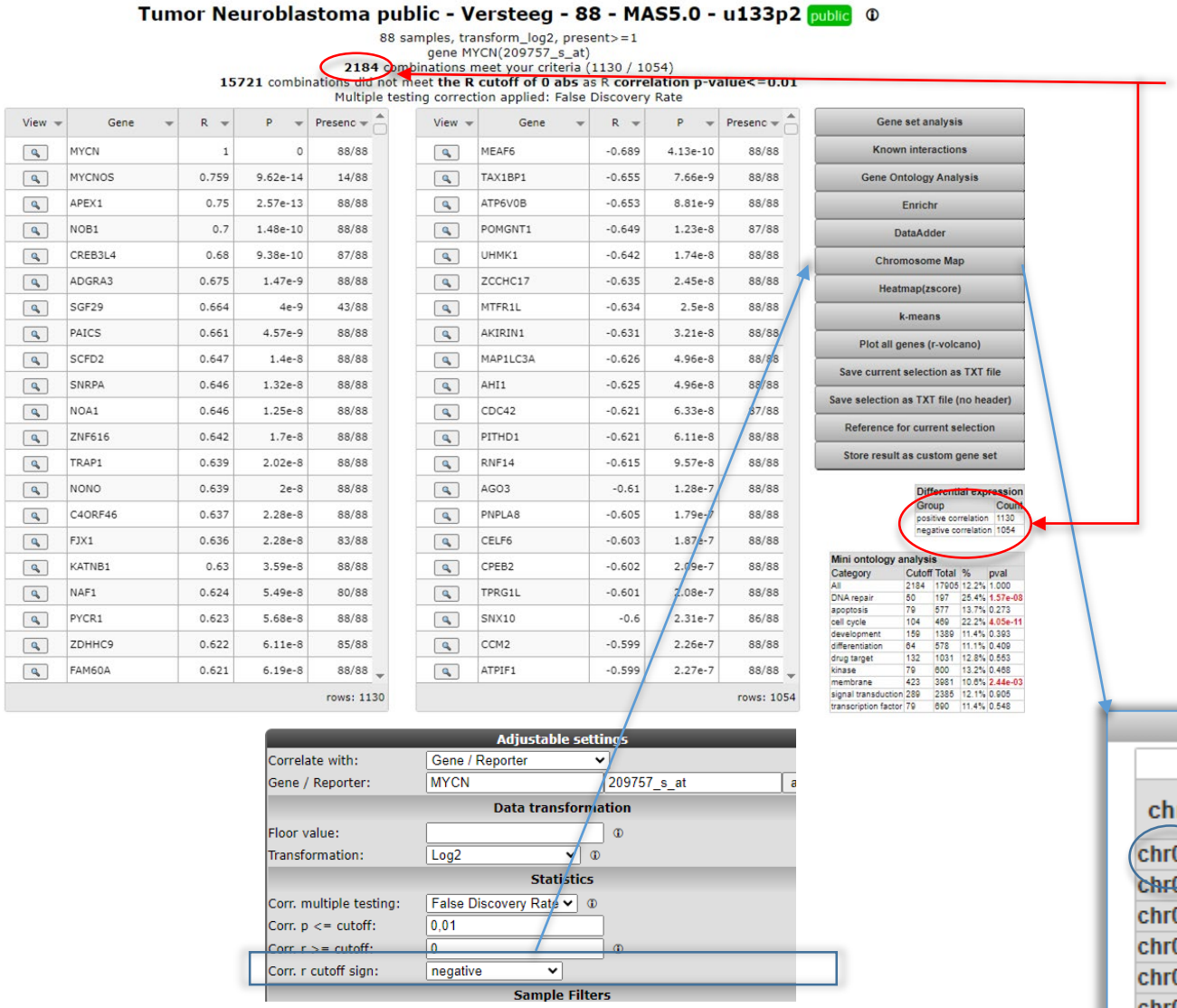
- **Do you think it is a wise idea to average the signal for MYCN over all the reporters and why?**
  - The large difference in the signals, with the lowest average signal being very low, this would not be a good idea. There might be something wrong with the low signal reporters.
  - In the reporter overview you can see indeed that the low signal reporters did not measure signal in most samples
  - If you follow the View all link, you will see the expression per reporter in a heatmap and a piece of the Genome Browser. The TView link brings you directly to the Genome Browser such that you can also zoom in and out (as is done here in the pic below) and look around on its genomic location. Both show you that one of the reporters has the opposite reading direction
    - 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)'. That means something was faulty with that reporter.
    - So, in this case where both the large signal difference and the opposite reading direction are indications that something is faulty with the reporter, it is not a good idea to average the signal of all reporters.



1.2.3 Finding correlating Genes | chromosomal overrepresentation

Want to know more?

https://r2-tutorials.readthedocs.io/en/latest/Correlating_Genes.html



View	Gene	R	P	Presenc
	MYCN	1	0	88/88
	MYCNOS	0.759	9.62e-14	14/88
	APEX1	0.75	2.57e-13	88/88
	NOB1	0.7	1.48e-10	88/88
	CREB3L4	0.68	9.38e-10	87/88
	ADGRA3	0.675	1.47e-9	88/88
	SGF29	0.664	4e-9	43/88
	PAICS	0.661	4.57e-9	88/88
	SCFD2	0.647	1.4e-8	88/88
	SNRPA	0.646	1.32e-8	88/88
	NOA1	0.646	1.25e-8	88/88
	ZNF616	0.642	1.7e-8	88/88
	TRAP1	0.639	2.02e-8	88/88
	NONO	0.639	2e-8	88/88
	C4ORF46	0.637	2.28e-8	88/88
	FJX1	0.636	2.28e-8	83/88
	KATNB1	0.63	3.59e-8	88/88
	NAF1	0.624	5.49e-8	80/88
	PYCR1	0.623	5.68e-8	88/88
	ZDHHC9	0.622	6.11e-8	85/88
	FAM60A	0.621	6.19e-8	88/88

rows: 1130

View	Gene	R	P	Presenc
	MEAF6	-0.689	4.13e-10	88/88
	TAX1BP1	-0.655	7.66e-9	88/88
	ATP6V0B	-0.653	8.81e-9	88/88
	POMGNT1	-0.649	1.23e-8	87/88
	UHMK1	-0.642	1.74e-8	88/88
	ZCCHC17	-0.635	2.45e-8	88/88
	MTFR1L	-0.634	2.5e-8	88/88
	AKIRIN1	-0.631	3.21e-8	88/88
	MAP1LC3A	-0.626	4.96e-8	88/88
	AHI1	-0.625	4.96e-8	88/88
	CDC42	-0.621	6.33e-8	87/88
	PITHD1	-0.621	6.11e-8	88/88
	RNF14	-0.615	9.57e-8	88/88
	AGO3	-0.61	1.28e-7	88/88
	PNPLA8	-0.605	1.79e-7	88/88
	CELF6	-0.603	1.87e-7	88/88
	CPEB2	-0.602	2.09e-7	88/88
	TPRG1L	-0.601	2.08e-7	88/88
	SNX10	-0.6	2.31e-7	86/88
	CCM2	-0.599	2.26e-7	88/88
	ATPIF1	-0.599	2.27e-7	88/88

rows: 1054

Category	Cutoff	Total	%	pval
All	2184	17905	12.2%	1.000
DNA repair	50	197	25.4%	1.57e-08
apoptosis	79	577	13.7%	0.273
cell cycle	104	469	22.2%	4.05e-11
development	159	1389	11.4%	0.393
differentiation	64	578	11.1%	0.409
drug target	132	1031	12.8%	0.563
kinase	79	600	13.2%	0.468
membrane	423	3981	10.6%	2.44e-03
signal transduction	289	2385	12.1%	0.905
transcription factor	79	690	11.4%	0.548

- **Approximately, how many genes were found by the test?**
  - 2184 to be exact:1130 positive/ 1054 negative
- **Where are overrepresented genes primarily located with respect to their chromosome location. ?**
  - Chrom 1p, which corresponds to the knowledge that with an mycn aplification often a loss is seen of the 1p arm.
    - In neuroblastoma, at the DNA level, MYCN amplification and loss of 1 copy of the chromosome 1p arm is a well established connection. It is described in literature that a number of tumor suppressor genes are located on chromosome 1p. 1p loss of heterozygosity (LOH) is frequently observed in MYCN amplified tumors. Interestingly, we can even 'see' this loss in the mRNA profiles, since at least a proportion of the genes show reduced expression in patients with elevated MYCN expression.

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.

R2: ChromosomeMap

	Whole chrom					p arm						
chr	Total Count	Reg Count	%	pval	p	Total Count	Reg Count	%	pval	q	Total Count	Reg Cou
chr01	2215	255	11.5%	1.15e-51	p	1197	195	16.3%	7.40e-80	q	1018	60
chr02	1489	62	4.2%	0.322	p	586	14	2.4%	8.06e-03	q	903	48
chr03	1261	28	2.2%	3.06e-05	p	580	15	2.6%	0.016	q	681	13
chr04	870	16	1.8%	6.49e-05	p	255	4	1.6%	0.018	q	615	12
chr05	1013	59	5.8%	0.093	p	198	10	5.1%	0.820	q	815	49
chr06	1199	80	6.7%	1.31e-03	p	683	31	4.5%	0.835	q	516	49
chr07	1120	100	8.9%	2.55e-11	p	377	38	10.1%	8.42e-07	q	743	62
chr08	796	36	4.5%	0.806	p	301	11	3.7%	0.389	q	495	25

1.2.4 Finding differentially expressed genes + pathways

Want to know more?

https://r2-tutorials.readthedocs.io/en/latest/Correlating_Genes.html?#step-7-gene-list-in-pathway-context

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

88 samples, transform_log2, present>=1
track alive
2620 combinations meet your criteria
15285 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
	AKR1C1	9.98e-13	1.85	alive: no < yes	88/88
	AKR1C2	1.14e-12	1.96	alive: no < yes	79/88
	FBXO30	2.26e-12	1.76	alive: no < yes	88/88
	MAP7	7.28e-12	1.8	alive: no < yes	87/88
	CCSER2	2.03e-11	0.89	alive: no < yes	88/88
	PGM2L1	2.43e-11	1.66	alive: no < yes	88/88
	IPO4	6.58e-11	-1.14	alive: no >= yes	85/88
	TTLL7	7.35e-11	1.59	alive: no < yes	88/88
	PIRT	1.19e-10	3.05	alive: no < yes	77/88
	PMP22	3.29e-10	1.53	alive: no < yes	88/88
	WDR47	6.35e-10	0.74	alive: no < yes	88/88
	TMOD2	7.33e-10	0.9	alive: no < yes	88/88
	SCN9A	7.75e-10	1.46	alive: no < yes	87/88
	PRKACB	7.95e-10	0.96	alive: no < yes	88/88
	DST	8.3e-10	0.75	alive: no < yes	88/88
	STXBP5	9.89e-10	1.13	alive: no < yes	88/88
	CCDC86	1.2e-9	-0.84	alive: no >= yes	86/88

Gene set analysis
Known interactions
Gene Ontology Analysis
Enrichr
DataAdder
Chromosome Map
Heatmap(zscore)
k-means
Plot all genes (xy, volcano etc)
Group change bar plot
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
alive: no >= yes	1523
alive: no < yes	1097

From your input (n=**2620**), **923** genes were also present in the current geneset collection.
Within the reference from the original search, 5656 genes were detected in the current geneset selection
The table below lists genesets where the number of genes from your list are present more than expected (**p<0.05** from correction)
alive: no < yes, alive: no >= yes,

set	R#	#	_p_value_	Genelist
over-representation KAH DNA_replication	36	20	7.06e-10	FEN1, LIG1, MCM2, MCM3, MCM4, MCM5, MCM6, MCM7, PCNA, POLA1, POLA2, RNASEH2A, SSBP1
over-representation KAH Ribosome_biogenesis_in_eukaryotes	64	27	4.71e-08	CSNK2A1, CSNK2A2, FBL, GAR1, GTPBP4, HEATR1, IMP4, NAT10, NHP2, NOB1, RPP25L, RPP40, SPATA5, TCOF1, UTP14A, UTP15, WDR3, WDR43, WDR75
over-representation KAH Pyrimidine_metabolism	94	34	3.20e-07	CAD, CMPK1, DCTPP1, DTYMK, ENTPD3, ENTPD4, NME1, NME4, NT5C2, NT5M, POLE3, POLR1B, POLR1D, POLR2C, POLR2H, POLR2I, POLR2L, POLR3D, POLR ZNRD1
over-representation KAH RNA_transport	139	45	3.97e-07	ALYREF, EIF2B2, EIF2S1, EIF2S2, EIF3B, EIF3F, EIF3G, EIF4E2, EIF4EBP1, EIF4G NUP188, NUP205, NUP37, NUP62, NUP93, NXT1, NXT2, PABPC1, PABPC3, PABP SEC13, SENP2, SNUPN, TACC3, THOC3, TPR, TRNT1, UBE2I, UPF1, UPF3B, XPO
over-representation KAH Cell_cycle	121	40	8.99e-07	ANAPC11, ATR, BUB1, BUB1B, CCNA2, CCNB1, CCNB2, CCNE1, CDC45, CDK1, E2F1, E2F3, ESPL1, FZR1, MAD2L1, MCM2, MCM3, MCM4, MCM5, MCM6, MCM7,

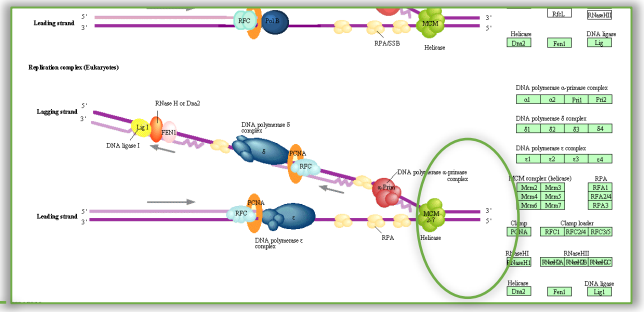
- **How many genes are differentially expressed between the alive “no” and “yes” group?**
  - 2620
- **Which KEGG pathway is the most significant?**
  - DNA replication. All genes in this process have a consistent positive correlation (as can be seen by the green color).

Columns: set = KEGG pathway name | R# = number of genes annotated to be in that geneset | # = number of genes from our list of Diff expr genes that were annotated to be part of this KEGG pathway | p_value =

An overrepresentation analysis is performed on all gene members of the pathways in the KEGG database

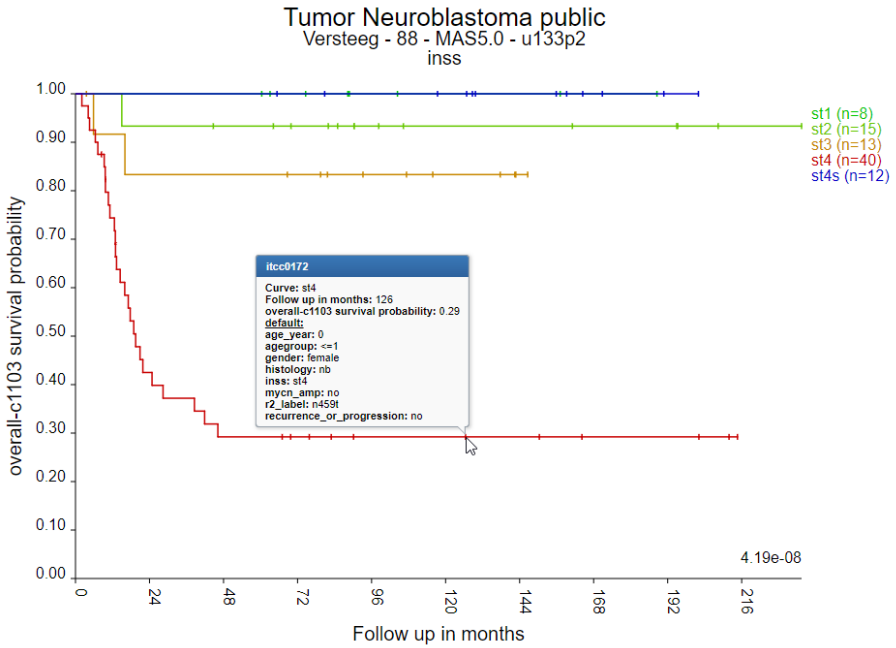
- **Can you find the function of the MCM2-7 complex in the picture? For which group of our analysis are these genes upregulated?**

MCM2-7 genes are in the figure for eukaryotes, and are involved in helicase. These genes show upregulation in the alive: no group.Minichromosome maintenance 2 (MCM2) is a member of the minichromosomal maintenance family of proteins that mainly regulates DNA replication and the cell cycle and is involved in regulating cancer cell proliferation in various cancers. Previous studies have reported that MCM2 plays a pivotal role in cell proliferation and cancer development



1.2.5. Using annotation and the Kaplan Meier curve

Want to know more?
https://r2-tutorials.readthedocs.io/en/latest/Kaplan_Meier.html



- **What does a drop in the curves mean? And the little vertical tick-mark on the horizontal parts of the curves? Scroll over the drops and the tick-marks of the curves to see clinical details of patients.**

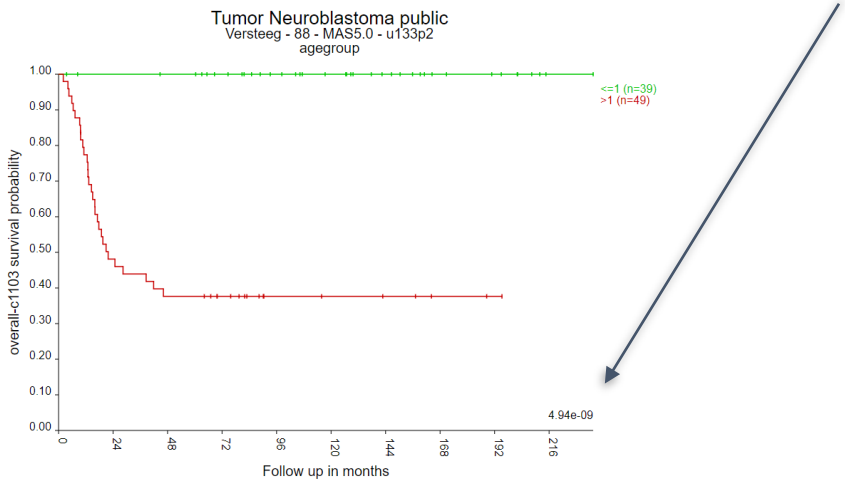
The horizontal axis (x-axis) represents time in months, and the vertical axis (y-axis) shows the probability of surviving or the proportion of people surviving. The lines represent survival curves of the five groups. A vertical drop in the curves indicates an event (here:death). The vertical tick mark on the curves means that a patient was censored at this time.

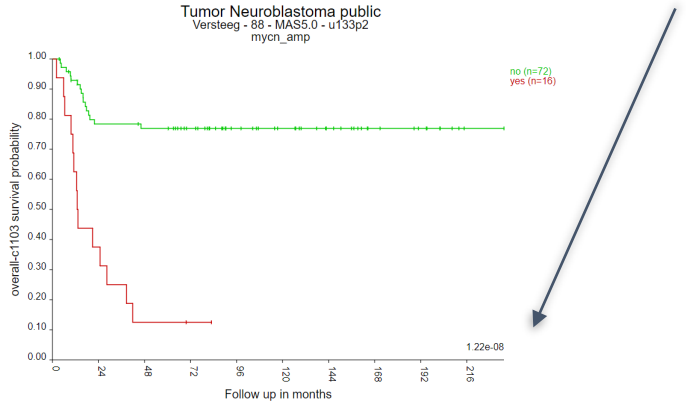
N.B. Hovering over vertical drops or ticks shows the annotation information of that patient. In this case we see that stage 4 has the worst prognosis. Stage 1 and stage 4s both have a survival probability of 1, i.e. they are drawn on top of each other.

- **Do you observe a significant difference between the groups?**

Yes, age_group <=1 has a good prognosis, agegroup > 1 has a bad prognosis, p=4.94e-09. Children that were older than 1 year when they got diagnosed with neuroblastoma, had a significantly worse prognosis

For mycn_amp, the difference in survival chances is significant as well: 1.22^{e}-08

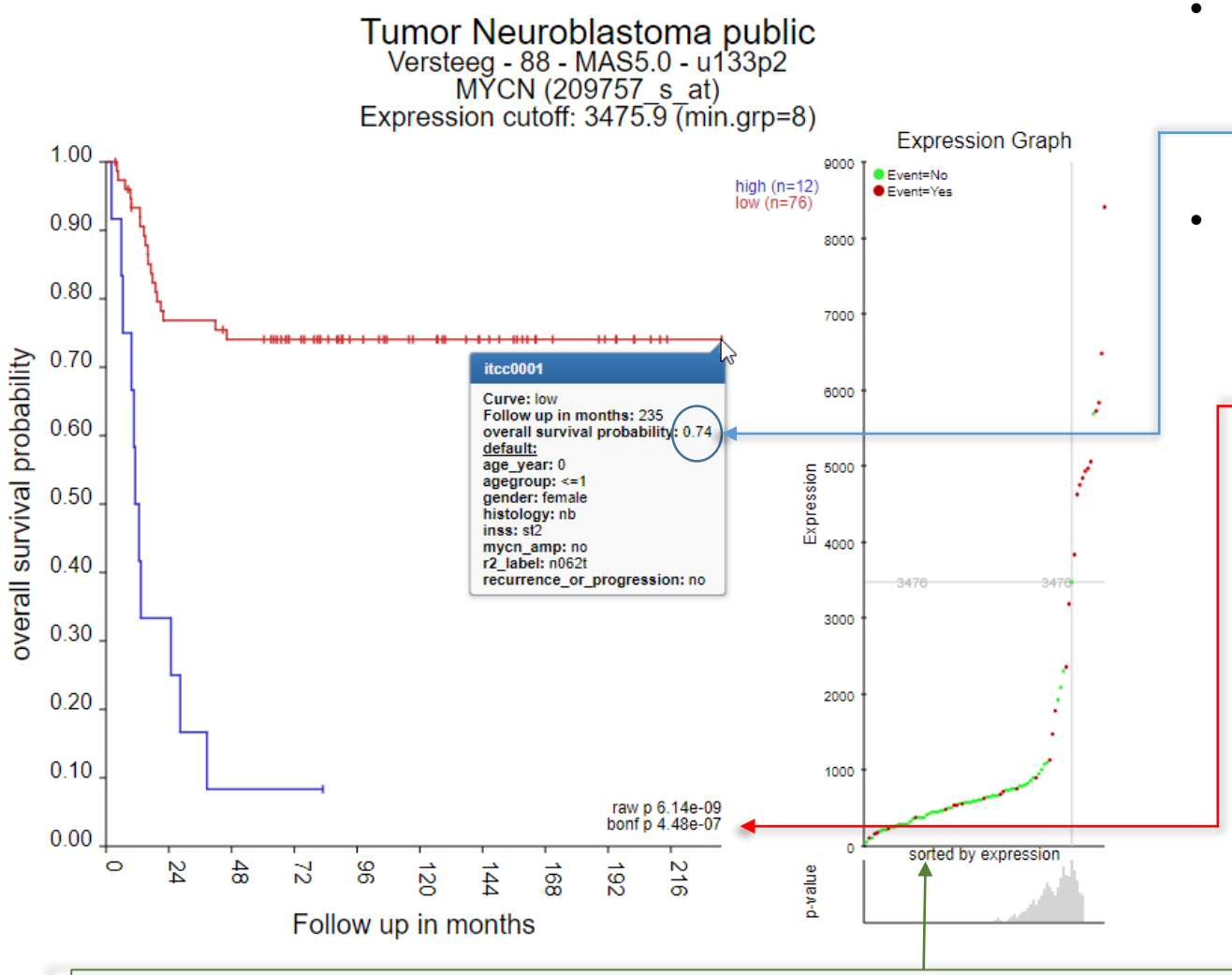




1.2.6. Kaplan Meier: Validating prognostic factors such as gene expression

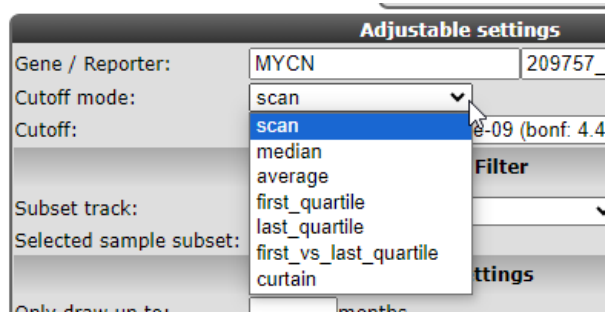
Want to know more?

https://r2-tutorials.readthedocs.io/en/latest/Kaplan_Meier.html?highlight=kaplan%20scan#step-2-kaplan-meier-by-gene-expression-the-kaplan-scan



- **What is the survival chance for high and low MYCN expression, as assessed by the kaplan scan (use the extreme right values)?**
  - 0.08 for the high group, and 0.74 for the low group

- **Which method gives the clearest prognostic groups? Is there a consequence for the statistics and P-value of the scanning method?**
  - The p-value is the most significant for the scan cut-off (scan p=6.1e-09/ bonf p 4.48e-07, median p=0.037, average p=1.08e-05) with most distinct prognostic groups
  - N.B. If you want to report the p-value of a Kaplan Scan, use the bonferroni p-value due to correction for multiple testing

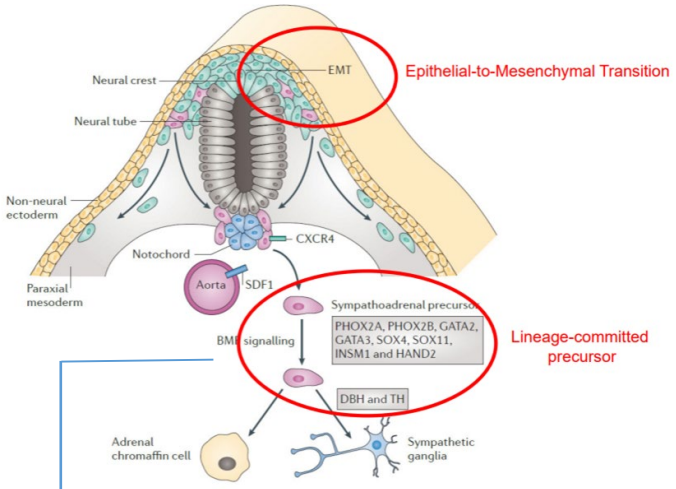


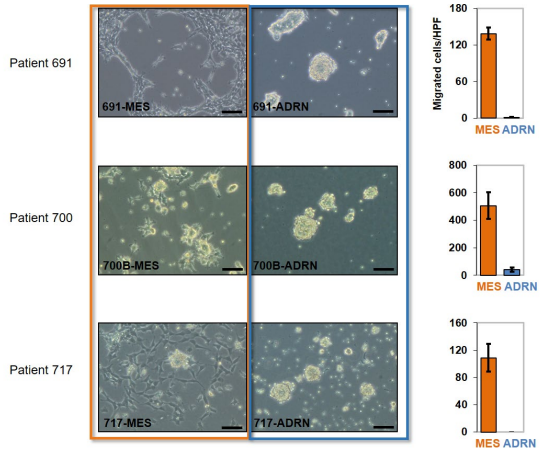
N.B. The Kaplan scanner separates the samples of a dataset into two groups based on the gene expression of one gene. In the order of expression, it will use every increasing expression value as a cutoff to create 2 groups and test the p-value in a logrank test. The highest value is then reported, accompanied by a Kaplan Meier picture

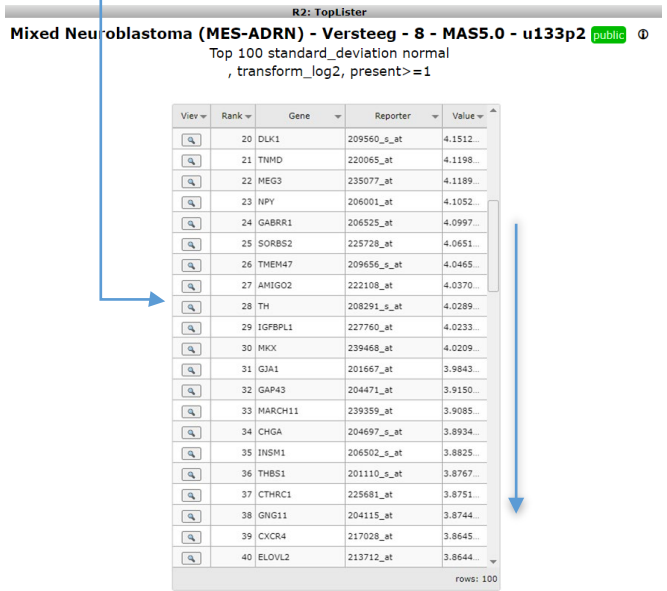
1.3 Different expression patterns between subgroups and the underlying biology

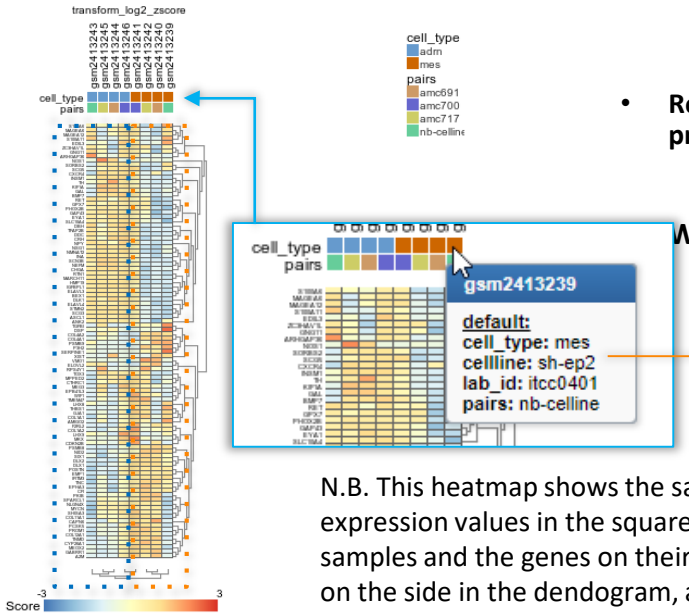
Want to know more?

https://r2-tutorials.readthedocs.io/en/latest/Using_Genesets_creating_Heatmaps.html#step-4-unsupervised-hierarchical-clustering-with-a-geneset









- **What do you note about the morphology of the cell lines?**
  - MES: cells grew attached, formed lamellipodia and were motile,
  - ADRN: More semi attached spheres and most likely less migration properties
    - The two types were found in each patient

- **Do you recognize any genes from figure 5 when you scroll down through the list? I.e. genes that come into play in the development of the sympatho-adrenal lineage from the neural crest?**
  - TH (Rank 28), INSM1 (Rank 35), PHOX2B (Rank 41), DBH (Rank 83)

N.B. **TopLister** is used to investigate the presence of subgroups *without using annotation* information in a dataset or to find a list of genes with the highest variation in gene expression

TopLister is an *unsupervised* algorithm, meaning it does not use any annotation to guide its analyses. It uses the gene expression values only. It is interesting that even without any guidance of what to focus on, several genes from the sympathoadrenal lineage are in the list of the top 100 genes with the highest variation in gene expression. Meaning, within this dataset, samples seem to behave in different ways with respect to these genes.

- **Roughly how many groups of samples do you see in the heatmap, showing similar expression profiles within that group? Is this what you expected?**
  - Although it's not a complete clear-cut black white division, you can see two groups in the heatmap.

What feature determines the clustering of the samples?

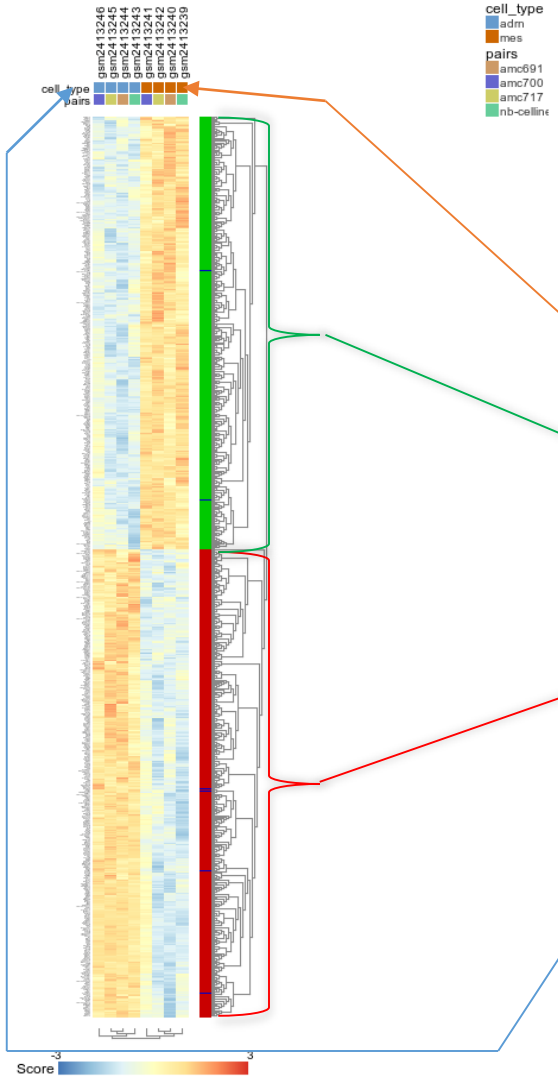
- The samples seem to cluster in agreement with the cell_type track, not with the patient track. This hints to two differing gene expression programs that are switched on per cell type. 1) Group Orange: Includes SHEP2, which clusters with mesenchymal origin tumors; migration capability; grows attached to the dish as loose cells 2) Group Blue: Includes SY5Y, clusters with neuro-ectodermal origin tumors; no migration; grows semi-attached in spheres together

N.B. This heatmap shows the samples on the columns, the genes on the rows and their color coded zscore of the expression values in the squares (see legend underneath). An unsupervised hierarchical clustering algorithm re-orders the samples and the genes on their similarity in z-score expression profiles and you can see the clustering trees of the genes on the side in the dendogram, and of the samples underneath. Thus, samples that show similar expression profiles over the geneset are grouping together.

r2-support@amsterdamumc.nl

1.3.2. Which genes make a difference? Creating signatures

Want to know more?
https://r2-tutorials.readthedocs.io/en/latest/Using_Signatures.html



- **How is this** *Differential expression between two groups* **heatmap figure different from the former?**
  - The heatmap is very cleanly divided. That is not strange because we asked for the most differentially expressed genes between these two subsets of samples. A *supervised search*, in which the cell type labels were provided to the algorithm, next to the expression data. It is therefore only logical that we get such a clean division of upregulated genes in the one group where they are downregulated in the other group.

    N.B. The heatmap itself is still clustered *unsupervised*, so you can see that the cells indeed group together with their respective groups, and this that these genes make a clear distinction between the two groups.
  - You also see these two neat genesets that now characterise the cell types:
    - one geneset is upregulated in the mesenchymal cell types, and downregulated in the adrenergic cell types.
    - and the other way around: one geneset is upregulated in the adrenergic cell types, and downregulated in the mesenchymal.
  - The result of this analysis was previously stored in R2: these two genesets we will use as gene signatures:
    - the genes that are upregulated in the mesenchymal like cells is stored as a mesenchymal signature geneset and was named r2_mesadrn_mes
    - and the genes that are upregulated in the other andrenergic like cells is stored as a andrenergic signature geneset and was named r2_mesadrn_adrn

1.3.2. Which genes make a difference? Gene set analyses

Want to know more?

https://r2-tutorials.readthedocs.io/en/latest/Correlating_Genes.html?highlight=gene%20ontology#step-5-establishing-overrepresentation-in-other-domains

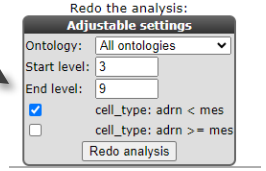
cell_type: adrn < mes, cell_type: adrn >= mes,

GoPath	R#	#	_p_value_	Goid-Desc	GeneSymbols
12505	4060	248	6.87e-13	3: endomembrane system (3:4)	over 100 entries (2, 151, 95,)
5518	61	13	1.36e-10	2: collagen binding (4:4)	ECM2, NID2, FN1, PCOLCE2, RELL2, ITGB1, NID1, PPIB, SPARC, ANTXR1, SERPINH1, ADAM9, MRC2
62023	102	17	9.45e-10	3: collagen-containing extracellular matrix (4:5)	COL3A1, CCDC80, ECM2, NID2, RELL2, EFEMP2, HSPG2, LAMB1, LAMB2, LAMB3, LAMC1, LOXL2, NID1, SPARC, COL18A1, HMCN1, COL27A1
22008	1446	104	1.73e-09	1: neurogenesis (5:7)	over 100 entries (1, 47, 56,)
7399	2154	141	2.82e-09	1: nervous system development (5:6)	over 100 entries (1, 63, 77,)
48699	1353	98	3.62e-09	1: generation of neurons (6:8)	CDH4, TUBB3, YAP1, SEMA4D, SEMA3C, DLL3, NRG3, STMN2, GPRIN1, CCSAP, COL3A1, CTNNA1, DLX1, DLX2, ECM2, EPHA3, EXT1, GPC2, FKBP4, TRAK1, UNC13A, AGTPBP1, FN1, CLCF1, ALK, GABRB3, TENM4, FBXW8, CYFIP2, NPTN, GLI3, GNAQ, RACGAP1, HES1, HSPA5, ID1, NANOS1, APOD, IL6ST, ITGB1, LAMB1, LAMB2, PHOX2A, LRP4, MAN2A1, MAP1B, MYB, NAGLU, NEDD4, NEFM, NOTCH3, RNF165, MINK1, MYEF2, HDGFRP3, PCM1, NGRN, PEX7, PIK3R1, FEV, KIRREL, PRKG1, MAPK8, B2M, PTCHD2, SDC4, SEMA3F, SGK1, BMPR1B, SNAP25, SPTB, SYT1, SYT4, CNTN2, TEAD3, VCL, WNT5A, WNT7A, DBNDD1, FZD3, CALR, RND2, STMN4, NDEL1, CAMK2B, FZD1, SCRT1, SYT3, RUNX2, IRS2, CBLN1, CDK5R1, CDK5R2, CCDC64, KLF4, ONECUT2, ZNF536, LZTS3
5788	287	32	4.61e-09	3: endoplasmic reticulum lumen (5:10)	CRTAP, CHGB, FSTL1, COL3A1, COL6A2, ERP44, FN1, SCG3, HSPA5, CYR61, LAMB1, LAMB2, LAMC1, GPX8, P4HB, DNAJC10, PPIB, UGGT2, PDGFC, B2M, QSOX1, PDIA2, WNT5A, WNT7A, ARSJ, COL18A1, CALR, CALU, EVA1A, COL27A1, SERPINH1, PDIA4
9653	2396	152	7.08e-09	1: anatomical structure morphogenesis (3:4)	over 100 entries (94, 58,)
43202	89	15	8.03e-09	3: lysosomal lumen (6:11)	NPC2, CSPG4, GPC2, GALNS, NSG1, NAAA, HEXA, HSPG2, LIPA, NAGLU, NEU1, HMP19, CTSA, BGN, SDC4
5925	378	38	1.18e-08	3: focal adhesion (5:6)	DLC1, CNN2, CSPG4, CTNNA1, SYNPO2, PALLD, CLASP2, NCSTN, GSN, ANXA1, HMGA1, HSPA5, HSPG2, IGF2R, ITGA4, ITGB1, LPP, P4HB, PPIB, DCAF6, B2M, SDC4, SDCBP, SLC9A1, SNTB2, TGFB1I1, TNS1, CAPN5, VCL, PEAK1, ATAT1, CALR, FZD1, ADAM9, PHLDB2, STARD8, GIT2, MRC2
5604	81	14	1.42e-08	3: basement membrane (5:6)	CCDC80, NID2, RELL2, EFEMP2, HSPG2, LAMB1, LAMB2, LAMB3, LAMC1, LOXL2, NID1, SPARC, COL18A1, HMCN1
5924	381	38	1.61e-08	3: cell-substrate adherens junction (4:5)	DLC1, CNN2, CSPG4, CTNNA1, SYNPO2, PALLD, CLASP2, NCSTN, GSN, ANXA1, HMGA1, HSPA5, HSPG2, IGF2R, ITGA4, ITGB1, LPP, P4HB, PPIB, DCAF6, B2M, SDC4, SDCBP, SLC9A1, SNTB2, TGFB1I1, TNS1, CAPN5, VCL, PEAK1, ATAT1, CALR, FZD1, ADAM9, PHLDB2, STARD8, GIT2, MRC2
16864	18	6	1.83e-08	2: intramolecular oxidoreductase activity, transposing S-S bonds (5:5)	ERP44, P4HB, TMX4, QSOX1, PDIA2, PDIA4
3756	18	6	1.83e-08	2: protein disulfide isomerase activity (4:6)	ERP44, P4HB, TMX4, QSOX1, PDIA2, PDIA4
30055	386	38	2.70e-08	3: cell-substrate junction (3:3)	DLC1, CNN2, CSPG4, CTNNA1, SYNPO2, PALLD, CLASP2, NCSTN, GSN, ANXA1, HMGA1, HSPA5, HSPG2, IGF2R, ITGA4, ITGB1, LPP, P4HB, PPIB, DCAF6, B2M, SDC4, SDCBP, SLC9A1, SNTB2, TGFB1I1, TNS1, CAPN5, VCL, PEAK1, ATAT1, CALR, FZD1, ADAM9, PHLDB2, STARD8, GIT2, MRC2
5775	160	21	3.19e-08	3: vacuolar lumen (5:10)	NPC2, CSPG4, GPC2, GALNS, NSG1, NAAA, GRN, HEXA, HSPG2, C6ORF120, LIPA, NAGLU, NEU1, HMP19, CTSA, PRSS2, BGN, SDC4, SDCBP, DSN1, CREG1
31589	305	32	4.45e-08	1: cell-substrate adhesion (4:4)	DLC1, COL3A1, CCDC80, ECM2, EPHA3, NID2, CLASP2, FN1, GBP1, RELL2, ID1, APOD, CYR61, ITGA4, ITGB1, LAMB1, LAMC1, BCAM, NID1, MINK1, PIK3R1, PRKX, SDC4, SLC9A1, VCL, PEAK1, CALR, ANTXR1, ADAM9, PHLDB2, MYADM, ONECUT2
5783	1677	112	5.58e-08	3: endoplasmic reticulum (4:8)	over 100 entries (2, 77, 33,)
97458	1359	95	5.72e-08	3: neuron part (3:4)	TUBB3, TUBB4A, PIAS3, CETN2, CPLX2, CPLX1, PPARGC1A, STMN2, CHRM2, SYNPO, GPRIN1, RAB3C, ADCYAP1R1, TPH2, CCSAP, NRSN1, NQO1, EEF1A2, SCAMP5, ENO2, FKBP4, TRAK1, PALLD, UNC13A, FKBP15, SEZ6L, TENM4, GCH1, RGS17, CYFIP2, NPTN, NSG1, NAAA, GNAQ, GNAZ, GRM8, TMOD2, HNMT, ICA1, APOD, IL6ST, ITGB1, KCNB1, RHOC, LRP4, MAP1B, NEDD4, NEFM, ATP1A3, ATP2B4, MINK1, PAM, PDE2A, ACTL6B, HMP19, PLCB4, TMEM57, SLC38A7, KIRREL, SVOP, MAPK8, PTPRN, QDPR, LRRC4, SLC6A2, SLC18A1, SLC18A3, BMPR1B, SNAP25, SYP, SYPL1, SYT1, SYT4, CNTN2, WNT7A, CACNA1B, DBNDD1, FZD3, ATAT1, AP3B2, STMN4, NDEL1, CAMK2B, CAMK2D, CASP8, SYT3, PRSS12, PKP4, AP3B1, CDK5R1, SYNJ2, BSN, SYT12, KCNB2, LZTS3
					RTN3, FAM3C, KIF1C, MMP24, TMED2, KDELR3, STMN2, SEC23IP, PRRC1, CSPG4, TMC8, MANEAL, CTNNA1, SCAMP5, EXT1, GPC2, ATF6, CLASP2, DNMBP, POFUT2, NCSTN, GLCE, FBXW8, GBP1, NSG1, MUC19, ZDHHC22, DSE, HLA-F, HSPG2, ICA1, ID1

cell_type: adrn < mes, cell_type: adrn >= mes,

GoPath	R#	#	_p_value_	Goid-Desc	GeneSymbols
5788	287	29	1.51e-21	3: endoplasmic reticulum lumen (5:10)	CRTAP, FSTL1, COL3A1, COL6A2, ERP44, FN1, HSPA5, CYR61, LAMB1, LAMB2, LAMC1, GPX8, P4HB, DNAJC10, PPIB, UGGT2, PDGFC, B2M, QSOX1, WNT5A, WNT7A, ARSJ, COL18A1, CALR, CALU, EVA1A, COL27A1, SERPINH1, PDIA4
5925	378	34	2.37e-21	3: focal adhesion (5:6)	DLC1, CNN2, CSPG4, CTNNA1, PALLD, NCSTN, GSN, ANXA1, HSPA5, HSPG2, IGF2R, ITGA4, ITGB1, LPP, P4HB, PPIB, DCAF6, B2M, SDC4, SDCBP, SLC9A1, SNTB2, TGFB1I1, TNS1, CAPN5, VCL, PEAK1, CALR, FZD1, ADAM9, PHLDB2, STARD8, GIT2, MRC2
5924	381	34	4.17e-21	3: cell-substrate adherens junction (4:5)	DLC1, CNN2, CSPG4, CTNNA1, PALLD, NCSTN, GSN, ANXA1, HSPA5, HSPG2, IGF2R, ITGA4, ITGB1, LPP, P4HB, PPIB, DCAF6, B2M, SDC4, SDCBP, SLC9A1, SNTB2, TGFB1I1, TNS1, CAPN5, VCL, PEAK1, CALR, FZD1, ADAM9, PHLDB2, STARD8, GIT2, MRC2
62023	102	16	4.31e-21	3: collagen-containing extracellular matrix (4:5)	COL3A1, CCDC80, ECM2, NID2, EFEMP2, HSPG2, LAMB1, LAMB2, LAMB3, LAMC1, LOXL2, NID1, SPARC, COL18A1, HMCN1, COL27A1
30055	386	34	1.05e-20	3: cell-substrate junction (3:3)	DLC1, CNN2, CSPG4, CTNNA1, PALLD, NCSTN, GSN, ANXA1, HSPA5, HSPG2, IGF2R, ITGA4, ITGB1, LPP, P4HB, PPIB, DCAF6, B2M, SDC4, SDCBP, SLC9A1, SNTB2, TGFB1I1, TNS1, CAPN5, VCL, PEAK1, CALR, FZD1, ADAM9, PHLDB2, STARD8, GIT2, MRC2
5518	61	12	1.37e-20	2: collagen binding (4:4)	ECM2, NID2, FN1, PCOLCE2, ITGB1, NID1, PPIB, SPARC, ANTXR1, SERPINH1, ADAM9, MRC2
12505	4060	152	2.03e-18	3: endomembrane system (3:4)	over 100 entries (1, 151,)
5604	81	13	1.30e-17	3: basement membrane (5:6)	CCDC80, NID2, EFEMP2, HSPG2, LAMB1, LAMB2, LAMB3, LAMC1, LOXL2, NID1, SPARC, COL18A1, HMCN1
1903561	2583	109	3.13e-17	3: extracellular vesicle (4:5)	over 100 entries (109,)
43230	2585	109	3.37e-17	3: extracellular organelle (3:4)	over 100 entries (109,)
43062	376	31	3.76e-17	1: extracellular structure organization (4:4)	COL3A1, COL6A2, CCDC80, OLFML2A, ECM2, NID2, FN1, GSN, EFEMP2, HSPG2, CYR61, ITGA4, ITGB1, LAMB1, LAMB2, LAMB3, LAMC1, LOXL2, NID1, P4HB, HTRA1, BGN, SDC4, SDCBP, SPARC, COL18A1, ANTXR1, COL27A1, SERPINH1, ADAM19, PHLDB2
70062	2566	108	5.99e-17	3: extracellular exosome (4:6)	over 100 entries (108,)

- **What can you say about the function of the differentially expressed genes?**
  - The table is color coded, and the color coding is seen above the table. Be aware that this color coding is independent from the color coding on the previous slide. The red – green color scheme in the Gene Ontology table depends on your chosen cell type as Group 1 and which group is put in Group 2 in the *Differential Expression Between Two Groups* analysis and thus could be swapped. Always check above the table for the color coding.
  - In our case red colored genes are higher expressed in Mesenchymal type cells and green colored genes are higher expressed in Adrenergic type cells.
  - In the table you see many biological functions popping up that you might have heard before in the context of neuroblastoma: collagen binding (most genes higher expressed in mes type), focal adhesion (important for cell motility, most genes higher expressed in mes type) and neuron related gene sets (most genes higher expressed in adrn type)
- **What can you say about the function of the differentially expressed genes when looking only at adrn < mes genes?**
  - We see many gene sets that deal with the extracellular matrix and are motility related



- **Gene set analysis: Which hallmark category of genes pops up as most important? Can you explain this?**
  - HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION: the difference between these two subgroups seems to have to do with the epithilial to mesenchymal trasition

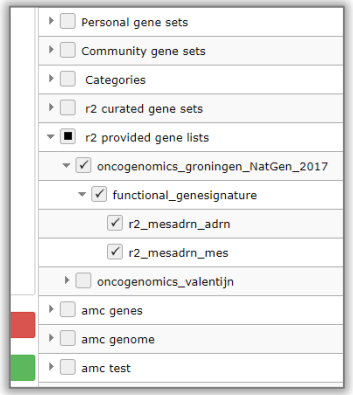
cell_type: adrn < mes, cell_type: adrn >= mes,

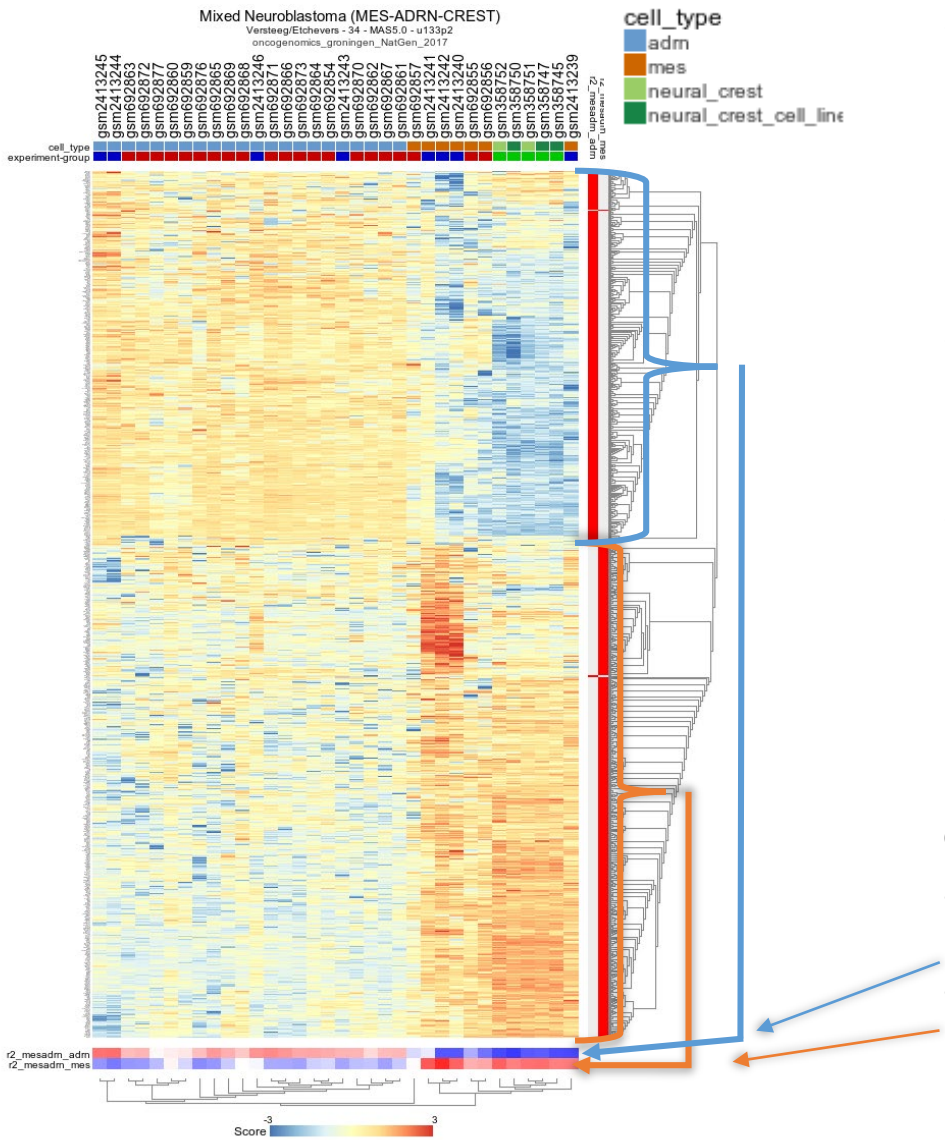
set	R#	#	_p_value_	
over-representation HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	199	26	2.56e-06	BGN, CALD1, CALU, COL3A1, COL6A2, CYR61, ECM2, EFEMP2, ENO2, FN1, FSTL1, H
over-representation	144	19	4.22e-05	ATP2B4, ATXN1, COL3A1, CYR61, DLC1, ICA1, ID1, INPP4B, LAMC1, MAP1B, MIOS, N

1.3.2. Which genes make a difference? Corroborate in another dataset, also two groups?

Want to know more?

https://r2-tutorials.readthedocs.io/en/latest/Using_Signatures.html





- Reflection:
  - Using a supervised (i.e. labeled) approach we found a set of genes that is specifically expressed in each cell type. We call this a signature (geneset, category). We stored them as r2_mesadrn_adrn and as r2_mesadrn_mes
  - The MES signature set seems to associate with the EMT transition
    - => the genes higher expressed in the MES type associate with motility
  - Now we want to corroborate these signatures in another dataset
- **What cell_type of the samples are assocciated with the 2 groups?**
  - In this dataset, again you have the annotation of the morphological phenotype of each cell. In the heatmap of the two signature gene sets together, the samples cluster by cell type again.
- **In which cluster are the neural crest cells positioned, and does that make sense?**
  - The expression profiles of the mesenchymal type cells are similar to the neural crest cells. Thus you can see that the clustered groups relate to the stages of differentiation, in which the lineage commited andrenergic group separates from the undifferentiated mesenchymal and neural crest cells

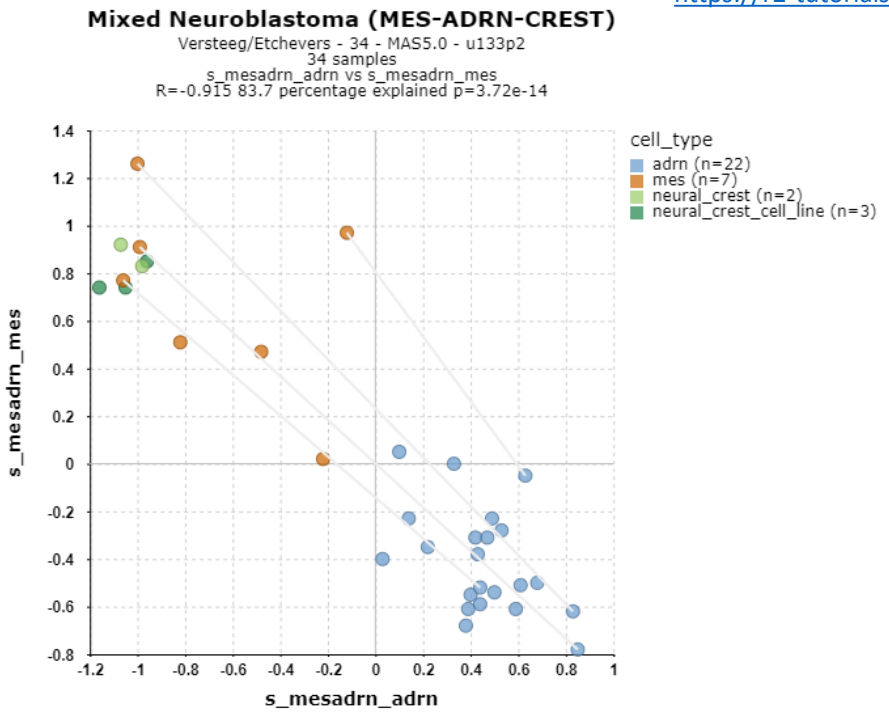
N.B. Under the heatmap, R2 shows two lines of values: one color coded line for r2_mesadrn_adrn and one for r2_mesadrn_mes. For each sample (so for each column):

- the average is calculated of the sample's z-scores of the r2_mesadrn_adrn genes (i.e. the adrn signature score)
- and the average of the sample's z-scores of the r2_mesadrn_mes genes (i.e. the mes signature score), resulting in two signature scores per sample.

1.3.2. Which genes make a difference? Corroborate in another dataset, also two groups?

Want to know more?

https://r2-tutorials.readthedocs.io/en/latest/Using_Signatures.html?highlight=signature%20score#step-4-plot-signature-scores-using-the-relate-2-tracks-module

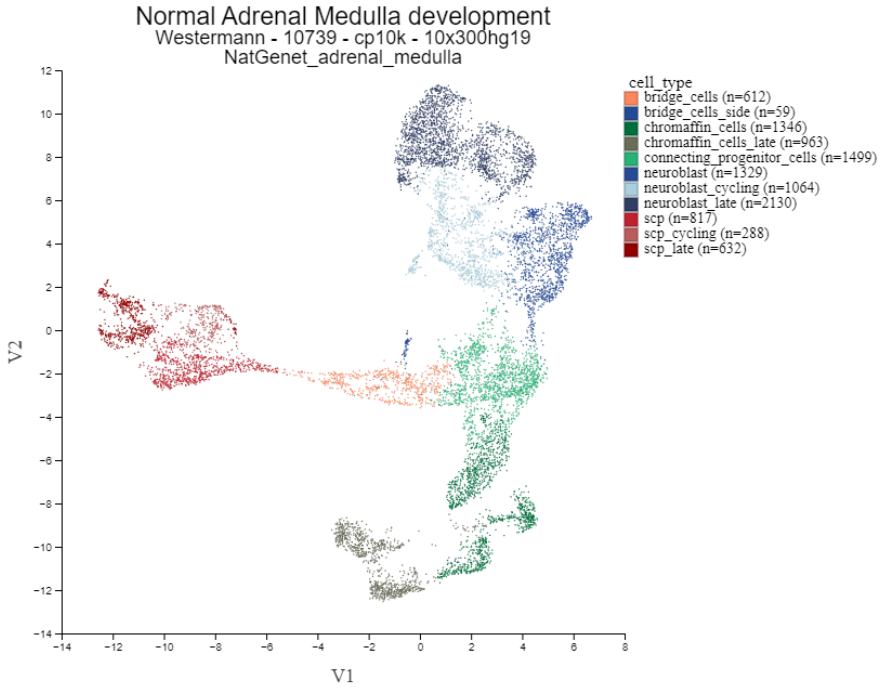


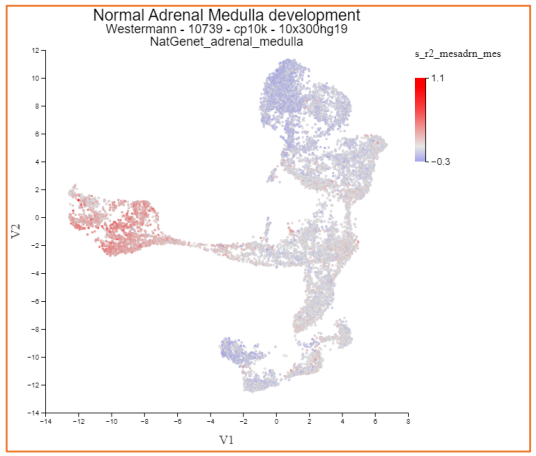
What is the relationship between MES and ADRN scores?

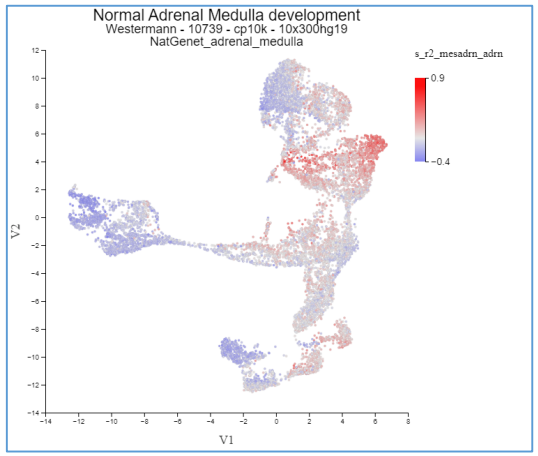
- Reflection:
  - Now that we have summarized the expression of two genesets in two single values – which therefore characterize the cells like these gene expression profiles did, we start to see a pattern
  - The genes in the mRNA signatures were used to calculate MES and ADRN scores for 22 adrenergic, 7 mesenchymal neuroblastoma cell lines, and 5 neural crest cell lines.
  - This graph is a very insightful representation that summarizes the 800 genes that we were looking at in the previous heatmap
- There is a negative correlation between the MES and ADRN scores. You can see a gradient of the ADRN and MES scores.
- We see that MES samples lay close to the neural crest samples in the graph, and the adrn samples further away. which is in agreement with our biological knowledge of these cell lines types.
- This suggests again that the MES cells show behavior more simlar to precursors of the adrenergic lineage
- The samples linked in this picture are originating from the same patient material. Therefore it shows heterogeneity in the tumor

1.3.2. Which genes make a difference? Corroborate in a Single cell dataset with UMAP

Want to know more?
https://r2-tutorials.readthedocs.io/en/latest/tSNE_dimensionality_reduction.html





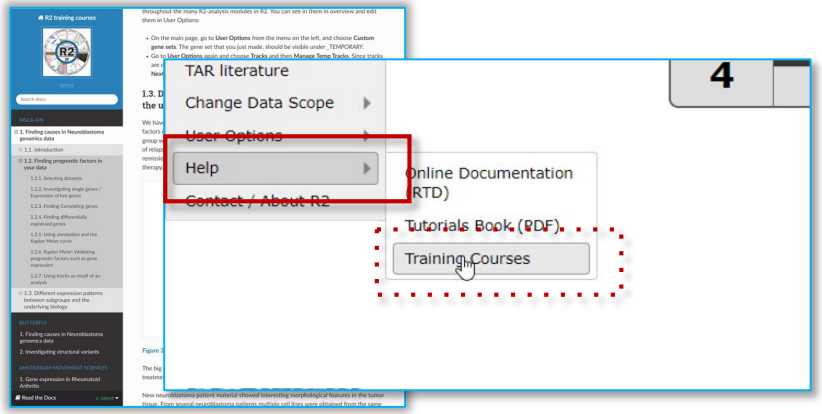


- **Can you see the route(s) of development in the representation?**
  - Dataset https://communities.springernature.com/posts/on-a-quest-to-identify-the-origin-of-neuroblastoma :"For the present study, we analysed 17 human adrenal glands from different developmental time points (seven weeks - 17 weeks postconception) by single-nucleus RNA sequencing. We annotated all cells based on their expression of marker genes"
  - In the UMAP you can see the identified cells cluster according to their different developmental time points. A starting point of Schwann cell precursors (SCPs), and two end points of chromaffin cells and sympathetic neuroblasts. "In addition, some cells from the early time points seven and eight weeks postconception formed transient populations of bridge cells (connecting SCPs and chromaffin cells) and connecting progenitor cells (located between the chromaffin cell, neuroblast and bridge cell clusters). Cells from later developmental time points clustered separately, allowing discrimination between more differentiated late and cycling cell populations. "
- **Do the signatures 'light up' the expected regions in the UMAP ?**
  - Yes. We overlay this UMAP with the signature scores of this dataset for the earlier discussed r2_mesadrn_adrn and r2_mesadrn_mes genesets
  - The Schwann cell precursor cluster has a high r2_mesadrn_mes signature score
  - The more differentiated cell types have a high r2_mesadrn_adrn signature score

What's next?

R2 Training Courses

- Help => Training Courses
- http://r2platform.com/courses

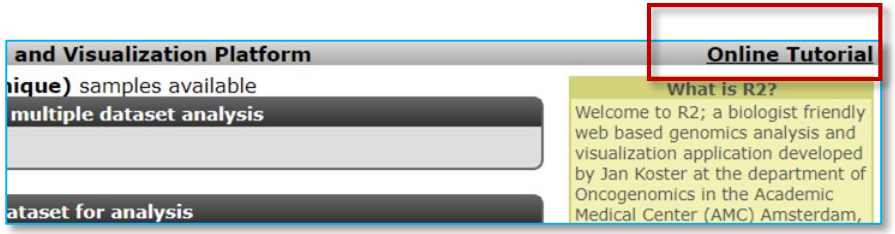


Do your own research with available datasets in the grid

- Find a public dataset of your interest in R2 dataset grid
- Think about a good biological research question
- Try to find the analyses that can help you gain insight

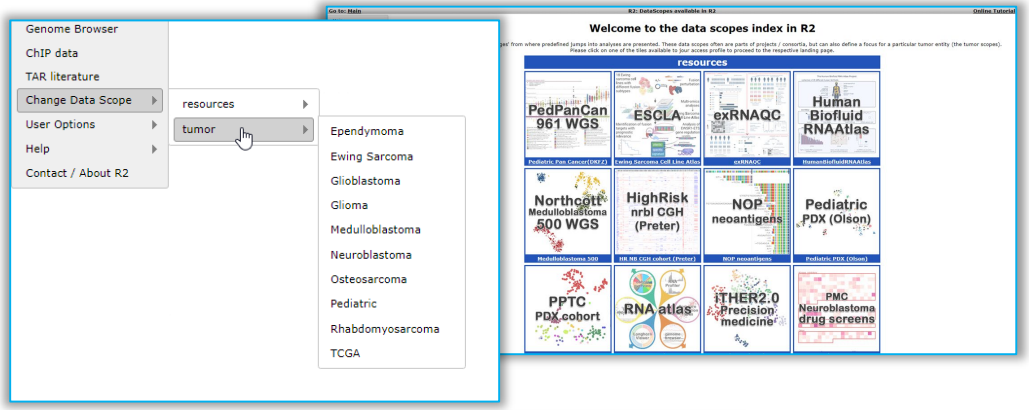
Tutorial

- Go to a chapter of interest and follow the examples
- Toy around with a similar pipeline on a different dataset
- Perform a follow up analysis from the tutorial
- Perform the same analysis on another dataset of your interest



Next:

- Look into other courses in the course material
- Datascopes
  - Look for datascopes that have similar type of data as yours and look at the kind of analyses that were done for inspiration
- **Find publicly available datasets and email us to get them uploaded**
- **Create you own data and email us to upload**



r2-support@amsterdamumc.nl